International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

© Global Society of Scientific Research and Researchers

http://ijcjournal.org/

Clustering Data Text Based on Semantic

Parisa Zandieh[a]*, Elham Shakibapoor[b]

[a,b]*Department of Computer Management, Marvdasht Branch , Islamic Azad University, Marvdasht, Iran*

[a]*Email: zandieh.parisa@yahoo.com*

[b]*Email: elhamshakibapour@miau.ac.ir*

Abstract

Clustering is one of the most important data mining techniques which categorize a large number of unordered text documents into meaningful and coherent clusters. Most of text clustering algorithms do not consider the semantic relationships between words and do not have the ability to recognize and use the semantic concepts.In this paper, a new algorithm has been presented to cluster texts based on meanings of the words. First, a new method has been presented to find semantic relationship between words based on Wordnet ontology then, text data is clustered using the proposed method and hierarchical clustering algorithm. Documents are preprocessed, converted to vector space model, and then are clustered using the proposed algorithm semantically. The experimental results show that the quality and accuracy of the proposed algorithm are more reliable than the existing hierarchical clustering algorithms.

Keywords: Text mining; Text clustering; Hierarchical clustering; Ontology; Semantic relationship.

1. Introduction

Concerning the enhancement of text documents and the increasing growth of text information, organizing large text sets into meaningful and controllable small groups plays important role in information retrieval, review, revision and perception. By the increasing number of digital texts available in the web, text mining techniques need to be take more consideration. Text mining discover and extract the unknown knowledge in a set of texts. The unknown knowledge has been stored in the text via semantic relationships between information.

----

* Corresponding author.

There are three major methods throughout the world to organize such large size of unstructured information as follows:

- Information retrieval
- information extraction
- knowledge discovery in the text

The first main phase in text mining is text pre-processing which aims to represent documents more appropriately. Document-based and concept-based formats are produced in this phase. In Document-based presentation format, the important thing is appropriate presentation of documents. It can be conversion of documents to an intermediate or semi-structured format, or can be using index on it, or can be any other presentations which make working on documents more efficient. In concept-based format, documents' presentation is improved. The concepts and meanings available in documents, the relationship between them and any conceptual information can be extracted from the text. The second phase is to extract knowledge from middle forms of document's representation. In this phase, the concepts and meanings available in the document, their relationships and any conceptual information can be extracted from the text. Clustering is one of the presented techniques in data mining which extracts groups that are very similar to the elements inside the groups and that are different from elements of other groups. In fact, clustering is an unsupervised classification in which classes have not been predetermined and in that sets of data that are usually vectors in multidimensional space, they are divided into a given number of clusters based on a similarity or dissimilarity measure. But any similarity cannot be estimated numerically such as the distance between concepts of two words. If clustering is performed based on concept and meaning of the object or words, clusters are not only a set of objects with numerical similarity but also, they are a group of objects that represent the same concept whit each other [1]. This technique model data via clusters, however, presentation of data with a low number of clusters cause the loss of details but it leads to simplification. High efficiency and accuracy of data clusters are two main and important goals of document clustering. In this paper, first, a new technique is presented to find similarity and semantic relationship between words using TF-IDF matrix and Wordnet ontology. Then, to increase accuracy and quality, a new method is presented based on hierarchical clustering algorithms using Wordnet ontology that categorizes data more accurately and qualitatively than existing hierarchical clustering algorithms. The rest of this paper is organized as follows.in section 2, existing algorithms and suggested methods are reviewed. In section 3, the suggested method is described. Section 4 presents the implementation and analysis of results obtained from algorithm for different parameters and compares it with other algorithms. Conclusion and suggestions is described in section 5**.**

2. Related work

Tar [2]presented a conceptual weight for text clustering system based on K-means algorithm and principles of ontology such that the importance of words of a cluster can be determined by given weighted values. Moh'd Alia[3] presented an optimized Harmony Search algorithm to solve the problem of finding cluster centers in C-means and Fuzzy C means algorithms. The algorithm look for proper cluster centers locally and globally in search space. Uysal[4] used a genetic algorithm in text classification to extract latent semantic features in a text.Wei [5] presented a new method using Wordnet and lexical chain for clustering. In this method, hierarchical structure of ontology is used to recognize similarity measure and to extract semantic features of text documents. Gabrilovich [6]calculated semantic relationships using conceptual analysis based on Wikipedia that was used to improve the calculation of words and text relationships. Witten and Milne [7] developed Wikipedia as external knowledge for text clustering that efficiency was improved using the concept and information for interpretation of texts. Song and his colleagues [8] suggested a genetic algorithm for text mining that it finds related concepts by indexing hidden concepts using similarity ontology.

In clustering text data, it has been tried to extract hidden semantic features from the text using ontologies or combination of the clustering algorithms with optimization algorithms that, text clustering is performed more accurately, qualitatively and rapidly such that data of different clusters have maximum differences with each other and data of a cluster are very similar to each other**.**

3. Suggested method

The main idea of proposed method is addition of semantic relation between the words in the input matrix for text clustering that in this paper, Complete-Linkage clustering algorithm clusters text data using the proposed matrix.

The input of problem is a set of text documents that are clustered based on semantic via suggested algorithm after pre-processing and applying essential changes. The stages of suggested algorithm are as follows:

1- Deleting numbers and punctuations

2- Tokenizing documents and converting them to the smallest semantic unit (token)

3- Deleting stop words

4- Words' stemming

5- Representing documents in form of vector space model, creation of TF-IDF matrix

6- Making semantic TF-IDF matrix

7- Creating distance matrix

In order to cluster the texts, Firstly, they should be preprocessed and also converted to apprehensible format for the system. In this paper, for the preprocessing of texts, first, numbers and punctuation marks are removed from the text. Then, documents are tokenized into meaningful small units. In the next step, the defined stop words in that language are removed from the text. Finally, stemming is performed by removing prefixes and suffixes. The output of this stage is Bag of words (BOW) that is displayed using vector space model. BOW is in form of a document-word matrix which determines the number of words' iterations in addition to the presence or absence of a word in each document.

Since the number of word iterations is not a good criterion for measuring similarity between documents, the words are weighed using TF-IDF as follows:

$$TF \ \times IDF(t_k\,, d_j) = Occ(t_k\,, d_j) \times \log \frac{Nb\_doc}{Nb\_doc(t_k)} \qquad (1)$$

Where Occ(t_k,d_j) is the number of words' iterations t_k in document d_j, Nb_doc is total number of existing documents and Nb_doc (t_k) is the number of documents where the word t_k has been used at least one time. In such method, words available in a small set of documents get higher weights and those available in most documents get lower weights [9]. TF-IDF is used as a weighted factor in information retrieval and text mining, it only shows importance of a word in a document and it doesn’t provide any semantic relationship between words. Two documents may be semantically similar but they don’t have any same words, while there are many words in documents that have semantic relationships, TF-LDF matrix do not be able to recognize such relations. Therefore, the matrix has been changed in this research such that semantic relationships and similarities between words and documents can be extracted. First, semantic similarity of all words in TF-IDF matrix is calculated using wup similarity function which finds semantic relationship between words using wordnet ontology[10]. Then, the score of words which similarity is above the threshold limit is assimilated. Therefore, in entry of a document, two words with semantic similarity and different TF-IDF score get higher TF-IDF score. The matrix is called as Semantic TF-IDF. The basis of all hierarchical algorithms is a distance matrix which usually uses a similarity criterion, such as Cosine or Euclidean similarity, in order to calculate the distance, in proposed method, Semantic TF-IDF matrix is used to obtain the distance between documents. In the matrix, each document has its own row which is considered as a vector. Then, the distance of all documents is calculated using cosine or Euclidean criterion thus a distance matrix is made.

This is a symmetrical square matrix. If there is N documents, the matrix's dimensions is N×N. Each entry of matrix shows the distance between two documents. The main diameter of the matrix is zero. According to the Complete Linkage Algorithm is used in the Proposed Technique, when the distance matrix is created, two documents which have the maximum distance are chosen in the distance matrix and combine as a cluster in this matrix. Then, the distance matrix is updated and the distance between all other documents and the new cluster is calculated. The process is repeated until all clusters are combined with each other and only one cluster is remained. It is predicted that use of the Semantic TF-IDF matrix assists to identify words which are more similar and have stronger semantic relationships, and consequently, in documents clustering based on their words, documents which have contents close to each other are placed in one cluster.

4. Implementation and results

In this section, the proposed algorithm has been implemented on the database 20Newsgroups and the results have been evaluated. MATLAB version R2015b is used to implement algorithm and system specifications are Windows 8.1, 64-bits, Intel core i7, 250/2 GHz, Memory RAM: 8 GB.

A. Dataset

The 20Newsgroups dataset contains 20000 text documents of newsgroups which have been categorized into 20 different groups. This collection is a famous public dataset for text analysis and machine learning techniques such as classification and clustering. In some versions of this dataset, groups that associate with each other are combined and five groups have created.

B. Implementation proposed Algorithm

First, the pre-processing has been performed on the database to prepare data for clustering. After pre-processing, documents are presented as vector space model leading to TF-IDF and Semantic TF-IDF matrices.

Proposed algorithm is executed on 100 documents by using SemanticTF-IDF matrix and Cosine and Euclidean similarity measure. Since the database has been categorized in five categories, the number of clusters has been considered as five in proposed algorithm.

C. Clustering quality evaluation

There are several criteria for clustering quality evaluation and they are divided into two categories:

1. Unsupervised evaluation indices known as internal criteria in scientific texts are responsible for determining the quality of clustering operations according to the information available in the data set.

2. Supervised evaluation criteria known as external criteria evaluate the performance of clustering algorithms using information outside the datasets under study.

In this paper, unsupervised evaluation criterion of Adjust Rand Index is used to evaluate the quality of clustering and it is calculated as follows[11]:

$$Adjusted\ Rand\ Index = \frac{TP - \frac{(TP+FN)(TP+FP)}{TP+FN+FP+TN}}{\frac{(TP+FN)+(TP+FP)}{2} - \frac{(TP+FN)(TP+FP)}{TP+FN+FP+TN}} \qquad (2)$$

TP: number of data pairs that should be placed in a cluster.

TN: number of data pairs that should be placed in separate clusters and they have been included correctly in separate clusters.

FP: number of data pairs that should be placed in a cluster but they have been included in separate clusters

FN: number of data pairs that should be placed in different clusters but they have been included in a cluster.

The possible values of Adjust Rand Index are [-1,+1]. The closer the value to one, the better the quality of clustering.

D. Result

Proposed algorithm has been implemented on 100 documents from the 20 newsgroups dataset in 2 states as follows

- Use of SemanticTF-IDF matrix and Euclidean similarity criterion

- Use of SemanticTF-IDF matrix and cosine similarity criterion.

Then it has been evaluated by Adjust Rand Index.

E. Comparing proposed algorithm to similar algorithms

In this section, the results obtained by implementation of the proposed algorithm in Section III are compared and analyzed with the results of similar existing algorithms. Proposed algorithm clusters text documents with considerable accuracy and quality. As seen in Table 1, Adjusted Rand Index of proposed algorithm with input matrix of semantic TF-IDF and cosine similarity is 0.2124, which is highly desirable.

The results of the proposed algorithm and the results of the implementation of clustering algorithms of Complete-Linkage and Average-Linkage were obtained by the same dataset using TF-IDF matrix then, their clustering qualities were calculated using Adjusted Rand Index(ARI) and they were compared with each other. Furthermore, the results of the proposed algorithm with use of SemanticTF-IDF matrix and cosine and Euclidean criteria were compared to [11] in which a clustering method has been presented on documents of 20Newsgroups dataset based on the Fuzzy C-means algorithm for clustering text data using the TF -IDF Matrix and cosine and Euclidean similarity criteria and the quality of clustering has been determined by Adjusted Rand Index (ARI).

The results are shown in following tables and figure:

Table 1: Comparing hierarchical algorithms, proposed algorithm and algorithms of (11) in terms of ARI

	Proposed Algorithm	**Average Linkage**	**Complete Linkage**	**KFC**	**MKFC**
Euclidean	0.2083	0.0364	0.1202	0.131	0.139
Cosine	0.2124	0.0048	0.1202	0.14	0.143

As seen in above table, proposed algorithm with use of the SemanticTF-IDF matrix as well as the TF-IDF matrix and both cosine and Euclidean similarity criteria, had better results than hierarchical algorithms of Complete –Linkage, Average-Linkage and such algorithm with cosine and Euclidean criteria as well as TF-IDF matrix outperformed KFC and MKFC methods presented in [17].

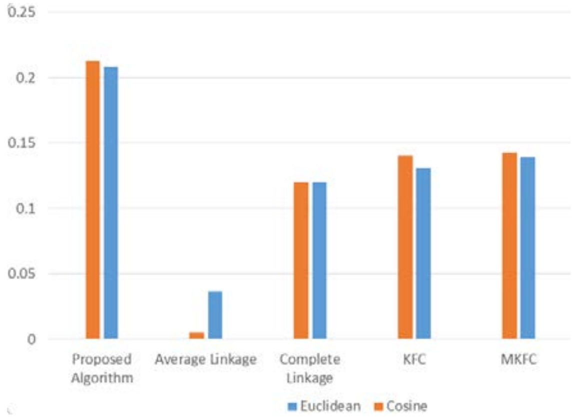

Figure 1: Comparing ARI and Euclidean and cosine criteria in hierarchical algorithms, the proposed algorithm and [17].

5. Conclusions

Proposed algorithm clusters automatically text documents and its clustering quality is better than conventional hierarchical clustering algorithms. Therefore its quality and accuracy of clustering is increased by using proposed Semantic TF-IDF matrix due to taking into consideration the meanings and concepts of words. Since clustering is done based on meaning and concept, problems such as labeling nodes, quality, unpredictability of quality and results, irrationality of clusters in the hierarchy in conventional clustering methods are settled.

6. Recommendations

The proposed algorithm has used one-point cross over operator. Other types of operators can be used such as two-point, multi-point, map - reduce, order, cycle and uniform cross overs and then the proposed algorithm can be implemented on them. Then, results obtained from above cross overs are compared to one-point cross over thus the best cross over is found for the proposed algorithm. The proposed algorithm is the first combination of hierarchical clustering algorithms with heuristic algorithms. It is suggested to combine other heuristic algorithms with hierarchical algorithms such as bee, cuckoo, harmony search and to evaluate the results in improvement of text clustering.

Acknowledgements

The present paper, text data clustering based on meaning, has been derived from M.A thesis of Department of Computer Engineering-Software from Marvdasht Branch, Islamic Azad University, Marvdasht,Iran.

References

[1] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern recognition letters, vol. 31, no. 8, pp. 651-666, 2010.

[2] H. H. Tar and T. T. S. Nyunt, "Ontology-based concept weighting for text documents," world Academy of Science, engineering and Technology, vol. 57, pp. 249-253, 2011.

[3] O. Moh'd Alia, M. A. Al-Betar, R. Mandava, and A. T. Khader, "Data clustering using harmony search algorithm," in International Conference on Swarm, Evolutionary, and Memetic Computing, 2011, pp. 79-88: Springer.

[4] A. K. Uysal and S. Gunal, "Text classification using genetic algorithm oriented latent semantic features," Expert Systems with Applications, vol. 41, no. 13, pp. 5938-5947, 2014.

[5] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains," Expert Systems with Applications, vol. 42, no. 4, pp. 2264-2275, 2015.

[6] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in IJcAI, 2007, vol. 7, pp. 1606-1611.

[7] I. Witten and D. Milne, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," in Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA, 2008, pp. 25-30.

[8] W. Song, C. H. Li, and S. C. Park, "Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures," Expert Systems with Applications, vol. 36, no. 5, pp. 9095-9104, 2009.

[9] W. K. Gad and M. S. Kamel, "Enhancing text clustering performance using semantic similarity," in International Conference on Enterprise Information Systems, 2009, pp. 325-335: Springer.

[10] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 1994, pp. 133-138: Association for Computational Linguistics.

[11] H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen, "Multiple kernel fuzzy clustering," IEEE Transactions on Fuzzy Systems, vol. 20, no. 1, pp. 120-134, 2012.