ISSN 2278-3091

Volume 9, No.1, January – February 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse25912020.pdf

https://doi.org/10.30534/ijatcse/2020/25912020

Understanding User's Behavior by Social Media Data Clustering

Mohamed Shenify[1]

[1]A Department of Computer Science, Albaha University, Albaha, Saudi Arabia, maalshenify@bu.edu.sa

ABSTRACT

Mining Social Media data is an important data mining problem. There are a couple tools used for these purposes and document clustering is one of them. Clustering of documents is the automatic division of text documents into a group of clusters in such a manner that the documents within a cluster have high similarity values, but dissimilar to documents in the other clusters. Document clustering is quite popular among the database community and NLP people and it has been studied by a host of researches of diverse fields. It has wide range of application areas like search engines, web mining, information retrieval and topological analysis. Recently, it is used in analysis of Social Media data to understand the Social Human behavior of users. In this article we propose a clustering algorithm for categorizing the users and hence understanding their behaviour. Our proposed algorithm is a hierarchical agglomerative one which uses fuzzy concept, where each cluster is viewed as a fuzzy set over some finite universal set. A Canberra metric based similarity measure is used to measure inter-cluster similarity. The efficiency of the algorithm has been established with the help of complexity analysis

Key words: Document Clustering, Semi-structured Data, Unstructured Data, Fuzzy Set, Graph, Sim function, Agglomerative Hierarchical Algorithm.

1. INTRODUCTION

Clustering [1] method is defined as the unsupervised classification of patterns into clusters. The objective of clustering method is to make a set of clusters of data from a given dataset such that the data points within a cluster are more similar to each other than to those outside the cluster. Due to widespread use of Social Media and other type of electronic publications such as emails and the WWW text databases are rapidly expanding and most of the information are available in soft form. These data are semi-structured or unstructured. Mining such data can be a challenging work. Till now most of the data mining tasks were focused on structured data sets. However recently semi-structured or unstructured data mining such as mining social media data is getting importance to the researchers and a couple of people have started working on this. Data published in social media like facebook, twitter and instagram are semi-structured or unstructured and information extracted from such data will help us understand the mood of the users e.g. how users react on a particular issue of importance or what is the opinion of users on a particular issue. Document clustering can be used as one of the tools for these purposes. Given a collection of unlabelled documents, document clustering can help in organizing the collection based on some criterion, thereby classifying the users based on their writing or post. Traditionally document clustering was used in many information retrieval works such as browsing of document, organization and viewing of retrieval results, generation of hierarchies of documents in search engines etc. But nowadays it is also used in other purpose like Social Media Data analysis. In [2], authors have done a comparative study on document clustering techniques. Social Media Data mining has been studied by different researches in detail [3, 4, 5, 6].

As the data accumulating in the different Social Platform are found to be semi structured or unstructured, most of the document clustering methods perform several preprocessing steps (e.g. stop word removal and stemming on the document set). Each document is represented by a vector of frequencies of the remaining terms within that document. By clustering one can identify dense and sparse regions and therefore, discover distribution patterns and interesting correlations among data attributes.

Fuzzy introduced by Zadeh [7] used successfully in several field of human knowledge. Accordingly it has been used in clustering of text data [8]. The similarity measure used for the clustering is defined in [8]. The clustering algorithm discussed [8] is an agglomerative hierarchical algorithm already available in the literature.

In this paper, a new document clustering algorithm is proposed to identify the patterns among Social Media data which uses the concept of fuzzy sets. The proposed algorithm is an agglomerative hierarchical algorithm and at any given stage of the algorithm there are smaller clusters and the decision at the current stage is to merge the incoming document with the cluster that satisfies a user defined threshold. The clusters obtained are represented as fuzzy sets over a finite universal set. For this purpose we have used a similarity measure derived from Canberra metric. The algorithm begins by considering each input data point as a cluster, compares it with the existing clusters at that stage of the algorithm and is merged with the cluster that satisfies a user defined threshold.

The paper is organized as follows. In section 2, we discuss some recent and similar works. Definitions and notations used in this paper are given in section 3. In section 4, the proposed algorithm is discussed. In section 5, the complexity analysis of the algorithm is discussed.

2. RELATED WORK

Social Networking has become very important part of human life. Nowadays lots of people have started using different social media platform like facebook, Twitter, Instagram etc. So, the amount of data accumulating in different social media platform is expanding rapidly. The problem is to utilize such huge amount of data for understanding user's behaviour. This is done by different data mining methods. In [3], authors have made a detailed study on various text mining methods used for finding patterns from the social Web. In [4], the authors have described a method of analysis of Twitter data based on different models, which are U-T model and U-T-C models, as heterogeneous information network. In [5], the authors have reviewed different data mining methods in social media. In [6], the authors have used clustering techniques on social media contents. In [9], the authors have used an optimized K-means clustering method for the analysis of social media. In [10], the authors have presented a nice method based on Fine-Grained document clustering via Ranking approach which leverages the search engine capability of handling big data efficiently. Authors in [23] and [.24] use social network data to analyze productivity and spam video, respectively.

During the last few years the concept of fuzzy sets has been used in different areas including clustering or pattern recognition ([11], [12], [13], [14], [24]). Conventional clustering techniques assume that an object or data point can belong to one and only one cluster. However there may be overlapping of clusters and thus the separation of clusters is a fuzzy notion and hence the concept of fuzzy sets has come into picture. In fuzzy clustering each data point is associated with each cluster using a membership value. Larger membership values indicate higher confidence in the assignment of the object to the cluster. So in this approach each cluster is a fuzzy set of all data points.

In the paper [15] the authors proposed an approach of fuzzy clustering of web documents. The documents are represented as vectors of variable lengths. Each element of the vector is a pair of key phrase and an importance weight associated with this key phrase in a particular document. Using this representation of documents, fuzzy clustering algorithm was applied. In [16] the authors proposed a fuzzy set approach for clustering large categorical data. There are so many similarity measures used for clustering. Canberra metric [17, 18, 19, 20] based similarity is one of them. It has been for many purposes namely anomaly detection [21] in network data. In this paper, we have used it for document clustering in social media data.

3. DEFINITIONS AND NOTATIONS USED

In below, we review some definitions and notations used in the algorithm.

3.1 Basic Definitions related to Fuzzy Set

Let us assume that A_i=(x, A_i(x)); i=1, 2,…m are fuzzy sets where $A_i(x)\in[0, 1]$ is the membership functions of A_i. We denote the universe of discourse of each A_i as $X_i=\{x_1, x_2,\ldots x_m\}$. When fuzzy sets are represented by their membership functions, then

$$A_i = \{A_i(x),\ a_{1i} \le x \le a_{2i}\},\ \forall i=1, 2,\ldots.m.$$

(1)

Measures a_{1i} and a_{2i} are such that $a_{1i} \le a_{2i}$, $A_i(x)=0\ \forall\ x< a_{1i} \wedge x> a_{2i}$.

3.2 Fuzzy Set Representation of Clusters

Social media contents are basically post of the users. These are the documents; we are going to use these as input for mining process. Such documents reflect the opinions, ideas, and views of the different users. Clustering such documents would help us to understand mood of a group of people. As the social media content are semi-structured or unstructured, some preprocessing methods must be applied.

Each document d is a finite list containing elements of the form (w, n), where w=keyword and n=number of occurrence of w in the document. Let W be set of distinct keywords appearing in the documents and $|W|=m$. Also $W=\{w_1, w_2, w_3,\ldots..w_m\}$ appears in some sequence. Any document can be represented by $(n_1, n_2, n_3,\ldots..n_m)$, where n_i =number of occurrence of w_i in d. If w_i is absent in d then n_i=0, for d. In this situation, a fuzzy set is defined over W. The fuzzy set representation of the cluster C having one document say d is $(n_1, n_2, n_3,\ldots..n_m)$ and is calculated as follows. Let A_C be the fuzzy set representing cluster C with associated membership function A_C:$W\rightarrow[0,1]$ and is defined as (2).

$$A_C(w_i) = \frac{n_i}{\sum_{i=1}^{m} n_i} \qquad (2)$$

Obviously $0\le A_C(w_i)\le 1$ for each i. The fuzzy set A_C represented by its membership function with $n_{num} = \sum_{i=1}^{m} n_i$ is the compact representation of the cluster C i.e. (A_C, n_{sum}).

3.3 Similarity Measure on Documents

For clustering the documents, we have used a measure called *Canberra metric* [17, 18, 19, 20] and is defined as follows.

Let d_1 and d_2 are two documents represented by $(n_1, n_2,..,n_m)$ and $d_2 = (p_1, p_2,..,p_m)$ respectively. Obviously d_1 and d_2 are two m-dimensional vectors. Then the *Canberra metric* [17, 18, 19, 20], $d(d_1, d_2)$ is given by the formula (3).

$$d(d_1,d_2)=\sum_{i=1}^{m}\frac{|n_i-p_i|}{|n_i|+|p_i|} \tag{3}$$

Here the range of equation (3) is [0, m]. For making it, [0, 1], it is to be divided by m. Therefore, a new formula is obtained, which is our similarity measure S on the documents and is expressed as in (4).

$$S=\frac{1}{m}\sum_{i=1}^{m}\frac{|n_i-p_i|}{|n_i|+|p_i|} \tag{4}$$

3.4 Similarity Measure on Clusters

For inter-cluster similarity, we replace the documents with fuzzy sets, which represent the corresponding clusters. Let C_1 and C_2 two clusters represented by the membership functions $A_{C1}(w_i)$ and $A_{C2}(w_i)$ respectively. The membership functions are computing using the formula given by equation (2). Then the similarity between C_1 and C_2 is denoted by $S(C_1, C_2)$ and is given by the formula in (5).

$$S(C_1,C_2)=\frac{1}{m}\sum_{i=1}^{m}\frac{|A_{C_1}(w_i)-A_{C_2}(w_i)|}{|A_{C_1}(w_i)|+|A_{C_2}(w_i)|} \tag{5}$$

3.5 Merging of Clusters

Let C_1 and C_2 be the two clusters having n_{sum1} and n_{sum2} terms respectively. Let (A_{C1}, n_{sum1}) and (A_{C2}, n_{sum2}) are compact representation of C_1 and C_2 respectively. Let C be the cluster formed by merging C_1 and C_2 with fuzzy representation of C as A_C. Then the membership function of A_C is given by

$$A_C(w_j)=\frac{(n_{sum1}A_{C1}(w_j)+n_{sum2}A_{C2}(w_j))}{(n_{sum1}+n_{sum2})} \tag{6}$$

4. THE PROPOSED ALGORITHM

At the beginning of the clustering process, each document is allocated to a separate cluster as a compact representation. Thereafter for every pair of clusters the *Canberra metric* based similarity is computed and then the *merge* function is used to obtain larger clusters if and only the *similarity* value is found to be within certain limit (the definition of Canberra metric and *merge* function is given in section 3). Accordingly fuzzy membership function of the new cluster is computed using the formula given in section 3. At any level, for any two clusters say C_1 and C_2, the corresponding *Canberra metric* is calculated using the formula given in section-3, to check whether they can be *merged* or not. If the value is found to be within a certain pre-determined threshold then C_1 and C_2 are *merged* using *merge* function to form a new bigger cluster with a new membership function. The process of *merging* of clusters continuous till no *merger* is possible or there is only one cluster at the top. In bellow we present the pseudo code for the proposed algorithm.

```
Algorithm DataInstanceClustering(n,$\sigma$)
Input: The number of documents d[i]; i=1,2,…n, and
       threshold $\sigma$
Output: A set of cluster S
Step1. The set of clusters S, where each cluster C of S
       having one document d[i]
Step2. If for any cluster $C_1 \in S$ and $S(C_1,C) \leq \sigma$, then
       merge($C_1$,C) to form a new cluster $C_2$ consisting of $C_1$ and C.
Step3. Remove $C_1$ from S.
Step4. Continue Step2 and Step3 till no merger of clusters
       is possible.
Step5. Return S
Step6. Stop
```

The algorithm supplies the set of clusters *S* of documents.

5. COMPLEXITY OF THE ALGORITHM

Let *n* be the total size of input data sets. The complexity of computing cluster of size-1 is O(*m*), where *m*=dimension of the feature vectors to which the input data are converted. The overall complexity is O(*mn*). The complexity of computing the similarity value of cluster pairs is O(*m*). Maximum such computations n^2 such computations are required. Thus the total complexity is O(mn^2). Again the complexity of merging cluster is O(*m*). The maximum number of execution is *n*-1. Therefore the overall complexity is O(mn^2+m(n-1)). i.e. O(mn^2).

6. CONCLUSION

In this paper, an algorithm for clustering Social Media Data is discussed. The algorithm is agglomerative hierarchical algorithm. As the Social Media Data are semi-structured or unstructured, the algorithm cannot be used directly so the data preprocessing is required to remove noises and undesirable contents from the datasets. The clusters are compactly represented as fuzzy sets. At each stage two clusters are merged if their similarity value is within certain specified limit. The similarity used for merging is *Canberra metric* based similarity. The execution of algorithm will stop if there is no merging possible or a particular level becomes empty. The efficacy of the algorithm is established by complexity analysis. The patterns extracted by the algorithm can be used for several purposes e.g. understanding social behavior of users, understanding mood of a group of users, classifying the users based on their posts or reactions on social media platform etc.

In the future attempts will be made to apply approaches other than agglomerative hierarchical. Attempt will also be made to apply some other tools like association rules mining, sequential patterns mining, etc.

REFERENCES

1. A. K. Jain, M. N. Murty, and P.J. Flynn. **Data clustering: A review**, *ACM Computing Surveys*, 31(3): 264-323, 1999. https://doi.org/10.1145/331499.331504
2. M. Steinbach, G. Karypis, and V. Kumar, **A comparison of document clustering techniques**, *in Proc. KDD-2000 Workshop on TextMining, Aug. 2000.*
3. Rizwana Irfan, Christine K. King, Daniel Grag Es, Samewen, Samee U. Khan, Sajjad A. Madani, Joanna Kolodziej, Lizhewang, Danchen, Ammar Rayes, Nikolaostziritas, Cheng-Zhong Xu, Alberty. Zomaya, Ahmed Saeed Alzahrani, and Hongxiang Li, **A Survey on Text Mining in Social Networks**,*The Knowledge Engineering Review*, Vol. 00:0, 1–24. 2004, Cambridge University Press.
4. N. Prangnawarat, I. Hulpus, C. Hayes, **Event Analysis in Social Media Using Clustering of Heterogeneous Information Networks**, *in Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference*-2015, 294-298.
5. G. Barbier, and H. Liu, **Data Mining in Social Media**, *Book chapter, Social Network Data Analytics, Springer Science Business Media, LLC 2011*, 327-351. https://doi.org/10.1007/978-1-4419-8462-3_12
6. I A Rytsarev, A V Kupriyanov, D V Kirsh and K S Liseckiy, **Clustering of social media content with the use of BigData Technology**, *The IV International Conference on Information Technology and Nanotechnology, IOP Conf. Series: Journal of Physics: Conf. Series* 1096 (2018), 1-7.
7. L. A. Zadeh, **Fuzzy Sets as Basis of Theory of Possibility**, *Fuzzy Sets and Systems I*, 1965, 3-28.
8. K. Thaoroijam, and A. K. Mahanta, **A Fuzzy based Document Clustering Algorithm**, *International Journal of Computer Applications (0975 – 8887)* Volume 151 – No.10, October 2016. https://doi.org/10.5120/ijca2016911923
9. A. Alsayat, H. El-Sayed, **Social media analysis using optimized K-Means clustering**, *2016 IEEE 14th International Conference on Software Engineering Research, Management.*
10. T. Sutanto, R. Nayak, **Fine-grained document clustering via ranking and its application to social media analytics**, *Social Network Analysis and Mining 1*, 2018. https://doi.org/10.1007/s13278-018-0508-z
11. J. C. Bezdek, ***Pattern Recognition with Fuzzy Objective Function Algorithm***, Plenum Press, New York.
12. R. N. Dave, **Generalized fuzzy C-shells clustering and detection of circular and elliptic boundaries**, *Pattern Recognition*, 25,713-722.
13. S. K. Pal, **Fuzzy tools for the management of uncertainty in pattern recognition, image analysis, vision and expert systems**, *in Proc. of International J. System Sc*, Vol 22,No 3, pp 511-549, 1991.
14. W. Pedrycz, **Fuzzy sets in pattern recognition: Methodology and methods**, *Pattern Recognition*, Vol 23, No ½, pp121-146, 1990. https://doi.org/10.1016/0031-3203(90)90054-O
15. M. Friedman, M. Last, O. Zaafrany, M. Schneider, and A. Kandel, **A New Approach for Fuzzy Clustering of Web Documents**, *in Fuzzy Systems, Proceedings. 2004 IEEE International Conference*, Vol 1, 377- 381, July 2004.
16. M. Dutta and A. K. Mahanta, **An algorithm for clustering large categorical databases using a Fuzzy set based approach**, *in Proceedings of NWTAC (National Workshop on Trends in Advanced Computing)* 2006, Tezpur University.
17. G. N. Lance, and W. T. Williams, **Computer programs for hierarchical polythetic classification ("similarity analysis")**, *The Computer Journal*. 9 (1), (1966) 60–64. https://doi.org/10.1093/comjnl/9.1.60
18. G. N. Lance, and W. T. Williams. **Mixed-data classificatory programs I Agglomerative Systems**, *Australian Computer Journal*, (1967) 15–20.
19. T. H. Clifford, and W. Stephenson, ***An Introduction to Numerical Classification***, Academic Press. New York-San Fransisco – London,1975.
20. S. M. Emran, and N. Ye, **Robustness of Canberra Metric in Computer Intrusion Detection**, *in Proceedings of 2001 IEEE Workshop on Information Assurance and Security, US Military Academy, NY (June 2001)* 80-84.
21. F. A. Mazarbhuiya, and M. Y. AlZahrani, L. Georgieva; **Anomaly Detection Using Agglomerative Hierarchical Clustering Algorithm**, *ICISA 2018, Lecture Notes on Electrical Engineering (LNEE)*, Volume 514, Springer, Hong Kong, pp 475-484.
22. J.R. E. Fernandez and G. M. Pascua. **What features emerge when asian countries are matched across productivity and social media use? : A basis for complex adaptive analysis**, *IJATCSE Journal*, 2019;8(6): 2998 – 2992. https://doi.org/10.30534/ijatcse/2019/51862019
23. F. El Mendili And Y..El Bouzekri El Idrissi, **Detection of video spam in social network based neural network convolution**, *IJATCSE Journal*, 2019;8(4): 1372 – 1381. https://doi.org/10.30534/ijatcse/2019/53842019
24. S. Handoyo, A. Widodo, W.H. Nugroho and I. N. Purwanto. **The implementation of a hybrid fuzzy clustering on the public health facility data**, *IJATCSE Journal*, 2019;8(4): 3549 – 3554. https://doi.org/10.30534/ijatcse/2019/135862019