EAI Endorsed Transactions
on Scalable Information Systems Research Article **EAI.EU**

Unsupervised Machine Learning based Documents Clustering in Urdu

Atta Ur Rahman[1,*], Khairullah Khan[1], Wahab Khan[2], Aurangzeb Khan[1] and Bibi Saqia[1]

[1]Department of Computer Science, University of Science & Technology Bannu, Pakistan
[2]Department of Computer Science & Software Engineering, IIU, Islamabad 44000, Pakistan

Abstract

The volume of data on the web is growing rapidly, due to the proliferation of news sources, contents, blogs and journals etc. Like other languages, the Urdu language has also observed tremendous growth on the internet. As the volume of data is expanding, information retrieval (IR) is becoming complicated. Document clustering is an unsupervised ML approach, employed to group a huge number of dispersed documents into a small number of significant and consistent clusters, thus providing a base for indexing, IR and browsing mechanisms. Documents clustering has a long tradition in English as well as English like western languages, but Urdu lags behind in terms sophisticated natural language processing (NLP) tools and resources for documents clustering. Documents clustering becomes a challenging task in Urdu language having a rich morphology, particular structure, syntax peculiarities and cursive nature. In this study, we have developed a framework of document clustering and analysed various similarity measures for Urdu documents. We have also checked the effect of stop words removal in the process of Urdu document clustering.

Keywords: Urdu; Documents clustering; Similarity Measures; K-Means Algorithm

Received on 30 September 2018, accepted on 04 December 2018, published on DD MM YYYY

doi: 10.4108/_______________

1. Introduction

The extent of data on the cyberspace is expanding quickly due to the large-scale and rapid expansion of web technologies [1-4]. These databases are continuously upgraded for growth of documents and possess a high query stack. This unstructured information has asked a few new examinations to investigate this gigantic information, sort related data and to subsequently enhance the association of the content existing on the web [5]. Nearly every information one requests are currently accessible on the internet [6]. English and European languages have mainly dominated the web since its beginning [7]. However, in the past few years, a widespread range of information in the Indian local languages such as Urdu, Hindi, Bengali, Oriya, Tamil, and Telugu have been observed on the internet [6, 8]. The richness of data along with the vibrant and diverse nature of the Web makes Information retrieval (IR) a challenging task [9, 10]. Document clustering presents a structure for categorizing a large collection of documents [11, 12]. Document clustering is exploited to consequently find the intrinsic characteristics and native grouping amongst documents, to sort out them into various clusters [13, 14]. Documents clustering is an exciting approach, since it groups the documents exclusive of human intercession and exempts organizations from the prerequisite of manual categorization of documents, which might be an arduous and tedious process [15]. Various studies regarding document clustering, exploiting English language documents as input have been presented [16]. However, each language can generate distinct levels of exactness, depending on each natural language shapes and characteristics, like morphological and syntax peculiarities, use of antonyms and synonyms, and utilization of native expressions etc [17, 18].
Structure of this paper is organized as: section 2 highlights the importance and challenges of Urdu, section 3 describes

*Corresponding author. attacs9@gmail.com

un-supervised learning approach. In section 4 related work is presented, Section 5 describes proposed work, section 6 tells about the adopted unsupervised clustering algorithm while section is about the experimentation work and section 8 provides conclusion.

2. Urdu Language

Urdu is a lingua franca and national language of Pakistan [25, 26]. As per Wikipedia statistics there exist 100 million native speakers of Urdu in Pakistan and India and an additional 300 million speakers around the globe [25]. The development of computational sources is the elementary step in any Natural Language Processing (NLP) task. Urdu is broadly communicated languages of Asian sub-continent, though due to sources scarceness, not sufficient effort has been accomplished aimed at Urdu language processing [7]. The "daily Jang" stayed the leading newspaper which generates the Urdu scripts digitally in the Nastaliq script design. Currently, many Urdu journals and magazines are issued in Pakistan on daily bases. There exists a bulk quantity of tweets in the shape of Geo News, Jang News, Roznama Dunya, Dawn News, BBC, ARY, AJJ, and Abb Takk News etc. Moreover, India also distributed more than 3,000 Urdu publications on regular basis.

2.1 Challenges in Urdu Document Clustering

The great number of issues identified in Urdu language causes document clustering a difficult task. This section describes a few major constraints which diminish the execution of proposed structure.

- Resource Scarceness

A large number of complexities related with Urdu content makes it a rare dialect to be studied for NLP. A benchmark and an extensive corpus is the essential prerequisite for any NLP associated task. However there is no standard corpus accessible for Urdu language processing. The accessible Urdu NE labeled corpora are: Backer-Riaz (2002) and Emile (2003) corpus. Currently, there is not any dataset accessible for Urdu documents clustering.

- Context sensitive and Cursive Nature

In Urdu, the state of a character isn't just influenced by its position but additionally by its adjacent characters. Urdu characters have distinctive shapes at beginning, middle and end of the word. For example in the word (Love, محبت), the state of (te,ت) is changed in the word (gift,تحفہ). Thus the character "ل" has a diverse shape in the words (slave,غلام), (Electricity, بجلی), (long, طویل), and (but, لیکن).

- Words segmentation problem

Segmentation is far problematic in Urdu dialect in light of the fact that here space is utilized for word limit. Space enclosure and exception are caused by utilization of space in Urdu content. For example space inclusion happens, such as, "خوب صورت" (hobsorat, beautiful) is a single word however because of space insertion the framework will assume it as two words like "خوب" and "صورت". Space omission issues happen, for example, "اس لیے" (aslye, therefore) is two words but because of space exclusion, the framework will consider it"اسلیے" like a single word.

- Compound Named issues

A compound named are made out of various words like (ولادی میر پوتنVladimir putin ,عمران احمد نیازی, Imran Ahmad Niazi), here both words refer to a single word but the system will consider each one as a three separate words. Such as "پوتن", "میر", "ولادی", and "نیازی", "احمد", "عمران".

- Large number of Synonyms

Urdu language possess a large number of synonyms like (جنت, heaven) has synonyms such as (فردوس، باغ ، بہشت) which create a great problem in documents clustering.

- Conjunction issues

Some entities are formed by utilizing conjunction word such as (پاکستان اور چین, Pakistan and China) and (علم و دانش, Knowledge and wisdom) etc.

- Acronym ambiguities

In English dialect acronym can be easily distinguished because of the upper casing principle, however in Urdu, it is very hard to perceive acronym, for example, (بی بی سی , BBC, سی پیک , CPEC ,یو این او, UNO) and so on.

The rest of the paper is structured as follows: Section 2 describe an extensive detailed of related work, Section 3 clarifies the proposed architecture employed for Urdu documents clustering, and Section 4 shows experimental analysis, result and evaluation metric while Section 5 concludes the paper.

3. Unsupervised Machine Learning

Supervised learning algorithm is typically used in subjective classification. This algorithm depends on manually labelled datasets and domain dependent. For that reason supervised algorithm is time consuming, required manual expertise and relatively difficult to understand a words of the human discourse. The few familiar examples of supervised learning algorithm are support vector machine (SVM), K-nearest neighbour (KNN) and Naive Bayesian classifier etc. While unsupervised algorithm working regardless of training data sets and its domain independent.

deal with distinct data, which returns the adjacent points as the cluster centroids [11]. The commonly used clustering algorithms depend on partition also include PAM define by [13], CLARA describe by [14], and CLARANS introduce by [15]. In density-based clustering, the core objective of the clustering algorithm is, the document which is in the section with a prominent density of the document space is counted to fit in the similar cluster [16]. A density-based k-means algorithm is suggested to improve the performance of DBSCAN and K-means algorithms. They utilized a dataset of 250 documents and observed that DBK-means has outperforms the k-means and DBSCAN algorithms [17].

Clustering algorithm founded on density and distance is also utilized, which calculates the distance and the density of every data points and combined those data objects which have minimum distance and highest density, using a decision graph [18]. The COBWEB expresses by [19] and GMM outline by [20] depend on statistical learning and neural network. In Kernel-based clustering algorithms, the input space is converted into a feature of high dimension. The classic algorithms of this type of clustering are kernel K-means explains by [21], kernel FCM mark by [22], kernel SOM specify by [23], and SVC characterize by [24]. A clustering algorithm known as affinity propagation (AP) is offered in 2007, which relies upon "message passing" amongst information objects. In this kind of algorithm, the client can't assign the quantity of groups as an input, such as a k-means algorithm. However, like a k-medoids, it can locate "exemplars", fellows of the input set that are illustrative of clusters [25]. Various strategies have been acquired to achieve semantic correlations amongst documents [26]. A famous tool such as WordNet has been utilized to improve the semantic association amongst words, such as synonyms etc [27]. Additional ontology made research are also incorporated [28, 29], which focuses on words semantic relationship. Chinese news-based clustering approach is proposed by utilizing a Neural network language model [30]. K-nearest neighbour, k-means and support vector machine are employed for Marathi news clustering [31]. Agglomerative hierarchical clustering is proposed for Urdu ligature recognition and they also utilized Naïve Bayes, decision tree, K-nearest neighbour and linear discernment analysis for classification [32]. A detailed study on Urdu document images has been conducted by utilizing various clustering algorithms such as Self organizing map, K-means and hierarchical clustering [33]. Urdu ligatures organization is accomplished using a deep neural network. They exploit a corpus of 2430 ligatures and achieved an accuracy of 73.13 % [34]. Table 1 shows the most related work about Urdu document clustering.

Table 1 Summaries of Related Work

Study	**Application**	**Clustering Algorithm**	**Dataset**	**Result**	**Language**
(Ghwanmeh, 2007)	Information Retrieval System	Hierarchical K-means (HKM), Traditional IR	242 Arabic documents	Precision of HKM for 2 cluster 0.62 and for 5 clusters 0.59, Traditional IR 0.49	Arabic
(Mumtaz & Duraiswamy, 2010)	Document Clustering	DBSCAN, K-means, and DBK-means	250 documents	Rand Index Such as DBSCAN 0.37, K-means 0.60, and DBK-means 0.73	English
(Kumar, Santosh, & Varma, 2011)	Multilingual Document Clustering	Bisecting K-means	FIRE Dataset	F measure 0.57 and Purity 0.68	English and Hindi
(Alkoffash, 2012)	Arabic text clustering	k-means and k-mediods	242 documents	Precision of K-means 0.56 and K-mediods 0.69	Arabic
(Afonso & Duque, 2014)	Clustering of Newspaper and Scientific Text	K-means, sIB and EM	Scientific and Newspaper Corpus of 36 documents	sIB correctness, Scientific 77.8 % and Newspaper 68.9 % , EM 53 %	Brazilian Portuguese
(Fan, Chen, Zha, & Yang, 2016)	A Chinese news based Clustering Approach	Neural network language model	Size of data 600 MB	F measure 0.93	Chinese
(Dangre, Bodke, Date, Rungta, & Pathak, 2016)	System for Marathi News Clustering	K-means, KNN, and SVM	Marathi Text	Not Available	Marathi
(Khan, Adnan, & Basar, 2017)	Multi-level Agglomerative Clustering for	Decision Tree, LDA, Naïve	A corpus of 2430 ligature	Accuracy of Decision Tree 62, LDA 61,	Urdu

Table 3 Example of Urdu Stop words and Key words

Urdu Documents	Key words	Stop words
جدید تحقیق کے مطابق وٹامن ڈی کی گولیاں ہڈیوں کی صحت بہتر نہیں کرتیں اور نہ ہی ان سے ہڈی ٹوٹنے سے بچاؤ میں مدد ملتی ہے۔	جدید، تحقیق، وٹامن ڈی، گولیاں، ہڈیوں، صحت، ، ہڈی، ٹوٹنے، بچاؤ، مدد	کے، مطابق، کی، بہتر ، نہیں، کرتیں، اور، نہ، ہی، اان، سے، میں، ملتی، ہے
ایک نئی تحقیق میں کہا گیا ہے کہ دوسری جنگ عظیم میں اتحادی افواج کی جانب سے جو بم برسائے گئے وہ اتنے طاقتور تھے کہ ان سے کرۂ ہوائی کو نقصان پہنچا ہے۔	تحقیق، دوسری جنگ عظیم، اتحادی افواج، بم، برسائے، طاقتور، کرۂ ہوائی، نقصان، پہنچا	ایک ، نئی، میں، کہا، گیا، ہے، کہ، کی، جانب، سے، جو، گئے، وہ، اتنے، تھے، ان، کو،
برطانیہ میں قائم لیورپول یونیورسٹی کے پروفیسر فرانسس مکلیون نے بتایا کہ 'مچھر یا دیگر کیڑے مکوڑے اور پودے انسانی جلد پر ٹاکسن یعنی حیوانی یا نباتاتی زہر چھوڑتے ہیں۔ جسکی وجہ سے اعصاب کی جانب سے دماغ کو خارش کا سگنل ملتا ہے اور ہم خارش کرنے لگتے ہیں۔	برطانیہ، قائم، لیورپول یونیورسٹی، پروفیسر، فرانسس مکلیون، مچھر، کیڑے مکوڑے، پودے، انسانی، جلد، ٹاکسن، حیوانی، نباتاتی، زہر، چھوڑتے، اعصاب، دماغ، خارش، سگنل	میں، کے، نے، بتایا، کہ، یا، دیگر، اور، پر، یعنی، ہیں، جسکی، وجہ، سے، کی، جانب، کو، کا، ملتا، ہے، ہم، کرنے، لگتے

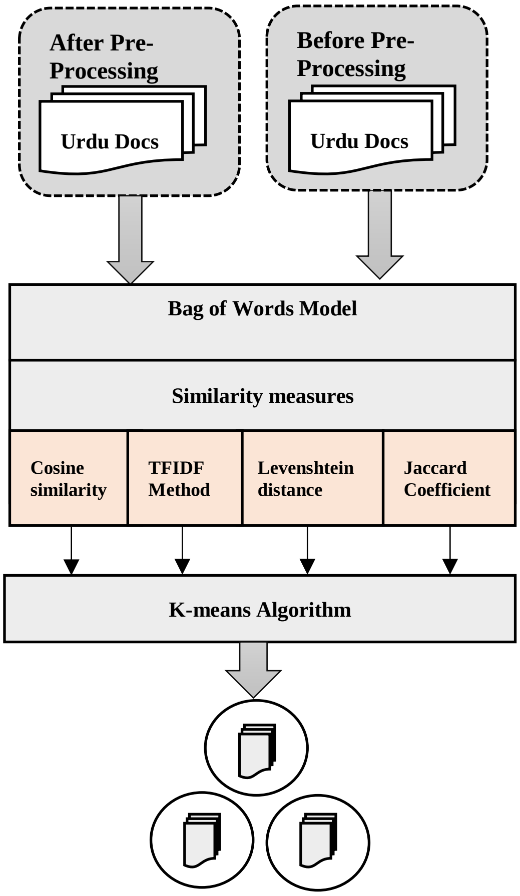

Figure 2 Proposed Architecture

5.3 Bag-of-Words Model

We have represented each document as a Bag-of-words model in this study. This technique is employed for name entity extraction, opinion targets and documents clustering. The Bag-of-words model make good use of Term Frequency Inverse Document Frequency for document clustering and describes their frequencies without any contextual as well syntactic association of words in documents. The Bag-of words model is well known and frequently used method for object classification, information retrieval (IR), similarity measuring and natural language processing (NLP).This model extract a document as pack of its words ignoring word sequence. [62].

5.4 Similarity Measures

Document clustering needs a correct description of the proximity amongst a set of documents, concerning of both, the pairwise likeness or space [63]. The similarity measure is utilized to implicitly capture the alikeness amongst

Table 5 Cosine Similarity

Terms	ملاقات	بھی	سے	وزیراعلیٰ	اور	گورنر	گے	کریں	دورہ	کا	خیبرپختونخوا	آج	عمران خان	وزیراعظم	Cosine similarity = 0.57
doc 1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	
doc 2	1	1	1	1	1	1	1	1	0	0	1	0	1	1	
doc 1.doc 2 = 5				doc 1_SQ 7.0 = 2.64			doc 2_SQ 11 = 3.31				Cosine = doc 1.doc 2/ doc1_SQ x doc 2_SQ				

Table 6 TF-IDF

Terms	ملاقات	بھی	سے	وزیراعلیٰ	اور	گورنر	گے	کریں	دورہ	کا	خیبرپختونخوا	آج	عمران خان	وزیراعظم
doc 1	0	0	0	0	0	0	1	1	1	1	1	1	1	1
doc 2	1	1	1	1	1	1	1	1	0	0	1	0	1	1
Idf	0.47	0.47	0.47	0.47	0.47	0.47	0.30	0.30	0.47	0.47	0.30	0.47	0.30	0.30
Tf-Idf doc 1	0	0	0	0	0	0	0.30	0.30	0.47	0.47	0.30	0.47	0.30	0.30
Tf-Idf doc 2	0.47	0.47	0.47	0.47	0.47	0.47	0.30	0.30	0	0	0.30	0	0.30	0.30

Table 7 Representation of String in Levenshtein distance

Source String: درحواست	String representation						
	ت	س	ا	و	ح	ر	د
Target String: درست	ت	س				ر	د

By applying Levenshtein distance we will performed three deletion operations such as (delt 1, delt 2 and delt 3 on highlighted characters) to convert source string into target string.

Table 8 Applying Deletion Operation of Strings

ت	س	ا	و	ح	ر	د
ت	س				ر	د

After applying Levenshtein distance we will obtained the strings such as

Table 9 Levenshtein Distance

Source String = Target String: درست	ت	س	ر	د

similarity measures in each cluster, utilizing a K-means algorithm are shown in table 11.

Table 11 Result of Similarity Measures before Pre-Processing

Number of Clusters	Similarity Measures			
	Cosine Similarity	TF-IDF	Levenshtein distance	Jaccard Coefficient
Cluster 1	0.45	1	1	0.05
Cluster 2	0.60	0.20	0.35	0.95
Cluster 3	0.40	0.65	1	0.75
Cluster 4	0.85	0.25	0.95	0.80
Cluster 5	1	0.1	0.05	1
Average	0.66	0.44	0.67	0.71

Table 11 described the result of five clusters by different techniques of similarity measures before pre-processing. The average accuracy of Cosine Similarity, TF-IDF, Levenshtein distance and Jaccard Co-efficient is 0.66 0.44, 0.67 and 0.71 respectively.

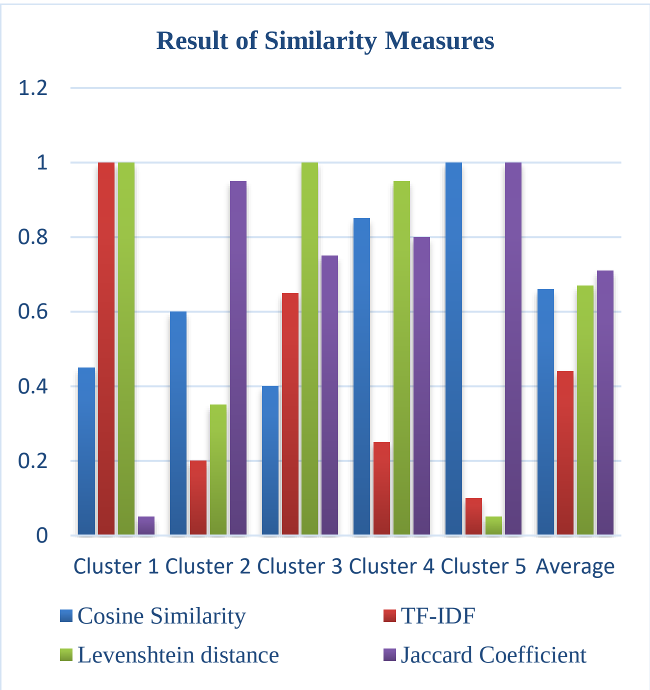

Figure 3 Result of Similarity Measures Before Pre-Processing

Figure 3 demonstrates the average result of five different clusters through several similarity measures techniques before pre-processing.

Table 12 Result of Similarity Measures after Pre-Processing

Number of Clusters	Similarity Measures			
	Cosine Similarity	TF-IDF	Levenshtein distance	Jaccard Coefficient
Cluster 1	0.55	1	0.90	0.95
Cluster 2	0.45	0.20	1	1
Cluster 3	0.95	0.65	0.15	1
Cluster 4	0.95	0.25	1	0
Cluster 5	1	0.1	0	0
Average	0.78	0.44	0.61	0.59

Table 12 shows the average result of five various clusters after pre-processing via different techniques of similarity measures. . The average accuracy of Cosine Similarity, TF-IDF, Levenshtein distance and Jaccard Co-efficient is 0.78 0.44, 0.61 and 0.59 correspondingly

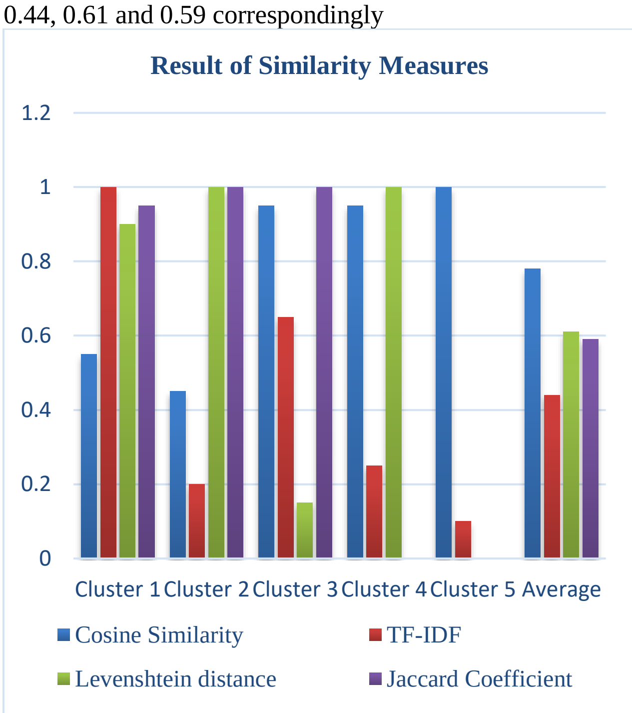

Figure 4 Result of Similarity Measure after Pre-processing

Figure 4 represents the average result of five clusters by utilized four various similarity measures after pre-processing.

8. Conclusion

Now a days, the progressive feelers that are widely adopted around the globe for the development of NLP framework, in almost all languages including Urdu, are machine learning approaches. The core reason behind its

measure," *Journal of systems science and systems engineering,* vol. 15, pp. 474-492, 2006.
[21] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text mining: predictive methods for analyzing unstructured information*: Springer Science & Business Media, 2010.
[22] P. Bide and R. Shedge, "Improved Document Clustering using k-means algorithm," in *Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on*, 2015, pp. 1-5.
[23] M. Peng, J. Huang, Z. Sun, S. Wang, H. Wang, G. Zhuo, *et al.*, "Improving distant supervision of relation extraction with unsupervised methods," in *International Conference on Web Information Systems Engineering*, 2016, pp. 561-568.
[24] R. Ali, M. A. Khan, M. Bilal, and I. Rabbi, "Reciprocal anaphora resolution in pashto discourse," in *Emerging Technologies, 2008. ICET 2008. 4th International Conference on*, 2008, pp. 1-5.
[25] K. Riaz, "Baseline for Urdu IR evaluation," in *Proceedings of the 2nd ACM workshop on Improving non english web searching*, 2008, pp. 97-100.
[26] M. Usman, Z. Shafique, S. Ayub, and K. Malik, "Urdu Text Classification using Majority Voting," *International Journal of Advanced Computer Science and Applications,* vol. 7, pp. 265-273, 2016.
[27] S. Karol and V. Mangat, "Evaluation of text document clustering approach based on particle swarm optimization," *Open Computer Science,* vol. 3, pp. 69-90, 2013.
[28] S. H. Ghwanmeh, "Applying Clustering of hierarchical K-means-like Algorithm on Arabic Language," *International Journal of Information Technology IJIT,* vol. 3, pp. 168-172, 2007.
[29] S. Al-Anazi, H. AlMahmoud, and I. Al-Turaiki, "Finding similar documents using different clustering techniques," *Procedia Computer Science,* vol. 82, pp. 28-34, 2016.
[30] H. Froud and A. Lachkar, "Agglomerative hierarchical clustering techniques for arabic documents," in *Advances in Computational Science, Engineering and Information Technology*, ed: Springer, 2013, pp. 255-267.
[31] J. Prakash and P. K. Singh, "Partitional algorithms for hard clustering using evolutionary and swarm intelligence methods: a survey," in *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*, 2013, pp. 515-528.
[32] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, pp. 281-297.
[33] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert systems with applications,* vol. 36, pp. 3336-3341, 2009.
[34] M. S. Alkoffash, "Automatic Arabic Text Clustering using K-means and K-mediods," *International Journal of Computer Applications,* vol. 51, 2012.
[35] L. Kaufman and P. J. Rousseeuw, "Partitioning around medoids (program pam)," *Finding groups in data: an introduction to cluster analysis,* pp. 68-125, 1990.
[36] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis* vol. 344: John Wiley & Sons, 2009.
[37] R. T. Ng and J. Han, "CLARANS: A method for clustering objects for spatial data mining," *IEEE transactions on knowledge and data engineering,* vol. 14, pp. 1003-1016, 2002.
[38] H. P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 1, pp. 231-240, 2011.
[39] K. Mumtaz and K. Duraiswamy, "A novel density based improved k-means clustering algorithm–Dbkmeans," *International Journal on computer science and Engineering,* vol. 2, pp. 213-218, 2010.
[40] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science,* vol. 344, pp. 1492-1496, 2014.
[41] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Machine learning,* vol. 2, pp. 139-172, 1987.
[42] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Advances in neural information processing systems*, 2000, pp. 554-560.
[43] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation,* vol. 10, pp. 1299-1319, 1998.
[44] Z.-d. Wu, W.-x. Xie, and J.-p. Yu, "Fuzzy c-means clustering algorithm based on kernel method," in *Computational Intelligence and Multimedia Applications, 2003. ICCIMA 2003. Proceedings. Fifth International Conference on*, 2003, pp. 49-54.
[45] D. MacDonald and C. Fyfe, "The kernel self-organising map," in *Knowledge-Based Intelligent Engineering Systems and Allied Technologies, 2000. Proceedings. Fourth International Conference on*, 2000, pp. 317-320.
[46] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *Journal of machine learning research,* vol. 2, pp. 125-137, 2001.
[47] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science,* vol. 315, pp. 972-976, 2007.
[48] S. Arch-int, "Web document clustering using semantic link analysis," in *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, 2005, pp. 13-18.
[49] U. Sridevi and N. Nagaveni, "Semantically enhanced document clustering based on pso algorithm," *European Journal of Scientific Research,* vol. 57, pp. 485-493, 2011.
[50] L. Jing, L. Zhou, M. K. Ng, and J. Z. Huang, "Ontology-based distance measure for text clustering," in *Proc. of SIAM SDM workshop on text mining, Bethesda, Maryland, USA*, 2006.
[51] U. Sridevi and N. Nagaveni, "Ontology based semantic measures in document similarity ranking," in *Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference on*, 2009, pp. 482-486.
[52] Z. Fan, S. Chen, L. Zha, and J. Yang, "A text clustering approach of Chinese news based on neural network language model," *International Journal of Parallel Programming,* vol. 44, pp. 198-206, 2016.