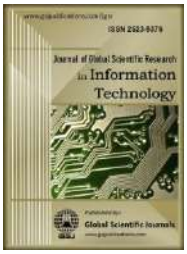
Journal of Global Scientific Research 8 (5) 2023/ 3075-3084

Contents lists available at www.gsjpublications.com

Journal of Global Scientific Research in Information Technology

journal homepage: www.gsjpublications.com/jgsr



Text Mining of Iraqi Law Using Clustering Technique

Mohammed R. M.[1], Zaid R. M.[2], Mohammed I. D.[3]

[1]*Collage of Biotechnology, University of Al-Qadisiyah, Al-Diwaniyah, Iraq.*
[2]*Jabber ibn Haiyan Medical University, An Najaf, An Najaf, Iraq.*
[3]*Collage of Information Technology and Computer Science, University of Al-Qadisiyah, Al-Diwaniyah, Iraq.*

ARTICLE INFO

Received: 21 Mar 2023,
Revised: 23 Mar 2023,
Accepted: 3 Apr 2023,
Online: 3 May 2023

Keywords:
Data mining, Stemming, ISRI, supervised clustering, K-mean

ABSTRACT

The quantity of online text data has expanded hugely in contemporary years because of the rise in demand for the world wide web. Consequence, there is a need to accommodate efficient content-based retrieval, search, and filtering for these tremendous and online repositories. It can be applied the proposal system to the articles of iraqi law. In this paper, it can apply the clustering approaches and data sets that have been randomly collected from iraqi scientific journals. This proposed scheme applies the arabic stemmed of the information science research institute (isri) as an initial processing step and supervised techniques of clustering. We have used the equivalent metrics euclidean and cosine in the proposed system. Our proposed system was evaluated by using (precision ،recall and f-measure) certainly, the valuation study is hand-me-down. The testing showed that the proposed system got accepts results in distributed papers according to the fields of Iraqi law. The method is easily understandable to non-mathematicians.

1. Introduction

Data mining is the way toward obtaining irregularities, models and association inside tremendous data collections to predict results. Employing a wide field from rules. In the corpus, Text Mining discovers new knowledge using appropriate devices, which surpasses the ability of humans to recognize information models and gain significance from corpus due to contrast with human thoughts and desires. Clustering one of the unsupervised training approaches. It uses the documents clustering in Texts Mining employment, which gathers records in the relative concentrations. The clustering of documents is one of the approaches used to aggregate records holding relevant data in groups, which encourages the distribution of related data. This method can improve the hunting procedure of a recovery framework productively and can help decide covered or obscure connections within a social network. [1,2].

However, since the Arabic language is rich and challenges high-quality treatment, for instance, request action terms, morphological analysis, few study attempts were conducted out through Arabic corpus. Primarily, In Arabic morphology, terms have royal associations and include a lot of grammatical and lexical information. More than

Corresponding author:
E-mail addresses: mohammed.rajih@qu.edu.iq (Mohammed, R.), zaid.rajah@jmu.edu.iq (Zaid), Mohammed.iqbal@qu.edu.iq (Mohammed, I.)
doi: 10.5281/jgsr.2023.7885180

300 million people in the world use the Arabic language as one of the universal languages. It differs from a different right-to-left language, in which they include it. The character set of the Arabic language comprises 28 letters and includes several symbols reaching a minimum of 90 letters. [3].

Arabic is a complex language for information retrieval (IR) for several causes. First, it accepts orthographic changes in Arabic; it can compose concluded mixtures of characters in many forms. Second, Arabic has a highly difficult morphology. Third, they well know irregular plurals. Irregular plurals are moderately similar to irregular plurals in English, without rarely resembling the singular figure, as irregular plurals resemble the singular figure in English. Since irregular plurals do not comply with normal morphological rules, exiting stemmers are not controlled. Fourth, because of the trilateral root system, Arabic words are usually ambiguous. In Arabic, a word is generally derived from a root that usually includes three letters. We may drop one or more of the root letters in some roots, breaking many highly ambiguous Arabic words with one another. Fifth, in written Arabic, it drops brief letters. Six, There are extensive synonyms, maybe because we appreciate a difference in appearance as an element of a great writing characteristic by Arabic speakers. [4,5].

This research uses the Clustering Technique to take part in Arabic Text Mining. It uses the system of proposals for articles that concentrate on Iraqi law. This research aimed to cluster these articles into collections according to the fields of Iraqi law and to support researchers in their processing of investigations. Those papers have been published in Iraqi scientific journals, and a lot of articles are used to do this research. Iraqi scientific journals randomly collect the articles that study Iraqi law because we use the clustering approaches in the proposed system and there are no data sets that have been examined in Iraqi law before. We organize this study as follows, The next section Related work, section 3 explains the used techniques, section 4 contains the proposed system, section 5 contains experimental result and evaluation, finally, section 6 contains conclusions.

2-Related work

Different researchers have studied how to combine information significantly within a clustering process.

In 2004 Charu et al. [6] by applying managed clustering techniques, it analyzes the benefits of building content organization structures, examines the use of grouping to get agreements and its use for document characterization. Fully unsupervised clustering has the difficulty of separating acceptably fine-grained types of records with a rational topic by experiencing issues. Where to use the data from a previous scientific organization to monitor the configuration of many relevant groups, but with some opportunity to identify and create the classes. Explain that the advantage of using inadequately managed bunching is that it is understandable to have some authority over the range of subjects that might require the classification structure to mark how each classification is characterized with an exact mathematical meaning.

In 2005 Kazem et al.[7] A root-extraction stemmer was performed for Arabic, similar to the Khoja[8] stemmer, but externally, a root dictionary. Tasks performed on the Arabic documents group in monolingual document retrieval tasks, that stemmer was got to produce equivalently to the Khoja stemmer and so-called light stemmers. Therefore, a root dictionary does not enhance retrieval of Arabic monolingual documents.

In 2017, Salma et al. [9] introduced a study to restrict the top portion of the papers and develop Arabic content mining using clustering. A co-reference objects system using the bunching calculations k-medoids and k-mean means was used to achieve this goal. It employs the representation measures of Euclidean and Cosine in this research. Using a corpus that includes 200 Arabic game news, the structure satisfies. Finally, to evaluate our framework, evaluation measures (Precision-Recall and F-measure) are used.

3- Used Techniques

We have used different kinds of methods in this scheme. These methods can be described as a method used in document clustering's, such as Stemmer (ISRI), a method that helps to determine

the relevance of document clustering quality in unsupervised clustering's, such as k-means, and a set of estimates. This section will provide a skimpy description of each method:-

3.1 *Stemming* is the reverse scheme of the derivative experiment wherever a root should be extracted from a presented term. ISRI Stemmer [7] is an additional simulation of the same Khoja Stemmer grammatical method[8]. This begins with the normalization of the term info, the removal of diacritics and the irrelevance of Arabic characters. The key to regularity is to collectively produce the distinct characters of Hamza to A, which opposes Khoja stemmer. The regulated term succeeds in a series of alternatives to expel understandable prefixes that are three or fewer letters and then outline it as shown by its length to a collection of samples. ISRI. If there is none an equal, scans for likely equals within the purposes of a company; it starts by eliminating comprehensible post-fixes. When the rest of the information term range is three or fewer letters, I must test the stemming mechanism. The key reverse of Khoja stemmer added is that ISRI does not admit roots opposite to references to a word. To achieve the unimportant representation of an info term that can be used for data retrieval, ISRI is progressively set up. For example, there are some signs, the lack of word mention, the eliminated roots are not true, the root could be a negligible component of the letter. Roots would be variable for further qualifying, unusual for phonetic based assignments[8].

3.2 *K- Mean* is one of the easiest unsupervised Education algorithms which clarify the well-popular clustering matter method usually used, which is an easy and quick way. It is simple to perform and has a tiny number of iterations. [10,11].

3.3 *In the* ***TF-IDF*** exhibition, the term frequency for every term is normalized by the inverse document repetition or IDF. The inverse document frequency normalization decreases the weight of words which happen more regularly in the group. This decreases the weight of basic words in the group, guaranteeing that the matching of texts is further affected by that of more distinctive terms which have approximately low frequencies in the group. [6].

3.4 Weight ***wi*** In the method of vector space, any text is supposed to be interpreted as a term vector for the figure ***ã = (a1, a2, . . . An)***. The weight ***wi*** is correlated with all the words ***ai***, where ***wi*** means the normalized repetition of the term in the vector space. Cosine normalization is a well-popular normalization technique. The weight wi of the term I is computed as illustrated in Eq. 1 for cosine normalization [3,6]:

$$w_i = \frac{tf_i * idf_i}{\sqrt{\sum_{i=1}^{n} (tf_i * idf_i)^2}} \qquad (1)$$

The amount of ***tf*** here means the term frequency of ***ai***, whereas the amount of ***idfi*** means the frequency of the inverse document.

3.5 The similarity of the two texts can be estimated by measuring the correlation of the cosine between the documents. The similarity of cosine between two weight vector documents computed as illustrated in Eq. 2 for cosine [1,6] :
U =($u_1 \ldots U_n$) and V =($v_1 \ldots V_n$) is given by:

$$\boldsymbol{cosine}\ (\boldsymbol{U}, \boldsymbol{V}) = \frac{\sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{v_i}) * \boldsymbol{f}(\boldsymbol{u_i})}{\sqrt{\sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{v_i})^2} * \sqrt{\sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{u_i})^2}} \qquad (2)$$

Here, f(.) is a damping function, such as a logarithmic function or a square root. We note that the normalization applied from the method of cosine normalization, which is typically applied in the study of text retrieval.

3.6 Semi-Supervised Clustering earlier information may be ready around the classes of clusters that are possible in the underlying data. This earlier information may take on the scheme of names associated with the documents, which means its underlying issue we refer the method of employing such labels to supervise the clustering method to as semi-supervised clustering. This form of learning is a connection between the clustering and classification difficulty because it applies the underlying class formation, but it is not fully tied down by the special formation. As a result, such a way gets applicability both to the clustering and classification scenarios. [6].

3.7 A projection of a document is specified by arranging the term frequencies (or weights) of some terms in the vector design of the document to zero. These are the terms that are related to being projected out. Apply the process of projection frequently in the subject of the supervised clustering algorithm. Each cluster is

denoted by a seed vector holding just a determined height number of projected words. Projection aims to separate a nearly small vocabulary that represents the subject material of a cluster completely while filtering out the non-relevant characteristics for that class. Use an incremental process of increasingly discovering the most suitable set of projected words, while concurrently cleaning the clusters, to meet the best feature set for each cluster. The iterative way of our method is like the K-means algorithm, although in our case, we also used a progressive merging method [6,9].

3.8 Evolution metric: Precision and recall mean that fundamental criteria applied in evaluating research approaches. Those criteria consider: That exists a collection of documents in the corpuses that is related to the study subject that documents are found either well be related or unrelated (these criteria do not admit for levels of relation). the genuine recovery collection might not ideally equal the collection of related documents. The understanding of recall and accuracy fits explicitly into Figure (1). Four different collections are available: documents that have been retrieved, documents that have not been retrieved, related documents and unrelated documents (as explained in the test set). The number of unrelated documents not reversed (true negatives), B is the number of unrelated documents reversed (false positives), C is the number of unrelated documents not reversed (false negatives) and D is the number of related documents reversed at the intersections of these sets (A, B, C, D) expressed according to A. (true positives). [9,11]. It defines the recall as illustrated in Eq. 3

$$\boldsymbol{Recall} = \frac{\boldsymbol{tp}}{\boldsymbol{tp} + \boldsymbol{fn}} \quad \textbf{(3)}$$

During applying the Figure(1), that equalisation will interpret as :

$$\boldsymbol{Recall} = \frac{\boldsymbol{D}}{\boldsymbol{D+C}} \quad \textbf{(4)}$$

It defines precision as illustrated in Eq. 5:

$$\boldsymbol{Precision} = \frac{\boldsymbol{tp}}{\boldsymbol{tp+fp}} \quad \textbf{(5)}$$

During utilizing the Figure (1), that equalization will interpret

as:

$$\boldsymbol{Precision} = \frac{\boldsymbol{D}}{\boldsymbol{D} + \boldsymbol{B}} \quad \textbf{(6)}$$

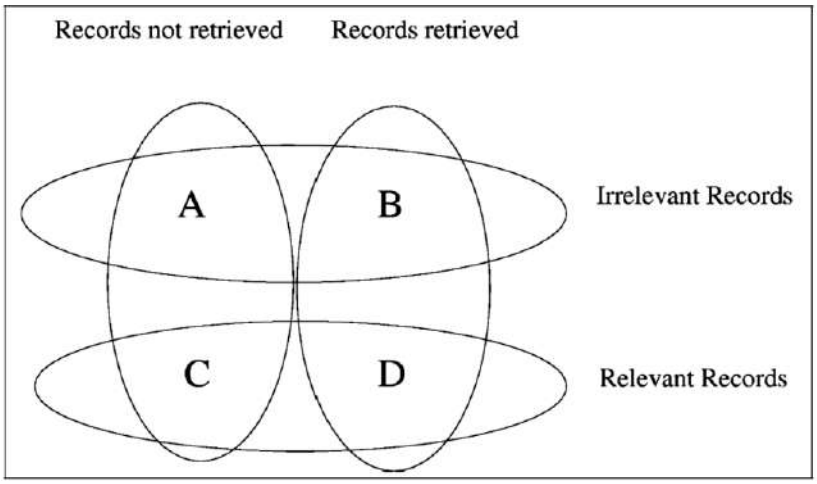


Figure 1. Defined Recall And Precision

An usually used measure in information restoration the so-called F-measure as illustrated in Eq. 7. Rijsbergen beginning proposed this measurement, and it mixes recall (r) and Precision (p) with equivalent weight in the next figure:

$$\boldsymbol{F(r,p) = \frac{2rp}{r+p}} \quad \textbf{(7)}$$

4-The proposal system :

This system included two phases as shown in the following training phase treatment: the documents and initial seeds to a got grouping of this document and sets of more frequency terms that have been used in the next phase(test phase). Below explain this two-phase in more details.

4.1 The first phase(training phase): this phase included many stages as shown in figure (2) :

4.1.1 Preprocessing stage: in this stage work will be divided into two parts First part treat documents and convert them to terms, second part converts the stored words in the text file that represent the initial case of seeds. As explain in figure (2).

First, we will process the text file that represents the data set for this paper and convert it to terms, This step is completed by using (ISRI) refer to section 3.1. But (ISRI) For example, the term (قانون) was given as a word (قان) and the term (قان) was given as a word (قان) and the word (قان) was given as a word (قان). This article suggested a special dictionary in Iraqi law to solve this weakness by sending the term in texts and the roots that were extracted to this dictionary and restoring the true root. And then, find features of every term. Then extracting features (TF, IDF, w) refer to section 3.3 and 3.4. For each term inside the text file and the relational of terms inside a corpus. Second, converted the stored words of a text file that represents the data set for this paper to terms vectors that represent the initial seeds for each cluster.

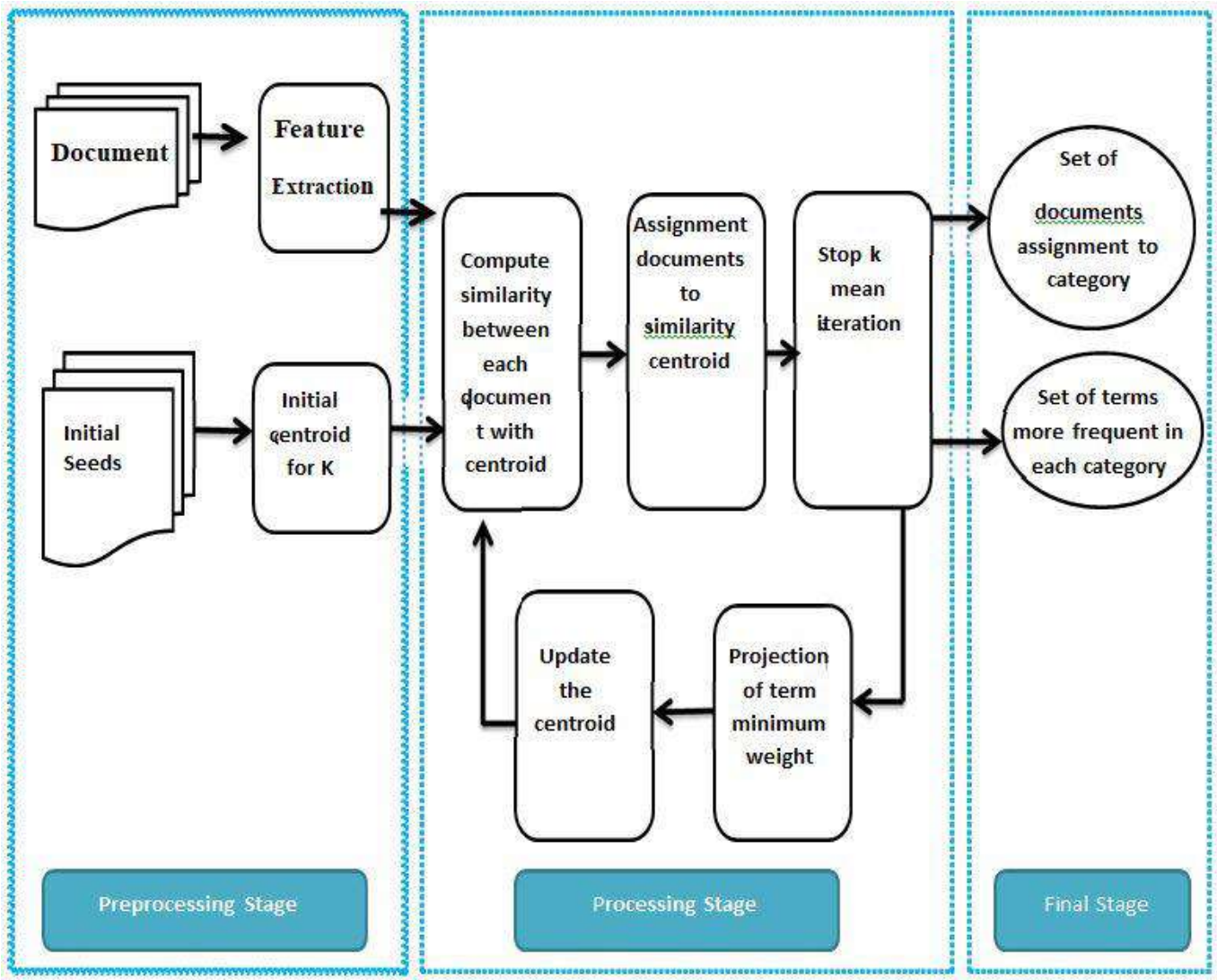


Figure 2. law documents training process

4.1.2 *Processing stage*: in this stage, they will do us all the processing steps first, the following steps will perform the clustering process :

1. *Compute the similarity*:- in this step we compute the similarity between every document with every seed refer to section 3.5.
2. *Allocation the documents*:- in this step we assign the document to seed depending on the high similarity refer to section (3.2 and 3.6), for example, allocate the document (X) that have the high similarity with a seed (Y). Finally, we will be every document allocate to one seed that closet to it, and mirage the terms for every document in one seed in the same vector.
3. *Compute the weight*:- we compute the weight for every term in one seed that found in the same vector by applying the Eq.(8) :

$$W_{ts} = \frac{\sum_0^m w_{ij}}{n} \qquad (8)$$

Where (W_{ts}) represent the term weight in the seed, (Wij) the weight of term (i) in a document (j), (n) the number of seeds and (m) the number of documents inside one seed that contain the term.

4. *Descending order:* - we rearrange the terms inside one seed by extra weight that compute in the last step.
5. *Projection technique*: - refer to section 3.7 in each iteration, separate the terms with minimum weight from the pseudo centroids of the former importance. This guarantee that we use just the terms which are many of the time appearing inside a collection of documents for the assignment method. The number of terms that are

separated in each iteration is with the terminal The aim is to decrease the number of non-zero weight terms in each repetition by a geometric factor. By (0.7), we mean this geometric factor, the use of an iterative projection approach assists get the terms. Which are commonly descriptive of the topic element of a collection? This is because, at the beginning, not several emphases when the collections are not so filtered, a higher number of measures should be included in the projection to save away from an unfortunate waste of data. In a later iteration, the clusters become gradually filtered, and it is understandable to stretch down to a shorter number of terms.

6. *Stopped the processing*:- we use the following procedure to stop the iteration of k mean to refer to section (3.2) by computing the ratio of unchanged in each seed between the last iteration and current iteration will be over 95%.

*4.1.3 Last Stage***:** in this stage, we will get the output of our proposed system in two forms :

3.1 First Output: will be assigned each document in the training set to the specific seed and can consider each seed as clusters.
3.2 Second Output: will be the terms more frequency in each cluster depend on the documents allocated.

4.2 The second phase (testing phase): testing that will be done through applying the same steps in the training phase but exchange the initial random seeds by the output of the training phase (Second Output) for applying as initial seeds and the documents that choice as a test set.

5. Experimental result and evaluation

This module applies to data from published scientific papers in law collected from Iraqi universities. There is not an existing bank of similar data to compare results with. We based the collected data on personal effort. A collection of (100) publishes we decided the paper into two groups, the first group included (30) it used published a paper during the training phase. Whereas the second group included (70) used during the test phase.

When applying the proposed system to this research which represents (the data set) in the training phase, I got the following results :

1. The compilation of research papers used in the training phase into groups similar in some terms, where each group of research can be considered as a specific class.
2. Getting terms that represent the most frequent words in the group, and therefore these words can be keywords for a specific set of research, i.e. any type of law, we can benefit from these results in information retrieval technology.

When applying the proposed system to special research in the testing phase and using the words produced from the training stage, refer to section (4.1.1) extracting features (TF, IDF, w) for all documents in test set and used terms more frequency refer to section (4.1.3) second output of training phase. After that, refer to section (4.1.2) applying steps of processing on test set.it distributed the research in the testing phase to clusters refer to section (4.1.3) and compute number of documents in each clusters . As shown in the table (1).

Table 1. explains distributed documents for each cluster

Text categories	No. of documents in each cluster
The Constitutional law	6
Administrative Law	7
international law	10
Criminal Law	18
civil law	7
Commercial Law	11
Personal Status Law	11

Table (2) show the distributed documents for each cluster ,When applying the proposed system and applying one of clusters approach (k-mean) on test set .

Table 2. comparing the distributed documents for each cluster

Text categories	No. of documents in each cluster	
	k-mean	proposed system
The Constitutional law	5	6
Administrative Law	8	7
international law	12	10
Criminal Law	16	18
civil law	7	7
Commercial Law	12	11
Personal Status Law	10	11

We compared these distribution results with the opinion of the experts and refer to section 3.8 got (TP, FP, FN, TN) for each cluster. As shown in the table (2)**.**

Table 3. explains (TP, FP, FN, TN) for each cluster.

Text categories	TP	FP	FN	TN
The Constitutional law	4	0	1	26
Administrative Law	5	1	1	24
international law	3	1	0	26
Criminal Law	6	0	1	24
civil law	5	1	1	24
Commercial Law	3	1	0	26
Personal Status Law	4	0	1	26

Refer to section 3.8 and the result in the table (3) find the evaluation metric for each cluster. As shown in the table (4)**.**

Table 4. explains the evaluation metric for each cluster.

Text categories	Recall	Precision	Accuracy	F-measure
The Constitutional law	0.8	1	0.967742	0.888889
Administrative Law	0.833333	0.833333	0.935484	0.833333
international law	1	0.75	0.966667	0.857143
Criminal Law	0.857143	1	0.967742	0.923077
civil law	0.833333	0.833333	0.935484	0.833333
Commercial Law	1	0.75	0.966667	0.857143
Personal Status Law	0.8	1	0.967742	0.888889

The result above shown the highest value of recall (**1**) in a cluster (**القانون التجاري, القانون الدولي**) where the lowest value was in (**0.8**) in the cluster (**الاحوال الشخصي, القانون الدستوري**) that mean the value of recall between (**0.8 – 1**). Where the value of Precision Ranging (**0.75—1**) An efficient system has to attain as high a recall value as possible without having to wastage the precision. While accuracy metric value between **(0.93 – 0.96**) that mean the accuracy of the proposed system is height. The highest value of the F-measure (**0.92**) reflecting the efficiency of the proposed system. Table (5) show the average of all metric ,When applying the proposed system and applying one of clusters approach (k-mean) on test set .

Table 5. explains average of evaluation metric for cluster method .

Cluster method	Recall	Precision	Accuracy	F-measure
proposed system	0.869	0.958	0.881	0.875
k-mean	0.839	0.978	0.851	0.845

Through the results, we note that the proposal module got accept results in distributed papers.

6. Conclusion

We can see that the proposed system gives comparable results even though it doesn't require complicated computations or manipulations. This paper introduced an enhancement version of supervised clustering. The enhancement refers to accuracy clustering, where the enhancement comes from the considerable increased in the accuracy clustering of traditional supervised clustering. They performed experiments on the data set. The testing showed that the proposed system got accepts results in distributed papers. The method is easily understandable to non-mathematicians.

7. References

[1]. Hanan M. Alghamdi a,l, Ali Selamat b, 2019 "*Arabic Web page clustering: A review"*, Journal of King Saud University – Computer and Information Sciences 31,p 1–14.

[2]. Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A, others. (2015). A *survey on text mining in social networks*. The Knowledge Engineering Review, 30(2), 157–170.

[3]. Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. book, Springer Science & Business Media.

[4]. Rogério dos Santos Alves; Alex Soares de Souza, et all. (2014). *Data Mining*. Igarss. Elsevier.

[5]. Al-Anzi, F.S., AbuZeina, D., 2016. *Big data categorization for Arabic text using latent semantic indexing and clustering.* In: International Conference on Engineering Technologies and Big Data Analytics. IIE, Bangkok, Thailand, pp. 1–4.

[6]. C. C. Aggarwal, S. C. Gates, P. S. Yu. 2004 *"On Using Partial Supervision for Text Categorization", IEEE Transactions on Knowledge and Data Engineering*, 16(2), 245–255,

[7]. Kazem Taghva, Rania Elkhoury, and Je_rey S Coombs. 2005. *Arabic stemming without a root dictionary*. pages 152–157.

[8]. Shereen Khoja and Roger Garside. 1999. *Stemming Arabic text.* Lancaster, UK, Computing Department, Lancaster University.

[9]. Salma M. and Faiez M. 2017 "*Arabic Text Mining Based On Clustering And Coreference Resolution* " International Conference on Current Research in Computer Science and Information Technology (ICCIT), Slemani – Iraq.

[10]. Y. H. Ali, M. R. Mohammed 2017," *Background modelling in video surveillance by using parallel computing", Iraqi Journal of Science, Vol. 58, No. 3B, pp: 1516-1522, Iraq*.

[11]. M. R. Mohammed, 2018," *Improved K-mean algorithm for background subtraction in video surveillance* " M.Sc. thesis, Computer Science, University of Technology,Iraq.

[12]. Alhammadi, N. A. M., Zaboon, K. H. .2022. A Review of IoT Applications, Attacks and Its Recent Defense Methods. *Journal of Global Scientific Research.* 7(3), 2128-2134.