www.ijecs.in
*International Journal Of Engineering And Computer Science ISSN: 2319-7242*
*Volume 6 Issue 1 Jan. 2017, Page No. 19890-19910*
*Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i1.08*

# A Novelistic Querying procedure for clustering the Legal Precedents

*[1]B.Basaveswar Rao, [2]B.V.RamaKrishna, [3]K.Gangadhara Rao, [4]K.Chandan*
3 Department of Computer Science & Engineering
1,2 Computer Centre
4 Department of Statistics
Acharya Nagarjuna University, Guntur

## Abstract

In this paper an attempt has been made to propose five query formation methods with professional usage point of view for clustering the dowry crime related legal precedents. K-mean clustering method is used on Tanagra open source for all the methods. The query formation methods are used to generate five types of queries to find the cosine similarity measure between the query Term Frequency Matrix (TFM) and Repository TFM. The repository TFM consists of 500 judgments related to dowry crimes and used in [27]. After the formation of clusters the performance metrics are computed and the results are analyzed in a twofold

i) Cluster analysis for within the query formation method
ii) Comparison of clustering results of different query formation methods

Finally the conclusions and the future scope of the research presented at the end of the paper.

***Keywords*****:** *Clustering, Bag-of-Words, Term Frequency, Document Matrix, Dowry-Death (DD), Dowry Harassment (DH), Dowry Acceptance (DA).*

## 1. Introduction

The goal of a legal documents clustering is to identify the coherence group of documents for given query by the users. The clustering methods normally quantify the similarity/closeness/relevance between the query and the set of documents by similarity measures. There are large number of collections in digital libraries and repositories related to precedent judgments etc. By using clustering techniques to organize these collections into a much smaller number of coherent groups.

In the Legal domain document clustering [1][8] organizes legal documents into clusters with best inter topic similarity. The organization of legal documents into a hierarchical clusters [2] based on topic segments improves the performance of document ranking. Traditional methods of classifying documents based on Bag-of-Words concept. This approach is suited for large corpora of texts whereas short text doesn't support sufficient word occurrences. Therefore semantic knowledge based classification increase accuracy in document classification. Lexical Chaining [3] tracks the semantic information in documents supporting clustering. Latent Semantic Indexing (LSI) is an Information Retrieval technique which reduces vector space and represents document as mixture of topics**.** Latent Direchlet Allocation (LDA) [4] is one of the latest models to represent document as mixture of topics. There has been a substantial improvement in improving the granularity from a document level similarity to an inter passage similarity [6] which obviously improves the clustering accuracy.

During web information retrieval text segmentation and inter document similarities, sentiment analysis and ontology survey techniques [19][20] applied over hierarchical collection of text documents. A novel approach Latent Dirichlet performs similarity measure between topics and documents also frequent concepts based document clustering algorithm. The efficacy of which depends on concepts of documents. Improvements in K-mean algorithm with shared nearest neighbor method [13][14] has been found to improve clustering efficiency. The concept of term based similarity measures [17] incorporate linguistic and semantic structures using syntactic dependencies. Semantic background knowledge is a backbone to these types of methods. This concept based clustering improves classification and clustering accuracy over web document text. In transactional

databases the clustering performed on frequent patterns of item selection. Topic segmentation and Sentiment analysis [11] plays an important role in hierarchical topic cluster algorithms. They help to group words into tree structured chains of topics.

For all these methods the words in the given Query happens to be the heart of the entire clustering mechanism. This would obviously help to create most coherent clusters. The legal domain requires very good approach of query preparation, so that the legal professionals would be at a greater ease. The objective of this paper is to suggest a method which would help to choose the best query preparation method amongst the suggested five methods to fulfill the needs of the legal professionals.

In section 2 various author's contribution to Legal Document clustering is presented. In section 3 Data description described. Section 4 Query Formation Methods and clustering process explained. In section 5 Preliminaries of cluster performance metrics are given the results are provided in section 6. Finally conclusions and further scope of research is given in section 7.

## 2. Literature Review

A. Devender et. al [1] proposed a new correlation based document clustering instead of TF-IDF vector purely based on semantic similarity measure. His work concentrated on concept extraction, semantic associations and meronymy.

S. Sowmya and M. Kanakaraj [2] applied NLP techniques to identify topic based summarized news from all sources to support user query. Semantic based BoW constructed using WordNet synsets. K-mean clustering provided better results with 90% accuracy with this approach.

S. Joshi, M.Prasad Deshpande and Thomas Hampp [3] proposed Electronic Stored Information Discovery model. Employing NDD (*Near Duplicate Detection*), Automatic Classification techniques to create coherent groups of documents. Syntactic grouping of documents detects duplicates in groups and semantic similarity organizes concept based groups. In terms of precision and recall a significant performance noticed by their experiments.

Jack G. Conard et. al [4] performed research to classify and cluster law firms where there are no taxonomies or labeled training documents available. In their work hierarchical and multiple assignment contexts based clustering techniques holds good performance for above mentioned legal data.

Chao-Lin Liu [5] et. al implemented a system to identify criteria to classify legal judgment summaries. Lexical knowledge to identify keyword-based and case-based classification applied in this system. This system achieved 20% quality over human provided cases summaries.

Eui-Hong Han [6] et. al proposed a new association rules based clustering which clusters related items using clusters of items. Their experiments n training data (stock-market, voting data) successfully grouped items belonged to same group. Their clustering shown better results over existed Auto class clustering algorithm.

Zichao Dai [8] et.al proposed a topical relevance model vector where topics are derived from knowledge embedded in short text collections organized using hierarchical clustering with purity control. His experiments over SVM classifier based web snippets shown significant improvement in short text classification.

Dipti Deodhare [9] et.al developed a soft clustering algorithm. Each document turned into lexical chains using WordNet which is beneficial than BoW. A semantic similarity matrix generated based on which lexical chain graph constructed. Documents associated with same cluster would have semantically similar lexical chains. This approach best suited for topic detection among corpus of documents. This approach also resulted good soft clustering of documents.

K.Raghuveer and Ravikumar [12] proposed hierarchical Latent Dirichlet Allocation (HLDA) approach to organize, analyze and present legal information. They clustered legal documents based on topics obtained from HLDA. This model is capable for grouping Legal judgments into different clusters and generate summary of each judgment in an effective manner.

B. Sindhiya and N. Tajunisha [17] introduced a novel method to represent meaning of texts in

dimensional space of concepts derived from Word Net. This is a two-way term → concept semantic relatedness model. Their experiments showed improved performance over existed SMTP model for clustering and classification.

Shshank Paliwal [9] et.al proposed a sub-topic structure of text documents and investigates whether clustering text documents can be improved if text segments of two documents utilized. Using inter document similarity approach over sub topics of documents improved clustering of documents.

Rupali Sunil Wagh [22] performed document analysis over legal domain where plain text documents are subdivided into groups to identify relevant and abstract based keywords. Using cosine similarity domain ontology applied with linguistic preprocessing increases the performance of clustering and better results obtained also his experiments improved the quality of clusters.

J.G. Konard [24] et.al experimented with clustering algorithms over legal firms to cluster legal documents. Clustering based hierarchical and multiple context assignments holds good resulting clusters over legal documents which are unlabeled without taxonomies.

## 3. Data Description

For clustering the legal documents are represented as in the form of term frequency matrix after preprocessing the pdf legal documents. Normally clustering methods are applied based on similarity measure calculated between the repository data and a query given by the user. For this experiment a term frequency matrix of 500 precedent judgment related to dowry cases, generated with the Extended Bag-of-Words EBoW [26] is treated as a repository data.. The EBoW contains 211 dowry related legal terms which are generated in [27]. For the query data the new (not in 500 set) dowry related president judgments (for each one judgment related to dowry death, dowry acceptance and dowry harassment and combine them in to one document) and represented as term frequency matrix with same EBoW [26, 27]. For the crime facts and frequent concepts clustering procedures the query data are prepared with the dowry related 10 FIR's. The FIR's are also represented as a term frequency matrix with the same EBoW [26, 27].

## 4. Query Formation and Clustering Mechanisms

In order to obtain the reliable clusters on dowry related precedents, five types of Query Formation Methods(QFM) are proposed. The methods are Comprehensive Clustering (CC), Crime Specific Mean Clustering (CSMC), Crime Bounded Interval Clustering (CBIC), Crime Facts Clustering (CFC) and Frequent Concepts Clustering (FCC). The dowry related crimes are classified in to three types of crimes they are Dowry Acceptance (DA), Dowry Harassment (DH) and Dowry Death (DD). The queries are generated from the FIR's investigation views of different dowry crimes except CC. In CC the query generated from the precedents with randomly selection of one from each type of crime. The EBoW [27] is adopted for finding the TFMs of CC, CSMC and CBIC. For the FCC the BoW is created through Apriori algorithm and for CFC the BoW are created form FIRs investigation views with the help of legal experts. The clustering mechanism depicted in figure-I.

DOI: 10.18535/ijecs/v6i1.08

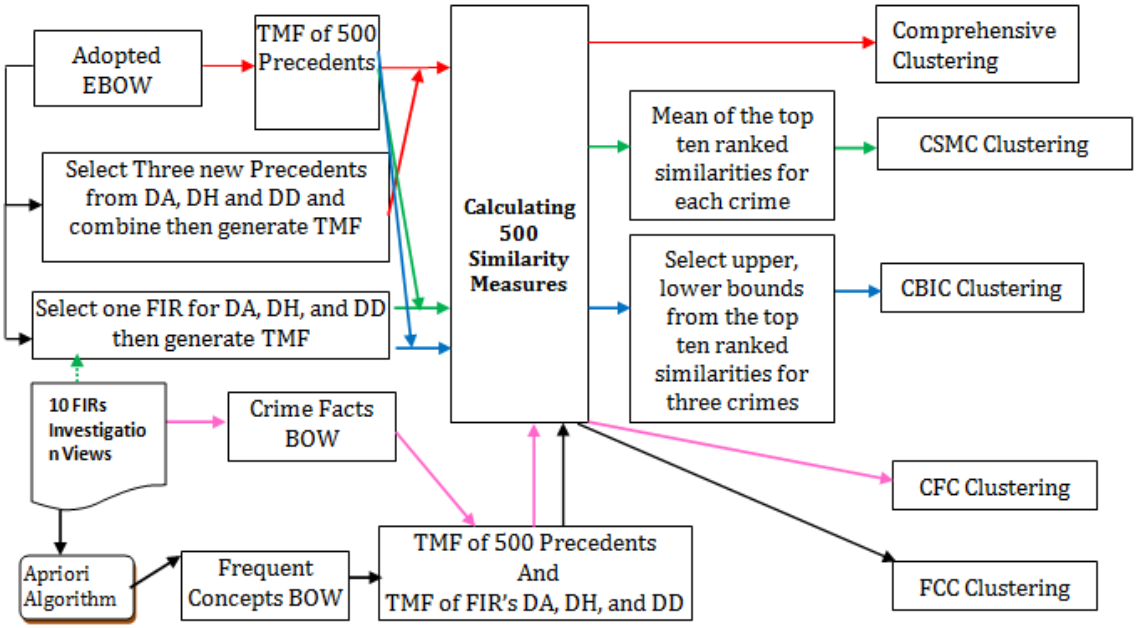



Figure 1. Clustering Mechanism

The various clustering methods are explained below.

4.1 Comprehensive *Clustering (CC)*

The objective of this method to generate a query with a combination of three dowry related crimes i.e. the user query is a comprehensive query of dowry related crimes. In this method the items are clustered based on cosine similarity values between the precedents TFM and the *comprehensive query TFM*. The query created with the combination of new (not included 500 precedents) dowry related crimes one from each type DD, DH and DA president case notes. After preprocessing for finding the query TFM we adopt EBoW [26]. The same EBoW is used to find for each repository TFM. Then calculate the cosine similarity measure between repositories TFM and query TFM using Tanagra [28] data mining open source package to perform K-mean clustering over these cosine similarities with default cluster settings. Four clusters formed with default optimal settings. The means of clusters are identified by inter cluster similarity and intra cluster similarity measures by tool.

4.2 Crime Specific *Mean Clustering (CSMC)*

In this method the repository TFM is same as k-means clustering but the query data TFM is different. The investigation views from FIR is taken and converted in to TFM for three dowry related crimes DD, DH and DA. To find the three types of cosine similarity matrices between the repository TFM and three TFMs. Then identify the top ten ranked similarities of three types of crimes and find the mean of these 10 similarities. These three mean values are taken as a cluster mean and then apply k-means clustering independently for getting three types of crime specific clusters they are DD, DA, DH clusters. These clusters are more helpful to identify the crime specific documents among the repository.

4.3 Crime Specific Bounded *Interval Clustering (CBIC)*

In this method the repository TFM is same as CC and CSMC. But the three query data TFMs is taken in CMC but the difference is in CSMC mean as taken as a cluster mean where as CBIC to fix the boundaries for the clusters. In statistical computations bounded interval function identifies the values between given bounding values maxima and minima. The function defined as follows

$$f(x) = x \forall x \in [a, b] where\ x \geq a\ and\ x \leq b$$

For three types of dowry crimes to calculate a, b boundary values from the top ten ranked similarity scores of DA, DH, and DD. The ten FIR's investigation views are taking formation of query data TFM and find cosine similarities. From ten FIR's DA is 3 , DH is 3 and DD is 4. For finding

the DA maximum and minimum values of top thirty similarities where is each FIR has top ten similarities. For finding the DH maximum and minimum values of top thirty similarities where is each FIR has top ten similarities. . For finding the DD maximum and minimum values of top forty similarities where is each FIR has top ten similarities. These Maxima and Minima bounding values treated as threshold for clustering. When the K-mean clustering based clusters undergoes BIC mechanism clusters become more compact and maintain high similarity precedents per each three dowry crimes. The similarity measure that can't fit in given threshold bounds excluded from cluster.

4.4 Crime *Facts Clustering (CFC)*

In this approach the repository TFM and query data TFM also different from CMC and CBIC. For generation of repository TFM and query data TFM the new BoW is created with only crime facts related words. These BOW words are selected from the investigation views of the 10 FIRs of three type's dowry crimes. Then this BoW also modified by the legal expert. This BoW is used to find the TFM of repository data of 500 precedents. For formation of three types of query and data TFMs for each dowry crimes DA, DH and DD then find the three types of cosine similarity measures for 500 precedents. The K-mean clustering applied on three types of similarity matrixes to get the DA, DH and DD clusters. This crime fact based clustering grouped presidents with crime intensity. Figure 4(b) shows the clear separation of DD, DH and DA based judgments into specific clusters.

4.5 *Frequent Concept Clustering (FCC)*

In this approach the BOW is created with frequent concepts of the 10 FIRs. The Apriori algorithm applied from Tanagra over the 10 FIRs investigation views to identify the three frequent item sets. The item sets are depicted in the following table.

Table 1: Frequent item sets for 10 FIRs

| | |
|---|---|
| ***Concept 1*** | Dowry, act, section, crime, domestic, forensic |
| ***Concept 2*** | Murder, death, violence, harassment, arrest |
| ***Concept 3*** | Police, witness, culprit, victim, local, accuse |

The new Bow is created with only 16 words. From this BOW to generate repository TFM of 500 precedents and three query data TFMs of 10 FIRs of DA, DH and DD. The DA, DH and DD cosine similarities are calculated and apply K-means clustering. The clusters are formed on the bases of frequent concept made up of documents that contain words related to a frequent concept.

## 5 Preliminaries of Metrics

The legal documents clusters are formed by the above approaches are evaluated by the widely accepted measures and the basics are explained below and they are defined in [7][10][18][22][29].

i) ***Purity***:

The coherence of a cluster is evaluated by purity that is the degree to which a cluster contains documents of single category. Purity $P(C_r)$ defined as the number of documents of the largest category in a cluster divided by cluster size.

$$P(C_r) = \frac{1}{n_r} max_i(n_r^i)$$

Where $C_r$ is a particular cluster of size $n_r$ and $max_i(n_r^i)$ is the number of documents that are formed the dominant category in cluster $C_r$ and $n_r^i$ represents the number of documents from cluster $C_r$ assigned to category i. Purity is a function of the relative size of the largest category in the resulting clusters. The overall purity of clustering is obtained by taking weighted sum of the individual cluster purities.

$$\text{Purity} = \sum_{r=1}^{k} \frac{n_r}{n} P(C_r)$$

Where k is total number of clusters and n is total number of documents. For an ideal cluster the purity value is high nearer to one.

ii) ***Precision and Recall***:

On the basis of common cluster documents precision of a cluster given as

$$Precision(i,j) = \frac{n_{ij}}{n_j}$$

Recall is the probability of class relevant information supported by a cluster with set of documents relevant to that category. It is given as

$$Recall\ R(i,j) = \frac{N_{ij}}{N_i}$$

Where $n_{ij}$ is the number of documents of category $i$ in cluster $j$, $n_j$ is the number of documents of cluster.

iv) ***F - Measure***: F-measure is a score to evaluate the quality of clustering. It is the harmonic combination of precision and recall. Let there be a category i and cluster j, then F-Measure calculated as

$$F(i) = \frac{2X(Precision)X(Recall)}{Preicisin+Recall}$$

The overall F-measure for the clustering result is the weighted average of the F-measure for each category i

$$F_C = \frac{\sum_{i=1}^{k}((n_i)F(i))}{\sum_{i=1}^{k} n_i}$$

Where $n_i$ is number of documents in $i^{th}$ category.

v) ***WSS***: Within Sum of Squares (WSS) is the total distance of data points from their respective cluster centroids. It decides the cohesion of cluster items within a cluster. The lower score shows the optimality of specific cluster. An ideal clustering maintains WSS score less than 40.

vi) ***R-Square***: This measure assumes that all the points belonged to a cluster fitted to a Regression Line. The values lie between (0-1). The R-Square value more than 60% is worthy clustering. Higher the value indicates more similarity exists between items in a cluster. The overall items going to be clustered also affect the R-Square value.

vii) ***Homogeneity***: It measures the homogeneity of precedents present in a specific cluster. For any Cluster $C_k$, k= [1... n] clusters. L is the total number of precedents in cluster. S is the subset of $L$ shared by other clusters. Homogeneity (H) calculated as $\frac{L-S}{L}$ defined. If the measure, H, is high value then it shows that all the precedents have the uniqueness and also stability of the cluster.

## 6. Results & Analysis

The five QFM's are used to group the relevant precedents, to buttress any query prepared by a professional user. The basis for the queries happens to be the precedents and FIR's too. All the methods used Tanagra® an open source tool for generating clusters. The cluster performance metrics are calculated and presented in Table-2.

### 6.1 Within QFM Clusters Analysis

In Comprehensive Clustering there are four clusters formed and the results are shown in Fig. 3(a), 3(b) of the Annexure-I. From the four clusters the third cluster has a centroid similarity value for the both repository and the dowry comprehensive query i.e. 69 and 64 respectively. The precedents in the third cluster are scattered around the centroid where as the proximity of the presidents happens to be very close or even intersecting at times. The third cluster size is very less i.e 38, to compare to other clusters and homogeneity is also high, so the legal professional may retrieve and study these precedents for any type of dowry related crimes.

In CSMC three queries generated for individual dowry crimes DD, DH and DA. For each separate dowry crime top ten ranked similarities are considered to evaluate mean of the similarity scores. Hence for DD, DH and DA mean values are 73.21, 74.79 and 72.99 respectively. These mean values are treated as potential centroids for a clustering, for each DD, DH and DA queries independently. The results are depicted in figures 4[a...f]. Each cluster hardly holds 9 to 12 precedents only and homogeneity is also > 70% for all the clusters Legal professionals can retrieve limited set of precedents for specific dowry crimes.

In CBIC similarity measures calculated as similar to CSMC. Applying the bounded Index value to fit cluster within given range [71 to 77] for each DD, DH and DA three independent clustering results presented in annexure figure 5[a…f]. Each bounded index cluster crops the precedents within the bounded limits. Legal professionals interested to analyze the precedents within a range of similarity to specific dowry crime can choose this method. In this experiment all the three DD, DH and DA clustering maintained an average of 11 precedents in bounded index clusters and homogeneity is >84%.. from these results this type of querying be produced better cluster.

In CFC legal expert selected crime facts BoW used to generate query TFM for DD, DH and DA. The clusters generated for each dowry crime query separately. All the precedents related to a specific dowry crime related FIR query are grouped into a cluster with similarity score range

of 65-80. The query TFM generated by combining the specific dowry crimes independently, so three DD, DH and DA TFMs are generated. This method is useful to Legal professionals searching precedents for a specific dowry crime.

In FCC a new BoW based on frequent concepts used to generate TFM for DD, DH and DA respectively. Applying theses three TFMs independently on precedents TFM similarity scores generated. The precedents are grouped basing on the concepts density as shown in figures 7[a...f]. For DD query TFM generated clustering the cluster with high similarities represents the precedents with DD crime related concepts high. Similarly the remaining two clustering's groups the precedents with high concept similarity to the cluster with high scores.

Table 2. Performance Metrics for various Clustering approaches

| Clustering Mechanism | Cluster No. | Purity | Precision | Recall | F-Measure | Homogeneity | WSS |
|---|---|---|---|---|---|---|---|
| Comprehensive Clustering (CC) | $C_1$ | 0.6729 | 0.6966 | 0.1879 | 0.5637 | 61.3% | 55.18 |
| | $C_2$ | 0.6833 | 0.7392 | 0.1982 | 0.5946 | 53.6% | 26.10 |
| | $C_3$ | 0.6336 | 0.7107 | 0.4748 | 0.5692 | 71.4% | 29.88 |
| | $C_4$ | 0.6128 | 0.6861 | 0.3378 | 0.4527 | 67.3% | 43.05 |
| CSMC (DD) | $C_1$ | 0.9456 | 0.8972 | 0.2455 | 0.7365 | 70.3% | 9.34 |
| | $C_2$ | 0.9421 | 0.9025 | 0.2623 | 0.7869 | 72.6% | 10.12 |
| | $C_3$ | 0.9433 | 0.9016 | 0.2724 | 0.8172 | 72.3% | 7.12 |
| CSMC (DH) | $C_1$ | 0.9124 | 0.9198 | 0.2643 | 0.7929 | 73.7% | 10.24 |
| | $C_2$ | 0.9205 | 0.9177 | 0.2589 | 0.7767 | 71.4% | 4.32 |
| | $C_3$ | 0.9112 | 0.9244 | 0.2711 | 0.8133 | 75.6% | 14.12 |
| CSMC (DA) | $C_1$ | 0.9256 | 0.8674 | 0.2517 | 0.7551 | 72.1% | 16.22 |
| | $C_2$ | 0.9203 | 0.8799 | 0.2581 | 0.7743 | 70.8% | 9.78 |
| | $C_3$ | 0.9189 | 0.8644 | 0.2507 | 0.7521 | 76.3% | 13.11 |
| CBIC (DD) | $C_1$ | 0.8012 | 0.8726 | 0.3076 | 0.9228 | 89.3% | 2.49 |
| | $C_2$ | 0.8029 | 0.8854 | 0.3112 | 0.9336 | 92.8% | 2.98 |
| | $C_3$ | 0.8033 | 0.8792 | 0.3276 | 0.9828 | 94.7% | 2.75 |
| CBIC (DH) | $C_1$ | 0.8206 | 0.8524 | 0.2972 | 0.8916 | 85.6% | 2.39 |
| | $C_2$ | 0.8194 | 0.8479 | 0.3124 | 0.9372 | 93.1% | 1.53 |
| | $C_3$ | 0.8189 | 0.8563 | 0.2993 | 0.8979 | 86.2% | 1.43 |
| CBIC (DA) | $C_1$ | 0.8102 | 0.8526 | 0.3048 | 0.9144 | 89.1% | 1.45 |
| | $C_2$ | 0.7997 | 0.8493 | 0.3174 | 0.9522 | 94.3% | 2.88 |
| | $C_3$ | 0.8156 | 0.8511 | 0.2912 | 0.8736 | 84.7% | 2.35 |
| CFC (DD) | $C_1$ | 0.6921 | 0.7123 | 0.2258 | 0.6774 | 63.8% | 15.11 |
| | $C_2$ | 0.6543 | 0.6933 | 0.1883 | 0.5649 | 72.5% | 52.71 |
| | $C_3$ | 0.6897 | 0.6827 | 0.1944 | 0.5832 | 57.3% | 48.50 |
| CFC (DH) | $C_1$ | 0.6124 | 0.6503 | 0.1758 | 0.5274 | 51.8% | 22.14 |
| | $C_2$ | 0.6577 | 0.6642 | 0.2012 | 0.6036 | 72.6% | 34.98 |
| | $C_3$ | 0.6432 | 0.6433 | 0.1876 | 0.5628 | 61.9% | 11.54 |
| CFC (DA) | $C_1$ | 0.6021 | 0.6188 | 0.1674 | 0.5022 | 72.1% | 19.14 |
| | $C_2$ | 0.6009 | 0.6101 | 0.1704 | 0.5112 | 69.3% | 26.27 |
| | $C_3$ | 0.6141 | 0.6203 | 0.1686 | 0.5058 | 58.2% | 36.81 |
| FCC (DD) | $C_1$ | 0.8936 | 0.8085 | 0.2397 | 0.7191 | 76.3% | 35.11 |
| | $C_2$ | 0.9183 | 0.8367 | 0.2562 | 0.7686 | 71.2% | 22.71 |
| | C3 | 0.8260 | 0.8695 | 0.2173 | 0.6519 | 72.6% | 38.50 |
| FCC (DH) | $C_1$ | 0.8936 | 0.8283 | 0.2412 | 0.7236 | 75.3% | 31.80 |
| | $C_2$ | 0.9243 | 0.8016 | 0.2542 | 0.7626 | 71.4% | 32.91 |
| | $C_3$ | 0.8231 | 0.8176 | 0.2114 | 0.6342 | 68.6% | 47.48 |
| FCC (DA) | $C_1$ | 0.7106 | 0.7913 | 0.2152 | 0.6456 | 74.8% | 38.53 |
| | $C_2$ | 0.8243 | 0.8116 | 0.2242 | 0.6726 | 71.9% | 23.46 |
| | $C_3$ | 0.7941 | 0.8051 | 0.2346 | 0.7038 | 69.3% | 34.90 |

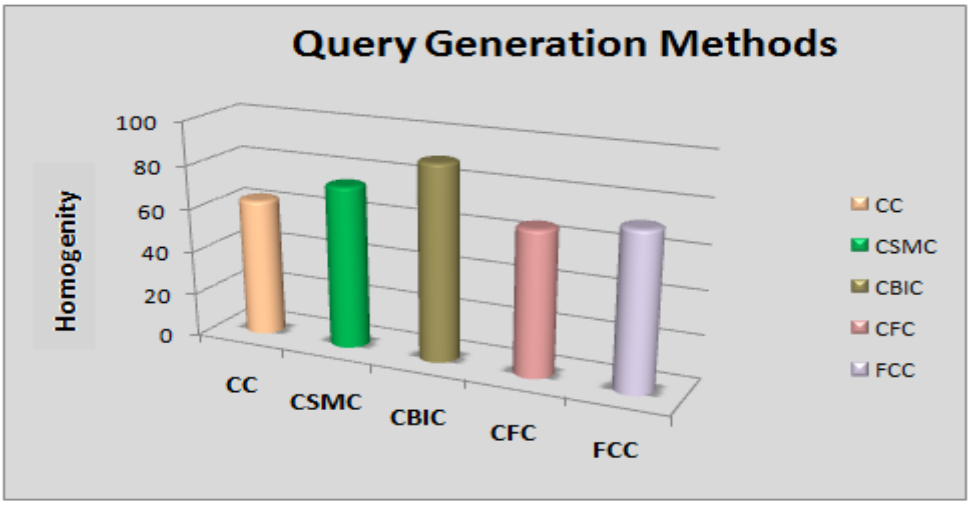



Figure 2(a). Homogeneity for various Query Formation Methods

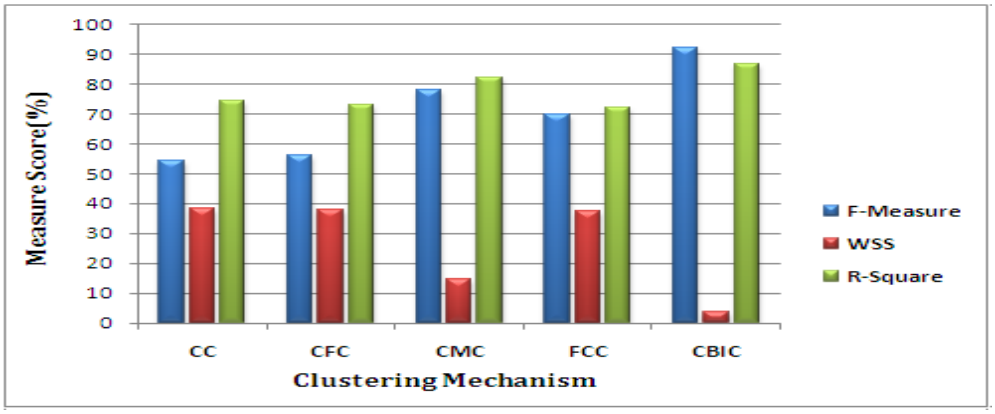



Figure 2(b). Metrics for various QFMs(DD, DH and DA values)

6.2 Comparative analysis among the different QFMs.

By looking at table 2, Fig 2(a) and fig (2b) the following observations are drawn. Among the five methods CBIC has the value of highest F-measure which indicates that quality of the cluster for this is very high. The higher the F measure is the higher the accuracy of the cluster. So CBIC happens to be the method with highest quality because of the reason that it is based on solid statistical fundamentals. This is followed by CSMC which also has statistical backdrop. The minimum WSS for CBIC indicates that the cluster centroid and the items around the centroid have high degree of homogeneity which is also visible from the figure 2(b). Because of the inverse relation between the R-square and WSS , CBIC has the highest value of the R-square followed by the CSMC.

By observing Figure 2(a) the Homogeneity value of the CBIC is high compared to other methods. This indicates that the documents are highly coherent in this method because it is a bounded interval dependent method. The second method with highest homogeneity value is CSMC since it maintains the documents around the mean value of highest ranked document similarities.

Hence both CBIC and CSMC methods are good clustering approaches compared to rest of the methods graded as FCC, CFC and CC with 3, 4 and 5 ranks .measures.

## Conclusion

The experiments show that the clusters with high value of the homogeneity, F- measure and lowest value of the WSS are CBIC and CSMC methods. Both these techniques make use of statistical measures. Among the five the other two namely FCC and CFC are based on concepts BOW. The two techniques will have second footing compared to CBIC and CSMC. The fifth method is a crude one without any foundation hence with poor values of quality metrics. The legal stake

holders with reasonable statistical knowledge and concepts can straight away make use of the first two methods namely CBIC and CSMC. The application of CBIC reduced the documents in each cluster and greatly reduced ambiguity in document retrieval. The CBIC process is an improvement over traditional clustering approach which supports cutting edge performance in minimizing document set and maintaining high coherence among specific crime. Since even best matched cluster holds large set of precedent judgment documents in real world. But clustering provides some idea about distribution of documents over crime facts. This five query formation methods are not sufficient for legal proceedings, there is a need to identify better query with intelligence and soft computing techniques. The future scope of the research is twofold one is to develop an intelligent query and another is to perform other clustering techniques like fuzzy c-means clustering etc.

## References


[1] A. Devender, B. Srinivas and A. Ashok, "Efficient Incremental Clustering of Documents based on Correlation", "Efficient Incremental Clustering of Documents based on Correlation", IJECS, ISSN: 2319-7242, vol. 4, issue 8, 2015.

[2] S. Sowmya Kamath and Monisha Kanakaraj, "Natural Language Processing –based e-News recommender system using information Rxtraction and Domain Clustering", International Journal of Image Mining, Vol. 1, No. 1, NITK, 2015.

[3] S. Joshi, M. P. Deshpande and Thomas Hampp, "Improving the Efficiency of Legal e-Discovery Services using Text Mining Techniques", IEEE, DOI 10.1109, Annual SRII Global Conference, 2011.

[4] G.J. Conard, Ying Zaho, G. Karypis and Khalid Al-Kofahi, "Effective Document Clustering for Large Heteogeneous Law Firm Collections", ICAIL-05, ACM Journal, 1-59593-081-7/05/0006, Iytaly.

[5] Yung-Shen Lin, Jung-Yi Jiang, Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering", IEEE Knowledge & Data Engineering Trans., Vol. 26, July, 2014.

[6] Eui-Hong Han, George Karypis,Vipin Kumar, Bamshad Mobasher,"Clustering Based on Association Rule Hypergraphs", NSF-Grant, CDA-9414015, 1998.

[7] Yanjun Li, Congnan Luo and Soon M. Chung,"Text Clustering with Feature Selection by Using Statistical Data", IEEE Transactions on Knowledge and Data Engineering, VOL. 9, NO. 26, 2008.

[8] G. Suresh Reddy, T.V. Rajini kanth, Anandrao, "Design and Analysis of Novel Similarity Measure for Clustering and Classification of High Dimensional Text Documents", ICCST-Conference, CompSysTech'14.

[9] Zichao Dai, Axin Sun, XU-Ying Liu, "CREST: Cluster based Representation Enrichment for Short Text Classification", NTUDIRP-2010, Singapore

[10] D. Deodhare, G.Sharma, A. Srivastava, Alind Sharma,"Semantically Driven Soft-Clustering of Documents using Lexical Chains",8th International Conference on Natural language processing,2010.

[11] Shashank Paliwal and Vikram Pudi,"Investigating Usage of Text Segmentation and Inter-Passage Similarities to Improve Text Document Clustering",8th International Conference on Machine Learning and Data Mining, 2012, Springer,pp:555-565, Berlin.

[12] Mohit Sharma, Pranjal Singh, "Text Document Clustering and Similarity Measures", IIT-Kanpur, University press, 2013

[13] G. Saidi Reddy, Dr. R.V.Krishnaiah, "Clustering Algorithm with a Novel Similarity Measure", IOSRJCE, ISSN: 2278-0661, Vol. 4, Issue 6, : 37-42, 2012.

[14] Ravi Kumar V, K. Raghuveer, "Legal Documents Clustering and Summarization using Hierarchical Latent Dirichlet Allocation", IAESIJ-AI, ISSN:2252-8938, pp: 27-35, 2013.

[15] J. Pallavi, D. Dharmadhikari, "Clustering with Multi-Viewpoint Based Similarity Measure: An Overview", International Journal of Engineering Inventions, ISSN: 2278-7461, Vol. 1, pp: 01-05, 2012.

[16] L. Ertoz, Micheal Steinbach, Vipin Kumar, "A New Shared Nearest Neighbor Clustering Algorithm and its Applications", University Grant Project, Minnesota, USA, 2012.

[17] P. Ramesh babu, M. Nagabhushana Rao, "Difference Vectors Similarity Measure in Text Clustering", Journal of Computer Science & Software Engineering, Vol. 1, 2014.

[18] S. Vijayalakshmi, D. Manimegalai, "Text Document Clustering with Flocking Algorithm using Secific Crimes Judgment Corpus", AJIT, ISSN:1682-3915, pp:21-28, 2014.

[19] Khaled M. Hammouda and Mohamad S. Kamel, "Incremental Document Clustering Using Cluster Similarity Histograms", University Press, University of Watreloo, Canada - 2001.

[20] Pawan Lingras, Min Chen and Duoqian Miao, "Precision of Rough Set Clustering", RSCTC-Conference, pp: 369-378, Canada,2008.

[20] Khaled Hammouda and Mohamad Kamel, "Collaborative Document Clustering", MIRC-Project, University Press, University of Waterloo, Canada, 2006.

[21] S. Paliwal and V. Pudi, "Investigating Usage of Text Segmentation and Inter-Passage Similarities to Improve Text Document Clustering", 8th International Conference on MLDM, pp: 555-565, 2012.

[22] R. Deshpande, K. Vaze, S. Rathod, T. Jarhad, "Comparative Study of Document Similarity Algorithms and Clustering Algorithms for Sentiment Analysis", IJETTCS, ISSN: 2278-6856, Vol. 3, 2014.

[23] S. V. Goal Rao and A. Bhanu Prasad, "Space and Cosine Similarity measures for Text Document Clustering", IJERT, ISSN: 2278-0181, Vol. 2, 2013.

[24] J.G. Konard, Ying Zaho, G. Karypis, "Effective Document Clustering for Large Heterogeneous Law Firm Collections", ICAIL Conference, Italy-2005.

[25] Rupali Sunil Wagh, "Exploratory Analysis of Legal Documents using Unsupervised Text Mining Techniques", IJERT, Vol. 3, Issue 2, Feb-2014.

[26] B.V.RamaKrishna, B.Basaveswar Rao, K. Gangadhar Rao and K. Chandan, "An Enumarative Framework for Extraction of Bag-of-Words from Legal Documents", AJCSIT, ISSN: 2249-5126, PP: 62-66, DEC-2015.

[27] B.V.RamaKrishna, B.Basaveswar Rao, K. Gangadhar Rao and K. Chandan, "Enhancement of Bag-of-Words for Legal documents", IOSR, AUG- 2016.

[28] B.V.RamaKrishna, B.Basaveswar Rao, K. Gangadhar Rao and K. Chandan, "Judicial Precedents Search Process supported by Similarity Measures", 2016.

[29] *http://www. eric.univ-lyon2.fr/~Ricco/Tanagra/en/tanagra.html*, Tanagra Open Software, developed by Ricco Rakotomalala, Lumière University Lyon

[30] *https://www.dezyre.com/data-science-in-r-programming-tutorial/k-means-clustering-techniques-tutorial*, DeZyre® , Industrial Experts online tutorial service, USA

## Annexure - I

1. Comprehensive Clustering of Precedents

K-Means 1

Parameters

| K-Means parameters | |
|---|---|
| Clusters | 4 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

Results

### Global evaluation

| | |
|---|---|
| Within Sum of Squares | 154.2284 |
| Total Sum of Squares | 1000.0000 |
| R-Square | 0.8458 |

### R-Square for each attempt

| Number of trials | 5 |
|---|---|
| Trial | R-square |
| 1 | 0.845769 |
| 2 | 0.845769 |
| 3 | 0.845713 |
| 4 | 0.845357 |
| 5 | 0.845772 |

### Cluster size and WSS

| Clusters | 4 | | |
|---|---|---|---|
| Cluster | Description | Size | WSS |
| cluster n°1 | c_kmeans_1 | 185 | 55.1818 |
| cluster n°2 | c_kmeans_2 | 82 | 26.1095 |
| cluster n°3 | c_kmeans_3 | 38 | 29.8832 |
| cluster n°4 | c_kmeans_4 | 195 | 43.0540 |

### Cluster centroids

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 | Cluster n°4 |
|---|---|---|---|---|
| similarity | 54.080708 | 37.870829 | 69.261218 | 45.697841 |
| LD | 47.043349 | 30.420073 | 63.930578 | 39.010224 |

*Use GROUP CHARACTERIZATION for detailed comparisons*

Figure 3(a): Comprehensive Clustering of precedents

DOI: 10.18535/ijecs/v6i1.08

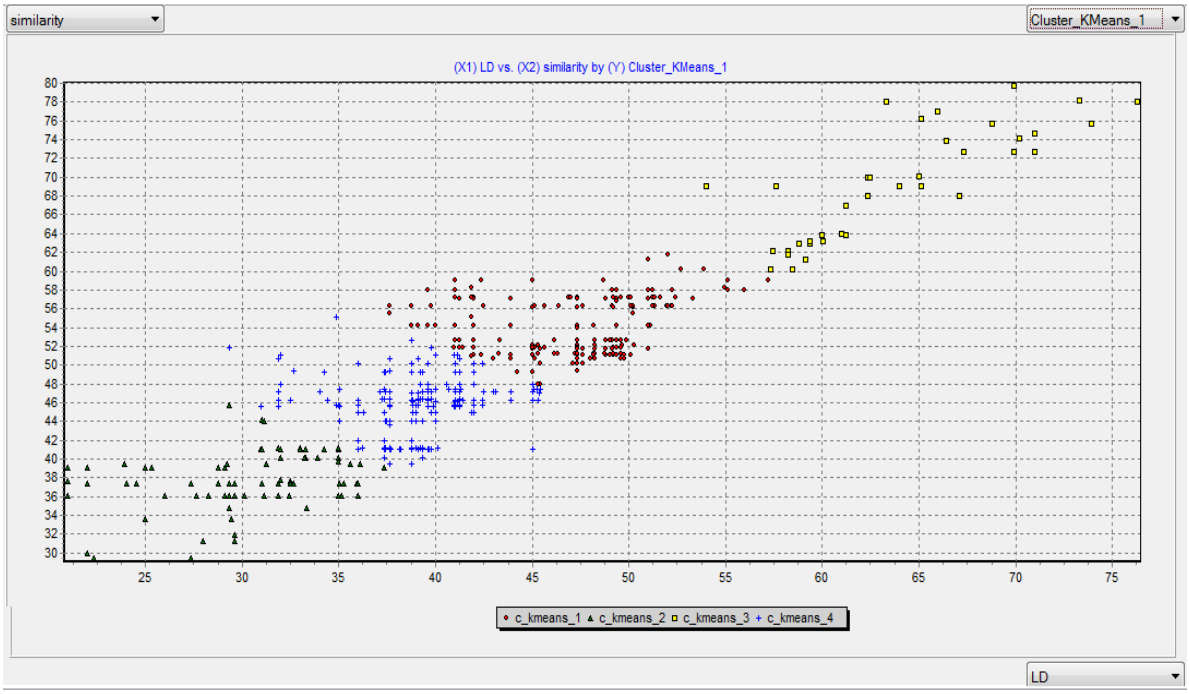
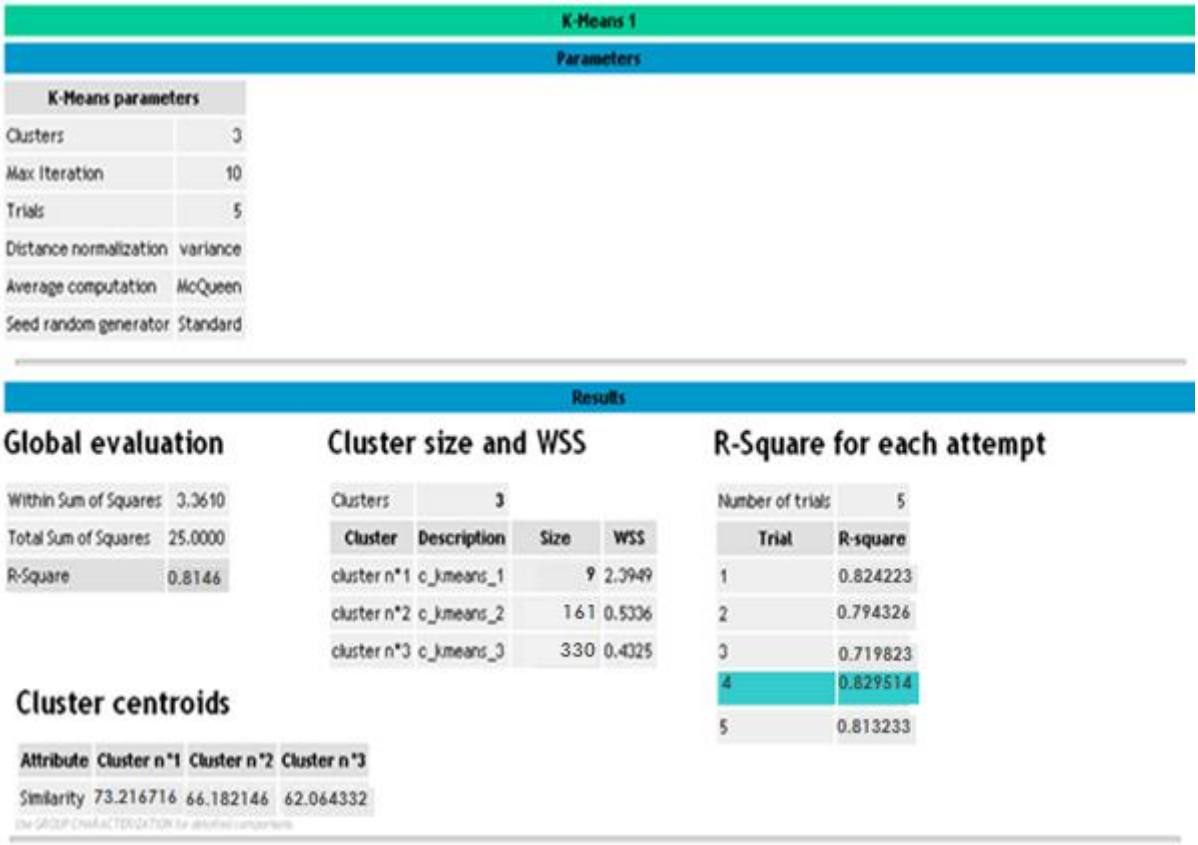



Figure 3(b): Comprehensive Clustering Graph for Precedents

2. Crime specific Mean Clustering (CMC)



K-Means 1

Parameters

| K-Means parameters | |
|---|---|
| Clusters | 3 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

Results

Global evaluation

| | |
|---|---|
| Within Sum of Squares | 3.3610 |
| Total Sum of Squares | 25.0000 |
| R-Square | 0.8146 |

Cluster size and WSS

| Clusters | 3 | | |
|---|---|---|---|
| Cluster | Description | Size | WSS |
| cluster n°1 | c_kmeans_1 | 9 | 2.3949 |
| cluster n°2 | c_kmeans_2 | 161 | 0.5336 |
| cluster n°3 | c_kmeans_3 | 330 | 0.4325 |

R-Square for each attempt

| Number of trials | 5 |
|---|---|
| Trial | R-square |
| 1 | 0.824223 |
| 2 | 0.794326 |
| 3 | 0.719823 |
| 4 | 0.829514 |
| 5 | 0.813233 |

Cluster centroids

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 |
|---|---|---|---|
| Similarity | 73.216716 | 66.182146 | 62.064332 |

Figure 4(a). Clustering with Mean of top 10 ranked similarity measures [CF1]

DOI: 10.18535/ijecs/v6i1.08

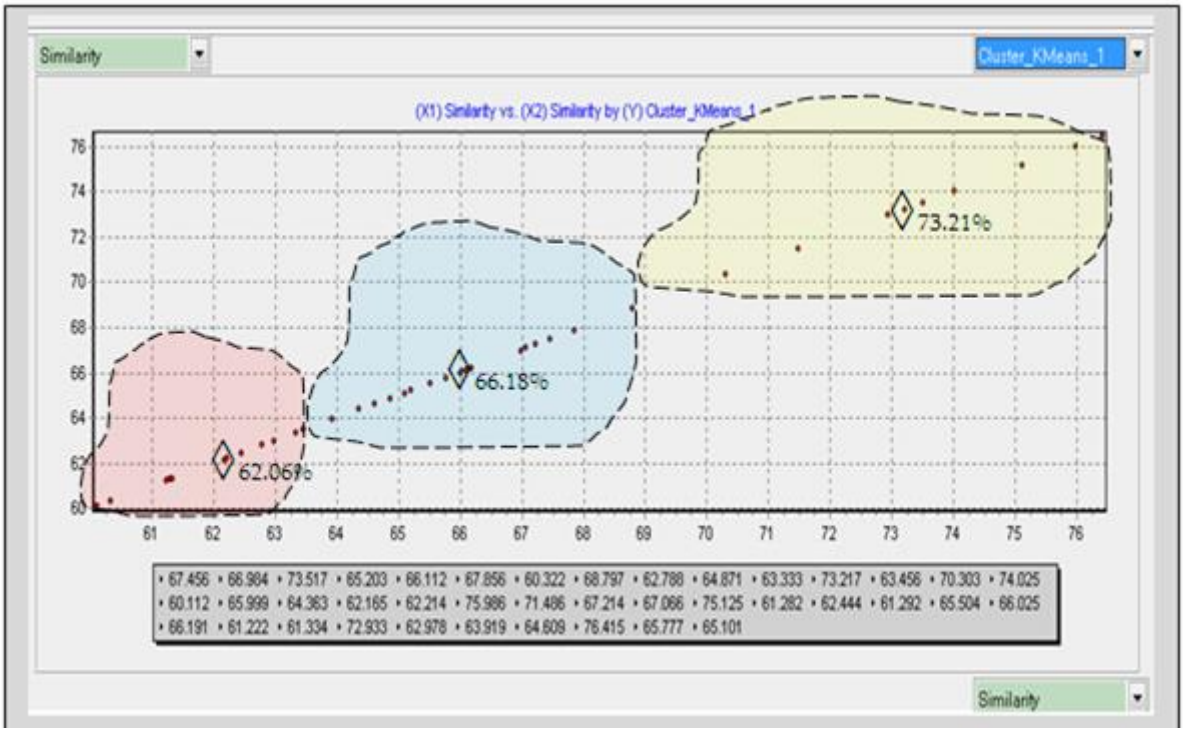



Figure 4(b). Clustering with Mean of top 10 ranked similarity measures [CF1]

**K-Means 1**

**Parameters**

| K-Means parameters | |
|---|---|
| Clusters | 3 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

**Results**

**Global evaluation**

| | |
|---|---|
| Within Sum of Squares | 3.3610 |
| Total Sum of Squares | 25.0000 |
| R-Square | 0.8067 |

**Cluster size and WSS**

| Clusters | 3 | | |
|---|---|---|---|
| Cluster | Description | Size | WSS |
| cluster n°1 | c_kmeans_1 | 22 | 14.422 |
| cluster n°2 | c_kmeans_2 | 218 | 6.4223 |
| cluster n°3 | c_kmeans_3 | 260 | 3.4223 |

**R-Square for each attempt**

| Number of trials | 5 |
|---|---|
| Trial | R-square |
| 1 | 0.724243 |
| 2 | 0.794226 |
| 3 | 0.819823 |
| 4 | 0.829514 |
| 5 | 0.803233 |

**Cluster centroids**

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 |
|---|---|---|---|
| Similarity | 55.169999 | 65.173346 | 56.164332 |

Figure 4(c). Clustering with Mean of top 10 ranked similarity measures [CF2]

DOI: 10.18535/ijecs/v6i1.08

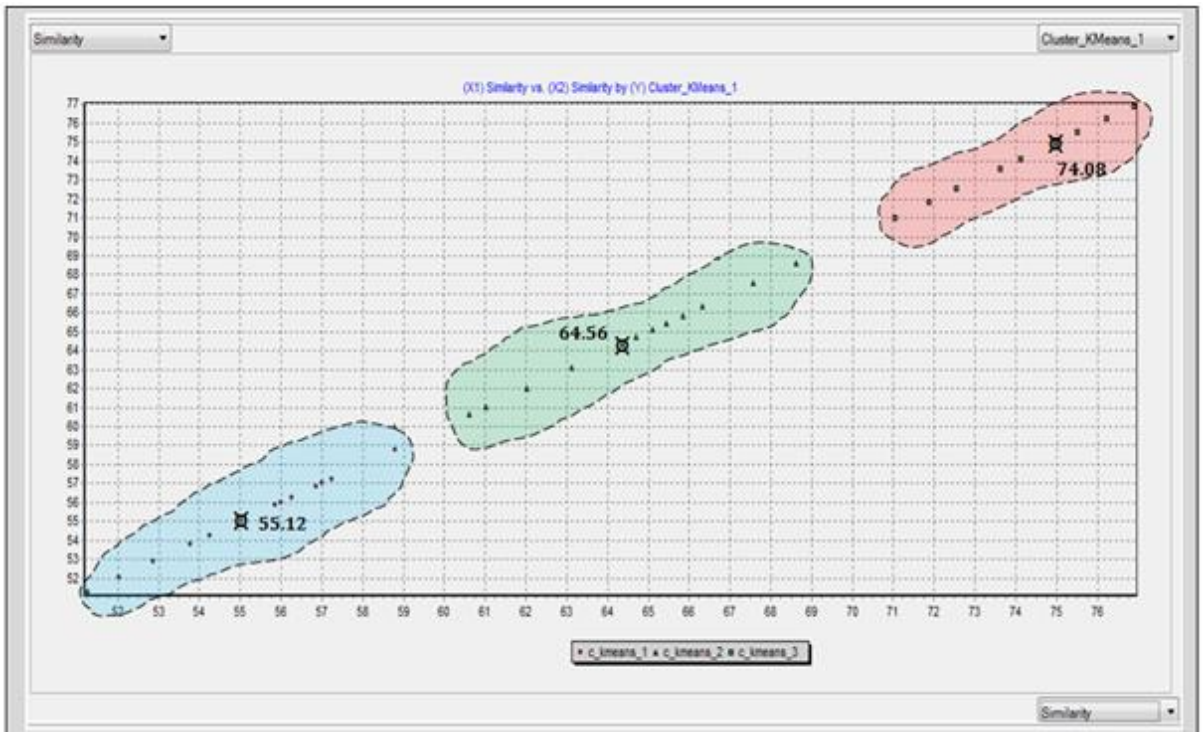



Figure 4(d). Clustering with Mean of top 10 ranked similarity measures [CF2]

**K-Means 1**

**Parameters**

| K-Means parameters | |
|---|---|
| Clusters | 3 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

**Results**

**Global evaluation**

| | |
|---|---|
| Within Sum of Squares | 3.3610 |
| Total Sum of Squares | 25.0000 |
| R-Square | 0.7436 |

**Cluster size and WSS**

| Clusters | 3 | | |
|---|---|---|---|
| Cluster | Description | Size | WSS |
| cluster n°1 | c_kmeans_1 | 20 | 2.3949 |
| cluster n°2 | c_kmeans_2 | 164 | 0.5336 |
| cluster n°3 | c_kmeans_3 | 316 | 0.4325 |

**R-Square for each attempt**

| Number of trials | 5 |
|---|---|
| Trial | R-square |
| 1 | 0.764223 |
| 2 | 0.714326 |
| 3 | 0.779823 |
| 4 | 0.749514 |
| 5 | 0.865560 |

**Cluster centroids**

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 |
|---|---|---|---|
| Similarity | 72.994326 | 63.897667 | 57.307751 |

Figure 4(e). Clustering with Mean of top 10 ranked similarity measures [CF3]

DOI: 10.18535/ijecs/v6i1.08

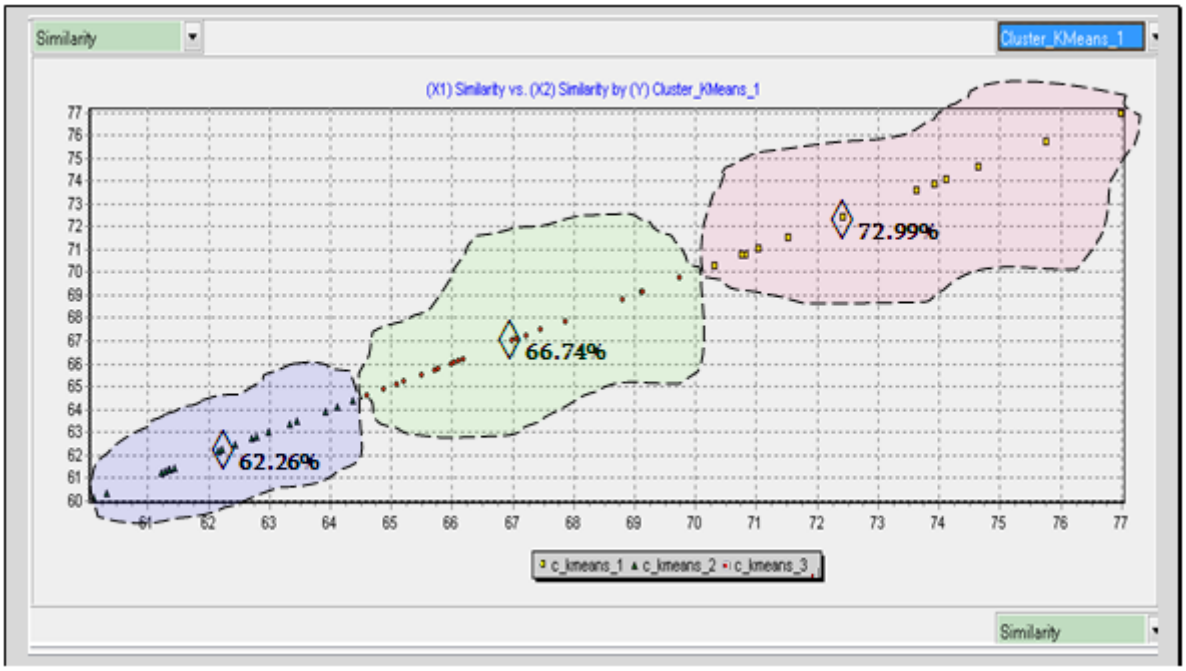
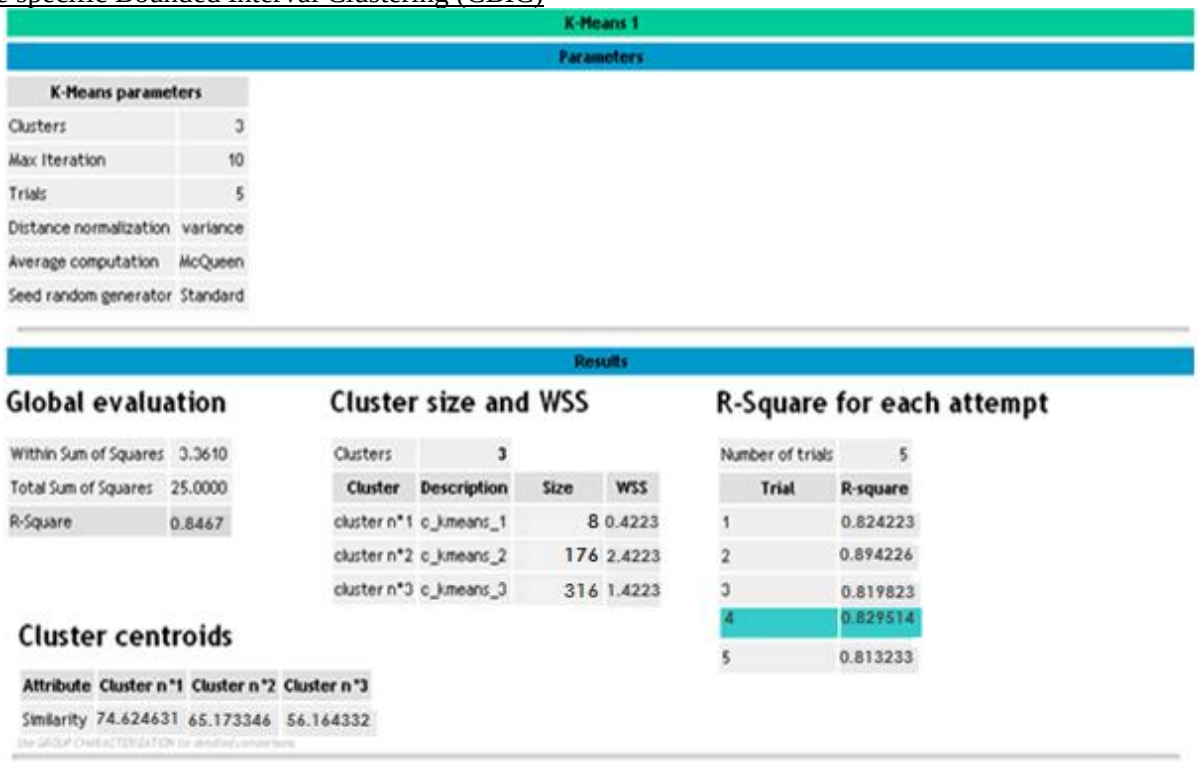



Figure 4(f). Clustering with Mean of top 10 ranked similarity measures [CF3]

3. Crime specific Bounded Interval Clustering (CBIC)



K-Means 1

Parameters

| K-Means parameters | |
|---|---|
| Clusters | 3 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

Results

**Global evaluation**

| | |
|---|---|
| Within Sum of Squares | 3.3610 |
| Total Sum of Squares | 25.0000 |
| R-Square | 0.8467 |

**Cluster size and WSS**

| Clusters | 3 | | |
|---|---|---|---|
| Cluster | Description | Size | WSS |
| cluster n°1 | c_kmeans_1 | 8 | 0.4223 |
| cluster n°2 | c_kmeans_2 | 176 | 2.4223 |
| cluster n°3 | c_kmeans_3 | 316 | 1.4223 |

**R-Square for each attempt**

| Number of trials | 5 |
|---|---|
| Trial | R-square |
| 1 | 0.824223 |
| 2 | 0.894226 |
| 3 | 0.819823 |
| 4 | 0.829514 |
| 5 | 0.813233 |

**Cluster centroids**

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 |
|---|---|---|---|
| Similarity | 74.624631 | 65.173346 | 56.164332 |

Figure 5(a) BIM Cluster for Dowry-Death judgments

DOI: 10.18535/ijecs/v6i1.08

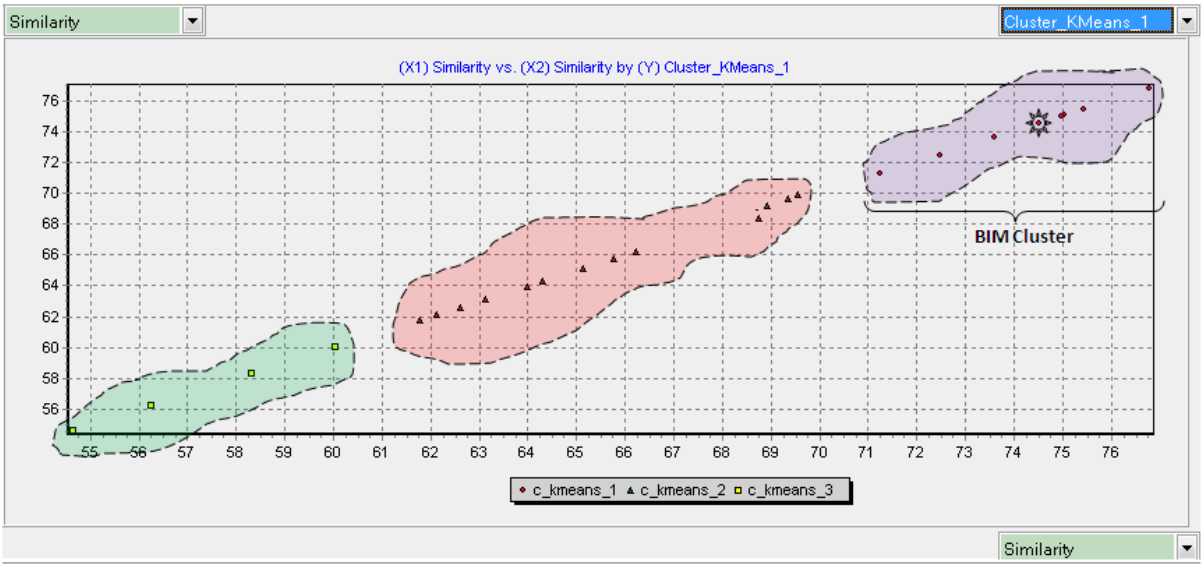



Figure 5(b) BIM Cluster for Dowry-Death judgments

**K-Means 1**

**Parameters**

| K-Means parameters | |
|---|---|
| Clusters | 3 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

**Results**

**Global evaluation**

| | |
|---|---|
| Within Sum of Squares | 3.3610 |
| Total Sum of Squares | 25.0000 |
| R-Square | 0.8656 |

**Cluster size and WSS**

| Clusters | 3 | | |
|---|---|---|---|
| Cluster | Description | Size | WSS |
| cluster n°1 | c_kmeans_1 | 12 | 2.3949 |
| cluster n°2 | c_kmeans_2 | 296 | 0.5336 |
| cluster n°3 | c_kmeans_3 | 192 | 0.4325 |

**R-Square for each attempt**

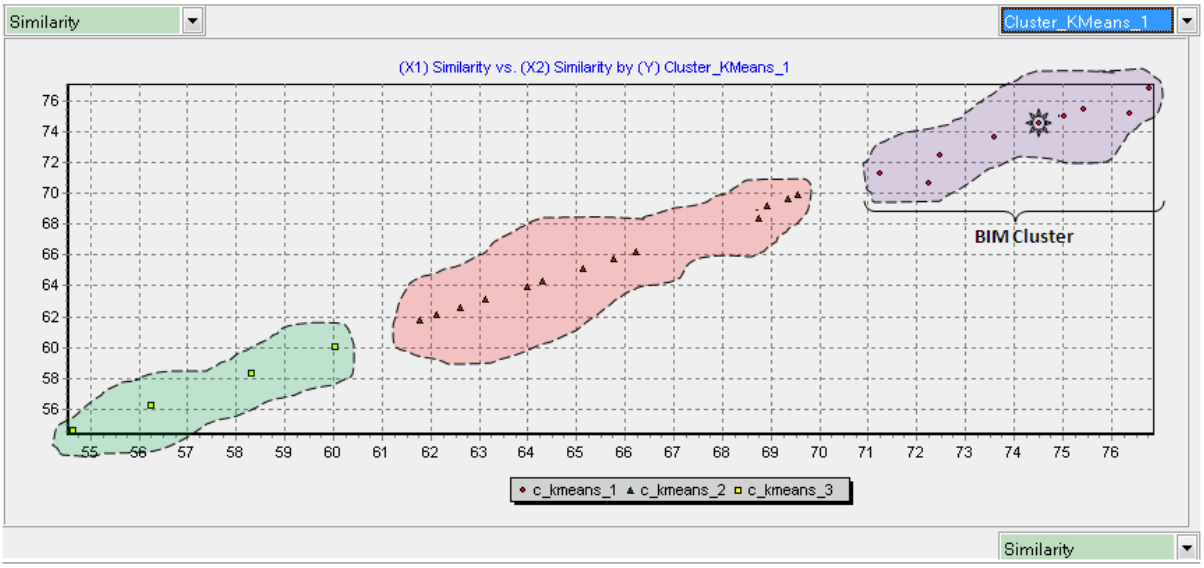
| Number of trials | 5 |
|---|---|
| Trial | R-square |
| 1 | 0.851518 |
| 2 | 0.861760 |
| 3 | 0.862940 |
| 4 | 0.865560 |
| 5 | 0.865560 |

**Cluster centroids**

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 |
|---|---|---|---|
| Similarity | 74.601237 | 63.897667 | 57.307751 |

Use GROUP CHARACTERIZATION for detailed comparisons

Figure 5(c) BIM Cluster for Dowry-Harassment judgments



**DOI: 10.18535/ijecs/v6i1.08**

Figure 5(d) BIM Cluster for Dowry-Harassment judgments

**K-Means 1**

**Parameters**

| K-Means parameters | |
|---|---|
| Clusters | 3 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

**Results**

**Global evaluation**

| | |
|---|---|
| Within Sum of Squares | 6.6871 |
| Total Sum of Squares | 50.0000 |
| R-Square | 0.8663 |

**Cluster size and WSS**

| Clusters | 3 | | |
|---|---|---|---|
| **Cluster** | **Description** | **Size** | **WSS** |
| cluster n°1 | c_kmeans_1 | 12 | 1.4513 |
| cluster n°2 | c_kmeans_2 | 212 | 2.8850 |
| cluster n°3 | c_kmeans_3 | 276 | 2.3508 |

**R-Square for each attempt**

| Number of trials | 5 |
|---|---|
| **Trial** | **R-square** |
| 1 | 0.866040 |
| 2 | 0.866258 |
| 3 | 0.866040 |
| 4 | 0.866040 |
| 5 | 0.866040 |

**Cluster centroids**

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 |
|---|---|---|---|
| Similarity | 72.550667 | 52.346867 | 62.783696 |

Figure 5(e) BIM Cluster for Dowry-Acceptance Judgments

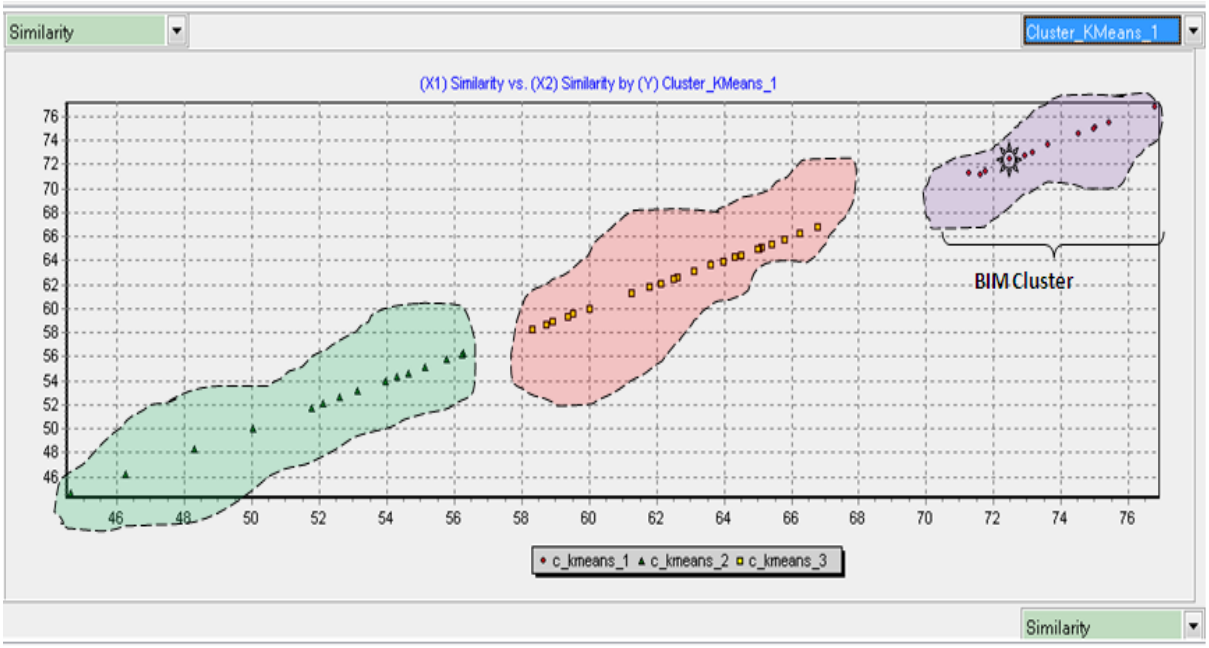



Figure 5(f) BIM Cluster for Dowry-Acceptance Judgments

4. Crime Facts Clustering (CFC)

DOI: 10.18535/ijecs/v6i1.08

**K-Means 1**

**Parameters**

| K-Means parameters | |
|---|---|
| Clusters | 3 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

**Results**

### Global evaluation

| | |
|---|---|
| Within Sum of Squares | 238.6555 |
| Total Sum of Squares | 1000.0000 |
| R-Square | 0.7613 |

### Cluster size and WSS

| Clusters | 3 | | |
|---|---|---|---|
| **Cluster** | **Description** | **Size** | **WSS** |
| cluster n°1 | c_kmeans_1 | 237 | 125.9789 |
| cluster n°2 | c_kmeans_2 | 39 | 31.7412 |
| cluster n°3 | c_kmeans_3 | 224 | 80.9354 |

### R-Square for each attempt

| Number of trials | 5 |
|---|---|
| **Trial** | **R-square** |
| 1 | 0.761282 |
| 2 | 0.761262 |
| 3 | 0.758479 |
| 4 | 0.761276 |
| 5 | 0.761345 |

### Cluster centroids

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 |
|---|---|---|---|
| Similarity | 42.448283 | 68.998147 | 53.134703 |
| Repository | 35.569860 | 63.758828 | 46.058794 |

Use GROUP CHARACTERIZATION for detailed comparisons

Figure 6(a) CFC Cluster for Dowry-Death Judgments

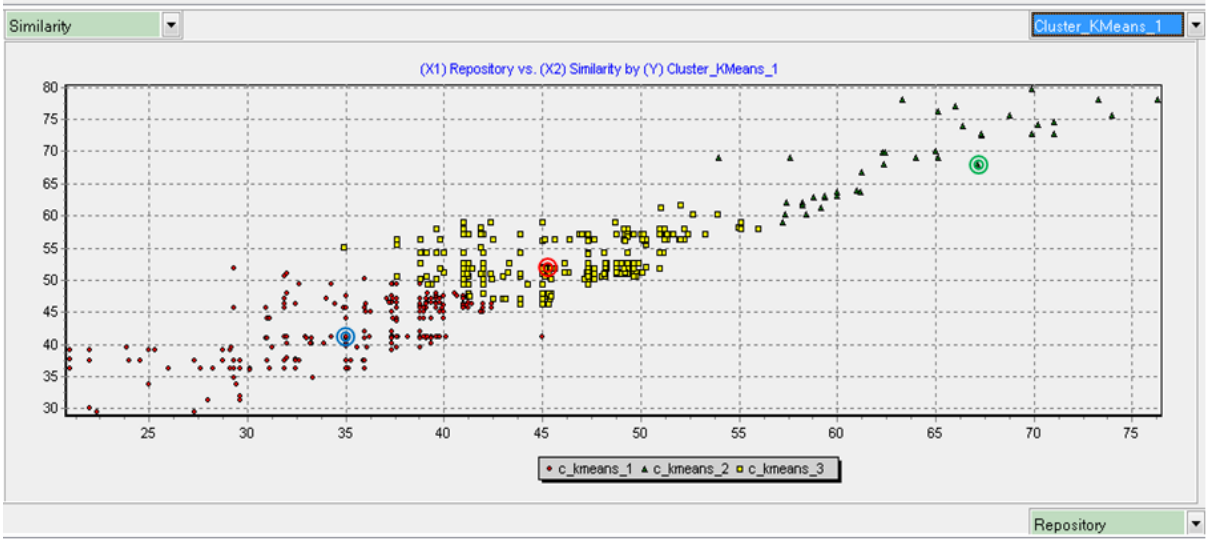



Figure 6(b) CFC Cluster for Dowry-Death Judgments

**K-Means 1**

**Parameters**

| K-Means parameters | |
|---|---|
| Clusters | 3 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

**Results**

### Global evaluation

| | |
|---|---|
| Within Sum of Squares | 236.0763 |
| Total Sum of Squares | 1000.0000 |
| R-Square | 0.7639 |

### Cluster size and WSS

| Clusters | 3 | | |
|---|---|---|---|
| **Cluster** | **Description** | **Size** | **WSS** |
| cluster n°1 | c_kmeans_1 | 221 | 77.4220 |
| cluster n°2 | c_kmeans_2 | 50 | 48.3154 |
| cluster n°3 | c_kmeans_3 | 229 | 110.3389 |

### R-Square for each attempt

| Number of trials | 5 |
|---|---|
| **Trial** | **R-square** |
| 1 | 0.763808 |
| 2 | 0.763884 |
| 3 | 0.763787 |
| 4 | 0.763800 |
| 5 | 0.763924 |

### Cluster centroids

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 |
|---|---|---|---|
| Similarity | 53.089989 | 67.722929 | 42.560306 |
| Repository | 46.067095 | 64.592001 | 35.691673 |

Use GROUP CHARACTERIZATION for detailed comparisons

Figure 6(c) CFC Cluster for Dowry-Harassment Judgments

DOI: 10.18535/ijecs/v6i1.08

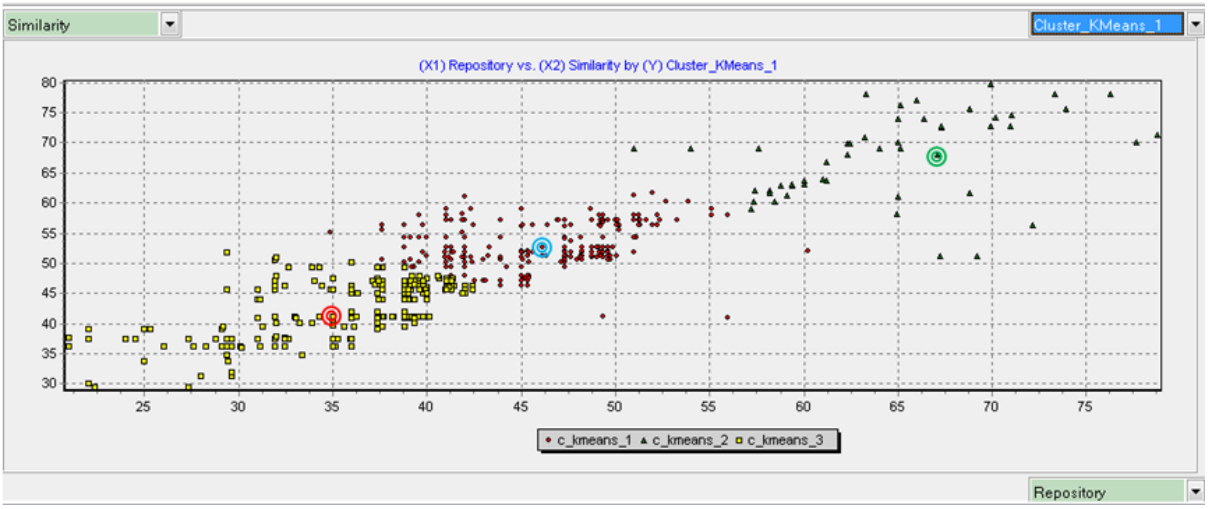



Figure 6(d) CFC Cluster for Dowry-Harassment Judgments

**K-Means 1**

**Parameters**

| K-Means parameters | |
|---|---|
| Clusters | 3 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

**Results**

**Global evaluation**

| | |
|---|---|
| Within Sum of Squares | 255.7475 |
| Total Sum of Squares | 1000.0000 |
| R-Square | 0.7443 |

**Cluster size and WSS**

| Clusters | 3 | | |
|---|---|---|---|
| **Cluster** | **Description** | **Size** | **WSS** |
| cluster n°1 | c_kmeans_1 | 41 | 36.9021 |
| cluster n°2 | c_kmeans_2 | 224 | 88.1128 |
| cluster n°3 | c_kmeans_3 | 235 | 130.7326 |

**R-Square for each attempt**

| Number of trials | 5 |
|---|---|
| **Trial** | **R-square** |
| 1 | 0.743967 |
| 2 | 0.744252 |
| 3 | 0.739278 |
| 4 | 0.743103 |
| 5 | 0.744157 |

**Cluster centroids**

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 |
|---|---|---|---|
| Similarity | 68.847070 | 53.380109 | 42.482851 |
| Repository | 62.944057 | 46.091701 | 35.866271 |

*Use GROUP CHARACTERIZATION for detailed comparisons*

Figure 6(e) CFC Cluster for Dowry-Acceptance Judgments

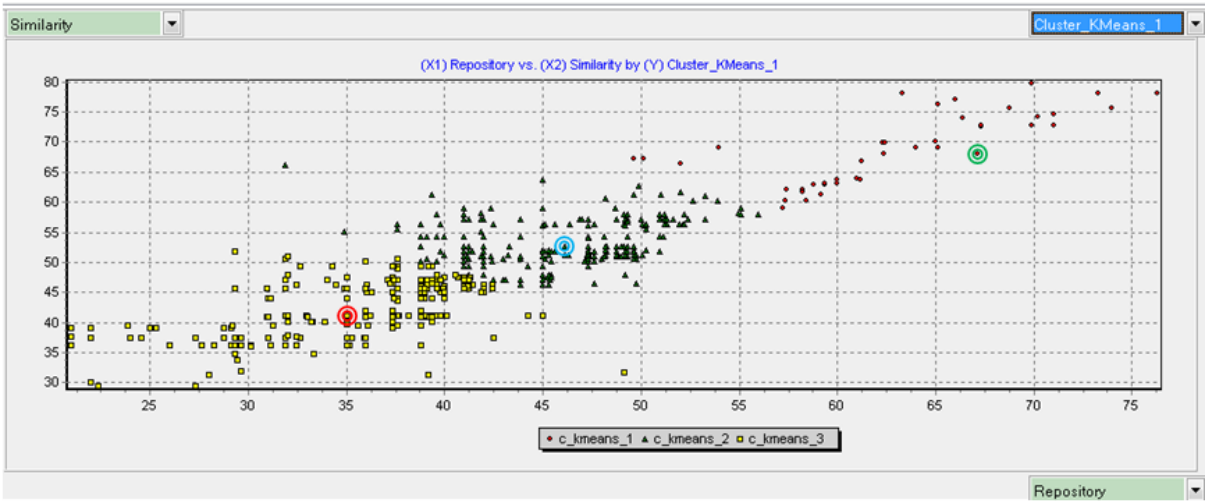



Figure 6(f) CFC Cluster for Dowry-Acceptance Judgments

5. FCC Clustering results for FIRs (DD, DH and DA)

DOI: 10.18535/ijecs/v6i1.08

**K-Means 1**

**Parameters**

| K-Means parameters | |
|---|---|
| Clusters | 3 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

**Results**

## Global evaluation

| | |
|---|---|
| Within Sum of Squares | 116.3416 |
| Total Sum of Squares | 392.0000 |
| R-Square | 0.7032 |

## Cluster size and WSS

| Clusters | 3 | | |
|---|---|---|---|
| **Cluster** | **Description** | **Size** | **WSS** |
| cluster n°1 | c_kmeans_1 | 309 | 56.1182 |
| cluster n°2 | c_kmeans_2 | 126 | 52.7139 |
| cluster n°3 | c_kmeans_3 | 65 | 48.5095 |

## R-Square for each attempt

| Number of trials | 5 |
|---|---|
| **Trial** | **R-square** |
| 1 | 0.701784 |
| 2 | 0.703210 |
| 3 | 0.701592 |
| 4 | 0.669521 |
| 5 | 0.701667 |

## Cluster centroids

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 |
|---|---|---|---|
| ItemSet | 4.981132 | 1.269231 | 4.107692 |
| SM | 57.178585 | 59.782167 | 71.737600 |

Use GROUP CHARACTERIZATION for detailed comparisons

Figure 7(a): Clustering Dowry-Death legal documents using FCC

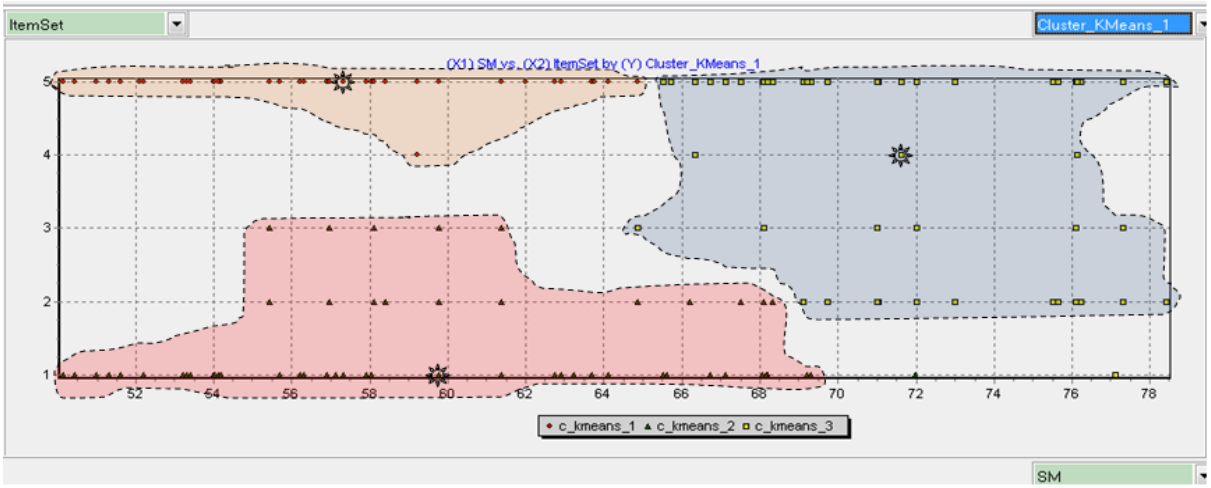



Figure 7(b): Clustering Dowry-Death legal documents using FCC graph

DOI: 10.18535/ijecs/v6i1.08

**K-Means 1**

**Parameters**

| K-Means parameters | |
|---|---|
| Clusters | 3 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

**Results**

### Global evaluation

| | |
|---|---|
| Within Sum of Squares | 102.2043 |
| Total Sum of Squares | 366.0000 |
| R-Square | 0.7208 |

### Cluster size and WSS

| Clusters | 3 | | |
|---|---|---|---|
| **Cluster** | **Description** | **Size** | **WSS** |
| cluster n°1 | c_kmeans_1 | 61 | 41.8067 |
| cluster n°2 | c_kmeans_2 | 209 | 12.9154 |
| cluster n°3 | c_kmeans_3 | 230 | 47.4822 |

### R-Square for each attempt

| Number of trials | 5 |
|---|---|
| **Trial** | **R-square** |
| 1 | 0.720753 |
| 2 | 0.720753 |
| 3 | 0.720753 |
| 4 | 0.720729 |
| 5 | 0.646296 |

### Cluster centroids

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 |
|---|---|---|---|
| ItemSet | 4.262295 | 4.980769 | 1.242857 |
| SM | 71.456524 | 56.873269 | 59.743143 |

Use GROUP CHARACTERIZATION for detailed comparisons

Figure 7(c): Clustering Dowry-Harassment legal documents using FCC

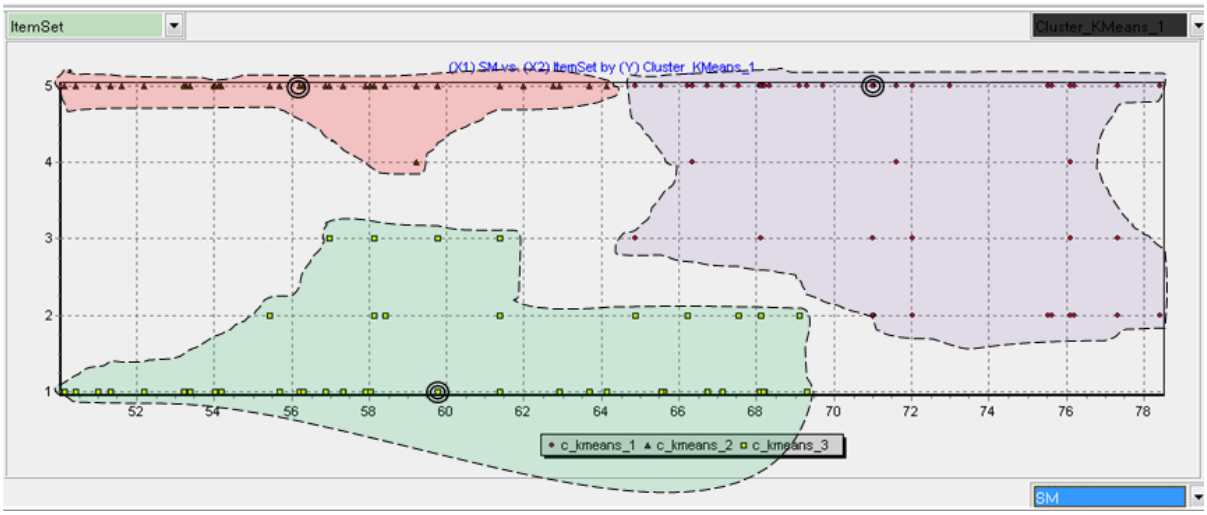



Figure 7(d): Clustering Dowry-Harassment legal documents using FCC

DOI: 10.18535/ijecs/v6i1.08

**K-Means 1**

**Parameters**

| K-Means parameters | |
|---|---|
| Clusters | 3 |
| Max Iteration | 10 |
| Trials | 5 |
| Distance normalization | variance |
| Average computation | McQueen |
| Seed random generator | Standard |

**Results**

## Global evaluation

| | |
|---|---|
| Within Sum of Squares | 94.9113 |
| Total Sum of Squares | 292.0000 |
| R-Square | 0.6750 |

## R-Square for each attempt

| Number of trials | 5 |
|---|---|

| Trial | R-square |
|---|---|
| 1 | 0.654295 |
| 2 | 0.654926 |
| 3 | 0.674433 |
| 4 | 0.654926 |
| 5 | 0.674961 |

## Cluster size and WSS

| Clusters | 3 | | |
|---|---|---|---|
| **Cluster** | **Description** | **Size** | **WSS** |
| cluster n°1 | c_kmeans_1 | 256 | 38.5314 |
| cluster n°2 | c_kmeans_2 | 190 | 13.4671 |
| cluster n°3 | c_kmeans_3 | 54 | 42.9127 |

## Cluster centroids

| Attribute | Cluster n°1 | Cluster n°2 | Cluster n°3 |
|---|---|---|---|
| ItemSet | 1.363636 | 4.756757 | 3.962963 |
| SM | 59.392255 | 56.864865 | 71.197703 |

Figure 7(e): Clustering Dowry-Acceptance legal documents using FCC

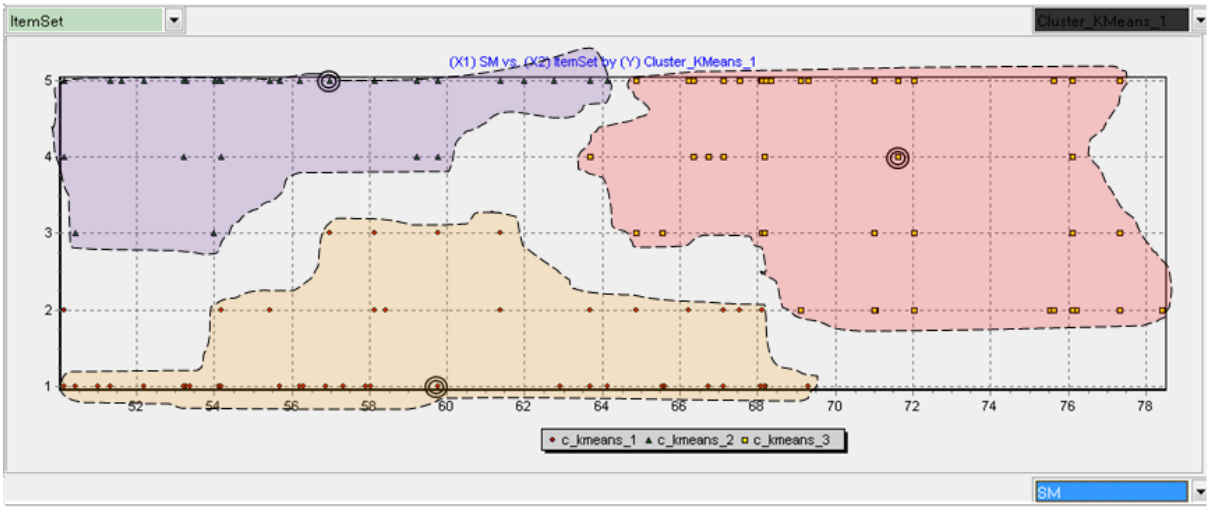



Figure 7(f): Clustering Dowry-Acceptance legal documents using FCC