Statistical and Semantic Feature Selection for Text Clustering

Asmaa Benghabrit[1], Brahim Ouhbi[1], Hicham Behja[1], Bouchra Frikh[2]
[1]LM2I laboratory ENSAM
Moulay Ismaïl University, Marjane II
B.P. 4024 Meknès, Morocco
[2]LTTI laboraory EST-Fès
Moulay Abdellah University, B.P. 1796 Atlas Fès
Fès, Morocco

ABSTRACT: *Organizing textual documents by categorizing them is important and beneficial for information retrieval; but when it comes to clustering documents containing a huge number of terms, the task become challenged. Therefore, selecting effective features is essential for reducing the feature space dimensionality and improving the clustering performances. While numerous methods have been developed for this purpose, fewer techniques considered the semantic knowledge that can be incorporate into the clustering process. This paper proposes first a new semantic feature selection method SIM based on the mutual information metric, and second a novel two phase clustering mechanism. The statistical feature selection method CHIR integrates into the frequency clustering stage and then our technique SIM is used in the second stage to pilot the semantic categorization. The content based analysis allows enhancing the frequency clustering by taking the semantic relationships between the features into account. The successful evaluation of our approach demonstrates its relevancy in catching statistical and semantic pertinent features that enable better clustering accuracy in terms of F-measure and purity.*

Keywords: Document Clustering, Feature Selection Methods, Semantic Analysis, Performance Measures

Received: 2 April 2013, Revised 5 May 2013, Accepted 9 May 2013

1. Introduction

Nowadays thanks to the availability of online information, retrieving information about any need can be automatically performed by just a keystroke or a mouse click. This textual revolution would be truly beneficial for the web users only if this information is organized. To do this, approaches such as text mining have been developed. Text mining is the analysis of natural language texts that allows an efficient extraction of relevant information and knowledge [1]. This process includes many technologies; we are interested in one of its powerful methods which is document clustering. Text clustering is the task of automatically structuring documents into homogonous categories in a way that analogous documents are grouped in the same cluster; which means that texts in each group are similar to each other and dissimilar to texts belonging to the rest of the groups. Hence, mixed documents written in natural language are represented in a structured way; this is helps and guides the users during their information searching and understanding [2].

In text clustering, words and phrases in unstructured documents have to be converted into numerical values to enable the application of the clustering process. To do this, documents are generally represented using the vector space model [3].

Nevertheless, this representation causes a high dimensionality problem which renders the employment of clustering algorithms questionable. Firstly, distance measures become meaningless in such case. The more the feature space has dimensions the more the documents spread out until they are all almost equidistant from each other [4]. As a result, clustering methods can't work efficiently. Another reason, that clustering methods struggle with high dimensionality, is the existence of aberrant features. In fact, the feature space contains some irrelevant and redundant elements which can degrade the clustering results especially when they outnumber the relevant ones [3] [4] [5]. To deal with this problem, dimension reduction is imperative for efficiently manipulating the massive quantity of data.

Dimensionality reduction is an effective approach for downsizing the feature space into a manageable size. It consists on transforming an initial dataset X with D as dimension size into a new dataset Y with a dimension size $d < D$ (often $d << D$); the transformation is carried whereas conserving the maximum amount of the original information. Many techniques have been researched on for this purpose. Even if all these methods have the same goal, they are divided into two categories according to the nature of the resulting features. More specifically, feature extraction and feature selection are the two approaches allowing performing the dimensionality reduction. Feature extraction refers to the mapping of the original high-dimensional space into a lower dimensional space; in other words the method creates a subset of new features by combinations of the existing ones. Consequently, the information provided by the irrelevant features is maintained and the interpretation of the new attributes is difficult since they are complexes features. In contrast to this, feature selection picks out an optimal subset containing the more informative features according to an objective function. This techniques not only reduces the high dimensionality of the feature space, but also makes the application of learning methods significantly easier [6], provides more accurate and efficient clustering [5] [7], and can avoid the overfitting phenomenon [3] which weakens the prediction capacities of the model. All these benefits lead to conclude that feature selection is the most appropriate and suitable method for solving efficiently the problem of high dimensionality.

Various feature selection methods have been developed including document frequency, Chi-square, mutual information, information gain, GSS coefficient, correlation coefficient and Odds ratio. However fewer take the fact that features in the dataset may be semantically related into consideration. We focus in this work on this issue and hence present a novel method SIM for selecting features using the semantic relations between the words. In the second part of this study, we present a new sequential clustering that combines a statistical and semantic analysis. It is an extinction of a clustering algorithm TCFS performed in [5] which is piloted by a frequency method CHIR based on the chi-square statistic. Once the TCFS clustering is achieved, we incorporate our SIM method which is based on the well known mutual information measure into the semantic clustering.

In the next section we discuss some related works and introduce our approach. Then a description of the CHIR method precedes the explanation of our semantic feature selection technique SIM in the third section. Section 4 contains the details of our proposed sequential clustering algorithm. Finally the experimental results obtained are analyzed in the fifth section.

2. Related Works

Since feature selection is a powerful tool for simplifying the application of categorizing algorithms and improving their performances, several document clustering studies have focused on its methods. A recently introduced metric, third moment having the ability to reinforce rare features was used in [8] as feature selection method. Similarly, Yang et al. [9] proposed a new method called CMFS that comprehensively compute the significance of a term; it is means calculating its significance from both inter-category and intra-category. An extension of the well known chi-square measure was introduced in [5]. In addition to the capacity of testing the dependence between a term and a category, the new statistic CHIR measures the nature of this dependence and keeps only terms that are positively dependant to the categories. This technique was used by Meena et al. [10] into a novel feature selection method relying on the Meta heuristic algorithm Ant Colony Optimization. The CHIR helps updating the pheromone value that possesses each term in the corpus.

In spite of their efficiency all the above methods still insufficient since they treat only the statistical relations between the dataset terms and do not exploit the semantic relations that may exist between them. Therefore, an interesting research topic is to perform a clustering document using not only a statistical analysis but also a semantic one. Based on this assumption, Meng et al. [11] proposed a two phase feature selection method. The first step consists on selecting features using a novel method named feature contribution degree FCD and then in the second step a new semantic vector space model is constructed by means of LSI. Thangamani et al. [12] take also into consideration the content relationships between features in addition of the frequency ones, but differently to [11], they use two separate feature selection methods instead of a sequential one. The frequency and content feature selection are made respectively using the standard tf-idf function and the support of ontology.

The second method allows performing a semantic clustering that improves the statistical clustering already performed using the first method.

Motivated by the studies utilizing the combination of co-occurrence terms techniques and the semantic similarity between words [11] [12] [13] [14], we propose a clustering algorithm that extends and performs the text clustering algorithm TCFS proposed in [5] by a semantic analysis. More precisely, we propose a document clustering through two stages: a statistical clustering followed by a semantic one. The statistical weight of each term is estimated by the CHIR-statistic [5] and the semantic weight is estimated using a novel measure SIM based on the mutual information metric. The features selection and the clustering procedure are performed iteratively until convergence. The mechanism not only reduces the number of dimensions drastically, but also significantly improves the clustering performances in term of *F*-measure and purity by selecting statistically and semantically relevant features. Our approach is different from [12] by first using a new efficient method SIM that respond to the semantic analysis by relying on mutual information based similarity measure; and second by employing a powerful statistical measure that can extract frequency information better than do the *tf-idf* function they used; the CHIR method is an improved extension of the chi-square metric, one of the most effective feature selection techniques [6].

3. Materials and Methods

3.1 Background Knowledge

In text clustering, documents are widely represented as bag of words using the vector space model [3] [15]. In this model, each document is considered as a term-weight vector where the terms are the features that characterize the document and are generally single words. We use the *tf-idf* function as weighting schema which is the base one for the vector space model [3]. The term frequency tf indicates the importance of a term in the document and is a document specific statistic. The idf factor is a global weighting and measures how widely a term is distributed over the dataset. In this case, the representation of a document *d* is:

$$d_{tf\text{-}idf} = [tf_1 \log\left(\frac{n}{df_1}\right), tf_2 \log\left(\frac{n}{df_2}\right),..., tf_D \log\left(\frac{n}{df_D}\right)] \quad (1)$$

where tf_i is the frequency of the term *i* in the document d_i, df_i is the number of documents that contain the term *i*, *n* is the total number of documents in the collection, and *D* is the dimension of the text database. Finally, since the document have different lengths; we normalized each term vector by the Euclidean length, so that all document vectors turned into unit vectors. On doing so, we eliminate all information on the length of the documents.

In each clustering algorithm, the similarity between two documents must be measured. Different measures exist for this purpose, but the most common one is the cosine function [15]. In fact, for high-dimensional, all Euclidean techniques performed poorly, contrary to cosine function that is successful in capturing the similarities between documents [15]. In addition, cosine similarity is easy to interpret and simple to compute especially when the term vectors are normalized to a unit length. For two documents d_i and d_j the cosine similarity between them is calculated as:

$$\cos(d_i, d_j) = \frac{d_i \blacksquare d_j}{\|d_i\| \|d_j\|} \quad (2)$$

where ▪ refers to the vector dot product and *d* to the length of the vector *d*. When two documents are identical, the cosine value is equal to 1, and it takes the value 0 when the documents are totally dissimilar and have nothing in common. Since we calculated the similarity between a document *d* that is normalized to a unit length and a centroïd of a cluster cen, the formula become as following:

$$\cos(d, cen) = \frac{d \blacksquare cen}{\|cen\|} \quad (3)$$

Large cosine values indicate that the centroïd and the document are likely similar since they share most of their words.

3.2 CHIR Method

CHIR is a new feature selection technique introduced in [5] and based on the χ^2 statistic. The chi-square measure between a word *w* and a category *c* is defined as:

$$\chi^2_{w,c} = \sum_{i \in \{w, \bar{w}\}} \sum_{j \in \{c, \bar{c}\}} \frac{(o\{i, j\} - E(i, j))^2}{E(i, j)} \quad (4)$$

where $o(w, c)$ is the observed frequency of the documents that belong to the category c and contain w. $E(w, c)$ is the expected frequency of the category c and the term w.

In order to select only relevant terms that have strong positive dependency on certain categories in the dataset, Li et al. [5] propose a new measure $R_{w,c}$ that determines whether the dependency between two attributes is positive or negative. It is defined as:

$$R_{w,c} = \frac{o(w,c)}{E(w,c)} \tag{5}$$

If the term w and the category c are independent, $R_{w,c}$ is close to 1. If there is a positive dependency, $R_{w,c}$ is larger than 1 and when there is a negative dependency $R_{w,c}$ is smaller than 1. To define the term goodness of a term w in a dataset with m classes, they combine the $R_{w,c}$ measure and the $\chi^2_{w,c}$ statistic as follows:

$$r\chi^2 = \sum_{j=1}^{m} p(R_{w,c_j}) \tag{6}$$

where $p(R_{w,c})$ is the weight of $\chi^2_{w,c}$ in the dataset and it is defined as:

$$p(R_{w,c}) = \frac{R_{w,c}}{\sum_{j=1}^{m} R_{w,c_j}} \; with \; R_{w,c_j} > 1 \tag{7}$$

A larger value of $r\chi^2$ indicates that the term w is more relevant to the category.

The steps of the CHIR algorithm to select l terms are:

1. Calculate the $r\chi^2$ statistic for each term in the dataset.
2. Sort the terms in descending order of their term goodness.
3. Select the top l terms from the list.

3.3 SIM Method

The mutual information is a widely used measure in information theory. Two terms are considered similar if their mutual information with all terms in the vocabulary is nearly the same [16]. To measure the similarity between two terms w and w' we use the following metric [16]:

$$\text{sim}(w, w') = \frac{1}{2|V|} \sum_{i=1}^{|V|} \frac{\min(I(z_i, w), I(z_i, w'))}{\max(I(z_i, w), I(z_i, w'))} + \frac{\min(I(w, z_i), I(w', z_i))}{\max(I(w, z_i), I(w', z_i))} \tag{8}$$

where V is the vocabulary and $I(w,w')$ is the mutual information between the terms w and w' which is calculated by the following formula:

$$I(w, w') = P_d(w, w') \log\left(\frac{P_d(w, w')}{P(w)\,P(w')}\right) \tag{9}$$

where d is the withdrawal, $P(w)$ and $P_d(w, w')$ are respectively the a priori probability of the term w and the probability of succession of the terms w and w' in a window of (d +1) words.

The similarity between a term w and a document centroïd d is defined in [17] as the average of the similarities between the term w and the m terms of the document centroïd. This measure is given by:

$$\text{sim}(w, d) = \frac{\sum_{j=1}^{m} sim(w, w_j)}{\sum_{j=1}^{m}\sum_{i=1}^{m} sim(w_j, w_i)} \tag{10}$$

In order to determine the semantic relevance of a term w in a corpus of k clusters, we calculate the weighted sum of its similarities with the document centroïd of each cluster through the following formula:

$$sim(w) = \sum_{i=1}^{k} P(I(w, d_i))\, sim(w, d_i) \tag{11}$$

where $P(I(w, d_i))$ is the weight of the similarity between the term w and the document centroïd d_i and $I(w, d_i)$ is the mutual information between them. If one considers the contingency table of a term w and a centroïd d where A is the number of times w and d co-occur i.e. w occur in documents that belong to the cluster whose centroïd is d, B is the number of times w occurs without d, C is the number of times d occurs without w and N is the total number of documents, then the mutual information criterion between a term w and a document d is defined by:

$$I(w, d) = P(w, d) \log\left(\frac{P_d(w, d)}{P(w)\,P(d)}\right) \tag{12}$$

And it is estimated by:

$$I(w, d) = \frac{A}{N} \log\left(\frac{A * N}{(A + B) * (A + C)}\right) \tag{13}$$

If there is a genuine association between w and d then the joint probability $P(w, d)$ will be much larger than $P(w)\,P(d)$, and consequently $I(w, d) > 0$. If there is no interesting relationship between w and d, then $P(w, d) \simeq P(w)\,P(d)$ thus $I(w, d) \simeq 0$. If w and d are in complementary distribution, then $P(w, d)$ will be much less than $P(w)\,P(d)$ forcing $I(w, d) < 0$. We then defined $P(I(w, d))$ the weight of $sim(w, d)$ as:

$$P(I(w, d)) = \frac{(I(w, d_i))}{\sum_{i=1}^{k} (I(w, d_i))} \; with\; I(w, d_i) > 0 \tag{14}$$

The steps of the SIM algorithm to select l terms are:

1. Calculate the $sim(w)$ measure for each term in the dataset.

2. Sort the terms in descending order of their criterion function.

3. Select the top l terms from the list.

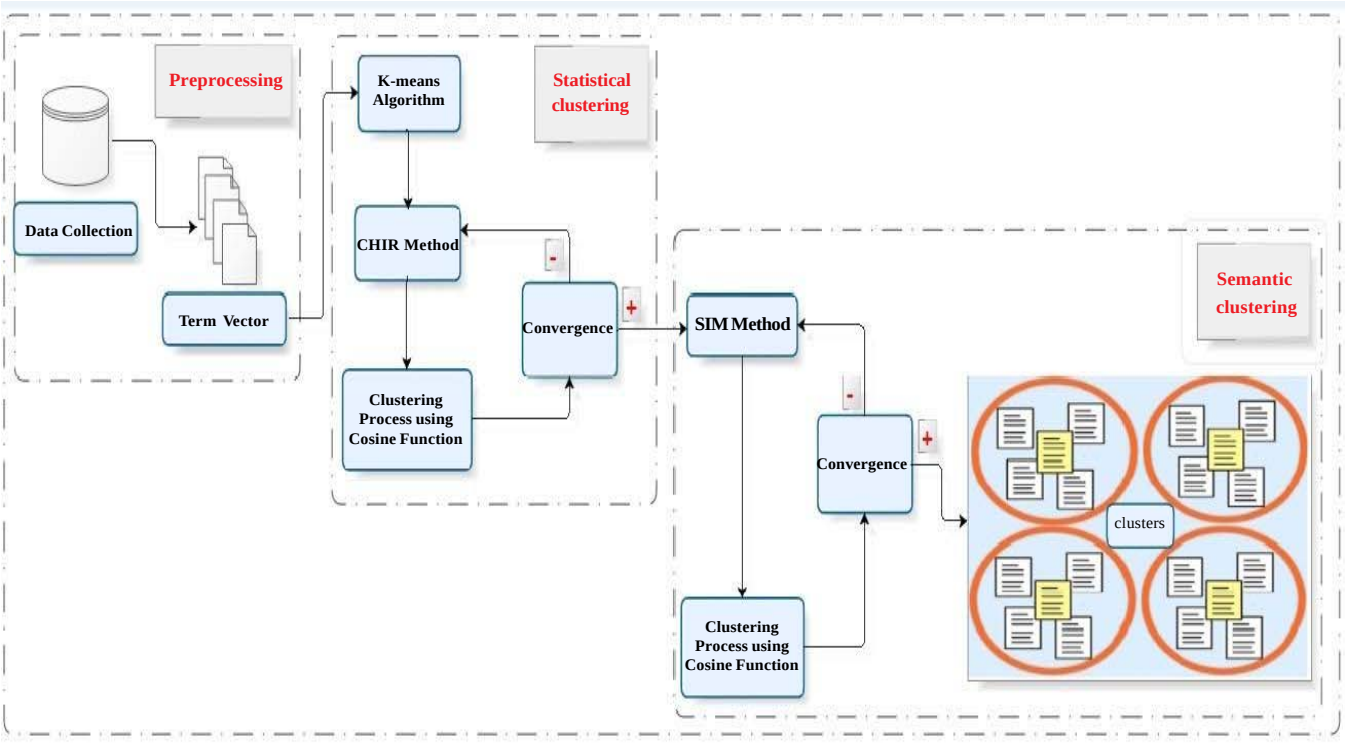

Figure1. The sequential clustering mechanism CHIRSIM

4. The CHIRSIM Method

Feature selection is an active research area used in several fields such as text clustering. The process typically involves metrics having the ability to select the most important features. Depending on whether external information is needed to implement the technique, feature selection methods are either supervised or unsupervised. Due to the simplicity of their employment, the unsupervised methods are commonly used. However, they are less efficient than the supervised techniques such as the CHIR and the SIM methods for text clustering [18]. The problem with these methods is that they cannot be directly applied because of the unavailability of the required class information. In case to adopt the supervised feature selection techniques in text clustering, we first perform the *K*-means algorithm to get initial clusters and centroïd; and then apply the feature selection method and the clustering iteratively. In this case, good clustering results will allow getting better features that will improve the clustering results and so on until convergence.

4.1 The K-means Algorithm

Traditionally clustering techniques are broadly divided into hierarchical and partition algorithms [19]; both methods were applied to document clustering. While hierarchical algorithms divides the given dataset into smaller subsets in hierarchical fashion [20] [21], the partitioning algorithms attempt a flat partitioning of a collection of documents into a predefined number of disjoint clusters [21] [5] [22] [23] [24]. They are further categorized into probabilistic clustering, *k*-medoids methods, and *k*-means methods, but the most commonly used is *k*-means algorithm. It was, on one hand, proven that *k*-means approach is as good as or better than the agglomerative hierarchical approaches in term of accuracy and efficiency [21]. On the other hand, the hierarchical clustering is extremely computational expensive as the size of data increases, in contrast, *k*-means algorithm is much faster [25]. Consequently, the *k*-means algorithm is well-suited for document categorization due to its relatively low computational requirement and high quality. For these reasons we use it as the basic of our clustering mechanism.

The steps of the *k*-means algorithm are the following:

1. Select randomly *k* documents from the dataset as initial centroïds.
2. Assign each document to the cluster that has the closest centroïd.
3. Recalculate the *k* centroïds.
4. Repeat steps 2 and 3 until convergence.

DataSets	Docs Num.	Classes Num.	Min. Class Size	Max. Class Size	Unique Terms Number	Avg. Terms per Doc	Avg. Pairwise Sim. by Cosine
CACM	205	13	9	31	2052	46	0.05
RTRS	445	7	20	138	7293	139	0.03
20NG	937	11	42	167	30883	323	0.03

Table 1. The Datasets

4.2 The CHIRSIM Algorithm

Before implementing our clustering approach, a preprocessing process was applied. It consists on performing steps that take as input a plain text document and as output a set of tokens to be included in the vector model. These steps typically consist of: firstly, removing special characters and punctuation that are not thought to hold any discriminative power under the vector model; secondly, splitting each sentence into individual words. Thirdly, reducing the words to their basic form i.e. stem. Finally, removing stop words which are terms that are not thought to convey any meaning.

After this preliminary step, we build clusters by performing the k-means algorithm as introductory step. We then categorize the documents starting by the *k* clusters guided by the k-means and using the CHIR-statistical measure (a co-occurrence method) to select relevant features. Finally, we use the clusters and centroids obtained from the frequency clustering as input to carry out the semantic clustering in which the mutual information measure SIM (a semantic measure) is used as feature selection method. The two feature selection methods use the current clusters and their centroids to estimate the relevancy of each term to the dataset as in shown in Figure 1. If a term is considered relevant, it is kept in the feature space. Otherwise, the term weight is

reduced by a factor f which is a predetermined factor in the range of [0, 1]. Then, in the new feature space and using the cosine function, the documents are reassigned to the cluster with the closest centroid.

Detailed steps of our algorithm are divided in three steps as follows:

4.2.1 Initial step

1. Perform the *k*-means algorithm on the dataset to get initial clusters and clusters centroids.

4.2.2 Statistical step: TCFS algorithm

2. Perform the CHIR method by using the current clustering result. The weight of each unselected feature is reduced by a predetermined factor $f \in [0,1]$.

3. Recalculate *k* centroids in the new feature space.

4. For each document in the dataset, calculate its similarity with each centroid and assign it to the closest cluster.

5. Repeat steps 2, 3, and 4 until convergence.

4.2.3 Semantic step: initial clusters and centroids are obtained from the TCFS algorithm

6. Perform the SIM similarity by using the current clustering result. The weight of each unselected feature is reduced by $f \in [0,1]$.

7. When all documents have been assigned, recalculate *k* centroids in the new feature space.

8. Reassign the documents to the closest cluster 9. Repeat steps 6, 7, and 8 until convergence.

5. Results and Discussion

5.1 The Data Set

The clustering mechanism is evaluated on two different types of text datasets; a compendium of abstracts CACM and two collections of articles Reuters-21578 and 20 Newsgroups. The three corpora are well known and frequently used on numerous information retrieval papers. We denote the sets extracted from CACM, Reuters and 20 Newsgroups by CACM, RTRS and 20NG. The information about these datasets is shown in Table 1.

5.2 Performance Measures

The effectiveness of the clustering can be evaluated by two types of measures. The first one, called internal quality measure, enables to compare different clusters without using external information. For this purpose, we use the cohesiveness of clusters which represents the average pairwise similarity between all documents in each cluster.

$$\begin{aligned} cohesiveness\,(C) &= \frac{1}{|C|^2} \sum_{d \in C,\, d' \in C} cosine\,(d, d') \\ &= \frac{1}{|C|} \sum_{d \in C} d \cdot \sum_{d' \in C} d' \\ &= \| cen \|^2 \end{aligned} \tag{15}$$

where *C* represents the cluster, cen is the centroid of the cluster, *d* and *d′* are documents in the cluster.

We notice that the cohesiveness of a cluster is the square of the cluster centroid vector length. This also includes the similarity of each document with itself, which is just one.

When a feature selection method is applied to text documents, the cohesiveness value of each cluster may change. A good feature selection method should eliminate irrelevant features, while obtaining large cohesiveness values of the clusters [5].

The second type of measure, called external quality measure, allows evaluating the reliability of the clustering mechanism by comparing the obtained clusters to known classes. We use to do this the F-measure and the purity. The F-measure is based on the fundamental informational retrieval parameters, "*precision*" and "*recall*". Precision is the ratio of the relevant documents retrieved number to the total number of the cluster documents. Recall is the ratio of the relevant documents retrieved number to the total number of the class documents. The formulas of the precision $P(i, j)$ and recall $R(i, j)$ are defined as:

$$P(i,j) = \frac{n_{ij}}{n_j} \tag{16}$$

and

$$R(i,j) = \frac{n_{ij}}{n_j} \tag{17}$$

where n_i is the number of the documents of the class i , n_j is the number of the documents of the cluster j and n_{ij} is the number of the documents of the class i in the cluster j.

The F-measure of a cluster j and a class i is given by:

$$F(i,\ j) = \frac{2 * P(i,\ j) * R(i,\ j)}{P(i,\ j) + R(i,\ j)} \tag{18}$$

For the entire clustering result, the F-measure is computed as following:

$$F = \sum i \frac{n_i}{n} max_j (F(i,\ j)) \tag{19}$$

where n is the total number of documents. A better clustering result is measured by the largest F-measure [26].

The purity of a cluster represents the ratio of the dominant class in the cluster to the size of the cluster. Thus the purity of the cluster j is defined as:

$$Purity\ (j) = \frac{1}{n_j} max_i\ (n_{ij}) \tag{20}$$

The global value for the purity is the weighted average of all purity values as given by the following formula:

$$Purity = \sum_j \frac{n_j}{n} Purity\ (j) \tag{21}$$

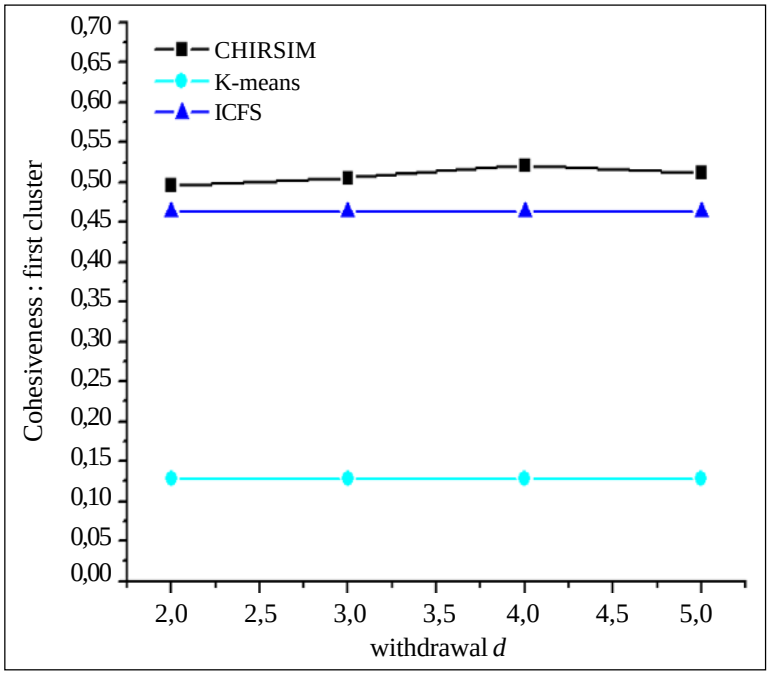

Figure 2. The first cluster cohesiveness value according to the withdrawal d for the CACM dataset

A better clustering result is measured by the largest purity value [27].

5.3 Results

In this section we compare our CHIRSIM clustering algorithm with the k-means and the TCFS algorithms. Since K-means

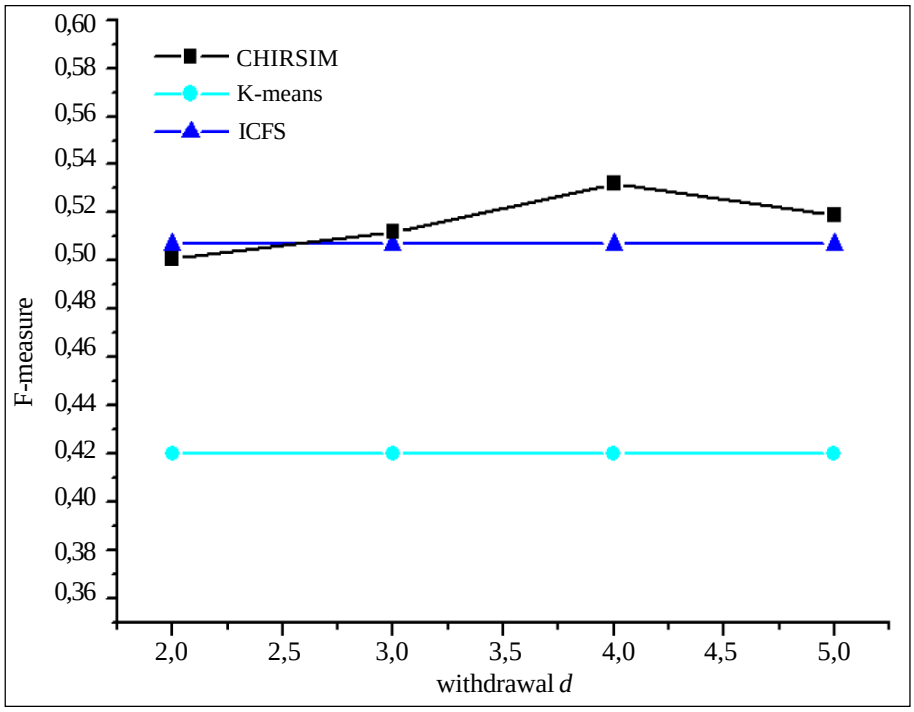

Figure 3. The F-measure value according to the withdrawal *d* for the CACM dataset

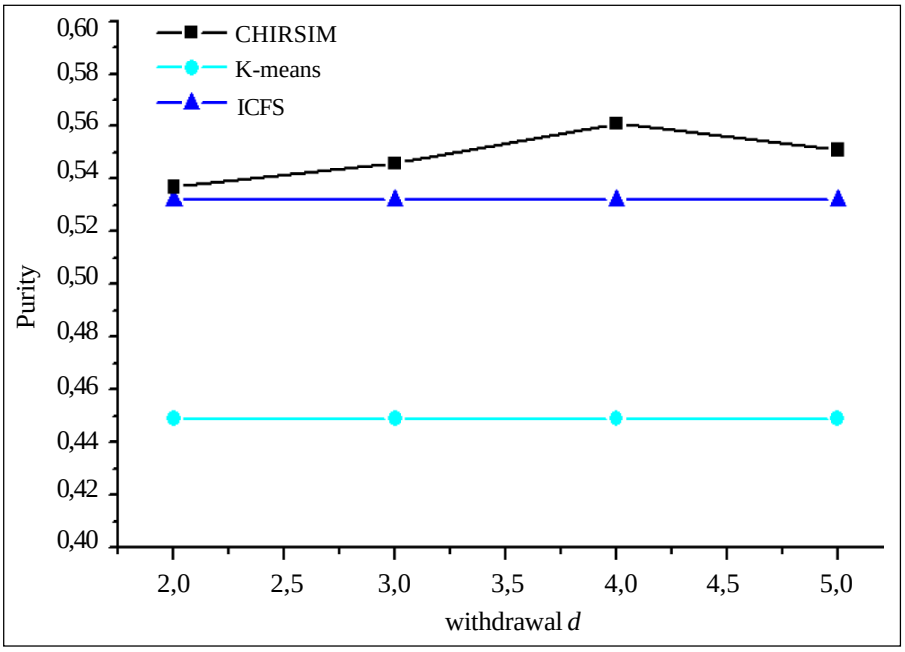

Figure 4. The purity value according to the withdrawal *d* for the CACM dataset

algorithm is easily influenced by the selection of initial centroids, we tested the algorithm with a set of 20 random initial centroids from the data set. We used the averages of results for the comparison. The experimental results show that our sequential clustering algorithm has the best clustering accuracy.

Since the parameters we use can significantly affect the clustering mechanism, the three algorithms should be performed with

different values of these parameters. However, because we compare our algorithm with the TFCS clustering mechanism we only varied the withdrawal d from 2 to 5 and keep the values of the percentage of the selected features t and the factor f that reduced the term weight of irrelevant features as they were taken in [5]. Hence $t = 25\%$ and $f = 0.5$.

Our proposed feature selection method SIM is based on the mutual information measure that is calculated on a sliding window *d*. To varied the withdrawal d we based on the fact that 98% of the occurrences of lexical relations relate words are separated by at most five words within a single sentence [15]. Therefore, we execute our sequential algorithm with $d = 2, 3, 4$ and 5. As shown in Figure 2, Figure 3, and Figure 4, the best results are obtained for $d = 4$.

Categorizing unstructured textual documents containing billion of terms is very complex; hence, the usefulness of feature selection. Many methods have been proposed for this purpose, but most of them focus on the frequency relations that exist between the features. However, the clustering mechanism requires not only the frequent informative features of the dataset but also the content relationship ones: the semantic analysis. Our proposed system offers a document clustering with the support of statistical and semantic analyses. The results show that the statistical clustering TCFS is optimized with our proposed semantic one. In fact, the semantic clustering reduces the irrelevant features in a considerable manner as shown by the improved cohesiveness values. The sequential algorithm with a semantic feature selection method can achieve a better accuracy and efficiency factors than the TCFS and much better than the *k*-means. Table 2 shows F-measure and purity values of the clusters obtained by running the three clustering when 25% of the terms were selected, the weight of unselected features is reduced by 50% and the withdrawal is 4. The semantic feature selection process produces more accurate results.

Datasets	F-measure			Purity		
	K-means	TCFS	CHIRSIM	K-means	TCFS	CHIRSIM
CACM	0.42	0.507	0.532	0.449	0.532	0.561
RTRS	0.63	0.687	0.739	0.681	0.742	0.787
20NG	0.601	0.633	0.648	0.644	0.674	0.692

Table 2. The Cohesiveness, F-Measure and Purity Values for the Three Algorithms **(**$t = 25\%$, $f = 0.5$ and $d = 4$)

6. Conclusion

The purpose of this study was to perform an efficient text document clustering while dealing with the curse of dimensionality. To address this problem, we proposed a new clustering approach based on the feature selection mechanism. Firstly, we introduced a novel semantic feature selection technique SIM; and then we extended the statistical text categorization TCFS by a semantic one that uses the SIM method. For a mutual beneficiation, the clustering step and the feature selection step were done in parallel order. Therefore, selecting pertinent features aimed to facilitate the improvement of the clustering which allowed obtaining more relevant features. Our sequential clustering algorithm showed through the experimental results, that considering the semantic relationships between features improves the frequency analysis and optimize the clustering accuracy by selecting statistical and semantic relevant features.

References

[1] Thangamani, M., Thangaraj, P. (2010). Survey on Text Document Clustering, *International Journal of Computer Science and Information Security*, 8 (2) 174-178.

[2] Sathiyakumari, Manimekalai, G., Preamsudha, V. (2011). A Survey on Various Approaches in Document Clustering, *Int. J. Comp. Tech. Appl*., 2 (5)1534-1539.

[3] Sebastiani, F. (2002). Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, 34 (1) 1-47.

[4] Parsons, L., Haque, E., Liu, H. (2004). Subspace clustering for high dimensional data: a review, *ACM*, 6 (1) 90-105.

[5] Li, Y., Luo, C., Chung, S.M. (2008). Text Clustering with Feature Selection by using Statistical Data Knowledge and Data Engineering, *IEEE Transactions on Know and Data Eng*., 20 (5) 641–651.

[6] Yang, Y., Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization, *presented at the ICML*.

[7] Zheng, Z., Rrihari, R. (2003). Optimally Combining Positive and Negative Features for Text Categorization, *In*: Proceedings of the ICM, Workshop for Learning from Imbalanced Datasets II.

[8] Peleja, F., Lopes, G. P., Silva, J. (2011). Text Categorization: A Comparison of Classifiers, Feature Selection Metrics and Document Representation, *In*: Proceedings of the 15th Portuguese Conference in Artificial Intelligence, p.660-674.

[9] Yang, Y., Liu, Y., Zhu, X., Liu, Z., Zhang, X. (2010). A New Feature Selection Base on Comprehensive Measurement both in Inter-category and Intra-category for text categorization, *Information Processing & Management*, 48 (4) 741-754.

[10] Meena, M. J., Chandran, K. R., Brinda, J. M. (2010). Integrating Swarm Intelligence and Statistical Data for Feature Selection in Text Categorization, *International Journal of Computer Applications*, 1 (11) 16-21, 2010.

[11] Meng, J., Lin, H., Yu, Y. (2010). A two stage feature selection method for text categorization. *In*: Proceedings of Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).

[12] Thangamani, M., Thangaraj, P. (2010). Integrated Clustering and Feature Selection Scheme for Text Documents, *Journal of Computer Science*, 6 (5) 536-541.

[13] Frikh, B., Djaanfar, A. S., Ouhbi, B. (2011). A New Methododlogy for Domain Ontology Construction from the Web, *International Journal on Artificial Intelligence Tools*, p. 1157-1170.

[14] Termier, A., Rousset, M., Sebag, M. (2001). Combining Statistics and Semantics for Word and Document Clustering Alexandre Termier, *In*: Ontology Learning Workshop, IJCAI'01, p. 49-54.

[15] Strehl, A., Ghosh, J., Mooney, R. (2000). Impact of Similarity Measures on Web-page Clustering, AAAI Workshop on A.1 for Web Search, p. 58-64.

[16] Dagan, I., Marcus, S., Markovitch, S. (1995). Contextual Word Similarity and Estimation from Sparse Data, *Computer Speech and Language*, 9 (2)123-152.

[17] Djaanfar, A.S., Frikh, B., Ouhbi, B. (2012). A Hybrid Method for Improving the SQD-PageRank Algorithm, Second International Conference on the Innovative Computing Technology, Casablanca, Morocco.

[18] Liu, T., Liu, S., Chen, Z., Ma, W. (2003). An Evaluation on Feature Selection for Text Clustering, Proc. of Int'l Conf. on Machine Learning.

[19] Berkhin, P. (2006). Survey Of Clustering Data Mining Techniques, Grouping Multidimensional Data, Cl (c) 25-71.

[20] Sotoca, J. M., Pla, F. (2010). Supervised feature selection by clustering using conditional mutual information-based distances, *Pattern Recognition*, 43 (6). 2068-2081, June.

[21] Steinbach, M., Karypis, G., Kumar, V. (2000). A Comparison of Document Clustering Techniques, KDD Workshop on text Mining.

[22] Thangamani, M., Thangaraj, P. (2010). Integrated Clustering and Feature Selection Scheme for Text Documents, *Journal of Computer Science*, 6 (5) 536-541.

[23] Zhong, S. (2005). Efficient online spherical k-means clustering, *In*: Proceedings 2005 IEEE International Joint Conference on Neural Networks, 5, p. 3180-3185.

[24] Ahmad, R., Khanum, A. (2010). Document Topic Generation in Text Mining by Using Cluster Analysis with EROCK, *International Journal of Computer Science & Security*, 4 (2) 176-182.

[25] Xiao, Y. (2010). A Survey of Document Clustering Techniques & Comparison of LDA and moVMF, CS 229 Machine Learning Final Projects, December.

[26] Steinbach, M., Karypis, G., Kumar, V. (2000). A Comparison of Document Clustering Techniques, KDD Workshop on text Mining.

[27] Zhao, Y., Karypis, G. (2004). Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering, *Machine Learning*, 55 (3) 311–331.