Reducing the Loss of Information through Annealing Text Distortion

Ana Granados, Manuel Cebrian, David Camacho, and Francisco de Borja Rodríguez

***Abstract*— Compression distances have been widely used in knowledge discovery and data mining. They are parameter-free, widely applicable, and very effective in several domains. However, little has been done to interpret their results or to explain their behavior. In this paper we take a step towards understanding compression distances by performing an experimental evaluation of the impact of several kinds of information distortion on compression-based text clustering. We show how progressively removing words in such a way that the complexity of a document is slowly reduced helps the compression-based text clustering and improves its accuracy. In fact, we show how the non-distorted text clustering can be improved by means of annealing text distortion. The experimental results shown in this paper are consistent using different data sets, and different compression algorithms belonging to the most important compression families: Lempel-Ziv, Statistical and Block-Sorting.**

***Index Terms*— Information Distortion, Data Compression, Normalized Compression Distance, Clustering by Compression, Kolmogorov Complexity.**

I. Introduction

Compression distances are currently a hot topic of research in several areas such as data mining [1], question-answering systems [2], [3], plagiarism detection [4], bioinformatics [5], [6], [7], [8], [9], philology [10], neural networks [11], prediction [12], software metrics [13], [14], [15], clustering [16], [17], [18], information retrieval [19] and text categorization [20], [21]. Many other computational problems can be reduced to compression oriented concepts, such as pattern discovery, regression, outlier detection and forecasting [22], [23]. This success relies on its parameter-free nature, wide applicability, and leading efficacy in several domains. Also, this methodology is benefiting from the very mature and diverse research field on compression algorithms, whose only target so far has been the detection and reduction of redundancy in stored digital information (see the comprehensive reference [24]).

Despite all this, little has been done to interpret compression distances results or to explain their behavior. The main reason for this, is the immense gap between their theoretical foundation - the Kolmogorov complexity in several flavors - and the state-of-the-art compression algorithms used in applications. Whenever some analytical work on compression distances is carried out, it is usually focused on the algebraic manipulation of algorithmic information theory concepts [2], [25], [26]. Even though those concepts are really supporting the use (and the optimality) of compression distances, they cannot help in interpreting the behavior of state-of-the-art compression algorithms like BZIP2 [27], LZMA [28], PPMZ [29] and many others. The idiosyncrasy and specificity of the wide diversity of compression algorithms cannot be captured by these universal - and uncomputable - concepts (see e.g. [30]).

Text distortion has been used to study the behavior of compression distances. For example, some theoretical (and experimental) basis to describe the behavior of the (normalized) compression distance-clustering when it is applied in a set of elements which have been perturbed by a certain amount of uniform random noise can be found at [31]. The impact of sporadic erasures on the limits of lossless data compression from a theoretical perspective can be found at [32]. On the other hand, word substitution has been suggested as a kind of text protection, based on the subsequent automatical detection of such substitutions by looking for discrepancies between words and their contexts [33].

In this paper we take a small step towards understanding compression distances by performing an experimental evaluation of the impact of several kinds of word removal on compression-based text clustering. In order to do so, we analyze how the information contained in the documents and our estimation of an upper bound for the Kolmogorov complexity progress as we remove words from the documents. We show how the conclusions of this analysis can be used to improve the accuracy of the clustering. It is worth highlighting that the results are consistent across the most important compression families: Lempel-Ziv, statistical and block-sorting, and across different data sets (see the Appendix for details of the data sets).

The main contributions of this paper can be briefly summarized as follows:

- New insights for the evaluation and explanation of the behavior of the compression distance-driven clustering algorithms.
- A technique to reduce our estimation of an upper bound for the Kolmogorov complexity of the documents while preserving most of the relevant information. This technique, that we called annealing text distortion, produces both a smooth decrease of the non-relevant information in the set of documents considered, and a smooth decrease of the documents complexity estimation.

Manuscript received April 30, 2009; revised xxxxxx xx, xxxx. This work was supported by the Spanish Ministry of Education and Science under TIN 2004-04363-CO03-03, TIN 2007-65989, CAM S-SEM-0255-2006, S-0505/TIC/000267 and TSI 2005-08255-C07-06.

A. Granados, M. Cebrián, D. Camacho and F. Rodríguez are with the Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, Spain (E-mail: ana.granadosf@uam.es; manuel.cebrian@uam.es; david.camacho@uam.es;f.rodriguez@uam.es).

M. Cebrian is with the MIT Media Laboratory, Cambridge, MA, USA (E-mail: cebrian@media.mit.edu)

- Experimental evidence of how to fine-tune the annealing text distortion so that better results are obtained when using the NCD-driven text clustering. This annealing text distortion can be seen as a change in the representation of the texts that allows the compressor to obtain more reliable similarities, which leads to an improvement of the the non-distorted NCD-driven text clustering.

This paper is structured as follows. Section II reviews the Normalized Compression Distance, one of the most successful similarity distances in the family of compression distances. Section III describes the data sets, the distortion techniques, and the clustering assessment. Section IV gathers and analyzes the obtained results. Section V summarizes the conclusions and describes ongoing research. Detailed information about the data sets used for the experiments can be found in the Appendix.

II. Evaluating text distortions: The Normalized Compression Distance

A natural measure of similarity assumes that two objects x and y are similar if the basic blocks of x are in y and vice versa. If this happens we can describe object x by making reference to the blocks belonging to y, thus the description of x will be very simple using the description of y.

This is what a compressor does to code the concatenated xy sequence: a search for information shared by both sequences in order to reduce the redundancy of the whole sequence. If the result is small, it means that the information contained in x can be used to code y, following the similarity conditions described in the previous paragraph.

This was studied by [25], [26], giving rise to the concept of *normalized compression distance* (NCD). This quantity is based on the use of compressors to provide a measure of the similarity between the objects. This distance may then be used to cluster those objects.

The definition is as follows

$$NCD(x,y) = \frac{\max\{C(xy) - C(x), C(yx) - C(y)\}}{\max\{C(x), C(y)\}}, \quad (1)$$

where C is a compression algorithm, $C(x)$ is the size of the C-compressed version of x, and $C(xy)$ is the compressed size of the concatenation of x and y. NCD generates a non-negative number $0 \leq NCD(x,y) \leq 1$. Distances near 0 indicate similarity between objects, while distances near 1 reveal dissimilarity.

The theoretical foundations for this measure can be traced back to the notion of Kolmogorov complexity $K(X)$ of a string X, which is the size of the shortest program able to output X in a universal Turing machine [34], [35], [36]. As this function is incomputable due to the Halting problem [37], the most usual estimation is based on data compression: $C(X)$ is considered a good upper estimate of $K(X)$, assuming that C is a reasonably good compressor for X [26].

The NCD is just one of the many similarity distances that use compression algorithms. Others [2], [10], [17], are small variations and can be easily reduced to it, as it is possible to prove that this distance minorizes (is as good as) any other that can be computed by a universal Turing machine. We use the NCD to evaluate how the information contained in the texts depends on the text distortions that we perform.

III. Experimental Design

A. Data sets

We have used the CompLearn Toolkit [38] which implements the clustering algorithm described in [26] to carry on the experiments. This clustering algorithm has an asymptotical cost of $O(n^3)$ from version 1.1.3. onwards [26]. Consequently, we have used a reduced number of documents for each data set.

First we have studied how the different distortion techniques affect the loss of information. Thus, we have used six different replacement methods and three different compression algorithms over four data sets comprised of texts written in English. We briefly describe here the data sets, but more detailed information can be found in the Appendix.

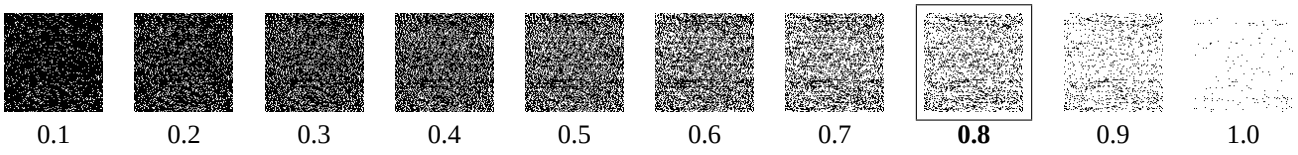
- Fourteen classical books from universal literature, to be clustered by author.
- Sixteen messages from a newsgroup (UCI-KDD) [39], to be clustered by topic.
- Twelve documents from the MedlinePlus repository [40], to be clustered by topic.
- Fourteen plots of movies from the Internet Movie Data Base (IMDB) [41], to be clustered by saga.

After observing that the non-distorted NCD-based text clustering could be improved using a specific distortion method, we have studied the behavior of this distortion method in bigger data sets. Thus, we have progressively increased the number of documents of the MedlinePlus data set to make it as big as possible while still using the clustering algorithm developed in [38]. We have used data sets of 50 and 60 documents from the MedlinePlus repository [40].

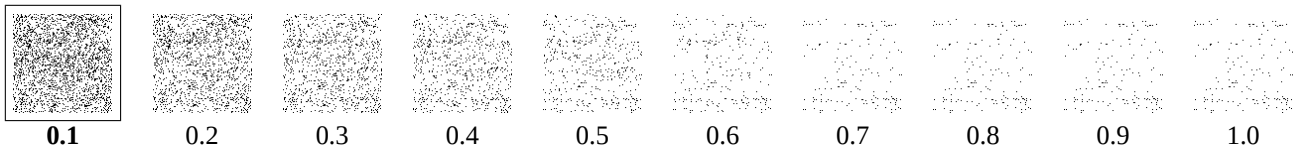
B. Distortion Techniques: Word Removal

Other works have shown that distorting the documents by removing the stop-words may have beneficial effects both in terms of accuracy and computational load when clustering documents or when retrieving information from them[42]. There are two main approaches to word removal, one in which a generic fixed stop-word list is used [43], [44], and other in which this list is generated from the collection itself [45]. The first approach is 'safer' in terms of maintaining the most relevant information of the documents. That is, the replaced words are not specific enough so as to lose important information. The second approach generates the stop-words list from the collection of documents, obtaining a much larger list which eventually produces a more aggressive word removal.

In this work we use the first approach to preprocess the documents. That is, we use an external and well-known dictionary, the BNC [46], to select the words that will be removed from the documents. The way in which we select and remove the words from the documents is different for each experiment. Thus, we use six different replacement methods,



(a) Most frequent word selection



(b) Least frequent word selection

Fig. 1. Visual representation of the information loss. Each binary image represents the information contained in all the documents belonging to the MedlinePlus data set. Black pixels represent remaining words and white pixels represent substituted words. First row corresponds to the images when the most frequent word selection method is used, while the second row corresponds to the least frequent word selection method. An image is created for every experiment, i.e., for every cumulative sum of word-frequencies (0.1, 0.2, and so on until 1.0, where all the words are selected). Note that even when all the words included in the BNC are replaced from the texts, the words that are not included in the BNC remain in the documents (observe black pixels in the images corresponding to the cumulative sum of 1.0). Although the amount of black pixels of the images in the boxes is quite similar, there will be shown that there is a big difference in means of clustering error in the experimental results.

which are pairwise combinations of two factors: *word selection method* and *substitution method*.

- *Word selection method*: the frequencies of words in English are estimated using the British National Corpus (BNC) [46], and then the list of words is sorted in *decreasing*/*increasing*/*random* order of frequency. We select the words to be removed by calculating the cumulative sum of the word-frequencies. Thus, we select the words that accumulate a frequency of 0.1, 0.2, 0.3, and so on, until 1.0, where finally all the words are selected. Depending on the way in which the list of words of the BNC is sorted before calculating the cumulative sum of frequencies, we will have one or another word selection method. Consequently, we have used three word selection methods: *most frequent word* selection method (MFW selection method), *least frequent word* selection method (LFW selection method) and *random word* selection method (RW selection method). For the RW selection method we repeat ten times the experiments and we calculate the mean and the standard deviation of the obtained results.
- *Substitution method*: when a word has to be removed from a text, each character of the word is replaced by either a random character, or an asterisk. Thus we have two substitution methods: *random character* substitution method and *asterisk* substitution method.

Note that all six combinations maintain the length of the document. This is enforced to ease the comparison of our estimation of an upper bound for the Kolmogorov complexity among several methods.

In order to gain an insight into how the information is decreased, we have created binary images that represent the information contained in a document. Each pixel can be either white or black. Black pixels represent remaining words and white pixels represent substituted words. As a consequence, a non-distorted document will be a completely black image, whereas a highly distorted document will have only some spurious black pixels. Fig 1 shows the information distortion progress for the MedlinePlus repository as a function of the cumulative sum of word-frequencies being replaced in each experiment. It can be observed that as the number of replaced words increases, the image has a higher number of white pixels. Note that even when all the words included in the BNC are replaced from the texts, the words that are not included in the BNC remain in the documents (see images on the right, which correspond to a cumulative sum of 1.0).

C. Text Clustering

We use text clustering to quantitatively measure how the relevant information of the documents remains in them as we incrementally remove words from the documents. Thus, after the distortion of the data sets, we execute the NCD clustering algorithm on each distorted test set and we quantitatively measure the error of the clustering results obtained. We use the CompLearn Toolkit [38], which implements the clustering algorithm described in [26]. This clustering algorithm comprises two phases.

First, the NCD matrix is calculated using a compression algorithm. We have tested the behavior of three different compression algorithms, LZMA, BZIP2 and PPMZ, each of them from a different family of compressors [24]. LZMA compressor, is a Lempel-Ziv-Markov chain algorithm [28]. BZIP2 compressor is a block-sorting compressor based on the Burrows-Wheeler transform and Huffman codes [27], [47], [48]. PPMZ compressor is an adaptive statistical data compression algorithm based on context modeling and prediction [29].

Second, the NCD matrix is used as input to the clustering phase and a dendrogram is generated as output. A dendrogram is an undirected binary tree diagram, frequently used for

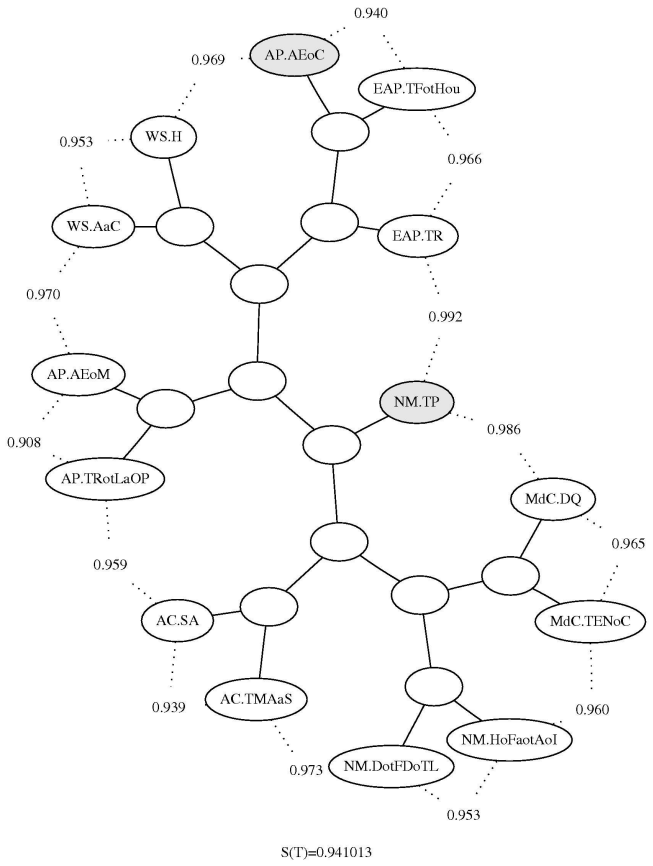


Fig. 2. Example of dendrogram for the Books repository. Each leaf of the dendrogram corresponds to a document. The numbers in the image represent the NCD average between two nodes. We measure the clustering error associated to a dendrogram adding all the pairwise distances between nodes starting with the same string. The distance between two nodes is the minimum number of internal nodes needed to go from one to the other. For example, the distance between the nodes with label $WS.H$ and $WS.AaC$ would be one, since both nodes are connected to the same internal node. After calculating this addition, we subtract the clustering error that corresponds to the perfect clustering from the total quantity obtained. The clustering error corresponding to this dendrogram is 11 because the nodes with label $NM.TP$ and $AP.AEoC$ have not been correctly clustered.

hierarchical clustering, that illustrates the arrangement of the clusters produced by a clustering algorithm. In Fig 2 we can observe a representative example of a dendrogram. Each leaf of the dendrogram corresponds to a document and its label helps us to easily analyze the quality of the dendrogram obtained, because each label starts with the name of the cluster in which the document should be included.

Once the CompLearn Toolkit [38] has been used to cluster the documents and the dendrograms are generated, we need to quantitatively measure the error of the dendrograms obtained. We define the distance between two nodes as the minimum number of internal nodes needed to go from one to the other. For example, in Fig 2 the distance between the nodes with label $WS.H$ and $WS.AaC$ would be one, since both nodes are connected to the same internal node. We use this concept to measure the clustering error of a dendrogram.

First, we add all the pairwise distances between nodes starting with the same string, i.e. we add all the pairwise distances between the documents that should be clustered together. For example, in Fig 2 there are three nodes whose label starts with NM. They correspond to the three books by Niccolò Machiavelli we are working with (see the Appendix for more details). Therefore, we add the distance between $NM.TP$ and $NM.DotFDoTL$, between $NM.TP$ and $NM.HoFaotAoI$, and between $NM.DotFDoTL$ and $NM.HoFaotAoI$. We repeat this procedure with all the nodes obtaining a certain total quantity. Then, after calculating this addition, we subtract the addition that corresponds to the perfect clustering, from the total quantity obtained. Consequently, if a dendrogram clusters perfectly all the documents, the clustering error would be 0, and in general, the bigger the clustering error, the worse the clustering would be. The clustering error corresponding to the dendrogram shown in Fig 2 is 11, because the sum of all the pairwise distances is 25, and the sum of all the pairwise distances in a perfect dendrogram for these documents is 14.

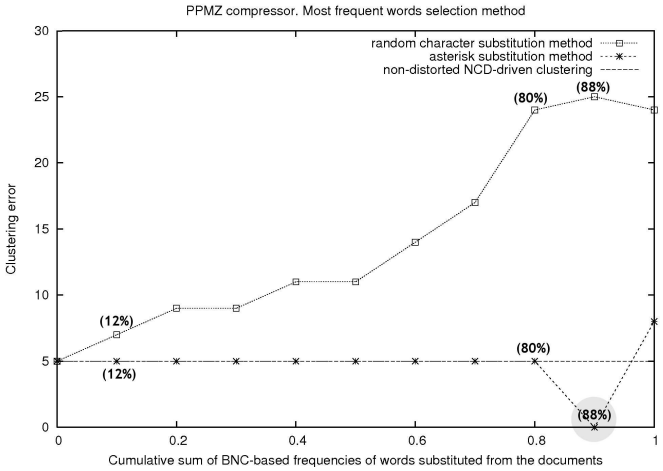
Finally, to get an insight into how the structure of the documents is affected by the distortion techniques, the Kolmogorov complexity of the distorted documents is estimated, based on the concept that data compression is an upper bound of the Kolmogorov complexity. That is, we estimate the upper bound of the Kolmogorov complexity as the length of the compressed file in bytes. We use the same three compression algorithms that we have used in the clustering phase, to estimate the upper bound of the Kolmogorov complexity, and we observe that the complexities are qualitatively similar for all of them.

IV. Experimental Results

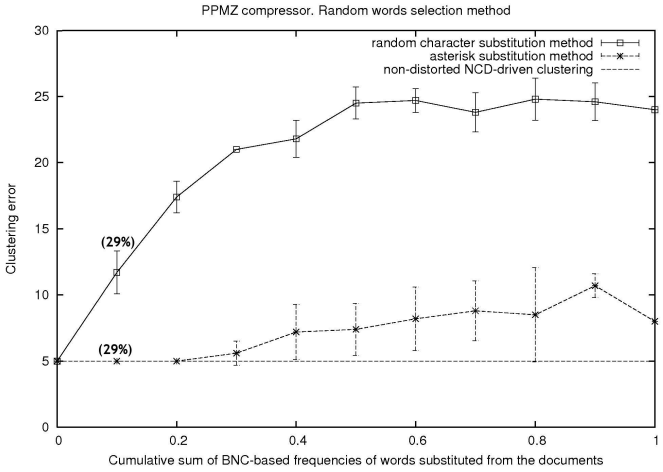
In this work we have studied the effects of distorting the information contained in different sets of documents using different distortion methods (see section III). We have used NCD-based text clustering to quantitatively evaluate the information loss. In terms of implementation, we have used the CompLearn Toolkit [38] to perform the clustering.

First, we have applied the NCD-based clustering algorithm over four different sets of texts written in English to study how the different distortion techniques affect the loss of information. We have tested the behavior of three different compression algorithms when calculating the NCD to cluster the documents. The results obtained for all the compression algorithms are similar. Consequently, we only show graphically the results that correspond to a specific compression algorithm (PPMZ), although the results corresponding to all compression algorithms are summarized in several tables. Analyzing these tables, it can be observed that all the compression algorithms provide similar results. A detailed description using different removal techniques and the LZMA compression algorithm in a particular data set can be found in [49].

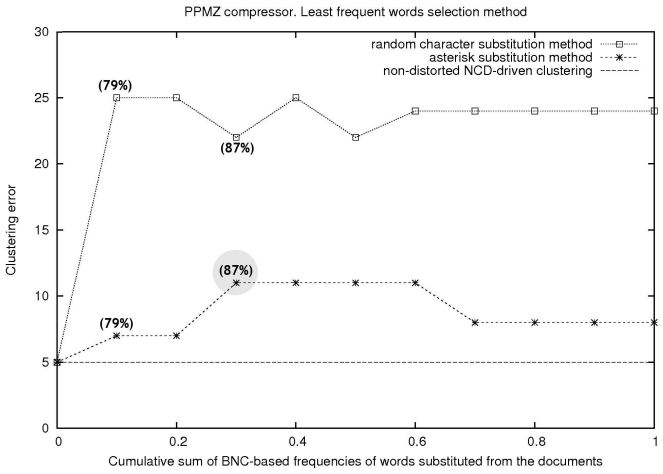
Second, after observing that the non-distorted NCD-based text clustering can be improved when one distortion method is used, we have studied the behavior of this distortion method in bigger data sets. Thus, we have used sets of 50 and 60 documents from the MedlinePlus repository. The results obtained in bigger data sets are consistent with the ones obtained in the previous ones.



(a) Most frequent word selection method



(b) Random word selection method



(c) Least frequent word selection method

Fig. 3. Books data set using the PPMZ compressor. Clustering error obtained for all the word selection methods. The numbers between brackets correspond to the percentage of substituted words in the documents. The asterisk substitution method performs better than the random character substitution method in all cases. The best results are obtained for the MFW selection method and asterisks replacement method (see curve with asterisk markers in (a)). Some points are highlighted inside a circle for further discussion.

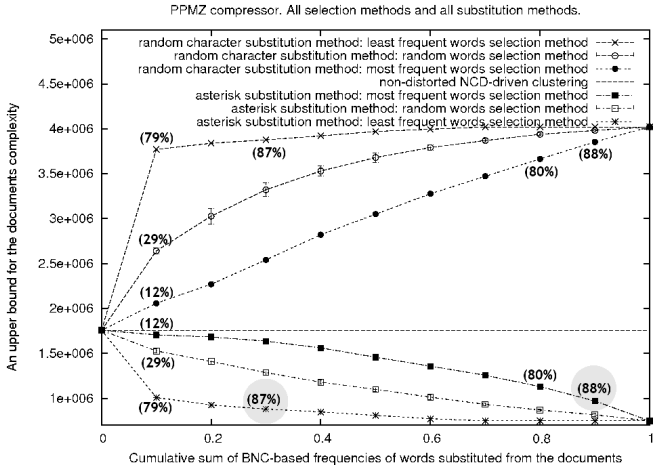


Fig. 4. Books data set using the PPMZ compressor. Estimation of an upper bound for the documents complexity for all the substitution methods and all the word selection methods. The values associated to the asterisk substitution method decrease for all the word selection methods, as the ones associated to the random character substitution method grow for all the word selection methods. The same percentages of substituted words included in Fig 3 are included in this figure to ease the comparison of both figures.

A. The Books Data set: a case study on PPMZ

For this data set, we show two different figures, Fig 3 depicts the clustering error and Fig 4 shows our estimation of an upper bound for the complexity of the documents. In both figures, the value on the horizontal axis corresponds to the cumulative sum of the BNC-based frequencies of the words substituted from the documents. For some relevant points, percentages of substituted words are included in the curve points between brackets. These percentages are calculated by dividing the number of substituted words in the documents by the total number of words contained in the documents. These percentages are useful to understand how important the choice of the words to be substituted from the documents is.

There are three different panels in Fig 3, corresponding each one of them to a different word selection method. In each panel, the curve with asterisk markers corresponds to the asterisk substitution method, while the one with square markers corresponds to the random character substitution method. The constant line corresponds to the non-distorted NCD-driven clustering error. We depict the non-distorted NCD-driven clustering as a constant line although it only has sense for a cumulative sum of frequencies of 0, because it is easier to see the difference between the line and the clustering error curves.

We can observe by looking at Fig 3, that the asterisk substitution method is always better than the random character substitution method. This was to be expected because substituting a word with random characters adds noise to the documents, and therefore most likely increases the Kolmogorov complexity of the documents and makes the clustering worse. On the other hand, we can realize that the best clustering results correspond to the MFW selection method (see Fig 3(a)), the worst results correspond to the LFW selection method (see Fig 3(c)), and the results corresponding to the RW selection method are maintained in between of them (see

Fig 3(b)).

Our estimation of an upper bound for the complexity of the documents for all selection methods and all substitution methods is depicted in Fig 4. The same relevant point percentages of distorted words included in Fig 3 are included in Fig 4. These percentages ease the comparison of Fig 3 and Fig 4. The values associated to the asterisk substitution method always decreases, while the values associated to the random character substitution method always increases.

Looking at the points highlighted inside a circle in Figs 3 and 4 we can observe that, for these points, although the complexity values and the percentages of removed words are similar, there is a significant difference in terms of clustering error. Consequently, we can realize that not only the substitution method is important, but also the word selection method. Thus, the best way to distort the documents is combining the MFW selection method and the asterisk substitution method. This is consistent across the different compression algorithms and the different data sets, how it will be shown in the next section.

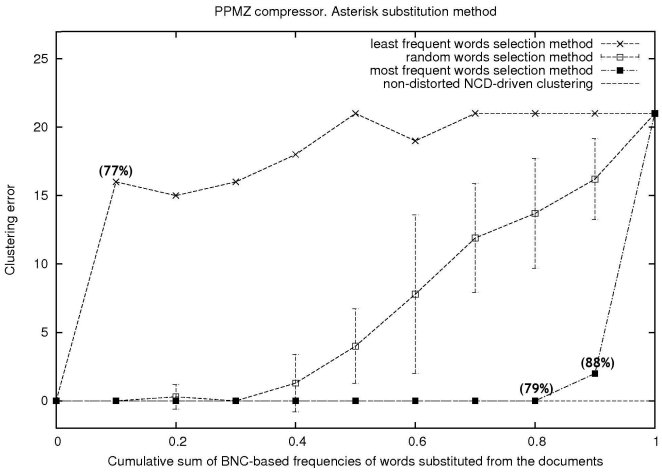
B. The Medline, UCI-KDD and IMDB data sets: graphical results for PPMZ

For each set of documents we show two figures. One depicts our estimation of an upper bound for the documents complexity while the other one depicts the clustering error. Only the results corresponding to the asterisk substitution method are shown. The results corresponding to the random character substitution method are not shown because these clustering results always get worse, as was to be expected. In every graph, the curve with black square markers corresponds to the MFW selection method, the graph with white square markers corresponds to the RW selection method, and the graph with star markers corresponds to the LFW selection method. On the other hand, the constant line corresponds to the non-distorted NCD-driven clustering.

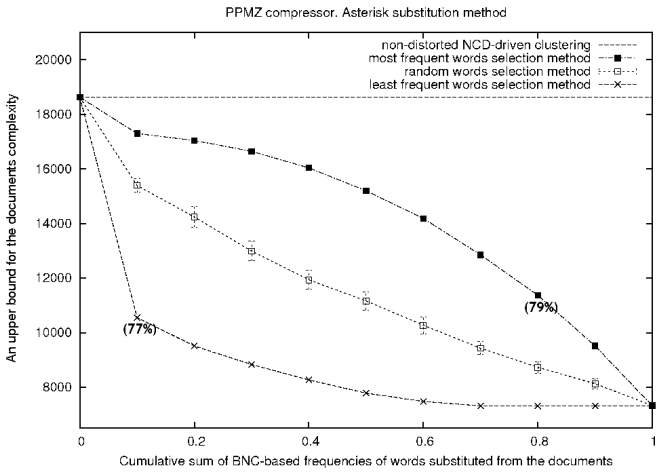
Figs 5(a), 6(a), and 7(a) depict the clustering error obtained when the asterisk substitution method is used to replace the words from the UCI-KDD, MedLinePlus and IMDB data sets respectively. The figures plot the clustering error of all the word selection methods. Analyzing the figures, we can notice that the best clustering results are obtained when the MFW selection method is applied (see curves with black square markers). In fact, for the UCI-KDD data set, these results correspond to the perfect clustering for the cumulative sum of frequencies from 0 to 0.8. For the MedLinePlus data set, the non-distorted NCD-driven clustering is improved from 0.5 to 0.8, although the perfect clustering is not achieved. For the IMDB data set, the perfect clustering is achieved for the cumulative sum of frequencies of 0.3-0.7 and 0.9.

Looking at complexity figures (Figs 4, 5(b), 6(b), and 7(b)) we can observe that the qualitative behavior of our estimation of an upper bound for the complexity is similar for all the data sets.

Analyzing and comparing the results for all the data sets, we can realize that the combination of the selection method and the substitution method is the key factor. Substituting the most frequent words using the asterisk substitution method is always the best option to maintain the most relevant information. That is, performing an annealing text distortion using asterisk to distort the documents is the best option, because this distortion preserves the most relevant information. In this case, our estimation of an upper bound for the documents complexity is slowly reduced and therefore the clustering error remains stable even though a considerable percentage of words were substituted from the documents. Its worth mentioning that the non-distorted clustering error can even be improved (see Figs 3(a) and 6(a)).

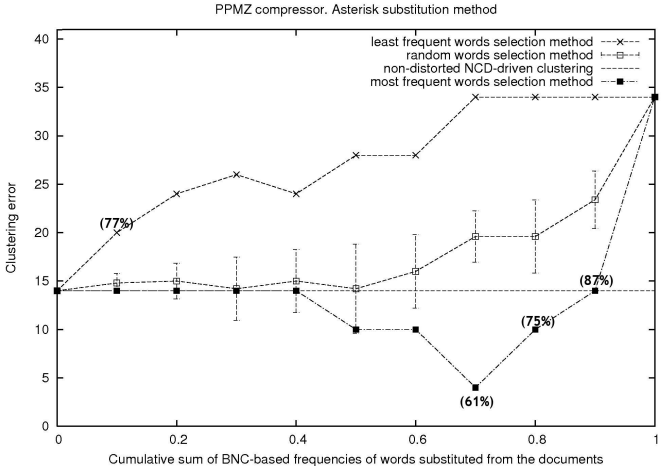


(a) UCI-KDD data set. Clustering error.

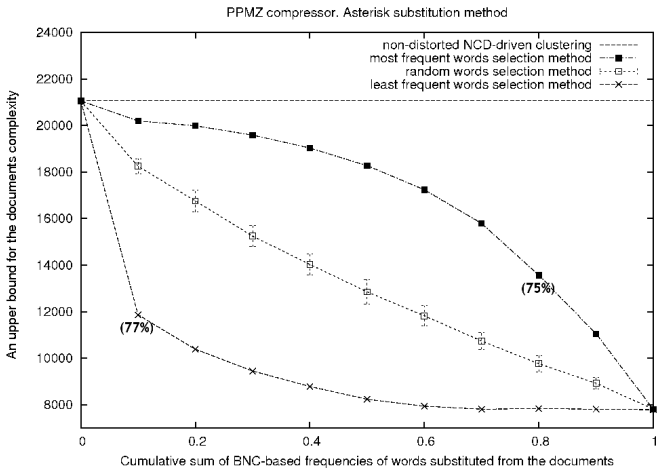


(b) UCI-KDD data set. Estimation of an upper bound for the documents complexity.

Fig. 5. UCI-KDD data set using the PPMZ compressor. Clustering error and estimation of an upper bound for the documents complexity obtained for all the word selection methods when the asterisk substitution method is used. The best clustering results correspond to the MFW selection method. Note that there is a big difference between the clustering errors obtained substituting a similar percentage of the words of the documents using different word selection methods (observe percentages 77% and 79% between brackets). The complexity plot is similar to the one obtained for the case study (see Fig 4).

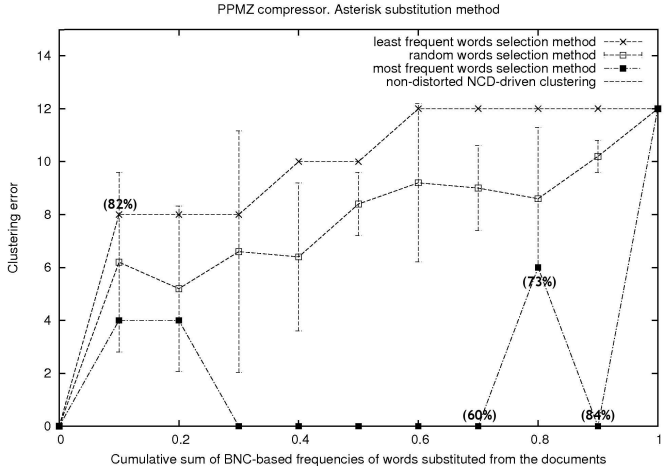


(a) MedlinePlus data set. Clustering error.

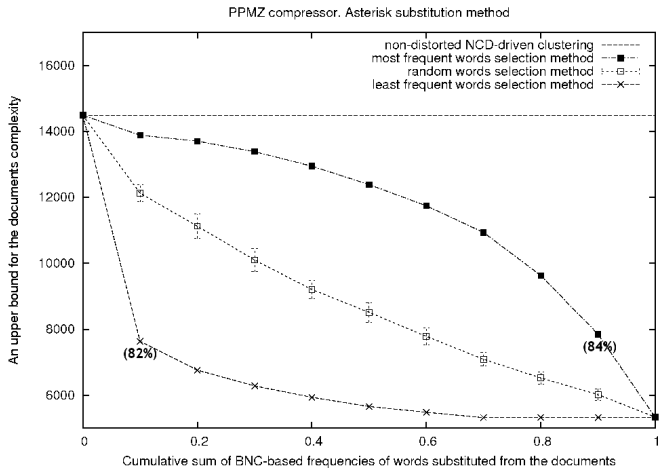


(b) MedlinePlus data set. Estimation of an upper bound for the documents complexity.

Fig. 6. MedlinePlus data set using the PPMZ compressor. Clustering error and estimation of an upper bound for the documents complexity obtained for all word selection methods when the asterisk substitution method is used. The best clustering results correspond to the MFW selection method (observe that from 0.5 to 0.8 the non-distorted NCD-driven clustering is improved). Note that there is a big difference between the clustering errors obtained substituting a similar percentage of the words of the documents using different word selection methods (observe percentages 77% and 75% between brackets).



(a) IMDB data set. Clustering error.



(b) IMDB data set. Estimation of an upper bound for the documents complexity.

Fig. 7. IMDB data set using the PPMZ compressor. Clustering error and estimation of an upper bound for the documents complexity obtained for all the word selection methods when the asterisk substitution method is used. The best clustering results correspond to the MFW selection method (observe that the results obtained for the cumulative sum of frequencies of 0.3-0.7 and 0.9, correspond to the perfect clustering). The complexity plot is similar to the one obtained for the case study (see Fig 4).

C. Results for all compression algorithms

We have shown all the clustering error curves for the PPMZ compression algorithm. Now, we show in three tables a summary of the experimental results for every compression algorithm and every data set when the asterisk substitution method is applied. We only show the clustering error, because as it can be observed looking at Figs 4, 5(b), 6(b), and 7(b) the complexity graphs are always qualitatively similar, therefore we do not represent the values of our estimation of an upper bound for the complexity in any table.

Each table contained in Fig 8 corresponds to a word selection method. Consequently, there is a table for the MFW selection method, another one for the RW selection method, and another one for the LFW selection method. In these tables, each column corresponds to a specific data set, and each row corresponds to a specific compression algorithm. The tables show for every data set and every compression algorithm three different clustering errors (clustErr) and the cumulative sum of frequencies where these clustering errors are obtained (cumSumFreq). These three clustering errors are: the clustering error obtained with no distortion (that is, the clustering error obtained clustering the original documents), the minimum clustering error obtained, and the maximum

clustering error obtained. As well as including these important measures, we mark with a double-box the results that improve the clustering error obtained with no distortion, and with a simple-box the results that maintain this clustering error. These boxes are included to focus the attention on the clustering error improvement.

Comparing Fig 3(a) and the first table showed in Fig 8 (a) can help us to better understand the table. In Fig 3(a) we can observe that the clustering error obtained with no distortion is 5. We can observe as well, that this clustering error is maintained for cumulative sum of frequencies from 0 to 0.8. We can see, as well, that the minimum clustering error is 0 (cumulative sum of frequencies 0.9), and the maximum is 8 (cumulative sum 1.0). All these data are included in the first table showed in Fig 8: this table has two cells for Books and PPMZ. One contains the above mentioned clustering error values (5, 0 and 8), and the other contains the cumulative sum of frequencies in which these clustering error values have been obtained (0.1-0.8, 0.9 and 1.0). Note that we do not take into account the clustering error with no distortion to create the table, because it is obvious that the clustering error corresponding to the cumulative sum of frequencies of 0 will always be the same, and we want to study the effect of the distortion. Therefore, we only consider the results obtained from 0.1 to 1.0.

The table shown in Fig 8 (a) has many boxes because when the MFW selection method is applied, the best results are obtained. This is due to the fact that using this word selection method, the clustering is improved or maintained for every repository and every compression algorithm. These results are consistent with the observed when analyzing the Figs 3, 5(a), 6(a) and 7(a), where it can be observed that the best clustering results correspond to the MFW selection method when applied with the asterisk substitution method. These empirical results demonstrate that this particular combination of selection method and substitution method allows to perform an annealing text distortion.

D. An example of bigger data sets: MedlinePlus

Once we have observed that the non-distorted NCD-based text clustering performance can be maintained or even improved for high cumulative sum of frequencies of substituted words when combining the MFW selection method and the asterisk substitution method, we want to study if the same behavior is observed in bigger data sets. In order to do so, among the previously studied data sets, we have selected the one in which worst non-distorted clustering results were obtained. Thus, we have progressively increased the number of documents of the MedlinePlus data set to make it as big as possible while still using the clustering algorithm developed in [38].

Figs 9(a), and 9(b) show the clustering results obtained when clustering 50 and 60 documents from the MedlinePlus repository, respectively. On the other hand, Fig 10 depicts the best dendrogram that we have obtained when clustering 50 documents. It corresponds to the cumulative sum of frequencies 0.7 in Fig 9(a). Analyzing this dendrogram, it can be noticed that only three documents are incorrectly clustered. These are the ones highlighted in gray.

Looking at figures 9(a), and 9(b) it can be observed that the non-distorted NCD-driven clustering results are not only maintained but even improved in all cases for high cumulative sum of frequencies of substituted words. This suggests that replacing the most frequent words of the language with asterisks helps the compressor to obtain more reliable similarities and therefore improves the clustering results.

Although there is still a limitation in the size of the data sets, due to the fact that the CompLearn uses the quartet tree method to generate the dendrogram and that algorithm has an asymptotical cost of $O(n^3)$ from version 1.1.3. onwards, in the future we want to use the heuristic approach for the quartet method described in [50] to increase the number of documents per data set. Furthermore, we want to use different clustering algorithms so that we can compare all the results to study whether or not the behavior observed in this work is consistent among different clustering algorithms. More details about this can be found in section V.

V. Conclusions and Future Work

In this paper we have moved a small step towards understanding compression distances by performing an experimental evaluation of the impact of several kinds of word removal on compression-based text clustering. Three main contributions have been presented. First, we have given new insights for the evaluation and explanation of the behavior of the compression distance-driven clustering algorithms. Second, we have presented a technique which reduces our upper bound estimation for the Kolmogorov complexity of the documents while preserving most of the relevant information. Third, we have observed experimental evidence of how to fine-tune the representation of the documents using annealing text distortion, in order to obtain better clustering results when using the NCD-driven text clustering.

We have used a clustering method [38] based on the NCD [26] to measure the amount of information contained in the distorted documents. We have used six different replacement methods to distort the documents (see section III-B). These replacement methods are pairwise combinations of two factors: *word selection* and *substitution method*. We have three word selection methods, depending on what words are chosen to be removed from the documents: MFW selection method, LFW selection method and RW selection method. We have two substitution methods, depending on the way in which the words are removed from the documents: *random character* substitution method and *asterisk* substitution method.

We have applied the clustering method over four different data sets repeating the clustering three times using each time a different compression algorithm to calculate the NCD. The same compression algorithms have been used to estimate the Kolmogorov complexity of the documents. The Kolmogorov complexity has been estimated based on the concept that data compression is an upper bound for it. That is, we estimate the upper bound for the Kolmogorov complexity as the length of the compressed file in bytes.

		Books		UCI-KDD		MedlinePlus		IMDB	
		clustErr	cumSumFreq	clustErr	cumSumFreq	clustErr	cumSumFreq	clustErr	cumSumFreq
LZMA	NonDistorted	4	0.1-0.7	0	0.1-1.0	14	0.1-0.6	18	0.1-0.6
	Minimum	2	0.8,0.9	0	0.1-1.0	10	0.7-0.8	6	0.7-0.8
	Maximum	9	1.0	0	0.1-1.0	28	1.0	22	1.0
PPMZ	NonDistorted	5	0.1-0.8	0	0.1-0.8	14	0.1-0.4,0.9	0	0.3-0.7,0.9
	Minimum	0	0.9	0	0.1-0.8	4	0.7	0	0.3-0.7,0.9
	Maximum	8	1.0	21	1.0	34	1.0	12	1.0
BZIP2	NonDistorted	7	0.2,0.4-0.6	0	0.1-0.6	14	0.1-0.4,0.6	0	0.1-0.6
	Minimum	5	0.7	0	0.1-0.6	10	0.5,0.7-0.8	0	0.1-0.6
	Maximum	12	0.3	15	1.0	24	1.0	12	1.0

(a) Most frequent word selection method. When this selection method is applied using the asterisk substitution method, the best clustering results are obtained.

		Books		UCI-KDD		MedlinePlus		IMDB	
		clustErr	cumSumFreq	clustErr	cumSumFreq	clustErr	cumSumFreq	clustErr	cumSumFreq
LZMA	NonDistorted	4	0.1-0.6	0	0.1-0.6	14	-	18	-
	Minimum	4	0.1-0.6	0	0.1-0.6	13.8	0.3	13.3	0.7
	Maximum	9	1.0	0.8	0.8	28	1.0	22	1.0
PPMZ	NonDistorted	5	0.1-0.2	0	0.1-0.3	14	-	0	-
	Minimum	5	0.1-0.2	0	0.1-0.3	14.2	0.3,0.5	5.2	0.2
	Maximum	10.7	0.9	21	1.0	34	1.0	12	1.0
BZIP2	NonDistorted	7	-	0	-	14	-	0	-
	Minimum	5.9	0.8	1.6	0.2	14.6	0.3	8.4	0.1
	Maximum	10.5	0.1	17.2	1.0	24	1.0	15.3	0.4

(b) Random word selection method. It can be observed that these clustering results are worse than the ones obtained when the MFW selection method is applied.

		Books		UCI-KDD		MedlinePlus		IMDB	
		clustErr	cumSumFreq	clustErr	cumSumFreq	clustErr	cumSumFreq	clustErr	cumSumFreq
LZMA	NonDistorted	4	-	0	0.1-0.2,0.5-1.0	14	-	18	0.1,0.5
	Minimum	9	0.1-0.4,0.6-1.0	0	0.1-0.2,0.5-1.0	20	0.1-0.2	10	0.6
	Maximum	12	0.5	2	0.3-0.4	28	0.5,0.7-1.0	22	0.2,0.4,0.7-1.0
PPMZ	NonDistorted	5	-	0	-	14	-	0	-
	Minimum	7	0.1-0.2	15	0.3	20	0.1	8	0.1-0.3
	Maximum	11	0.3-0.6	21	0.5,0.7-1.0	34	0.7-1.0	12	0.6-1.0
BZIP2	NonDistorted	7	-	0	-	14	-	0	-
	Minimum	4	0.3-0.4	8	0.3	16	0.1	10	0.6
	Maximum	9	0.1-0.2	16	0.1,0.6	26	0.3-0.4	32	0.1

(c) Least frequent word selection method. These are the worst clustering results obtained.

Fig. 8. These tables show a summary of the experimental results for every compression algorithm and every data set when the asterisk substitution method is used. They show three relevant clustering errors and the cumulative sum of frequencies where these clustering errors are obtained. These three clustering errors are: the clustering error obtained with no distortion, the minimum clustering error obtained, and the maximum clustering error obtained. The results that improve the clustering error obtained with no distortion are highlighted inside a double-box. The results that maintain the non-distorted clustering error are highlighted inside a simple-box.

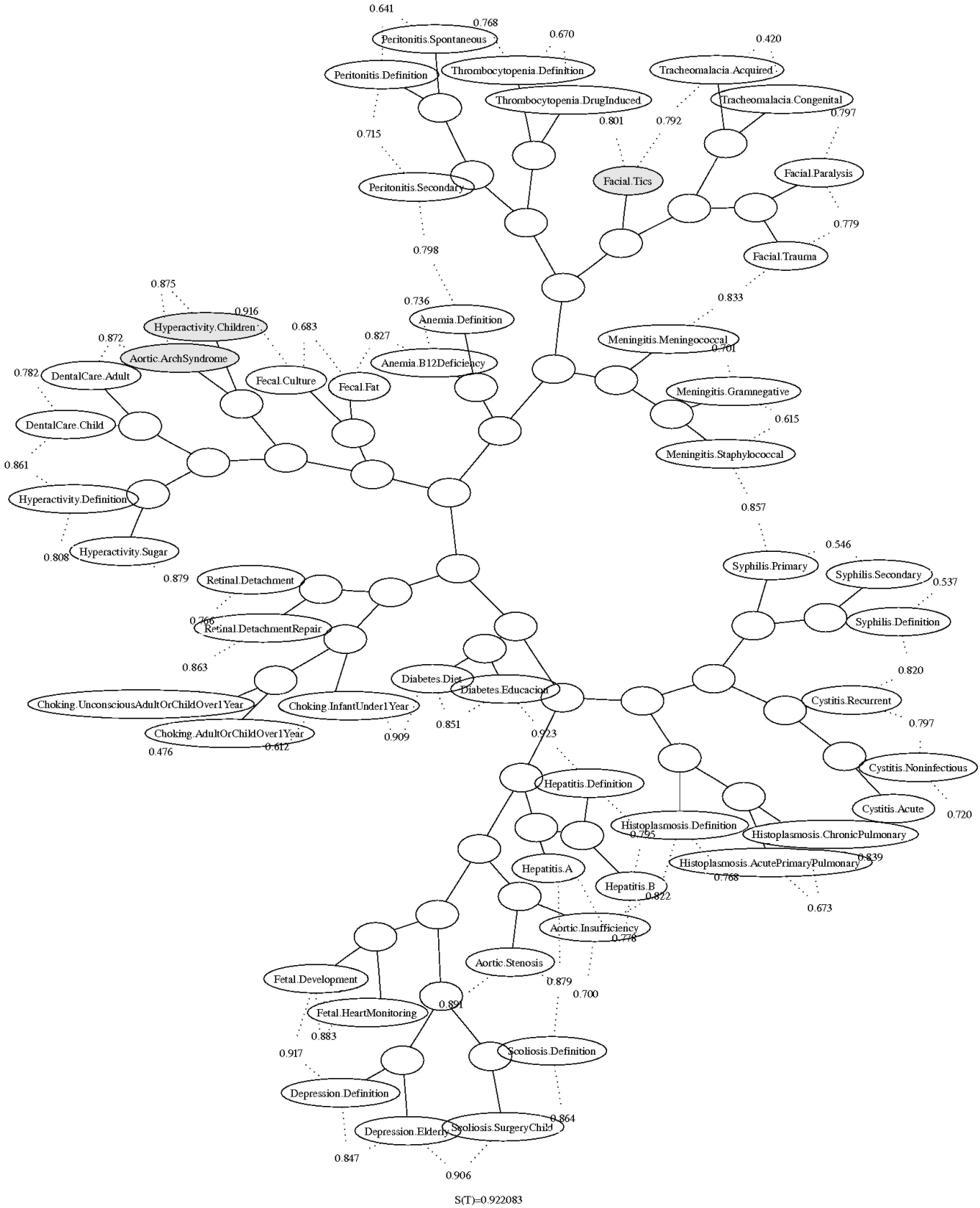
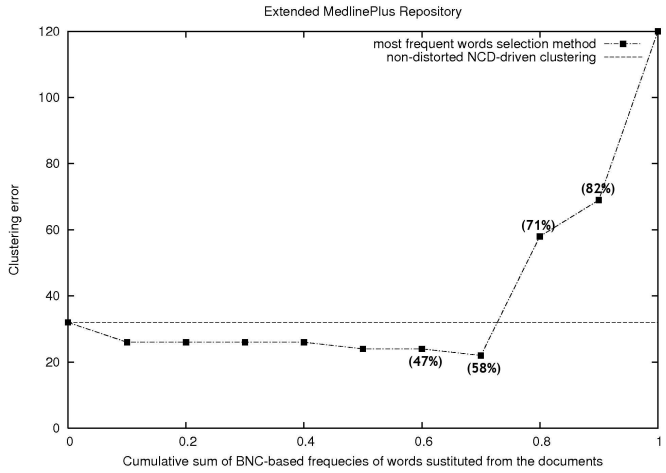
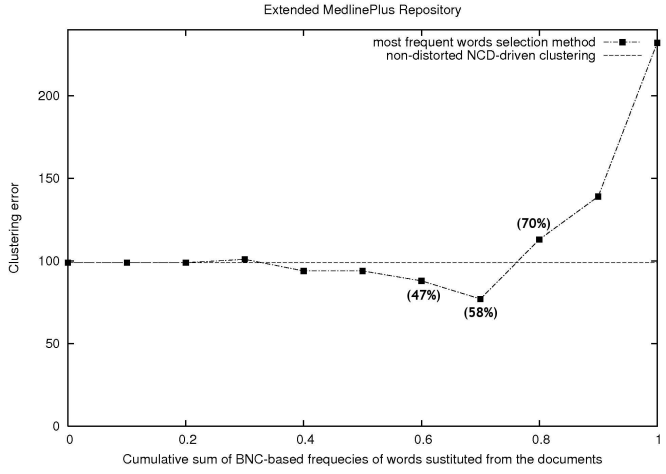


Fig. 10. Example of dendrogram for the extended MedlinePlus repository. The numbers in the image represent the NCD average between two nodes. The dendrogram corresponds to the figure 9(b), in particular, it corresponds to the 0.7 cumulative sum of frequencies for the most frequent words selection method graph. The documents which have not been correctly clustered are highlighted in gray. Note that only three documents have not been correctly clustered: $Facialtics$, $Aorticarchsyndrome$ and $Hyperactivityandchildren$.



(a) Number of documents = 50



(b) Number of documents = 60

Fig. 9. Extended MedlinePlus data set. Clustering error obtained when clustering the original documents, and when clustering the documents using the MFW selection method and the asterisk substitution method. The numbers between brackets correspond to the percentage of substituted words in the documents. It is worth mentioning that the non-distorted NCD-driven clustering can be improved when the MFW selection method and the asterisk substitution method are applied together to preprocess the documents.

We have observed that carrying out an annealing text distortion is the best way to distort a text, in order to decrease its complexity while conserving its most relevant information. When we remove the most frequent words of the language the distortion is annealed, in the sense that the percentage of removed words increases slowly with respect to the cumulative sum of BNC-frequencies. Thus, the best clustering results are obtained when combining the MFW selection method and the asterisk substitution method. In that case, the experimental results show how the clustering error is maintained even when the percentage of replaced words is strongly increased. It seems that we are replacing precisely the least relevant parts of the documents. This, in turn, helps the compression algorithm to estimate the complexity of the documents in an accurate manner. Consequently, the compressor obtains more reliable similarities. In fact, the clustering can be improved by removing the non-relevant information from the documents, because this removal helps the compressor to better find the relevant similarities among the documents.

For the other word selection methods (LFW and RW selection methods), the clustering error increases faster even though the documents complexity is also reduced. Thus, it seems that the information that has been replaced is relevant in the clustering process, and consequently we are losing important information. As a consequence, the similarities among the documents are not being correctly measured.

We have observed that the obtained results are consistent across the most important compression families (Lempel-Ziv, Statistical and Block-Sorting) and across different data sets.

An important issue for further investigation could be doing an analytic study of our experimental work. We can address this issue using the information bottleneck method [51], which has been already applied to document clustering [52], [53]. The information bottleneck method is a technique for finding the best tradeoff between accuracy and complexity (compression) when summarizing (clustering) a random variable X, given a joint probability distribution between X and an observed relevant variable Y. There are two challenges that have to be addressed to apply the information bottleneck method. First, the rate distortion function has to be calculated. Second, the probability density function has to be related with a particular algorithmic transformation of the source. The above mentioned rate distortion function filters the information on the documents. In our opinion, the conclusions of our work could be used to specify a rate distortion function in terms of the frequency of the words. We will study this issue in further investigations.

In the future, we want to apply the annealed distortion to non-textual data sets using the idea created for time series in [54] in which the authors propose a way to represent any kind of information to be discrete. After representing the information in a discrete way, we will estimate the frequency of every symbol in order to be able to apply the annealed distortion, since the annealed distortion presented in our work is based on the frequencies of the symbols (words in the English language). Using these ideas, we will apply the annealed distortion to images and music.

On the other hand, we plan to study other distortion methods that do not maintain the initial length of the documents. We also want to compare the NCD with other similarity distances, like Vector Space Model [55] or Kullback-Leibler distance [56], [57], and with the Compression-Based Dissimilarity Measure as well [54]. As we said in section III-A the number of documents per experiment is relatively small. We will use other clustering algorithms like heuristic approach to quartet tree method [50], K-means [58] or Support Vector Machines [59] to test our approach in larger data sets.

APPENDIX

Here, we briefly enumerate the different data sets used in the first phase of our experiments. All of them comprise texts written in English.

- Fourteen classical books. We try to cluster them by author. We have two books by Agatha Christie: *The Secret Adversary*, and *The Mysterious Affair at Styles*. Three books by Alexander Pope: *An Essay on Criticism*, *An Essay on Man*, and *The Rape of the Lock, an heroic-comical Poem*. Two books by Edgar Allan Poe: *The Fall of the House of Usher*, and *The Raven*. Two books by Miguel de Cervantes: *Don Quixote*, and *The Exemplary Novels*. Three books by Niccolò Machiavelli: *Discourses on the First Decade of Titus Livius*, *History of Florence and of the Affairs of Italy*, and *The Prince*. Two books by William Shakespeare: *The tragedy of Antony and Cleopatra*, and *Hamlet*.
- Sixteen messages from a newsgroup (UCI-KDD) [39]. We try to cluster them by topic. We have three documents on atheism, three documents on Christian religion and homosexuality, two documents on Christian religion and reincarnation, two documents on politics and guns, three documents on cryptography, governs and communications, and three documents on inherent problems of cryptography.
- Twelve documents from the MedlinePlus repository [40]. We try to cluster them by topic. We have three documents related with alcohol: alcohol use, alcoholic neuropathy, and alcoholism. Three documents on diabetes: diet, education, and definition. Three documents on meningitis: gramnegative, meningococcal, and staphylococcal. Three documents on tumors: hepatocellular carcinoma, spinal tumor, and thyroid cancer.
- Fourteen plots of movies from the Internet Movie Data Base (IMDB) [41]. We try to cluster them by saga. We have the saga of Indiana Jones: *Raiders Of The Lost Ark*, *Temple Of The Doom*, and *The Last Crusade*. We have the saga of Pirates Of The Caribbean: *The Curse of the Black Pearl*, *Dead Man's Chest*, and *At World's End*. We have the initial saga of Star Wars: *A New Hope*, *The Empire Strikes Back*, and *Revenge of the Jedi*. We have the saga of The Matrix: *The Matrix*, *Matrix Reloaded*, and *Matrix Revolutions*. We have the saga of The Mummy: *The Mummy*, and *The Mummy Returns*.

In the second phase, we have used two different sets of documents from the MedlinePlus repository [40]:

- Fifty documents: there are two documents related with anemia: definition and B12 deficiency. Three documents on aortic: arch syndrome, aortic insufficiency, and aortic stenosis. Three documents on choking: choking adult or child over one year, choking infant under one year, and choking unconscious adult or child over one year. Three documents on cystitis: cystitis acute, cystitis noninfectious and cystitis recurrent. Two documents on dental care: adult and child. Two documents on depression: definition and depression elderly. Two documents on diabetes: diabetes diet, and diabetes education. Three documents on facial: facial paralysis, facial tics and facial trauma. Two documents on fecal: fecal culture and fecal fat. Two documents on fetal: fetal development and fetal heart monitoring. Three documents on hepatitis: hepatitis A, hepatitis B and definition. Three documents on histoplasmosis: histoplasmosis acute primary pulmonary, histoplasmosis chronic pulmonary and definition. Three documents on hyperactivity: hyperactivity and children, hyperactivity and sugar and definition. Three documents on meningitis: gramnegative, meningococcal and staphylococcal.Three documents on peritonitis: peritonitis secondary, peritonitis spontaneous and definition. Two documents on retinal: retinal detachment and retinal detachment repair. Two documents on scoliosis: scoliosis surgery child and definition. Three documents on syphilis: syphilis primary, syphilis secondary and definition. Two documents on thrombocytopenia: thrombocytopenia drug induced and definition. Two documents on tracheomalacia: tracheomalacia acquired and tracheomalacia congenital.
- Sixty documents: we have added ten new documents to the data set of fifty documents described previously. These new documents are: three documents on immunoelectrophoresis: immunoelectrophoresis plasma and urine, immunoelectrophoresis serum, and immunoelectrophoresis urine. Three documents on paroxysmal: paroxysmal cold hemoglobinuria, paroxysmal nocturnal hemoglobinuria, and paroxysmal supraventricular tachycardia. Two documents on methylmalonic: methylmalonic acidemia, and methylmalonic acid test. Two documents on necrotizing: necrotizing enterocolitis, and necrotizing vasculitis.

ACKNOWLEDGMENT

We would like to thank Francisco Sánchez for his useful comments on the draft. We would like to thank the anonymous referees for their constructive comments on the manuscript.

REFERENCES

[1] R. L. Cilibrasi and P. M. Vitanyi, "The google similarity distance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, 2007.

[2] X. Zhang, Y. Hao, X. Zhu, and M. Li, "Information distance from a question to an answer," in *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2007, pp. 874–883.

[3] D. Ravichandran and E. Hovy, "Learning surface text patterns for a Question Answering system," *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 41–47, 2001.

[4] X. Chen, B. Francia, M. Li, B. McKinnon, and A. Seker, "Shared information and program plagiarism detection," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1545–1551, 2004.

[5] C. Ané and M. Sanderson, "Missing the Forest for the Trees: Phylogenetic Compression and Its Implications for Inferring Complex Evolutionary Histories," *Systematic Biology*, vol. 54, no. 1, pp. 146–157, 2005.

[6] H. Otu and K. Sayood, "A new sequence distance measure for phylogenetic tree construction," *Bioinformatics*, vol. 19, no. 16, pp. 2122–2130, 2003.

[7] A. Kocsor, A. Kertesz-Farkas, L. Kajan, and S. Pongor, "Application of compression-based distance measures to protein sequence classification: a methodological study," *Bioinformatics*, vol. 22, no. 4, pp. 407–412, 2006.

[8] N. Krasnogor and D. Pelta, "Measuring the similarity of protein structures by means of the universal similarity metric," *Bioinformatics*, vol. 20, no. 7, pp. 1015–1021, 2004.

[9] H. Pao and J. Case, "Computing Entropy for Ortholog Detection," *ICCI 2004: International Conference on Computational Intelligence*, 2004.

[10] D. Benedetto, E. Caglioti, and V. Loreto, "Language Trees and Zipping," *Physical Review Letters*, vol. 88, no. 4, p. 48702, 2002.

[11] M. Cuturi and J. Vert, "The context-tree kernel for strings," *Neural Networks*, vol. 18, no. 8, pp. 1111–1123, 2005.

[12] K. Emanuel, S. Ravela, E. Vivant, and C. Risi, "A combined statistical-deterministic approach of hurricane risk assessment," *Bulletin of the American Meteorological Society*, vol. 87, no. 3, pp. 299–314, 2006.

[13] T. Arbuckle, A. Balaban, D. Peters, and M. Lawford, "Software documents: comparison and measurement," in *SEKE '07: Proceedings of the 18th Int. Conf. on Software Engineering and Knowledge Engineering*, 2007.

[14] E. Allen, T. Khoshgoftaar, and Y. Chen, "Measuring Coupling and Cohesion of Software Modules: An Information-Theory Approach," *Seventh International Software Metrics Symposium*, 2001.

[15] W. T. Scott, "A new approach to data mining for software design," in *CSITeA '04: Proceedings of the Int. Conf. on Computer Science, Software Engineering, Information Technology, e-Business, and Applications*, 2004.

[16] R. Cilibrasi, P. Vitanyi, and R. de Wolf, "Algorithmic clustering of music," *Web Delivering of Music, 2004. WEDELMUSIC 2004. Proceedings of the Fourth International Conference on*, pp. 110–117, 2004.

[17] A. Kraskov, H. Stoegbauer, R. Andrzejak, and P. Grassberger, "Hierarchical clustering using mutual information," *Europhysics Letters*, vol. 70, no. 2, pp. 278–284, 2005.

[18] C. Santos, J. Bernardes, P. Vitanyi, and L. Antunes, "Clustering fetal heart rate tracings by compression," in *CBMS '06: Proceedings of the 19th IEEE Symposium of Computer-Based Medical Systems*, 2006, pp. 685–690.

[19] D. Parry, "Use of kolmogorov distance identification of web page authorship, topic and domain," *Open Source Web Information Retrieval*, 2005.

[20] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," Dortmund University, Tech. Rep., 1997.

[21] E. Leopold and J. Kindermann, "Text categorization with support vector machines. how to represent texts in input space?" *Machine Learning*, vol. 46, no. 1-3, pp. 423–444, 2002.

[22] C. Faloutsos and V. Megalooikonomou, "On data mining, compression, and Kolmogorov complexity," *Data Mining and Knowledge Discovery*, vol. 15, no. 1, pp. 3–20, 2007.

[23] R. Martínez, M. Cebrián, F. de Borja Rodríguez, and D. Camacho, "Contextual information retrieval based on algorithmic information theory and statistical outlier detection," in *IEEE Information Theory Workshop*, 2007.

[24] D. Salomon, *Data Compression: The Complete Reference*. Springer, 2004.

[25] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, "The similarity metric," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[26] R. Cilibrasi and P. Vitanyi, "Clustering by compression," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1523–1545, 2005.

[27] J. Seward, *BZIP2*, [Online] Available: http://bzip.org/.

[28] I. Pavlov, *LZMAX*, [Online] Available: http://www.7-zip.org/sdk.html.

[29] C. Bloom, *PPMZ*, [Online] Available: http://www.cbloom.com.

[30] M. Cebrián, M. Alfonseca, and A. Ortega, "Common pitfalls using the normalized compression distance: What to watch out for in a compressor," *Commun. Inf. Syst.*, vol. 5, no. 4, pp. 367–384, 2005.

[31] M. Cebrian, M. Alfonseca, and A. Ortega, "The normalized compression distance is resistant to noise," *IEEE Transactions on Information Theory*, vol. 53, no. 5, pp. 1895–1900, 2007.

[32] S. Verdú and T. Weissman, "The information lost in erasures," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 5030–5058, November 2008.

[33] S. Fong, D. Roussinov, and D. Skillicorn, "Detecting word substitutions in text," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 8, pp. 1067–1076, 2008.

[34] A. Turing, "On computable numbers, with an application to the entscheidungsproblem," *Proceedings of the London Mathematical Society*, vol. 2, no. 42, pp. 230–265, 1936.

[35] A. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems Information Transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[36] M. Li and P. Vitányi, *An introduction to Kolmogorov complexity and its applications*, 2nd ed. Springer-Verlag, 1997.

[37] M. Sipser, *Introduction to the Theory of Computation*, 2nd ed. PWS Publishing, 2006.

[38] R. Cilibrasi, A. L. Cruz, S. de Rooij, and M. Keijzer, *CompLearn Toolkit*, [Online] Available: http://www.complearn.org/.

[39] *UCI Knowledge Discovery in Databases Archive*, Information and Computer Science, University of California, Irvine. [Online] Available: http://kdd.ics.uci.edu/.

[40] *MedlinePlus Health Information, MedlinePlus website*, U.S. National Library of Medicine and National Institutes of Health. [Online] Webpage: http://medlineplus.gov/.

[41] *IMDB, Internet Movie Database*, [Online] Available: http://www.imdb.com/.

[42] Y. Yang, "Noise reduction in a statistical approach to text categorization," in *Proceedings of SIGIR*, 1995, pp. 256–263.

[43] C. Van Rijsbergen, *Information Retrieval*. Butterworth-Heinemann Newton, MA, USA, 1979.

[44] G. Salton, *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989.

[45] W. Wilbur and K. Sirotkin, "The automatic identification of stop words," *Journal of Information Science*, vol. 18, no. 1, p. 45, 1992.

[46] B. N. C. Consortium, *British National Corpus*, Oxford University Computing Services [Online] Available: http://www.natcorp.ox.ac.uk/.

[47] D. W. M. Burrows, "A block-sorting lossless data compression algorithm," *Digital Systems Research Center Research Report*, vol. 124, 1994.

[48] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proc. Inst. Radio Engineers*, vol. 40(9), pp. 1098–1101, 1952.

[49] A. Granados, M. Cebrián, D. Camacho, and F. de Borja Rodríguez, "Evaluating the impact of information distortion on normalized compression distance," in *Proceedings of the 2nd International Castle Meeting on Coding Theory and Applications (ICMCTA)*, ser. Lecture Notes in Computer Science, A. Barbero, Ed., vol. 5228. Springer-Verlag, 2008, pp. 69–79.

[50] S. Consoli, K. Darby-Dowman, G. Geleijnse, J. Korst, and S. Pauws, "Heuristic approaches for the quartet method of hierarchical clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, no. 1, 5555.

[51] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377.

[52] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *ACM SIGIR 2000*. ACM press, 2000, pp. 208–215.

[53] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised document classification using sequential information maximization," in *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2002, pp. 129–136.

[54] E. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Towards parameter-free data mining," in *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2004, pp. 206–215.

[55] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley, 1999.

[56] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[57] S. Kullback, "The kullback-leibler distance," *The American Statician*, vol. 41, pp. 340–341, 1987.

[58] J.A. and M. Wong, "A k-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[59] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, 2001.