Introducing Biases in Document Clustering

Yunior Ramírez-Cruz

Center for Pattern Recognition and Data Mining,
Content Management Systems Division, DATYS, Santiago de Cuba,
Cuba

yunior@cerpamid.co.cu

Abstract. In this paper, we present three criteria for introducing biases in document clustering algorithms, when information characterizing the document collections is available. We focus on collections known to be the result of a document categorization or sample-based document filtering process. Our proposals rely on profiles, i.e., document samples known to have been used for obtaining the collection, to extract statistics which determine the biases to introduce. We conduct an experimental evaluation over a number of collections extracted from the widely used corpus RCV1, which allows us to confirm the validity of our proposals and determine a number of situations where biased clusterings, according to different criteria, outperform their unbiased counterparts.

Keywords. Document clustering, introduc biases.

Introducción de sesgos en el agrupamiento de documentos

Resumen. En este artículo se presentan tres criterios para la introducción de sesgos en algoritmos de agrupamiento de documentos, cuando se dispone de información que caracteriza las colecciones de documentos. Nos concentramos en colecciones de las que se conoce que son el resultado de un proceso de categorización o filtrado de documentos basado en muestras. Nuestras propuestas utilizan perfiles, es decir muestras de documentos de las que se conoce que han sido utilizadas para obtener la colección, para extraer estadísticos que determinan los sesgos a introducir. Llevamos a cabo una evaluación experimental sobre un conjunto de colecciones extraídas del corpus ampliamente utilizado RCV1, que nos permiten confirmar la validez de nuestras propuestas y determinar un número de situaciones donde los agrupamientos sesgados según diferentes criterios superan a sus contrapartes no sesgadas.

Palabras clave. Agrupamiento de documentos, introducción de sesgos.

1 Introduction

As the World Wide Web grows the amount of available digital information increases exponentially. This overabundance has brought about the necessity of relying on automated techniques to adequately handle, process, access, organize and present this information in order to aid users in satisfying their information needs. Central to these automated techniques is Data Mining, which enables users to sift through large data repositories to find concise pieces of relevant information according to their interests.

Since a significant part of the information currently available consists in text documents, Text Mining has become a particularly important branch of Data Mining. In this article we focus on one Text Mining task, document clustering, which consists in dividing a document collection into a set of groups reflecting some aspects of its inner structure. Because of its usefulness in facilitating navigation of large document collections, thus reducing the burden of human information analysts or computational systems performing other highly time-consuming Text Mining tasks, document clustering has earned an enormous importance.

Generally, clustering algorithms assume not to have any information about the inner properties of the collection. However, practical situations arrive where some sort of information is available, such as the origin of the collection, the areas it covers, etc. The sole particular case that has attracted significant attention is that of clustering a subset of the results of Web searches [1]. In general terms, these systems, called *clustering engines*, view clustering as a source of complementary information rather than the desired output.

http://dx.doi.org/10.13053/CyS-18-1-2014-024

Clustering engines cluster the results of query-based Web searches as a by-product of a process aimed to generate short, meaningful, readable labels which serve as potential query reformulations. Being the generation of these labels the main purpose, special algorithms have been developed which usually treat final clusters simply as the sets of documents containing these phrases, presented as the potential results of the new reformulated query.

In this paper, we focus on a different particular case: clustering documents known to belong to a certain category as the result of a document categorization and/or sample-based document filtering process. There exist an important number of real life cases where this situation arises, especially in large scale document classification or filtering systems, where some supervised classification algorithm is used for filtering relevant documents or choosing documents covering a number of topics or belonging to specific classes.

Since information streams become increasingly larger, the number of documents delivered by the classifier, while being a very small percentage of the general stream, may still contain tens or hundreds of thousands of documents, which are not directly easy to process by human analysts, thus calling for the need to apply supplementary text mining techniques, for instance, applying clustering to structure and organize the filtered documents. In general, clustering algorithms make no assumptions regarding the document collection, thus treating unfiltered highly heterogeneous collections in the same manner as considerably more homogeneous collections like those that have been determined to belong to a class by a classifier.

Here, we explore several ways in which the information that should have been initially used to build the classifiers may be additionally used to bias the results of clustering algorithms when applied on the results of the classification process and assess the extent to which these biases allow the clustering algorithms to better structure focused, topic-centered collections, uncover previously implicit information, etc. In this paper, we build on preliminary results reported in [2], and present an extended analysis by conducting new experiments that highlight a number of important facts regarding the behavior of our proposals, especially those concerning their potential use-value.

The ideas discussed in this work cover one of the ways in which document clustering and document categorization may interact. Several authors have addressed some other forms of interaction. For example, Kyriakopoulou and Kalamboukis [3] use clustering as a pre-processing step for categorization, thus applying the classifiers to sets of documents (clusters) rather than to individual documents. Kalton *et al*. [4] address the relation between clustering and classification by generalizing the philosophy of the *k*-means algorithm [5] into a framework where clustering is treated as a process of iteratively optimizing a supervised classifier.

We do not attempt to integrate clustering and categorization algorithms into a single set of techniques, so the kind of interrelation between document clustering and categorization addressed in this paper should not be confused with these approaches. Our focus is on profiting from the fact that some set of common subjects are very likely to be treated in the collections to be clustered (which is in turn due to the fact that the collections are known to be the result of a categorization process) to obtain biased clustering which somehow take this commonalities into account. Finally, we should note that, despite the coincidentally common use of the term *biased*, there is no relation between our work and clustering algorithms referred to as *density-biased* [6], which are algorithms that, at some point, randomly draw a sample from the dataset using density-biased sampling.

The remainder of the paper is structured as follows. In Section 2, we describe the biasing criteria that we introduce, along with an intuitive argumentation of the sort of new information they are expected to uncover. In Section 3, we describe a series of experiments conducted to assess the effect of these biases on the outputs of several clustering algorithms. Finally, we present our conclusions and discuss attractive directions for future work.

2 Biasing Criteria

The fact that the documents belonging to a collection are the result of a categorization or sample-based filtering process strongly implies that these documents must have satisfied certain criteria that led one or several classifiers to put them together. Moreover, the fact that one or several supervised classifiers have been trained in order to perform this task implies that a training set was used, which in a number of cases may be available.

Here, we work on the premise that such a set of documents, different from those in the collection obtained as the result of classification, is available, and that information from these documents may be effectively used to introduce biases in clustering the documents in the collection. We will refer to this set of documents as the *profile*.

In what follows, we consider documents to be represented in the vector space model [7]. That is, a document *d* is represented as a vector $d = \left(w_1^d, w_2^d, \ldots, w_N^d\right)$ where *N* is the number of terms of the vocabulary used in the collection and w_i^d is the weight assigned to term t_i in document *d*. Our sole initial assumption regarding the weighting scheme to be used is that the more important, or useful, a term is for adequately describing the documents where it occurs, the higher it must be weighted.

For every collection *C*, we assume the availability of the profile *P*, composed by a set of *K* documents p_1, p_2, … , p_K, different from those in the collection.

Biases are introduced following two types of criteria, both built on the idea that most classifiers somehow rely on coinciding terminology between the documents to classify and those in the training set, whether by using similarity or distance functions to directly compare documents or by estimating parameters based on these coincidences.

The first type of bias aims to partially mitigate the effect of class-specific common terms in rendering documents highly similar. We will refer to this as the *marginal information biasing* (*MIB*) criterion. The second type of bias pursues the opposite effect, that is, terms found to be more descriptive of the underlying common information are allowed to exert a greater influence in the results of clustering by contributing more to similarity measures. We consider two variants for accounting for how descriptive a term may be considered. We will refer to both variants jointly as the *highly descriptive information biasing* (*HDIB*) criteria.

2.1 Marginal Information Biasing Criteria

In a heterogeneous collection, high similarity is a very strong reason for placing two documents in the same cluster. However, collections obtained as the result of categorization or filtering are in principle known to cover some specific subject, thus being less heterogeneous and featuring some degree of background collection-specific similarity. Our first biasing criterion aims to eliminate a part of this background similarity in order to lead clustering algorithms to concentrate in the heterogeneous aspects still present in the collection.

We refer to this criterion as marginal information bias because, by eliminating a part of the contribution of highly descriptive terms, the rest of the terms, i.e., those that would otherwise be considered as marginal, are allowed to play a more influential role in results obtained by clustering algorithms.

Let $d = \left(w_1^d, w_2^d, \ldots, w_N^d\right)$ be the representation of a document belonging to the collection and let $r = \left(w_1^r, w_2^r, \ldots, w_N^r\right)$ be a vector representing the profile *P*, such that

$$w_i^r = \frac{1}{K} \sum_{j=1}^{K} w_i^{p_j} . \qquad (1)$$

That is, the representative is the average of all vectors in the profile.

The new, biased representation of *d* will be a vector $d_b = \left(w_1^{d_b}, w_2^{d_b}, \ldots, w_N^{d_b}\right)$, where

http://dx.doi.org/10.13053/CyS-18-1-2014-024

$$w_i^{d_b} = \begin{cases} w_i^d - w_i^r & \text{if } w_i^d > w_i^r \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

According to this criterion, the weight of a term in the biased representation is diminished by an amount determined by the average weight of the term in the profile. Thus, the more influential a term may be expected to be in the decision of classifying a document as belonging to the class represented by the profile, the more it is penalized when constructing the biased vector. As a result of this, the influence of such terms in making biased document representations similar is lower, so intuitively we expect their influence in the results of clustering algorithms to be also lower.

2.2 Highly Descriptive Information Biasing Criteria

As we mentioned previously, the purpose of these criteria is to modify the weights of terms in the vectors representing the document in such a way that those terms that best describe the collection, according to information extracted from the profile, are allowed to contribute more to similarity between documents.

In general, both criteria rely on probabilistic statistics often used in language modeling. Let $d = \left(w_1^d, w_2^d, \ldots, w_N^d\right)$ be the representation of a document belonging to the collection. The new, biased representation will be a vector $d_b = \left(w_1^d b_1, w_2^d b_2, \ldots, w_N^d b_N\right)$, where the coefficients b_i introduce term-specific biases.

In the first variant, the bias coefficient is the probability of observing the term t_i in the profile, calculated by *adding-one*-smoothed maximum likelihood [8] as follows:

$$b_i = \Pr\left(t_i \mid P\right) = \frac{count(t_i, P) + 1}{\sum_{j=1}^{N} count(t_j, P) + N} \tag{3}$$

where *count*(t_i | P) is the number of occurrences of term t_i in documents belonging to the profile and N is the size of the vocabulary, i.e., the number of different terms occurring in the language.

We will refer to this variant as the *term probability variant* of the HDIB criterion, *HDIB-Prob* for short.

In document categorization and sample-based filtering tasks, profiles are often likely to be composed by a relatively small number of documents. Because of this, we introduce smoothing to probability calculation so terms not occurring in the profile do not yield zero-valued biased weights. In these cases, the small but non-zero bias coefficient $\frac{1}{N}$ will cause these terms to play a diminished role in rendering documents similar, but will still allow the original unbiased weights to have a contribution to document similarity. Besides, as the biased weight depends on both the bias coefficient and the original unbiased weight, different terms are still allowed to have individual behaviors, which would be lost if a zero-valued bias were applied.

According to this first variant, terms that are not very probable in the profile will have their original weights considerably diminished. It should be noticed that, since probabilities are always in the range (0, 1), all terms will have their weights diminished to some extent according to this criterion. The difference lies in the fact that the diminution applied to high probability terms is considerably smaller.

The second variant of the HDIB criterion aims to favor terms that may be considered distinctive of the profile. We assess the distinctiveness of a term by comparing the probability of observing it in the profile to that of observing it in a model of the general language, that is

$$b_i = \frac{\Pr\left(t_i \mid P\right)}{\Pr\left(t_i \mid L\right)} \tag{4}$$

where Pr(t_i | P) is the probability of observing the term in the profile and Pr(t_i | L) is the probability of observing the term in the general language. Both probabilities are calculated using *adding-one*-smoothed maximum likelihood as in Equation 3. In our case, by *general language* we understand a global document collection from where all potential collections and its associated profiles may be extracted. Alternatively, the model of the general language may be estimated from a separate large collection whose terminology is reasonably general.

According to this variant, distinctive terms, i.e., those that are more probable in the profile than in the general language, have their weights increased by an amount proportional to this distinctiveness measure. On the other hand, terms that are more common in the general language will have their weights diminished accordingly, whereas the weights of terms whose distribution is very similar in both will suffer little or no modification.

We will refer to this variant as the *term distinctiveness variant* of the HDIB criterion, *HDIB-Dist* for short.

3 Experimental Evaluation

We established an experimental evaluation to determine the effect of the different biasing criteria. To facilitate reproducibility, we used a standard document collection, on which we created a simulated environment reflecting the characteristics of the problem we treat and defined a rationale for determining whether the effect of introducing some bias is beneficial, use-valuable, etc.

3.1 Experimental Setup

To construct our experimental environment, we used the standard corpus *Reuters Corpus Volume 1* (RCV1) [9]. This corpus is composed by news-stories published by the press agency Reuters between August 20th, 1996 and August 19th, 1997, which were released for research purposes. The corpus totals 806,792 documents and is currently administered and distributed by the American National Institute for Standards and Technology (NIST). The corpus we used, Volume 1, is composed exclusively by English language documents, whereas its counterpart Volume 2 contains documents in thirteen languages: Dutch, French, German, Chinese, Japanese, Russian, Portuguese, Spanish, Latin American Spanish, Italian, Danish, Norwegian, and Swedish. Although these documents cover the same time period, they are not translations of the documents in Volume 1 and no guarantee is provided that any document in Volume 1 has equivalents in any particular languages of Volume 2.

In this experimentation, we limited ourselves to English documents only, leaving multi-lingual and cross-lingual issues for further research.

In RCV1, documents are labeled with one or several *Topics* categories, which are organized into a taxonomic tree. Documents belonging to a particular category are divided into a training set and a test set.

In order to create an evaluation environment that simulated the real life situations we are interested in, we selected 10 out of the 17 non-leaf *Topics* categories such that all their subcategories are leaves of the category tree. Each selected category determined a collection composed by the documents in the test set of the corresponding category. The associated profile consisted of the documents in the category's training set. We used the information regarding the subcategories for structuring the selected collections into subcollections. Additionally, an extra pseudo-subcollection, composed by all documents belonging to the selected category but to none of its subcategories, was considered. We selected the 10 smallest categories (in terms of number of documents) that fulfill the desired conditions. Despite having selected the 10 smallest categories, the constructed collections span a wide range of sizes, as shown in Table 1.

In our preliminary work [2], we conducted experiments on 5 relatively small collections from RCV1 and compared pairs of clusterings using Jaccard's coefficient [10], a symmetric measure of the degree of coincidence between two clusterings. By doing so, we intended to assess the degree of variation introduced by the biases in order to decide whether further study was worth conducting, but disregarded the notion of determining what biasing option was better, or more use-valuable, according to some rationale. Here, we explore further into the latter idea, by introducing the rationale that the best clustering is the one that better fits the subcategory structure of the selected collections, which is thus taken as the gold standard.

Table 1. Description of the document collections from RCV1 used in the experiments

Collection	Collection size	Profile size	Subcollections	Subcollection size
E14	2,112	65	E141	364
			E142	192
			E143	1172
			None	416
E31	2,349	66	E311	1658
			E312	52
			E313	108
			None	571
E13	6,416	187	E131	5492
			E132	922
			None	126
G15	20,309	363	G151	3258
			G152	2072
			G153	2301
			G154	8266
			G155	2086
			G156	258
			G157	1991
			G158	4248
			G159	38
			None	1492
E51	20,639	641	E511	2831
			E512	12234
			E513	2236
			None	3915
C4	22,478	653	C41	11043
			C42	11535
C31	39,451	1,058	C311	4133
			C312	6452
			C313	1074
			None	28402
C17	40,983	1,172	C171	17876
			C172	11202
			C173	2560
			C174	5625
			None	4609
E21	41,875	1,255	E211	15361
			E212	26552
			None	920
C18	51,355	1,462	C181	42169
			C182	4529
			C183	7204
			None	29

Here, we evaluate to what extent the clustering obtained by each biased variant fits the subcategory structure of the selected collection, and how this fit compares to the unbiased variant.

To determine the best fit, we set an evaluation scheme using the standard IR measures precision, recall and F_1 [11]. For a cluster c_{eval} and a subcategory c_{gold}, precision accounts for the ratio of documents correctly placed in c_{eval}, that is, documents in c_{eval} also found in c_{gold}, with respect to the total number of documents in c_{eval}, as follows:

$$precision = \frac{|c_{gold} \cap c_{eval}|}{|c_{eval}|} \tag{5}$$

For c_{eval} and c_{gold}, recall accounts for the ratio of documents in c_{gold} that are placed in c_{eval} as follows:

$$recall = \frac{|c_{gold} \cap c_{eval}|}{|c_{gold}|} \tag{6}$$

Finally, the F_1 measure combines precision and recall in such a way that, if similar values are obtained for precision and recall, the value obtained for the F_1 measure is close to their average; but largely sacrificing one measure in favor of the other is penalized by making the value of the F_1 measure closer to the lowest value. The F_1 measure is defined as follows:

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{7}$$

For evaluating a clustering, a greedy strategy is used to establish a pairing between clusters and subcategories in such a way that the best F_1-scored *subcategory–cluster* pair is determined, then the second best scored pair, and so on until all subcategories have been a assigned a match. If the number of subcategories exceeds the number of clusters, the remaining subcategories are paired with empty pseudo-clusters. The final score assigned to the clustering is the average of F_1 scores over all pairs. Notice that excess clusters are not paired to empty pseudo-subcategories, since the overall largely diminished F_1 values thus obtained mainly reflect inherent defects of clustering algorithms themselves, rather than showing the differences between the ability of the different biasing criteria to enable an algorithm to create a number of clusters that match the known subcategories better than their unbiased counterparts.

To evaluate the effect of introducing biases in a wide range of clusterings, we chose the radius-α-β_0-compact sets algorithm [12], setting a large number of value combinations for the pair of parameters α and β_0, the Single-Pass algorithm [13] and the *k*-means [5] algorithm.

The radius-α-β_0-compact sets algorithm works by constructing the so-called *radius α maximum β_0-similarity graph*. In this graph, the set of documents $d_1, d_2, \ldots, d_{|C|}$ is mapped into the set of nodes $n_1, n_2, \ldots, n_{|G|}$, in such a way that each node n_i represents a document d_i. A directed edge (n_i, n_j) is inserted if $\text{sim}(d_i, d_j) \geq \beta_0$ and $\text{sim}(d_i, d_j) \geq \max\{\text{sim}(d_i, d_k) \mid \forall d_k \neq d_i\} - \alpha$. That is, for every document d_i, the highest similarity value to any other document(s) is determined, then edges are inserted from the node representing d_i to the nodes representing these documents, along with the nodes representing documents whose similarity to d_i differs from the highest value by at most α. Once the graph is constructed, the algorithm proceeds by removing orientation from the edges, finding the connected components of the undirected graph thus obtained, and associating a cluster to every connected component containing the documents represented by its nodes.

The radius-α-β_0-compact sets algorithm is independent of the presentation order of objects and does not require the number of clusters to be provided as a parameter.

In our previous work [2], we had considered three clustering algorithms, the β_0-connected components algorithm [14], the β_0-compact sets algorithm [14] and the extended stars algorithm [15]. In the aforementioned work, we pointed out that the β_0-connected components algorithm shows a tendency of finding excessively large, uncohesive clusters, whereas the β_0-compact sets algorithm and the extended stars algorithm show a tendency of finding cohesive but excessively small clusters. Here, by using the radius-α-β_0-compact sets algorithm, we obtain a tunable combination of both behaviors, which has been experimentally proven to outperform each individual behavior when appropriate values are given to the α parameter [16]. Two extreme cases may be pointed out: if $\alpha = 0$ the algorithm behaves like the classic β_0-compact sets algorithm; whereas for $\alpha \geq \beta_{max} - \beta_0$, where $\beta_{\max}$ is the maximum similarity value in the collection, the algorithm behaves like the classic β_0-connected components algorithm.

The Single-Pass algorithm works incrementally by comparing each new document to the centroids of the currently existing clusters and assigning it to the cluster(s) to whose centroid(s) it is most similar, provided that the similarity value is above a given threshold. If no similar enough centroids are found, a new cluster is created containing the new document. Every time a document is added to a cluster, its centroid is recalculated. Despite existing criticism on the Single-Pass algorithm, such as the fact of being dependent on the presentation order of objects, it was included in the experimentation due to its wide utilization, which is a consequence of its simplicity, as well as its low temporal and spatial complexity.

For its part, the *k*-means algorithm works by randomly creating an initial clustering and applying an iterative optimization process over it. For creating the initial clustering, *k* points are randomly generated, which are assumed to be approximations of the clusters' centroids. In our case, we did not generate random vectors to represent the initial centroids. Instead, we randomly selected *k* different documents from the collection and used them as the initial centroids. Every step of the iterative optimization process consists in reassigning every document to its most similar centroid to obtain a new approximate clustering and recalculating all centroids. This process is repeated until convergence is achieved.

In our experiments, term weights in the unbiased representation are calculated by the standard *tf-idf* weighting scheme [7, 17] as follows:

$$w_i^d = \frac{count(t_i, d)}{|\, d \,|} \log\left(\frac{|\, C \,|}{df(t_i)} \right) \tag{8}$$

where *count*(t_i, *d*) is the number of occurrences of term t_i in document d_i, *df*(t_i) is the number of documents in the collection that contain the term t_i, |*d*| is the size (number of term occurrences) of document *d*, and |*C*| is the size (number of documents) of the collection. We calculated |*C*| and *df*(t_i) on the entire RCV1 corpus, not in the particular collections. Likewise, we used the entire corpus to estimate the model of the general language required by the HDIB-Dist criterion (Equation 4).

Profile representatives used for the MIB criterion are calculated as the average of the standard *tf-idf*-weighted vectors representing all documents in the profile.

For applying every biasing criterion, the unbiased vectors are modified according to Equation 2, 3 or 4, as corresponds, and normalized. Similarity is determined using the cosine measure, which is defined as ignoring, when appropriate, the norms of vectors that are known to be normalized.

As we mentioned before, parameters β_0 and α are required by the radius-α-β_0-compact sets algorithm. Here, we automatically determined three different β_0 values to be used for each collection: a low value, a medium value and a high value.

These values were calculated using bootstrap resampling [18] as follows. For each collection, 10000 resampling iterations were performed. Each iteration consisted on randomly selecting, with replacement, 10000 pairs of documents, and calculating the similarities between their unbiased representations. The obtained similarity values were sorted incrementally and the values on the first, second and third quartiles selected. Thus, every iteration yielded a low β_0 estimate (first quartile), a medium β_0 estimate (second quartile) and a high β_0 estimate (third quartile). After all iterations were completed, all low (medium, high) values were incrementally sorted, and the medians selected as the low (medium, high) estimates to use for the β_0 parameter over the

Table 2. Estimates determined for the three β_0 parameters calculated for each collection

Collection	Low β_0	Medium β_0	High β_0
E14	0.0226739954535	0.050303832055	0.101295564842
E31	0.0391369721497	0.0765094339114	0.131134881948
E13	0.0399003697181	0.0768210021483	0.129021235989
G15	0.0197349026598	0.0380730593604	0.0700967924837
E51	0.00976769336522	0.0202294008107	0.0378787838493
C4	0.00543866348549	0.0124250158598	0.0265710642448
C31	0.00536329468531	0.0124423759324	0.0246880172546
C17	0.00291446092089	0.0103199043968	0.0243165751003
E21	0.00236039805877	0.011660162816	0.0324985687286
C18	0.00610379640282	0.0116080228745	0.0201699824237

collection. Table 2 shows the values obtained for each collection.

For each selected β_0 parameter value, 9 values were used for the α parameter, namely those corresponding to 5%, 10%, 20%, 25%, 1/3, 50%, 2/3, 75% and 100% of the value of β_0.

Considering all combination of choices, 108 radius-α-β_0-compact-sets clusterings were obtained for each collection. Every clustering was characterized by the choice of a biasing option (to use no bias or to use one of the three biasing criteria described in Section 2), a β_0 value and an α value. Once all clusterings were obtained, averaged F_1 values were calculated in order to establish the comparisons.

The β_0 values calculated for the radius-α-β_0-compact sets clusterings were also used as threshold value choices for the for the Single-Pass clusterings.

In the case of the *k*-means algorithm, the value of *k* was set to the number of subcollections that the collections are known to have, including the special case of the pseudo-subcollection containing the documents belonging to the associated category but to none of its subcategories, as shown in Table 1.

Notice that there are no parameter combinations to consider for this algorithm. However, due to the random nature of the algorithm's initial centroid selection phase, for each combination of a collection and a biasing choice we obtained 10 runs, calculated their corresponding averaged F_1 scores and averaged them to obtain the final score.

3.2 Results and Discussion

Figures 1 and 2 depict the behavior of the radius-α-β_0-compact-sets clusterings obtained using the unbiased representation compared to those obtained using the three biased criteria. In Fig. 1, the averaged F_1 values obtained over every collection for the best α, β_0 parameter combination using the unbiased representation is compared to those corresponding to the biased variants for the same parameter combination. In Fig. 2, the best unbiased results are compared to the best biased result for every biased criterion, regardless of the α, β_0 parameter combination for which each one was obtained. In both figures, collections are showed in increasing order according to their size (number of documents) and F_1 values are showed as percentages.

Analyzing these results, several remarks can be made. Firstly, it may be observed that the cases for which the unbiased clusterings fit best the subcategory structures do not always coincide with the cases for which each unbiased variant does.

http://dx.doi.org/10.13053/CyS-18-1-2014-024

Concerning the MIB criterion, Fig. 1 shows that the best performing unbiased clusterings were outperformed by their equivalent MIB-biased clusterings in 6 out of 10 collections; whereas Fig. 2 shows that the best performing MIB-biased clusterings outperformed the best unbiased results in 9 out of 10 collections. Moreover, for larger classes this behavior became more consistent, either comparing the best unbiased results to their equivalent MIB-biased results or to the overall best MIB-biased results, which points to the usefulness of coupling the MIB criterion to

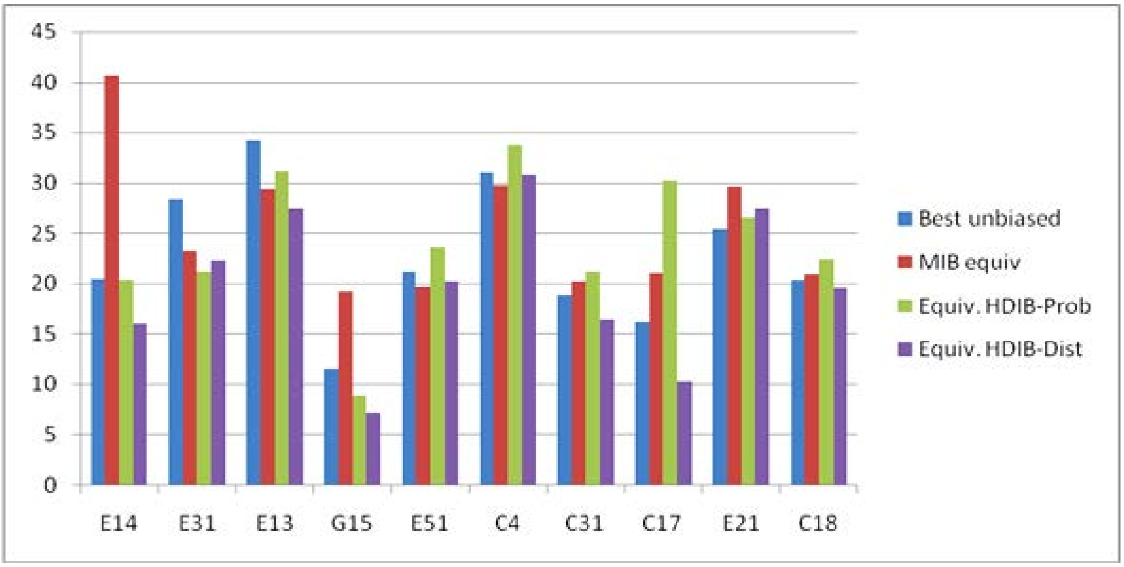

Fig. 1. Best unbiased F_1 values per class and corresponding biased F_1 values on radius-α-β_0-compact-sets clusterings, all shown as percentages

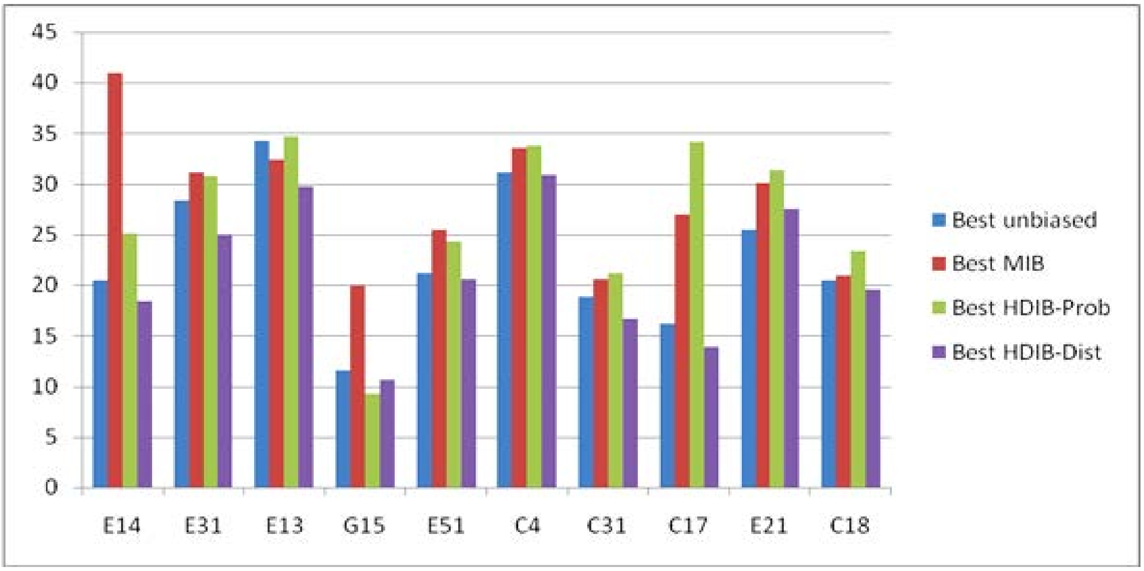

Fig. 2. Best F_1 values per class for all biasing criteria on radius-α-β_0-compact-sets clusterings, all shown as percentages

the radius-α-β_0-compact-sets clustering algorithm in real-life applications, where very large collections are commonplace.

In our opinion, these results back our hypothesis that diminishing the contribution of profile terms to similarity values allowed some marginal features to better guide the algorithm into finding a better fit to the known subcategory structure of these collections.

A more thorough examination allowed us to notice an unexpected side-effect. We estimated the first, second and third quartiles of the distribution of similarity values between pairs of documents, applying a bootstrap resampling procedure as we did for automatically setting the three values for parameter β_0 in Subsection 3.2. Analyzing these values, we observed that biased similarity values showed a tendency to increase, being the first quartile value about 3 times greater than the first quartile value for unbiased similarities, the second quartile value about 6 times greater and the third quartile value about 9 times greater. Initially, we expected the MIB criterion to induce overall smaller similarity values, but this observation suggests that the transfer of weight mass from profile-specific terms to marginal terms, and the subsequent vector normalization is not only causing marginal terms to individually exert a greater influence on the similarity values, but also to collectively make similarity values increase.

Concerning the HDIB-Prob criterion, Fig. 1 shows that the best performing unbiased clusterings were outperformed by their equivalent HDIB-Prob-biased clusterings in 6 out of 10 collections; whereas Fig. 2 shows that the best performing HDIB-Prob-biased clusterings outperformed the best unbiased results in 9 out of 10 collections. In this case, we consider that the weight mass transfer from a large number of terms to fewer terms, which is usually the effect of applying the criterion, is behaving as a term frequency-based feature selection heuristics, thus helping the algorithm to better find the collection inner structure. A similar effect on overall similarity values, as the one described for the MIB criterion, was also observed. Here, the bootstrap resampling first quartile biased similarity value was about 24 times greater than the unbiased first quartile value, the second quartile value about 19 times greater and the third quartile value about 15 times greater.

Comparing overall results, it should be noticed that, while the results obtained by the MIB criterion and HDIB-Prob are numerically equivalent, the collections where each variant worked best did not coincide. The best performing MIB-biased clusterings outperformed the best unbiased results for collection G15, for which the HDIB-Prob criterion worked poorly, being outperformed not only by the best unbiased clustering, but even by the overall worst performing variant, HDIB-Dist. Similarly, for class E13, where the best performing HDIB-Prob-biased clustering outperformed the unbiased results, the best MIB-biased results obtained poorer results. According to their best results, the HDIB-Prob criterion was the overall best variant for six collections and the MIB criterion was the best ranked variant in the remaining four collections. The four collections over which the MIB criterion was the best performing variant are among the five smallest collections, whereas the five largest collections are among those over which the HDIB-Prob criterion is the best performing variant. We consider that this fact does not necessarily show a behavior degradation of the MIB criterion for increasing collection sizes, but rather a better ability of the HDIB-Prob criterion to profit from larger profiles (in all cases, profiles grow as collection sizes grow) for more accurately estimating term probabilities.

A very important remark is that the unbiased clusterings were not the overall best for none of the collections, being outperformed by the HDIB-Prob criterion and the MIB criterion in two distinct sets of 9 collections. Moreover, there were 8 collections where both the HDIB-Prob-biased and the MIB-biased best performing clusterings outperformed the best unbiased result, including collection E21, where all three biased variants outperformed the unbiased variant. These results support the initial hypothesis under which the HDIB-Prob and the MIB criteria were formulated, and the idea of introducing biased representations as a whole, for the case of the radius-α-β_0-compact-sets clustering algorithm.

As a final remark, Fig. 2 shows that there was only one collection, G15, where the HDIB-Dist criterion was not the worst performing variant. This result came as a surprise, as we expected the notion of distinctiveness to be in general more useful for determining the descriptive power of terms. The observed behavior is due to the fact that the best-ranked terms, according to the distinctiveness coefficient of Equation 4, obtained these higher scores from being infrequent in the general language as a whole rather than from being more frequent in the collections than in the general language. Thus, most of these terms turned out to be unlikely to occur in a document, even more in a pair of documents, causing pairwise document-document similarity values to drop. Calculating the bootstrap resampling estimates for the first, second and third quartiles of the distribution of similarity values, and comparing them to unbiased values, we observed a very strong diminution, being the first quartile value only 1.21% of the unbiased first quartile value, the second quartile value 4.32% and the third quartile value 14.5%. This situation caused the radius-α maximum β_0-similarity graphs to be considerably less connected than for other biasing choices, thus producing a larger number of very small clusters.

Figures 3 and 4 show the behavior of the Single-Pass clusterings obtained using the unbiased representation compared to those obtained using the three biased criteria, in a manner homologous to that of Figures 1 and 2.

In this case, although the trends observed for the radius-α-β_0-compact sets algorithm do not hold as clearly, Fig. 4 shows that some variant of biased clustering was the best choice in 6 out of 10 collections. Interestingly, the HDIB-Dist criterion, which was the overall worst option when coupled with the radius-α-β_0-compact sets algorithm, turned out to be the best choice in one of these cases, whereas the MIB criterion was the best choice in 4 cases and the HDIB-Prob criterion was the best choice for one collection. No clear relation may be observed here between collection size and the behavior of biasing choices.

Fig. 6 shows the results obtained for the *k*-means algorithm. Here, since there are no parameter combinations, a single score is obtained for each collection, which is the average of the 10 runs performed to account for the randomness of the algorithm's initialization.

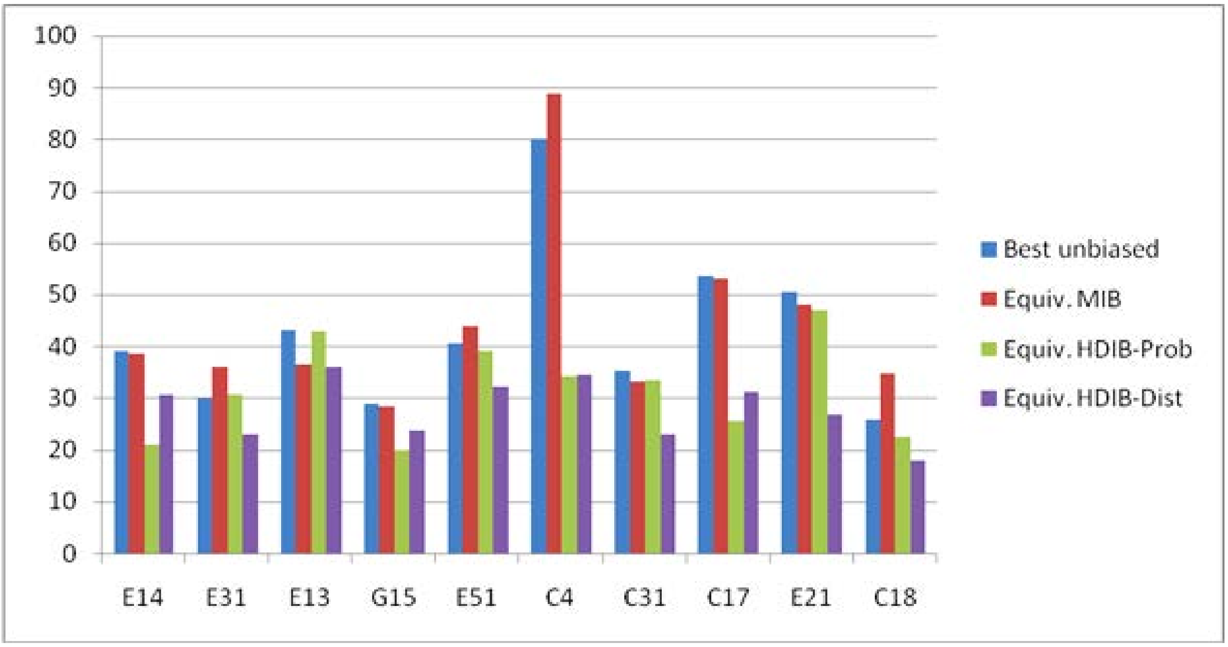

Fig. 3. Best unbiased F_1 values per class and corresponding biased F_1 values on Single-Pass clusterings, all shown as percentages

http://dx.doi.org/10.13053/CyS-18-1-2014-024

Here, some variant of biased clustering was the best choice in 6 out of 10 collections; the same amount as for the Single-Pass algorithm, although not over the same 6 collections. The most interesting remark is that, coupled with the *k*-means algorithm, the HDIB-Dist criterion turned out to be the best choice over 4 collections, being the MIB criterion the best choice for the remaining

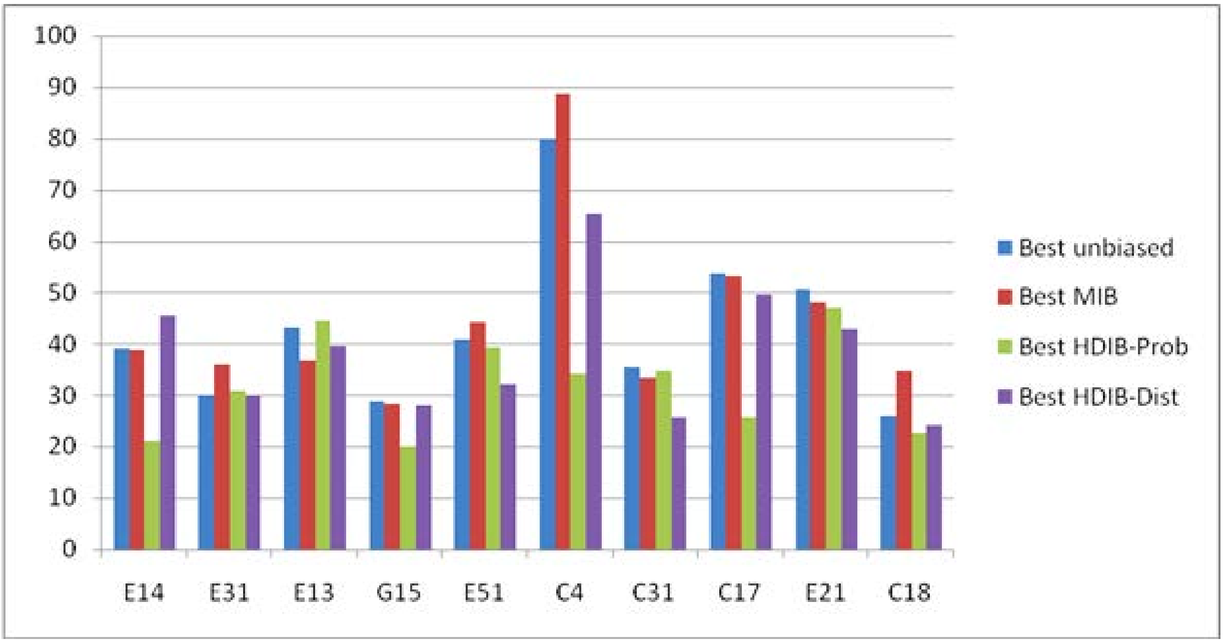

Fig. 4. Best F_1 values per class for all biasing criteria on Single-Pass clusterings, all shown as percentages

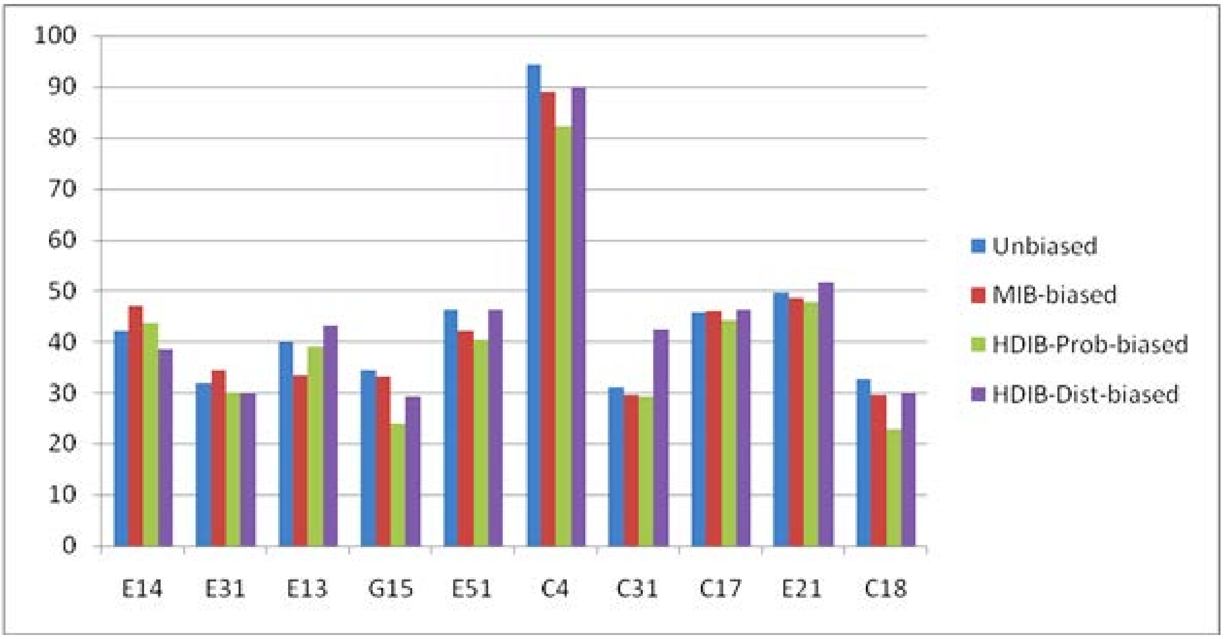

Fig. 5. Averages of F_1 values per class (each averaged over 10 runs) for all biasing criteria on *k*-means clusterings, all shown as percentages

two. This result, along with the case where it was the overall best choice coupled with the Single-Pass algorithm, contrasts with the behavior observed for this criterion when coupled with the radius-α-β_0-compact sets algorithm. The main reason for this difference is that only document to document similarities are used in the radius-α-β_0-compact sets algorithm, which, as we discussed previously, show a trend of being extremely low when the HDIB-Dist criterion is applied. For the case of the Single-Pass and *k*-means algorithms, individual documents are compared to cluster centroids, which are usually much less disperse than vectors representing individual documents, thus yielding higher similarity values and allowing the notion of distinctiveness of the criterion to be better expressed.

Considering overall results, notable differences may be observed between experimental results obtained when coupling the biasing choices with three different clustering algorithms, which clearly indicate that the selection of biasing criteria for practical cases must depend on the algorithms to use, thus taking into account how the modifications introduced in biased representations may affect the working of the clustering algorithms *per se*.

Finally, it should be highlighted that, summing up all results, some variant of biased clusterings performed better than their unbiased counterparts in 22 out of 30 cases, which, in our opinion, provides good evidence of the convenience and practical use-value of the application of biases on document representations.

4 Conclusions

In this paper, we have presented three criteria for introducing biases in document clustering algorithms for the particular case of document collections known to be the result of a document categorization or sample-based document filtering process. These criteria rely on profiles, the document samples used for training classifiers or otherwise obtaining the collection, to apply biases.

When applied to document clustering algorithms, these biases lead the algorithms to obtain different clusterings for the collections, which arguably enables information analysts to discover latent, previously implicit information by analyzing different versions of the collection structure, each of which reflects some principled criterion in making certain subsets of terms exert different degrees of influence in obtaining the clustering.

To the best of our knowledge, this is the first time this particular type of biased clustering has been addressed.

We conducted an experimental evaluation, where a standard corpus was used to simulate a real-life situation where a number of document collections are available, along with their profiles and information regarding their subclass structure. Using that information, we introduced the rationale that a biased clustering may be considered to be better, or somehow more useful, than its unbiased counterpart, if it fits better the collection subclass structure. In this environment, different variants of biased clusterings consistently outperformed their unbiased counterparts in a high number of cases.

We consider that the observed results provide good evidence of the convenience and practical use-value of the application of biases on document representations, thus calling for further experimentation to analyze the coupling between the proposed criteria and other popular algorithms to determine to what extent the conclusions obtained here continue to hold.

Beyond the considerations explicitly stemming from the experimental results shown here, we argue for the intuitive convenience of using some of the proposed biasing criteria for achieving certain useful goals in clustering document collections known to globally cover some common topics, either by obtaining only the biased clusterings or by obtaining both biased and unbiased clusterings and analyzing their differences. For example, in the case of the MIB criterion, biased term weights might be useful for detecting specific terminology describing subtopics in the collection. Furthermore, since a common behavior of all three proposed biasing criteria is that of redistributing the weight mass over terms after normalizing the vectors, by somehow introducing thresholds to these new weights, a subset of the most important terms of a collection may be chosen, thus making our biasing criteria work as a feature selection method.

References

1. **Carpineto, C., Osinski, S., Romano, G. & Weiss, D. (2009).** A Survey of Web Clustering Engines. *ACM Computing Surveys* 41(3).
2. **Ramírez-Cruz, Y. (2013).** Assessing the Effect of Introducing Biases in Document Clustering, *Proceedings of the XV International Convention and Fair Informática 2013*.
3. **Kyriakopoulou, A. & Kalamboukis, T. (2006).** Text Classification Using Clustering. *Proceedings of the Discovery Challenge Workshop at ECML/PKDD 2006*, 28–38.
4. **Kalton, A., Wagstaff, K. & Yoo, J. (2001).** Generalized Clustering, Supervised Learning, and Data Assignment. *Proceedings of the ACM SIGKDD Seventh International Conference on Knowledge Discovery and Data Mining*, 299–304.
5. **J. Hartigan & Wong, M. (1979).** Algorithm AS136: A K-Means clustering algorithm. *Applied Statistics* 28, 100–108.
6. **Palmer, C. & Faloutsos, C. (2000).** Density biased sampling: An improved method for data mining and clustering. *Proceedings of the ACM SIGMOD 19th International Conference on Management of Data*, 82–92.
7. **Salton, G., Wong, A. & Yang, C. S. (1975).** A Vector Space Model for Automatic Indexing. *Communications of the ACM* 18(11), 613–620.
8. **Lindstone, G.J. (1920).** Note on the General Case of the Bayes-Laplace Formula for Inductive or a Posteriori Probabilities. *Transactions of the Faculty of Actuaries* 8, 182–192.
9. **Lewis, D.D., Yang, Y., Rose, T. & Li, F. (2004).** RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research* 5, 361–397.
10. **Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2001).** On Clustering Validation Techniques. *Journal of Intelligent Information Systems* 17(2–3), 107–145.
11. **van Rijsbergen, C.J. (1979).** *Information Retrieval*, London: Butterworths.
12. **López-Caviedes, M. & Sánchez-Díaz, G. (2004).** A New Clustering Criterion in Pattern Recognition. *WSEAS Transactions on Computers* 3(3), 558–562.
13. **Hill, D.R. (1968).** A vector clustering technique. *Mechanized Information Storage, Retrieval and Dissemination*.
14. **Martínez-Trinidad, J.F., Ruiz-Shulcloper, J. & Lazo-Cortés, M. (2000).** Structuralization of Universes. *Fuzzy Sets and Systems* 112(3), 485–500.
15. **Gil-García, R., Badía-Contelles, J.M. & Pons-Porrata, A. (2003).** Extended Star Clustering Algorithm. *Lecture Notes on Computer Science* 2905, 480–487.
16. **Pons-Porrata, A., Sánchez-Díaz, G., Lazo-Cortés, M. & Alfonso-Ramírez, L. (2005).** An Incremental Clustering Algorithm based on Compact Sets with Radius alpha. *Lecture Notes on Computer Sciences* 3773, 302–310.
17. **Sparck-Jones, K. (1972).** A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation* 28(1), 11–21.
18. **Efron, B. & Tibshirani, R. (1993).** *An Introduction to the Bootstrap*. London: Chapman and Hall/CRC Press.

Yunior Ramírez-Cruz received a B.S. in Computer Science (2006) and a M.S. in Computer Science (2008) at Universidad de Oriente (UO), Cuba. He has authored or co-authored 15 scholarly papers and has directed four B.S. theses and three Master’s theses. He received the Best Paper Award of the Computational Linguistics Track of the X International Symposium on Social Communication (2007) and the Best Student Paper Award at the VIII Mexican International Conference on Artificial Intelligence, MICAI 2008. He is a member of the Cuban Society for Mathematics and Computing, and the Cuban Association for Pattern Recognition.

Article received on 07/05/2013, accepted on 10/07/2013.