Improving Alternative Text Clustering Quality in the Avoiding Bias Task with Spectral and Flat Partition Algorithms

M. Eduardo Ares, Javier Parapar, and Álvaro Barreiro

IRLab, Department of Computer Science, University of A Coruña, Spain
{maresb,javierparapar,barreiro}@udc.es

Abstract. The problems of finding alternative clusterings and avoiding bias have gained popularity over the last years. In this paper we put the focus on the quality of these alternative clusterings, proposing two approaches based in the use of negative constraints in conjunction with spectral clustering techniques. The first approach tries to introduce these constraints in the core of the constrained normalised cut clustering, while the second one combines spectral clustering and soft constrained k-means. The experiments performed in textual collections showed that the first method does not yield good results, whereas the second one attains large increments on the quality of the results of the clustering while keeping low similarity with the avoided grouping.

1 Introduction

Data analysis plays nowadays a central role in several fields of science, industry and business. With the ever-growing size of the data collections being compiled and used by public institutions and private firms alike a great need for automatic data analysis tools has arisen, in order to provide a way to exploit those collections in an effective and timely manner.

Clustering is the most popular non-supervised automatic data analysis tool. Given a data collection, the clustering algorithms try to form a meaningful grouping of the data, categorising the data instances (text documents, in our case) in various groups (clusters), such that the instances in the same cluster bear high similarity between them and low similarity with the instances that have been put in the other clusters.

Unfortunately, the concepts of "meaningful grouping" and "high" and "low" similarity are very subjective. Sometimes, and even though the grouping of the data found by a certain clustering algorithm can make sense from a purely mathematical point of view, it might be completely useless or even meaningless to the user. Gondek and Hofmann illustrate in [1] several examples this situation, such as the clustering of news corpora which have been already annotated by a certain criterion (such as region) or the clustering of users' data with gender or income information. The outcome of the algorithm might reflect a grouping of the data

P. García Bringas et al. (Eds.): DEXA 2010, Part II, LNCS 6262, pp. 407–421, 2010.

which is well known, or which would be easy to find with a manual examination. Consequently, it will be of little use to the user of the data analysis tool.

Thus, sometimes mechanisms are needed to find alternative clusterings to the one proposed by the clustering algorithm. If we are trying to avoid the tendency (bias) of the clustering algorithm to fall in a certain grouping of the data that is being clustered the task is called Avoiding Bias. This problem has been tackled by several authors in the last years, which have proposed a wide range of approaches, ranging from distance learning [2] to using constraints [3]. However, it should be underlined that avoiding bias is still a clustering process, where the main focus is providing the user with a meaningful grouping of the data. For instance, the easiest way to find a very different grouping from the one given would be assigning randomly documents to clusters, which would be obviously a very bad solution in terms of clustering quality. Thus, a compromise has to be reached between the quality of the clustering and the distance to the avoided grouping when devising an avoiding bias algorithm.

In this paper we study various ways to obtain an alternative clustering with high quality while keeping the objective of avoiding the known clustering. Concretely, we test two different approaches which use a strategy similar to the one in [3] (using negative constraints to steer the clustering process away from the known clustering), making use of spectral clustering techniques to try to attain that high quality. The first one is introducing negative constraints in the constrained normalised clustering approach proposed by Ji et al. in [4]. The second one is introducing the soft constrained k-means algorithm proposed by Ares et al. in [3], which has been shown to have good results, in the second phase of a normalised cut clustering algorithm [5]. The experiments carried out with these approaches showed that, while the first approach does not yield good results, the combined one (normalised cut plus soft constrained k Means) outperforms soft constrained k-means in terms of quality of the results while keeping a good avoidance of the known clustering.

This paper is organised as follows: in Section 2 the clustering algorithms on top of which the proposed approaches are built are introduced. In Section 3 we tackle the problem of introducing negative constraints in normalised cut, while in Section 4 we introduce the experiments which were carried out and their results. Finally, Sections 5 and 6 are respectively devoted to the related work and the conclusion and future works.

2 Clustering Algorithms

In this section we describe the clustering approaches which we have used in the methods proposed in this paper. Firstly, we survey normalised cut, a very effective spectral clustering algorithm introduced by Shi and Malik in [5], and its constrained counterpart, constrained normalised cut, introduced by Ji et al. in [4]. Afterwards, we outline soft constrained k-means, a constrained clustering algorithm based on k-means introduced by Ares et al. in [3].

2.1 Normalised Cut

The spectral clustering algorithms [6,7] are a family of algorithms which use results from graph spectral theory to perform the clustering of data. Concretely, normalised cut tries to tackle a clustering problem by transforming it into a graph cut problem.

The first step is creating a graph $G = (V, E, W)$ in which the documents to be clustered are the vertices ($V = \{v_1, v_2, ..., v_n\}$), and the weights ($W = \{w_{1,1}, w_{1,2}, ..., w_{n,n}\}$) of the edges ($E$) are related to the similarity between the documents joined by each edge, such that the more similar the documents are, the higher the weight of the edge. Hence, the aim of the clustering process, creating groups of documents such that the documents in the same cluster are very similar and documents in different clusters have low similarity, can be reformulated as cutting this new graph G in connected components in a way that the weights of the edges which join vertices in different connected components are low and the ones of the edges which join vertices in the same connected component are high.

To measure this, Shi and Malik introduced the normalised cut (NCut) value of a cut of a graph in [5]. For a graph $G = (V, E, W)$ and a cut $\{A_1, A_2, ...A_k\}$ of that graph, NCut is defined as:

$$\mathrm{NCut}(A_1, ...A_k) = \sum_{i=1}^{k} \frac{\mathrm{cut}(A_i, \bar{A}_i)}{\mathrm{vol}(A_i)} \tag{1}$$

$$\mathrm{cut}(A, B) = \sum_{i \in A, j \in B} w_{ij} \tag{2}$$

$$\mathrm{vol}(A) = \sum_{i \in A} \sum_{j=1}^{n} w_{ij} \tag{3}$$

where w_{ij} is the weight of the edge that joins vertices i and j, and $\bar{A}_i$ are the vertices which are not included in A_i (i.e., $\bar{A}_i = V \setminus A_i$).

As it follows from (1), the NCut of a graph cut is minimised when the sum of the weights of the edges joining documents in different connected components are low, while keeping the sizes of the different connected components, which are measured using their volume, as high as possible. This last condition tries to ensure a certain balance between the connected components, to avoid trivial solutions with connected components comprising only very few vertices. Thus, a graph cut with a low NCut value would fulfil the requisites of a good clustering.

Finding a cut $\{A_1, A_2, ..., A_k\}$ of a certain graph G which minimises the NCut value can be transformed [7] into a trace minimisation problem (4), where H is a $n \times k$ matrix (where n is the number of documents to be clustered) which encodes the membership of vertices to connected components as indicated in (5), $D = (d_{ij})$, is a diagonal matrix with $d_{ii} = \mathrm{degree}(v_i)$ and L is the Laplacian matrix ($L = D - W$) of G.

$$\min_{A_1, ...A_k} \mathrm{Tr}(H^T L H) \text{ subject to } H^T D H = I \tag{4}$$

$$H = (h_{ij}) = \begin{cases} \frac{1}{\sqrt{vol(A_j)}} & \text{if vertex i} \in A_j \\ 0 & \text{else} \end{cases} \quad (5)$$

Unfortunately, the condition imposed by (5) on the values of H makes the minimisation in (4) NP-hard. If that discreteness condition is dropped and a simple variable substitution is performed ($Y = D^{\frac{1}{2}}H$), the minimisation can be rewritten in the standard form of a trace minimisation problem (6):

$$\min_{Y \in \mathbb{R}^{n \times k}} \mathrm{Tr}(Y^T \left[D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \right] Y) \text{ subject to } Y^T Y = I \quad (6)$$

It can be shown that (6) is minimised by the matrix Y which contains as columns the eigenvectors corresponding to the smallest eigenvalues of $D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$. However, as the values of Y are not constrained, this matrix is no longer composed of indicator vectors for the connected components. Instead, each of the documents has been projected into $\mathbb{R}^k$, and a further step has to be taken (such as applying a clustering algorithm like k-means) in order to find a discrete segmentation of the points in that space. Once this segmentation has been found, we can transpose it to the original documents, providing a clustering of the original collection.

2.2 Constrained Normalised Cut

Based on the same principles of normalised cut, Ji et al. proposed in [4] a constrained clustering algorithm which makes some changes in the function to be minimised in order to introduce *a priori* knowledge in the clustering process, specifically which pairs of documents the user wants to be grouped by the clustering algorithm into the same cluster.

To achieve this, they introduced a new matrix U with n columns and a row for each constraint used in the algorithm. Thus, a constraint which establishes that data points i and j should be in the same cluster will be encoded as a row of zeroes with the exception of positions i and j, which will be set to 1 and -1 (or vice-versa, as these constraints are non-directional). If membership to connected components is encoded in a matrix H as in (5), the Frobenius norm of the product of matrices U and H will be smaller as more constraints are respected in the clustering, with a minimum of zero when none of them is disregarded. Thus, a new minimisation problem can be written involving both NCut and the supplied constraints:

$$\min_{A_1, \ldots A_k} (\mathrm{NCut}(A_1, ..., A_k) + ||\beta U H||^2) \quad (7)$$

where $\beta > 0$ is a parameter which controls the degree of enforcement of the constraints. The higher that β is, the tighter the enforcement of the constraints is. This minimisation problem, following a derivation similar to the one used in the non constrained case, can be written as:

$$\min_{Y \in \mathbb{R}} \mathrm{Tr}(Y^T \left[D^{-\frac{1}{2}} (L + \beta U^T U) D^{-\frac{1}{2}} \right] Y) \quad (8)$$

again subject to $Y^TY = I$. As this problem is in the standard form of a trace minimisation problem, the same theoretical result used in the unconstrained case can be used here. Thus, this equation is minimised by a matrix Y which contains as columns the eigenvectors which correspond to the smallest eigenvalues of matrix $D^{-\frac{1}{2}}(L + \beta U^TU)D^{-\frac{1}{2}}$. Again, these columns are not proper indicator vectors, so a segmentation of the projected data points has to be performed in order to produce a clustering of the data.

2.3 Soft Constrained k-Means

Batch k-means [8] is one of the most popular flat clustering algorithms. The first step of the algorithm is the initialisation, where some points in the representation space are taken as seeds of the clustering process. Typically these seeds are chosen randomly between the documents to be clustered. Afterwards, the main core of the algorithm is a loop in which documents are assigned to clusters depending on its similarity with clusters' centroids. Once all of them have been assigned the centroids are recalculated and the process starts again. This loop is repeated until a given convergence condition is met (typically when the change in the centroids between a iteration and the next is very small).

Based on batch k-means skeleton, Wagstaff et al. introduced in [9] a constrained clustering algorithm which enables the use of domain knowledge in the clustering process. This domain knowledge can be introduced in the form of two kinds of instance level pairwise constraints: Must-links, which indicate that two documents must be in the same cluster, and Cannot-links, to indicate that two documents must be in different clusters. To honour these constraints they modified the cluster assignment policy, assigning the documents to the closest (most similar) centroid such that this assignment does not violate any constraints. That is, if a document with which the document being assigned has a Must-link constraint has been assigned to a cluster in the current iteration, the document will be assigned directly to that cluster. Otherwise, the document is assigned to the cluster with the closest centroid, excluding those containing documents with which the document being assigned has a Cannot-Link constraint, in order to enforce that kind of constraints. In that paper, the authors show that these constraints can effectively affect the clustering process, leading it towards a better solution. On the other hand, the authors admit as well that the absolute nature of the proposed constraints can make sometimes the presence of this constraints harmful. For instance, Cannot links can lead the clustering process to a dead end, if a document has a Cannot link with at least one document in each cluster.

In order to address these limitations, Ares et al. introduced in [3] two kinds of non absolute constraints: May-Links and May-Not-Links, which indicate that two documents are, respectively, likely or not likely to be in the same cluster. The implementation of these constraints alters again the assignment process of the documents. After the absolute constraints introduced by Wagstaff et al. are accounted for, each cluster is given a score which is initialised with the similarity between the document and its centroid. Then, the score of a given cluster will be increased in a certain factor w for each document with which that document has a

May-Link and was last assigned to that cluster. Conversely, the score of a cluster will be decreased by the same factor for each document with which the document has a May-Not-Link and was last assigned to the cluster. The authors claim that these new constraints overcome the drawbacks of the absolute constraints, while maintaining good effectiveness. Namely, the May-Not-Links are shown to be effectively better than their absolute counterparts (Cannot-links), because their efficacy seems to be similar and the May-Not-Links are not affected by the dead end problem, as it is always possible to find a suitable cluster for all the data points. Anyway, the algorithm proposed in the paper allows as well the introduction of domain knowledge in form of absolute constraints, following the same strategy proposed by Wagstaff et al.

3 Negative Constraints in Normalised Cut

As it was previously explained, Ji et al. proposed in [4] an addition to normalised cut which allowed introducing domain knowledge in the clustering process. However, the method that they propose only allows the introduction of *positive* information, i.e., pairs of documents that the user thinks that they ought to be in the same cluster. But this is not the only kind of information that a user might have available about the documents to be clustered. For instance, it is also very likely that the user has some intuition about which pairs of documents might not (or must not) be in the same cluster (this is what we will call *negative* information). Actually, this negative information is less informative to the clustering algorithm than the positive constraints, as with the positive information we are actually providing the algorithm with fragments of the desired final grouping (or at least we hope to be doing so). However, is precisely this lesser informativeness (and the less restrictions that they impose on the algorithm) which makes the negative constraints more likely to be elicited from the domain knowledge, or even the only information that can be provided, in cases where the nature of the task being tackled does not allow the obtaining of positive information at all. For instance, this is the case of the Avoiding Bias task, which is the main focus of this paper.

In the Avoiding Bias task, the only information available is the grouping of the documents that we are trying to avoid. We can not obtain any positive clues from it, as neither the fact that two documents are in the same cluster, nor the fact that they are in different ones gives us any positive evidence about if they should be in the same cluster in an alternative grouping.

However, if two documents are in the same cluster in the grouping that we are trying to avoid, it is sensible to make some indication (using non absolute negative constraints) to the clustering algorithm that these documents might not be in the same cluster in other grouping, expecting that the distorsion induced by these constraints is on the one hand enough to break the bias of the algorithm to fall in the avoided clustering and on the other hand not strong enough to break completely the structure of the similarities between documents, so that the final clustering of the data is still meaningful. This is precisely the intuition that sustains Ares et al. Avoiding Bias approach in [3].

Obviously, the same point could be made about using positive non absolute constraints on documents which are not in the same cluster in the avoided grouping. However, bringing closer these documents will not have the effect of avoiding the bias of the clustering algorithm to fall in the given grouping. To do so, these constraints should be very strong, but this will likely compress the representation space too much, providing clusters of bad quality.

In this section we will tackle the problem of introducing the negative constraints into the normalised cut clustering algorithm.

3.1 Negative Constraints in Constrained Normalised Cut

In Sect. 2.2 we have explained the approach used by Ji et al. [4] to transform the classic normalised cut algorithm into a constrained clustering one, allowing the use of positive constraints. Intuitively, a similar scheme could be used to try to introduce negative information as well.

In their paper, the authors introduce a matrix U which encodes the positive constraints, such that the Frobenius norm of the product of that matrix and the indicator matrix is in inverse proportion with the number of constraints which are respected by the clustering represented by the indicator matrix, having a minimum of zero when all of them are honoured. Thus, introducing this factor into the function minimised at the core of the normalised cut algorithm (7,8) causes a change in the nature of the solution, now having to find a clustering of good quality (minimising NCut) while respecting as well the constraints (minimising the new term). The influence of the constraints is controlled by a parameter (β), being the enforcement of the constraints greater as the value of β increases, with a minimum in $\beta = 0$, where the the constraints are not taken into account at all.

With that in mind, an apparently easy and intuitive way to introduce the negative constraints would be using a new matrix U_N, which would encode the negative constraints in the same way as the positive ones were encoded in U. Again, the Frobenius norm of the product of U_N with the indicator matrix will be lower as more of the pairs of documents linked by a constraint are in the same cluster, and, vice versa, higher as more of them are not in the same cluster, which is precisely the objective of the negative information. In order to introduce this new term in the minimisation a new parameter (β_N) is needed to control the enforcement of the negative constraints. As this new factor is in direct proportion to the number of negative constraints which are respected in the clustering, it must be introduced in the formula with a minus sign (9,10). Again, the value of β_N is equal or greater than 0, with a harder enforcement of the constraints as its value increases.

$$\min_{A_1,...A_k}(\mathrm{NCut}(A_1, ..., A_k) - ||\beta_N U_N H||^2) \tag{9}$$

$$\min_{Y \in \mathbb{R}} \mathrm{Tr}(Y^T \left[D^{-\frac{1}{2}}(L - \beta_N {U_N}^T U_N) D^{-\frac{1}{2}} \right] Y) \tag{10}$$

Even though this approach seems theoretically sound, it does not yield good results in the Avoiding Bias task. Our explanation about why this happens is given in Sect. 4.5.

3.2 Combining Soft Constrained k-Means and Normalised Cut

As it has been previously explained (Subsect. 2.1), the normalised cut algorithm is based on transforming the clustering problem into a graph cut problem. The aim of the process is finding a cut of the graph which minimises its normalised cut value. Being this a NP-hard problem, a certain relaxation of the conditions imposed on the solution has to be performed in order to reduce its complexity and make it computationally accessible. Thus, the outcome of this minimisation is a projection of the data points into $\mathbb{R}^k$, instead of the grouping itself, and a last step should be performed to reach the final clustering of the data. In order to perform this last phase, Shi and Malik propose using k-means on the projected data points.

Our proposal in this paper is using the soft constrained k-means algorithm proposed by Ares et al. instead of batch k-means, enabling the introduction of domain knowledge in form of absolute (Must and Cannot-Link) and non-absolute (May and May-Not-Link) constraints. Even though they would be defined over the initial documents, the one to one correspondence between them and the projected documents (the document which was represented by the vertex v_i of the graph is now encoded in the i^{th} row of matrix Y) enables us to apply these same instance level constraints over the corresponding projected documents.

From the point of view of soft constrained k-means, the normalised cut acts as a kind of document preprocessing phase, where the documents are transformed from the chosen document representation to a representation in $\mathbb{R}^k$ based on the normalised cut criterion. The effect of this "preprocessing" is twofold: not only we are benefiting from the increment of cluster quality caused by using the normalised cut algorithm, but also we are likely to experiment an increase in the effect of the pairwise constraints. As documents which are close to constrained ones are affected as well by the changes in the destination of the later ones induced by the constraints, our intuition is that the effectiveness of the constraints in this new data space is increased, as similar documents (over which the same constraints tend to be true) are brought together and dissimilar ones are separated (thus avoiding some non desired "interferences" of the constraints over non related documents).

In terms of performance, the computational cost of this combined approach is the same of that of the normalised cut algorithm, as the cost of the soft constrained k-means and of batch k-means is the same. Consequently, being the costliest operation of the whole algorithm still by a wide margin the calculation of eigenvectors, the total cost will depend on the method chosen to perform that calculus. This cost can be kept fairly moderated if a standard algorithm is used. For instance, using Lanczos algorithm, the time complexity would be $O(kN_{Lanczos}nnz(M))$, where k is the desired number of clusters (i.e. of eigenvectors), $N_{Lanczos}$ is the number of iteration steps of the algorithm and $nnz(M)$

is the number of non zero elements of the matrix $D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$ (see Sect. 2.1), whose eigenvectors are being calculated.

4 Experiments

4.1 Methodology

In order to test the practical behaviour of the algorithms we have set an avoiding bias experiment following the standard methodology of the papers on that subject. Thus, we will use text document collections in which documents have been categorised according to two different criteria. Using the standard methodology of avoiding bias experiments, we will assume alternatively that one of them is the known grouping and we will try to avoid it, evaluating the results of the process comparing the resulting grouping with both the known one (to assess the avoidance that has been achieved) and the "unknown" one (as a way to measure the quality of the results).

We have used as baseline the original soft constrained k-means approach to Avoiding Bias introduced by Ares et al. in [3], where the authors show that it improves an algorithm specially tailored for Avoiding Bias such as Conditional Information Bottleneck [1]. Thus, we have replicated the same experimental conditions used in that paper. The set of constraints was created introducing a constraint for each pair of documents which are in the same cluster in the known grouping of the data (the only *a priori* information available). In the case of the baseline and of the combined (NC+SCKM) approach, which support bidirectional and unidirectional constraints, we have used the bidirectional ones. Moreover, we will assume that the number of clusters is known, setting it to the number of clusters of the non avoided grouping of the data. Finally, as the clustering seeds were also chosen randomly from the documents, and the outcome of the processes is really dependant on the quality of the initial seeds, several repetitions of the clustering process have to be performed in order to have a faithful representation of the performance of the algorithms. We report the average of these initialisations.

Following this approach, the only parameters which should be initially set are w, the strength of the constraints in the baseline and in the approach based on the combination of normalised cut and soft constrained k-means and β_N, the tightness of the observance of the negative constraints in the approach based on constrained normalised cut. Besides, in our experiments we have detected that the clustering algorithms yielded better results when the number of dimensions of the projection of the documents performed in the spectral phase is greater than the wanted number of clusters. Typically, the best performance was obtained when the number of eigenvectors ranged from 10 to 20 (in opposition to the number of desired clusters, which ranges from 2 to 5), a fact that is likely caused by the combination of two circumstances. Firstly, the high topicality of the collection compared with the number of expected clusters, and, secondly, this relatively small number of desired clusters, which would cause a great loss of information in the projection if we take the same number of eigenvectors. However,

taking too many dimensions could result in adding noise to the documents, which would worsen the quality of the clustering. Thus, after some preliminary tests, we have used to create the projection of the documents the first 15 eigenvectors, a value which we have found that performs well in all collections.

4.2 Datasets

To perform the experiments we have used the two datasets used in the baseline experiments, which were originally defined in [1].

Dataset (i) was created from WebKB's Universities Dataset, which was made collecting webpages from the websites of different U.S. universities (Cornell, Texas, Washington, Wisconsin and others). These webpages have been manually tagged according to two aspects: university and topic ("course", "department", "faculty", "project", "staff", "student" and "other"). The dataset used in the experiments is created taking the documents from the Universities of Cornell, Texas, Washington and Wisconsin which were as well tagged as "course", "faculty", "project" "staff", "student", which yields a total of 1087 documents.

Dataset (ii) was created from Reuters RCV-1, a huge document collection composed of about 810,000 news stories from Reuters, one of the most important news agencies. These documents have been manually tagged according to three aspects: topic, geographical area and industry. The dataset used in the experiments is created taking the documents with have been labelled with respectively only one topic and region label and whose topic is "MCAT" or "GCAT" and whose region is "UK" or "INDIA". This yields a total of 1600 documents.

4.3 Document Representation

As in the baseline experiments, we have used Mutual Information as the original representation of the documents (i.e., the one used to build the graph G), as it has been shown to perform consistently better than other $tf \cdot idf$ approaches [10]. Thus, the representation of a document d in a collection of m terms and d documents is a vector (11) where the components are the mi values of the terms (12),13), calculated used the frequency of the each term t in the document d ($tf(d, f)$).

$$mi(d) = [mi(d, t_1); mi(d, t_2); \ldots; mi(d, t_m)] \tag{11}$$

$$mi(d, t) = \log\left(1 + \frac{\frac{tf(d,t)}{N}}{\frac{\sum_i^D tf(d_i,t)}{N} \times \frac{\sum_j^m tf(d,t_j)}{N}}\right) \tag{12}$$

$$N = \sum_i \sum_j tf(d_i, t_j) \tag{13}$$

The similarity between two documents d_1 and d_2 was computed using the cosine distance between their vectors, which was also also the distance function used to compare the projected documents after the spectral phase.

4.4 Metrics

In order to evaluate the results of our tests we have used two different metrics, which compare the clustering of a collection of n documents yielded by the algorithm $\Omega = \{\omega_1, \omega_2, ...\omega_k\}$ with a certain ground truth $\mathbb{C} = \{c_1, c_2, ...c_j\}$.

Purity (P) [11] measures how well the clustering outcome matches the target split in average. Higher Purity values mean more similarity between Ω and $\mathbb{C}$.

$$\mathrm{P}(\Omega, \mathbb{C}) = \frac{1}{n} \sum_k \max_j |\omega_k \cap c_j| \tag{14}$$

On the other hand, Mutual Information (MI) [12] measures how much information about a grouping is conveyed by another. Again, higher values of Mutual Information mean more agreement between Ω and $\mathbb{C}$.

$$MI(\Omega; \mathbb{C}) = \sum_k \sum_j \frac{|\omega_k \cap c_j|}{n} \log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|} \tag{15}$$

4.5 Results

In order to set the parameters of the algorithms we have used a crossvalidation strategy. This strategy involved tuning the value of these parameters in one of the avoiding bias problems, specifically in collection (i) avoiding the grouping by "Topic", and using that value in the other problems. The value w chosen for the baseline (soft constrained k-means) was 0.0025, the value which obtained the best compromise between quality and avoidance. In the combined approach (NC+SCKM), as the focus of this paper is improving the quality of the grouping, the value ($w = 0.05$) was chosen as the one which yielded the best similarity (MI) with the non avoided grouping of the documents ("University") while maintaining a similarity with the avoided grouping ("Topic") less or equal to the one achieved by the baseline, which was itself quite low. As for the constrained normalised cut with negative constraints, the tuning process showed poor quality values and a great instability of the algorithm with respect to the values of β_N. Our explanation about why this happens is given at the end of this section.

The results of the performed experiments are shown in Table 1. As in the experiments in [3] and in [1], for each dataset and avoided grouping we report the values of Mutual Information (MI) with the avoided and the non-avoided groupings, to see to which of them the outcome of the clustering process is mostly leaning, and Purity (P) with the non-avoided grouping, to measure the quality of the clustering. Hence, a good result would have high values of MI and P with the non-avoided grouping and a low value of MI with the avoided one. The results reported are the average of the ten different initialisations of seeds and document inspection order tested in each combination of dataset and avoided grouping.

As a preliminary note, it is worth remarking that the results show the expected increase in the quality of clustering of normalised cut with respect to batch k-means. Moreover, they also point out a tendency in the non constrained

Table 1. Results for the avoiding bias experiment with the defined datasets for batch k-means, soft constrained k-means (SCKM), normalised cut and the combined approach (NC+SCKM)

Dataset (i)	Avoiding Topic (k=4)			Avoiding University (k=5)		
	MI(Topic)	MI(Univ.)	P(Univ.)	MI(Univ.)	MI(Topic)	P(Topic)
Batch k-means	0.5069	0.2304	0.4364	0.2972	0.5682	0.6874
SCKM ($w = 0.0025$)	0.0052	0.2789	0.4772	0.0031	0.4499	0.6484
Normalised cut	0.4801	0.4097	0.4994	0.5822	0.5606	0.6794
NC+SCKM ($w = 0.05$)	0.0032	0.9340	0.7684	0.0011	0.6569	0.7163

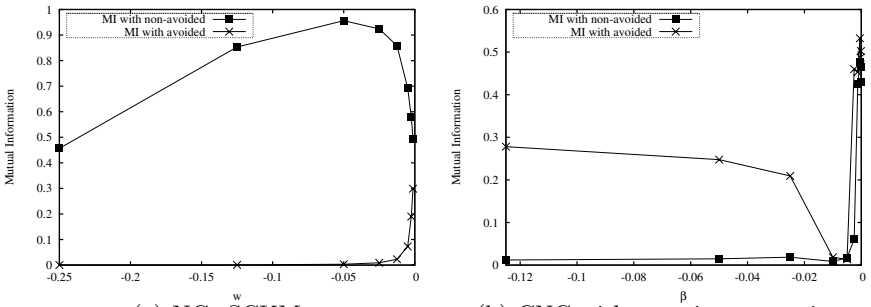
Dataset (ii)	Avoiding Topic (k=2)			Avoiding Region (k=2)		
	MI(Topic)	MI(Region)	P(Region)	MI(Region)	MI(Topic)	P(Topic)
Batch k-means	0.0075	0.0874	0.8253	0.1400	0.0093	0.9838
SCKM ($w = 0.0025$)	0.0003	0.1194	0.8253	0.0004	0.0075	0.9838
Normalised cut	0.0075	0.1510	0.8253	0.1862	0.0106	0.9838
NC+SCKM ($w = 0.05$)	<0.0001	0.1643	0.8253	<0.0001	0.0164	0.9838

algorithm (in our case, normalised cut) to fall in one of the two groupings of the collections, even though this tendency is sometimes less clear than in the case of the batch k-means.

The similarity of the outcome of the proposed algorithm (NC+SCKM) with the non avoided clustering (which, as it has been said before, is used as a indication of the quality of the clustering) is in all cases greatly increased over the soft constrained k-means results. Moreover, the results show how the introduction of this constrained phase has not any detrimental effect over the quality of the normalised cut results, and in fact improves them in all cases. As for the avoided grouping, the similarity of the results of our technique is still reduced, keeping it in values equal or less than those of the baseline, which were already low.

It should be also noted that the reason for the repeated values of P for the four methods in dataset (ii) is the structure of the dataset, where in each of the possible groupings one of the clusters is much bigger than the other (still, the MI values for that dataset attest the improvements attained using the combined method). Finally, it is also worth remarking that further tests on the training collection have shown that the parameter w of this combined approach is quite stable. This can be seen in Fig. 1(a), which shows that the MI with the avoided and non-avoided groupings are not affected to a greater extent by wide variations around the chosen value of 0.05.

The results of the tests performed with the approach introduced in Sect. 3.1 (which introduces the negative constraints in the core of the constrained normalised cut algorithm) are not included in Table 1 as the quality values achieved were poor and the value of the parameter β_N was very unstable. This is shown in Fig. 1(b): for almost all values of the parameter the similarity with the avoided grouping is much higher than with the non-avoided one, and for the values of β_N in which the two similarities come closer the quality of the result is very low and a small variation of the parameter produces an abrupt change in the quality values. Our intuition is that the cause of this behaviour has to do with the function which is minimised. With positive constraints, the function in (7) has its lower bound in zero, a value which, if obtained, would mean both that the clustering has good quality (NCut = 0) and that all the



(a) NC+SCKM (b) CNC with negative constraints

Fig. 1. Stability of the parameters of the two proposed algorithms in the training collection (Dataset (i), avoiding TOPIC)

constraints are respected ($||\beta UH||^2 = 0$). However, this is not what happens in the minimised function when negative constraints are involved (9). Here, a low value can be obtained if all the constraints are respected, regardless of the quality of the clustering, as one value is subtracted from the other. This makes tuning the value of β_N very hard, as a small change can alter dramatically the balance between those two factors.

5 Related Work

In the constrained clustering field [13], the problems of avoiding bias and finding alternative clusterings have gained popularity in the last years, with several authors looking into them and proposing different approaches. Bae and Bailey proposed in [14] a method similar to the one used in this paper, using negative constraints to try to steer the clustering away from the avoided grouping. They incorporate these constraints in a Average Link clustering algorithm, controlling with a parameter the compromise between obtaining a clustering of quality and honouring the constraints. However, they only report results in synthetic and numeric data collections with a very limited number of features.

Gondek and Hoffman introduced in [1] another strategy to find alternative clusters using Conditional Information Bottleneck clustering. Their approach tries to optimise an objective function which combines the objectives of yielding clusters of good quality and which should be different from the given clustering. To do so they need the complete distribution of each variable, which is one of the main drawbacks of the method.

In [15], Davidson and Qi present an approach to finding alternative clustering which also uses constraints, in this case to characterise the grouping to be avoided. A distance function matrix is learnt from these constraints, which is decomposed afterwards using Singular Value Decomposition (SVD). Finally, the matrices yielded by SVD are used to build an alternative distance function that is used to created transformed versions of the original data points, over which

the clustering algorithm would be applied. Thus, this method has the advantage of being quite general, not being tied to any clustering clustering. Again, they tested their approach only in non-textual collections.

Cohn et al. introduced in [16] an algorithm to iteratively alter the grouping found by a clustering process according to the user feedback. They incorporate the user preferences altering the KL-divergence measure between the documents marked by the user, introducing a new factor to measure the importance of a term for distinguishing the documents. Even though they conduct their tests over textual documents, the collections are again very small.

Obviously, the avoiding bias method which is most related to the ones proposed in this paper is the one introduced by Ares et al. [3], which uses the soft constrained k-means algorithm, described in Sect. 2.3. It was used as baseline in our experiments (Sect. 4.5), and in one of the approaches proposed in this paper we have combined it with normalised cut 3.2. Another general constrained clustering algorithm which is also related to this paper is constrained normalised cut by Ji et al. [4], as it is the core of one of the Avoiding Bias methods proposed in this paper (Sect. 3.1). The unsuitability of that algorithm for the Avoiding Bias problem was discused in Sect. 4.5.

6 Conclusions

In this paper we have studied two approaches based on the use of negative constraints in conjunction with spectral clustering techniques to tackle the Avoiding Bias problem. While one of them, based in introducing the negative constraints in the core of constrained normalised clustering, did not yield good results, the second one, which combines normalised clustering and soft constrained clustering gave very good results in the experiments carried out, as it increased (in some cases dramatically) the quality of the clustering while maintaining a good avoidance of the known grouping. On a more general level, it should be noted that the possible fields of application of this approach are not limited to the Avoiding Bias problem on text. This algorithm can be applied in any general constrained clustering situation, where, opposed to constrained normalised cut (which would only allow the use of one kind of information), it lets the user use different kinds of knowledge (negative and positive, absolute and non absolute,...).

Acknowledgements. This work was co-funded by FEDER, Ministerio de Ciencia e Innovación, Xunta de Galicia and Ministerio de Educación under projects TIN2008-06566-C04-04 and 07SIN005206PR and FPU grant AP2007-02476.

References

1. Gondek, D., Hofmann, T.: Non-redundant data clustering. In: ICDM 2004: Proceedings of the Fourth IEEE International Conference on Data Mining, pp. 75–82. IEEE Computer Society, Los Alamitos (2004)
2. Davidson, I., Qi, Z.: Finding alternative clustering using constraints. In: ICDM 2008: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. IEEE Computer Society, Los Alamitos (2008)

3. Ares, M.E., Parapar, J., Barreiro, A.: Avoiding bias in text clustering using constrained k-means and may-not-links. In: Azzopardi, L., Kazai, G., Robertson, S., Rüger, S., Shokouhi, M., Song, D., Yilmaz, E. (eds.) ICTIR 2009. LNCS, vol. 5766, pp. 322–329. Springer, Heidelberg (2009)
4. Ji, X., Xu, W., Zhu, S.: Document clustering with prior knowledge. In: SIGIR 2006: Proceedings of the 29th Annual international ACM SIGIR conference on Research and development in information retrieval, pp. 405–412. ACM, New York (2006)
5. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 888–905 (2000)
6. Ding, C.: A tutorial on spectral clustering. In: Tutorial presented at ICML 2004: 21st International Conference on Machine Learning (2004)
7. von Luxburg, U.: A tutorial on spectral clustering. Technical Report TR-149, Max Planck Institute for Biological Cybernetics (2006)
8. McQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967)
9. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S.: Constrained k-means clustering with background knowledge. In: ICML 2001: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577–584, Morgan Kaufmann Publishers Inc., San Francisco (2001)
10. Pantel, P., Lin, D.: Document clustering with committees. In: SIGIR 2002: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 199–206. ACM Press, New York (2002)
11. Rosell, M., Kann, V., Litton, J.E.: Comparing comparisons: Document clustering evaluation using two manual classifications. In: Proceedings of the International Conference on Natural Language Processing (2004)
12. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
13. Basu, S., Davidson, I., Wagstaff, K.: Constrained Clustering: Advances in Algorithms, Theory, and Applications. Chapman & Hall/CRC, Boca Raton (2008)
14. Bae, E., Bailey, J.: COALA: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In: ICDM 2006: Proceedings of the Sixth International Conference on Data Mining, pp. 53–62. IEEE Computer Society, Los Alamitos (2006)
15. Davidson, I., Qi, Z.: Finding alternative clustering using constraints. In: ICDM 2008: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. IEEE Computer Society, Los Alamitos (2008)
16. Cohn, D., Caruana, R., McCallum, A.: Semi-supervised clustering with user feedback. Technical Report TR-2003-1892, Cornell University (2003)