Language Modelling of Constraints for Text Clustering

Javier Parapar and Álvaro Barreiro

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

Abstract. Constrained clustering is a recently presented family of semi-supervised learning algorithms. These methods use domain information to impose constraints over the clustering output. The way in which those constraints (typically pair-wise constraints between documents) are introduced is by designing new clustering algorithms that enforce the accomplishment of the constraints. In this paper we present an alternative approach for constrained clustering where, instead of defining new algorithms or objective functions, the constraints are introduced modifying the document representation by means of their language modelling. More precisely the constraints are modelled using the well-known Relevance Models successfully used in other retrieval tasks such as pseudo-relevance feedback. To the best of our knowledge this is the first attempt to try such approach. The results show that the presented approach is an effective method for constrained clustering even improving the results of existing constrained clustering algorithms.

1 Introduction and Motivation

Clustering is an important data mining tool in order to exploit the knowledge present in the document collections. Lately it has been also demonstrated as an useful tool not only by itself but also for other Information Retrieval (IR) tasks such as cluster-based retrieval [14] or clustering of search results [23]. Recently a new family of constrained clustering algorithms [7] has achieved great importance because they enabled the introduction of domain knowledge in the clustering process. In these semi-supervised methods the domain knowledge is introduced as rules in a generalized framework making the algorithm itself still domain-independent. In this way knowledge that was unused in traditional clustering algorithms is exploited to improve the grouping of data.

Till this moment, the way in which this new clustering task was carried out was by designing new specifically tailored algorithms. Due to the popularity of the task, several new algorithms appeared based on traditional clustering algorithms: partitional algorithms [20, 2], hierarchical algorithms [12, 3], probabilistic approaches [6, 25], matrix decomposition based methods [10, 21], etc. All these algorithms force the accomplishment of the constraints in the document to cluster

assignment or by modifying the objective functions, in contrast, we propose an approach based on maintaining the simplicity of the clustering algorithms. The idea explored in this paper is to avoid the creation of new constrained clustering algorithms and keep using the well-known and tested clustering algorithms for this new semi-supervised clustering task. So the question is how unsupervised clustering algorithms can be used for constrained clustering? To the best of our knowledge this is the first time that this question is answered: our proposal is by introducing the constraints directly in the document representation by means of their Language Modelling.

The main contributions of this paper are on one hand the design of a new approach to constrained clustering which allows the use of unsupervised clustering algorithms instead of the specially tailored new ones and on the other hand to allow so by modifying the document representation by means of the language modelling of the constraints. More precisely our proposal is to expand de documents that are affected by constraints using Relevance Models [13].

Language Modelling (LM) is a high performance theoretical retrieval framework very used in IR. Relevance Models (RM) [13], presented under the LM framework, is a technique for the pseudo relevance feedback (PRF) task and has been proven very successful to improve retrieval effectiveness. Since it was originally presented in [13] it has been used for cluster based retrieval [14], passage retrieval [15] or sentence retrieval [4]. In this paper we will use the RM framework to alter the original document representations. In RM the query and the documents in the relevance set are assumed as samples of the same Relevance Model, in our proposal we assume that there exists a Relevance Model which generates a document and the set of documents that share constraints with the given document. Therefore, for every document we can estimate the Relevance Model given the documents that constrain it. Meanwhile in the PRF task a query is expanded with the best terms of the relevance model obtained from the relevance set, in the clustering task, every document which is affected by a set of constraints, will be expanded with the best terms of the relevance model obtained from the set of documents which it shares constraints with.

The rest of the paper is organized as follows. Section 2 presents the proposed method for the language modelling of the constraints. Section 3 explains the clustering algorithms with which the approach is tested with some considerations about distance functions. In Section 4 the evaluation and its results are reported. Section 5 describes the related work and, finally, conclusions and future work are reported in Section 6.

2 Modelling of Constraints in the Language Modelling Framework

As previously exposed constrained clustering algorithms use the background knowledge to drive the clustering process. Constrained clustering is different from a classification task, where it is exactly known which groups exist in the data and examples of those categories are provided to the algorithm. In con-

strained clustering the domain knowledge gives the clustering algorithm rules over documents. These rules reflect some preferences about whether or not the documents should be in the same cluster, being still the algorithm which finds the groups in the data.

The most of existing constrained clustering algorithms relay over the so called instance level constraints [19]. Instance level constraints can be defined as rules between two documents referring to whether (positive constraints) or not (negative constraints) they must be part of the same clustering. Depending on the algorithm design and the enforcement desired for the constraints they are commonly classified in: absolute constraints, constraints that the algorithm can not violate and must mandatory honour at the end of the clustering process (Must-Link and Cannot-Link for positive and negative constraints respectively); and soft constraints, non absolute constraints that the algorithm could not honour at the end of the clustering process (May-Link and May-Not-Link for positive and negative constraints respectively).

When working in real scenarios dealing with non categorical information is the most common situation, so soft constraints are commonly used taking advantage of the parameters that controls the enforcement of the soft constraints in the algorithms that support those kind of constraints. From now, when talking about constraints we will refer to positive soft constraints, i.e, May-Links.

2.1 Relevance Models

The use of RM for PRF was designed in the LM theoretical framework. In LM the probability of a document given a query, $P(d|q)$, is estimated using the Bayes' rule as presented in Eq. 1.

$$P(d|q) = \frac{P(q|d) \cdot P(d)}{P(q)} \stackrel{rank}{=} \log P(q|d) + \log P(d) \tag{1}$$

In practice $P(q)$ is dropped for document ranking purposes. The prior $P(d)$ encodes a-priori information on documents and the query likelihood, $P(q|d)$, incorporates some form of smoothing, one of the most used forms of smoothing is Dirichlet smoothing [24] as defined in Eq. 2.

$$P(q|d) = \prod_{i=1}^{n} P(q_i|d) = \prod_{i=1}^{n} \frac{tf(q_i, d) + \mu \cdot P(q_i|C)}{|d| + \mu} \tag{2}$$

where n is the number of query terms, $tf(q_i, d)$ is the raw term frequency of q_i in d, $|d|$ is the document length expressed in number of terms, and μ is a parameter for adjusting the amount of smoothing applied. $P(q_i|C)$ is the probability of the term q_i occurring in the collection C that is usually obtained with the maximum likelihood estimator computed using the collection of documents.

The RM approach builds better query models using the information given by the pseudo relevant documents. Two estimations were originally presented in [13]: RM1 and RM2. In RM the original query is considered a very short sample

of words obtained from the relevance model (R). If more words from R are desired then it is reasonable to choose those words with highest estimated probability when considering the words for the distribution already seen. So the terms in the lexicon of the collection are sorted according to that estimated probability, which after doing the assumptions using the RM1 method, is estimated as in Eq. 3.

$$P(w|R) \propto \sum_{d \in C} P(d) \cdot P(w|d) \cdot \prod_{i=1}^{n} P(q_i|d) \tag{3}$$

Usually $P(d)$ is assumed to be uniform. $\prod_{i=1}^{n} P(q_i|d)$ is the query likelihood given the document model, which is traditionally computed using Dirichlet smoothing (see Eq. 2). Then for assigning a probability to the terms in the relevance model we have to estimate $P(w|d)$; in order to do so it is also common to use Dirichlet smoothing. In order to obtain the expanded query a certain top r documents from the initial retrieval are taken for the estimation instead of the whole collection C, conforming the pseudo relevance set RS, and relevance model probabilities $P(w|R)$ are then calculated using the estimate presented in Eq. 3. To build the expanded query the e terms with highest estimated values for $P(w|R)$ are selected. The expanded query is used to produce a second document ranking.

RM3 [1] is a later extension of RM1 which is the most effective estimation of RM [16]. RM3 linearly interpolates the terms selected by RM1 with the original query as in Eq. 4 instead of using them directly. The final query is used in the same way as in RM1 to produce a second ranking using negative cross entropy.

$$P(w|q') = (1 - \lambda) \cdot P(w|q) + \lambda \cdot P(w|R) \tag{4}$$

2.2 Introducing the Constraints in the Document Representation

One important point in every clustering algorithm is the way in which the documents are represented. Over that representation will relay the computation of the similarity/distance functions among documents and/or centroids. When dealing with textual documents, they are usually represented according to the Vector Space Model, assigning one dimension to each term in the lexicon. The way in which each term is weighted for every document varies being the TF·IDF and the pointwise Mutual Information the most used weighting schemas due to their good performance.

In this paper we want to introduce the constraints in the document representation under the LM framework. In order to do so, we have to consider the document representations as probability distributions. So we decided to weight the terms by means of the maximum likelihood estimator. Once that the original document representation is defined we proceed to the constraint modelling. Let us define $C(d) = \{\hat{d}^1, \cdot, \hat{d}^{|C(d)|}\}$ as the set of documents that share a constraint with the document d. In order to introduce those constraints in the document representation our proposal is:

1. Suppose that for every affected document d and its $C(d)$ there exists a supporting Relevance Model.
2. That Relevance Model can be estimated under the RM framework.
3. From the estimated Relevance Model the e best terms are selected to alter the original representation of the document d. Then a linear interpolation is done as in RM3, being λ in this case the parameter that weight the importance of the constraints in the interpolated representation.
4. Use the altered document representation in the clustering process with unsupervised algorithms.

$$P(w|R) \propto \sum_{\hat{d} \in C(d)} P(\hat{d}) \cdot P(w|\hat{d}) \cdot \prod_{i=1}^{|d|} P(d_i|\hat{d}) \tag{5}$$

In Eq. 5 the reformulation of the Eq. 3 for our task is presented. Equation 5 gives the estimation of probabilities in the Relevance Model underlying d and the set of documents $C(d)$ that constrains it. In practice $P(\hat{d})$ can be considered to be uniform. In our task the role of q in the query likelihood presented in Eq. 3 is played by the document affected by constraints d meanwhile the role of the RS is played by $C(d)$. Results as the way of how $C(d)$ is constructed $\prod_{i=1}^{|d|} P(d_i|\hat{d})$ should be considered uniform because the constraints are defined explicitly having every one the same weight. Talking in terms of relevance each time a constraint is explicitly established between two documents d^x and d^y it is equivalent to assess that the document d^x is relevant for the document d^y and *vice versa*, non existing any grading in the relevance assessment. Therefore the final estimation used in this approach is presented in Eq. 6, the final document representation is then computed as in Eq. 7.

$$P(w|R) \propto \sum_{\hat{d} \in C(d)} P(w|\hat{d}) \tag{6}$$

$$P(w|d') = (1 - \lambda) \cdot P(w|d) + \lambda \cdot P(w|R) \tag{7}$$

3 Clustering Algorithms

Before presenting the clustering algorithms that we will use to asses our proposal (K-Means family and Normalized Cut family) we have to do some consideration about the similarity/distance functions. As previously stated when working in the LM framework we will work with probability distributions, so in order to be scrupulous with that fact we have to work with similarity/distance functions according to that. In IR usually Kullback Leibler Divergence (KLD) as in Eq. 8 is used in such cases. Unfortunately KLD is only defined when $Q(i) > 0$ for any i such that $P(i) > 0$ and also is a non-symmetric measure. The Normalized Cut algorithm requires a symmetric function so we decided to use the I-Divergence to the mean (IDM). This is a symmetric version of the I-Divergence (both previously succesfully used in the clustering task [6]) that is a Bregman divergence,

a family of divergence functions including the KLD and squared Euclidean distance that guarantees the decreasement of the K-Means objective function [5]. So the distance function between two documents, d^x and d^y, used in every algorithm is IDM defined as in Eq. 9.

$$KLD(P \parallel Q) = \sum_i P(i) \cdot \log \frac{P(i)}{Q(i)} \tag{8}$$

$$IDM(d^x, d^y) = \sum_{i=1}^{n} d_i^x \log \frac{2d_i^x}{d_i^x + d_i^y} + d_i^y \log \frac{2d_i^y}{d_i^x + d_i^y} \tag{9}$$

In Section 4 a preliminary experiment is presented comparing the presented set-up (MLE as document representation with IDM as distance function) with the traditional set-up for text clustering (TF·IDF and cosine distance) in the unsupervised algorithms, showing that our proposal is not only competitive but also significantly improves the traditional set-up. In this paper we will asses our proposal with two clustering families: partitional and spectral algorithms. Next we will briefly revise the algorithms:

3.1 Partitional algorithms

The batch K-Means (KM, [17]) algorithm is a well-known efficient iterative clustering algorithm. It is one of the most popular ones due to its simplicity and good performance, which enables its use in large datasets.

A constrained counterpart of KM is the Soft Constrained K-Means (SCKM) [2]. SCKM is an extension to KM which allows the introduction of soft constraints in the clustering by altering the similarity values between documents and centroids: the similarity score is initialised with the similarity between the document and the centroid of the cluster, and it will be modified depending on the soft constraints affecting the data instance. Namely, the score of a cluster is increased a certain amount w for each document which was last assigned to that cluster and has a constraint with the document being assigned.

3.2 Spectral Algorithms

Spectral Clustering algorithms use graph spectral techniques to tackle the clustering problem transforming it into a graph cut problem. Thus, finding a *good* clustering of the data in k clusters can be reformulated in terms of finding a *good* cut of a weighted graph where each vertex corresponds to a data point and the weight of an edge is proportional to the similarity between data points. One of the most popular is Normalised Cut (NC, [18]), defined in a way such a cut of the graph with a low NC value corresponds to a good (as defined above) clustering of the data. Hence, the Normalised Cut (NC) algorithm proceeds building the graph from the data and finding a cut of it with a small NC value.

It can be shown that the minimisation of NC can be presented as a matrix trace minimisation problem [18], which, if subject to some constraints, would

yield the exact solution. Unfortunately this is NP-hard problem, and so the constraints have to be relaxed in order to make the algorithm computationally affordable. With this relaxation, the documents are projected in a reduced space ($\mathbb{R}^k$, where k is the desired number of clusters) using the smallest k eigenvectors of a Laplacian matrix of the graph. Given these projections, K-Means is used to find a discrete segmentation of this space. Once this segmentation has been performed, we can backtrace each projected document to the original one, obtaining the final outcome of the NC clustering algorithm.

In [10] the authors proposed a Constrained Normalised Cut (CNC) algorithm which introduces soft constraints in NC. In order to do so, they altered the function minimised in the NC algorithm to obtain a new one, such that the cut of the graph which minimises this function would convey a grouping which is still a good one but also tries to respect the constraints supplied by the user. To achieve this, they built a new matrix which encodes positive constraints and introduced it in the core of the minimisation problem, controlling the degree of enforcement of the constraints with a parameter β, with higher values of this parameter meaning a tighter enforcement. The results of the minimisation is a projection of the points in $\mathbb{R}^k$, and so a segmentation of the projected documents has to be performed in order to produce the final clustering of the data.

Two considerations have to be done about the spectral methods. It is very common to pre-process the similarity matrix between documents with a Gaussian Filter. When using IDM as distance function its form is:

$$e^{\left(\frac{-IDM(d^x,d^y)}{2\sigma^2}\right)} \tag{10}$$

Also in practice the dimension of the reduced space is taken greater than k (the number of desired clusters) because it performs considerably better [11], let us call this dimension δ, the number of eigenvectors keep in the projection phase.

4 Experiments and Results

The primary objective of this paper is to assess the use of unsupervised clustering algorithms for the constrained clustering task by modifying the document representation. So in the experiments we will compare the performance of two different family of clustering algorithms, partitional and spectral ones, by their traditional formulation (KM and NC), the constrained counterparts (SCKM and CNC), and the traditional formulation with the constraints modelled in the document representation (KM_{RM} and NC_{RM}).

4.1 Constraints and Seed Initialisations

All the presented algorithms are affected by the seed initialization problem of the KM algorithm. In order to reduce that problem, for every algorithm we did ten runs with different seeds, the same seeds in each collection for the six different algorithms. The results reported in the table 2 are the average for the ten different initialisations.

KM and NC are not affected by constraints (their values are reported as baselines), for SCKM, CNC, KM_{RM} and NC_{RM} we have to consider also the constraint generation. So for every seed initialisation, we did five different randomly chosen constraints sets. These constraints represent the 1% of all the possible constraints and the same constraints are used in each collection for the four different algorithms. The result for every seed initialisation in the algorithms affected by constraints is the average of the five different constraints sets. The constraints were created from the reference grouping used as clustering ground truth by randomly selecting pairs of documents which belonged to the same cluster, as it is traditionally done in constrained clustering evaluation.

4.2 Collections

We run experiments with publicly available datasets that have been widely used in the evaluation of clustering algorithms:

1. ModApte10: a split of Reuters-21578 with documents belonging to one of the biggest ten categories considering only the documents categorised in only one group (7282 documents, 10 groups)
2. WebKBUniversities: the WebKB dataset with the golden truth corresponding to universities, and taking only the documents from Cornell, Texas, Washington and Wisconsin universities and removing those corresponding to "misc", "other" and "department" (1087 documents, 4 groups).
3. WebKBTopics: the same dataset as (2), but this time distributed in five groups, corresponding to the topics "course", "faculty", "project", "staff", and "student" (1087 documents, 5 groups).
4. News3Related: a sample of three categories of the 20 Newsgroups collection. Following the same approach in [6], we have chosen 300 documents randomly from each of the categories `talk.politics.misc`, `talk.politics.guns`, and `talk.politics.mideast` (900 documents, 3 groups).

We decided to choose the WebKB collection and both of its categorization because in this collection the bias problem occurs, tending the clustering algorithms to follow one of the categorizations. Dealing with this problem is a very common task for the constrained clustering algorithms (avoiding bias task), therefore it is an interesting collection for the evaluation of constrained clustering algorithms. The use of small datasets comprised by sparse high-dimensional data is interesting because the clustering task is notably difficult, as the clustering algorithms are more prone to fall in local minima [6].

4.3 Metrics

In order to assess the effectiveness of the different clustering algorithms we have compared the outcomes of the algorithms with the reference groupings using three metrics: Adjusted Rand Index (ARI), Purity and Entropy. However, as the results for the three metrics show the same trends, only the results for Adjusted

Rand Index [9] are presented in this paper. This metric measures the ratio of good decisions made by the algorithm over a collection of n data points on a pairwise basis correcting certain deficiencies of the Rand Index. Higher values of Adjusted Rand Index indicate a greater similarity between the results and the reference.

4.4 Parameter Training

To deal with the values of the parameters involved in the different approaches we decided to use traditional training and test methodology. We tuned the parameters for ARI in the ModApte collection, and the trained values were used in the other collections. However the parameters σ (the Gaussian filter parameter) $\sigma \in \{0, 0.05, 0.10, 0.15...0.90, 0.95, 1\}$ and δ (the number of eigenvectors keep in the projection phase) $\delta \in \{1, 5, 10, ...|C|\}$ involved in the spectral algorithms had to be tuned for every collection because they are very sensitive, they were tuned in the NC algorithm. So the parameters tuned were: the parameters w and β for the enforcement of the constraints in the SCKM and CNC algorithms take values in $\{0.00250, 0.00500, 0.0125, 0.0250, 0.0500\}$ and $\{5, 10, 20, 30\}$ respectively. The parameters involved in the RM estimation namely, the Dirichlet smoothing parameter μ which takes values in $\{5, 10, 15, 25, 50, 100, 500, 1000\}$, was trained in the KM algorithm and the same values used in the NC algorithm, the parameter e (the number of terms selected from the Relevance Model) was set to 500 without tuning it. Furthermore, the interpolation parameter λ which takes values in $\{0, 0.1, 0.2, ...0.9, 1\}$ was tuned using the same strategy as with μ.

In the experiments as usually we have considered that the number of clusters (k) in the grouping used as reference was known, and so the number of desired clusters was set to that amount in each of the tested clustering algorithms

4.5 Statistical Significance

Finally, we have assessed the statistical significance of the results of the experiments using the Sign Test [8], a choice which was motivated by its reduced number of assumptions about the data in comparison with other tests such as Wilcoxon's or Student's t. The results of each approach were compared with the rest of the methods for every collection. For each test ten observations $(ARI_{xi}, ARI_{yi}), i \in [1..10]$ were considered, one for each initialisation of the seeds, where ARI_{xi} is the ARI of the method X and ARI_{yi} is the ARI for Y. Over these observations we performed a Lower-Tailed test, where the null hypothesis was $H_0 : P(+) \geq P(-)$, i.e. , that the values ARI_{xi} were greater or equal to ARI_{yi} (meaning that the quality of the results of the method X was greater or comparable to that of the Y method), and the alternative hypothesis was $H_1 : P(+) < P(-)$.

4.6 Results

In order to clarify the competitiveness of the baselines given the experimental conditions in terms of document representations and distance measures, a

Table 1. Adjusted Rand Index values, statistical significant improvements w.r.t to the alternative set-up for each algorithm according with the Sign Test are starred (the null hypothesis is rejected for a p-value ≤ 0.0547).

	ARI	
Set-up	*KM*	*NC*
TF·IDF and Cosine	0.319	0.311
MLE and IDM	$0.446^{\star}$	$0.648^{\star}$

preliminary experiment was carried out in the ModApte collection comparing for both KM and NC the averaged ARI values when using classical TF·IDF document representation and cosine distance function and when using the experimental conditions designed in this paper. Results are reported in Table 1 showing not only that the probabilistic representation in combination with the IDM measure performs well but it also significantly outperforms the classical clustering set-up.

Table 2. Adjusted Rand Index values, statistical significant improvements w.r.t KM, SCKM, KM_{RM}, NC, CNC and NC_{RM} according with the Sign Test are marked as k, s, κ, n, c, η respectively (the null hypothesis is rejected for a p-value ≤ 0.0547). Best values bolded.

	ARI					
Collection	*KM*	*SCKM*	KM_{RM}	*NC*	*CNC*	NC_{RM}
ModApte (Training)	0.446	$\mathbf{0.983}^{k\kappa nc\eta}$	$0.820^{knc\eta}$	0.648^{k}	0.771^{kn}	0.781^{kn}
WebKBUniversities	0.073	0.311^{kn}	$\mathbf{0.581}^{ksnc\eta}$	0.009	0.342^{kn}	0.377^{ksnc}
WebKBTopics	0.230	0.574^{kn}	0.505^{kn}	0.331^{k}	$\mathbf{0.734}^{ks\kappa n\eta}$	$0.668^{ks\kappa n}$
News3Related	0.183	$0.712^{kn\eta}$	$\mathbf{0.833}^{ksnc\eta}$	0.258^{k}	$0.783^{ks\eta}$	0.617^{kn}

In Table 2 an effectiveness comparison between the different approaches is presented in terms of ARI. When analysing the results the first consideration is that, as expected, the presented approach performs significantly much better than the unconstrained algorithms, showing that it is a valid approach for the constrained clustering task. When comparing with the ad-hoc constrained clustering algorithms we have to remark that in WebKBUniversities and News3Related collections the best method is one based on the language modelling of the constraints and performs significantly better than both constrained algorithms, meanwhile only the CNC can achieve significant improvements over both RM based approaches in only one testing collection (WebKBTopics). The evaluation, as commented before, showed similar trends for the other metrics (Purity and Entropy). These numbers show that the proposed approach is valid for the constrained clustering task, achieving results comparable or even better than specially tailored clustering algorithms.

In additional experiments not reported here, it is also showed that the parameters of the constrained clustering algorithms are much less stable than the

λ parameter of the presented approach. Also it is an advantage that the λ parameter behaviour has been widely studied in other retrieval tasks. On the other hand the interpretability of the role of the parameter λ is very easy and it will only depend on the importance that we want to assign to the constraints in the interpolated model.

5 Related Work

So far with the approach presented in this paper three alternatives exists for the introduction of constraints in the clustering process. (a) The presented approach to introduce the constraints directly in the document representation. (b) The design of new specially tailored algorithms as the ones commented in Section 1 based on forcing the accomplishment of the constraints in the document to cluster assignment or by modifying the objective functions. (c) An alternative approach introduces the constraints in the clustering process through the use of distance learning methods. In [22] Xing et al. present an algorithm that given some constraints learns a distance metric over $\mathbb{R}^n$ respecting those constraints, however this requires to solve a convex optimization problem.

6 Conclusions and Future Work

In this paper we have proposed the use of unsupervised clustering algorithms for the constrained clustering task. The main contributions are two: the use of the document representation to code the constraints and the use of Relevance Models under the LM framework to model those constraints. The evaluation showed that the use of our proposal with the traditional clustering algorithms achieves comparable and even better results than specially tailored constrained clustering algorithms, allowing in this way the use of the unsupervised algorithms for the constrained clustering task. The proposal has been build upon a strong and well-studied theoretical base as is the Language Modelling framework which allows the interpretability of the elements involved in the approach pretty straightforward. As future work we want to test our approach in additional clustering frameworks and test other PRF techniques such as the Rocchio's framework and also to study how to accommodate other kind of constraints (absolute and negative) in this framework.

Acknowledgements: This work was funded by *Secretaría de Estado de Investigación, Desarrollo e Innovación* under project TIN2008-06566-C04-04.

References

1. N. Abdul-jaleel, J. Allan, W. B. Croft, O. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade. UMass at trec 2004: Novelty and hard. In *In Proceedings of TREC-13*, 2004.
2. M. E. Ares, J. Parapar, and A. Barreiro. Avoiding bias in text clustering using constrained k-means and may-not-links. In *ICTIR 2009*, pp. 322–329.

3. E. Bae and J. Bailey. Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *ICDM 2006*, pp. 53–62.
4. N. Balasubramanian, J. Allan, and W. B. Croft. A comparison of sentence retrieval techniques. In ACM SIGIR 2007, pp. 813–814.
5. A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, December 2005.
6. S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In ACM KDD 2004, pp. 59–68.
7. S. Basu, I. Davidson, and K. Wagstaff. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, 2008.
8. W. J. Conover. *Practical nonparametric statistics*. John Wiley & Sons, New York, third edition, 1971.
9. L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.
10. X. Ji and W. Xu. Document clustering with prior knowledge. In ACM SIGIR 2006, pp. 405–412.
11. R. Jin, C. Ding, and F. Kang. A probabilistic approach for optimizing spectral clustering. In *In Advances in Neural Information Processing Systems 18*, 2005.
12. D. Klein, S. Kamvar, and C. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In ICML 2002, pp. 307–314.
13. V. Lavrenko and W. B. Croft. Relevance based language models. In ACM SIGIR 2001, pp. 120–127.
14. K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In ACM SIGIR 2008, pp. 235–242.
15. X. Li and Z. Zhu. Enhancing relevance models with adaptive passage retrieval. In ECIR 2008, pp. 463–471.
16. Y. Lv and C. Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In ACM CIKM 2009, pp. 1895–1898.
17. J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297, 1967.
18. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.
19. K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In ICML 2000, pp. 1103–1110.
20. K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In ICML 2001, pp. 577–584.
21. F. Wang, T. Li, and C. Zhang. Semi-supervised clustering via matrix factorization. In SDM 2008, pp. 1–12.
22. E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance Metric Learning, with Application to Clustering with Side-information. In *Advances in Neural Information Processing Systems* 15, pp. 505–512, 2002.
23. H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In ACM SIGIR 2004, pp. 210–217.
24. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.
25. Z. Zhai, B. Liu, H. Xu, and P. Jia Constrained LDA for grouping product features in opinion mining. In PAKDD 2011, pp. 448–459.