Improving Text Clustering with Social Tagging

M. Eduardo Ares and **Javier Parapar** and **Álvaro Barreiro**
Information Retrieval Lab, Computer Science Department, University of A Coruña
Facultade de Informática, Campus de Elviña S/N, 15071 A Coruña, Spain
{maresb,javierparapar,barreiro}@udc.es

Abstract

In this paper we study the use of social bookmarking to improve the quality of text clustering. Recently constrained clustering algorithms have been presented as a successful tool to introduce domain knowledge in the clustering process. This paper uses the tags saved by the users of Delicious to generate non artificial constraints for constrained clustering algorithms. The study demonstrates that it is possible to achieve a high percentage of good constraints with this simple approach and, more importantly, the evaluation shows that the use of these constraints produces a great improvement (up to 91.25%) of the clustering algorithms effectiveness.

Introduction and Motivation

Lately several web-based tagging systems such as Technorati, Flickr or Delicious have become very popular. In this paper we will exploit the information created by the community in Delicious: a social bookmarking service where the users can save the URLs of their favourite web-pages offering also the possibility of associating tags to them.

On the other hand the clustering methods are a very important data mining tool in order to exploit the knowledge present in data collections. In the last years a new family of clustering algorithms, constrained clustering (Basu, Davidson, and Wagstaff 2008), has achieved great importance because they enable the introduction of domain knowledge in the clustering process. The work presented in this paper uses the Delicious tags to generate positive soft constraints between documents (documents that share some tags are likely to be in the same cluster) and evaluates the effect of using those constrains in two different constrained clustering algorithms (Constrained Normalized Cut (Ji and Xu 2006) and Soft Constrained K-Means (Ares, Parapar, and Barreiro 2009)). The evaluation carried out showed large improvements over their non-constrained counterparts (Normalized Cut (Shi and Malik 2000) and K-Means (MacQueen 1967)) when using these "social-constraints".

To the best of our knowledge, this is the first time in which the information in social tags is used in the form of constraints to improve the outcome of a clustering process. Previous efforts to incorporate that information (Ramage et al. 2009) have been oriented to use tags in an extended vector space model that includes tags and page text or to model jointly words and tags with latent Dirichlet allocation.

Social Tags and Constrained Clustering

Given the tags associated to the documents by the users of Delicious, the most straightforward option to translate that information into constraints could be creating a positive constraint between two documents d_i and d_j (meaning that they should be in the same cluster) if they share some tag. This simple approach is quite naive because some common tags can produce a lot of non valid constraints. Hence, the approach we have followed was generating a constraint between two documents if they have in common at least t tags.

Another important question is the absoluteness of the constraints. Even if we use this approach to turn tags into constraints, a fair amount of them are bound to be inaccurate (i.e., linking documents which should not be in the same cluster) until a high value of the parameter t, due to the polysemy of the terms used as tags or to differences in the criteria of the taggers. Consequently, we have used soft positive constraints, meaning that the documents affected by one of them are likely to be in the same cluster, without forcing the clustering algorithm to actually put them so.

In this paper we have used two constrained clustering algorithms: a spectral one, Constrained Normalized Cut (CNC) and a partitional one, Soft Constrained k-Means (SCKM). CNC, introduced in (Ji and Xu 2006), alters the eigenproblem at the core of the Normalised Cut (NC) method (Shi and Malik 2000) adding a new term which encodes positive constraints. SCKM, introduced in (Ares, Parapar, and Barreiro 2009), extends the Constrained k-Means algorithm (Wagstaff et al. 2001) to allow the use of soft constraints. The assignment policy is similar to that of k-Means, but the similarity score between a document and a centroid is altered depending on the nature of the constraints which affect the document. In both algorithms the strength of the constraints is controlled by a parameter (β in CNC and w in SCKM), with higher values of the parameter meaning a greater strength of the constraints.

Evaluation Methodology

We have used the classic methodology in the evaluation of clustering experiments. Starting from a set of documents

Category	# docs
Computers	3401
Regional	1645
Arts	1215
Science	891
Society	865
World	632
Reference	594
Business	563
Shopping	528
Home	361
Games	328
Recreation	278
Health	110
News	105
Sports	73
Total	11589

Top 10 tags	# docs
reference	3790
tools	2706
software	2674
design	2397
web	2108
blog	2088
free	2076
programming	1791
development	1790
resources	1687

Table 1: Dataset description

that have been categorised by hand, we will apply the proposed approach and compare its outcome with the manual reference using certain metrics.

In our experiments we have used a subset from the DeliciousT140 dataset[1], which contains 144,574 web documents tagged with a total of 67,104 tags (Zubiaga et al. 2009). As this collection does not contain a categorisation of the documents which could be used as a reference in the experiments, we have created one of our own, using the Open Directory Project[2] (ODP). The intersection between the URLs in DeliciousT140 and ODP (using the dump made on 2010-10-25) yielded, after removing those in which the text extracted (using HTML parser) was empty, 11,589 documents. Those were the documents used in the experiments, in which they were represented using Mutual Information, as it has been shown to outperform any other $tf \cdot idf$ based approach (Pantel and Lin 2002). The golden truth[3] was created assigning each document to its corresponding top-level category in the ODP hierarchy. The final dataset is described in Table 1.

We have compared the results of our method with those of Normalised Cut (NC) and k-Means (KM), the non constrained counterparts of the algorithms tested for our approach, using three document representations: only documents, documents+tags (i.e., tags appended to the document) and only tags. We also report the results of an upperbound model which uses the set of constraints yielded by filtering with a perfect oracle the constraints resulting from linking two documents if they share one or more tags.

To compare the output of the algorithms with the reference we have used Adjusted Rand Index (ARI) (Hubert and Arabie 1985). Based on Rand Index, ARI measures the amount of good decisions made by the algorithms on a pairwise basis correcting certain deficiencies of that metric.

[1] http://nlp.uned.es/social-tagging/delicioust140/

[2] www.dmoz.org

[3] Available on www.dc.fi.udc.es/~edu/DT140dmozRef.tar.gz

	Only Docs	Only Tags	Docs + Tags
NC	0.1781	**0.1899**	0.1660
(best d)	(95)	**(19)**	(21)
KM	0.1465	**0.1864**	0.1817

Table 2: Results of the baselines (best values in **bold**)

t	# accurate constraints	# constraints	% accurate constraints
1	6,530,518	30,477,693	21.43
2	4,489,184	15,524,394	28.92
3	3,011,039	8,231,157	36.58
4	1,953,132	4,461,885	43.77
5	1,227,851	2,450,620	50.10
6	750,981	1,355,464	55.40
7	445,971	747,349	59.67
8	259,020	410,263	63.14
9	148,549	225,360	65.92
10	84,553	123,906	68.24
11	47,426	67,688	70.07
12	26,693	37,216	71.72

Table 3: Evolution of the number and ratio of accurate constraints as t increases

Experimentation and Results

In all the experiments reported the number of clusters that the algorithm should look for in the data (k) was assumed to be know, setting it to 15, the number of classes of the golden truth. Also, for the spectral algorithms, NC and CNC, a special consideration has to be made. Even though in the literature the number of eigenvectors used in both algorithms (which we would call d) is almost always assumed to be equal to the number of desired clusters (i.e., k), we have detected (in consonance with (Jin, Ding, and Kang 2005)) that better results can be obtained if a bigger number of eigenvectors is used. Consequently, we have considered d a parameter, testing values for it between 15 and 300.

Finally, as the results of the clustering performed by the four algorithms (NC, CNC, KM and SCKM) are very dependent on the choice of the clustering seeds, the same 10 random sets of seeds were tested in order to have a good representation of the effectiveness of each algorithm; we report the average of these ten initialisations. Furthermore, in the case of SCKM in each of the seed initialisations the order in which the documents are inspected and assigned to a cluster was also randomised, to minimise any possible influence over the results.

The results obtained by the baselines introduced above are shown in Table 2. It can be seen that the best results are obtained by the second approach, in which the pages are only represented by the tags associated with them, improving the results obtained by using only the documents contents (first approach). That is to say, tags seem to be good representations of what a document is about, and maybe even better its content itself.

We will centre the remaining of the paper on the clustering

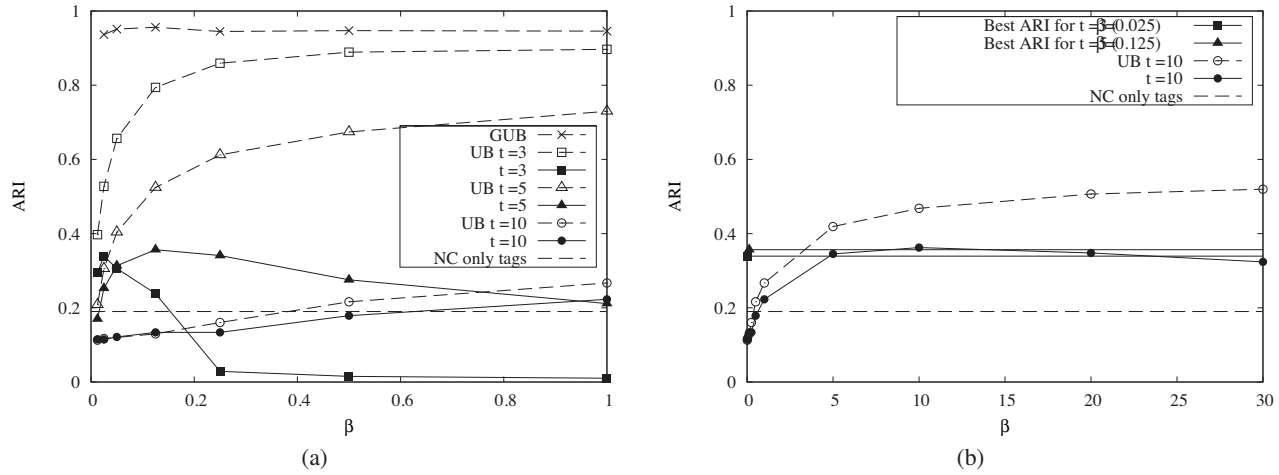

Figure 1: Results using Constrained Normalised Cut (CNC) with d=225

results with the constraints sets resulting from setting t to 3, 5 and 10, as they showcase (Table 3) an interesting array of situations: lots of constraints, but very noisy ($t = 3$), moderate number of constraints with moderate noise ($t = 5$) and (relatively) small amounts of constraints and noise ($t = 10$). To give a better understanding of the results, apart from the results obtained by the global upper bound introduced in the previous (referenced as "GUB" in the labels), we will also show the results when using only the accurate constraints in those subsets (marked with a leading "UB" in the labels).

Figure 1 shows the results obtained using CNC as the constrained clustering algorithm in the core of our approach. In our experiments we have detected that the clustering can not be successfully performed for some constraints sets (specially for the Upper-bound model) in several initialisations of the seeds when low values of d (<225) were used, because one or more clusters became empty in the middle of the process (effectively preventing the clustering process to continue). As with higher values of d we have seen that the changes in the ARI were negligible we report only the results with d=225. Figure 1(a) shows the results for β in (0,1] while Figure 1(b) focuses on the interval (0,30]. The first important result is the high improvement potential of using constraints created from social tags. In this example, the global upper bound model reaches an ARI of about 0.95, much higher than that of the best baseline (0.19). However, this is a theoretical result, showing how much the clustering results could improve if we had a perfect way to filter the noisy constraints. As for the filtering method proposed in this paper (requiring some number of tags in common to create a constraint), the results show that the constraints which pass that filter are able to improve the results of the baseline in the three scenarios tested, almost doubling it (a summary of the results is shown in Table 4). However, the method is not perfect: if we compare the results obtained when only the accurate constraints surviving the filter are used (those beginning with "UB") it can be seen how the inaccurate constraints that evade the cut are still able to harm noticeably the

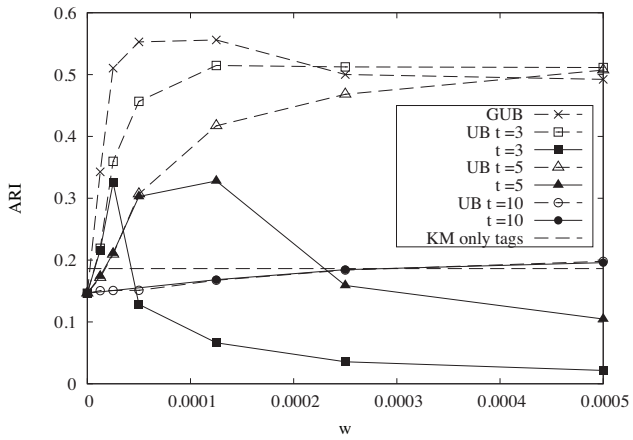

Figure 2: Results using Soft Constrained k-Means (SCKM)

quality of the results.

Finally, some interesting insights can be obtained analysing the behaviour of the parameter β for the three sets of constraints. Beginning from β=0, at first increasing its value provides an improvement of the quality of the clustering until a certain peak value, from which the ARI starts to decrease slowly. In our opinion that peak point is where the influences of the similarity between documents and the constraints are balanced, and thus the information in each one is put to its best use. Regarding this, it is important to note how not only this best β is higher when the set of constraints is more accurate, but also how in that case that best value is also more stable. Indeed, as it can be seen in Figure 1(b), with t=10 (which provides a ratio of accurate constraints of about 68%) wide variations of the best β (10, much higher than those for t=3 and t=5, with ratios of 50% and 37%) do not decrease much the quality of the results. This same phenomenon, albeit in a lesser scale, can be appreciated in Figure 1(a) when comparing the results for t=3 and t=5.

Table 4: Comparison of the best ARI for each constraints set and best baseline for each algorithm (best values in **bold**)

	CNC	SCKM
t = 3	0.3393 (β=0.025)	0.3253 (w=2.5 E-5)
t = 5	0.3410 (β=0.125)	**0.3281** (w=1.25 E-4)
t = 10	**0.3632** (β=10.0)	0.2935 (w=2.5 E-3)
Only tags	0.1899 (using NC)	0.1864 (using KM)

The results when using SCKM (Figure 2) have a global behaviour very similar to those obtained when using CNC. When using this algorithm the best result of the global upper bound is 0.82 for w=0.025 (this point is not shown in the figure, which is focused on the interval [0, 5E-4] due to space constraints). This value is again a great improvement over the baseline (0.19), reinforcing the idea of creating constraints as an effective way to exploit social tags. Even so, the difference with the best value when using CNC (0.95) is quite patent, which we attribute to the expected difference of effectiveness between using a partitional and a spectral clustering algorithm (also noticeable in Table 2 in the difference of ARIs when using only documents).

With respect to the results with the three tested sets of constraints it is interesting to see how, despite the aforementioned difference in the global upper bounds, the best values of SCKM are close to those of CNC (Table 4). Also, the evolution of the results when moving the parameter w mimics what was observed on CNC: an initial rise in quality, a peak point and a slower decrease, with more accuracy in the constraints entailing a higher best w and more stability for that parameter (the vicinity of the best w for t=10 is again not shown in Figure 2 due to space constraints). However, it should be noted that the parameter w in SCKM is globally much less stable than β in CNC; the Figure 2 shows how really small variations of the best values of w (note the range of the x axis) cause strong drops in the quality of the clustering.

Conclusions and Future Work

In this paper we have proposed a method to use in a clustering process the information contained in social tags with the aid of a constrained clustering algorithm, turning the tags into constraints between documents.

This proposal was evaluated with standard clustering test methodology and using two different algorithms, a partitional one and a spectral one. The results showed in both cases substantial improvements over the methods used as baselines, some of which use the tags in the representation of the documents. For instance, Table 4 shows that a 91% improvement is attained over a high performing baseline such as Normalised Cut. Hence, the proposed method is shown to be a valid approach to improve the clustering of web documents with social tagging. To the best of our knowledge, this is the first attempt at using constraints to incorporate social tag information in a clustering process.

Acknowledgements

This work was funded by *Ministerio de Ciencia e Innovación* under project TIN2008-06566-C04-04. The first author also wants to acknowledge the support of *Ministerio de Educación* with FPU grant AP2007-02476.

References

Ares, M. E.; Parapar, J.; and Barreiro, A. 2009. Avoiding bias in text clustering using constrained k-means and may-not-links. In *ICTIR '09: Proceedings of the 2nd International Conference on Theory of Information Retrieval*, 322–329. Berlin, Heidelberg: Springer-Verlag.

Basu, S.; Davidson, I.; and Wagstaff, K. 2008. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC.

Hubert, L., and Arabie, P. 1985. Comparing partitions. *Journal of Classification* 2:193–218.

Ji, X., and Xu, W. 2006. Document clustering with prior knowledge. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 405–412. ACM Press.

Jin, R.; Ding, C. H. Q.; and Kang, F. 2005. A probabilistic approach for optimizing spectral clustering. In *Neural Information Processing Systems Foundation*.

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. In Cam, L. M. L., and Neyman, J., eds., *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 281–297. University of California Press.

Pantel, P., and Lin, D. 2002. Document clustering with committees. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 199–206. ACM Press.

Ramage, D.; Heymann, P.; Manning, C. D.; and Garcia-Molina, H. 2009. Clustering the tagged web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, 54–63. New York, NY, USA: ACM.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22:888–905.

Wagstaff, K.; Cardie, C.; Rogers, S.; and Schrödl, S. 2001. Constrained k-means clustering with background knowledge. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, 577–584. Morgan Kaufmann Publishers Inc.

Zubiaga, A.; García-Plaza, A. P.; Fresno, V.; and Martínez, R. 2009. Content-based clustering for tag cloud visualization. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*, 316–319. Washington, DC, USA: IEEE Computer Society.