Multi-view Clustering of Multilingual Documents

Young-Min Kim[†] Massih-Reza Amini[‡] Cyril Goutte[‡] Patrick Gallinari[†]

[†]Université Pierre et Marie Curie
Laboratoire d'Informatique de Paris 6
104, avenue du Président Kennedy
75016 Paris, France
First.Last@lip6.fr

[‡]National Research Council Canada
Institute for Information Technology
283, boulevard Alexandre-Taché
Gatineau, J8X 3X7, Canada
First.Last@nrc-cnrc.gc.ca

ABSTRACT

We propose a new multi-view clustering method which uses clustering results obtained on each view as a voting pattern in order to construct a new set of multi-view clusters. Our experiments on a multilingual corpus of documents show that performance increases significantly over simple concatenation and another multi-view clustering technique.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering; I.5.3 [**Clustering**]: Algorithms

General Terms

Algorithms, Experimentation

Keywords

Multilingual document clustering, Multi-view learning, PLSA

1. INTRODUCTION

Much data is now available in multiple representations, or *views*, for example multimedia content or web pages translated into several languages. Multi-view learning is a principled approach to handle these kinds of documents using the relations between the multiple views. The key is to leverage each view's characteristics in order to do better than simply concatenating views. Recently, multi-view clustering methods have been proposed that address this situation and have been shown to improve over traditional single-view clustering. [2] proposed an extension of k-means and EM for a dataset with two views, and [5] presented a late fusion approach which re-estimates the relationship between documents from single-view clustering results. [3] and [6] show that dimensionality reduction via canonical correlation between views gives better results for document clustering than via principal components analysis or random projections.

This paper introduces a novel multi-view clustering method. Our approach consists of two steps. First we find robust topics for each view using the PLSA approach. The topic pattern over the multiple views defines cluster signatures for each document. We use those to prime a second-stage clustering process over all the views. Experiments carried out on a large five language corpus of Reuters documents show that we consistently improve over competing techniques.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

2. MUTLI-VIEW CLUSTERING

We consider a multilingual document as $\mathbf{d} \stackrel{def}{=} (d_1, ..., d_V)$ where each version or view $d_v, v \in \{1, ..., V\}$ provides a representation of document $\mathbf{d}$ in a different language, with feature space $\mathcal{X}_v$. Our algorithm operates in two steps.

2.1 Stage I - Single-view clustering

At the first stage of our multilingual clustering, we apply Probabilistic Latent Semantic Analysis (PLSA) [7] independently over each of the V languages, constraining each model to have the same number of unobserved topics. For every view v, the probability that document d_v arises from topic $z \in Z$ is given by $p(z|d_v)$, estimated by PLSA. Documents are then assigned to each topic using the maximum posterior probability. We hence obtain a set of V estimated topics $(z^1_{\mathbf{d}}, ..., z^V_{\mathbf{d}})$ for each document $\mathbf{d}$, which we call the *voting pattern* in the following. Each $z^v_{\mathbf{d}}$ indicates the estimated topic index of $\mathbf{d}$ on the v^{th} view according to the view-specific PLSA model.

2.2 Stage II - Voting & Multi-view clustering

Once a *voting pattern* is obtained for each multilingual document, we attempt to group documents such that in each group, documents share similar voting patterns. As documents belonging to each of these groups received by definition similar votes from the view-specific PLSA models, the voting pattern representing each of these groups is called the *cluster signature*. We keep the C largest groups with the most documents as initial clusters. Documents that have voting patterns with at least $V - 1$ in common with a *cluster signature* are pre-assigned to that cluster. The remaining documents have *voting patterns* different from any of the selected *cluster signatures*. They are matched to one of these C groups by applying a PLSA model on the concatenated document features.

The parameters of the final PLSA model are first initialized using the documents that have been pre-assigned to the selected *cluster signatures*. For these documents $p(c \mid \mathbf{d})$ has a binary value equal to 1 if $\mathbf{d}$ belongs to cluster c and 0 otherwise. For the remaining documents, posteriors are estimated at each iteration as in the traditional `E-step`. In the `M-step`, after updating model parameters, we keep the val-

ues of $p(c \mid \mathbf{d})$ fixed for the pre-assigned documents. After convergence, documents are assigned to the clusters using the posteriors $p(c \mid \mathbf{d})$. Note that any generative model giving $p(c \mid \mathbf{d})$ may be employed instead of PLSA, such as Latent Dirichlet Allocation [4].

3. EXPERIMENTS

We perform experiments on a publicly available multilingual multi-view text categorization corpus extracted from the Reuters RCV1/RCV2 corpus [1].[1] This corpus contains more than 110K documents from 5 different languages, (English, German, French, Italian, Spanish), distributed over 6 classes. The multilingual collection is originally a comparable corpus as it covers the same subset of topics in all languages. In order to produce multiple views for documents, each original document extracted from the Reuters corpus was translated in all other languages using a phrase-based statistical machine translation system. The indexed translations are part of the corpus distribution.

Experiments are repeated 10 times on the whole dataset, using different random initializations of the PLSA models. The number of topics in each single-view PLSA model as well as the number of clusters C are fixed to 6, the number of classes in the collection. We used the micro-averaged precision (micro-AvgPre) as well as the Normalized Mutual Information (NMI) to measure clustering results [8]. In order to use these evaluation measures, the predicted label for each cluster is the label of the most dominant class in that cluster. The reported performance is averaged over the 10 different runs. To validate our approach we compare our algorithm (denoted by voted-PLSA in the following) with a PLSA model operating over the concatenated feature representations of documents (conc-PLSA) and the late fusion approach (Fusion-LM) for multi-view clustering [5].

First, we are interested in the clustering results after the first step of our algorithm on the $C = 6$ largest clusters containing each the same *voting pattern* documents. Table 1 shows the micro-AvgPre performance of the clustering results per language as well as the percentage of documents being grouped with our voting strategy. We observe that partitions formed using the votes of single-view models contain more than half of the documents in the collection and that these groups are highly homogeneous with an average precision of 0.76.

Table 1: Proportion of pre-assigned documents and average precision on those, obtained from the first stage single-view PLSA models.

Language	% of documents	micro-AvgPre
English	51.18	0.79
French	63.85	0.78
German	67.44	0.80
Italian	58.03	0.60
Spanish	73.73	0.81
Average	62.84	0.76

Table 2 summarizes results obtained by conc-PLSA, Fusion-LM and voted-PLSA averaged over five languages and 10 different initializations. We use bold face to indicate the highest performance rates, and the symbol $\downarrow$ indicates that performance is significantly worse than the best result, according to a Wilcoxon rank sum test used at a p-value threshold of 0.05. Note that in our approach, the second stage multi-view clustering model relies on a PLSA on the concatenated views, just as in conc-PLSA. This suggests that the difference of 2 to 3 points in micro-AvgPre and NMI (respectively) between voted-PLSA and conc-PLSA shows the real impact of the first stage voting process. In addition, both voted-PLSA and conc-PLSA perform at least as well as Fusion-LM.

[1] http://multilingreuters.iit.nrc.ca/

Table 2: micro-AvgPre and NMI of different clustering techniques averaged over 10 initialization sets and 5 languages.

Strategy	micro-AvgPre	NMI
conc-PLSA	$0.63^{\downarrow}$	$0.41^{\downarrow}$
Fusion-LM	$0.61^{\downarrow}$	$0.41^{\downarrow}$
voted-PLSA	**0.65**	**0.44**

4. CONCLUSIONS

We presented a multi-view clustering approach for multilingual document clustering. The proposed approach is an incremental algorithm which first groups documents having the same *voting patterns* assigned by view-specific PLSA models. Working in the concatenated feature spaces the remaining unclustered documents are then assigned to the groups using a constrained PLSA model. Our results have brought to light the positive impact of the first stage of our approach which can be viewed as a voting mechanism over different views. The effect of the length of these *voting patterns* and the number of latent variables in view-specific PLSA models are interesting avenues for future research.

Acknowledgments

This work was supported in part by the IST Program of the EC, under the PASCAL2 Network of Excellence.

References

[1] M.-R. Amini, N. Usunier, and C. Goutte. Learning from multiple partially observed views - an application to multilingual text categorization. In *NIPS 22*, pages 28–36, 2009.

[2] S. Bickel and T. Scheffer. Multi-view clustering. In *ICDM-04*, pages 19–26, 2004.

[3] M. B. Blaschko and C. H. Lampert. Correlational spectral clustering. In *CVPR-08*, 2008.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, pages 993–1022, 2003.

[5] E. Bruno and S. Marchand-Maillet. Multiview clustering: A late fusion approach using latent models. In *SIGIR-09*, pages 736–737, 2009.

[6] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *ICML-09*, pages 129–136, 2009.

[7] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, January 2001.

[8] N. Slonim, N. Friedman, and N. Tishby. Unsupervised document classification using sequential information maximization. In *SIGIR-02*, pages 129–136, 2002.