An Extension of PLSA for Document Clustering

Young-Min Kim, Jean-François Pessiot, Massih R. Amini, Patrick Gallinari
Computer Science Laboratory of Paris 6, University Pierre and Marie Curie
104, Avenue du President Kennedy, 75016 Paris, France
{kim,pessiot,amini,gallinari}@poleia.lip6.fr

ABSTRACT

In this paper we propose an extension of the PLSA model in which an extra latent variable allows the model to co-cluster documents and terms simultaneously. We show on three datasets that our extended model produces statistically significant improvements with respect to two clustering measures over the original PLSA and the multinomial mixture MM models.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

Document Clustering, PLSA

1. INTRODUCTION

With the ever-increasing volume of on-line textual information, an efficient partitioning of documents into clusters can constitute a real saving in terms of efficiency for various information retrieval or enterprise portal applications. Document clusters can for example, help users to quickly evaluate classical search engines results, or navigate through huge document collections [2]. They can also be useful for distributed search or extractive text summarization [1, 5].

Probabilistic Latent Semantic Analysis (PLSA) is a well known algorithm to model document collections. When applied to document clustering, latent topics in PLSA are identified to document clusters which may be too restrictive in cases where there are more topics than document clusters in a collection.

Our paper is organized as follows, in section 2, we extend the PLSA model (Ext-PLSA) by incorporating into its graphical representation an additional latent variable corresponding to document clusters. In section 3, we present experimental results showing that our proposed approach outperforms the initial PLSA and the Multinomial Mixture model (MM) over three text collection.

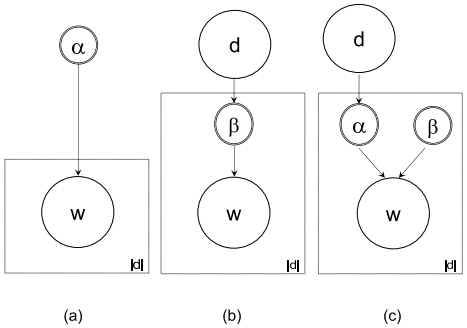

Figure 1: Graphical model representation of the Multinomial Mixture model (a), PLSA (b) and its extended version (c). The *plates* indicate the repeated sampling of the enclosed variables.

2. EXTENDED PLSA

Probabilistic Latent Semantic Analysis. The PLSA model introduced by [4] is a probabilistic model which characterizes each word in a document as a sample from a mixture model, where mixture components are conditionally-independent multinomial distributions. This model associates an unobserved latent variable (called aspect or concept) $\beta \in B$ to each observation corresponding to the occurrence of a word $w \in \mathcal{V}$ within a document $d \in \mathcal{D}$. The underlying generation process of this aspect model is shown in figure 1 (b).

When document clustering is performed with the PLSA model, the latent topics play the role of the document clusters. In this case, the probability of observing the topic β given the document d, $p(\beta \mid d)$ is interpreted as the probability that the document d belongs to the cluster β. The clustering is performed using : $\text{cluster}(d) = \text{argmax}_{\beta \in B}\, p(\beta \mid d)$

An Extended version of PLSA. When the number of latent topics of a document collection and the number of desired document clusters are different, the PLSA model can not be used for document clustering. We propose an exten-

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

sion of PLSA, which includes two latent variables α, β to model the topics on two different levels.
The corrresponding generative process is as follows :
- Pick a document d with probability $p(d)$,
- Choose a document topic α with probability $p(\alpha|d)$,
- Choose a word topic β with probability $p(\beta)$
- Generate a word w with probability $p(w|\alpha, \beta)$

Figure 1 (c) depicts this process. Words are in this case generated not only by latent topics (as it is the case with the PLSA model) but also by document clusters. This assumption hence enables the generative model to capture the discourse on two different semantic levels: on the general topics dealt in the collection expressed by variables α and on different sub-topics represented by variables β. In this case the generation of a word w within a document d can be expressed by the joint probability:

$$p(d, w) = \sum_{\alpha \in A} \sum_{\beta \in B} p(d)p(\alpha|d)p(\beta)p(w|\alpha, \beta)$$

Model parameters in this case are $\Phi = \{p(d), p(\alpha \mid d), p(\beta), p(w \mid \alpha, \beta) : d \in \mathcal{D}, w \in \mathcal{V}, \alpha \in A, \beta \in B\}$ and which are estimated by maximizing the log-likelihood function using an EM-like algorithm [3]. In the E-step we estimate the posterior probabilities of the latent variables :

$$p^{(t+1)}(\alpha, \beta|d, w) \propto p^{(t)}(\alpha|d)p^{(t)}(\beta)p^{(t)}(w|\alpha, \beta) \qquad (1)$$

In the M-step, we re-estimate the model parameters which maximize the expectation of the log-likelihood

$$p^{(t+1)}(\alpha|d) \propto \sum_{w \in \mathcal{V}} \sum_{\beta \in B} n(w, d)p^{(t)}(\alpha, \beta|d, w)$$

$$p^{(t+1)}(\beta) \propto \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{V}} \sum_{\alpha \in A} n(w, d)p^{(t)}(\alpha, \beta|d, w)$$

$$p^{(t+1)}(w|\alpha, \beta) \propto \sum_{d \in \mathcal{D}} n(w, d)p^{(t)}(\alpha, \beta|d, w)$$

$$p(d) \propto \sum_{w \in \mathcal{V}} n(w, d)$$

Once the model parameters are learnt each document $d \in \mathcal{D}$ is assigned to the cluster which maximizes the posterior probability : $\text{cluster}(d) = \text{argmax}_{\alpha \in A}\, p(\alpha|d)$

3. EXPERIMENTS

In order to evaluate our model for the document clustering task, we used standard labeled text classification corpora using the class labels as an objective knowledge reflecting the datasets implicit structure. In the following, we describe the datasets we used as well as the measures we carried out to evaluate the clustering performance.

Data sets and Evaluation Criteria. We conducted our experiments on the Reuters, 20Newsgroups and WebKB data sets. General preprocessing steps for these three collections consist in converting all words to lowercase, mapping digits to a single *digit* token, removing non alpha-numeric characters and removing words occurring in less than 3 documents or appearing in a stop list. After preprocessing, the Reuters collection contains 4335 news articles divided into 7 classes. The 20Newsgroups collection contains 16010 Usenet messages divided into 5 classes. The WebKB collection contains 4196 web pages divided into 4 classes. In the following, the reported performance are averaged over 10 random subset splits of each initial collection while preserving the proportions between different classes in each subset.

Table 1: Best average precision and the corresponding average NMI on the Reuters, 20Newsgroups and WebKB datasets.

Measure	Collection	MM	PLSA	Ext-PLSA
Average Prec.	Reuters	0.61	0.64	**0.71**
	20Newsgroups	0.62	0.71	**0.77**
	WebKB	0.48	0.61	**0.68**
NMI	Reuters	0.27	0.38	**0.42**
	20Newsgroups	0.36	0.49	**0.54**
	WebKB	0.11	0.28	**0.36**

In order to compare the performance of the algorithms, we used the micro-averaged precision and recall [6] as well as the Normalized Mutual Information [7].

Results. In table 1, we present a comparison of MM, PLSA and Ext-PLSA. We compared clustering performance on the 10 subsets of the three data collections. The Ext-PLSA model is significantly better than the two models according to a Wilcoxon rank sum test used at a p-value threshold of 0.01.

4. CONCLUSION

In this article, we proposed an extended version of the PLSA model which separates the latent topics and the document clusters, allowing the model to simultaneously cluster documents and terms. Experiments conducted on the Reuters, 20Newsgroups and WebKB datasets have shown that the proposed Ext-PLSA performs significantly better than the MM model and the original PLSA model.

5. REFERENCES

[1] M.-R. Amini and P. Gallinari. The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization. In *Proceedings of ACM SIGIR*, pages 105-112, 2002.

[2] D.-R. Cutting, J.-O. Pedersen, D. Karger and John W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In *Proceedings of ACM SIGIR*, pages 318-329, 1992.

[3] A.-P. Dempster, N.-M. Laird and D.-B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1:1-38, 1977.

[4] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of ACM SIGIR*, 50-57, 1999.

[5] K. Kummamuru, R. Lotlikar, AS. Roy, K. Signal and R. Krishnapuram. A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results. In *Proceedings of World Wide Web*, 658-665, 2004.

[6] N. Slonim and N. Tishby. Unsupervised Document Classification using Sequential Information Maximization. In *Proceedings of ACM SIGIR*, 129-136, 2002.

[7] A. Strehl and J. Ghosh. Cluster Ensembles, A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal Machine Learning Research* 3:583-617, 2002