Improving Document Clustering in a Learned Concept Space

Jean-François Pessiot, Young-Min Kim, Massih R. Amini, Patrick Gallinari
Universit Pierre et Marie Curie (Paris 6)
4, place de Jussieu
75252, Paris Cedex 05

Abstract

Most document clustering algorithms operate in a high dimensional bag-of-words space. The inherent presence of noise in such representation obviously degrades the performance of most of these approaches. In this paper we investigate an unsupervised dimensionality reduction technique for document clustering. This technique is based upon the assumption that terms co-occurring in the same context with the same frequencies are semantically related. On the basis of this assumption we first find term clusters using a classification version of the `EM` algorithm. Documents are then represented in the space of these term clusters and a multinomial mixture model (`MM`) is used to build document clusters. We empirically show on four document collections, `Reuters-21578`, `Reuters RCV2-French`, `20Newsgroups` and `WebKB`, that this new text representation noticeably increases the performance of the `MM` model. By relating the proposed approach to the Probabilistic Latent Semantic Analysis (`PLSA`) model we further propose an extension of the latter in which an extra latent variable allows the model to co-cluster documents and terms simultaneously. We show on these four datasets that the proposed extended version of the `PLSA` model produces statistically significant improvements with respect to two clustering measures over all variants of the original `PLSA` and the `MM` models.

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 [6]. They can also be useful in distributed search [24] or in extractive text summarization where topically related documents facilitate the search of relevant text-spans to be extracted for the summary [1, 15].

Most existing text clustering approaches rely on the bag-of-words representation [6]. Using words as features, each document is represented in a high dimensional vocabulary space as a vector of (normalized) word frequency counts [20]. The *sparsity* (most documents contain less than 5% of the vocabulary terms [10]) and the *noise* (text data extracted from internet pages, chat logs or e-mails may often contain spelling errors and abbreviations [14]) in this representation indeed affect the final clustering performance.

These difficulties have motivated the development of dimensionality reduction techniques as a pre-processing step to determine a more compact and relevant document representation. Examples of such approaches are the singular value decomposition used in the Latent Semantic Indexing (`LSI`) [7] or other matrix factorization approaches like random projections [3] or non-negative matrix factorization [16, 25]. Matrix factorization approaches have successfully been applied to the clustering of text data, including web access log pages [18]. Other approaches for dimensionality reduction of text data include probabilistic models and co-clustering approaches. The former include the popular Probabilistic Latent Semantic Analysis [13] and Latent Dirichlet Allocation [4]. Those two models have successfully been used for the task of topic discovery [13, 12]. Co-clustering approaches aim at simultaneously clustering documents and words. Recent advances include formulations in the bipartite graph framework [9] and in the matrix factorization framework [2]. Non informative words can also be removed by simple heuristics based for example, on their document frequencies [21]. These unsupervised approaches though less efficient than supervised feature selection methods allow to find less noisy representation space than the initial bag-of-words space.

In this paper we propose two methods for unsupervised dimensionality reduction in the context of document clustering. Both approaches rely on the idea of replacing the usual `BOW` document representation by a condensed semantic representation in a concept space. Concepts (also referred as *word topics* in the following) are identified by a probabilistic model. The first technique makes the hypothesis that words occurring with the same frequencies in the same documents are semantically related. Based on this assumption words are first partitioned into word-topics. Each document collection is then represented in the bag-of-concepts space by a vector where each feature corresponds to a word-topic representing the number of occurrences of words from that word-topic in the document. Documents are then clustered in this concept space. We empirically show on `Reuters-21578`, `Reuters RCV2-French`, `20Newsgroups` and `WebKB` document collections that the clustering performance of a multinomial mixture (`MM`) model in the concept space is significantly better than its clustering performance on the original vocabulary space. The second approach is an extension of the `PLSA` model (denoted by `Ext-PLSA` in the following) and it jointly performs topic identification and document clustering. In this case, we use two latent variables which respectively identify the word topics and the document clusters. This allows using `PLSA` for both multiple topic identification and clustering in a principled framework. Empirical results indicate that this extended version of the aspect model produces statistically significant improve-

ments with respect to two clustering measures compared to the original PLSA and the MM models operating in the original and the induced concept spaces.

The remainder of this paper is organized as follows. Section 2 presents different probabilistic models for dimensionality reduction and document clustering we propose and use in our experiments. We present experimental results in section 3 and summarize our contribution in section 4.

2 Models

This section presents four probabilistic frameworks for modeling the nature of documents in an unsupervised setting. Each framework defines a generative model for documents and encompasses different probabilistic assumptions for their generation. The first two models are state of the art and correspond to the multinomial mixture model (section 2.2) and the Probabilistic Latent Semantic Analysis (section 2.3) [13]. We bridge the gap between these models by our first contribution which represents documents in a reduced word-topic space (section 2.4) and in section 2.5, we present our extension of the PLSA model.

2.1 Notations

We assume that the collection consists of a set of n unlabeled documents $\mathcal{D} = \{d_i\}_{i\in\{1,...,n\}}$ containing words from a vocabulary $\mathcal{V} = \{w_j\}_{j\in\{1,...,m\}}$. Each word $w \in \mathcal{V}$ is represented by a vector of its occurrences in documents of the collection $\vec{w} =< n(d_i, w) >_{i\in\{1,...,n\}}$, and each document $d \in \mathcal{D}$ is represented by the vector of word frequencies $\vec{d} =< n(d, w_j) >_{j\in\{1,...,m\}}$. We further assume that the collection contains K latent document clusters $A = \{\alpha_1, ..., \alpha_K\}$ and L latent word topics (or concepts) $B = \{\beta_1, .., \beta_L\}$.

2.2 Multinomial Mixture

In this framework each document is assumed to be generated by a mixture model:

$$p(\vec{d} \mid \Theta) = \sum_{k=1}^{K} p(\alpha_k \mid \Theta) p(\vec{d} \mid \alpha_k, \Theta) \quad (1)$$

We further assume that there is an univocal correspondence between each document cluster $\alpha \in A$ and each mixture component. A document d is therefore generated by first selecting a mixture component according to the prior cluster probabilities $p(\alpha_k \mid \Theta)$ and then generating the document from the selected mixture component, with probability $p(\vec{d} \mid \alpha_k, \Theta)$ (Figure 1 (a)).

We further assume that the word occurrences within each document are independent. This assumption corresponds to the Naive Bayes model in the supervised case and is referred to the Mixture of Multinomials in unsupervised learning [19]. In this case, the probability of a document d given cluster α_k can be expressed as

$$p(\vec{d} \mid \alpha_k, \Theta) \propto \prod_{j=1}^{|\mathcal{V}|} p_{jk}^{n(d,w_j)} \tag{2}$$

Where, p_{jk} is the probability of generating word w_j in document cluster α_k. The complete set of model parameters consists of multinomial parameters for the cluster priors $p(\alpha_k)$ and word generation probabilities p_{jk}:

$$\Lambda = \{p(\alpha_k) : \alpha_k \in A; p_{jk} : w_j \in \mathcal{V}, \alpha_k \in A\}$$

We estimate the parameters Λ by maximizing the complete data log-likelihood using the Expectation Maximization algorithm (`EM`) [8]. The algorithm can be sketched out as follows. The initial set of parameters $\Lambda^{(0)}$ is obtained randomly. We then iteratively estimate the probability that each mixture component $\alpha_k \in A$ generates each document $d \in \mathcal{D}$ using the current parameters $\Lambda^{(t)}$ (`E-step`) and update the Multinomial mixture parameters $\Lambda^{(t+1)}$ by maximizing the complete data log-likelihood (`M-step`).

2.3 Probabilistic Latent Semantic Analysis

The `PLSA` model introduced by [13] 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 (figure 1 (b)):

- Choose a document d with probability $p(d)$,
- Choose a topic β according to $p(\beta \mid d)$,
- Generate a word w with probability $p(w \mid \beta)$

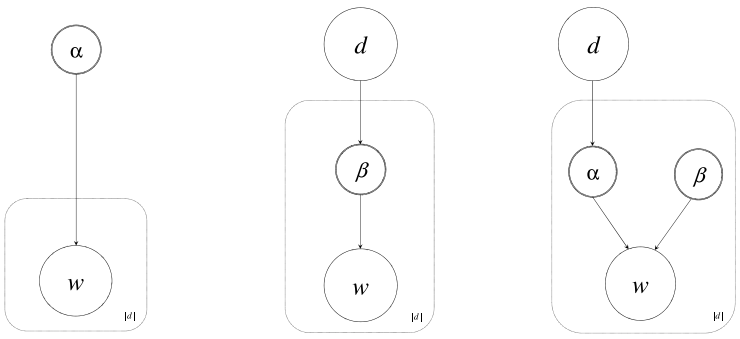
The generation of a word w within a document d can then be translated by the following joint probability model:

$$p(w, d) = p(d) \sum_{\beta \in B} p(\beta \mid d) p(w \mid \beta) \tag{3}$$

This model overcomes the simplifying assumption of the multinomial mixture model where all words are supposed to be generated from the same topic (equation 2). In `PLSA`, a topic is drawn independently from $p(\beta \mid d)$ each time that a new word is generated in a document. This provides a much more natural way to handle unusual words or multi-topicality.

The model parameters in this case are

$$\Delta = \{p(d), p(\beta \mid d), p(w \mid \beta) : d \in \mathcal{D}, \beta \in B, w \in \mathcal{V}\}$$



(a) (b) (c)

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

and they are obtained by maximizing the (log-)likelihood,

$$\mathcal{L}_1 = \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{V}} n(d, w) \log p(d, w) \tag{4}$$

using the Expectation Maximization (EM) algorithm [8]. The iterative update rule consists of first computing posterior probabilities of latent variables using the current parameters $\Delta^{(t)}$ (`E-step`), and then to update $\Delta^{(t+1)}$ by maximizing the log-likelihood function (4) in the `M-step`.

Once the model parameters have been estimated, each vocabulary word $w \in \mathcal{V}$ can be assigned to a word topic $\beta \in B$ following the Bayes rule:

$$\text{cluster}(w) = \underset{\beta \in B}{\text{argmax}}\, p(\beta \mid w) \equiv \underset{\beta \in B}{\text{argmax}}\, p(w|\beta) \sum_{d} p(d)p(\beta|d) \tag{5}$$

In `PLSA`, latent word topics and document clusters are identical [11]. The probability of sampling from topic β given document d, $p(\beta \mid d)$, is interpreted as the posterior cluster probability for document d. Clustering is performed using Bayes decision rule: When document clustering is performed with the `PLSA` model, the latent word topics play the role of the document clusters [11]. In this case, the probability of observing the topic β given the document d, $p(\beta \mid d)$ is interpreted as the posterior probability of the document d and clustering is performed from the following Bayes decision rule

$$\text{cluster}(d) = \underset{\beta \in B}{\text{argmax}}\, p(\beta \mid d)$$

2.4 Concept learning

In this section, we present our first dimensionality reduction model which finds latent word topics, or *concepts*, by directly grouping words of the vocabulary. The inherent assumption which leads to this partitioning is that *words occurring with the same frequencies in the same documents are semantically related.* This assumption takes into account the presence of synonym terms in the context of a discourse. Once concepts are found, documents are then represented in the deduced concept space where clustering is performed.

Formally, we assume that each word $w \in \mathcal{V}$ is generated by a mixture density:

$$p(\vec{w}|\Theta) = \sum_{l=1}^{L} \pi_l p(\vec{w}|\beta_l, \theta_l) \tag{6}$$

Where, as previously noted, L is the number of latent topics to be found and Θ is the set of all model parameters (mixing rates and density parameters). We further suppose that each word belongs to exactly one word topic. This assumption can be formalized using a topic indicator vector $C_j = \{C_{hj}\}_h$ for each word $w_j \in \mathcal{V}$ defined as:

$$\forall w_j \in \mathcal{V}, \exists \beta_l \in B; w_j \in \beta_l \Leftrightarrow C_{lj} = 1 \text{ and } \forall h \neq l, C_{hj} = 0$$

Word clustering is then performed by searching the parameters Θ maximizing the complete data log-likelihood:

$$\mathcal{L}_2(C, \Theta) = \sum_{w_j \in \mathcal{V}} \sum_{l=1}^{L} C_{lj} \log p(\vec{w}_j, \beta_l, \Theta) \tag{7}$$

Here, the cluster indicator vectors C are estimated together with model parameters Θ. In our experiments, we assumed that words are independently generated by the mixture density (6) where each mixture component $p(\vec{w}|\beta)$ obeys a Naive Bayes model. Hence by denoting, for each document $d_i \in \mathcal{D}$ and cluster $\beta_l \in B$, the probability of generation of d_i in β_l as q_{il}; the complete set of model parameters consists of multinomial parameters of mixing components $\pi_l = p(\beta_l)$ and document generation probabilities q_{il}:

$$\Theta = \{p(\beta_l) : \beta_l \in B; q_{il} : d_i \in \mathcal{D}, \beta_l \in B\}$$

From this assumption, the probability of a word $w \in \mathcal{V}$ given a latent topic β_l can be expressed as

$$p(\vec{w} \mid \beta_l) \propto \prod_{i=1}^{n} q_{il}^{n(d_i, w)} \tag{8}$$

We used a classification version of the `EM` algorithm proposed by [5] to estimate the model parameters Θ. This algorithm is depicted on the right (Algorithm 1) and it may be sketched out as follows. The initial set of model

Algorithm 1: The `CEM` algorithm for word clustering

Input :

- A partition $P^{(0)}$ is randomly initialized and the conditional probabilities $p(w \mid \beta = l, \theta_l^{(0)})$ are estimated on the corresponding clusters.
- $t \leftarrow 0$

repeat

- `E-step`: Estimate the posterior probabilities that each word w_j in the vocabulary belongs to each of the partitions $P_l^{(t)}$:

$$\forall w_j \in \mathcal{V}, \forall l \in \{1, ..., L\}, \mathbb{E}[C_{lj}^{(t)} \mid w_j; P^{(t)}, \Theta^{(t)}] = \frac{\pi_l^{(t)} p(w_j \mid \beta_l)}{p(w, \Theta^{(t)})}$$

- `C-step`: Assign to each word $w_j \in \mathcal{V}$ the cluster $P_k^{(t+1)}$ with the highest posterior probability $\mathbb{E}[C \mid w]$. Let $P^{(t+1)}$ be the new partition.
- `M-step`: Update the parameter estimates $\Theta^{(t+1)}$ by maximizing equation (7).
- $t \leftarrow t + 1$

until *convergence of* $\mathcal{L}_2$;

Output : Word Clusters, $P^{(t)}$

parameters $\Theta^{(0)}$ are estimated on the basis of randomly obtained word partitions $P^{(0)}$. Three steps are then repeated until the convergence of the complete data log-likelihood (7). In the `E-step`, a conditional expectation of each word $w_j \in \mathcal{V}$ with respect to each of its associated cluster indicators $C_{lj}; l \in \{1, ..., L\}$ is computed. As each C_{lj} is a binary random variable, this conditional expectation is equal to the posterior probability of the word w_j within the latent topic β_l which can be estimated by the Bayes rule and the current model parameters $\Theta^{(t)}$. In the `C-step`, words are assigned to each cluster with the highest posterior probability estimated previously and new model parameters $\Theta^{(t+1)}$ are estimated by maximizing the complete data log-likelihood. Lagrange multipliers are used to enforce $\sum_l \pi_l = 1$ and $\forall l, \sum_{i=1}^{n} q_{il} = 1$ constraints. In the `M-step`, the model parameters are updated as:

$$\pi_l^{(t+1)} = \frac{\sum_{j=1}^{|\mathcal{V}|} C_{lj}^{(t+1)}}{|\mathcal{V}|}, q_{il}^{(t+1)} = \frac{\sum_{j=1}^{|\mathcal{V}|} C_{lj}^{(t+1)} \times n(d_i, w_j)}{\sum_{j=1}^{|\mathcal{V}|} \sum_{i=1}^{n} C_{lj}^{(t+1)} \times n(d_i, w_j)}$$

Finally, we obtain word clusters (or topics) which serve as a new feature space representation for each document in the collection, where each feature corresponds to a word cluster and represents the total number of occurrences of its words appearing in the document.

From Eq. 8, it is clear that words occurring in the same documents with similar frequencies will have similar $p(\vec{w} \mid \beta_l)$ and therefore, using the Bayes decision rule, belong to the same word topic. Note also that combining Eq. (5) and the `M-step` equation for $p(\beta|d)$ shows that `PLSA` also tends to assign words with similar co-occurrence patterns to the same topic/cluster.

2.5 An Extended version of PLSA

The problem of the aforementioned clustering in the concept space is that the clustering process passes through two mixture models and it may be altered from successive generative assumptions. In this section we propose an extension of the `PLSA` model in which document and word clustering are performed simultaneously in a single aspect model. In this case the underlying generation process is as follows:

- Pick a document d with probability $p(d)$,
- Choose a document cluster 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: document clusters and word topics.

In this case the generation of a word w within a document d can be expressed by the following probability:

$$p(d, w) = \sum_{\alpha \in A} \sum_{\beta \in B} p(d)p(\alpha|d)p(\beta)p(w|\alpha, \beta) \tag{9}$$

Following the maximum likelihood principle, the model parameters are hence,

$$\Phi = \{p(d), p(\alpha \,|\, d), p(\beta), p(w \,|\, \alpha, \beta) : (d, w, \alpha, \beta) \in \mathcal{D} \times \mathcal{V} \times A \times B\}$$

and they are estimated by maximizing the log-likelihood function (4) using an `EM`-type algorithm [8].

In the `E-step`, conditional probabilities of latent variables given documents and words are estimated from the current model parameters:

$$p^{(t+1)}(\alpha, \beta|d, w) = \frac{p^{(t)}(\alpha|d)p^{(t)}(\beta)p^{(t)}(w|\alpha, \beta)}{\sum_{\alpha' \in A} \sum_{\beta' \in B} p^{(t)}(\alpha'|d)p^{(t)}(\beta')p^{(t)}(w|\alpha', \beta')} \tag{10}$$

Algorithm 2: Extended version of PLSA

Input :

- A document collection $\mathcal{D}$,
- Random initial model parameters $\Phi^{(0)}$.
- $t \leftarrow 0$

repeat

- Re-estimate model parameters using multiplicative update rules (10–13)
- $t \leftarrow t + 1$

until *convergence of* $\mathcal{L}_1$ *(eq. 4)* ;

Output : A generative classifier with parameters $\Phi^{(t)}$

In the `M-step`, new model parameters maximizing the expectation of the log-likelihood (4) are estimated:

$$p^{(t+1)}(\alpha|d) = \frac{\sum_{w\in\mathcal{V}}\sum_{\beta\in B} n(d,w)p^{(t)}(\alpha,\beta|d,w)}{\sum_{w\in\mathcal{V}}\sum_{\alpha'\in A}\sum_{\beta\in B} n(d,w)p^{(t)}(\alpha',\beta|d,w)} \tag{11}$$

$$p^{(t+1)}(\beta) = \frac{\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{V}}\sum_{\alpha\in A} n(d,w)p^{(t)}(\alpha,\beta|d,w)}{\sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{V}}\sum_{\alpha\in A}\sum_{\beta'\in B} n(d,w)p^{(t)}(\alpha,\beta'|d,w)} \tag{12}$$

$$p^{(t+1)}(w|\alpha,\beta) = \frac{\sum_{d\in\mathcal{D}} n(d,w)p^{(t)}(\alpha,\beta|d,w)}{\sum_{d\in\mathcal{D}}\sum_{w'\in\mathcal{V}} n(d,w')p^{(t)}(\alpha,\beta|d,w')} \tag{13}$$

This algorithm (Algorithm 2) is also an `EM`-like algorithm, and the iterative use of equations 10, 11, 12 and 13 corresponds to alternating the `E-step` and `M-step`. Convergence is therefore guaranteed to a local maximum of the likelihood.

Notice that the maximum likelihood estimate of parameter $p(d)$ has an exact analytical expression (and does not need iterative estimation):

$$p(d) = \frac{\sum_{w\in\mathcal{V}} n(d,w)}{\sum_{d'\in\mathcal{D}}\sum_{w\in\mathcal{V}} n(d',w)}$$

Once model parameters are obtained, each document $d \in \mathcal{D}$ is assigned to a cluster using the Bayes decision rule:

$$\text{cluster}(d) = \underset{\alpha \in A}{\text{argmax}}\, p(\alpha|d)$$

3 Experimental Setup

We conducted a number of experiments aimed at evaluating how the representation in the concept space can help to improve the clustering performance. The multinomial model was used both in its basic version (section 2.2) and for the concept learning method (section 2.4). `PLSA` was used both for direct document clustering and for finding concept spaces (word topics). In the following, these models are denoted by `MM` and `PLSA` when they perform clustering in the original vocabulary space. When clustering is performed in the concept space, we use the following notations. A first letter identifying the word topic algorithm (respectively `C` or `P` for `CEM` and `PLSA`) followed by the acronym of the clustering approach. Thus `C-MM` denotes document clustering with `MM` on the concept space induced by `CEM`, `P-MM` denotes document clustering with `MM` on the concept space induced by `PLSA`, `C-PLSA` is the document clustering with `PLSA` on the concept space of `CEM` and finally `P-PLSA` denotes two successive applications of `PLSA`, first for word topics identification and then for document clustering onto the induced concept space. In the light of these results we compare in the second part of our experiments the performance of the extended `PLSA` (`Ext-PLSA`) against all the previous clustering models as well as the `Kmeans` and `LDA` algorithms [4] and a `NMF` model obtained by minimizing the Frobenius norm. As KL-minimal `NMF` is essentially equivalent to `PLSA` [11], we do not include it in our comparison. Latent Semantic Indexing [7] is another popular matrix factorization method for the analysis of text data. But as NMF outperforms LSI for document clustering [25], we do not include LSI in our comparison. In our experiments, we used standard labeled text classification corpora with the class labels representing an objective knowledge reflecting the datasets implicit structure. The following sections describe the corpora we used in our experiments as well as the accuracy measures we used to evaluate the performance of the proposed models.

3.1 Data sets

We conducted our experiments on `Reuters-21578`, `Reuters RCV2-French`, `20Newsgroups` and `WebKB` datasets.

The `Reuters-21578`[1] collection contains Reuters news articles from 1987. We selected documents in the collection that are assigned to at least one class. Each document in the corpus can have multiple labels, but in practice more than 80% of articles are associated to a single topic. In addition, for multiply-labeled documents, only the first class from the <TOPIC> field was retained. We kept

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

the documents associated with the 7 most frequent classes, which resulted in 4335 documents, each with a unique label.

The `Reuters RCV2-French`[2] (denoted by `RCV2-French` in the following) is the French part of the multilingual Reuters corpus which contains over 487,000 Reuters News stories in thirteen languages. We focused on 6 relatively populous classes and for each class we sampled up to 5000 documents from RCV2. Documents belonging to more than one of our 6 classes were assigned the label of their smallest class ending to a monolingual case.

The `20Newsgroups`[3] dataset is a collection of newsgroup postings from 20 different Usenet discussion forums. We deleted cross-posted and duplicate messages, and we regrouped the documents into one of the 5 following classes (alt., comp., sci., rec. talk.) which resulted in 16010 documents. We ignored in this case file headers and subject lines.

The `WebKB`[4] contains 8145 web pages gathered from computer science departments. The collection includes web pages of four departments divided into seven categories. In this paper, we used the 4 most populous ones all together containing 4196 pages. Our specific pre-processing for this collection consists in filtering the text by removing html tags.

Table 1: Class proportions in `Reuters-21578`, `Reuters RCV2-French`, `20Newsgroups` and `WebKB` datasets.

`Reuters-21578`	
Class	perc. %
earn	46.2
acq	24.9
money	8.4
crude	7
grain	6.5
trade	4.8
interest	2.4

`RCV2-French`	
Class	perc. %
`C15`	16.7
`CCAT`	16.7
`E21`	16.7
`ECAT`	16.7
`GCAT`	16.7
`M11`	16.5

`20Newsgroups`	
Class	perc. %
comp.	30.5
sci.	24.7
rec.	20.3
talk.	19.5
alt.	5.0

`WebKB`	
Class	perc. %
student	39.1
faculty	26.8
course	22.1
project	12.0

General preprocessing steps for these four collections consist in converting all words to lowercase, mapping digits to a single *digit* token and suppressing non alpha-numeric characters. We also used a stop-list to remove very frequent words and also filtered words occurring in less than 3 documents. These preprocessing lead to an initial vocabulary size of 6990, 34272, 38630 and 11170 words for respectively the `Reuters-21578`, `RCV2-French`, `20Newsgroups` and `WebKB` data sets. Table 1 summarizes the characteristics of these four collections. Classes are shown in the decreasing order of their sizes in the respective collections.

We finally performed experiments over 10 random cross-validation splits of each initial collection while preserving the proportions between different classes

[2] http://trec.nist.gov/data/reuters/reuters.html
[3] http://kdd.ics.uci.edu/databases/20newsgroups/
[4] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

in each subset. This setting avoids bias caused by the inherent structure of each collection. It helps reduce the effects of the random initializations made in the generative models we considered. The reported performance are averaged over these 10 randomly selected subsets.

3.2 Evaluation Criteria

In order to compare the performance of the algorithms, we used the micro-averaged precision and recall [22] as well as the Normalized Mutual Information [23]. To estimate these measures we first assigned documents in the clusters to the majority class present in that cluster. From these assignments we hence estimate the performance measures as follows.

Micro-averaged Precision and Recall: For each class l in the collection, we first estimate its correctly and incorrectly assigned number of documents, respectively denoted by $\kappa(l)$ and $\vartheta(l)$; as well as $\gamma(l)$, the number of documents incorrectly not assigned to l. We then compute the precision and recall of the class l from the following:

$$\text{Prec}(l) = \frac{\kappa(l)}{\kappa(l) + \vartheta(l)}, \quad \text{Rec}(l) = \frac{\kappa(l)}{\kappa(l) + \gamma(l)}.$$

The micro-averaged precision and recall are thus defined by:

$$\text{Micro-averaged Precision} = \frac{\sum_l \kappa(l)}{\sum_l \kappa(l) + \vartheta(l)}$$

And,

$$\text{Micro-averaged Recall} = \frac{\sum_l \kappa(l)}{\sum_l \kappa(l) + \gamma(l)}$$

By noticing that $\sum_l \kappa(l) + \gamma(l)$ and $\sum_l \kappa(l) + \vartheta(l)$ are both equal to the total number of documents in a collection, we have in this case a perfect equality between the micro-averaged precision and recall. In our experiments we will refer to these terms by *Average precision*.

Normalized Mutual Information: The Normalized Mutual Information (NMI) is a measure which estimates the quality of a clustering with respect to the true classes of a dataset [23]. It corresponds to the normalized mutual information between the cluster assignments of instances and their underlying class labels and is given from the following expression:

$$\text{NMI} = \frac{\sum_{h=1}^{c} \sum_{l=1}^{c} n_{h,l} \log\left(\frac{n \times n_{h,l}}{n_h n_l}\right)}{\sqrt{\left(\sum_{h=1}^{c} n_h \log \frac{n_h}{n}\right)\left(\sum_{l=1}^{c} n_l \log \frac{n_l}{n}\right)}}$$

where n is the total number of documents, c the number of classes (or clusters), n_h and n_l are the number of documents in respectively cluster h and class l, and $n_{h,l}$ is the common number of documents in both cluster h and class l. As previously the range of NMI values is within $[0, 1]$ and it tends to 1 for increasingly better cluster qualities.

3.3 Experimental Results

We begin our experiments by first examining the effect of changing the document representation from the initial vocabulary space into the new learned concept space (induced by the CEM algorithm). We analyze this effect on the final clustering results of the multinomial mixture model. The two clustering schemes involved here are the multinomial mixture model operating in the vocabulary space (MM) and in the concept space (C-MM). We recall that documents are represented in the latter space by a vector where each component corresponds to a word topic and represents the total number of words from this topic occurring in the document.

The number of document clusters is fixed to the original number of class labels of each dataset, we have then varied the number of word topics from 10 to 100. Table 3, shows the precision, recall and Average precision performance over the four Reuters-21578, RCV2-French, 20Newsgroups and WebKB databases. In these experiments, the number of word topics on each collection was fixed to the number which provided the best clustering results of the MM model in the concept space. This corresponds to $|B| = 10$ on the Reuters-21578, RCV2-French and 20Newsgroups datasets and $|B| = 20$ on the WebKB collection. The table 2 illustrates the ability of CEM to identify word topics on the 20Newsgroups and WebKB datasets.

Table 2: An example of term clusters found with CEM in 20Newsgroups(top) and WebKB (down) data collections.

Cluster i**:** wrong christian words christians truth meaning paul john bible word faith fact reason men **Cluster** j**:** suicide suicides deaths stat selfdefense risks guns statstic accidents homicides nejm
Cluster k**:** course courses homework project projects umd class assignment due assignments exam fall berkeley lecture **Cluster** l**:** computer research systems cs science programming acm sciences engineering software system algorithms program

It should be noticed that the value of $|B|$ influences the generality level of the resulting word topics. When $|B|$ is small, the word topics are constrained to be quite general. A high value will allow more detailed and specialized words topics. Hence the optimal $|B|$ values for Reuters-21578, RCV2-French, 20Newsgroups and WebKB correspond to rather general word topics. There are two explanations for this observed tendancy for fewer word topics. First, if there are lots of specialized word topics, then two documents dealing with the same topics may appear dissimilar in the induced concept space. This phenomenon degrades the clustering quality. Second, a high number of word topics implies a higher number of parameters to be learnt. Hence the learning task is more dif-

ficult, resulting in poor clustering performance. As a consequence, the optimal number of word topics $|B|$ tends to be small in our experiments.

The figures 2 and 3 show the evolution of performance of the MM and PLSA algorithms for a varying number of word topics (induced by the CEM and the PLSA models).

In the tables, 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 0.01 [17]. On these four collections and for about all classes, the precision, recall, and the Average precision of C-MM are significantly better than the MM algorithm. We notice here that in the case where classes are unbalanced (especially on Reuters-21578, 20Newsgroups and WebKB), small classes are absorbed by larger ones. This phenomenon is however attenuated when clustering is performed in the concept space, the class trade in Reuters-21578 becomes even visible. These results suggest that the independence hypothesis of the MM model between document vector components is more likely to be true when documents are represented in a concept space. Indeed, it is more reasonable to assume that the discourse of a document consists of independent word topics, than assuming that vocabulary words occurring in a document be independent from one another.

Figure 2 shows the document clustering performance, on the four datasets, of the MM algorithm operating in the bag-of-words space and in the concept spaces induced by the CEM and PLSA algorithms, for varying numbers of word concepts ($L = |B|$). The figure also depicts the performance of the PLSA model for document clustering. We recall that in this case, latent topics in PLSA correspond to document partitions which are hence fixed to the number of class labels on each data set. In our experiments, the size of the concept space for Reuters-21578 is limited to 70 since we obtained empty partitions for higher values with the CEM algorithm on this collection.

The first observation is that in the original bag-of-words space, PLSA outperforms the MM model over all four datasets. The gap in performance between these two models are the largest on the WebKB collection reaching 17% and 14% for respectively the NMI and average precision measures. These results confirm the effectiveness of the document generation assumption made by the PLSA model and discussed in section 2.3. The MM model becomes however, more competitive in different concept spaces induced by the CEM and the PLSA algorithms (i.e. C-MM and P-MM). The latter two have significantly better results than the MM model performing in the initial bag-of-words space, for concept spaces with various sizes. These observations confirm our findings made previously, that the independence assumption made by the MM model is more accurate on a space where each direction corresponds to a group of similar words. We also notice that both C-MM and P-MM models achieve results comparable to PLSA in terms of NMI and Average Precision. Moreover, the C-MM model performs uniformly better than the P-MM model on the four collections, in both NMI and Average precision.

As the underlying assumption of the MM model is that each dimension axes are

Table 3: Precision, Recall clustering performance of the Naive-Bayes model learnt on the vocabulary space (MM) and on the concept space (C-MM) for the Reuters-21578, RCV2-French, 20Newsgroups and WebKB datasets.

Data set	$\|B\| = 10$	Precision		Recall	
		MM	C-MM	MM	C-MM
Reuters-21578	earn	$0.77^{\downarrow}$	**0.89**	**0.93**	$0.84^{\downarrow}$
	acq	$0.43^{\downarrow}$	**0.65**	$0.60^{\downarrow}$	**0.77**
	money	$0.35^{\downarrow}$	**0.41**	$0.26^{\downarrow}$	**0.48**
	crude	$0.34^{\downarrow}$	**0.46**	$0.57^{\downarrow}$	**0.48**
	grain	$0.43^{\downarrow}$	**0.52**	$0.13^{\downarrow}$	**0.54**
	trade	0	**0.37**	0	**0.22**
	interest	0	0	0	0
	Average	$0.61^{\downarrow}$	**0.70**	$0.61^{\downarrow}$	**0.70**
RCV2-French	C15	$0.38^{\downarrow}$	**0.45**	$0.10^{\downarrow}$	**0.31**
	CCAT	$0.53^{\downarrow}$	**0.77**	$0.79^{\downarrow}$	**0.85**
	E21	$0.63^{\downarrow}$	**0.68**	$0.43^{\downarrow}$	**0.79**
	ECAT	**0.48**	$0.43^{\downarrow}$	$0.18^{\downarrow}$	**0.34**
	GCAT	$0.40^{\downarrow}$	**0.51**	$0.61^{\downarrow}$	**0.71**
	M11	$0.46^{\downarrow}$	**0.65**	**0.76**	$0.57^{\downarrow}$
	Average	$0.48^{\downarrow}$	**0.60**	$0.48^{\downarrow}$	**0.60**
20Newsgroups	comp.	$0.62^{\downarrow}$	**0.76**	**0.92**	$0.87^{\downarrow}$
	sci.	$0.57^{\downarrow}$	**0.78**	$0.37^{\downarrow}$	**0.62**
	rec.	$0.60^{\downarrow}$	**0.67**	$0.72^{\downarrow}$	**0.89**
	talk.	$0.67^{\downarrow}$	**0.84**	$0.50^{\downarrow}$	**0.78**
	alt.	0	0	0	0
	Average	$0.62^{\downarrow}$	**0.75**	$0.62^{\downarrow}$	**0.75**
WebKB	student	$0.47^{\downarrow}$	**0.72**	**0.83**	$0.77^{\downarrow}$
	faculty	$0.41^{\downarrow}$	**0.55**	$0.21^{\downarrow}$	**0.64**
	course	$0.58^{\downarrow}$	**0.86**	$0.36^{\downarrow}$	**0.77**
	project	$0.37^{\downarrow}$	**0.45**	$0.10^{\downarrow}$	**0.29**
	Average	$0.48^{\downarrow}$	**0.68**	$0.48^{\downarrow}$	**0.68**

mutually independent, from the clustering results of C-MM and P-MM it becomes apparent that the CEM algorithm groups dependent terms more efficiently than PLSA. These results suggest that the simple assumption of CEM (section 2.4) is in better agreement with the assumption used by both PLSA and CEM to assign words to topics which, we recall, is: *words co-occurring with the same frequency in the same documents are similar*.

The performance curves of the PLSA model operating in the original bag-of-words space and in the concept space induced by the CEM algorithm (C-PLSA) as well as its extension Ext-PLSA introduced in section 2.5 are shown in Figure 3.

On the Reuters-21578 and 20Newsgroups datasets, Ext-PLSA outperforms PLSA on average precision for any number of concepts. The best performances of Ext-PLSA are obtained for 25 concepts on Reuters-21578 and 20 concepts on 20Newsgroups, and correspond to improvements of respectively 6% and 8% in average precision over PLSA. The NMI curves of Ext-PLSA on Reuters-21578, RCV2-French, 20Newsgroups and WebKB are very similar to the corresponding average precision curves. On the WebKB collection, the performance of the Ext-PLSA model is better than that of the PLSA model on both measures (except when the number of concepts is equal to 10).

We further notice that C-PLSA shows performances similar to Ext-PLSA on Reuters-21578 and WebKB. On the 20Newsgroups dataset, C-PLSA is consistently outperformed by Ext-PLSA for any number of concepts. On the other hand, the Ext-PLSA model captures the cluster structure of the document collection and the topics or concepts. The joint clustering of Ext-PLSA reduces the bias that might result from the two successive clusterings performed in the first model described in section 2.4.

Finally we notice that there is a big gap between the performance of PLSA and NMF using the Frobenius norm especially on RCV2-French, 20Newsgroups and WebKB where initial dimension sizes are higher than the one of the Reuters-21578 collection. This difference might be due to the use of the quadratic norm which is not relevant in high dimensional spaces. This is consistent with the results of the Kmeans algorithm in which we used the Euclidean norm.

In table 5, we present a comparison of all the different algorithms operating in the initial bag-of-words space and in the different concept spaces. The number of concepts was fixed to the one for which each respective model obtained its best results in terms of Average precision. We compared clustering performance on the 10 subsets of the four data collections. The best Ext-PLSA model here is significantly better than almost all other models except the C-MM model which behaves nearly the same. Note also that the clustering performance of MM and PLSA in the concept space induced by the CEM algorithm is considerably increased. This is especially true on the NMI measure, where, on the WebKB corpus, the performance of C-MM is three times higher than for MM.

Table 4 shows the complexity and the average execution time of different algorithms on 3.16GHz Intel Core 2 Duo processor with 4G RAM. Comparatively, the proposed Ext-PLSA approach takes the same time in execution than LDA and NMF but it is more efficient than the two latter on all datasets.

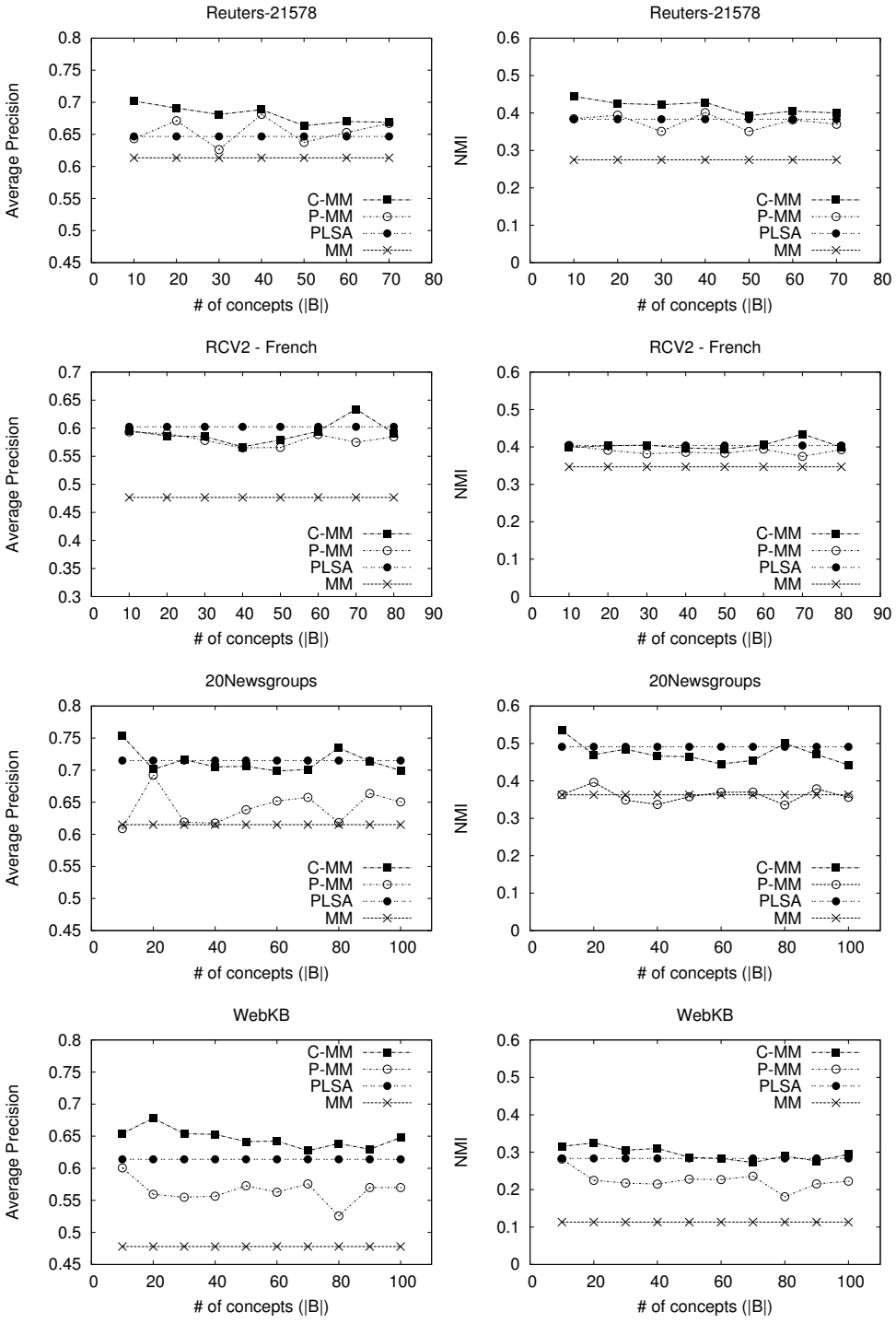


Figure 2: Average Precision (left) and NMI (right) of the clustering algorithms in the bag-of-words space and in concept spaces induced by the `CEM` and `PLSA` algorithms.

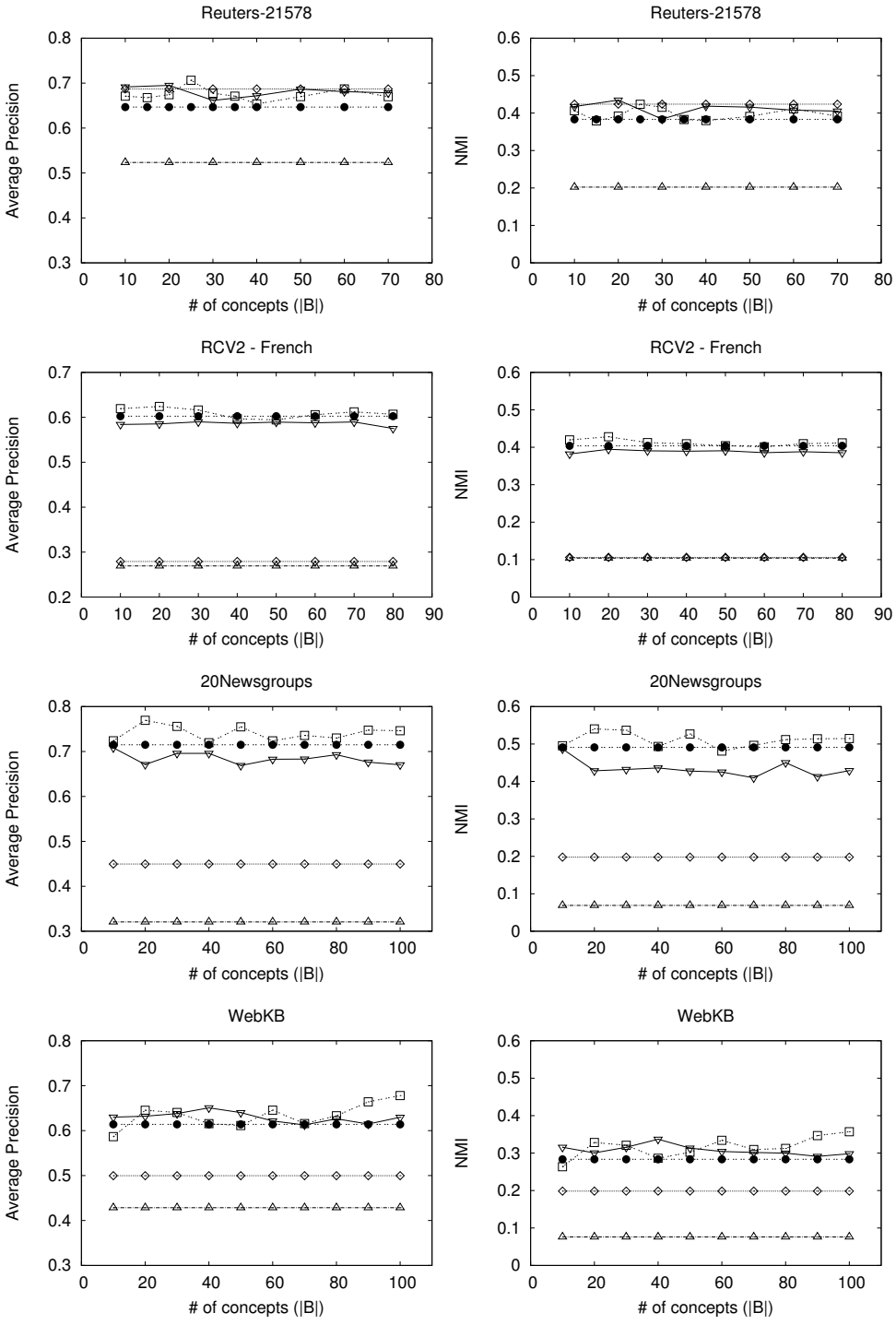


Figure 3: Average precision (left) and NMI (right) performance of the `PLSA` (•) and its proposed extension version, `Ext_PLSA` (□) compared to the performance of `Kmeans` (△), `NMF` (◇) and `C-PLSA` (▽).

Table 4: Computational and complexity comparisons on different datasets. M represents the number of (term, document) pairs observed in the corpus. Recall that n is the total number of documents, m the vocabulary size, K the number of document clusters and L the number of topics. For the execution time, the latter is fixed to 10.

Algorithm	Complexity	Execution time			
		Reuters	RCV2	News	WebKB
Kmeans	$O(K \times M)$	1 sec	2 sec	1 sec	1 sec
NMF	$O(K \times n \times m)$	40 sec	28 min	5 min	40 sec
MM	$O(K \times n \times m)$	5 sec	10 min	3 min	20 sec
C-MM	$O(L \times m \times n)$	10 sec	20 min	7 min	1 min
P-MM	$O(L \times M)$	10 sec	5 min	2 min	25 sec
PLSA	$O(K \times M)$	5 sec	3 min	1 min	10 sec
Ext-PLSA	$O(K \times M \times L)$	1 min	30 min	11 min	1 min
LDA	$O(K \times n \times m)$	1 min	30 min	15 min	2 min

Table 5: Best average precision and the corresponding average NMI of different clustering algorithms on the Reuters-21578, RCV2-French, 20Newsgroups and WebKB datasets. The symbols $^{\downarrow}$ indicate the cases of algorithms significantly worse than that of the best algorithm which performance is shown in bold.

Algorithm	Reuters-21578		RCV2-French		20Newsgroups		WebKB	
	AP	NMI	AP	NMI	AP	NMI	AP	NMI
Kmeans	$0.52^{\downarrow}$	$0.20^{\downarrow}$	$0.27^{\downarrow}$	$0.10^{\downarrow}$	$0.32^{\downarrow}$	$0.07^{\downarrow}$	$0.43^{\downarrow}$	$0.08^{\downarrow}$
NMF	0.69	0.42	$0.28^{\downarrow}$	$0.11^{\downarrow}$	$0.45^{\downarrow}$	$0.20^{\downarrow}$	$0.50^{\downarrow}$	$0.20^{\downarrow}$
MM	$0.61^{\downarrow}$	$0.27^{\downarrow}$	$0.48^{\downarrow}$	$0.35^{\downarrow}$	$0.62^{\downarrow}$	$0.36^{\downarrow}$	$0.48^{\downarrow}$	$0.11^{\downarrow}$
P-MM	0.68	$0.40^{\downarrow}$	$0.59^{\downarrow}$	$0.40^{\downarrow}$	$0.69^{\downarrow}$	$0.40^{\downarrow}$	$0.60^{\downarrow}$	$0.28^{\downarrow}$
C-MM	0.70	**0.44**	0.60	$0.40^{\downarrow}$	0.75	0.53	0.68	$0.32^{\downarrow}$
PLSA	$0.64^{\downarrow}$	$0.38^{\downarrow}$	0.60	$0.40^{\downarrow}$	$0.71^{\downarrow}$	$0.49^{\downarrow}$	$0.61^{\downarrow}$	$0.28^{\downarrow}$
P-PLSA	0.68	$0.39^{\downarrow}$	0.60	$0.39^{\downarrow}$	$0.69^{\downarrow}$	$0.40^{\downarrow}$	$0.63^{\downarrow}$	$0.28^{\downarrow}$
C-PLSA	0.69	0.42	$0.59^{\downarrow}$	$0.39^{\downarrow}$	$0.71^{\downarrow}$	$0.49^{\downarrow}$	$0.65^{\downarrow}$	$0.34^{\downarrow}$
Ext-PLSA	**0.71**	0.42	**0.62**	**0.43**	**0.77**	**0.54**	**0.68**	**0.36**
LDA	0.69	$0.38^{\downarrow}$	$0.55^{\downarrow}$	$0.32^{\downarrow}$	$0.67^{\downarrow}$	$0.41^{\downarrow}$	$0.59^{\downarrow}$	$0.25^{\downarrow}$

4 Conclusion

In this paper, we have studied the problem of document clustering in a reduced concept space. Our first contribution sought to find this space by partitioning vocabulary words according to the hypothesis that words co-occurring in the same context with the same frequency are topically related. Experiments on four datasets have shown that the performance of a baseline MM model is significantly improved when it operates in the induced concept space. We further proposed an extended version of the PLSA model which learns jointly the word topics

and the document clusters. Experiments conducted on the `Reuters-21578`, `RCV2-French`, `20Newsgroups` and `WebKB` datasets have shown that the proposed `Ext-PLSA` performs significantly better than the original `PLSA` model. Compared to the `C-MM` which partitions documents in the concept space induced by `CEM`, the join clustering of documents and words performed by `Ext-PLSA` presents the advantage to reduce the bias of the successive clustering steps done previously.

ACKNOWLEDGEMENTS

This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence IST-2002-506778. This publication only reflects the authors view.

References

[1] M.-R. Amini and P. Gallinari. The use of unlabeled data to improve supervised learning for text summarization. In *Proceedings of the* 25^{th} *Annual International ACM SIGIR*, pages 105–112, 2002.

[2] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. *J. Mach. Learn. Res.*, 8:1919–1986, 2007.

[3] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the* 7^{th} *ACM SIGKDD*, pages 245–250, 2001.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] G. Celeux and G. Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3):315–332, 1992.

[6] D. R. Cutting, J. O. Pedersen, D. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the* 15^{th} *Annual International ACM SIGIR*, pages 318–329, 1992.

[7] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[8] 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)*, 39(1):1–38, 1977.

[9] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. pages 269–274, 2001.

[10] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1/2):143–175, 2001.

[11] E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. In *Proceedings of the* 28^{th} *Annual International ACM SIGIR*, pages 601–602, 2005.

[12] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

[13] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the* 22^{nd} *Annual International ACM SIGIR*, pages 50–57, 1999.

[14] C. Knoblock, D. Lopresti, S. Roy, and L. V. Subramaniam. Special issue on noisy text analytics. *International Journal on Document Analysis and Recognition*, 11:147–155, 2006.

[15] K. Kummamuru, R. Lotlikar, A. Roy, K. Signal, and R. Krishnapuram. A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *Proceedings of the* 13^{th} *International Conference on World Wide Web*, pages 658–665, 2004.

[16] D. D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[17] E. Lehmann. *Nonparametric Statistical Methods Based on Ranks*. McGraw-Hill, New York, 1975.

[18] S. Oyanagi, K. Kubota, and A. Nakase. Application of matrix clustering to web log analysis and access prediction. In *in: WEBKDD 2001Mining Web Log Data Across All Customers Touch Points, Third International Workshop*, pages 13–21, 2001.

[19] D. Pavlov, R. Balasubramanyan, B. Dom, S. Kapur, and J. Parikh. Document preprocessing for naive bayes classification and clustering with mixture of multinomials. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 829–834, 2004.

[20] K. V. Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

[21] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[22] N. Slonim and N. Tishby. Unsupervised document classification using sequential information maximization. In *Proceedings of the* 25^{th} *Annual International ACM SIGIR*, pages 129–136, 2002.

[23] A. Strehl and J. Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research*, 3:583–617, 2002.

[24] J. Xu and W. Croft. Cluster-based language models for distributed retrieval. In *Proceedings of the* 22^{nd} *Annual International ACM SIGIR*, pages 254–261, 1999.

[25] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the* 26^{th} *annual international ACM SIGIR*, pages 267–273, 2003.