# Une extension du modèle sémantique latent probabiliste pour le partitionnement non-supervisé de documents textuels

Young-Min Kim[1], Jean-François Pessiot[1], Massih R. Amini[2], Patrick Gallinari[1]

[1] Laboratoire d'Informatique de Paris 6
104 Avenue du Président Kennedy 75016 Paris, France

[2] Centre National de Recherche au Canada
283 Bd. Alexandre-Taché Gatineau, QC J8X 3X7, Québéc, Canada

**Résumé** : Dans cet article, nous proposons une extension du modèle sémantique latent probabiliste (PLSA) pour la tâche de partitionnement de documents (clustering). Nous montrons que ce modèle étendu est équivalent à une combinaison linéaire de modèles de factorisation matricielle non-négative au sens de la fonction objective KL-divergence. Nous validons notre modèle sur les trois collections de documents et, montrons empiriquement que notre approche est statistiquement plus performante que le modèle PLSA de base pour la tâche de clustering.

**Mots-clés** : Apprentissage non-supervisé, Partitionnement de documents, Analyse sémantique latent probabiliste

## 1 Introduction

La tâche de *clustering* de documents est un problème important en fouille de données et en apprentissage. Un des méthodes pour cette tâche, PLSA (pour *Probabilistic Latent Semantic Analysis*) (Hofmann, 1999), est à présent l'algorithme standard largement employé pour cette tâche de partitionnement dans la communauté de Recherche d'Information. Dans ce papier, nous étendons le modèle PLSA en distinguant les variables latentes de partitions de celles des thématiques. Dans ce cas, nous supposons que les mots sont à la fois générés par un discours général représenté par les partitions de documents ainsi que par différentes thématiques présentes dans chacune de ces partitions. Nous prouvons de plus que notre modèle PLSA étendu est équivalent à une combinaison linéaire de modèles de factorisation matricielle non-négative (FMN) au sens de la fonction objective KL-divergence. En outre, les résultats empiriques de partitionnement de documents sur trois collections de documents textuelles montrent que notre approche est statistiquement plus performante que le modèle de PLSA de base et la FMN dans la majorité des cas.

# 2 Modèles de partitionnement de documents

## 2.1 Probabilistic Latent Semantic Analysis (PLSA)

Avec le modèle (PLSA) (Hofmann, 1999), chaque document d'une collection $\mathcal{D}$ est représenté par une distribution de probabilité sur les $K$ valeurs de la variable thématique latente $\alpha \in \mathcal{A} = \{\alpha_1, ..., \alpha_K\}$, où chaque valeur de $\alpha$ correspond à une distribution de probabilité sur l'ensemble des mots de la collection. Dans le processus génératif correspondant à ce modèle (figure 1, a), un document est d'abord choisi suivant la probabilité $P(d)$, ensuite une thématique $\alpha$ est générée avec une probabilité $P(\alpha|d)$, et finalement un mot $w$ est émis suivant la probabilité $P(w|\alpha)$. Dans le cas où les partitions de documents sont confondues aux thématiques, l'algorithme PLSA peut être utilisé comme un algorithme de clustering de documents.

## 2.2 Extension de PLSA

Dans notre extension du modèle PLSA, nous supposons que les mots du vocabulaire $\mathcal{V}$ associé à la collection $\mathcal{D}$ sont générés à la fois par les clusters de documents et les thématiques de mots. Cette supposition peut s'interpréter de la manière suivante : dans une collection, il existe différentes thématiques de documents correspondant aux différents discours et à l'intérieur de chacune de ces thématiques, il peut y avoir des sous-thématiques. Ainsi, la différence principale entre notre modèle et le modèle PLSA est l'utilisation d'une variable latente supplémentaire $\beta$, représentant les sous-thématiques (ou concepts de mots dans la suite). Le processus génératif correspondant à notre modèle (figure 1, b) est le suivant : choisir un document $d$ suivant la probabilité $P(d)$, générer une thématique $\alpha$ d'après $P(\alpha|d)$, choisir un concept de mots $\beta$ suivant la probabilité $P(\beta)$, et générer un mot $w$ d'après $P(w|\alpha, \beta)$. Grâce aux deux variables latentes $\alpha$ et $\beta$, le nouveau modèle est capable de capturer le discours sur deux niveaux de granularités différents. Lorsque notre modèle est utilisé pour le clustering de documents, le nombre de clusters de documents (cardinalité de $\alpha$) peut être choisi indépendamment du nombre de concepts de mots (cardinalité de $\beta$) présents dans la collection. Pour apprendre les paramètres du modèle, nous utilisons l'algorithme EM en maximisant la log-vraisemblance comme avec le modèle PLSA.

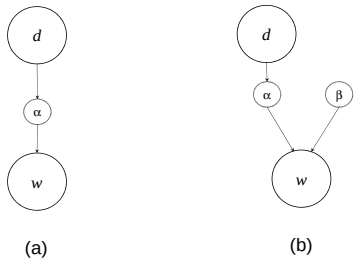


FIGURE 1 – Modèles graphiques correspondants à (a) PLSA et (b) PLSA étendu.

### 2.3 PLSA étendu et Combinaison linéaire de FMN

La factorisation en matrices non-négatives (FMN) permet d'approximer une matrice non-négative $F$ par un produit de matrices $C$ et $H$ : $F \approx HC^t$. Recemment, (Ding *et al.*, 2008) ont montré que les modèles PLSA et FMN sont équivalents au sens de la divergence de Kullback-Leibler (KL) en utilisant des lois de mise-à-jour de matrices $C$ et $H$ décrites dans Lee & Seung (1999). À partir de cette preuve nous décomposons les probabilités jointes $\forall i, \forall j, P(d_i, w_j)$ suivants les probabilités des variables latentes correspondantes aux thématiques et aux concepts de mots dans le modèle PLSA étendu et on trouve

$$P(d_i, w_j) = \sum_l P(\beta_l) \sum_k P(\alpha_k) P(d_i \mid \alpha_k) P(w_j \mid \alpha_k, \beta_l) \tag{1}$$

D'après (Ding *et al.*, 2008), la maximisation de la log-vraisemblance des données est équivalente à la minimisation de la divergence KL. En plus chacun des termes $\sum_k P(\alpha_k)P(d_i \mid \alpha_k)P(w_j \mid \alpha_k, \beta_l)$ est équivalent à une factorisation $\tilde{H}S\tilde{C}_l^t$. Le modèle PLSA étendu peut ainsi être vu comme une combinaison linéaire des FMN suivantes :

$$F \approx \sum_l p(\beta_l) H C_l^t \tag{2}$$

## 3 Résultats

Dans nos expériences nous avons utilisé les collections `Reuters` [1], `20Newsgroups` [2] et `WebKB` [3]. Afin d'évaluer la pertinence des partitions obtenues, nous suivons l'approche de (Slonim & Tishby, 2002) en attribuant à chaque partition la classe majoritaire qu'elle contient. Comme critères d'évaluations nous avons utilisé les mesures, micro-moyenne de précision, et l'Information Mutuelle Normalisée (IMN) et les performances obtenues sont une moyenne à partir de 10 sous-ensembles aléatoires des collections de départ.

Pour le modèle PLSA étendu, les expériences avec des nombres variés de concepts de mots sont nécessaires pour trouver le nombre de concepts qui donne le meilleur résultat, et pour vérifier l'influence sur les résultats du nombre de concepts. Nous avons ainsi varié le nombre de concepts de mots ($L$) de 10 à 70 pour la collection `Reuters`, et de 10 à 100 pour `20Newsgroups` et `WebKB`. Nous présentons au tableau 1, une comparaison entre les différentes méthodes de clustering. Les précisions moyennes affichées sont calculées sur les 10 sous-ensembles de chaque collection. Pour le modèle PLSA étendu, elles sont calculées pour différentes valeurs du nombre de concepts de mots. Nous prenons ici la meilleure précision moyenne parmi ces moyennes (tableau 1). Ceux à droite représentent les valeurs IMN moyennes. Comme les modèles PLSA, FMN et K-means n'utilisent pas de concepts de mots, nous prenons seulement la précision moyenne et

1. http ://www.daviddlewis.com/resources/testcollections/reuters21578/
2. http ://kdd.ics.uci.edu/databases/20newsgroups
3. http ://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

l'IMN moyenne sur les dix sous-ensembles pour ces modèle. Le symbole $^{\downarrow}$ indique que la performance est significativement pire que la meilleure performance selon le test de Wilcoxon à un seuil de $5\%$. Pour les meilleures précisions moyennes, les performances du modèle PLSA étendu sont supérieures aux autres méthodes sur les trois collections de documents.

TABLE 1 – (Meilleure) Précision moyenne et l'IMN correspondante des différents modèles de clustering sur les corpus `Reuters`, `20Newsgroups` et `WebKB`.

| Modèles | `Reuters` | | `20Newsgroups` | | `WebKB` | |
|---|---|---|---|---|---|---|
| | Prec. moy. | IMN | Prec. moy. | IMN | Prec. moy. | IMN |
| `Kmeans` | $0.52^{\downarrow}$ | $0.20^{\downarrow}$ | $0.32^{\downarrow}$ | $0.07^{\downarrow}$ | $0.43^{\downarrow}$ | $0.08^{\downarrow}$ |
| `FMN` | $0.69$ | $0.42$ | $0.45^{\downarrow}$ | $0.20^{\downarrow}$ | $0.50^{\downarrow}$ | $0.20^{\downarrow}$ |
| `PLSA` | $0.64^{\downarrow}$ | $0.38^{\downarrow}$ | $0.71^{\downarrow}$ | $0.49^{\downarrow}$ | $0.61^{\downarrow}$ | $0.28^{\downarrow}$ |
| `PLSA-Eten` | **0.71** | **0.42** | **0.77** | **0.54** | **0.68** | **0.36** |

Nous avons proposé une extension du modèle PLSA en distinguant les variables latentes correspondantes aux thématiques de documents et aux concepts de mots. Cette distinction nous a permis d'avoir une modélisation plus fine du processus générative des données. Nous avons en outre montré l'équivalence entre le PLSA étendu et une combinaison linéaire de modèles de factorisation en matrices non-négatives. Et, nous avons aussi validé empiriquement notre approche sur trois collections de documents textuels pour la tâche de clustering. Nos résultats empiriques montrent que la dissociation des thématiques et des sous-thématiques permettent en effet d'améliorer substantiellement les résultats du partitionnement par rapport au modèle de PLSA de base.

# Références

DING C., LI T. & PENG W. (2008). On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics and Data Analysis*, **52**, 3913–3927.

HOFMANN T. (1999). Probabilistic latent semantic indexing. In *SIGIR '99 : Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*.

LEE D. D. & SEUNG H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755), 788–791.

SLONIM N. & TISHBY N. (2002). Unsupervised document classification using sequential information maximization. In *ACM SIGIR*, p. 129–136.