Utilisation de PLSI en recherche d'information
Représentation des requêtes[†]

Jean-Cédric Chappelier Emmanuel Eckard
Laboratoire d'Intelligence Artificielle
École polytechnique fédérale de Lausanne, Suisse
{jean-cedric.chappelier, emmanuel.eckard}@epfl.ch

Résumé. Le modèle PLSI (« *Probabilistic Latent Semantic Indexing* ») offre une approche de l'indexation de documents fondée sur des modèles probabilistes de catégories sémantiques latentes et a conduit à des applications dans différents domaines. Toutefois, ce modèle rend impossible le traitement de documents inconnus au moment de l'apprentissage, problème particulièrement sensible pour la représentation des requêtes dans le cadre de la recherche d'information. Une méthode, dite de « *folding-in* », permet dans une certaine mesure de contourner ce problème, mais présente des faiblesses. Cet article introduit nouvelle une mesure de similarité document-requête pour PLSI, fondée sur les modèles de langue, où le problème du « *folding-in* » ne se pose pas. Nous comparons cette nouvelle similarité aux noyaux de Fisher, l'état de l'art en la matière. Nous présentons aussi une évaluation de PLSI sur un corpus de recherche d'information de près de 7500 documents et de plus d'un million d'occurrences de termes provenant de la collection TREC–AP, une taille considérable dans le cadre de PLSI.

Abstract. The PLSI model ("*Probabilistic Latent Semantic Indexing*") offers a document indexing scheme based on probabilistic latent category models. It entailed applications in diverse fields, notably in information retrieval (IR). Nevertheless, PLSI cannot process documents not seen during parameter inference, a major liability for queries in IR. A method known as "folding-in" allows to circumvent this problem up to a point, but has its own weaknesses. The present paper introduces a new document-query similarity measure for PLSI based on language models that entirely avoids the problem a query projection. We compare this similarity to Fisher kernels, the state of the art similarities for PLSI. Moreover, we present an evaluation of PLSI on a particularly large training set of almost 7500 document and over one million term occurrence large, created from the TREC–AP collection.

1 Introduction

Depuis dix ans, le modèle PLSI (« *Probabilistic Latent Semantic Indexing* ») (Hofmann, 1999; Hofmann, 2000; Hofmann, 2001) offre une approche de l'indexation de documents fondée sur des modèles probabilistes de catégories sémantiques latentes. Ce modèle a conduit à plusieurs applications (Ahrendt *et al.*, 2005; Gaussier *et al.*, 2002; Jin *et al.*, 2004; Mei & Zhai, 2006; Steyvers *et al.*, 2004; Vinokourov & Girolami, 2002), notamment dans le domaine de la recherche d'information (RI). Toutefois, une limitation majeure de ce modèle vient du fait qu'il n'est pas génératif vis à vis de documents dont le modèle est inconnu, et qu'il tend à sur-apprendre (Blei

[†] Ce travail a été financé dans le cadre du projet 200020–119745 du Fond National Suisse.

et al., 2003; Popescul *et al.*, 2001). Un certain nombre d'extensions et d'alternatives ont été proposées pour y remédier : *Latent Dirichlet Allocation* (Blei *et al.*, 2003), *undirected PLSI* (Welling *et al.*, 2005), *correlated topic models* (Blei & Lafferty, 2007), *rate adapting Poisson models* (Gehler *et al.*, 2006) ; mais ces améliorations restent coûteuses en terme de complexité.

Dans le cadre de la RI, la nature non générative de PLSI vis-à-vis des modèles de document inconnus conduit à un traitement spécifique des requêtes, appelé « *folding-in* », qui consiste à estimer leurs paramètres spécifiques, non vus pendant la phase d'apprentissage (Hofmann, 1999; Hinneburg *et al.*, 2007). Le but de cet article est d'introduire une nouvelle similarité document–requête théoriquement fondée, présentée en section 3, qui évite le « *folding-in* » : on considère les requêtes comme de nouvelles instances générées par des modèles de documents *déjà* connus. Cette nouvelle approche est comparée à l'état de l'art pour PLSI basé sur les noyaux de Fisher (Chappelier & Eckard, 2009). Pour finir, la section 4 apporte des résultats expérimentaux obtenus sur une grande collection créée à partir du corpus d'évaluation TREC–AP. PLSI n'étant pas génératif, ses paramètres doivent être effectivement appris sur toute la collection utilisée, et non seulement sur un échantillon d'apprentissage. À notre connaissance, il n'avait jamais été tenté d'appliquer PLSI à une base d'une telle envergure, plus de 7000 documents et d'un million d'occurrences de termes.

2 Le modèle PLSI

Dans le modèle PLSI, les documents sont représentés comme des occurrences successives de paires d'indices (d, w) pour une catégorie $z \in Z$ donnée, d étant l'indice d'un document et w, celui d'un terme. De plus, w et d sont supposés indépendants sachant z, de sorte que le modèle s'écrit : $P(d, w) = \sum_{z \in Z} P(z)\, P(w|z)\, P(d|z)$.

Les paramètres de PLSI sont $\theta = (P(z), P(w|z), P(d|z))$, pour tous les z, w et d possibles dans le modèle. Ces paramètres s'estiment pour une collection de documents donnée en utilisant une variante de l'algorithme *expectation-maximisation* (EM) (Hofmann, 1999; Hofmann, 2001).

Le modèle de similarité document-requête utilisé dans PLSI repose sur les noyaux de Fisher (Hofmann, 2000). Plusieurs variantes existent en fonction des approximations effectuées (Chappelier & Eckard, 2009), mais chacune se compose de deux termes additifs qui traduisent respectivement la contribution directe des catégories latentes, notée K_z, et celle des termes, notée K_w. La différence la plus significative entre ces variantes réside dans la façon d'approcher la matrice d'information de Fisher, soit pas la matrice identité comme fait initialement, soit par la diagonale (variantes DFIM, pour « *Diagonal Fisher Information Matrix* »). La prise en compte des termes DFIM pondère les composantes K_z et K_w, évitant une sur-représentation de K_z, dont les performances sont faibles (Chappelier & Eckard, 2009).

3 Éviter la projection des requêtes

La projection des requêtes (« *folding-in* ») est une technique qui permet de contourner la nature non générative de PLSI en estimant les paramètres des documents inconnus tels que les requêtes : les paramètres $P(q|z)$ d'une requête q sont estimés par un processus EM simplifié où les valeurs des $P(w|z)$ et $P(z)$ sont fixées sur celles initialement apprises sur le corpus.

Cette méthode a ses inconvénients, notamment pour l'estimation de la vraisemblance du corpus d'apprentissage (Welling *et al.*, 2008) et la cohérence avec les $P(d|z)$ connus.

Nous présentons ici une nouvelle mesure de similarité document–requête qui s'inspire des méthodes à base de modèles de langue (Ponte & Croft, 1998; Zhai, 2008) qui évite entièrement la phase de projection pour les requêtes et les problèmes liés à l'apprentissage des paramètres $P(q|z)$: on représente les requêtes non comme des nouveaux modèles de documents pour lesquels les paramètres $P(q|z)$ sont à apprendre, mais comme de nouvelles occurrences des modèles des documents déjà connus. On réduit ainsi la RI à un problème d'identification de modèle : pour une requête q donnée, quels sont les modèles d déjà connus les plus représentatifs de q ?

Une solution classique à une telle question consiste à maximiser la log-vraisemblance de la requête par rapport au modèle $P(d, w)$ (Ponte & Croft, 1998) :

$$\mathcal{S}_{\text{LogL}}(d, q) = \sum_{w \in q \cap d} n(q, w) \log P(d, w), \tag{1}$$

où $n(q, w)$ est le nombre d'occurrences du terme w dans la requête q, et où « $w \in q \cap d$ » représente les termes qui apparaissent dans q (i.e. $n(q, w) > 0$) et tels que $P(d, w) > 0$.

Une autre solution courante pour l'identification de modèle est la minimisation de la divergence de Kullback-Leibler entre la distribution empirique (q) et la distribution du modèle (d) (Lafferty & Zhai, 2001) :

$$\mathcal{S}_{\text{KL}}(d, q) = -\text{KL}\left(\widehat{P}(w|q) \middle\| P(w|d)\right) = \sum_{w \in q \cap d} \widehat{P}(w|q) \log \frac{P(w|d)}{\widehat{P}(w|q)}, \tag{2}$$

avec $\widehat{P}(w|q) = n(q, w)/|q|$ le nombre d'occurrences du terme w dans la requête q divisé par sa longueur $|q|$.

Ces deux approches, bien que liées, ne sont pas équivalentes :

$$\mathcal{S}_{\text{KL}}(d, q) = \frac{1}{|q|}\Big(\mathcal{S}_{\text{LogL}}(d, q) - |q| \, \log P(d)\Big) - \underbrace{\sum_{w} \widehat{P}(w|q) \log \widehat{P}(w|q)}_{f(q)}.$$

Lors de la maximisation de $\mathcal{S}(d, q)$ par rapport à d pour une requête q donnée, les deux approches se distinguent par un facteur additif $|q| \log P(d)$: cela revient à prendre ou non en compte la longueur des documents via $|q|$ et $P(d)$, qui est en pratique très proche de $|d|/|C|$ (où $|C|$ est la taille de tout le corpus).

On peut aussi généraliser les démarches précédentes à tout estimateur $\widetilde{P}(w|q)$ de $\widehat{P}(w|q)$. Par exemple, le lissage de Jelinek-Mercer (JM) (Zhai & Lafferty, 2004) donne $\widetilde{P}(w|q) = (1 - \lambda)\, \widehat{P}(w|q) + \lambda\, P_{\text{GE}}(w)$, avec une constante de lissage λ comprise entre 0 et 1, et $P_{\text{GE}}(w)$ la probabilité a priori («*General English*») du terme w, typiquement estimée par $P_{\text{GE}}(w) = \sum_{d \in C} \widehat{P}(w, d)$.

Une autre façon de construire un estimateur $\widetilde{P}(w|q)$ de $\widehat{P}(w|q)$ consiste à prendre les documents les plus pertinents d'une première phase de recherche, comme il est fait dans le Modèle de Pertinence (Lavrenko & Croft, 2001) et dans le *pseudo-feedback* (Zhai & Lafferty, 2001) : une

	CACM	CRAN	TIME	CISI	MED	AP89_01XX
Nb. de termes	4 911	4 063	13 367	5 545	7 688	13 379
Nb. d'occurrences ($\|C\|$)	90 927	120 973	114 850	87 067	76 571	1 321 482
Nb. de documents	1 587	1 398	425	1 460	1 033	7 466
$\|d\|$ moyen	56.8	85.1	268.6	56.7	73.8	177.2
Nb. de requêtes	64	225	83	112	30	50
$\|q\|$ moyen	12.7	8.9	8.2	37.7	11.4	79.3

TAB. 1 – Données des collections de documents utilisées pour l'évaluation.

première recherche est effectuée en utilisant la similarité ci-dessus (Eq. (2), avec $\widehat{P}(w|q)$ ou sa version lissée) ; on utilise alors les N documents les plus pertinents pour estimer $\widetilde{P}(w|q)$ par

$$\widetilde{P}(w|q) = \frac{1}{N} \sum_{i=1}^{N} P\left(d_i(q), w\right),$$

avec $d_i(q)$ le $i$ème meilleur document pour la requête q. Une deuxième phase est ensuite effectuée en utilisant $\mathcal{S}(d,q) = -\mathrm{KL}(\widetilde{P}(w|q), P(w|d))$.

On obtient donc finalement huit schémas de recherche sans projection de requêtes : la log-vraisemblance (Eq. 1) ou la divergence de Kullback-Leibler (Eq. 2), avec pour chacune la possibilité d'appliquer un lissage de Jelinek-Mercer, le pseudo-feedback, ou les deux.

4 Expériences

Afin d'évaluer l'approche proposée ici, nous considérons 14 mesures de similarité : les huit basées sur les modèles de langage (décrite en section précédente) et les 6 meilleures variantes du noyau de Fisher (Chappelier & Eckard, 2009) : le modèle d'origine de Hofmann K^H, sa version DFIM $K^{\text{DFIM-H}}$, ainsi que leurs composantes « termes » K_w et « catégories » K_z. Les questions suivantes se posent alors : 1) comment se comporte la nouvelle approche sans « *folding-in* » par rapport aux meilleurs noyaux de Fisher ? 2) comment ces mesures se comparent-elles à l'état de l'art, le modèle BM25 (Robertson *et al.*, 1994) ? 3) la ré-estimation de $P(w|q)$, que ce soit à l'aide du lissage de Jelinek-Mercer ou du pseudo-feedback, améliore-t-il les résultats ?

La nature non générative de PLSI oblige à estimer les paramètres sur l'entièreté de la collection évaluée, et il est donc impossible d'utiliser des collections aussi grandes que celles de TREC dans leur totalité. En cohérence avec les travaux précédemment publiées sur PLSI, nous utilisons les collections d'évaluation de SMART[1] : CACM, CISI, MED, CRAN et TIME pour répondre à ces questions. De plus, nous utilisons un plus grand corpus constitué d'une partie de la collection TREC–AP 89. Pour les mêmes raisons, nous n'avons gardé que les 7466 premiers documents de cette collection, et les requêtes 1 à 50.[2] Les caractéristiques de ces collections sont données dans le tableau 1.

[1] `ftp://ftp.cs.cornell.edu/pub/smart/`

[2] Les documents AP890101-0001 à AP890131-0311. La phase d'apprentissage de EM pour $|Z| = 128$ a pris 45 heures de temps CPU, et utilisé 6.7 Gb de RAM, sur un serveur de calcul Intel Xenon octo-cœur de 2 GHz avec 32 Gb de mémoire.

		CACM	CRAN	TIME	CISI	MED	AP89
Résultats	MAP de BM25	**31.4**	**42.4**	**69.2**	12.3	52.3	19.7
	MAP du meilleur modèle PLSI	30.0	39.6	60.8	**20.2**	**53.8**	**21.6**
	Meilleur modèle PLSI :	K_w^{H}	$\mathcal{S}_{\mathrm{KL}}$	$K_w^{\text{DFIM-H}}$	K_w^{H}	K^{H}	$K_w^{\text{DFIM-H}}$
	Obtenu pour $\lvert Z\rvert =$	16	128	8	8	32	48
	MAP de $\mathcal{S}_{\mathrm{KL},\lvert Z\rvert=128}$	22.9	39.6	49.1	19.5	52.8	11.4
Concl.	PLSI $>$ BM25 ?	Non	Non	Non	**OUI**	**oui**	**oui**
	$\mathcal{S}_{\mathrm{KL},\lvert Z\rvert=128}$ vs noyaux de Fisher	$<$	$>$	$<$	$\simeq$	$\simeq$	$<$

TAB. 2 – Principaux résultats des 14 modèles sur les 6 collections.

Pour les collections SMART, chaque expérience a été effectuée 6 fois avec des conditions initiales d'apprentissage différentes, pour chaque modèle, et pour différentes quantités de catégories latentes : $|Z| \in \{1, 2, 8, 16, 32, 64, 128\}$; soit 2940 expériences en tout. Pour TREC-AP, les expériences ont été effectuées avec une seule condition initiale d'apprentissage, pour chaque $|Z| \in \{1, 32, 48, 64, 80, 128\}$, soit 84 expériences en tout.

Pour toutes ces expériences, le stemmer de Porter implémenté dans Xapian[3] a été utilisé. Les résultats de l'évaluation ont été obtenus par l'outil standard `trec_eval`[4]. Nous présentons ici les résultats en termes de *Mean Average Precision* (MAP), mais les conclusions sont similaires si l'on utilise la précision à 5 points ou la R-précision. A l'exception de la figure 2, les figures publiées représentent la MAP en fonction du nombre $|Z|$ de catégories latentes, moyennée sur 6 expériences, ainsi que les barres d'erreur correspondant à un écart-type. Les principales conclusions de ces 3024 expériences, résumées dans le tableau 2, sont :

1. Figure 1 : $\mathcal{S}_{\mathrm{KL}}$ (Eq. 2) donne de meilleures performances que $\mathcal{S}_{\mathrm{LogL}}$ (Eq. 1).
   Les deux mesures ($\mathcal{S}_{\mathrm{KL}}$ et $\mathcal{S}_{\mathrm{LogL}}$) améliorent leurs performances au fur et à mesure que $|Z|$ grandit. Nous nous sommes arrêtés à $|Z| = 128$ pour des raisons pratiques.
2. Figures 3 et 4 : $\mathcal{S}_{\mathrm{KL}}$ surpasse les meilleurs noyaux de Fisher sur CRAN et obtient des performances similaires sur MED et CISI. Rappelons qu'il est aussi supérieur parce qu'il ne requiert de phase spécifique pour les requêtes (« *folding-in* »).
3. Figures 3 et 4 : le lissage de $\widehat{P}(q|w)$ n'améliore pas les performances, ni par lissage JM, ni par le *pseudo-feedback*. De plus, bien que l'implémentation ait fait l'objet d'un effort particulier pour en limiter la complexité, le lissage augmente considérablement le temps d'évaluation (il n'a pas d'effet sur le temps d'apprentissage) : contrairement à la variante non lissée, ce ne sont pas seulement les termes présents dans la requête, aussi ceux du document, voire de toute la collection, qui entrent en ligne de compte ; aussi l'évaluation d'une requête est-elle bien plus lente que sans lissage : entre 2 (CISI) et 20 (TIME, MED) fois plus lents pour le lissage de Jelinek-Mercer, et entre 30 (MED) et 150 (CRAN) fois plus lent avec la recherche en deux passes avec $N = 3$.
4. Figures 3 : sur des collections « sémantiquement difficiles », les meilleurs noyaux PLSI ont de meilleures performances que le modèle BM25 : CISI, où les requêtes et les documents ne partagent que de rares termes (et donc particulièrement utile pour mesurer dans quelle mesure un modèle de recherche est robuste à la synonymie.)[5], MED (vocabulaire spécialisé), et TREC–AP.

[3] `http://xapian.org/`
[4] `http://trec.nist.gov/trec_eval/`
[5] CISI est remarquable pour avoir des requêtes censées retourner des documents avec lesquelles elles ne partagent *aucun* terme significatif.

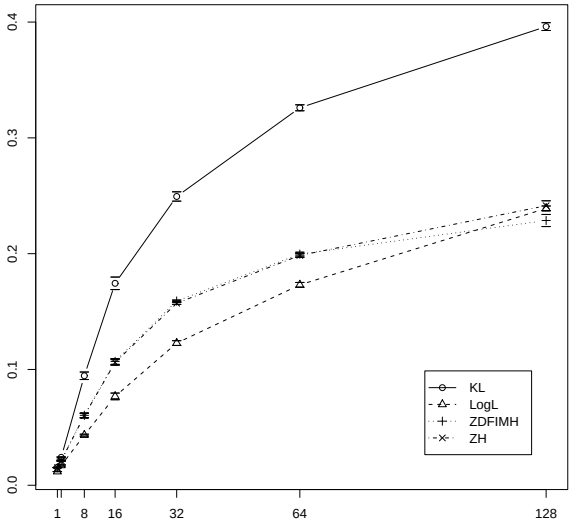

FIG. 1 – Exemple typique (ici sur CRAN) qui montre comment la similarité $\mathcal{S}_{\text{KL}}$ basée sur $P(w|d)$ dépasse $\mathcal{S}_{\text{LogL}}$ basée sur $P(d,w)$. Les composantes K_z^{H} (ZH) et $K_z^{\text{DFIM-H}}$ (ZDFIMH) des noyaux de Fisher sont également représentées.

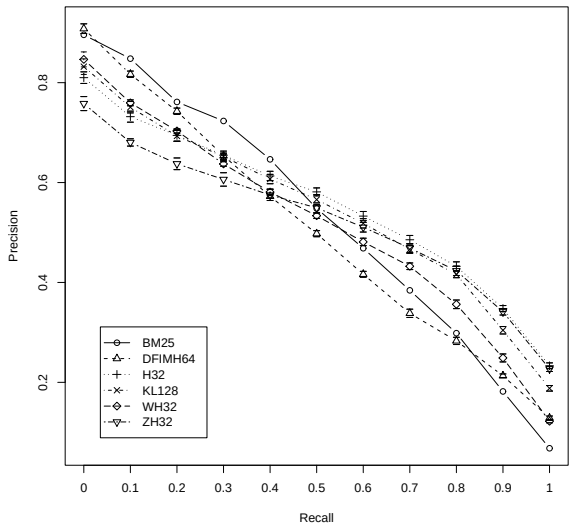

FIG. 2 – Courbes précision-rappel sur MED pour BM25, $K^{\text{DFIM-H}}$ à $|Z|$ =64 (DFIMH64), K^{H} à $|Z|$ =32 (H32), $\mathcal{S}_{\text{KL}}$ à $|Z|$ =128 (KL128), K_w^{H} à $|Z|$ =32 (WH32), et K_z^{H} à $|Z|$ =32 (YH32).

Les conclusions doivent être plus nuancées pour MED : les différents modèles n'ont pas le même comportement à différentes valeurs de rappel (figure 2) ; certains sont meilleurs pour un rappel bas et d'autres meilleurs à un rappel haut. Des mesures de performance globales comme la MAP ou la R-précision ne rendent pas compte de ces subtilités.

5 Conclusions

Cet article apporte un modèle de similarité pour PLSI théoriquement fondée évitant entièrement les écueils de la représentation des requêtes ; par ailleurs, il fournit une évaluation des performances de PLSI sur une collection plus grande que les collections SMART sur lesquelles les expériences ont été faites jusqu'à présent. Aux questions qui se posent, nous pouvons répondre :

1. que la nouvelle approche sans « *folding-in* » des requêtes se compare favorablement aux meilleures variantes du noyau de Fisher, particulièrement pour les plus grands nombres de catégories latentes ;
2. et que ces modèles se comparent favorablement avec BM25 pour les collections sémantiquement difficiles comme CISI, MED et TREC–AP.
3. Figures 3 et 4 : le lissage de $\widehat{P}(w|q)$ n'améliore les résultats dans aucun des cas testés. Sur les huit variantes proposées, seule la similarité de Kullback-Leibler avec $P(w|d)$ non lissé est valable.

Ainsi, nous confirmons expérimentalement que les modèles à catégories latentes comme PLSI pourraient s'avérer intéressants pour la recherche d'information sur des collections de taille raisonnable, mais sémantiquement difficiles, où les requêtes et les documents qui leur sont pertinents ne partagent que peu de termes. Dans ces cas, il est recommandé d'utiliser $K_w^{\text{DFIM-H}}$ ou $\mathcal{S}_{\text{KL}}$ (Eq. 2) s'il est possible de faire tourner l'apprentissage avec un nombre suffisant de

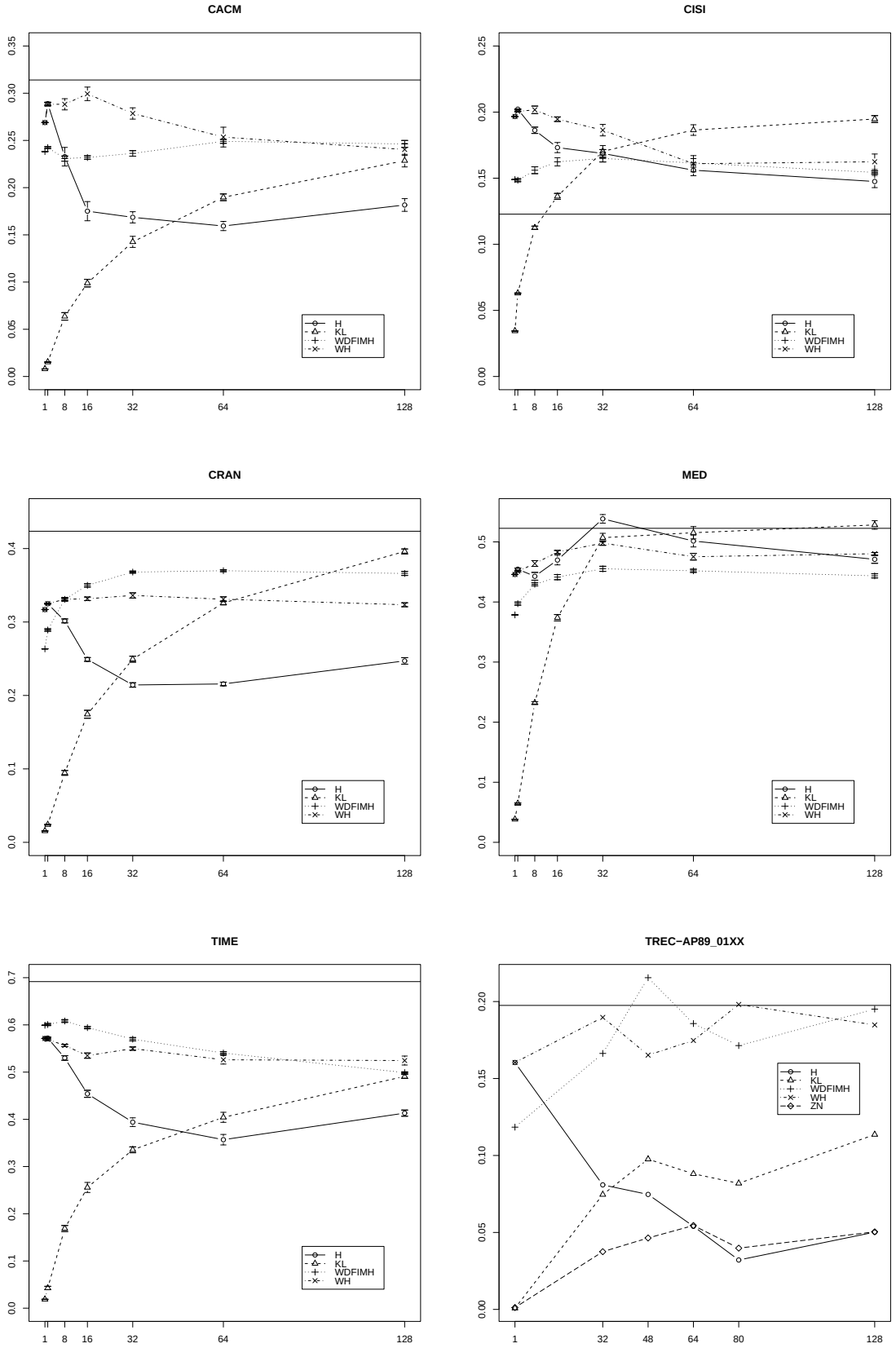

FIG. 3 – Résultats obtenus sur les 6 collections pour différents modèles : K^{H} (H), $\mathcal{S}_{\mathrm{KL}}$ (KL), $K_w^{\text{DFIM-H}}$ (WDFIMH), et K_w^{H} (WH). Les lignes horizontales représentent les performances de BM25, qui ne dépend pas de $|Z|$.

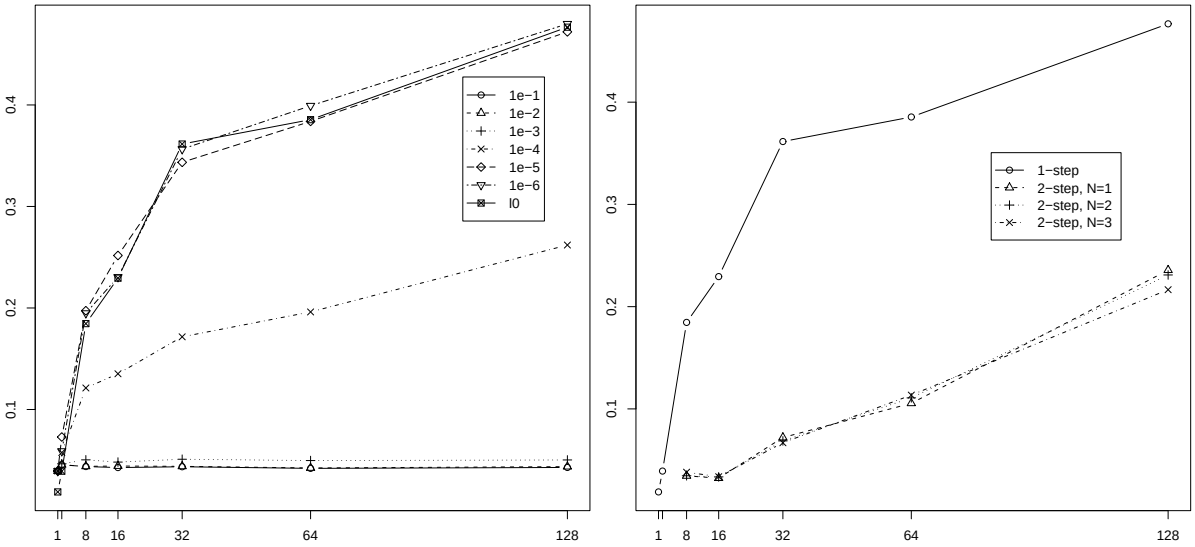

FIG. 4 – Exemple typique montrant (ici sur TIME) comment le lissage de Jelinek-Mercer de $\widehat{P}(q|w)$ (à gauche) ou par le pseudo-feedback (à droite) dégradent les performances, comparé à la variante non lissée $\widehat{P}(q|w)$ notée « l0 » à gauche, et « 1 step » à droite.

catégories latentes. $\mathcal{S}_{\text{KL}}$ présente de plus l'avantage de ne pas demander de ré-apprentissage pour la projection des requêtes (« *folding in* »).

Toutefois, la conclusion globale est que PLSI n'est pas adapté à la recherche documentaire sur de *grandes* collections : comme il comporte en paramètre un modèle des documents vus pendant l'apprentissage, il est pas nature non génératif, et ne passe tout simplement pas à l'échelle lorsque le nombre de documents atteint les dizaines de milliers. De plus, pour sophistiqué qu'il soit, PLSI surpasse à peine le modèle BM25, et cela au prix d'une complexité et d'un temps de calcul rédhibitoires.

On peut spéculer que PLSI pourrait s'avérer significativement meilleur que les modèles de l'état de l'art en utilisant un bien plus grand nombre de catégories latentes, mais les limitations induites par le nombre de paramètres à apprendre rendent ces cas impossibles à calculer en pratique. Il est fort probable qu'un très grand $|Z|$ améliore les performances de K_z ou de $\mathcal{S}_{\text{KL}}$, mais l'apprentissage de tels modèles s'avère en pratique impossible, tout particulièrement pour les grandes collections pour lesquelles de tels $|Z|$ seraient justement le plus appropriés.

Références

AHRENDT P., GOUTTE C. & LARSEN J. (2005). Co-occurrence models in music genre classification. In IEEE *Int. Workshop on Machine Learning for Signal Processing*.

BLEI D. & LAFFERTY J. (2007). A correlated topic model of *Science*. *An. of App. Stat.*, **1**(1), 17–35.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

CHAPPELIER J.-C. & ECKARD E. (2009). Rôle de la matrice d'information et pondération des composantes dans les noyaux de Fisher pour PLSI. In *Actes de la Sixième Conférence francophone en Recherche d'Information et Applications*, p. 267–282 : LSIS-USTV.

GAUSSIER E., GOUTTE C., POPAT K. & CHEN F. (2002). A hierarchical model for clustering and categorising documents. In *Proc. of 24th BCS-IRSG European Colloquium on IR Research*, p. 229–247.

GEHLER P. V., HOLUB A. D. & WELLING M. (2006). The rate adapting Poisson model for information retrieval and object recognition. In *Proc. of the 23rd Int. Conf. on Machine Learning*, p. 337–344.

HINNEBURG A., GABRIEL H.-H. & GOHR A. (2007). Bayesian folding-in with Dirichlet kernels for PLSI. In *Proc. of the 7th IEEE Int. Conf. on Data Mining*, p. 499–504.

HOFMANN T. (1999). Probabilistic latent semantic indexing. In *Proc. of 22th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, p. 50–57.

HOFMANN T. (2000). Learning the similarity of documents : An information-geometric approach to document retrieval and categorization. In *Adv. in Neural Inf. Proc. Sys. (NIPS)*, volume 12, p. 914–920.

HOFMANN T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, **42**(1), 177–196.

JIN X., ZHOU Y. & MOBASHER B. (2004). Web usage mining based on probabilistic latent semantic analysis. In *Proc. of 10th Int. Conf. on Knowledge Discovery and Data Mining*, p. 197–205.

LAFFERTY J. & ZHAI C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proc. of 24th Annual Int. Conference on Research and Development in Information Retrieval (SIGIR)*, p. 111–119.

LAVRENKO V. & CROFT W. B. (2001). Relevance based language models. In *Proc. of 24th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, p. 120–127.

MEI Q. & ZHAI C. (2006). A mixture model for contextual text mining. In *Proc. of 12th Int. Conf. on Knowledge Discovery and Data Mining*, p. 649–655.

PONTE J. M. & CROFT W. B. (1998). A language modeling approach to information retrieval. In *Proc. of 21st Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, p. 275–281.

POPESCUL A., UNGAR L. H., PENNOCK D. M. & LAWRENCE S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proc. of the 17th Conf. in Uncertainty in Artificial Intelligence*, p. 437–444.

ROBERTSON S. E., WALKER S., JONES S., HANCOCK-BEAULIEU M. & GATFORD M. (1994). Okapi at TREC–3. *Proc. of the 3rd Text REtrieval Conf.*

STEYVERS M., SMYTH P., ROSEN-ZVI M. & GRIFFITHS T. (2004). Probabilistic author-topic models for information discovery. In *10th Int. Conf. on Knowledge Discovery and Data Mining*, p. 306–315.

VINOKOUROV A. & GIROLAMI M. (2002). A probabilistic framework for the hierarchic organisation and classification of document collections. *Journ. of Intelligent Information Systems*, **18**(2/3), 153–172.

WELLING M., CHEMUDUGUNTA C. & SUTTER N. (2008). Deterministic latent variable models and their pitfalls. *SIAM Conference on Data Mining SDM 2008*.

WELLING M., ROSEN-ZVI M. & HINTON G. (2005). Exponential family harmoniums with an application to information retrieval. In *Ad. in Neural Inf. Proc. Sys. (NIPS)*, volume 17, p. 1481–1488.

ZHAI C. (2008). Statistical language models for information retrieval a critical review. *Found. Trends Inf. Retr.*, **2**(3), 137–213.

ZHAI C. & LAFFERTY J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *Proc. of 10th Int. Conf. on Information and Knowledge Management (CIKM)*, p. 403–410.

ZHAI C. & LAFFERTY J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, **22**(2), 179–214.