Rôle de la matrice d'information et pondération des composantes dans les noyaux de Fisher pour PLSI[†]

Jean-Cédric Chappelier — Emmanuel Eckard

Laboratoire d'Intelligence Artificielle
École Polytechnique Fédérale de Lausanne
CH–1015 Lausanne
{jean-cedric.chappelier,emmanuel.eckard}@epfl.ch

RÉSUMÉ. *Des similarités entre documents à base de catégories sémantiques latentes et de noyaux de Fisher ont été proposées pour la première fois il y a dix ans par T. Hofmann dans le contexte du "Probabilistic Latent Semantic Indexing", puis étendues par Nyffenegger et al. (2006). Le présent article présente une étude approfondie et une révision de ces modèles par (1) une description unifiée et simplifiée, (2) une étude du rôle de la matrice d'information de Fisher $G(\theta)$, et (3) une analyse de l'impact des paramètres associés aux catégories latentes. Il fournit de plus de nouveaux résultats expérimentaux sur une grande collection de document provenant du corpus d'évaluation TREC–AP.*

ABSTRACT. *An information-geometric approach for document similarities in the framework of "Probabilistic Latent Semantic Indexing" was first proposed by T. Hofmann (2000) and later extended ("revisited") by Nyffenegger et al. (2006). This paper presents an in-depth study and revision of these models by (1) providing a simpler unified description framework, (2) investigating the role of the Fisher Information Matrix $G(\theta)$, and (3) analyzing the impact of latent "topic" parameters in such models. It furthermore provides new experimental results on larger collections coming from the TREC–AP evaluation corpus.*

MOTS-CLÉS : *Recherche d'information, classification textuelle, représentation de documents, noyau de Fisher, PLSI.*

KEYWORDS: *Information Retrieval, Document Classification, document representation, Fisher kernel, PLSI.*

[†] Ce travail a été financé dans le cadre des projets 200021–111817 et 200020–119745 du Fond National Suisse.

1. Introduction

L'indexation automatique de grandes collections de documents en vue de leur analyse, de leur organisation ou de leur fouille demeure l'un des grands sujets de l'intelligence artificielle moderne. Dans ce contexte, la représentation des documents comme mélanges de catégories latentes s'est avérée prometteuse. Le modèle « Probabilistic Latent Semantic Indexing » (PLSI) (Hofmann, 1999, Hofmann, 2000, Hofmann, 2001) à base de représentations latentes probabilistes a conduit à diverses extensions et applications sur des données textuelles (Vinokourov *et al.,* 2002, Gaussier *et al.,* 2002, Steyvers *et al.,* 2004, Jin *et al.,* 2004, Mei *et al.,* 2006), sonores (Ahrendt *et al.,* 2005), ou graphiques (Monay *et al.,* 2004, Quelhas *et al.,* 2005, Bosch *et al.,* 2006, Monay *et al.,* 2007, Lienhart *et al.,* 2007).

Dans ce contexte, la « similarité cosinus » employée à l'origine, sans justification théorique, pour évaluer la proximité sémantique entre documents, a laissé la place à des similarités à base de noyaux de Fisher, mieux justifiée théoriquement (Hofmann, 2000). Cette approche a ensuite été étendue à plusieurs autres modèles de similarité résumés ci-dessous (Nyffenegger *et al.,* 2006).

Cet article propose une étude approfondie et une révision de ces différents modèles par (1) une description unifiée et simplifiée, (2) une étude du rôle de la matrice d'information de Fisher $G(\theta)$, et (3) une analyse de l'impact des paramètres associés aux catégories latentes. De plus, il fournit de nouveaux résultats expérimentaux sur une grande collection de document provenant du corpus d'évaluation TREC–AP.

La section suivante offre un rappel sur le modèle de documents et les mesures de similarité de PLSI. Le rôle de la matrice d'information de Fisher est ensuite examiné, suivi par une discussion sur les proportions entre les différentes composantes des noyaux. Pour finir, le cadre expérimental est présenté et les principaux résultats sont résumés, avant la conclusion.

2. Modèle de documents et mesures de similarité

2.1. *Le modèle PLSI de documents*

PLSI est un modèle à catégories latentes pour la classification de documents et la recherche d'information (Hofmann, 1999). Il modélise les documents comme des réalisations de tirages aléatoires successifs de couples document–terme (d, w) : itérativement, une catégorie sémantique $z \in Z$ est d'abord choisie, avec une probabilité $P(z)$, puis un terme w et un document d sont choisis avec respectivement des probabilités $P(w|z)$ et $P(d|z)$. Dans PLSI, w et d sont supposés indépendants pour z connu ; la probabilité d'une paire (d, w) s'écrit alors :

$$P(d,w) = \sum_{z \in Z} P(z) \underbrace{P(w|z)P(d|z)}_{d,w \text{ indép. à } z \text{ connu}} .$$

Les paramètres du modèle PLSI sont $\theta = (P(z), P(w|z), P(d|z))$, pour tous les z, w et d du modèle. Ces paramètres s'estiment par l'algorithme Expectation-Maximization (EM) pour une collection de documents C donnée (Hofmann, 1999, Hofmann, 2001).

2.2. *Variantes des noyaux de Fisher pour PLSI*

Rappelons que pour une famille $P(X|\theta)$ de modèles stochastiques paramétrée par θ, le noyau de Fisher fournit une mesure de similarité entre occurrences, lequel, pour deux occurrences X et Y de cette famille au point θ, est défini par

$$K(X,Y) = U_X(\theta)^T \; G(\theta)^{-1} \; U_Y(\theta), \qquad [1]$$

où $U_X(\theta)$ est le gradient par rapport aux paramètres de la log-vraisemblance du modèle : $U_X(\theta) = \nabla_\theta \log P(X|\theta)$, et où la matrice d'information de Fisher $G(\theta)$ est la covariance de $U_X(\theta)$: $G(\theta) = \mathbf{E}_X[U_X(\theta)\; U_X(\theta)^T]$.

Le noyaux de Fisher pour PLSI dérivé par Hofmann (Hofmann, 2000), mesurant la distance entre un document d et une requête q, s'écrit :

$$K^{\mathrm{H}}(d,q) = \sum_z \frac{P(z|d)P(z|q)}{P(z)} + \sum_w \widehat{P}(w|d)\widehat{P}(w|q) \sum_z \frac{P(z|d,w)P(z|q,w)}{P(w|z)},$$

où $\widehat{P}(w|d) = \frac{n(d,w)}{|d|}$, avec $n(d,w)$ le nombre d'occurrences du terme w dans le document d et $|d| = \sum_w n(d,w)$, sa taille (en nombre de termes).

Cette mesure améliore notablement les performances par rapport aux formulations originales qui utilisent la mesure cosinus. Toutefois, la dérivation de K^{H} néglige la contribution de la matrice d'information de Fisher $G(\theta)$,[1] et contient une normalisation par la taille du document $|d|$, dont la justification théorique n'est pas évidente[2] (Nyffenegger *et al.*, 2006). Différentes variantes pour le noyau de Fisher de PLSI sont ainsi introduites :

1) La remarque concernant la renormalisation par la longueur du document conduit au développement d'un noyau de Fisher K^F, non normalisé par $|d|$. Un noyau hybride nommé VS (« *vector space* ») est également proposé, dans lequel la composante K_z^H liée aux catégories sémantiques latentes n'est pas normalisé par $|d|$, mais où la composante K_w^H, liée aux termes, l'est.

2) D'autre part, l'observation concernant la matrice d'information de Fisher conduit au développement d'un noyau dit « *DFIM* » (*Diagonal Fisher Information*

1. Hofmann assimile $G(\theta)$ à la matrice identité par une reparamétrisation justifiée dans le cas des multinomiales ; cependant, PLSI n'est ni une multinomiale, ni dans une famille exponentielle, et $G(\theta)$ peut s'éloigner significativement de la matrice identité dans ce type de cas.
2. Elle pourrait cependant s'expliquer en envisageant le noyau sous l'angle d'un processus stochastique indépendant et identiquement distribué.

Matrix), où les composantes diagonales de la matrice d'information de Fisher $G(\theta)$ sont prises en compte en utilisant l'approximation

$$G(\theta)_{(ii)} \approx \sum_{d \in C} \left(U_d(\theta)_{(i)}\right)^2 . \qquad [2]$$

Par contraste et lorsque c'est nécessaire, on nommera « *IFIM* » (*Identity Fisher Information Matrix*) les noyaux qui ne tiennent pas compte de ces facteurs et font l'hypothèse d'identité pour $G(\theta)$.

Nyffenegger et al. (2006) proposent une étude comparative du noyau de Fisher K^{H}, du noyau VS K^{VS} et de sa version DFIM $K^{\text{DFIM-VS}}$. Le présent article va plus loin en introduisant un certain nombre d'autres variantes : le noyau de Fisher K^{F} ; sa version DFIM, $K^{\text{DFIM-F}}$, qui prend en compte les termes diagonaux de $G(\theta)$ dans le contexte de K^{F} ; et $K^{\text{DFIM-H}}$, la version construite de la même façon à partir de K^{H}. Il étudie aussi indépendamment les composantes liées aux termes et celles liées aux catégories latentes de chacun de ces noyaux.

3. Normalisations des différentes variantes du noyau

En appliquant l'équation [1] au modèle PLSI, on constate que tous les noyaux sont composés de deux termes additifs : l'un reflète la contribution des catégories sémantiques latentes $z \in Z$, et l'autre, la contribution des termes w :[3]

$$K(d,q) = K_z(d,q) + K_w(d,q) = \sum_z k_z(z,d,q) + \sum_w k_w(w,d,q) \qquad [3]$$
$$= \sum_z \frac{U_d(z)\,U_q(z)}{\alpha(z)} + \sum_w \sum_z \frac{U_d(w|z)\,U_q(w|z)}{\gamma(w,z)},$$

en notant $U_d(z)$ la composante de $U_d(\theta)$ qui correspond à $P(z)$, et de la même façon $U_d(w|z)$ pour $P(w|z)$, et où $\alpha(z)$ et $\gamma(w,z)$ sont soit égaux à 1 (cas IFIM), soit représentent les composantes diagonales de $G(\theta)$, estimées par l'équation [2] sur une collection de documents C :

$$\alpha(z) = \sum_{d \in C} U_d(z)^2, \qquad \gamma(w,z) = \sum_{d \in C} U_d(w|z)^2.$$

Le rôle normalisateur de $G(\theta)$ apparaît ici clairement : premièrement, tous les termes indépendants de d s'annulent dans k_w et k_z puisqu'ils se factorisent dans $\alpha(z)$ et $\gamma(w,z)$; deuxièmement, k_z et k_w ont une forme proche de $\frac{a \cdot b}{a^2+b^2+c^2}$, laquelle est majorée par 0.5. Cette forme est exacte lorsque d et q sont tous deux compris dans C ; on a alors, par exemple pour k_z, $a = U_d(z)$, $b = U_q(z)$ et $c = \sum_{\delta \neq d,q} U_\delta(z)$.

3. Les contributions des paramètres $P(d|z)$ se simplifient, puisque pour deux documents d'indices d_1 et d_2 différents, les vecteurs $U_{d_1}(\{P(d|z)\})$ et $U_{d_2}(\{P(d|z)\})$ sont orthogonaux.

Les expression détaillées de k_z et k_w pour les différents noyaux étudiés ici sont données dans le tableau 1.

$k_z(z,d,q)$	IFIM ($G(\theta)=I$)	DFIM
non normalisé (F)	$k_z^{\mathrm{F}}(z,d,q) = \frac{P(d,z)P(q,z)}{P(z)}$	$k_z^{\mathrm{F}}(z,d,q) \cdot \underbrace{\left(\sum_\delta \frac{P^2(\delta,z)}{P(z)}\right)^{-1}}_{\alpha^{\mathrm{F}}(z)^{-1}}$
Hofmann (H)	$k_z^{\mathrm{H}}(z,d,q) = \frac{P(z\|d)P(z\|q)}{P(z)} = \frac{k_z^{\mathrm{F}}(z,d,q)}{P(d)\,P(q)}$	$k_z^{\mathrm{H}}(z,d,q) \cdot \underbrace{\left(\sum_\delta \frac{P^2(z\|\delta)}{P(z)}\right)^{-1}}_{\alpha^{\mathrm{F}}(z)^{-1}}$
« *Vector space* » (VS)	$k_z^{\mathrm{VS}} = k_z^{\mathrm{F}}$	$k_z^{\text{DFIM-VS}} = k_z^{\text{DFIM-F}}$

$k_w(w,d,q)$ toutes versions	
$\widehat{P}(d,w)\widehat{P}(q,w)\sum_z \frac{P(d\|z)P(q\|z)}{P(d,w)P(q,w)}$	$\begin{cases} P^2(z)P(w\|z) & \leftarrow k_w^{\mathrm{F}} \\ \left[\sum_\delta \widehat{P}^2(\delta,w)\frac{P^2(\delta\|z)}{P^2(\delta,w)}\right]^{-1} & \leftarrow k_w^{\text{DFIM-F}} \end{cases}$
$\widehat{P}(w\|d)\widehat{P}(w\|q)\sum_z \frac{P(d\|z)P(q\|z)}{P(d,w)P(q,w)}$	$\begin{cases} P^2(z)P(w\|z) & \leftarrow k_w^{\mathrm{H}} \\ \left[\sum_\delta \widehat{P}^2(w\|\delta)\frac{P^2(\delta\|z)}{P^2(\delta,w)}\right]^{-1} & \leftarrow k_w^{\text{DFIM-H}} \end{cases}$
$k_w^{\mathrm{VS}} = k_w^{\mathrm{H}}$; $k_w^{\text{DFIM-VS}} = k_w^{\text{DFIM-H}}$	

Tableau 1. *Composante catégories $k_z(z,d,q)$ et composante termes $k_w(w,d,q)$ des différents noyaux de Fisher pour PLSI. $\widehat{P}(d,w) = \frac{n(d,w)}{|C|}$. Noter que $k_w^H = \frac{|C|^2}{|d||q|}k_w^F$.*

4. Rapports et proportions entre composantes des noyaux

La renormalisation par $|d|$ dans l'équation d'Hofmann entraîne les proportions suivantes entre K^{H} et K^{VS}, exprimées en fonction de K^{F} :

	K_z	K_w
non normalisé : K^{F}	K^{F}_z	K^{F}_w
Hofmann (2000) : K^{H}	$\frac{1}{P(d)P(q)} K^{\mathrm{F}}_z$	$\frac{\lvert C\rvert^2}{\lvert d\rvert\lvert q\rvert} K^{\mathrm{F}}_w$
« Vector Space » : K^{VS}	K^{F}_z	$\frac{\lvert C\rvert^2}{\lvert d\rvert\lvert q\rvert} K^{\mathrm{F}}_w$

Compte tenu de l'ordre de grandeur typique de la taille de $|C|$ (nombre total d'occurrences de tous les termes dans la collection de documents) par rapport aux valeurs typiques pour $|d|$ et $|q|$ (i.e. $|C|^2 \gg |d||q|$), il est évident que le noyau VS est fortement dominé par sa composante K_w. Ceci est confirmé expérimentalement, comme montré en section suivante.

En ce qui concerne les noyaux K^{F} et K^{H}, on peut remarquer que la log-vraisemblance non normalisée

$$l^{\mathrm{F}}(d) = \sum_w n(d,w) \log \sum_z P(z)P(w|z)P(d|z)$$

et celle utilisée par Hofmann (2000), l^{H}, sont reliées par $l^{\mathrm{H}}(d) = \frac{1}{|d|} l^{\mathrm{F}}(d)$. Lorsque la log-vraisemblance est multipliée par une constante, le noyau de Fisher résultant reste inchangé (cf Eq. 1). Toutefois, si la log-vraisemblance est multipliée par une fonction de d indépendante des paramètres, c'est-à-dire si $l'_d(\theta) = \lambda(d)\, l_d(\theta)$, avec $\nabla_\theta \lambda = 0$, alors

$$G'(\theta) = \mathbf{E}_d \left[(\nabla_\rho l'_d(\rho)) \, (\nabla_\rho l'_d(\rho))^T \right] = \mathbf{E}_d \left[\lambda(d)^2 (\nabla_\rho l_d(\rho)) \, (\nabla_\rho l_d(\rho))^T \right],$$

et

$$K'_\theta(d,q) = \lambda(d)\, \lambda(q) \left(\nabla_\theta l_d(\theta))^T \, G'(\theta)^{-1} \, (\nabla_\theta l_d(\theta) \right),$$

qui, dans le cas général, ne peut pas s'écrire en termes de $K_\theta(d,q)$.

Cependant, si $G(\theta)$ n'est pas prise en compte (IFIM), les deux noyaux sont alors bien en relation directe :

$$K'_\theta(d,q) = \lambda(d)\, \lambda(q)\, K_\theta(d,q).$$

Ainsi, entre les versions IFIM des noyaux K^{F} et K^{H}, on doit avoir : $K^{\mathrm{H}}(d,q) = \frac{1}{|d|\,|q|} K^{\mathrm{F}}(d,q)$. La raison pour laquelle cette relation n'est pas exactement vérifiée dans les formules du tableau 1 vient du fait que ces formules utilisent des hypothèses différentes dans leurs dérivations : Hofmann (2000) postule que $\sum_w \frac{\widehat{P}(w|d)}{P(w|d)} P(w|z) \approx 1$, là où Nyffenegger et al. (2006) postulent que $\sum_w \frac{\widehat{P}(d,w)}{P(d,w)} P(w|z) \approx 1$.

Remarquons toutefois que

$$\sum_w \frac{\widehat{P}(w|d)}{P(w|d)} P(w|z) = P(d)\,\frac{|C|}{|d|}\,\sum_w \frac{\widehat{P}(d,w)}{P(d,w)} P(w|z)\,.$$

Le passage de la première approximation à la seconde transforme donc $|d|/|C|$ en $P(d)$, ce qui explique les formules obtenues (tableau 1).

Il y aurait ainsi trois versions similaires possibles pour les noyaux d'Hofmann : K^{H}, le noyau initialement calculé et décrit plus haut, mais aussi $K^{\mathrm{H}_1} = \frac{|C|^2}{|d|\,|q|} K^{\mathrm{F}}$, qui a le même k_z que K^{H} mais un k_w différent, et $K^{\mathrm{H}_2} = \frac{1}{P(d)\,P(q)} K^{\mathrm{F}}$, qui a le même k_w que K^{H} mais un autre k_z.

Ces trois noyaux ont été comparés expérimentalement : on ne constate aucune différence dans les résultats finaux (le classement des documents), les différences dans les valeurs des scores de similarité étant de l'ordre de 10^{-4}%. Ceci vient de ce que $|d|/|C|$ est de fait un très bon estimateur de $P(d) = \sum_z P(d|z)P(z)$. Pour des raisons de cohérence avec la littérature existante, nous garderons ici K^{H} plutôt que K^{H_1} ou K^{H_2}.

Nous avons donc en tout 14 noyaux différents qui s'expriment tous sous la forme de l'équation [3] et sont résumés dans le tableau 1 : les trois modèles IFIM K^{F}, K^{H}, et K^{VS} ; leurs versions DFIM ; et les (huit) versions de K_w et K_z réparties entre $K^{(\text{DFIM-})\mathrm{F}}$ et $K^{(\text{DFIM-})\mathrm{H}}$. À noter qu'il n'y a pas de K_w ni K_z spécifiques pour $K^{(\text{DFIM-})\mathrm{VS}}$ puisque par construction $K_z^{(\text{DFIM-})\mathrm{VS}} = K_z^{(\text{DFIM-})\mathrm{F}}$ et $K_w^{(\text{DFIM-})\mathrm{VS}} = K_w^{(\text{DFIM-})\mathrm{H}}$.

5. Expériences

Nous avons évalué ces 14 noyaux sur les bases d'évaluation standard de la recherche d'information CACM, CISI, MED, CRAN et TIME provenant de la collection SMART[4]. Nous les avons de plus évalués sur un corpus nettement plus grand, constitué d'une partie du corpus TREC–AP 89 (Harman, 1995), une collection de nouvelles d'agence de l'Associated Press collectées sur l'année 1989. Pour des raisons pratiques (temps de calcul et taille mémoire), seuls les 7466 premiers documents de la collection et les 50 premières requêtes ont été conservés[5] ; en ce qui concerne les occurrences de termes, cela constitue une base plus de 10 fois plus grande que la plus grosse base de SMART, et près de 5 fois plus grande en terme de nombre de documents. Les caractéristiques principales de ces corpus d'évaluation sont présentées dans le tableau 2.

4. `ftp://ftp.cs.cornell.edu/pub/smart/`
5. Documents AP890101-0001 à AP890131-0311.

	CACM	CRAN	TIME	CISI	MED	AP89_01XX
Nb de termes	4 911	4 063	13 367	5 545	7 688	13 379
$\|C\|$	90 927	120 973	114 850	87 067	76 571	1 321 482
Documents						
Nb	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
Requêtes						
Nb	64	225	83	112	30	50
$\|q\|$ moyen	12.7	8.9	8.2	37.7	11.4	79.3

Tableau 2. *Caractéristiques principales des corpus d'évaluation.*

		CACM	CRAN	TIME	CISI	MED	AP89
Résultats	MAP de BM25	**31.4**	**42.4**	**69.2**	12.3	52.3	19.7
	K_w^{H} MAP	30.0	33.6	55.6	**20.2**	49.8	16.5
	$K_w^{\text{DFIM-H}}$ MAP	23.2	37.0	60.8	15.6	45.5	**21.6**
	Meilleure MAP des noyaux PLSI	30.7	37.6	60.8	**20.3**	**53.8**	**21.6**
	Meilleur noyau PLSI, pour $\|Z\| =$	K_w^{F} 16	$K^{\text{DFIM-H}}$ 64	$K_w^{\text{DFIM-H}}$ 8	K^{VS} 8	K^{H} 32	$K_w^{\text{DFIM-H}}$ 48
Conclusions	noyau PLSI $>$ BM25 ?	Non	Non	Non	**OUI**	**oui**	**oui**
	K_z contribue ?	Non	Non	Non	Non	peu	Non
	DFIM $G(\theta)$ contribue ?	Non	**Oui** (K_w)	**Oui** (K_w)	Non	peu	**Oui** (K_w)

Tableau 3. *Principaux résultats et conclusions des 3024 expériences sur les 14 modèles et 6 corpus.*

Pour les expériences sur les collections SMART, six expériences avec des conditions initiales d'apprentissage différentes ont été effectuées pour chacun des modèles, et pour différents nombres de catégories latentes : $|Z| \in \{1, 2, 8, 16, 32, 64, 128\}$, soit un total de 2940 expériences[6]. Pour le corpus TREC-AP, les expériences n'ont été faites que sur la base d'une seule condition initiale mais pour différents $|Z| \in \{1, 32, 48, 64, 80, 128\}$, soit 84 expériences en tout.

Pour toutes ces expériences, le stemming a été effectué à l'aide du stemmer de Porter de Xapian[7]. Les résultats ont été obtenus grâce à l'outil standard `trec_eval`[8]. Nous utilisons ici la mesure *Mean Average Precision* (MAP) pour les présenter, mais les conclusions se sont avérées être exactement les mêmes avec la précision à 5 points (P5) ou la R-précision.
Toutes les figures (excepté la 5) représentent la MAP en fonction de $|Z|$, avec des barres d'erreur verticales correspondant à un écart type.

Les résultats les plus importants de ces 3024 expériences, résumés dans tableau 3, sont :

1) Comme l'illustre par exemple la figure 1, $\{K^{\text{DFIM-VS}}, K_w^{\text{DFIM-H}}, K^{\text{H}}\}$ (resp. $\{K^{\text{DFIM-F}}, K_w^{\text{DFIM-F}}\}$) se comportent de façon semblable. La raison est que le rôle normalisateur de $G(\theta)$ rend $K_z \ll K_w$ pour les noyaux DFIM.

De plus, $K^{\text{VS}} \simeq K_w^{\text{H}}$ puisque $\frac{|C|^2}{|d||q|} \gg 1$, comme mentionné en section 4. Le modèle VS ne vaut donc pas la peine d'être considéré, puisqu'il imite de très près le comportement de la composante $K_w^{\text{(DFIM-)H}}$.

2) Comme l'illustre par exemple la figure 2, K_z détériore les performances, de façon générale : seul, il donne de mauvais résultats ; de plus, au fur et à mesure que son rôle devient plus important dans K^{H} et K^{F}, quand $|Z|$ s'accroît, les performances de ces noyaux se détériorent : partant de K_w pour une valeur faible de $|Z|$, les performances de K^{H} et K^{F} chutent jusqu'à celles de K_z pour $|Z|$ élevé.

3) K_w pris seul offre toujours de bons résultats, si ce n'est les meilleurs résultats.

4) On observe les mêmes effets sur le corpus TREC-AP corpus, de plus grande taille, comme illustré au bas de la figure 4.

5) Comparés au modèle BM25 (Robertson *et al.*, 1994), qui est l'état de l'art en la matière, les meilleurs noyaux basés sur PLSI donnent de meilleurs résultats sur les corpus les plus difficiles sémantiquement : CISI, où les documents et les requêtes partagent peu de termes, ce qui en fait un échantillon de choix pour évaluer les modèles de recherche robustes à la synonymie ou qui utilisent des catégories latentes[9], MED (vocabulaire spécialisé) et TREC–AP.

---

6. 2940 : 5 corpus, 6 expériences, 7 nombres de catégories latentes et 14 noyaux.
7. `http://xapian.org/`
8. `http://trec.nist.gov/trec_eval/`
9. CISI est remarquable en ceci que certaines requêtes sont supposées extraire des documents avec lesquelles elles ne partagent *aucun* terme significatif.

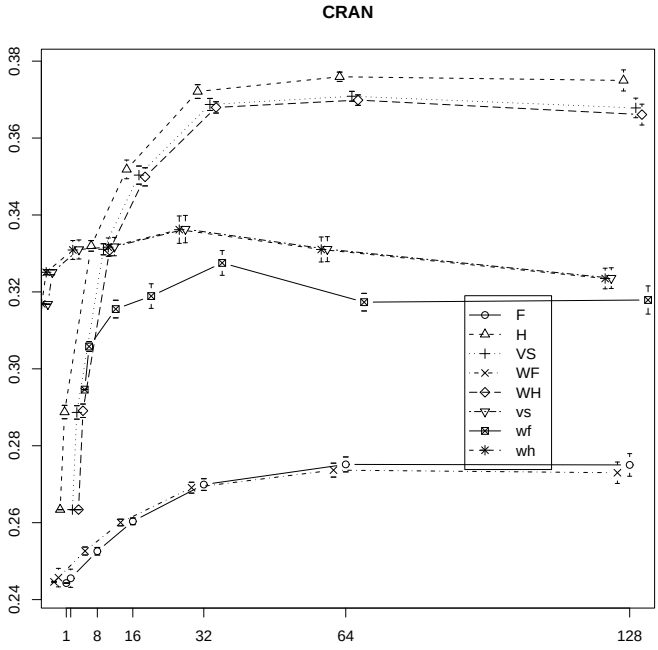

Figure 1. *Résultats typiques illustrés ici par la base CRAN, montrant que* $K^{DFIM\text{-}VS}(VS) \simeq K_w^{DFIM\text{-}H}(WH)$, $K^{VS}(vs) \simeq K_w^{H}(wh)$ *et* $K^{DFIM\text{-}F}(F) \simeq K_w^{DFIM\text{-}F}(WF)$.

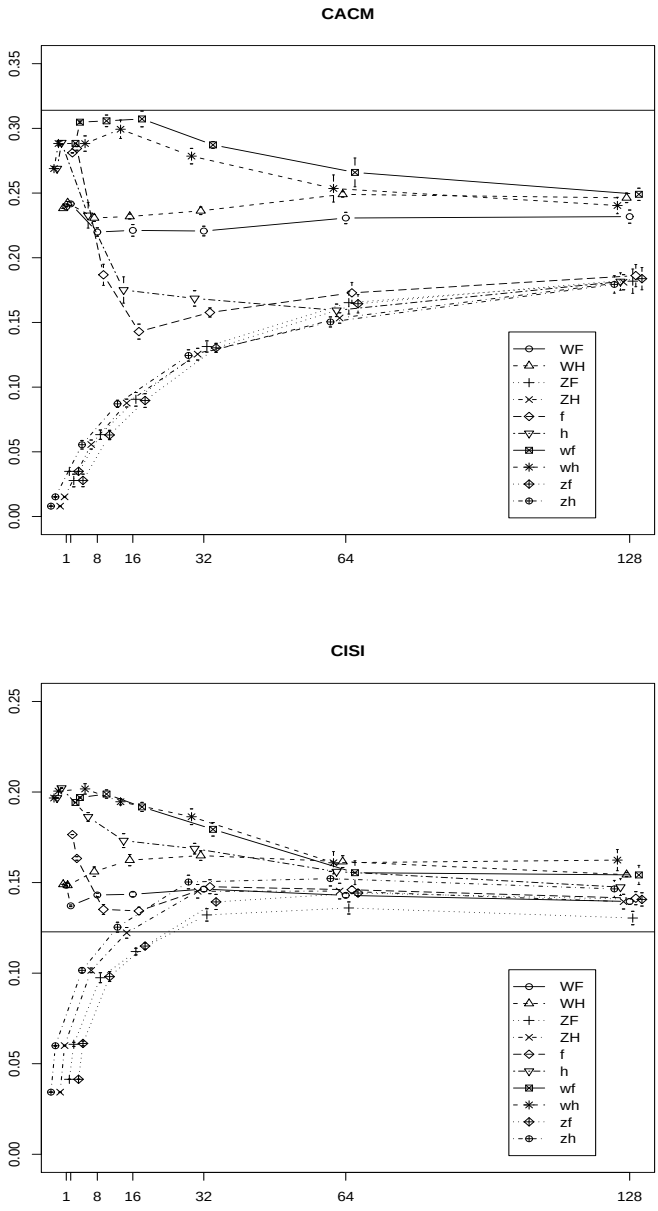

Figure 2. *Résultats obtenus sur les corpus CACM et CISI pour différents modèles :* $K_w^{DFIM\text{-}F}$ *(WF),* $K_w^{DFIM\text{-}H}$ *(WH),* $K_z^{DFIM\text{-}F}$ *(ZF),* $K_z^{DFIM\text{-}H}$ *(ZH),* K^H *(h),* K^F *(f),* K_w^F *(wf),* K_w^H *(wh),* K_z^F *(zf), et* K_z^H *(zh). La barre horizontale représente la performance du modèle BM25, indépendant de* $|Z|$*, qui constitue l'état de l'art. Autres corpus en figures 3 et 4.*

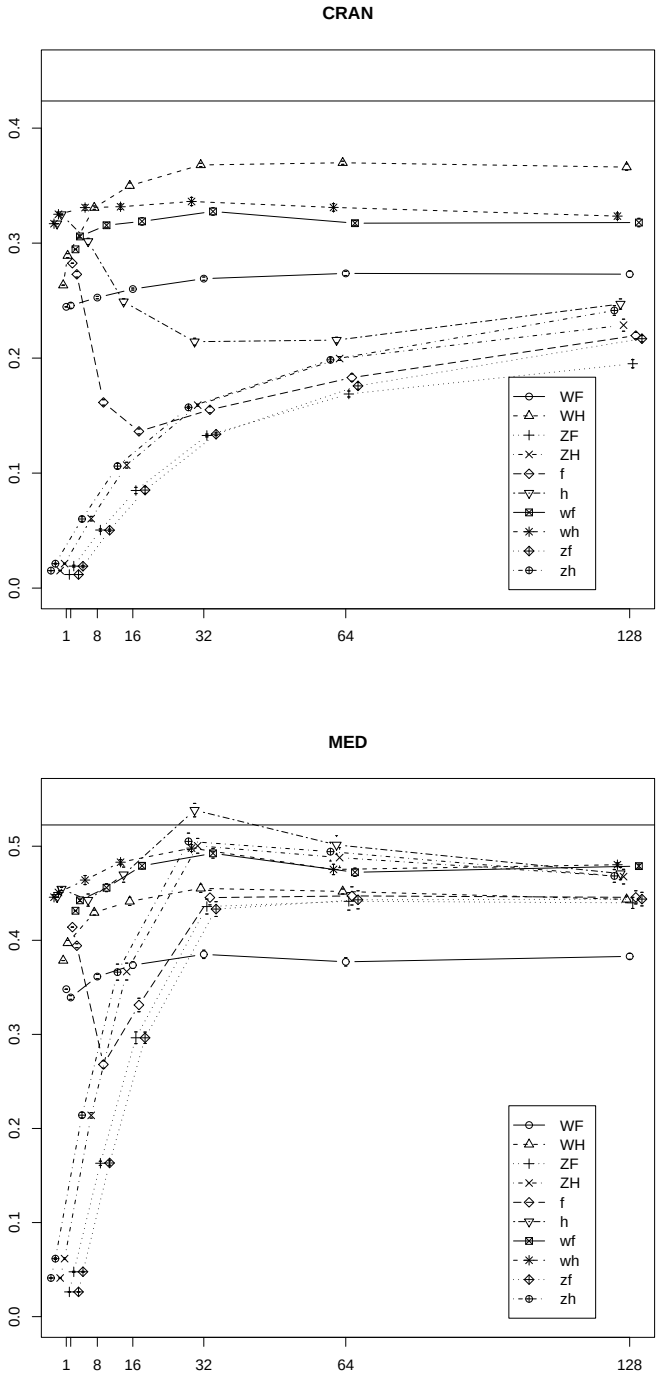

Figure 3. *Résultats obtenus sur les corpus CRAN et MED (légende figure 2).*

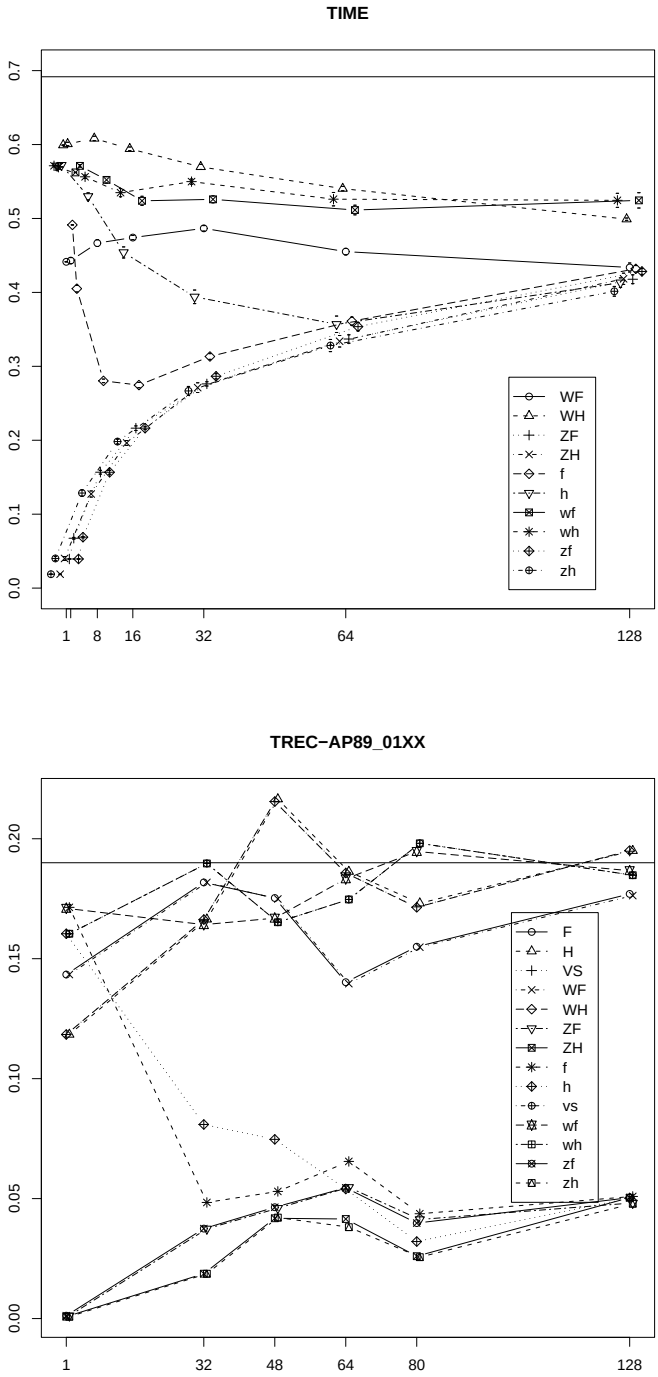

Figure 4. *Résultats obtenus sur les corpus TIME et TREC–AP (légende figure 2).*

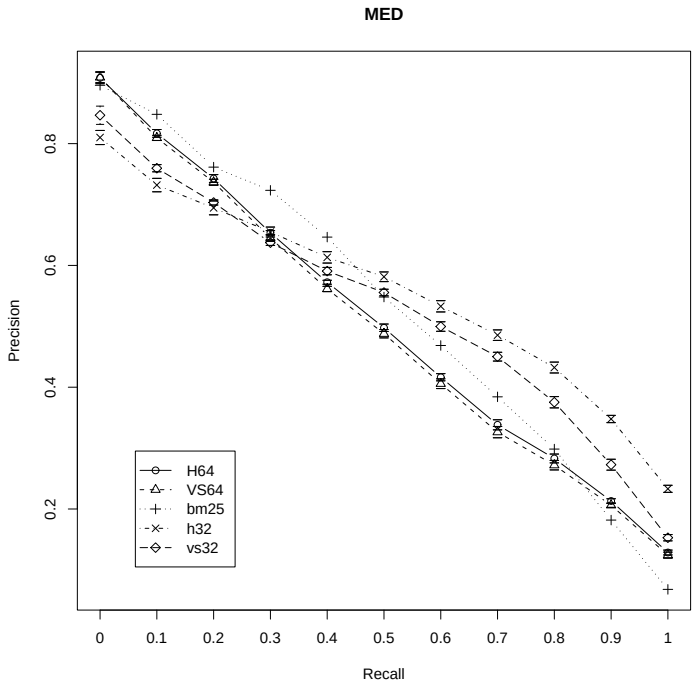

Figure 5. *Courbes précision-rappel sur le corpus MED pour BM25, $K^{DFIM\text{-}H}$, avec $|Z|$=64 (H64), $K^{DFIM\text{-}VS}$, avec $|Z|$=64 (VS64), K^{H}, avec $|Z|$=32 (h32), et K^{VS}, avec $|Z|$=32 (vs32).*

Le seul cas où les conclusions précédentes doivent être nuancées est le corpus MED. Sur ce corpus, les différents noyaux se comportent plus ou moins bien en fonction de la valeur de rappel considérée (Figure 5) : certains sont meilleurs à bas rappel et d'autres à haut rappel. Des mesures globales comme MAP ou R-Prec ne permettent pas de rendre compte de ce genre de nuances.

6. Conclusion

Le présent article analyse différentes dérivations de noyaux de Fisher pour le modèle PLSI. Il se concentre particulièrement sur le rôle de la matrice d'information de Fisher $G(\theta)$ et sur l'importance relative des composantes représentant les contributions des catégories sémantiques latentes et des termes, respectivement.

Nous avons pu confirmer expérimentalement que les modèles à sémantique latente comme PLSI peuvent se révéler intéressants dans des collections de documents sémantiquement difficiles, où documents et requêtes ne partagent pas nécessairement de termes significatifs, si tant est que des capacités de calcul suffisantes sont disponibles.

En ce qui concerne le rôle des composantes « catégories sémantiques latentes » et « termes », K_z peut clairement être négligé — tous du moins pour les nombres de catégories latentes qui permettent une utilisation pratique de tels modèles ($|Z|$ petit). Il est fort possible qu'un nombre beaucoup plus grand de catégories sémantiques latentes ($|Z|$ grand) améliore les performances de K_z, particulièrement pour les plus grandes collections de documents ; mais en pratique, il n'est pas possible d'entraîner de telles configurations, parce que PLSI ne passe pas à l'échelle, justement sur les grandes collections pour lesquelles l'idée serait prometteuse[10].

En ce qui concerne le rôle de la matrice d'information de Fisher, son rôle normalisateur améliore les résultats sur les collections les plus grandes (TIME, CRAN, TREC–AP89).

Globalement, pour les corpus où PLSI pourrait être avantageux par rapport aux modèles standards, il est recommandé d'utiliser $K_w^{\text{DFIM-H}}$ comme mesure de similarité.

7. Bibliographie

Ahrendt P., Goutte C., Larsen J., « Co-occurrence Models in Music Genre Classification », IEEE *Int. Workshop on Machine Learning for Signal Processing*, Sep, 2005.

Bosch A., Zisserman A., Munoz X., « Scene Classification via pLSA », *Proc. of the European Conf. on Computer Vision*, 2006.

10. Il a par exemple fallu 45 heures de temps de processeur et 6.7 Gb de RAM pour exécuter l'apprentissage EM pour la base TREC–AP sur un cœur d'un ordinateur octo-core Intel Xenon à 2 GHz.

Gaussier E., Goutte C., Popat K., Chen F., « A Hierarchical Model for Clustering and Categorising Documents », *Proc. of 24th BCS-IRSG European Colloquium on IR Research*, p. 229-247, 2002.

Harman D., « Overview of the Fourth Text REtrieval Conference (TREC-4) », *Proc. of Forth Text REtrieval Conf. (TREC–4)*, p. 1-23, 1995.

Hofmann T., « Probabilistic Latent Semantic Indexing », *Proc. of 22th Int. Conf. on Research and Development in Information Retrieval*, p. 50-57, 1999.

Hofmann T., « Learning the Similarity of Documents : An Information-Geometric Approach to Document Retrieval and Categorization », *Advances in Neural Information Processing Systems*, vol. 12, p. 914-920, 2000.

Hofmann T., « Unsupervised learning by probabilistic latent semantic analysis », *Machine Learning*, vol. 42, n° 1, p. 177-196, 2001.

Jin X., Zhou Y., Mobasher B., « Web usage mining based on probabilistic latent semantic analysis », *Proc. of 10th Int. Conf. on Knowledge Discovery and Data Mining*, p. 197-205, 2004.

Lienhart R., Slaney M., « PLSA on Large-scale Image Databases », *Proc. of the 2007 Int. Conf. on Acoustics, Speech and Signal Processing, IEEE, (ICASSP'2007)*, vol. 4, p. 1217-1220, 2007.

Mei Q., Zhai C., « A mixture model for contextual text mining », *Proc. of 12th Int. Conf. on Knowledge Discovery and Data Mining*, New York, NY, USA, p. 649-655, 2006.

Monay F., Gatica-Perez D., « PLSA-based Image Auto-Annotation : Constraining the Latent Space », *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, 2004.

Monay F., Gatica-Perez D., « Modeling semantic aspects for cross-media image indexing », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

Nyffenegger M., Chappelier J.-C., Gaussier E., « Revisiting Fisher Kernels for Document Similarities », *Proc. of 17th European Conference on Machine Learning*, vol. 4212 of *Lecture Notes in Computer Science*, Springer, p. 727-734, 2006.

Quelhas P., Monay F., Odobez J.-M., Gatica-Perez D., Tuytelaars T., Gool L. V., « Modeling scenes with local descriptors and latent aspects », *Proc. of ICCV 2005*, vol. 1, p. 883-890, 2005.

Robertson S. E., Walker S., Jones S., Hancock-Beaulieu M., Gatford M., « Okapi at TREC–3 », *Proc. of the Third Text REtrieval Conf. (TREC–3)*, 1994.

Steyvers M., Smyth P., Rosen-Zvi M., Griffiths T., « Probabilistic author-topic models for information discovery », *Proc. of 10th Int. Conf. on Knowledge Discovery and Data Mining*, p. 306-315, 2004.

Vinokourov A., Girolami M., « A Probabilistic Framework for the Hierarchic Organisation and Classification of Document Collections », *Journal of Intelligent Information Systems*, vol. 18, n° 2/3, p. 153-172, 2002.