Fusion de résultats en recherche d'information : application aux documents manuscrits en-ligne

Sebastián Peña Saldarriaga, Emmanuel Morin, Christian Viard-Gaudin

Fusion de résultats en recherche d'information

Application aux documents manuscrits en-ligne

Sebastián Peña Saldarriaga* — **Emmanuel Morin*** — **Christian Viard-Gaudin****

* *LINA - UMR CNRS 6241*
Université de Nantes
** *IRCCyN - UMR CNRS 6597*
École Polytechnique de l'Université de Nantes

{sebastian.pena-saldarriaga, emmanuel.morin, christian.viard-gaudin}@univ-nantes.fr

RÉSUMÉ. *Ce travail présente les résultats d'une étude sur la combinaison des deux approches majeures existantes pour la recherche de documents manuscrits en-ligne. La première approche consiste à appliquer des méthodes de recherche d'information (RI) aux documents issus d'un processus de reconnaissance. La deuxième, quant à elle, ne nécessite pas de reconnaissance explicite et utilise un algorithme de* word spotting. *La fusion permet d'améliorer les performances de la recherche. Les résultats montrent que pour des textes ayant un taux d'erreur au niveau mot inférieur à 23 %, les performances après fusion sont comparables à celles obtenues avec la vérité terrain. De plus, pour des textes fortement dégradés, des améliorations sont également observées.*

ABSTRACT. *This paper explores the fusion of two quite different approaches for retrieving online handwritten documents. The first approach is based on information retrieval techniques carried out on the noisy texts obtained through handwriting recognition, while the second approach is recognition-free using a* word spotting *algorithm. Results show that fusion can bring improvements in retrieval performances. For texts having a word error rate (WER) lower than 23%, the performances obtained with the combined system are close to the performances obtained on clean digital texts. In addition, for poorly recognized texts (WER $>$ 52%), important improvements can also be observed.*

MOTS-CLÉS : *Word spotting, recherche d'information, fusion de résultats, écriture en-ligne*

KEYWORDS: *Word spotting, information retrieval, rank fusion, on-line handwriting*

1. Introduction

Les documents manuscrits en-ligne constituent une nouvelle source d'information, ils résultent de l'émergence des dispositifs de saisie que sont les stylos électroniques couplés à des supports papiers. De cette façon, de véritables documents peuvent être produits, ils peuvent consister aussi bien en des prises de notes, de cours, des copies d'examens, des rédactions d'articles, de dépêches, etc. Cela élargit les champs d'application de la saisie d'écriture en-ligne cantonnés souvent à des terminaux de petites tailles (PDA, smartphone) où seule la reconnaissance des caractères se justifiait.

Il existe deux types de méthodes d'indexation et de recherche pour les documents manuscrits. La première approche consiste à appliquer des méthodes standard en recherche d'information (RI) aux transcriptions obtenues grâce à un moteur de reconnaissance de l'écriture (**?**). Dans ce cas, on parle alors de *RI bruitée* car les transcriptions contiennent des erreurs. La seconde approche évite un processus explicite de reconnaissance et tente d'identifier les mots-clés soumis par un utilisateur dans une distribution donnée par un automate de markov à états cachés (**?**, **?**) ou directement dans le tracé manuscrit (**?**, **?**, **?**, **?**). Dans ce cas on parle de *(key)word spotting*. Ces travaux ne concernent que le domaine manuscrit en-ligne, les travaux sur le manuscrit hors-ligne (**?**), sur les documents historiques (**?**) ou dactylographiés (**?**) étant hors du cadre de cette étude.

Chacune de ces approches possède ses avantages propres. Dans le cas du word spotting on met souvent en avant sa robustesse pour détecter les mots-clés. En revanche il est souvent reproché à ce type de méthodes d'avoir une approche binaire (présence/absence de mots-clés) de la RI (**?**). D'un autre côté, les méthodes standard de RI possèdent de schémas de pondération favorisant des termes en fonction de leur importance. De plus, elles peuvent considérer les variations morphologiques d'un même mot comme une seule entité, dans le cas du word spotting des tentatives pour imiter ce comportement ne sont apparues que très récemment (**?**). Cependant, les performances des méthodes standard risquent d'être pénalisées par une quantité importante d'erreurs de reconnaissance.

En pratique, il est difficile de déterminer *a priori* laquelle de ces approches est la meilleure. Ainsi, afin d'améliorer les performances de la recherche, ces deux approches ont été combinées grâce à des méthodes standard de fusion de résultats de RI. Les résultats de ces expérimentations sont présentés dans la suite de ce document.

Dans une première partie, les méthodes de recherche de documents manuscrits et les méthodes de fusion de résultats seront présentées (cf. sections **??** et **??**). Puis, la section **??** décrit la méthode expérimentale ayant conduit aux requêtes soumises aux différents algorithmes de recherche de documents manuscrits. Enfin, les résultats obtenus avec les différentes méthodes de fusion seront présentés (section **??**), puis discutés dans la section finale.

2. Recherche d'information et word spotting

Cette section présente les différents modèles de référence de RI classique et du word spotting utilisés dans les expériences décrites ici.

2.1. *Modèles de RI*

Étant donnée une requête q et une collection de documents $\mathcal{D}$, un modèle de RI peut être représenté par une fonction $f_q : \mathcal{D} \rightarrow \mathbb{R}$. Le score associé à un document est appelé *retrieval status value* (rsv). Chaque modèle possède une méthode pour calculer l'ensemble $\tau \in \mathbb{R}^n$ des scores d'un ensemble de documents, ces scores déterminent l'ordre de présentation des résultats.

Dans nos expériences, les documents du corpus sont représentés par une matrice $\mathbf{A}$ de $n \times m$ éléments, où n est la taille de la collection, et m celle du lexique. Chaque ligne de cette matrice correspond au vecteur $\mathbf{d}_j$ d'un document. Chaque élément du vecteur correspond à un terme. La valeur associée à cet élément indique son poids dans le document (cf. figure **??**).

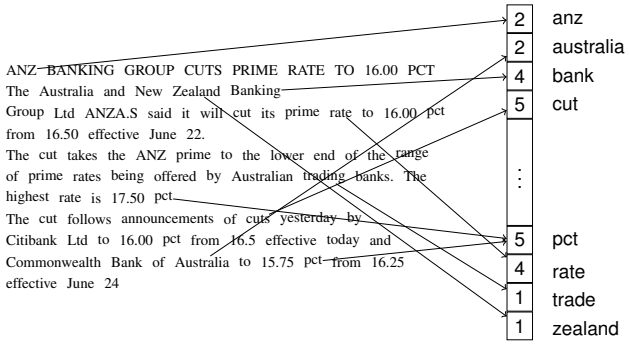


Figure 1. *Représentation vectorielle d'un texte.*

Les termes sont pondérés par la mesure $tf \times idf$ normalisée (**?**). Ainsi, le poids d'un terme i dans le document j est donné par la formule suivante :

$$\mathbf{A}_{j,i} = \frac{\mathrm{tf}(i,j) \times \mathrm{idf}(i)}{\|\mathbf{d}_j\|_2} \quad [1]$$

Avec $\mathrm{tf}(i,j)$ la fréquence du terme i dans le document j. Le facteur $\mathrm{idf}(i)$ est défini à partir du rapport entre la taille du corpus et le nombre de documents contenant le terme i (noté n_i).

$$\mathrm{idf}(i) = \log\left(\frac{n}{n_i}\right) \quad [2]$$

2.1.1. *Modèle vectoriel*

Dans le modèle vectoriel (**?**), la requête est considérée comme un court document. Ainsi q et $\mathcal{D}$ sont réprésentés dans le même espace à m dimensions. Le rsv d'un document est calculé par le produit scalaire de $\mathbf{q}$ et $\mathbf{d}_j$ sur la norme euclidienne des vecteurs. En considérant $\mathbf{A}$ telle que définie précédemment, τ peut être calculé grâce au cosinus de l'angle entre le vecteur de la requête et les lignes de $\mathbf{A}$:

$$\tau = \mathbf{A}\mathbf{q} \quad [3]$$

2.1.2. *Modèle probabiliste*

Le modèle probabiliste (**?**, **?**) estime la probabilité que le document j soit pertinent pour la requête q. Cela revient à remplacer la pondération $tf \times idf$ (équation **??**) par la pondération BM25, couramment appelée formule Okapi :

$$\mathbf{A}_{j,i} = \frac{\mathrm{tf}(i,j) \times (k+1)}{\mathrm{tf}(i,j) + k \times \left((1-b) + b\left(\frac{|d_j| \times n}{\sum_{x=1}^{n} |d_x|}\right)\right)} \times \mathrm{idf}(i) \quad [4]$$

Dans la formule **??**, k et b sont des hyperparamètres habituellement fixés à 2 et $0{,}75$ respectivement. L'hyperparamètre b contrôle l'impact de la normalisation par la longueur du document, tandis que k contrôle l'importance donnée à la fréquence des termes. $|d_j|$ est la longueur du document en nombre de termes. En considérant, que $\mathbf{q}$ et les lignes de $\mathbf{A}$ ont été préalablement normalisées, le calcul de τ s'effectue grâce à la formule **??**.

2.1.3. *Modèles de langage*

La modélisation probabiliste du langage est particulièrement appliquée au domaine de la recherche d'information depuis les travaux de (**?**). Un document est représenté par une distribution multinomiale estimée à partir des occurrences des termes, c'est-à-dire son modèle unigramme de langage, θ_d.

Une manière d'estimer θ_d consiste à appliquer le principe de maximum de vraisemblance (maximum likelihood estimate) :

$$\theta_d = \frac{\mathbf{d}}{\|\mathbf{d}\|_1} \quad [5]$$

L'inconvénient de cette estimation est qu'elle n'attribue de probabilité qu'aux termes apparaissant dans le document. Cela pose un problème lorsqu'un terme de la requête n'y apparaît pas. Il existe des techniques de lissage permettant de remédier à ce problème (**?**).

Calculer le rsv revient à mesurer une vraisemblance entre le modèle de la requête θ_q et celui du document θ_d. En choisissant la divergence de Kullback-Leibler comme distance, et selon l'intuition que plus la distance est importante, moins le document est pertinent, le rsv d'un document donné est calculé grâce à la formule suivante (**?**) :

$$\mathrm{rsv}(q,d) = -\sum_{t \in q} p(w|\theta_q) \log p(w|\theta_d) \quad [6]$$

En supposant que $\mathbf{d}_j$ est le vecteur des fréquences des termes, l'équation **??** peut être remplacée par :

$$\mathbf{A}_{j,i} = \frac{\mathrm{tf}(i,j)}{\|\mathbf{d}_j\|_1} \quad [7]$$

De plus, en considérant que le vecteur de la requête $\mathbf{q}$ a été normalisé par sa norme L_1, le calcul de τ s'effectue grâce à la formule suivante[1] :

$$\tau = -(\log \mathbf{A})\mathbf{q}$$

2.2. *Word spotting*

Étant donnée un motif κ, un algorithme de word spotting peut être représenté comme une fonction $f_\kappa : \mathcal{D} \rightarrow \mathbb{R}^n$. Le résultat de cette fonction est l'ensemble Υ des scores associés à chacune des occurrences du motif dans un document. Les expériences menées dans le cadre de ce travail utilisent un moteur industriel de word spotting : InkSearch®.

InkSearch® est in système de recherche de mots-clés dans un tracé en-ligne inclus dans un moteur de reconnaissance industriel (MyScript® Builder[2]).

Lors de la recherche, chaque occurrence du motif recherché se voit attribuer un indice de confiance. Cet indice reflète la similarité entre le motif et l'occurrence. Puisque f_κ est défini de $\mathcal{D}$ dans $\mathbb{R}^n$, le résultat de la fonction ne peut pas être interprété comme un rsv. Comme le word spotting réalise une recherche de motifs, il ne peut calculer de rsv que pour une occurrence d'un motif donné.

L'absence de rsv rend impossible la comparaison des méthodes venant du domaine de la RI et celles venant du domaine de la reconnaissance de formes. En effet, ces dernières doivent être évaluées par leur capacité à retrouver des motifs et non pas des documents. Ainsi, nous proposons de définir un rsv à partir de f_κ en additionnant les scores pour chaque occurrence de chacun des motifs composant la requête.

Soit la requête $q = \{\kappa_1, \kappa_2, \ldots, \kappa_n\}$ composée de n motifs, et Υ^i l'ensemble des scores des m occurrences du motif κ_i dans un document d, le rsv de d peut être calculé grâce à la formule suivante :

$$\mathrm{rsv}(q,d) = \sum_{i=1}^{n} \sum_{j=1}^{m} \Upsilon^i(j) \quad [8]$$

C'est cette formule qui est utilisée dans toutes les expériences présentées ici.

---

1. L'implémentation de cette méthode utilise le Lemur Toolkit : `http://www.lemurproject.org/`
2. `http://www.visionobjects.com/products/software-development-kits/myscript-builder/myscript-inksearch/`

3. Fusion de résultats en RI

Cette section décrit les méthodes de fusion de résultats utilisées dans nos expériences. Seules les méthodes les plus courantes, ne nécessitant pas de données d'entraînement, seront abordées. D'autres méthodes plus complexes existent mais montrent des résultats mitigés selon les corpus utilisés. Pour un état de l'art des méthodes existantes le lecteur peut se référer à (**?**) ou (**?**).

Tableau 1. *Éléments de base pour les méthodes de fusion*

Symbole	Définition
i	Un document
τ	Classement de documents
$\tau\left(i\right)$	Rang du document i
$\omega^{\tau}\left(i\right)$	Score normalisé du document i
$r^{\tau}\left(i\right) = 1 - \frac{\tau(i)-1}{\|\tau\|}$	Score basé sur le rang de i
$b^{\tau}\left(i\right)$	Score de Borda du document i
R	$\left\{\tau_1, \tau_2, \ldots \tau_{\|R\|}\right\}$; Ensemble de classements
$h(i, R)$	Nombre de classements contenant i
$s^{\tilde{\tau}}\left(i\right)$	Score combiné du document i

Les différentes méthodes définies ici se basent sur les éléments décrits dans le tableau **??**. La notation est celle adpotée également dans (**?**).

Les trois premières méthodes considérées se basent sur les scores des documents. Les opérateurs CombSUM et CombMNZ introduits par (**?**) sont basés sur une combinaison linéaire des scores. Dans cette étude, nous proposons également de combiner les scores en prenant la moyenne harmonique (CombHMEAN) qui permet de minorer l'effet des valeurs atypiques élevées.

Les trois dernières méthodes se basent sur les rangs des documents. La méthode de Borda (**?**) est une méthode de vote pondérée issue de la théorie des choix collectifs. Les méthodes de fusion par l'union et l'intersection sont, quant à elles, définies en fonction d'un score dérivé à partir du rang d'un document.

L'union considère tous les documents, tandis que l'intersection ne considère que les documents qui apparaissent dans tous les classements. La formalisation mathématique des différentes méthodes est présentée dans les tableaux **??** et **??**.

4. Méthodologie

Alors qu'il existe beaucoup de corpus de RI dans le domaine du texte électronique, ces ressources sont inexistantes dans le domaine manuscrit en-ligne. Dans une étude antérieure, un corpus de documents manuscrits en ligne à été collecté pour la catégorisation de textes (**?**). Ce corpus est composé de 2 029 dépêches de l'agence Reuteurs

Tableau 2. *Formalisation mathématique des méthodes de fusion basées sur les scores*

Méthode	Formule	
CombSUM	$s^{\hat{\tau}}(i) = \sum_{\tau \in R} s^{\tau}(i)$	[9]
CombMNZ	$s^{\hat{\tau}}(i) = h(i,R) \times \sum_{\tau \in R} s^{\tau}(i)$	[10]
CombHMEAN	$s^{\hat{\tau}}(i) = \frac{\lvert R \rvert}{\sum_{\tau \in R} \frac{1}{\omega^{\tau}(i)}}$	[11]

Tableau 3. *Formalisation mathématique des méthodes de fusion basées sur les rangs*

Méthode	Formule		
Borda	$s^{\hat{\tau}}(i) = \sum_{\tau \in R} b^{\tau}(i)$		[12]
Union	$s^{\hat{\tau}}(i) = \sum_{\tau \in R} r^{\tau}(i)$		[13]
Intersection	$s^{\hat{\tau}}(i) = \begin{cases} 0, & \text{si } h(i,R) \neq \lvert R \rvert \\ \sum_{\tau \in R} r^{\tau}(i), & \text{sinon} \end{cases}$		[14]

reproduites de façon manuscrite. En moyenne les documents comportent 120 mots, soit au total 146 350 mots manuscrits pour environ 21 505 lignes d'écriture. La figure **??** montre quelques exemples issus du corpus manuscrit.

Puisque ce corpus était à l'origine destiné à la catégorisation de textes, il ne possède pas un ensemble de requêtes prédéfini et la liste de documents pertinents correspondant à chacune, il a été adapté automatiquement afin de l'utiliser dans des expériences de RI. Cette méthodologie s'inspire de travaux précédents (**?**) menés sur l'intégralité du corpus Reuters-21578 (**?**).

Le corpus a été divisé de manière aléatoire en deux sous-ensembles de taille égale :

– soit Q : l'ensemble pour générer les requêtes (1 016 documents) ;

– soit T : l'ensemble de test (1 013 documents) ;

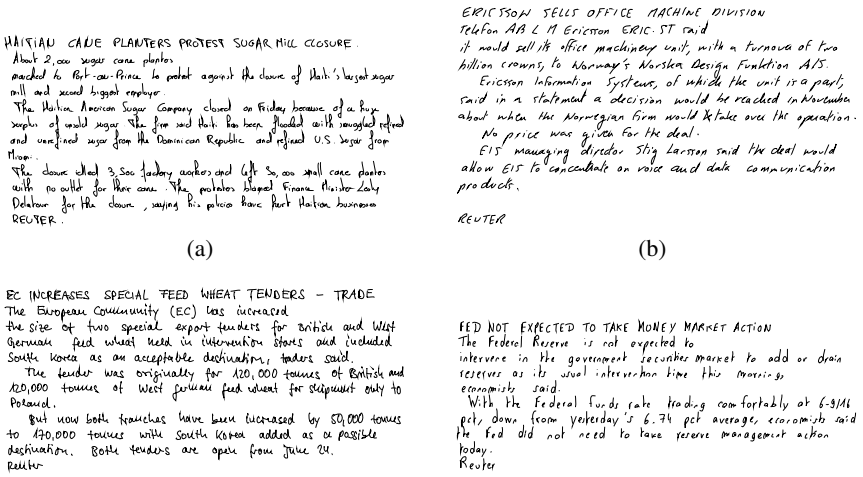


Figure 2. *Échantillons de documents manuscrits issus la base collectée.*

La première étape dans la génération de requêtes consiste à choisir les 100 termes les plus représentatifs de chaque catégorie grâce à la formule suivante :

$$score_{t,c} = \log \frac{p_{t,c} \times (1 - q_{t,c})}{(1 - p_{t,c}) \times q_{t,c}} \quad [15]$$

Il s'agit d'une adaptation de la formule de Robertson-Spärck Jones (**?**). Le score associé à chaque terme t est donné par le risque relatif rapproché[3] entre la probabilité de t dans les documents de la catégorie c ($p_{t,c}$) et la probabilité de t dans les documents des autres catégories ($q_{t,c}$).

Une fois la liste des termes obtenue, la requête prototypique associée à une catégorie donnée est calculée grâce à la formule de contrôle de pertinence Ide-dec-hi (**?**). Les requêtes générées pour chaque catégorie sont indiquées dans le tableau **??**. Les requêtes contiennent au maximum 5 termes. Ces termes sont obtenus après une étape de filtrage de mots outils et de racinisation. L'étape de génération utilise exclusivement la vérité terrain, c'est-à-dire les documents textuels correspondant au tracé manuscrit.

Les requêtes générées sont soumises aux différentes méthodes pour une recherche de documents manuscrits dans l'ensemble T. Les documents qui correspondent à la catégorie de la requête sont considérés comme pertinents. Ainsi, les performances des différentes systèmes et de leur fusion peuvent être mesurées. Dans l'étape de recherche sont utilisés les documents manuscrits ou les transcriptions données par un moteur de reconnaissance.

3. *Odds ratio*

Tableau 4. *Requêtes générées pour chacune des catégories du corpus.*

Catégorie	Requête
Earnings	vs ct net shr loss
Acquisitions	acquir stake acquisit complet merger
Grain	tonn wheat grain corn agricultur
Foreign Exchange	stg monei dollar band bill
Crude	oil crude barrel post well
Interest	rate prime lend citibank percentag
Trade	surplu deficit narrow trade tariff
Shipping	port strike vessel hr worker
Sugar	sugar raw beet cargo kain
Coffee	coffe bag ico registr ibc

Même si les requêtes semblent representatives des différentes catégories, d'un point de vue purement lexical, il est difficile d'estimer dans quelle mesure elles ont un sens d'un point de vue humain. Dans le cadre de ces expériences, il importe avant tout qu'un même ensemble de requêtes soit soumis aux différentes méthodes de recherche.

5. Discussion et résultats

Cette section présente et discute les résultats obtenus après fusion des différentes méthodes de recherche de documents manuscrits en-ligne.

5.1. *Méthodes de référence*

Les méthodes de RI Bruitée se basent sur deux versions transcrites du corpus manuscrit. La première version qu'on pourrait juger de relativement bonne qualité (BQ) enregistre un taux d'erreur au niveau mot de 22,19 %. La deuxième version qu'on pourrait juger de mauvaise qualité (MQ), enregistre quant à elle, un taux d'erreur de 52,47 %.

Les résultats individuels pour chacune des méthodes de référence sont donnés dans le tableau **??**. Les résultats sont présentés en termes de précision moyenne (Mean Average Precision, MAP).

L'impact des erreurs de reconnaissance dans la précision moyenne est évident. Selon la qualité du corpus, une baisse entre 5 % et 20 % peut être observée. Il faut également noter que les performances de InkSearch® ne dépendent pas de la qualité du corpus car cette méthode utilise le tracé manuscrit.

En ce qui concerne les mesures de haute précision (cf. figure **??**), l'impact des erreurs de reconnaissance est moins important que sur la précision moyenne pour les

Tableau 5. *Précision moyenne pour les méthodes de référence. Les colonnes indiquent la version du corpus transcrit, les performances obtenues avec la vérité terrain sont également données.*

	Vérité	BQ	MQ
Cosinus	0,6887	0,6385	0,4980
BM25	0,6989	0,6546	0,5005
LM	0,5589	0,4960	0,4101
IS	-	0,6547	0,6547

Tableau 6. *Précision moyenne après fusion des résultats. Les chiffres en gras indiquent une amélioration des performances tandis que ceux en italique indiquent une dégradation.*

(a) Version BQ

	Cosinus	BM25	LM
CombSUM	**0,6826**	**0,6933**	*0,6361*
CombMNZ	**0,6857**	**0,6933**	*0,6346*
CombHMEAN	**0,6852**	**0,6940**	*0,6393*
Borda	**0,6871**	**0,6933**	*0,6435*
Union	**0,6775**	**0,6785**	*0,6351*
Intersection	*0,6369*	*0,6433*	*0,5177*

(b) Version MQ

	Cosinus	BM25	LM
CombSUM	**0,6782**	**0,6741**	*0,6451*
CombMNZ	**0,6760**	**0,6721**	*0,6428*
CombHMEAN	**0,6710**	**0,6729**	*0,6410*
Borda	**0,6728**	**0,6733**	*0,6472*
Union	**0,6734**	**0,6692**	*0,6411*
Intersection	*0,4958*	*0,4896*	*0,4232*

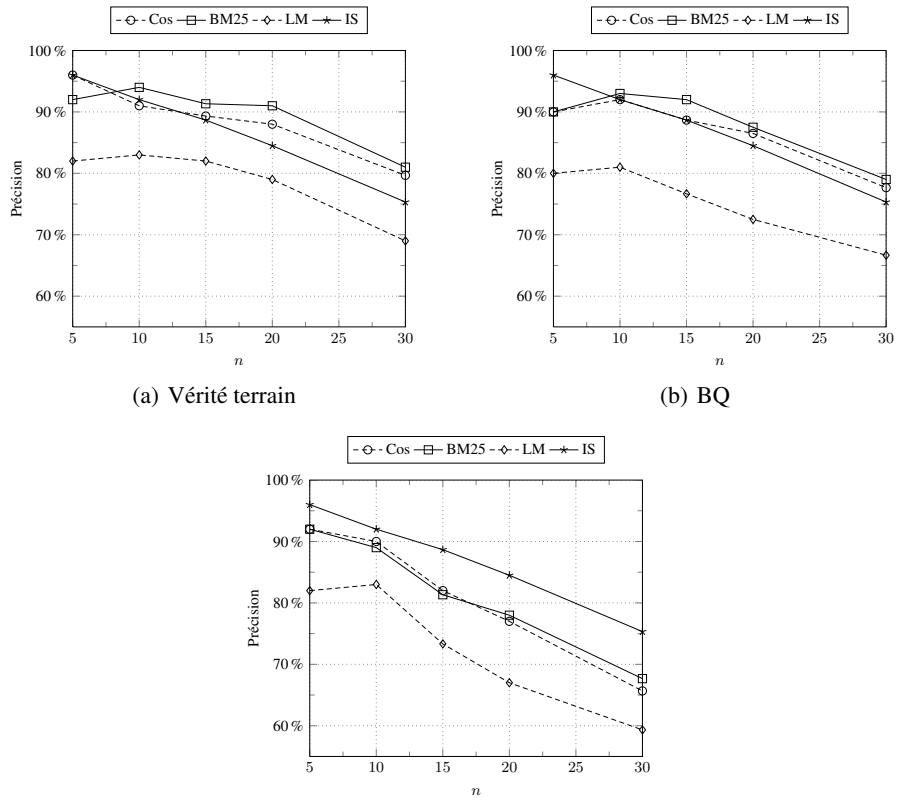
méthodes Cosinus et BM25. Dans le cas de la version BQ du corpus, une différence importante n'est observée qu'à partir de 15 documents. Concernant la version de mauvaise qualité, toutes les méthodes sont inférieures à InkSearch®. LM arrive toujours en dernière position indépendamment de la version du corpus utilisé.

Il est évident que la façon de générer (*absence/présence de mots-clés*) et d'exprimer le besoin d'information (*mots-clés racinisés*) peut favoriser l'approche booléenne du word spotting. Cela n'est pas sans conséquence dans les performances de InkSearch®.

5.2. *Fusion de résultats*

Les mêmes mesures d'évaluation utilisées pour les méthodes de référence, seront utilisées pour la présentation des résultats après fusion avec InkSearch®. La précision moyenne est présentée dans le tableau **??** et la précision à *n* documents dans la figure **??**.

Le tableau **??** montre que la fusion a des effets positifs sur la précision moyenne, à l'exception notable des résultats combinés par *Intersection* ou avec *LM*. Dans toutes



(a) Vérité terrain (b) BQ

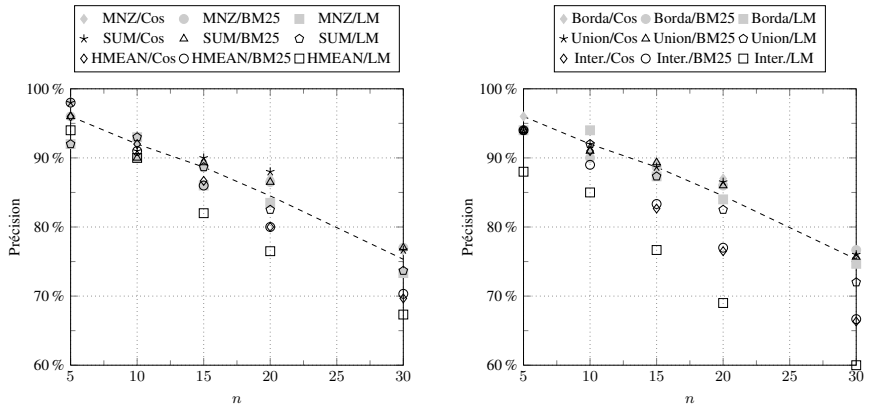
(c) MQ

Figure 3. *Précision à* n *documents pour les méthodes de référence*

les autres configurations la MAP après fusion est supérieure à celle des méthodes de référence prises individuellement. Il faut également noter qu'avec la version BQ du corpus (cf. tableau **??**), les résultats après fusion sont très proches de la MAP obtenues avec la vérité terrain.

Même si l'intersection ne permet pas d'améliorer les performances, elle permet d'estimer dans quelle mesure les documents communs influencent la MAP. En comparant les deux version du corpus, il est observé que la différence entre les résultats fusionnés est de l'ordre de 1 % à 2 % alors que celle entre les résultats individuels peut aller jusqu'à 15 %. Cela veut dire que dans le cas de la version de mauvaise qualité, même si InkSearch® domine le processus de fusion, la recherche dans les documents fortement bruités peut apporter de la pertinence aux résultats du word spotting.

En ce qui concerne les mesures de haute précision, l'impact positif de la fusion de résultats est moins évident. Les graphiques correspondant à la version MQ du cor-



(a) Fusion basée sur les scores/MQ

(b) Fusion basée sur les rangs/MQ

(c) Fusion basée sur les scores/BQ

(d) Fusion basée sur les rangs/BQ

Figure 4. *Précision à* n *documents après fusion. La ligne pointillée donne la courbe de tendance pour IS. Les points au-dessus indiquent une amélioration et les autres l'inverse.*

pus (cf. figures **??** et **??**) montrent qu'aucune configuration n'est systématiquement supérieure ou inférieure à InkSearch®. En revanche, avec la version de bonne qualité, les performances de la fusion avec les résultats de *LM* sont très pauvres comparées à celles de InkSearch®.

Dans toutes les autres configurations avec la version de bonne qualité, les points se retrouvent de manière quasi-systématique au-dessus de la courbe InkSearch®. Enfin, sauf pour $n = 5$, l'intersection a un effet positif sur les performances.

Compte tenu de l'inconstance des améliorations, il reste à analyser quelles sont les conditions nécessaires à la réussite de la fusion. Il a été mis en avant que le taux de chevauchement entre les documents pertinents et non-pertinents (**?**) est un aspect important lorsqu'il s'agit de prédire la fructuosité de la fusion. Pour deux ensembles de

documents τ_1 et τ_2 restitués par deux systèmes de recherche, le taux de chevauchement de documents pertinents (respectivement non pertinents) est calculé grâce à la formule suivante :

$$Chev. = \frac{|\tau_1 \cap \tau_2|}{|\tau_1 \cup \tau_2|} \quad [16]$$

Les taux de chevauchement données dans le tableau **??**, ainsi que les résultats obtenus avec l'intersection, montrent que ces indices ne sont pas des estimateurs fiables des bénéfices attendus après fusion.

Le coefficient de corrélation par rangs de Spearman (ρ) a également été évoqué auparavant (**?**). Il peut être objecté à cette mesure et à d'autres comme le τ de Kendall de ne pas prendre en compte la pertinence des documents dans leurs calculs. Il peut être remarqué dans le tableau **??** que le coefficient de corrélation de Spearman n'est pas un indicateur fiable.

Tableau 7. *Taux de chevauchement des différentes méthodes de référence par rapport à IS.*

(a) BQ

	Cosinus	BM25	LM
Chev. Pert.	95,71 %	95,81 %	78,20 %
Chev. Non-pert.	31,73 %	30,84 %	22,04 %
ρ	0,7259	0,7366	0,7378

(b) MQ

	Cosinus	BM25	LM
Chev. Pert.	78,08 %	76,90 %	66,83 %
Chev. Non-pert.	17,48 %	16,85 %	11,42 %
ρ	0,6439	0,6424	0,6753

Enfin, il a été également suggéré que les différents systèmes doivent avoir des performances similaires pour que la fusion soit efficace (**?**) mais les résultats obtenus avec le corpus MQ montrent que cette affirmation n'est pas toujours vérifiée empiriquement.

6. Conclusion

Cette étude montre les impacts de la fusion de résultats en recherche d'information appliquée au domaine des documents manuscrits en-ligne. Diverses expériences ont été menées sur un sous-ensemble du corpus Reuters-21578. Comme ce corpus n'a pas été créé à l'origine pour la RI ad-hoc, un algorithme d'adaptation a été conçu afin de le conformer aux besoins de ce type d'expériences.

Les conclusions que nous tirons de cette étude ne doivent pas être hâtivement généralisées hors du domaine manuscrit en-ligne, car les expériences décrites ici peuvent présenter plusieurs limites :

– La nature des requêtes. Générées automatiquement et qui peuvent favoriser une approche par absence/présence de mots-clés de la RI ;

– La taille du corpus. Qu'on peut considérer comme petit comparé aux corpus habituellement utilisés en RI (type TREC) ;

– Le nombre de requêtes. Il serait souhaitable de faire fonctionner les différentes méthodes sur un échantillon plus important de requêtes ;

– Le nombre de systèmes fusionnés. Car dans cette étude la fusion s'effectue toujours 2 par 2.

Malgré ces limites, ces premières expériences ont permis d'étudier certains aspects liés à la problématique de la recherche de documents manuscrits en-ligne.

D'une part, l'évaluation des méthodes de référence, se basant sur le résultat d'un moteur de reconnaissance de l'écriture en-ligne, a mis en évidence un impact négatif dans les performances de la recherche d'information. Cet impact est évident lorsque la précision moyenne est mesurée, la différence entre la MAP obtenue avec la vérité terrain et les versions transcrites pouvant aller de 5 % à 20 %. En revanche, cet impact est moins important lorsque des mesures de haute précision sont considérées.

Lorsque les méthodes de recherche sont évaluées individuellement, c'est la méthode de word spotting qui obtient les meilleures performances. Cependant, ces résultats doivent être considérés à la lumière de l'algorithme décrit en section **??**.

D'autre part, lorsque les méthodes de recherche sont combinées, les performances sont améliorées de façon substantielle dans la plupart des configurations étudiées (malgré les nombreuses erreurs de reconnaissance).

Dans un premier temps, cette étude pourrait être complétée en utilisant des requêtes préparées par un humain et dont la pertinence serait déterminée par des juges extérieurs. Il pourrait aussi être intéressant de confirmer ces premiers résultats sur d'autres corpus, mais l'indisponibilité de corpus de RI dans le domaine manuscrit en-ligne reste un frein puissant à cette éventualité.

Enfin, les résultats obtenus suggèrent qu'il y a une relation entre les erreurs de reconnaissance, la degradation des performances en RI bruitée et le bénéfice attendu après fusion qui n'est pas entièrement capturée par les mesures prédictives telles le taux de chevauchement ou les coefficients de corrélation. Des modèles et mesures qui prennent en compte l'influence du bruit dans tous le processus sont souhaitables. Cela constitue une autre voie de recherche pour nos futurs travaux.

Remerciements

Ces travaux ont été soutenus par la Région Pays de la Loire à travers le Projet DEPART et par l'Agence Nationale de la Recherche à travers le programme Technologies Logicielles (ANR-06-TLOG-009). Les opinions exprimées dans cet article n'engagent que les auteurs et ne reflètent pas nécessairement l'avis des mécènes.

7. Bibliographie

Beitzel S. M., Jensen E. C., Chowdury A., Grossman D., Frieder O., Goharian N., « Fusion of effective retrieval strategies in the same information retrieval system », *Journal of the American Society of Information Science & Technology*, vol. 50, n° 10, p. 859-868, 2004.

Borda J. C., *Mémoire sur les élections au scrutin*, Histoire de l'Académie Royale des Sciences, 1781.

Farah M., Vanderpooten D., « An outranking approach for rank aggregation in information retrieval », *SIGIR '07, proceedings of the 30th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 591-598, 2007.

Ide E., *The SMART retrieval system - Experiments in automatic document processing*, Prentice-Hall, Inc., chapter New Experiments in Relevance Feedback, p. 337-354, 1971.

Jain A. K., Namboodiri A. M., « Indexing and retrieval of on-line handwritten documents », *ICDAR 2003, proceedings of the 10th International Conference on Document Analysis & Recognition*, p. 655-659, 2003.

Jawahar C. V., Balasubramanian A., Meshesha M., Namboodiri A. M., « Retrieval of online handwriting by synthesis and matching », *Pattern Recognition*, vol. 42, n° 7, p. 1445-1457, 2009.

Kwok T., Perrone M. P., Russell G., « Ink retrieval from handwritten documents », *IDEAL 2000, Lecture Notes in Computer Science*, vol. 1983, p. 461-466, 2000.

Lafferty J., Zhai C., « Document language models, query models, and risk minimization for information retrieval », *SIGIR 2001, proceedings of the 24th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 111-119, 2001.

Lee J. H., « Analysis of multiple evidence combination », *SIGIR '97, proceedings of the 20th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 267-276, 1997.

Lewis D. D., « An evaluation of phrasal and clustered representations on a text categorization task », *SIGIR '92, proceedings of the 15th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 37-50, 1992.

Lopresti D., Tomkins A., « On the searchability of electronic ink », *IWFHR 1994, proceedings of the 4th International Workshop on Frontiers in Handwriting Recognition*, p. 156-165, 1994.

Marinai S. a., « Font Adaptive Word Indexing of Modern Printed Documents », *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 28, n° 8, p. 1187-1199, 2006.

Peña Saldarriaga S., Morin E., Viard-Gaudin C., « Impact de la reconnaissance de l'écriture en-ligne sur une tâche de catégorisation », *Actes de la 6ème Conférence en Recherche d'Information et Applications (CORIA 2009)*, p. 219-234, 2009.

Ponte J. M., Croft W. B., « A language modeling approach to information retrieval. », *SIGIR 1998, proceedings of the 21st Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 275-281, 1998.

Rath T. M., Manmatha R., « Word spotting for historical documents », *International Journal on Document Analysis & Recognition*, vol. 9, n° 2–4, p. 139-152, 2006.

Renda M. E., Straccia U., « Web metasearch : rank vs. score based rank aggregation methods », *SAC 2003, proceedings of the 18th Annual ACM Symposium on Applied Computing*, p. 841-846, 2003.

Robertson S. E., Spärck Jones K., « Relevance weighting of search terms », *Journal of the American Society for Information Science*, vol. 27, n° 3, p. 129-146, 1976.

Russell G., Perrone M. P., Chee Y.-m., Ziq A., « Handwritten document retrieval », *IWFHR 2002, proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition*, p. 233-238, 2002.

Salton G., Wong A., Yang C. S., « A vector space model for automatic indexing », *Communications of the ACM*, vol. 18, n° 11, p. 613-620, 1975.

Sanderson M., « Word sense disambiguation and information retrieval », *SIGIR '94, proceedings of the 17th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 142-151, 1994.

Schimke S., Vielhauer C., « Document retrieval in pen-based media data », *Proceedings of the 2nd International Conference on Automated Production of Cross Media Content for Multichannel Distribution*, p. 186-190, 2006.

Shaw J. A., Fox E. A., « Combination of multiple searches », *TREC-2, proceedings of the 2nd Text REtrieval Conference*, p. 243-252, 1994.

Spärck Jones K., « Experiments in relevance weighting of search terms », *Information Processing & Management*, vol. 15, n° 3, p. 133-144, 1979.

Spärck Jones K., Walker S., Robertson S. E., « A probabilistic model of information retrieval : development and comparative experiments, part 1 », *Information Processing & Management*, vol. 36, n° 6, p. 779-808, 2000a.

Spärck Jones K., Walker S., Robertson S. E., « A probabilistic model of information retrieval : development and comparative experiments, part 2 », *Information Processing & Management*, vol. 36, n° 6, p. 809-840, 2000b.

Srihari S., Huang C., Harish S., « A search engine for handwritten documents », *Document Recognition & Retrieval XII, proceedings of the SPIE-IS&T Electronic Imaging*, vol. 5676, p. 66-75, 2005.

Vinciarelli A., « Application of information retrieval techniques to single writer documents », *Pattern Recognition Letters*, vol. 26, n° 14, p. 2262-2271, 2005.

Vinciarelli A., « Indexation de documents manuscrits », *Actes du Colloque International Francophone sur l'Ecrit et le Document (CIFED 2006)*, p. 49-54, 2006.

Wu S., McClean S., « Data fusion with correlation weights », *ECIR 2005, Lecture Notes in Computer Science*, vol. 3408, p. 275-286, 2005.

Zhai C., Lafferty J., « A study of smoothing methods for language models applied to ad hoc information retrieval », *SIGIR 2001, proceedings of the 24th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 334-342, 2001.