Modèles de Documents Parcimonieux basés sur les annotations et les " word embeddings " -Application à la personnalisation

Nawal Ould Amer, Philippe Mulhem, Mathias Géry

Modèles de Documents Parcimonieux basés sur les annotations et les "word embeddings" - Application à la personnalisation

Nawal Ould Amer[*,**] — **Philippe Mulhem** [*] — **Mathias Géry** [**]

[*] *LIG - Université de Grenoble, {Nawal.Ould-Amer, Philippe.Mulhem}@imag.fr*
[**] *LaHC - Université de Saint-Étienne, Mathias.Gery@univ-st-etienne.fr*

RÉSUMÉ. *Nous présentons dans cet article des modèles de langues parcimonieux sociaux de documents qui permettent de détecter les termes les plus importants du document et d'éliminer les termes communs ou non significatifs. La détection de ces termes est guidée et renforcée par les liens entre les termes du document et ses annotations sociales (tags). En prenant le contre-pied des approches classiques de personnalisation qui généralement s'intéressent en priorité aux profils utilisateurs ou à la fonction de correspondance, notre proposition porte sur la mise en avant des termes les plus importants des documents afin de mieux personnaliser les réponses. Les évaluations effectuées sur le corpus Social Book Search 2016 montrent que nos propositions apportent dans certains cas une amélioration aux modèles de l'état de l'art.*

ABSTRACT. *In this paper, we define social parcimonious language models that emphasize the most important terms in documents, and lower less important terms. The detection of important terms relies on the document itsef and on the tags that were employed by users to describe the document. Conversely to classical personalization approaches that focus first on user's profiles or on the matching function, our proposal focuses on the documents representations. Evaluation achieved on the Social Book Search 2016 collection show that our proposal outperforms reference approaches in several cases.*

MOTS-CLÉS : *Profil utilisateur, modèles parcimonieux, plongement de mots*

KEYWORDS: *User profile, Parsimonious models, Word Embeddings*

1. Introduction

Cet article s'intéresse à la définition d'un modèle de représentation de documents annotés par des utilisateurs, pour la recherche d'information. L'idée est de proposer une manière de prendre en compte le contenu des documents et ses annotations dans un cadre intégré de recherche d'information (RI). Ceci veut dire que nous gardons à l'esprit que les modèles de documents doivent caractériser les termes qui sont potentiellement utiles pour la RI, notamment par leur pondération.

Notre proposition s'inspire des modèles parcimonieux de RI (Hiemstra *et al.*, 2004), en les utilisant non pas pour la définition des termes utilisés pour un bouclage de pertinence, mais pour adapter les index des documents afin d'y intégrer les annotations faites par les utilisateurs. Cette proposition repose également sur des similarité entre termes et annotations utilisant les plongements de mots[1] (Mikolov *et al.*, 2013). Nous utilisons cette intégration lors d'une étape de réordonnancement des réponses à la suite d'une première recherche classique. Dès lors, il est possible de vérifier si ces nouveaux modèles de langues fournissent de meilleurs résultats que les approches de référence.

Le plan que nous proposons dans cet article est le suivant. Dans un premier temps, nous dressons un état de l'art en section 2 sur les approches à base de modèles parcimonieux ainsi que sur l'utilisation en RI de plongements de mots ("word embeddings") et nous argumentons nos choix. Nous décrivons ensuite le modèle proposé en section 3. Notre proposition est ensuite appliquée à la recherche d'information personnalisée en partie 4. Les expérimentations sont présentées en section 5 et les résultats sont commentés en section 6. Nous concluons en section 7.

2. État de l'art

Les Modèles de Langue Parcimonieux : ML-Parcimonieux

Les Modèles de Langue Parcimonieux (notés ML-Parcimonieux dans la suite) ont été introduits en 2003 dans (Jones *et al.*, 2003), puis implémentés et utilisés dans (Hiemstra *et al.*, 2004) dans le contexte de la recherche d'information. Ces modèles ont pour but de faire une estimation précise et compacte de la distribution des termes dans les documents. Les ML-Parcimonieux se basent initialement sur une estimation, par un modèle de langue standard, de chaque document puis ré-estiment ce modèle afin que les termes non essentiels du modèle soient sous-pondérés (voire éliminés). Plus précisément, l'idée est d'obtenir pour les termes communs du modèle général des valeurs de probabilités qui seront négligées. Les ML-Parcimonieux visent donc à représenter un document avec les termes spécifiques qui permettent de le distinguer des autres documents dans la collection, et en pénalisant les termes qui sont bien représentés dans le modèle général (le modèle de la collection).

1. Nous choisissons d'utiliser dans la suite (ainsi que dans le titre de cet article) l'anglicisme "word embeddings" car il est utilisé tel quel dans la communauté francophone.

La ré-estimation des probabilités des termes de chaque document est effectuée en utilisant un algorithme de Maximisation d'Espérance (*Expectation Maximisation (EM)*), afin d'estimer le modèle de langue parcimonieux, noté θ_d, d'un document d. La probabilité de chaque terme sachant le modèle parcimonieux de langue θ_d, $P(t|\theta_d)$ est calculé par itération des deux étapes suivantes :

$$Etape - E : e_t = tf(t,d) \times \frac{\lambda P(t|\theta_d)}{\lambda P(t|\theta_d) + (1-\lambda)P(t|\theta_C)} \quad [1]$$

$$Etape - M : P(t|\theta_d) = \frac{e_t}{\sum_{t \in V} P(t|\theta_d)} \quad [2]$$

où V est l'ensemble des termes du vocabulaire de la collection, θ_C modèle de langue de la collection (modèle général). A l'initialisation de l'algorithme, la probabilité $P(t|\theta_d)$ est estimée en utilisant un modèle de langue standard.

Dans l'étape E (formule [1]), les termes avec une forte valeur de probabilité dans le document, et ayant une forte valeur de probabilité dans le modèle général, seront pénalisés. A l'issue de chaque itération, leur probabilités seront réduites.

A l'étape M (formule [2]), les probabilités des termes sont normalisées. A l'issue de cette étape, les termes avec des probabilités faibles (inférieures à un seuil prédéfini, généralement 0.0001 (Hiemstra *et al.*, 2004)), seront éliminés du modèle.

Le paramètre λ permet de contrôler le degrés de parcimonie du modèle. Les valeurs faibles de λ donnent un modèle très parcimonieux, c'est-à-dire que beaucoup de termes sont "éliminés" du modèle. Le modèle est estimé après un nombre fixé d'itérations ou jusqu'à convergence.

"Word Embeddings" : Skip-Gram

Les "word embeddings" permettent de représenter sous forme d'un vecteur chaque mot (ou terme) d'un corpus, en utilisant les termes qui apparaissent autour du mot, aussi appelés *contexte*. Cette représentation permet d'identifier les termes qui sont utilisés dans le même *contexte*. Cette technique est couramment utilisée dans le domaine de traitement automatique du langage naturel. Elle est encore relativement peu employée dans le domaine de la recherche d'information. Depuis les travaux de (Mikolov *et al.*, 2013) introduisant la technique de représentation vectorielle des termes appelée *word2vec*, plusieurs travaux se sont intéressés à utiliser cette technique pour la recherche d'information (ALMasri *et al.*, 2010 ; Zamani et Croft, 2016 ; Kuzi *et al.*, 2016), ou la classification (Kim, 2014 ; Balikas et Amini, 2016).

Word2vec est une approche basée sur un apprentissage de représentation vectorielle des termes en utilisant des réseaux de neurones. L'idée principale est : i) de prédire un terme sachant un contexte, ou bien ii) étant donné un terme, prédire son contexte. Plus précisément, les auteurs de *word2vec*, proposent une architecture de réseau de neurones appelée *skip-gram* qui consiste en trois couches : une couche d'entrée, une couche de projection et une couche de sortie pour la prédiction des termes

voisins ou proches. Chaque représentation vectorielle d'un terme est entraînée afin de maximiser les probabilités de ces termes proches du corpus.

Soit une séquence de termes $S = \{t_1, t_2, t_T\}$, l'objectif de l'approche *skip-gram* est de maximiser la moyenne logarithmique comme suit :

$$L(S) = \frac{1}{T} \sum_{i=1}^{T} \sum_{j \in C(i)} \log P(t_j|t_i) \quad [3]$$

où $C(i)$ est le contexte du terme t_i qui est représenté par un ensemble de termes voisins de t_i, $P(t_j|t_i)$ est calculée en utilisant la fonction *softmax* :

$$P(t_j|t_i) = \frac{exp(\vec{t_j}.\vec{t_i})}{\sum_{tj \in V_e} exp(\vec{t_j}.\vec{t_i})} \quad [4]$$

où $\vec{t_j}$ et $\vec{t_j}$ sont les représentations vectorielles des termes t_i et t_j respectivement et V_e est le vocabulaire du corpus sur lequel le "word embeddings" est calculé.

Une des applications classiques de *word2vec* dans le domaine de la recherche d'information est de calculer la similarité entre deux termes (Kenter et de Rijke, 2015) en se basant sur leur représentation vectorielle. Le modèle généré est capable de trouver les termes les plus proches d'un terme et cette technique a montré de bons résultats en recherche d'information (ALMasri *et al.*, 2010 ; Zamani et Croft, 2016).

Positionnement et application
Comme présenté dans la partie précédente, une des applications des modèles parcimonieux est le modèle de bouclage de pertinence, ou *feedback* (Hiemstra *et al.*, 2004 ; Meij *et al.*, 2008 ; Kaptein *et al.*, 2008). Dans les modèles de langue parcimonieux ou dans les modèles de feedback classique, la requête n'est pas utilisée dans l'estimation du modèle feedback. Le problème majeur est que s'il y a peu, ou pas, de documents pertinents dans l'ensemble de feedback, le modèle peu dégrader les résultats, car dans ce cas il n'utiliserait que des termes non pertinents en supposant qu'ils le sont. Pour pallier au problème, (Tao et Zhai, 2006) propose un modèle régularisé par la requête (Query-regularized mixture model), où les auteurs introduisent la requête dans l'estimation du modèle afin de garantir la sélection des termes relatifs à la requête. De ce travail, on peut souligner que deux points essentiels des modèles parcimonieux et des modèles de feedback sont : la nécessité de sélectionner des termes pertinents, et celle de garder un contrôle guidé par la requête. Nous tenons compte de cette remarque dans nos expérimentations.

D'autre part, l'approche *word2vec* tente de prédire les termes les plus proches d'un terme spécifique : étant donné un terme, le modèle permet de retrouver une liste de termes pondérés par leur similarité avec le terme de référence.

Notre intuition est que la combinaison de ces deux techniques pourrait de manière efficace éliminer les termes les moins importants. Nous proposons d'utiliser les tags d'un document comme étant le point de régularisation en appliquant *word2vec* dans

la sélection des termes proches des tags. Plus précisément, nous combinons les deux approches dans la re-estimation des termes du modèle du document. En effet, notre modèle donne plus d'importance aux termes spécifiques au document et ceux qui sont en relations avec les tags sociaux. Dans ce papier, nous allons répondre aux deux questions de recherches qui sont :

1) ***QR1 :*** *Comment estimer le modèle de langue du document parcimonieux en intégrant les tags et les "word embeddings" ?*

2) ***QR2 :*** *Quel est l'impact du modèle dans la performance des systèmes de recherche d'information ?*

3. Approche

Dans cette section, nous allons répondre à la question de recherche **QR1**. Pour répondre à cette question, nous allons d'abord présenter les différentes notations et ensuite présenter le modèle d'estimation du Modèle de Langue Parcimonieux et Social proposé, noté **MLPS** dans la suite.

3.1. *Notations*

Nous présentons ci-dessous les différentes notations employées dans la suite de l'article :

– $d = \{t_1, t_2, t_3, ..., t_N\}$: un document composé de mots ;

– $TG(d) = \{tg_1, tg_2, tg_3, ..., tg_M\}$: l'ensemble des tags du document d assignés par l'ensemble des utilisateurs de la collection ;

– $D(u)$: ensemble des documents annotés par l'utilisateur u ;

– $LTG(u)$: la liste des tags de l'utilisateur u.

– $PU(u) = \{< tg_1, w_1 >, < tg_2, w_2 >, < tg_3, w_3 >, ..., < tg_K, w_K >\}$: profil de l'utilisateur u, où tg_i est un tag de l'utilisateur u et w_i poids du tags tg_i ;

– θ_d : le modèle de langue du document d ;

– θ_C : le modèle de langue de la collection C ;

– V : le vocabulaire de la collection, basé sur les termes présents dans les documents.

3.2. *Estimation du Modèle de Langue Parcimonieux Social : MLPS*

Soit un document $d = \{t_1, t_2, t_3, ..., t_N\}$ et son ensemble de tags $TG(d) = \{tg_1, tg_2, tg_3, ..., tg_M\}$. En premier lieux, nous estimons le modèle du document θ_d,

où la probabilité de chaque terme t du document est estimée en utilisant le modèle de vraisemblance maximale comme suit :

$$P(t|\theta_d) = \frac{tf(t,d)}{|d|} \quad [5]$$

avec $tf(t,d)$ la fréquence du terme t dans le document d et $|d|$ la taille du document d.
Nous définissons la ré-estimation de la distribution de chaque terme du document en utilisant les tags du document. Le but est d'attribuer une importance plus grande aux termes qui sont en relation avec les tags du document, tout en éliminant les termes avec des probabilités faibles. Pour la ré-estimation de distribution des termes, nous employons une approche de maximisation d'espérance, où la nouvelle probabilité de chaque terme est calculée à chaque itération au travers des étapes suivantes :

$$Etape - E : e_t = tf(t,d) \times P(t|\theta_{d_g}) \times \frac{\lambda P(t|\theta_d)}{\lambda P(t|\theta_d) + (1-\lambda)P(t|\theta_C)} \quad [6]$$

$$Etape - M : P(t|\theta_d) = \frac{e_t}{\sum_{t \in V} P(t|\theta_d)} \quad [7]$$

Dans l'étape E (formule [6]), les termes avec une grande valeur de probabilité dans le modèle de la collection seront pénalisés.

La probabilité $P(t|\theta_{d_g})$ est estimée comme suit :

$$P(t|\theta_{d_g}) = \frac{1}{|TG(d)|} \sum_{tg \in TG(d)} \frac{P(t|tg)}{\sum_{t' \in d} P(t'|tg)} \quad [8]$$

La probabilité $P(t|tg)$ est estimée en calculant le cosinus entre les deux vecteurs de représentations (par "word embeddings") du terme t et du tag tg. La différence de notre proposition avec l'approche classique de parcimonie réside donc dans l'intégration des tags lors de la ré-estimation des probabilités de termes.

4. Application : Recherche d'information personnalisée

Nous étudions maintenant l'apport de notre modèle en l'appliquant à la recherche d'information personnalisée. Pour cela, nous choisissons d'utiliser une approche de ré-ordonnancement des résultats : nous reclassons les documents en fonction de leur similarité avec le profil de l'utilisateur qui pose la requête, par le calcul de score $RSV(PU(u), d)$. Dans notre cas, nous représentons un document par son modèle de langue parcimonieux social présenté en 3.2. Pour la modélisation de l'utilisateur, nous choisissons un modèle classique de l'état de l'art (Bouadjenek *et al.*, 2013 ; Noll et Meinel, 2007 ; Cai *et al.*, 2010 ; Vallet *et al.*, 2010), où l'utilisateur est représenté par un vecteur des tags qu'il a utilisés pour annoter des documents, avec leurs poids respectifs :

$$Pu(u) = \{< tg_1, w_1 >, < tg_2, w_2 >, < tg_3, w_3 >, ..., < tg_K, w_K >\} \quad [9]$$

où dans chaque couple $< tg_i, w_i >$, tg_i représente un tag utilisé par l'utilisateur pour annoter les documents et w_i le poids de ce tags défini comme suit :

$$w_i = \frac{tf(tg, LTG(u))}{|D(u)|} \quad [10]$$

Pour calculer la correspondance $RSV(PU(u), d)$, nous utilisons un calcul de négative de la divergence de Kullback-Leibler (Manning *et al.*, 2008) défini dans notre cas par :

$$P(d||u) = -\sum_{t_i \in PU(u)} P(t_i) \log \frac{P(t_i|u)}{P(t_i|d)} \quad [11]$$

5. Expérimentations

5.1. *Collection de test*

Nous avons utilisé le corpus de données Social Book Search (2016) (Koolen *et al.*, 2016) pour évaluer nos propositions. Ce corpus comporte 2,8 millions de documents (commentaires sur des livres) et 120 requêtes utilisateurs tirées du forum de discussions LibraryThing[2]. C'est donc une collection avec de vraies requêtes utilisateurs, et c'est l'une des raisons qui nous fait l'utiliser. Par ailleurs, cette collection intègre des annotations faites par les utilisateurs sur les ouvrages, ainsi que les catalogues (ouvrages achetés) par les utilisateurs qui posent des requêtes.

5.2. *Modèles de référence*

Nous rappelons que le but de notre proposition est de répondre à notre question de recherche **QR2** pour déterminer si la ré-estimation des probabilités des termes du document améliore le qualité des résultats. De ce fait, nous comparons notre approche à différentes modélisations de documents classiques qui sont :

1) **ML-Standard** : Modèle de Langue Standard, où la probabilité de chaque terme du document est estimée en utilisant le maximum de vraisemblance. Dans ce cas nous n'utilisons donc pas de lissage pour estimer la probabilité des termes ;

2) **ML-JM** : Modèle de Langue avec lissage de Jelinek-Mercer. Ce lissage réalise une combinaison linéaire (avec un paramètre $\lambda_{JM} \in [0, 1]$) entre le modèle de langue standard estimé sur le document et le modèle de langue standard estimé sur tout le corpus ;

3) **ML-Dirichlet** : Modèle de Langue avec lissage de Dirichlet. Ce lissage est assez similaire à Jelinek-Mercer, mais l'impact (via le paramètre $\mu \in R^+$) du modèle estimé sur le corpus dépend également de la taille du document ;

4) **ML-Parcimonieux** : Modèle de Langue Parcimonieux, présenté en section 2.

2. https ://www.librarything.com/

5.3. *Configuration des paramètres*

Paramètres du modèle MLPS

Notre modèle **MLPS** possède un paramètre de parcimonie λ. Nous conduisons des expérimentations suivant chaque valeur de $\lambda \in [0, 1]$ par pas de 0,1. Les résultats obtenus variant peu en fonction des valeurs de λ, nous choisissons une valeur de $\lambda = 0,5$.

Apprentissage du modèle *word2vec*

Nous avons réalisé un apprentissage des représentations vectorielles sur la collection Wikipedia en utilisant le modèle Skip-Gram (Mikolov *et al.*, 2013). Le corpus de Wikipedia est constitué de 20 151 102 documents avec une taille de vocabulaire de 2 451 307 mots. Les paramètres de l'apprentissage sont les suivants : la taille des vecteurs est de 100 dimensions, la fenêtre du contexte est de 8 mots et le nombre d'échantillonnage négatif est de 25. Pour les autres paramètres nous avons gardé les valeurs par défaut présentés dans le papier de référence (Mikolov *et al.*, 2013).

Paramètres des modèles de référence

Pour les modèles de référence nous avons choisi les paramètres par défaut classiques : la valeur $\mu = 2500$ pour le modèle de langue avec lissage de Dirichlet, la valeur de $\lambda_{JM} = 0,15$ pour le modèle de langue avec lissage de Jelinek-Mercer, et $\lambda = 0,5$ pour le modèle de langue parcimonieux.

6. Résultats et discussion

L'objectif de cette partie est de répondre à notre deuxième question de recherche ***QR2*** *: "Quel est l'impact du modèle de langue parcimonieux et social sur les résultats ?*

6.1. *Comparaisons des modèles des documents*

Dans cette partie, nous comparons notre modèle **MLPS** et les modèles de référence ; **ML-Standard**, **ML-JM**, **ML-Dirichlet** et **ML-Parcimonieux**. Nous définissons le ré-ordonnancement entre la requête et le document par le modèle *Kullback-Leibler*. Le modèle de la requête θ_Q est estimé avec le modèle de langue standard ($P(w|\theta_Q) = \frac{tf(w,Q)}{|Q|}$). Les résultats sont reportés dans le tableau 1.

D'après les résultats présentés dans ce tableau, le modèle de langue parcimonieux social MLPS obtient les meilleurs résultats en termes de valeurs de MAP et de MRR. En effet, les résultats en MAP du modèle **MLPS** sont inférieurs de respectivement 14%, 12% et 4% pour les modèles **ML-Dirichlet**, **ML-Parcimonieux** et **MK-JM** respectivement, et en MRR ces diminutions sont respectivement de 16%, 9% et 9% pour les modèles **MK-JM**, **ML-Dirichlet**, **ML-Parcimonieux**. Néanmoins, nous notons que le modèle de langue standard (**ML-Standard**), obtient les mêmes résultats

que le modèle de langue parcimonieux social (**MLPS**) en valeur de MAP mais en revanche obtient des résultats de 12% inférieurs au modèle **MLPS** en valeurs de MRR. Pour les résultats en valeur de P@5, le modèle parcimonieux (**ML-parcimonieux**) obtient les meilleurs résultats où les autres modèles sont presque équivalents. En utilisant des tests de Student bilatéraux pairés, nous n'obtenons aucune valeur p inférieure au seuil de 5%, nous ne pouvons donc pas conclure sur la significativité statistique de ces différences.

Tableau 1 – Comparaison avec les modèles de référence

Modèle	MAP	MRR	P@5
ML-Standard	0,0481 (0%)	0,2009 (-12%)	0,0804 (-5%)
ML-JM	0,0464 (-4%)	0,1907 (-16%)	0,0804 (-5%)
ML-Dirichlet	0,0417 (-14%)	0,2074 (-9%)	0,0824 (-2%)
ML-Parcimonieux	0,0422 (-12%)	0,2072 (-9%)	**0,0843**
MLPS	**0,0484**	**0,2282**	0,0824 (-2%)

Notre hypothèse de base était que l'utilisation des tags devrait permettre de ré-estimer les distributions des termes du document en donnant plus d'importance aux termes qui sont dans le même contexte que les tags du document et ainsi permettre une amélioration des résultats. Les résultats obtenus confortent cette hypothèse. On constate que cette amélioration est très faible.

Afin de mieux comprendre le comportement des résultats, nous avons analysé les résultats requête par requête. Sur l'ensemble des 120 requêtes, nous constatons que, pour 17 requêtes, notre modèle parcimonieux social (**MLPS**) obtient de meilleurs résultats et pour 34 requêtes les modèles dégradent les résultats et pour 50 requêtes les modèles sont équivalents.

L'une des raisons pour lesquelles notre modèle ne permet pas de nette amélioration des résultats peut provenir de la qualité des tags. En effet, pour que le modèle se comporte de façon correcte (et attendue par rapport à notre hypothèse) il faudrait que les tags attribués au document par les utilisateurs soient des termes qui décrivent ou caractérisent le contenu du document. En revanche, si les tags sont des tags erronés, très personnels, ou n'ont aucun sens, le modèle va ré-estimer de façon erronée les distributions des termes et ainsi dégrader les résultats. On remarque que, dans le corpus SBS, pour une majorité des documents les tags sont soit mal orthographiés (*America-fromthestacks*, ...), soit non significatifs (*books about books*, *first edition with stamped "Estate of AR" copyright noticek*, *If you really want to know what inspires me*, ...), soit sont une description du document (*On a NE college campus a young woman is found frozen in a snow bank, Her friends never reported her missing,start great* & then fizzles, ...). Il en résulte que notre approche prend en compte beaucoup d'éléments qui n'ont pas les caractéristiques attendues.

Une autre raison du comportement obtenu est que dans le cas où un document n'est décrit par aucun tag alors notre modèle social proposé se comporte comme un

modèle de langue parcimonieux (**ML-Parcimonieux**). Dans les expérimentations menées, 1582 documents sur 9000 (soit environ 18% des documents traités) n'ont pas de tags associés.

Le modèle de langue parcimonieux social (**MLPS**) permet d'estimer un modèle de langue social car il s'appuie sur les tags attribué par l'ensemble des utilisateurs. Quant aux modèles de langue classiques attribuent une importance aux termes fréquents dans les documents ou tient en compte d'autres paramètres (suivant les lissages de modèles). Cependant, une combinaison des deux modélisations des documents, le modèle de langue parcimonieux social (**MLPS**) et les modèles de langue classiques, devraient améliorer les résultats et tirant avantage de la complémentarité de ces modèles.

Nous étudions une telle combinaison en utilisant pour un document ré-ordonné une combinaison linéaire entre le score de RSV(Q,d) où le modèle du document d est représenté par le **MLPS** et un score de de RSV (Q,d) où le modèle du document d est représenté par l'un des modèles (**ML-Standard, ML-Parcimonieux, ML-Dirichlet et ML-Standard**) comme suit :

$$RSV_{MLPS\&ML-X}(Q,d) = (1-\alpha)RSV_{MLPS}(Q,d) + \alpha RSV_{ML-X}(Q,d) \quad [12]$$

Où X est l'un des modèles (Standard, JM, Dirichlet ou Parcimonieux) étudiés et les RSV sont calculés avec la formule de Kullback-Leibler décrite plus haut. Les résultats des combinaisons sont présentés dans le tableau 2.

Tableau 2 – Combinaison du modèle de langue parcimonieux social MLPS avec les autres modèles de langues. Les valeurs de MAP sont obtenues avec les meilleurs valeurs de α, ce qui représente les meilleurs résultats possibles. Les pourcentages avec le symbole '*' sont les différences relatives avec le modèle MLPS et les pourcentages sans * sont les différences relatives avec le meilleur résultat (en gras).

Modèles	MAP	α
MLPS	0.0484 (-8%) (/ *)	-
MLPS & ML-Standard	0,0511 (-2%) (+5% *)	0,7
MLPS & ML-Dirichlet	0,0491 (-6%) (+1% *)	0,2
MLPS & ML-Parcimonieux	**0,0524** (/) (+7% *)	0,4

Comme nous pouvons le voir dans les résultats obtenus, toutes les combinaisons améliorent les valeurs de MAP par rapport aux résultats du tableau 1. De plus, la combinaison des modèles parcimonieux classiques et les modèles de parcimonieux sociaux améliorent davantage les valeurs de MAP par rapport aux autres modèles, avec un bon équilibre, $\alpha = 0,4$ entre les modèles. Pour les combinaisons avec **ML-Standard** et **ML-JM,** α donne moins d'importance au modèle parcimonieux social, mais davantage lors de la combinaison avec **ML-Dirichlet**. Ce comportement devra être étudié plus précisément dans des travaux futurs.

De cette première série d'expérimentations, nous pouvons conclure que le modèle de langue parcimonieux social (**MLPS**) est capable de ré-estimer les modèles des documents en favorisant les termes importants et représentatifs d'un document. La qualité des tags est un paramètre important dans la ré-estimation des modèles de documents. De plus, la combinaison des modèles de langues classiques avec le modèle de langue parcimonieux social permet d'améliorer davantage les performances des systèmes de recherche d'information.

6.2. *Résultats sur la RI personnalisée*

Dans cette partie, nous appliquons notre modèle sur la RI personnalisée. Pour reclasser les $top - 100$ documents, nous calculons le score entre le modèle utilisateur et le document comme présenté dans la partie 4. Le but de cette série d'expérimentation est vérifier l'avantage procurée par notre hypothèse de départ.

Les résultats du tableau 3 montrent que le modèle que nous avons proposé, **MLPS**, obtient des valeurs plus élevées que les autres modèles en termes de MAP et P@5. Plus précisément, notre modèle **MLPS** dépasse le modèle **ML-Parcimonieux** de 18% et les modèles **ML-Dirichlet** de 17%, **ML-JM** de 10%, et le **ML-Standard** de 9 % en valeur de MAP. Pour les valeurs de valeur de P@5, **MLPS** dépasse le modèle de langue parcimonieux (**ML-Parcimonieux**) de 17%, le modèle de langue avec lissage de Dirichlet (**LM-dirichlet**) de 12,5%, le modèle de langue avec lissage de Jelinek-Mercer (**ML-JM**) de 20% et le modèle de langue standard (**ML-Standard**) de 17%. On remarque cependant que notre proposition a un résultat en MRR plus faible de 5% que le modèle de langue parcimonieux**ML-Parcimonieux**. Les valeurs de MRR se focalisent sur le premier résultat pertinent, alors que les valeurs de MAP et P@5 prennent en compte plusieurs résultats pertinents. Notre proposition favorise donc davantage une meilleure couverture (en terme de rappel) des réponses. Les résultats obtenus permettent donc de conforter notre hypothèse. En effet, notre modèle est capable d'estimer de manière plus précise la distribution des termes importants d'un document par rapport aux autres modèles de l'état de l'art. Cependant, nous pensons que le modèle pourrait obtenir encore de meilleurs résultats si la qualité des tags était meilleure. En effet, tout comme pour les expérimentations de la partie précédente (tableau 1), notre modèle est sensible à la qualité des tags.

De cette deuxième série d'expérimentations nous pouvons conclure que, d'une part, la prise en compte des tags dans l'estimation de termes important des document a un effet positif dans la qualité des résultats, et d'autre part, que la combinaison de l'approche de modèle de langue parcimonieux avec le "word embeddings" permet d'améliorer les résultats. Nous pouvons répondre positivement à la question de recherche **QR2**.

Tableau 3 – Résultats sur la personnalisation

Modèles	MAP	MRR	P@5
ML-Standard	0,0312 (-9%)	0,1252 (-6%)	0,0465 (-17%)
ML-JM	0,0310 (-10%)	0,1206 (-9%)	0,0442 (-21%)
ML-Dirichlet	0,0286 (-17%)	0,1146 (-14%)	0,0488 (13%)
ML-Parcimonieux	0,0282 (-18%)	**0,1334**	0,0465 (-17%)
MLPS	**0,0343**	0,1264 (-5%)	**0,0558**

7. Conclusion

Nous avons proposé dans cet article un modèle de langue parcimonieux social qui permet de représenter un document avec les termes les plus importants en prenant en compte les annotations des utilisateurs. Cette proposition repose sur une extension des modèles de langue parcimonieux de recherche d'information qui intègre ces annotations avec les plongements de mots ("word embeddings").

Nous avons évalué nos propositions sur la collection Social Book Search 2016, et elles améliorent les résultats par rapport aux approches de l'état de l'art testés. De plus, une combinaison de notre modèle avec les modèles de l'état de l'art, qui profite de la complémentarité des approches, améliorent encore davantage les résultats. Nous avons de plus testé notre modèle dans le cas de la recherche d'information classique et dans le cas de la recherche d'information personnalisée et nous avons obtenu de meilleurs résultats que les approches de l'état de l'art dans les deux cas.

Toutefois, notre modèle présente une limite, car notre proposition dépend grandement de la qualité des tags attribués par les utilisateurs. En effet, l'estimation des probabilités des termes du document avec notre modèle est très sensible à la qualité des tags. De fait, si la qualité des tags n'est pas bonne alors notre modèle estime de façon erronée les distributions des termes dans le document. Nous allons étudier à l'avenir des moyens de "nettoyer" les annotations de la collection SBS qui est très bruitée, et nous évaluerons ce travail sur d'autres collections.

Remerciements
Ce travail est soutenu par le projet ReSPIr de la région Auvergne Rhône-Alpes.

8. Bibliographie

ALMasri M., Berrut C., Chevallet J.-P., « A Comparison of Deep Learning Based Query Expansion with Pseudo-Relevance Feedback and Mutual Information », *European Conference on IR Research, ECIR*, Milton Keynes, UK, p. 709-715, 2010.

Balikas G., Amini M., « Multi-label, Multi-class Classification Using Polylingual Embeddings », *European Conference on IR Research, ECIR*, Milton Keynes, UK, p. 723-728, 2016.

Bouadjenek M. R., Hacid H., Bouzeghoub M., « Sopra : A New Social Personalized Ranking Function for Improving Web Search », *ACM SIGIR '13*, p. 861-864, 2013.

Cai Y., Li Q., Xie H., Yu L., « Personalized Resource Search by Tag-Based User Profile and Resource Profile », *WISE 2010*, p. 510-523, 2010.

Hiemstra D., Robertson S., Zaragoza H., « Parsimonious Language Models for Information Retrieval », *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, ACM, New York, NY, USA, p. 178-185, 2004.

Jones K. S., Robertson S., Hiemstra D., Zaragoza H., *Language Modeling and Relevance*, Springer Netherlands, Dordrecht, p. 57-71, 2003.

Kaptein R., Kamps J., Hiemstra D. D., « The impact of positive, negative and topical relevance feedback », *17th Text REtrieval Conference, TREC 2008*, vol. SP 500-277 of *NIST special publication*, NIST, Gaithersburg, MD, USA, November, 2008.

Kenter T., de Rijke M., « Short Text Similarity with Word Embeddings », *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, ACM, New York, NY, USA, p. 1411-1420, 2015.

Kim Y., « Convolutional Neural Networks for Sentence Classification », *CoRR*, 2014.

Koolen M., Bogers T., Gäde M., Hall M., Hendrickx I., Huurdeman H., Kamps J., Skov M., Verberne S., Walsh D., *Overview of the CLEF 2016 Social Book Search Lab*, Springer International Publishing, p. 351-370, 2016.

Kuzi S., Shtok A., Kurland O., « Query Expansion Using Word Embeddings », *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, ACM, New York, NY, USA, p. 1929-1932, 2016.

Manning C. D., Raghavan P., Schütze H., *Chapter 12. Language Models for Informatio Retrieval de Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

Meij E., Weerkamp W., Balog K., de Rijke M., « Parsimonious Relevance Models », *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, ACM, New York, NY, USA, p. 817-818, 2008.

Mikolov T., Chen K., Corrado G., Dean J., « Efficient Estimation of Word Representations in Vector Space », *CoRR*, 2013.

Noll M. G., Meinel C., « Web Search Personalization via Social Bookmarking and Tagging », *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, p. 367-380, 2007.

Tao T., Zhai C., « Regularized Estimation of Mixture Models for Robust Pseudo-relevance Feedback », *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, ACM, New York, NY, USA, p. 162-169, 2006.

Vallet D., Cantador I., Jose J. M., « Personalizing Web Search with Folksonomy-Based User and Document Profiles », *European Conference on IR Research, ECIR*, Milton Keynes, UK, p. 420-431, 2010.

Zamani H., Croft W. B., « Embedding-based Query Language Models », *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, ACM, New York, NY, USA, p. 147-156, 2016.