Correction de césures et enrichissement de requêtes pour la recherche de livres

Romain Deveaud — Florian Boudin — Eric SanJuan — Patrice Bellot

LIA - Université d'Avignon
339, chemin des Meinajariès
84 000 Avignon Cedex 9
prenom.nom@univ-avignon.fr

RÉSUMÉ. *Les livres numérisés accessibles sur Internet constituent une importante source d'information. Néanmoins, la Reconnaissance Optique des Caractères (ROC) introduit parfois des erreurs qui peuvent pénaliser la Recherche d'Information. Dans cet article nous proposons une méthode de correction des césures et nous en analysons l'impact sur une tâche de recherche de livres. Nous décrivons également une série d'expériences sur l'enrichissement de requêtes à partir de mots extraits de Wikipédia. Les résultats obtenus montrent qu'utiliser un grand nombre de mots ainsi qu'une répartition adéquate des poids entre la requête initiale et l'enrichissement apporte une amélioration significative par rapport à l'état de l'art.*

ABSTRACT. *Digitized books are now a common source of information on the Web, however OCR sometimes introduces errors that can penalize Information Retrieval. In this paper we propose a method for correcting hyphenations and we analyse its impact on a standard book retrieval task. We also experiment query expansion with words extracted from the Wikipedia page related to the query. We show that there is a significant improvement over the state-of-the-art when using a large weighted list of words.*

MOTS-CLÉS : *Livres numérisés, césures, enrichissement de requête, Wikipédia.*

KEYWORDS: *Book retrieval, hyphenation, query expansion, Wikipedia.*

1. Introduction

Actuellement, le nombre de livres disponibles au format électronique ne cesse d'augmenter. Cette tendance à la numérisation massive a donné naissance à de grandes bibliothèques numériques. On peut notamment citer Google Books[1], qui propose toujours plus d'ouvrages numérisés disponibles en libre consultation sur Internet, ou encore le projet Gutenberg[2] qui met librement à disposition des livres en différents formats (iPhone, iPad, Kindle...). Le développement de méthodes de Recherche d'Information (RI) spécialisées pour ce type de documents représente un véritable enjeu pour la communauté.

Les ouvrages sont numérisés à l'aide d'un procédé de Reconnaissance Optique des Caractères (ROC) qui transforme les images scannées des pages en texte exploitable par ordinateur. Néanmoins ce procédé est automatique et introduit généralement des erreurs qui peuvent diminuer les performances des systèmes de RI pour certaines requêtes (Taghva *et al.,* 1994). Par exemple, les césures naturellement présentes dans les ouvrages au format papier font partie de ces erreurs qui peuvent handicaper les modèles de RI. Elles sont utilisées pour contrôler les retours à la ligne mais elles sont interprétées comme plusieurs mots lors de l'indexation. De ce que nous savons, aucune étude n'a rapporté de résultats d'expérimentations mesurant l'impact d'une correction de ces césures sur les performances des moteurs de recherche. Nous proposons dans la Section 3.1 une approche simple et performante pour la correction des césures et nous l'évaluons sur une collection de livres numérisés issue de la *Book Track* d'INEX 2009 (Kazai *et al.,* 2009).

Nous nous intéressons dans cet article à des méthodes d'enrichissement de requêtes. Autrement dit, nous ajoutons automatiquement de l'information à la requête formulée par l'utilisateur dans le but d'améliorer les performances du système de recherche de livres. Dans notre cas, nous avons choisi d'utiliser Wikipédia comme source externe d'information en nous basant sur le travail mené par (Koolen *et al.,* 2009). Nous détaillons notre approche dans la Section 3.2 ainsi que les expérimentations et les résultats que nous obtenons dans la Section 4. Ces derniers suggèrent que l'utilisation d'un enrichissement composé d'un grand nombre de mots pondérés apporte une amélioration significative de la précision sur les 10 premiers livres de l'ordre de 10% par rapport à l'approche de Koolen *et al*.

2. Travaux connexes

Dans cet article nous présentons des approches visant à améliorer l'efficacité des moteurs de recherche de livres en utilisant une ou plusieurs ressources externes, dans notre cas Wikipédia, pour enrichir les requêtes. Nous allons détailler dans un premier temps les travaux effectués dans le cadre de la RI dans des livres numérisés, puis nous

1. `books.google.com`
2. `www.gutenberg.org`

nous attarderons dans un second temps sur les différentes méthodes d'enrichissement de requêtes.

2.1. *La Recherche d'Information dans les livres numérisés*

Les activités de numérisation d'ouvrages se sont accélérées au cours des dernières années, donnant naissance à de nombreuses collections libres de droits. La campagne d'évaluation INEX[3] propose notamment une *Book Track* visant à explorer de nouvelles techniques de RI propres aux livres numérisés – ou à adapter des méthodes déjà éprouvées. Plusieurs pistes sont privilégiées comme la recherche de pages ou de passages pertinents, l'exploitation de la structure des ouvrages, ou la recherche d'ouvrages entiers traitant d'un sujet prédéfini (Kazai *et al.,* 2009).

Plusieurs études ont été menées sur la RI dans les livres en général. Certaines cherchent à adapter les modèles existants tandis que d'autres prennent le parti d'exploiter les informations structurelles disponibles. Par exemple, (Wu *et al.,* 2008a) ont exploré la possibilité d'utiliser une combinaison de plusieurs index différents, chacun correspondant à une partie identifiable des livres. Ils ont notamment montré que la table des matières et l'index de fin d'ouvrage sont des éléments importants pour estimer la pertinence d'un livre. De la même façon, (Magdy *et al.,* 2008) montrent que les entêtes et les titres contiennent beaucoup d'informations pertinentes par rapport à leur taille réduite. Enfin, (Wu *et al.,* 2008b) ont présenté un travail reposant sur l'indexation des termes présents dans les index de fin d'ouvrages, dans le but d'améliorer la récupération de pages pertinentes par rapport à une requête utilisateur. Néanmoins nous n'utilisons aucun élément structurel des livres dans les approches présentées dans cet article. Nous nous basons uniquement sur Wikipédia comme source d'informations externe afin de pouvoir enrichir la requête initiale et ainsi élargir le champ de la recherche.

2.2. *Enrichissement de requêtes*

Les requêtes utilisateur sont généralement trop courtes ou incomplètes pour pouvoir décrire précisément un besoin d'information. Une des solutions pour pallier à ce problème est d'enrichir la requête en lui ajoutant des termes. Cela se fait traditionnellement en utilisant une méthode de type *Pseudo-Relevance Feedback* (PRF) dont le principe est de sélectionner les termes directement à partir de la collection cible (Harman, 1992, Xu *et al.,* 1996). Le processus consiste à effectuer une première recherche en utilisant la requête initiale, puis d'extraire les termes qui vont enrichir la requête à partir des N premiers documents renvoyés. Plus récemment, des variations de *Pseudo-Relevance Feedback* ont émergé comme par exemple la possibilité de sélectionner les termes de l'enrichissement à partir d'une collection externe dédiée. (He *et al.,* 2006) ont comparé les deux méthodes, à savoir l'enrichissement à partir de la collection

3. `www.inex.otago.ac.nz`

cible et l'enrichissement à partir d'une collection externe, et ont montré que leur efficacité est similaire malgré le fait que les termes sélectionnés soient majoritairement différents.

Le travail sur lequel nous nous basons est celui de Koolen *et al.* (2009). Ils proposent une approche basée sur un enrichissement de requêtes utilisant Wikipédia comme collection externe, qu'ils appliquent ensuite à la recherche de livres. Des pistes concernant le *Pseudo-Relevance Feedback* à partir de Wikipédia ont déjà été explorées, notamment par (Li *et al.,* 2007) qui proposent une méthode visant à enrichir les requêtes dites « faibles » qui ne permettent pas de récupérer suffisamment de documents pertinents lors de la première étape de recherche effectuée pour le *Pseudo-Relevance Feedback*. Dans le cadre d'une tâche de RI classique, ils montrent une amélioration notable de la précision sur les premiers documents renvoyés.

L'approche de Koolen *et al.* se démarque par le fait qu'un seul document est utilisé pour sélectionner les termes. En effet, à une requête correspond une seule page Wikipédia dans laquelle les termes utilisés pour l'enrichissement vont être extraits. Une page Wikipédia est associée à une requête utilisateur si les mots de son titre se retrouvent dans la requête. Tous les mots de la page Wikipédia en question sont alors classés selon une mesure $tf.idf$, où les mesures idf sont calculées au sein de toute la collection de livres. Les N mots possédant les plus forts $tf.idf$ sont sélectionnés pour être ajoutés à la requête originale. Généralement, un poids plus important est donné aux termes de la requête originale par rapport à ceux de l'enrichissement. Dans leur cas, une méthode de pondération simple est utilisée : un poids N fois plus important est affecté aux termes de la requête originale, où N est le nombre de termes présents dans l'enrichissement. Les livres sont enfin classés en utilisant une approche par Modèle de Langue pour la RI (Ponte *et al.,* 1998) avec un lissage de Dirichlet (Zhai *et al.,* 2001).

Nous reprennons ces idées dans la Section 3.2 en expérimentant d'autres mesures pour la sélection des mots de l'enrichissement. Nous analysons également l'impact de différentes répartitions de poids entre la requête originale et son enrichissement.

3. Approche proposée

Nous utilisons la collection INEX 2009 *Book Track* pour nos expériences. Elle est composée de 50 239 livres numérisés au format XML et de 16 requêtes. Nous utilisons également les jugements de pertinence officiels pour les évaluations ; ceux-ci ont été collectés par 9 juges différents (Kazai *et al.,* 2009).

3.1. *Correction de césures*

La collection présente peu d'erreurs dues à la Reconnaissance Optique des Caractères, mais il y a par contre beaucoup de césures. De plus, le trait d'union représentant le découpage d'un mot est parfois oublié par la ROC. Nous regroupons donc sous

le terme « césure » les mots séparés par un retour à la ligne, marqués ou non par un trait d'union. Ces césures peuvent être problèmatiques lors de l'indexation car les mots concernés sont vus comme deux termes différents et ne sont pas indexés correctement. C'est la raison pour laquelle nous avons décidé de reconstituer ces mots sectionnés. Nous utilisons dans ce but un lexique comprenant 118 221 mots uniques extraits du corpus *English Gigaword*[4]. L'algorithme de correction itère sur tous les couples de lignes successives de tous les livres de la collection. Pour chaque couple, les deux lignes sont découpées en suites de caractères séparées par des espaces. Un mot candidat est alors formé en concaténant la dernière sous-chaîne de la première ligne et la première sous-chaîne de la seconde. Le mot candidat est corrigé si il est présent dans le lexique.

Afin d'évaluer l'impact de cette correction, nous utilisons une approche par Modèle de Langue (ML) avec un lissage de Dirichlet en faisant varier le paramètre μ. L'implémentation de Indri est utilisée tout au long de nos expériences. Indri est un moteur de RI implémentant différents modèles, développé dans le cadre du projet Lemur[5]. Pour l'indexation, nous nous servons de la liste de 418 mots-outils fournie par Lemur ainsi que de l'algorithme de *stemming* de Porter.Dans les *topics* de l'INEX 2009 Book Track, les requêtes utilisateurs sont délimitées par les champs `<title>`. Nous construisons donc les requêtes à partir de ces champs et nous fixons le nombre de livres retournés à 100. La collection contient en tout 613 107 923 lignes parmis lesquelles 37 551 834 (6,125%) ont été modifiées par notre méthode. Les résultats sont présentés dans le Tableau 1.

Modèle	**Collection originale**		**Collection corrigée**	
	MAP	P@10	MAP	P@10
ML, μ = 2500	0.302	0.486	0.304	0.507
ML, μ = 1000	0.299	0.493	0.302	0.507
ML, μ = 0	0.244	0.443	0.243	0.450

Tableau 1. *Résultats de la recherche de livres sur la collection originale et sur la collection corrigée de l'INEX 2009 Book Track, en terme de précision moyenne (MAP) et de précision à 10 (P@10).*

On peut voir que l'amélioration mesurée est relativement faible (de l'ordre de 1%), malgré le nombre important de mots corrigés. La redondance d'information introduite par la grande taille des documents (approximativement 118 000 mots par livre) a tendance à réduire l'impact des erreurs introduites par les mots mal orthographiés.

4. LDC Catalog No. LDC2007T07, disponible à `www.ldc.upenn.edu`
5. `www.lemurproject.org`

3.2. *Enrichissement de requêtes avec Wikipédia*

(Koolen *et al.,* 2009) sélectionnent des termes importants pour l'enrichissement d'une requête à partir d'une unique page Wikipédia traitant du même sujet que la requête. Cette page est sélectionnée en faisant correspondre exactement les mots de la requête avec le titre de la page Wikipédia. Le faible nombre de requêtes présentes dans notre collection ne nous permet pas de procéder de la même façon. Quand une page ne peut pas être sélectionnée directement en faisant correspondre son titre avec les mots de la requête, nous lui associons le meilleur résultat renvoyé par le moteur de recherche de Wikipédia.

Une autre de nos modifications porte sur la répartition des poids entre les deux parties de la requête. En effet, si la pondération utilisée par Koolen *et al.* est efficace lorsque peu de termes sont ajoutés à la requête originale, elle peut rendre l'enrichissement inutile si celle-ci contient un grand nombre de termes. À titre d'exemple, si la requête est enrichie avec 50 termes ceux-ci auront un poids 50 fois moins important que les termes de la requête originale, ce qui limite grandement l'influence de l'enrichissement dans la fonction de classement. Une approche adaptative pour l'équilibrage des poids entre la requête originale et l'enrichissement a déjà été évoquée par (Lv *et al.,* 2009). Cependant la collection sur laquelle les évaluations ont été menées est issue du Web. Elle est principalement composée de documents courts, ce qui ne correspond pas avec notre collection très hétérogène. Nous proposons donc d'expérimenter différentes répartitions de poids fixées arbitrairement.

Dans leur approche, Koolen *et al.* sélectionnent les termes de la page Wikipédia en utilisant une mesure $tf.idf$. Nous pensons que cette mesure n'est peut être pas appropriée à cause de l'idf calculé sur la collection. En effet l'idf favorise les termes rares présents dans les livres ce qui peut mener à un enrichissement de requête très spécifique. Nous voulons précisemment éviter ce phénomène, c'est pourquoi nous avons implémenté une mesure uniquement basée sur la fréquence des termes, l'*entropie*. De plus, les termes extraits de la page Wikipédia n'ont pas tous la même importance par rapport au sujet de la requête. Nous avons donc utilisé les scores obtenus par les méthodes d'extraction ($tf.idf$ ou $entropie$) pour pondérer les différents termes à l'intérieur de l'enrichissement.

4. Expérimentations et résultats

Pour nos expérimentations, nous avons fait varier le nombre N de mots ajoutés dans l'enrichissement de la requête. En effet, Koolen *et al.* avaient observé que les meilleurs résultats étaient obtenus en ajoutant 10 termes, mais ils avaient également noté que l'obtention de cette valeur était certainement due à une particularité de leur système de pondération. Nous utilisons la collection corrigée par la méthode décrite dans la Section 3.1 pour nos expérimentations. À titre de comparaison, nous avons implémenté la méthode de Koolen *et al.* telle que nous l'avons décrite dans la Section 2.2. Le nombre de documents renvoyés pour chaque requête est fixé à 100, les résultats sont présentés dans le Tableau 2.

Méthode	**N = 5**		**N = 10**		**N = 20**		**N = 50**	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
entropie (1 :3)	0.301	0.489	0.346	0.564	0.330	0.529	0.353	0.564
entropie (2 :2)	0.327	0.557	0.348	0.564	0.361	0.592†	**0.363**	**0.593**‡
entropie (3 :1)	0.330	0.564	0.342†	0.564	0.349	0.564	0.347	0.557
tf.idf (1 :3)	0.245	0.479	0.249	0.450	0.257	0.464	0.246	0.486
tf.idf (2 :2)	0.277	0.486	0.290	0.521	0.289	0.140	0.295	0.514
tf.idf (3 :1)	0.310	0.536	0.311	0.543	**0.317**	**0.557**	0.314	0.536
Koolen *et al.*	0.308	**0.550**	**0.321**	0.536	0.301	0.521	0.306	0.507

Tableau 2. *Performances de l'enrichissement de requête avec les N meilleurs mots classés par* tf.idf *ou* entropie*, avec une répartition des poids (X :Y) (*† *: t.test* < 0.05 *;* ‡ *: t.test* < 0.01*). Ces expérimentations ont été effectuées sur la collection corrigée.*

On observe que les meilleurs résultats sont obtenus avec un enrichissement de 50 termes sélectionnés par *entropie* et en répartissant les poids de façon égale entre les termes de la requête originale et ceux de l'enrichissement ($X = Y$). Globalement, les résultats atteints par nos méthodes varient différemment de ceux obtenus par l'approche de Koolen *et al.*, notamment lorsque plus de 10 termes sont ajoutés. Comme nous l'avons dit précédemment, ces différences sont dues au système de pondération qu'ils utilisent et qui favorise largement la requête originale jusqu'à quasiment annuler l'effet des termes ajoutés lorsque ceux-ci sont très nombreux. Ceci est d'ailleurs confirmé par les scores obtenus par notre implémentation de leur méthode. En effet avec 50 termes dans l'enrichissement, les scores sont très proches de la *baseline* présentée dans la Section 3.1. Pour rappel, cette *baseline* n'utilisait que les mots de la requête utilisateur.

On voit également que la mesure $tf.idf$ pour la sélection des termes de l'enrichissement donne de meilleurs résultats avec 20 termes ajoutés. Néanmoins l'utilisation d'une mesure idf favorise les termes spécifiques qui sont donc présents dans peu de livres, c'est pourquoi les performances du système se dégradent quand un poids trop important est donné au termes sélectionnés. Les bons résultats obtenus en utilisant l'*entropie* confirment par ailleurs cette hypothèse. On observe une amélioration significative de la précision sur les 10 premiers livres renvoyés. Les meilleurs résultats sont obtenus avec l'enrichissement maximal et en donnant un poids égal à la requête originale et aux termes issus de Wikipédia.

5. Conclusion

Nous avons proposé une méthode de correction des césures dans l'optique d'améliorer les performances de RI dans les livres. Nous n'avons pas été en mesure de mettre en évidence des différences significatives entre la collection corrigée et la collection originale. Néanmoins nous avons remarqué que l'utilisation de la collection corrigée ne dégrade jamais les performances en termes de précision à 10, pour chacune des requêtes. C'est pourquoi nous pensons qu'un nombre plus important de requêtes serait plus favorable à nos expérimentations, nous prévoyons donc d'étendre cette étude à

l'édition 2010 de la *Book Track* d'INEX quand les relevés de pertinence seront disponibles. L'ajout des requêtes des éditions 2009 et 2010 devrait permettre de former un ensemble plus conséquent d'une centaine de requêtes.

De plus, nous pensons que ces corrections peuvent prendre toute leur importance dans le cas d'une recherche ciblée de pages ou de passages. En effet l'impact de la correction des césures est atténué par le grand nombre de mots présents dans les livres. Or, à l'échelle d'une page, la correction de quelques mots informatifs peut être décisive dans l'estimation de la pertinence d'un passage.

Nous avons également expérimenté des méthodes d'enrichissement de requêtes avec des termes issus de pages Wikipédia. Nous avons montré qu'un système de pondération approprié et adapté à la méthode de sélection des termes pouvait améliorer les résultats. Par ailleurs, nous avons vu que les pondérations entre la requête originale et l'enrichissement sont liées aux méthodes de sélection. Nous avons enfin remarqué que l'utilisation d'une mesure basée sur idf ne semblait pas approriée à la sélection des termes utilisés dans l'enrichissement, contrairement à l'*entropie* qui a montré qu'elle était une mesure intéressante dans le cadre d'une tâche de recherche de livres comme la notre.

6. Bibliographie

Harman D., « Relevance feedback revisited », SIGIR '92, 1992.

He D., Peng Y., « Comparing two blind relevance feedback techniques », SIGIR '06, 2006.

Kazai G., Doucet A., Koolen M., Landoni M., « Overview of the INEX 2009 Book Track », *INEX*, p. 145-159, 2009.

Koolen M., Kazai G., Craswell N., « Wikipedia pages as entry points for book search », WSDM '09, 2009.

Li Y., Luk W. P. R., Ho K. S. E., Chung F. L. K., « Improving weak ad-hoc queries using wikipedia asexternal corpus », SIGIR '07, 2007.

Lv Y., Zhai C., « Adaptive relevance feedback in information retrieval », CIKM '09, 2009.

Magdy W., Darwish K., « Book search : indexing the valuable parts », BooksOnline '08, 2008.

Ponte J. M., Croft W. B., « A language modeling approach to information retrieval », SIGIR '98, 1998.

Taghva K., Borsack J., Condit A., « Results of applying probabilistic IR to OCR text », SIGIR '94, 1994.

Wu H., Kazai G., Taylor M., « Book search experiments : Investigating IR methods for the indexing and retrieval of books », ECIR'08, 2008a.

Wu H., Mary Q., Kazai G., Roelleke T., Mary Q., « Modelling anchor text retrieval in book search based on back-of-book index », *In Proceedings of the SIGIR 2008 Workshop on Focused Retrieval*, 2008b.

Xu J., Croft W. B., « Query expansion using local and global document analysis », SIGIR '96, 1996.

Zhai C., Lafferty J., « A study of smoothing methods for language models applied to Ad Hoc information retrieval », SIGIR '01, 2001.