Indexation de sous-collections pour l'amélioration de la haute précision

Joëlson Randriamparany[*,**]

**IRIT, Institut de Recherche Informatique de Toulouse*
Université Paul Sabatier Toulouse III
118 Route de Narbonne, F-31062 Toulouse Cedex 9, France
***ENI, Ecole Nationale d'Informatique*
Université de Fianarantsoa
BP 1264, CP 301, Fianarantsoa, Madagascar
randriam@irit.fr

RESUMÉ : *Cet article présente une méthode de recherche d'information basée sur une indexation en deux étapes. L'objectif est de trouver si affiner l'indexation et la recherche sur une sous-collection homogène améliore la qualité de l'information recherchée. Nous évaluons l'intérêt d'une telle approche en termes de précision en utilisant les modèles de recherches okapi BM25 et TF-IDF et les collections TREC-7 et TREC-8 ad hoc. Les résultats montrent que cette méthode améliore significativement les hautes précisions au moins sur 44% des requêtes.*

MOTS CLÉS : *Indexation, Recherche d'information, Haute précision.*

ABSTRACT: *This paper presents an information retrieval method based on a two stages indexing. The objective of this work is to analyze the impact of refining indexing and search on a homogeneous sub-collection in the quality of the results. We evaluate the impact of this approach in terms of precision using okapi BM25 and TF-IDF models on TREC-7 and TREC-8 ad hoc collections. The results show that this method improves high precision at least on 44% of the queries.*

KEY-WORDS: *indexing, information retrieval, high precision.*

1. Introduction

Comparer la représentation des documents issue de l'indexation à celle de la requête permet à un Système de Recherche d'Information (SRI) de sélectionner et d'ordonner les documents à restituer à l'utilisateur en réponse à sa requête. Différents modèles ont été proposés dans la littérature pour améliorer la mise en correspondance de la requête et des documents (Salton, 1971), (Robertson et al, 1976), (Deerwester et al., 1990), (Ponte, 1998). Les modèles intègrent une méthode d'indexation dont de nombreuses variantes existent en particulier à travers de la fonction de pondération des termes choisie (TF-IDF, BM25).

Dans cet article, nous étudions l'impact d'une recherche en deux étapes sur les performances d'un système. La technique de réinjection de pertinence utilise aussi une recherche en deux temps (Rocchio, 1971). Selon cette approche, la requête initiale est soumise au système ; l'utilisateur juge la pertinence des documents retrouvés et la requête est reformulée en y ajoutant des termes issus des documents pertinents. Cette nouvelle requête est alors soumise au système. Notre approche est différente. La requête initiale n'est pas modifiée, mais c'est la collection à laquelle elle est soumise qui est modifiée. Ainsi, une première recherche permet de constituer un sous-ensemble de documents potentiellement intéressants. Ce sous-ensemble est alors indexé plus finement pour constituer la nouvelle collection de recherche.

De façon similaire, Champclaux (2008) utilise deux phases de recherche. La première consiste à restituer des documents en utilisant la mesure cosinus ou le modèle okapi BM25 (Robertson, 1994). La deuxième phase exploite le résultat de la première en utilisant la fonction Simrank (Champclaux, 2007). C'est une méthode de calcul de similarité basée sur la théorie des graphes. Les résultats montrent une amélioration de la précision moyenne ou MAP et la P10 de plus de 50% en faveur de OkaSim (Okapi puis Simrank) par rapport à Okapi seul (Champclaux, 2009). Mais cette approche n'est pas adaptée au traitement de grandes collections à cause de la complexité élevée de l'algorithme de l'ordre de $O(max(d,t)^3)$ (où d est le nombre des documents et t est le nombre des termes dans la collection) et qui entraine un temps de calcul trop élévé. Hearst et Pedersen (1996) regroupent les n-premiers documents retrouvés dans diverses classes selon leurs ressemblances. Selon eux, la meilleure classe regroupe au moins 50% de documents pertinents c'est-à-dire qu'au moins la moitié des documents de la classe sont pertinents. Les auteurs comparent ainsi les résultats avant classification avec ceux de la meilleure classe trouvée en se référant au nombre de documents qu'elle contient. Cette combinaison de méthodes améliore significativement la qualité de recherche moyenne (MAP) de l'ordre de 25% (resp. 27% et 31%) pour la précision à 5 (resp. à 10 et à 20) documents.

2. Description de notre approche et collection d'évaluation

L'objectif de la méthode de recherche d'information en deux temps que nous avons définie est d'affiner la recherche sur la collection. Elle se fait d'abord sur la collection entière et ensuite sur une collection plus ciblée et plus petite. La première étape consiste à soumettre la requête à un SRI sur une collection donnée. Une première indexation de la collection est effectuée avec des paramètres et une première recherche permet de restituer un ensemble de documents potentiellement en lien avec la requête. Nous avons limité à 1000 documents retrouvés par requête cette première recherche pour constituer les sous collections. La deuxième étape consiste à effectuer une nouvelle recherche sur le sous-ensemble des documents restitués lors de la première étape. Cet ensemble de documents est donc à nouveau indexé. Une nouvelle recherche est effectuée avec la requête initiale Les deux phases d'indexation et de recherche peuvent utiliser des paramètres différents.

Nous avons évalué cette méthode sur les collections TREC-7 et TREC-8 ad hoc avec les requêtes correspondantes[1]. TREC-7 et TREC-8 partagent les mêmes corpus. La taille totale de la collection est de 1,9 Go et elle est composée de 528 155 documents. Nous avons utilisé la totalité des requêtes des deux campagnes c'est-à-dire les 50 requêtes correspondantes à chacune d'elles. Nous avons utilisé Terrier (Ounis et al. 2006) comme moteur d'indexation et de recherche pendant toutes les expérimentations. Terrier implémente différents modèles d'indexation et de recherche. Pour toutes les expérimentations nous avons utilisé le modèle okapi BM25 et le modèle TF-IDF comme modèle de pondération de termes. Les paramètres de modèles sont détaillés dans le tableau 1.

Modèles de recherche	Paramètres utilisés					Requête
	k1	*k2*	*k3*	*b*	ignore.low.idf.terms	Court : Titre
BM25	1,2	0	8	0,5	FAUX	
TF-IDF	1,2			0,5	FAUX	

Tableau 1 : Détails des paramètres utilisés pendant notre expérimentation pour les deux collections

Afin d'évaluer notre approche, nous utilisons le logiciel trec_eval et les critères suivants : la précision moyenne ou MAP, la précision exacte ou R-Prec ainsi que les hautes précisions ou les précisions à *n* documents retrouvés.

3. Résultats globaux

Nous présentons ici le comportement des mesures d'évaluation sur la totalité des requêtes. Les deux tableaux ci-dessous permettent de comparer les résultats des traitements pour la collection TREC-7. A titre d'exemple, dans le tableau 2, la combinaison (BM25, TF-IDF) indique que BM25 est le modèle de recherche utilisé

[1] http://trec.nist.gov/data/intro_eng.html

lors de la première étape et TF-IDF pour la seconde. Lors de la comparaison, les valeurs des mesures de la première étape serviront comme base de références. Pour les deux tableaux, la colonne A correspond aux résultats de la première phase de recherche tandis que B et C correspondent à ceux de la deuxième. La colonne Comparaison (relative) montre les écarts relatifs entre les valeurs des mesures issues de la première phase et de la deuxième phase de traitements. Nous avons effectué le test de Student pour évaluer la significativité des valeurs avec un seuil de 5%.

	A	B	C	Comparaison (relative)	
	BM25	(BM25, BM25)	(BM25, TF-IDF)	B/A	C/A
MAP	**0,1906**	0,1366	0,1896	**-28,32%**	-0,52%
R-Prec	**0,2463**	0,1707	0,2303	**-30,68%**	-6,50%
P5	**0,468**	0,2971	0,452	**-36,51%**	-3,42%
P10	**0,43**	0,2657	0,446	**-38,21%**	-3,72%
P20	**0,364**	0,1957	0,354	**-46,23%**	-2,75%
P30	**0,31**	0,1819	0,2987	**-41,32%**	-3,65%

Tableau 2 : Résultats sur la collection TREC-7 : référence de base okapi BM25. Les valeurs en gras sont statistiquement significatives.

	A	B	C	Comparaison (relative)	
	TF-IDF	(TF-IDF, BM25)	(TF-IDF, TF-IDF)	B/A	C/A
MAP	**0,1897**	0,1432	0,1894	**-24,51%**	-0,16%
R-Prec	**0,2442**	0,1789	0,2303	**-26,74%**	-5,69%
P5	**0,46**	0,303	0,456	**-34,13%**	-0,87%
P10	**0,434**	0,2788	0,444	**-35,76%**	2,30%
P20	**0,366**	0,2061	0,352	**-43,69%**	-3,83%
P30	**0,31**	0,1919	0,2987	**-38,10%**	-3,65%

Tableau 3 : Résultats sur la collection TREC-7 : référence de base TF-IDF. Les valeurs en gras sont statistiquement significatives.

L'indexation en deux phases n'améliore pas la performance de la recherche en général. En termes de précision moyenne, les diminutions sont de l'ordre de 0,52% et de 0,16% pour les combinaisons (BM25, TF-IDF) et (TF-IDF, TF-IDF) au profit des valeurs de référence. Il en est de même pour la précision exacte et les hautes précisions. Par contre, la méthode en deux étapes favorise la P10 pour les deux combinaisons (amélioration de 3,72% pour (BM25, TF-IDF) et 2,30% pour (TF-IDF, TF-IDF).

Les résultats pour TREC-8 sont du même ordre. L'augmentation de la R-précision est de l'ordre de 2,63% pour la combinaison (BM25, TF-IDF) et de l'ordre de 2,32% pour (TF-IDF, TF-IDF). On trouve aussi une amélioration au niveau de la P10 (resp. P20) de l'ordre de 0,86% (resp. 0,25%) pour (BM25, TF-IDF). Pour (TF-IDF, TF-IDF) seule la P20 est améliorée d'environ 1%.

Une analyse plus fine des résultats montre que les détériorations des hautes précisions sont dues à la régression de position des documents pertinents retrouvés en tête de liste par la première recherche. Ceci peut être dû à la méthode de calcul du poids des termes et de la fréquence des termes dont le pouvoir de discrimination est mal représenté.

Le cas critique se situe dans la deuxième phase avec le modèle okapi BM25. Lors du calcul du score des documents, certaines requêtes ne retournent aucun document. En effet, les fréquences de documents contenant les termes de ces requêtes sont égales ou dépassent la moitié du nombre de documents retrouvés lors de la première phase. Le calcul de l'idf (log [(N-n+0,5)/(n+0,5)] où N est le nombre de documents retrouvés par la première phase et n le nombre de documents contenant le terme de requête) donne alors une valeur nulle ou négative et par suite la similarité entre la requête et le document devient nulle.

4. Analyse par requête

Cette analyse se porte sur des requêtes qui ont connu des améliorations lors de la deuxième étape par rapport à la première au niveau de la précision moyenne. La méthode en deux étapes basée sur TF-IDF pour la deuxième phase améliore la précision moyenne pour 29 requêtes (58%) pour TREC-7 et de 22 requêtes (44%) pour TREC-8. Nous avons analysé plus finement ces requêtes (cf. tableau 4). La lecture des tableaux est similaire à celle définie dans la section précédente.

	A	A'	B	B'	Comparaison (relative)		Nb requêtes
	BM25	(BM25, BM25)	(BM25, TF-IDF)	(TF-IDF, TF-IDF)	B/A	B'/A'	
MAP	0,2035	**0,2039**	0,2224	**0,222**	**9,29%**	**8,88%**	
R-Prec	0,2596	**0,2596**	0,2745	**0,2738**	**5,74%**	**5,47%**	
P5	0,5103	0,5034	0,5379	0,5379	5,41%	6,85%	29
P10	0,4793	**0,4828**	0,5276	**0,5241**	**10,08%**	**8,55%**	
P20	0,3948	**0,3966**	0,4172	**0,4138**	**5,67%**	**4,34%**	
P30	0,323	**0,3241**	0,3483	**0,3483**	**7,83%**	**7,47%**	

	A	A'	B	B'	Comparaison (relative)		Nb requêtes
	BM25	(BM25, BM25)	(BM25, TF-IDF)	(TF-IDF, TF-IDF)	B/A	B'/A'	
MAP	0,215	**0,2259**	0,2392	**0,2489**	**11,26%**	**10,18%**	
R-Prec	0,2626	**0,2745**	0,3087	**0,3169**	**17,56%**	**15,45%**	
P5	0,4727	0,4857	0,4455	0,4571	-5,75%	-5,89%	22
P10	0,4045	0,419	0,4182	0,4286	3,39%	2,29%	
P20	0,3568	0,3643	0,3773	0,3905	5,75%	7,19%	
P30	0,3182	**0,3286**	0,3364	0,3397	**5,72%**	3,38%	

Tableau 4 : Résultats sur la collection TREC-7 puis TREC-8. Les valeurs en gras sont statistiquement significatives. Les valeurs correspondent aux requêtes améliorées.

Nous constatons des améliorations significatives de la précision moyenne (+9,26% pour (BM25, TF-IDF) et 8,88% pour (TF-IDF, TF-IDF)). La précision exacte est également améliorée (+5,74% pour (BM25, TF-IDF) et 5,47% pour (TF-IDF, TF-IDF)). Les améliorations concernent aussi les hautes précisions, par exemple, l'amélioration de la P10 est de 10,08% pour la combinaison (BM25, TF-IDF) et de 8.55% pour la combinaison (TF-IDF, TF-IDF)

Les résultats sont similaires pour la collection TREC-8. Nous constatons une forte amélioration de la précision moyenne et de la précision exacte. De même, la MAP augmente de façon significative (11,26%) ainsi que la R-Prec (17,56%) pour

la combinaison (BM25, TF-IDF). Pour la combinaison (TF-IDF, TF-IDF), l'amélioration de la MAP et R-Prec sont de 10,18% et de 15,45%. Des améliorations concernent aussi les hautes précisions.

5. Conclusions et perspectives

Nous avons constaté que lorsque l'ensemble des requêtes est considérée notre approche amène globalement une détérioration au niveau des performances. D'un autre côté en se basant sur l'analyse par requête, l'indexation en deux temps s'avère performante tant au niveau de la précision moyenne qu'au niveau des hautes précisions sur un certain nombre de requêtes en améliorant significativement les précisions moyennes et les hautes précisions. Notre approche améliore la précision de 58% des requêtes pour TREC-7 et de 44% pour la TREC-8. Ce résultat est intéressant et nous amène à poursuivre notre recherche en cherchant une relation entre la structure de la requête elle-même et le niveau d'amélioration afin de dégager une possibilité d'une règle.

6. Références bibliographiques

G. Salton. *The Smart Retrieval System*, Prentice Hall, Englewood Cliffs, NJ, 1971.

I. Ounis, G. Amati, V. Plachouras, Ben He, C. Macdonald, C. Lioma. « Terrier: A high performance and scalable Information Retrieval platform », *ACM SIGIR Conference on Research and Development*, p. 465-471, 2006.

J.J. Rocchio, Relevance feedback in information retrieval, In: G. Salton (ed.), The Smart retrieval system: experiments in automatic document processing, p. 313-323, 1971.

J. M. Ponté, B. Croft, « A langage modeling approach to information retrieval », *ACM SIGIR, Conference and Research and Development in Information Retrieval*, p. 275-281, 1998.

M.A Hearst, J.O. Pedersen. « Reexamining the Cluster Hypothesis: Scatter/Gather on retrieval results ». *ACM SIGIR*, p. 330-337, 1996.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman. « Indexing by Latent Semantic Analysis ». *JASIST*. 41(6), p. 391-407, 1990.

S. E. Robertson, S. Walkery, M. Beaulieu. Okapi at TREC 7: automatic ad hoc, filtering, VLC and interactive track, 1998.

S. E. Robertson, Sparck Jones K.. « Relevance weighting of search terms ». *Journal of the American Society for Information Sciences, 27 (3)*, p 129-146, 1976.

Y. Champclaux, T. Dkaki, J. Mothe, « Enhancing high precision using structural similarities », *IADIS International Conference WWW/Internet*, p. 494-498, 2008.

Y. Champclaux, T. Dkaki, J. Mothe, « Enhancing high precision by combining okapi BM25 with structural similarity in an information retrieval system », *International Conference on Enterprise Information Systems (ICEIS 2009)*, p. 279-285, 2009.