Fusion de systèmes pour la recherche de passages dans des textes

Désiré Kompaoré — Emmanuel LeMoing* — Josiane Mothe*, ****

** ERT34, Institut Universitaire de Formation des Maîtres, 56 Avenue de l'URSS, 31400 Toulouse, France*

*** Institut de Recherche en Informatique de Toulouse, 118 Route de Narbonne, 31062 Toulouse Cedex 04, France*

{kompaore, mothe}@irit.fr

RÉSUMÉ *Les systèmes de recherche d'information (RI) présentent une grande variabilité dans la liste des documents qu'ils retrouvent en réponse à une même requête. Dans cet article nous étudions l'apport de la fusion des résultats des systèmes pour la RI. Ainsi, nous utilisons les résultats obtenus sur un ensemble de 50 requêtes par différents systèmes qui ont participé à la tâche TREC de sélection de phrases pertinentes. Nous étudions la fusion par union et par intersection des résultats des systèmes, et nous montrons que la fusion aveugle apporte des améliorations peu sensibles. Nous étudions également une fusion plus " intelligente " des systèmes qui prend en considération la difficulté des requêtes et la variabilité des réponses des systèmes.*

ABSTRACT *In this paper, we study different way of merging information retrieval system responses. Different methods are applied: while "blind" union and intersection does not improve the final results, we show that merging system responses according to query difficulty is a promising approach. We evaluate the methods on TREC Novelty track.*

MOTS-CLÉS : *recherche d'information, fusion de systèmes, recherche de passages, information textuelle*

KEYWORDS : *information retrieval, systems fusion, passage retrieval, textual information*

1. Introduction

Les systèmes de recherche d'information (SRI) montrent une certaine variabilité dans les résultats qu'ils produisent en réponse à un besoin d'information. Un premier aspect de la variabilité concerne la tendance d'un système à favoriser le rappel plutôt que la précision (ou inversement). Un autre aspect de la variabilité est lié aux documents sélectionnés par les systèmes en réponse à une requête. Enfin, la variabilité est liée aux différentes techniques utilisées par les systèmes lors du

processus de recherche. En effet, même si certains besoins sont difficiles à satisfaire pour une majorité de systèmes (Dkaki et al. 2004), (Buckley et al. 2004), certains systèmes ont des comportements spécifiques qui leur permettent de mieux répondre à certaines requêtes. (Fox et Shaw 1994) ont montré que combiner les résultats de plusieurs recherches améliore les performances par rapport au résultat d'une seule recherche, et que la combinaison la plus efficace (CombSUM) consiste à additionner les valeurs de similarité pour chaque document. Dans (Belkin et al. 1994), dix requêtes ont été formulées manuellement et combinées pour être soumises au système INQUERY (Callan et al. 1992). Ces expérimentations ont montré que les résultats sont fortement liés à la formulation des requêtes et que la combinaison des requêtes améliore la performance générale. De son côté, (Lee 1997) a montré que ComSUM peut être combinée de façon efficace en considérant la différence de chevauchement entre documents pertinents et non pertinents. Il a montré qu'il vaut mieux fusionner des systèmes qui ont un plus fort chevauchement de documents pertinents. (Beitzel et al. 2003) ont un peu contredit cette hypothèse en montrant que l'amélioration n'est pas tant liée au taux de chevauchement qu'au nombre de documents pertinents qui n'apparaissent que dans un résultat de recherche. Dans (He 2003), différentes fonctions de pondération de termes sont utilisées en fonction des requêtes et de leurs caractéristiques. Cette méthode, appliquée à la tâche 'Robuste' de TREC (Voorhees 2005) est efficace pour les requêtes aux performances faibles.

Les techniques ci-dessus s'appuient sur la similarité calculée par le système entre le document sélectionné et la requête. Cette information n'est pas nécessairement disponible. C'est le cas des systèmes de filtrage d'information ou des systèmes ayant participé à la sous-tâche " recherche de passages " de la tâche " Nouveauté " de TREC (Harman 2002). C'est dans ce contexte que nous avons mené notre étude. Dans nos expérimentations, les réponses ne sont pas ordonnées ; les techniques de fusion prenant en compte l'ordre des réponses sont alors impossibles.

2. Collections de test

Les collections TREC que nous utilisons dans notre étude sont composées des résultats (liste des documents retournés pour une requête donnée) issus des systèmes ayant participé à la sous-tâche1 (détection de la pertinence) de TREC " Nouveauté " 2002 et 2003. Les caractéristiques de ces collections sont fournies dans le tableau 1.

	NIST-2002	NIST-2003
Nombre de besoins d'information	49	50
Nombre moyen de documents pertinents par besoin	22,3	25
Nombre moyen de phrases issues des documents par besoin	1321	796,4
Nombre moyen de phrases pertinentes par besoin	27,9	311,14
% moyen de phrases pertinentes	2,1	39,1

Tableau 1. *Caractéristiques de la collection de test de TREC 2002 et 2003*

Les critères d'évaluation des systèmes sont ceux définis par TREC à savoir le rappel, la précision et la mesureF (Harman 2002). Treize groupes ont participé à TREC 2002, correspondant à un total de 43 systèmes ou ensembles de résultats, et quatorze groupes ont participés à TREC 2003 correspondant à 42 systèmes. En pratique, un groupe utilise généralement un seul outil pour lequel il teste différents paramètres ; nous appellerons donc 'système' un outil et les paramètres associés.

3. Méthodes de fusion de systèmes

Les méthodes proposées dans la littérature et relatives à la fusion de résultats s'appuient sur la prise en compte de la similarité entre la requête et les documents, information qui n'est pas disponible dans notre étude. Dans cet article nous proposons donc une méthode qui ne nécessite pas cette information et qui repose sur le concept très simple d'union ou d'intersection de résultats. Notre contribution porte d'avantage sur le choix des systèmes à fusionner en fonction des contextes. Fusionner des systèmes qui ont obtenus de mauvais résultats pourrait certes permettre d'améliorer les résultats qu'ils obtiennent, mais cette fusion aurait peu d'intérêt. Nous nous focalisons donc plutôt sur les systèmes ayant obtenu les meilleurs résultats.

3.1. Fusion aveugle des systèmes

Pour chaque année (2002 et 2003), nous sélectionnons les 5 meilleurs systèmes que nous fusionnons avec le meilleur. Le meilleur système pour 2002 (thunv3) a obtenu un rappel de *0.40* et une précision de *0.20*, pour 2003 THUIRnv0315 a obtenu une précision de *0.60* et un rappel de *0.79*. La première expérimentation a consisté à fusionner les résultats du meilleur système avec chacun de ses 4 suivants. Dans le tableau 2 nous mesurons la variation de performance moyenne par rapport aux performances initiales du meilleur système.

	2002 (Thunv3)		**2003 (THUIRnv315)**	
	Union	**Intersection**	**Union**	**Intersection**
Rappel	*17.5 %*	*-10 %*	*7.6 %*	*-10.2 %*
Précision	*-10 %*	*10 %*	*-5 %*	*3 %*
mesureF	*4.2 %*	*0.6 %*	*-0.3 %*	*-3.6 %*

Tableau 2. *Performance après fusion avec les meilleurs systèmes*

Il est bien évident que l'intersection favorise la précision (puisque les deux systèmes étaient d'accord pour sélectionner les unités d'information

correspondantes), alors que l'union favorise le rappel (puisque les unités pertinentes retrouvées par l'un ou l'autre des systèmes sont retrouvées après fusion).

De façon générale, le tableau 2 montre que lorsque les meilleurs systèmes sont considérés, leur fusion (par union ou par intersection) ne modifie pas sensiblement les performances observées en termes de mesureF (+0,6% pour l'intersection et 4 ,2% pour l'union en 2002, et -3,6% pour l'intersection et –0,3% pour l'union en 2003). Nous pouvons donc dire qu'il est difficile d'améliorer les performances du meilleur système avec ses suivants.

3.2. Fusion " intelligente " des systèmes

Les techniques de fusion que nous avons utilisées dans la section précédente ne prenaient pas en compte les spécificités des systèmes et des requêtes. Nous essayons de voir si la fusion peut être " orientée " en fonction de paramètres sur les requêtes et de paramètres sur les systèmes.

Nous distinguons deux catégories de requêtes : les requêtes faciles, et les requêtes difficiles (Dkaki 2004). Isoler ces deux types de requêtes permet de voir l'impact de ces catégories de requêtes sur les performances des systèmes. Le deuxième paramètre de test est issu de la catégorisation des systèmes utilisés. Le premier critère est le nombre d'unités (compris entre 1967 et 14752) retrouvé par les systèmes. Suite à nos analyses, ce critère influe peu sur les résultats. Le deuxième critère compare la ressemblance et la différence de l'ensemble des unités retrouvées par les systèmes. Nous avons identifié pour ce deuxième critère trois catégories de systèmes : *les systèmes opposés, les systèmes équivalents, et les systèmes différents.* Nous considérons que deux systèmes sont *opposés* s'ils ont une intersection nulle de documents pour une requête donnée. Les systèmes *équivalents* quant à eux retournent les mêmes documents pour une requête. La fusion (Intersection ou Union) n'a pas d'impact sur les performances lorsqu'elle est appliquée sur des systèmes équivalents. De même, les systèmes opposés voient leurs performances augmenter à l'issue d'une fusion par union. Par exemple, les systèmes cmu02t300rAs, fdut11n3, novcolmerg et nttcslabnvr2 retournent chacun 50 premiers documents différents. Dans ce cas, nous les considérons comme systèmes opposés. Le tableau suivant nous montre une amélioration des performances des systèmes après fusion par union. Pour chacun des systèmes, la 1ère colonne donne la valeur initiale et la 2ème colonne la valeur obtenue après fusion.

	Fdut11n3		Novcolmerg		Nttcslabnvr2	
	avant	après	Avant	après	avant	après
Mesure-F moyenne	0,161	**0,184**	0,103	**0,173**	0,166	**0,168**

Tableau 3. *Comparaison des performances initiales et après fusion de systèmes opposés*

En étudiant les résultats des systèmes, les requêtes 377, 420 et 432 ont les plus faibles valeurs de mesureF sur l'ensemble des requêtes. Nous essayons de voir le comportement de la fusion sur ce genre de requêtes, en particulier la requête 377 qui a la plus faible valeur de mesure-F.

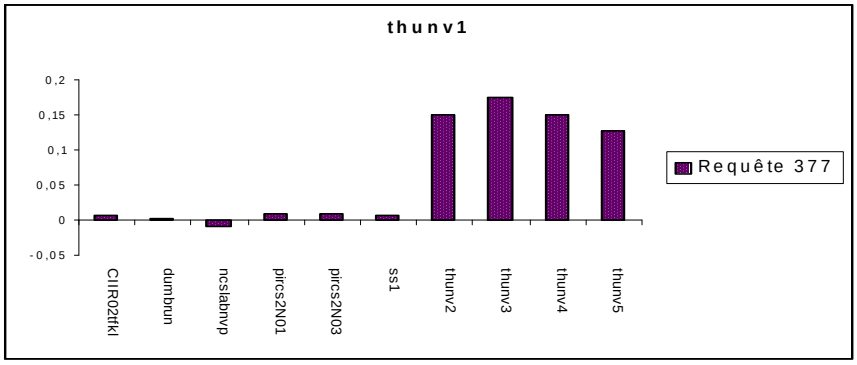

Figure 3. *Mesure-F après fusion avec thunv1 pour la requête 377*

Pour la requête la plus difficile (377) la fusion avec thunv1 est celle qui permet d'obtenir la meilleure augmentation de performance. De plus, le couple de système thunv1/thunv3 est celui qui permet d'obtenir la meilleure augmentation de performance. En effectuant une fusion intelligente on choisira de fusionner ces deux systèmes pour la requête 377 plutôt que d'autres.

4. Conclusions et perspectives

La fusion aveugle des systèmes apporte des améliorations mineures sur le meilleur système lorsque nous limitons les expérimentations sur les 5 meilleurs systèmes. La fusion " d'intelligente " permet de voir si certains caractères ont une incidence sur les performances des systèmes après fusion. Cela nous a permis de voir si le choix de paramètres peut *orienter* les résultats de la fusion. Deux classes de systèmes ont été trouvées à l'issue de la fusion: ceux qui améliorent et ceux qui dégradent les performances des systèmes. Nous avons ensuite affiné cette expérimentation en testant si le caractère facile ou difficile des requêtes avait un impact sur les résultats. Nous nous sommes ensuite intéressés à certaines requêtes et avons essayé de classifier les systèmes en fonction de ce nouveau paramètre. Les résultats que nous avons obtenus jusqu'à présent nous montrent que le choix du système à fusionner avec les autres est déterminant, et apporte des améliorations par rapport aux résultats initiaux.

Remerciements

Ces travaux s'inscrivent dans le cadre du projet ARIEL soutenu par le programme TCAN.

Bibliographie

N.J Belkin et al. "*The effect of multiple query representations on information retrieval performance*". ACM-SIGIR pp. 339-346 - 1993.

J.P. Callan et al. "*The INQUERY Retrieval System"*. Tjoa A. M. and Ramos I. editors, *Database and Expert Systems Applications, International Conference*. pp. 78–83, 1992.

E.A. Fox and J.A. Shaw. "*Combination of Multiple Searches*". TREC-2, NIST Special Publication 500-215, pp. 243-252 - 1994.

B. He, I Ounis "*A query-based Model Selection Approach for Poorly-performing Queries"* -2003.

J. Lee. "*Analysis of multiple evidence combination"*. ACM-SIGIR pp. 267-276 - 1997.

S. M. Beitzel et al. "*Disproving the fusion hypothesis: an analysis of data fusion via effective information retrieval strategies, SAC '03"*: ACM symposium on Applied computing, pp. 823-827 - 2003.

T. Dkaki et al. " *Recherche de la nouveauté dans les textes: une tâche difficile "*, Veille Stratégique Scientifique & Technologique pp 355-368 – 2004

C.Buckley, D. Harman. *" The NRRC reliable information access (RIA) workshop "*. ACM SIGIR pp. 528 - 529- 2004

J. Belkin and al. "*The Interactive Searching Behavior of Expert Online Searches using INQUERY"*. New Tools and Old Habits, 1994

D. Harman. *"Overview of the TREC 2002 Novelty Track"* TREC 2002

E. Voorhees, "The TREC Robust Retrieval track", ACM-SIGIR . June 2005