Fusion de résultats en recherche d'information : Mesure de l'impact de l'union et de l'intersection de résultats

Data fusion in information retrieval: measuring the impact of fusing by union and intersection

Désiré Kompaoré, Josiane Mothe

Institut de Recherche en Informatique de Toulouse
Université de Toulouse
(118 Route de Narbonne, 31062 Toulouse, cedex 4)
[kompaore/mothe] @irit.fr

RÉSUMÉ. Cet article présente une étude que nous avons menée sur la fusion de données en recherche d'information. Nous proposons deux stratégies de fusion "systématique" des systèmes, en nous basant sur des techniques simples de fusion par union et intersection. L'objectif de ce travail est de quantifier les améliorations en termes de rappel et de précision attendues par ces techniques. Les résultats sur la collection de TREC novelty 2002 (resp. 2003) montrent que la fusion par union améliore le rappel du meilleur système de 48% (resp. 18%) si l'on considère la meilleure combinaison, et en moyenne de 23% (resp. 6,5%) si l'on considère la combinaison avec les 9 autres meilleurs systèmes. De la même façon, la fusion par intersection améliore la précision du meilleur système de 15% (resp. 9%) si l'on considère la meilleure combinaison, et en moyenne de 10% (resp. 5,5%) si l'on considère la combinaison avec les 9 autres meilleurs systèmes. La tendance de ces résultats est identique lorsque plus de systèmes sont fusionnés deux à deux.

MOTS-CLÉS : recherche d'information, fusion de données, performance, évaluation, TREC.

ABSTRACT. The focus of this paper is data fusion in information retrieval. We investigate the effect of two simple fusion algorithms: union and intersection. Union improves recall, and intersection improves precision. The goal of this paper is to quantify the improvement through a set of experiments. We show that, considering TREC 2002 (resp. 2003) novelty benchmark collection, fusing using union improve the recall of the best system by 48% (resp. 18%) considering the best fusion and in average by 23% (resp. 6,5%) when considering the 9 following best systems. In the same way, when considering fusing by intersection, precision of the best system is improve by 15% (resp. 9%) when considering the best fusion and in average by 10% (resp. 5,5%) when considering the 9 next best systems. The trend is similar when considering more systems and fusing the results 2 by 2.

KEYWORDS: information retrieval, data fusion, performance, evaluation, TREC**.**

1. Introduction

De nombreux paramètres peuvent influencer les résultats qu'obtiennent les systèmes de recherche d'information : la méthode d'indexation des documents utilisée, le traitement de la requête, le modèle de recherche sous-jacent et la fonction d'ordonnancement des résultats. Les travaux dans le domaine de la recherche d'information visent donc à proposer des améliorations dans l'un ou l'autre des maillons de la chaine utilisée. D'autres travaux, partant du postulat que différents systèmes retrouvent les documents dans un autre ordre ou retrouve différents documents, visent à combiner ces systèmes. C'est dans ce dernier cadre que s'inscrivent nos travaux.

Les méthodes de fusion de résultats de systèmes [data fusion] proposées dans la littérature s'appuient sur la prise en compte de la similarité entre la requête et les documents, c'est-à-dire le degré de pertinence des documents restitués par les systèmes. Cependant, cette information est rarement disponible. Les moteurs du web par exemple ne la fournissent pas en même temps que les documents. Nous nous sommes donc intéressés aux méthodes qui ne nécessitent pas ce type d'information, de sorte que la méthode soit applicable quels que soient les systèmes à fusionner utilisés. Certaines autres

techniques de fusion se basent sur le rang des documents retrouvés ; il s'agit là d'un moyen de contourner le manque d'information sur la similarité entre les documents retrouvés et la requête. Dans ces méthodes, l'objectif principal est de mieux satisfaire l'utilisateur en lui restituant d'abord (c'est-à-dire en haut de la liste) les documents pertinents. Pourtant, dans certaines activités, l'objectif de l'utilisateur n'est pas d'obtenir quelques documents pertinents, mais de bien collecter un ensemble de documents répondant à un besoin et qui seront par la suite analysés (Dousset et Mothe, 2004). Les activités de veille scientifique et technologique répondent par exemple à ce cadre. Dans ce contexte, il est donc important de pouvoir s'assurer que l'ensemble des documents retrouvés contient un maximum de documents pertinents (rappel élevé, faible silence documentaire) et peu de documents non pertinents (précision élevée et faible bruit documentaire). Une méthode simple permettant d'augmenter le rappel consiste à fusionner les résultats obtenus en effectuant une union. Par cette méthode, le rappel ne peut être qu'augmenté (ou maintenu au même niveau si les deux systèmes retrouvent les mêmes documents pertinents). De la même façon, une méthode simple pouvant permettre d'augmenter la précision consiste à considérer l'intersection des ensembles retrouvés : un document retrouvé par les deux systèmes a plus de chances d'être pertinent. Il est cependant bien connu qu'appliquer une méthode qui augmente le rappel dégrade généralement la précision et inversement.

Dans cet article, nous nous intéressons à mesurer ces effets.

2. Travaux reliés

(Fox et Shaw, 1994) ont proposé des fonctions de fusion de résultats de plusieurs systèmes basées sur une combinaison linéaire des scores des documents. Parmi les formules proposées, la formule CombSUM calcule la somme des scores de tous les documents retournés par les SRI. La formule CombMNZ prend en compte le nombre de systèmes qui ont retrouvé le même document et multiplie la valeur de CombSUM par ce nombre. Les auteurs ont montré que la combinaison de plusieurs techniques de recherche augmente l'efficacité globale de la recherche. D'après leurs conclusions, la formule Comb-SUM appliquée à la collection TREC-2 apporte des améliorations de la R-Précision (précision lorsque R documents sont retrouvés, R étant le nombre de documents effectivement pertinents) de l'ordre de 13 %. Ces techniques de fusion ont également été utilisées avec succès par (Lee, 1997) qui a montré que CombMNZ permet d'obtenir de meilleures performances que CombSUM et produit de bons résultats dans le cas où le taux de chevauchement des documents pertinents est élevé (entre 0,75 et 0,82) et le taux de chevauchement des documents non pertinents bas (entre 0,30 et 0,40). (Beitzel et al., 2004) ont contredit ce résultat 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 (Vogt et Cottrell, 1998), les auteurs proposent d'utiliser un modèle linéaire pour combiner les différents scores obtenus. Les résultats de leurs travaux montrent que cette méthode est seulement efficace lorsque le taux de chevauchement des documents pertinents est très élevé et le taux de chevauchement des documents non pertinents est faible. Dans le cas où le score de similarité entre le document et la requête n'est pas disponible, il existe d'autres techniques qui se basent par exemple sur le rang des documents dans la fusion pour améliorer la recherche. (Voorhees et al., 1994) propose une technique simple de fusion, basée sur les rangs des documents, qui consiste à choisir les documents classés en première position dans les différentes listes retournées par les SRI, puis les documents classés en deuxième position, et ainsi de suite, après avoir supprimé les doublons. (Lee, 1997) utilise le rang des documents comme alternative aux scores de similarité, et il obtient de bons résultats en termes de précision moyenne. (Soboroff et al. 2001) se sont intéressés à la comparaison entre l'utilisation des scores de similarité et les rangs des documents, lors de la RI. Ils ont combiné les sous-listes des différents SRI en remplaçant les scores de similarité par une mesure qui prend en compte les rangs des documents Les résultats obtenus montrent que l'utilisation des rangs est plus performante que l'utilisation des scores de similarité. (Farah et Vanderpooten, 2007) propose une technique d'agrégation des rangs prenant en compte les documents ayant le même rang dans les listes de documents. Les expérimentations qu'ils ont effectuées montrent que leur méthode permet d'obtenir de meilleurs résultats que CombSUM et CombMNZ. Une des conclusions est que la fusion doit être effectuée sur les listes restituées par les meilleurs systèmes. (Lillis et al., 2006) utilise une approche de fusion probabiliste basée sur les performances passées des systèmes, pour un ensemble de requêtes de test. Les résultats qu'ils obtiennent sont supérieurs à ceux obtenus avec CombSUM. (Wu and McClean, 2006) analysent le comportement des méthodes de fusion les plus répandues (CombSUM et CombMNZ) et montrent qu'il est possible de prédire les performances des méthodes de fusion ; leurs expérimentations se basent sur les collections de TREC. (Spoerri, 2007) analysent en détail deux effets liés à la fusion de données: l'autorité (plus il y a de systèmes qui retrouvent un même document, plus le document est potentiellement pertinent) et l'effet du rang (plus un document est retrouvé haut dans les listes, plus il est potentiellement pertinent). En utilisant les données de TREC ils montrent que ces phénomènes se retrouvent quelque soit le nombre de documents retrouvés considérés, mais que si les systèmes retrouvent un grand nombre de documents, alors l'effet de rang ne débute que lorsque suffisamment de système ont retrouvé un document et/ou si le nombre de documents considéré augmente.

3. Données étudiées

3.1. Tâche TREC considérée, requêtes et caractéristiques des collections

Les méthodes de fusion sont évaluées sur les collections de TREC 2002 et 2003 utilisées dans la sous-tâche détection de passages de la tâche "nouveauté". Le choix de cette collection est motivé par le fait que cette tâche vise à sélectionner les documents ou parties de documents intéressantes, sans se soucier de l'ordre dans lequel ils sont restitués. Cela correspond à notre cadre d'étude.

La tâche détection de la nouveauté a été introduite en 2002 lors de la campagne TREC-11. Etant donné une requête et une liste ordonnée de documents pertinents, l'objectif de cette tâche est de retrouver les passages (phrases) pertinents et nouveaux répondant à la requête. Nous avons utilisé dans nos travaux les documents issus de la première sous-tâche qui consiste à détecter les passages pertinents dans les documents. La collection de départ est constituée de 50 requêtes provenant des campagnes *adhoc* TREC6, TREC7, et TREC8. Ces 50 requêtes ont été sélectionnées parmi celles pour lesquelles les jugements de pertinence comprenaient entre 10 et 70 documents pertinents. En 2002, TREC a choisi de sélectionner 50 requêtes parmi les besoins d'information identifiés par les numéros 300 à 450. Le NIST a sélectionné les documents effectivement pertinents pour chacun de ces besoins d'information (jugements de pertinence), avec un maximum de 25 documents par besoin, et les a fournis aux participants. Un exemple de requête est décrit dans la figure 1.

Une requête est composée d'un numéro qui identifie la requête, un titre (T) qui donne une description du sujet de la requête en quelques mots, une description (D) de la requête exprimée à travers une phrase et une partie narrative (N) qui explique plus en détail le type de documents pertinents et non pertinents qui est recherché.

Dans une seconde étape, des juges indiquent quelles phrases de ces documents sont effectivement pertinentes. Le même type de principe a été utilisé en 2003, avec 50 besoins. Les caractéristiques de ces collections sont fournies dans le tableau 1.

Topic : 35

Title : NATO, Poland, Czech Republic, Hungary

Descriptive : Accession of new NATO members : Poland, Czech Republic, Hungary, in 1999.

Narrative : Identity of current and newly-invited members, statements of support for and opposition to NATO enlargement and steps in the accession process and related special events are relevant. Impact on the new members, i.e., requirements they must satisfy, and their expectations regarding the implications for them are relevant. Progress in the ratification process is relevant. Future plans for NATO expansion, identification of nations admitted on previous occasions, and comments on future NATO structure or strategy are not relevant.

Figure 1 : *Exemple de requête TREC*

	TREC2002	TREC2003
Nombre de besoins d'information	49	50
Nombre moyen de documents pertinents par besoin	22,3	25
Nbre moyen de phrases 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 des collections TREC 2002 et 2003

3.2. Exécutions et évaluation

Chaque participant propose la liste des éléments que le système considère comme pertinent. Un même système, paramétré différemment peut être utilisé. Dans ce cas, deux exécutions seront soumises et évaluées.

Trec_eval est généralement utilisé pour évaluer une exécution. Ce logiciel permet de calculer plus d'une centaine de mesures. Dans cet article, nous nous sommes focalisés sur les mesures en lien avec notre étude : le rappel, la précision et

la mesure F. Le rappel exact mesure la proportion de documents pertinents effectivement retrouvés, en moyenne sur l'ensemble des requêtes considérées. La précision exacte (non interpolée) mesure la précision moyenne sur l'ensemble des documents retrouvés. Enfin, la mesure F combine les deux mesures précédentes (qui varient en sens inverse). Elle est calculée par la formule (2*RP/(R+P)) où R (resp. P.) est le rappel (resp. précision) exact ; faisant ainsi jouer la même importance au rappel et à la précision.

4. Etude préliminaire

L'objectif de cette analyse préalable est de détailler les caractéristiques des collections et des systèmes que nous utilisons dans les expérimentations. La figure 2 montre la distribution des performances des systèmes en termes de rappel pour les collections 2002 et 2003. Nous utilisons une représentation des données sous forme de boîte à moustaches (Tukey, 1977). Ce type de représentation permet de représenter une distribution de données. Elle utilise 5 valeurs qui résument des données : le minimum, les 3 quartiles Q1, Q2 (médiane), Q3, et le maximum. Les quartiles jouent un rôle important dans l'interprétation. La médiane Q2 divise la série en deux groupes d'effectif égaux. Le premier quartile divise le groupe de données situé en dessous de la médiane en deux groupes d'effectif égaux. Le troisième quartile divise quant à lui le groupe situé au delà de la médiane en deux groupes de taille égale. Les quartiles représentent 25 (Q1), 50 (Q2) et 75% (Q3) des données analysées.

Dans la figure 2, la première boite à moustaches décrit la répartition des mesures de rappel pour la campagne de 2002. Les valeurs de rappel qui sont présentées correspondent à la moyenne que chaque système obtient sur l'ensemble des requêtes pour lesquelles le système est évalué. En 2002 la valeur maximale de rappel est 0,597 (et de 0,999en 2003) et 75% des valeurs de rappel sont inférieures à 0,4. En 2003, plus de la moitié des systèmes obtiennent une valeur de rappel supérieure à 0,42%.

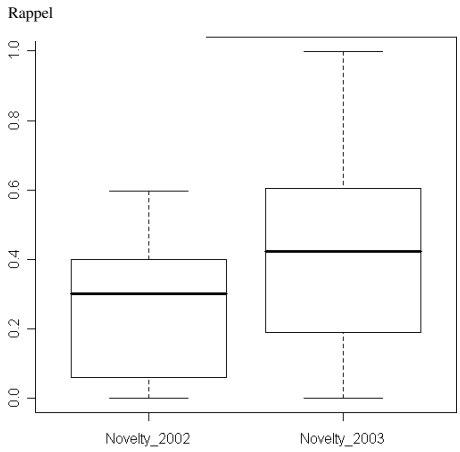

Figure 2 : Répartition du rappel pour les collections TREC-Nouveauté 2002 et 2003.

La figure 3 complète l'interprétation des boites à moustaches. Elle représente les variations entre le système qui obtient la plus grande valeur de rappel et les autres systèmes. Dans cette figure, les différences de rappel entre deux systèmes de rangs consécutifs ne sont pas grandes. Un certain nombre de paliers sont toutefois à remarquer aussi bien pour 2002 que pour 2003. Par exemple, en 2003, une grande différence de rappel existe entre le premier système (ISIALL03 0,999) et le deuxième système (MeijiHilF13 0,84). De la même façon, en 2003, une variation importante (0,066) du rappel est observée entre le cinquième et le sixième système en 2003 ; idem pour 2002 avec une variation entre le cinquième et sixième système de 0,049.

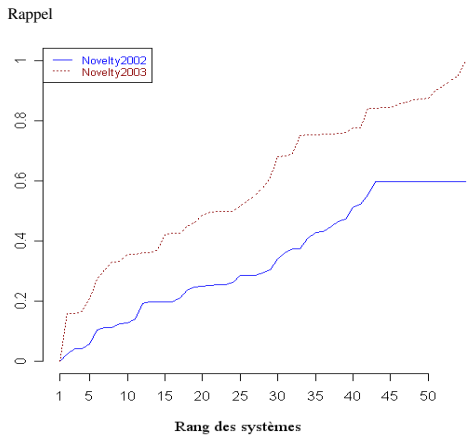

Figure 3 : Variation entre le rappel du meilleur système et les autres systèmes

Concernant la précision (Figure 4), quelques systèmes obtiennent des valeurs de précision en dessous de 0,5 en 2003. Par cela, ils se distinguent des autres systèmes. Il s'agit des systèmes suivants : lexiclone03 (0,484), ISIALL03 (0,411), ISIRAND03 (0,41), umbcrun2 (0,405), umbcrun3 (0,4), umbcrun1 (0,396).

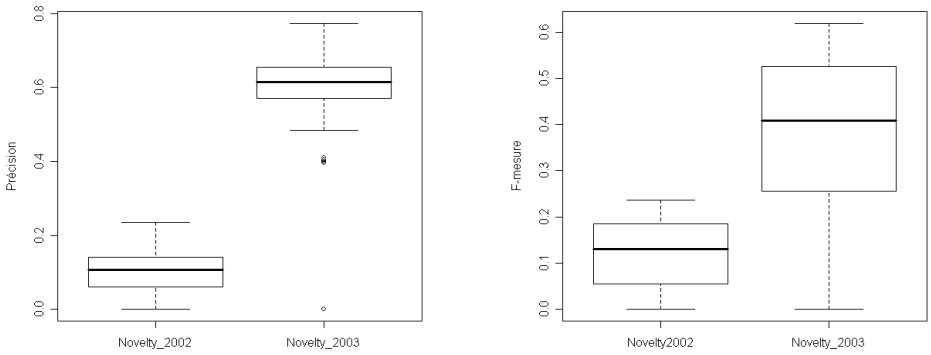

Figure 4. - Répartition des valeurs initiales de précision et de F-mesure

Le système ISINONE03 obtient quant à lui une valeur de précision nulle en 2003. Les performances des systèmes en 2003 sont très élevées par rapport à 2002 comme on peut le constater dans la figure 4 pour la précision et la mesure F.

La figure 5 montre la différence entre la précision du meilleur système et celle des autres systèmes ; cette différence est inférieure à 0,2 pour les 40 premiers systèmes de 2002 et 2003. Les plus grandes variations sont dues aux systèmes évoqués dans le paragraphe précédent et qui ont des valeurs de précision bien plus bases que les autres systèmes. De plus, les différences entre les performances des meilleurs systèmes et ceux qui obtiennent de mauvais résultats sont grandes. Par exemple la différence entre le meilleur et le dernier système en termes de rappel en 2003 est de 0,999.

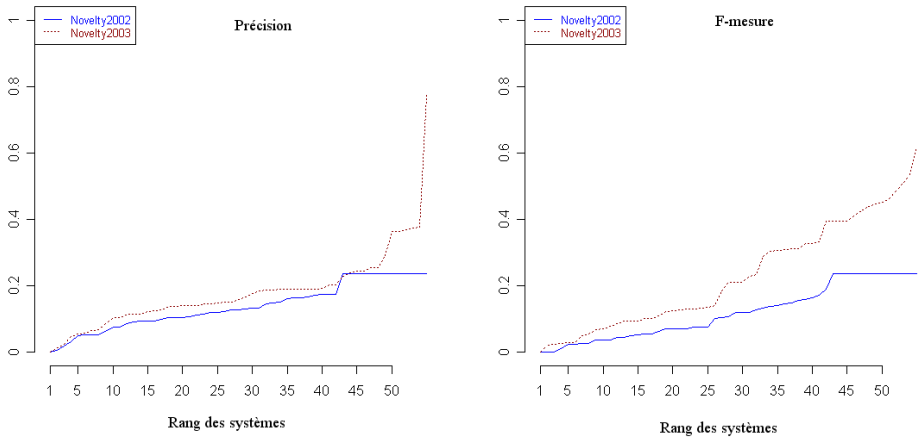

Figure 5 : Variation entre la précision et la mesure F du meilleur système et les autres systèmes

Cette analyse préalable indique que les performances moyenne entre les meilleurs systèmes sont proches, indépendamment des stratégies de recherche utilisées par les systèmes. Elle indique également que certains systèmes ont des performances très faibles. Ainsi, dans la suite des expérimentations, nous nous sommes focalisés sur les meilleurs systèmes (les 10 premiers en termes de la mesure combinée mesure F).

5. Fusion par union et intersection

L'objectif de cette étude est de mesurer l'amélioration en termes de précision que peut apporter la fusion par intersection, ainsi que celle en termes de rappel que peut amener la fusion par union.

5.1. Base de la comparaison

Les résultats obtenus par les techniques de fusion sont comparés à ceux obtenus par les systèmes simples. Le tableau 2 indique les performances obtenues par les 10 meilleurs systèmes.

Ainsi, en 2002, le système ayant obtenu les meilleures performances est le système thunv3 avec 0,237 (mesure F) avec un rappel de 0,404 et une précision de 0,204. En 2003, le meilleur système obtient une mesure F de 0,619 (rappel 0,792 et précision 0,597).

	Rappel	Précision	Mesure F
Thunv3	0,404	0,204	0,237
Thunv1	0,341	0,229	0,236
Thunv2	0,341	0,236	0,236
Thunv4	0,335	0,216	0,226
CIIR02tfkl	0,556	0,141	0,213
CIIR02tfnew	0,556	0,141	0,213
pircs2N01	0,486	0,161	0,211
pircs2N02	0,486	0,161	0,211
pircs2N03	0,4	0,184	0,199
pircs2N04	0,4	0,184	0,199

2002

	Rappel	Précision	Mesure F
THUIRnv0315	0,792	0,597	0,619
ISIDSCm203	0,832	0,534	0,597
Ulowa03Nov01	0,696	0,636	0,594
THUIRnv0.11	0,726	0,606	0,593
MeijiHilF13	0,84	0,52	0,589
MeijiHilF14	0,84	0,52	0,589
Ulowa03Nov02	0,637	0,649	0,568
THUIRnv0312	0,665	0,624	0,564
THUIRnv0313	0,637	0,633	0,552
THUIRnv0314	0,628	0,632	0,548

2003

Tableau 2 : Performances initiales des 10 meilleurs systèmes

5.2. Principe de fusion

Les méthodes de fusion que nous étudions se basent sur les principes simples d'intersection et d'union des ensembles de documents retrouvés (Kompaore et al., 2006).

L'union regroupe les documents sélectionnés par les systèmes fusionnés, après suppression des éventuels doublons. Le principe de l'union est celui de la théorie des ensembles appliqué aux ensembles de documents retrouvés. L'intersection quant à elle permet de regrouper les documents restitués en commun par les systèmes fusionnés.

5.3. Fusion par union

Dans cette section, nous étudions l'impact de la fusion par union des résultats des systèmes. Comme nous l'avons précisé précédemment, ce type de fusion a pour objectif d'améliorer le rappel. Nous indiquons toutefois les performances en termes de mesure F afin de mieux pouvoir comparer globalement les performances.

Le tableau 3 indique les variations mesurées. Les chiffres en gras marquent une amélioration par rapport aux résultats du meilleur système. Les chiffres en italique indiquent au contraire une dégradation. Les fusions sont ordonnées par rappel décroissant.

Le tableau 3 montre que pour les données de 2002, la fusion par union de thunv1 et thunv3 n'apporte aucune amélioration par rapport au rappel initial de thunv3. Ce tableau nous indique que la fusion par union de thunv3 avec les autres versions du même système (thunv1, thunv2, et thunv4) ne modifie pas les performances obtenues par thunv3 pour les mesures utilisées. En analysant les données, on constate que les réponses des 4 versions du système thunv constituent en fait des sous ensembles des réponses du système thunv3. Concernant la fusion de thunv3 avec les autres meilleurs systèmes, le rappel est amélioré. Cette amélioration du rappel peut aller jusqu'à presque 50%, montrant ainsi que les documents restitués par l'un et l'autre des systèmes fusionnés sont différents. La fusion de ces deux systèmes permet d'obtenir un rappel supérieur à ce qu'il aurait été si les systèmes avaient été considérés individuellement (Chacun de ces deux systèmes pris individuellement : Ciir02tfkl obtenait seul un rappel de 0,556 contre 0,597 pour la fusion (soit une augmentation de 7%) et Thunv3 obtenait seul un rappel de 0,404 (soit une augmentation de 48% pour la fusion).

Systèmes fusionnés	Rappel	Mesure F
thunv3	(0,404)	(0,237)
Ciir02tfkl-thunv3	**0,597 (+48%)**	*0,212 (-11%)*
Ciir02tfnew-thunv3	**0,597 (+48%)**	*0,212*
pircs2n01-thunv3	**0,543 (+34%)**	*0,211 (-11%)*
pircs2n02-thunv3	**0,543 (+34%)**	*0,211*
pircs2n03-thunv3	**0,496 (+23%)**	*0,216 (-9%)*
pircs2n04-thunv3	**0,496 (+23%)**	*0,216*
Thunv1-thunv3	0,404	0,237
Thunv2-thunv3	0,404	0,237
Thunv3-thunv4	0,404	0,237
2002		

Systèmes fusionnés	Rappel	Mesure F
thuirnv0315	(0,792)	(0,619)
meijihilf13-thuirnv0315	**0,938 (+18%)**	*0,613 (-1%)*
meijihilf14-thuirnv0315	**0,938 (+18%)**	*0,613*
isidscm203-thuirnv0315	**0,921 (+16%)**	*0,618*
thuirnv0311-thuirnv0315	**0,82 (+4%)**	*0,623 (-1%)*
thuirnv0312-thuirnv0315	**0,804 (+2%)**	*0,623*
thuirnv0315-uiowa03nov01	**0,797 (+1%)**	**0,62**
thuirnv0315-uiowa03nov02	**0,793 (+0,13%)**	0,619
thuirnv0313-thuirnv0315	0,792	0,619
thuirnv0314-thuirnv0315	0,792	0,619
2003		

Tableau 3 : Performance des meilleurs systèmes fusionnés par union avec le meilleur système

Les mêmes conclusions peuvent être faites en 2003 où par exemple, le système thuirnv0315 seul obtenait un rappel de 0,792 contre 0,938 pour la fusion avec meijihilf13 (soit une augmentation de 18%) ; le système meijihilf13 seul obtenait 0,84 (soit une augmentation de 12% pour la fusion). Ainsi, dans la mesure où l'on connaît le meilleur système, il est intéressant de le fusionner avec n'importe quel autre système performant. Il est cependant important de noter que cette information n'est connue qu'*a posteriori*. La section 5 présente donc une étude plus globale des résultats de fusion par union.

5.4. Fusion par intersection

Nous présentons dans cette section les résultats obtenus concernant la fusion par intersection des résultats des meilleurs systèmes. Ce type de fusion a pour objectif d'améliorer la précision. Comme précédemment, nous indiquons toutefois les performances en termes de mesure F ; les résultats sont ordonnés par ordre de précision décroissante ; les chiffres en gras indiquent une amélioration et ceux en italique une dégradation.

Systèmes fusionnés	Précision	Mesure F
thunv3	(0,204)	(0,237)
thunv2-thunv3	**0,234 (+15%)**	*0,235*
pircs2n03-thunv3	**0,232 (+14%)**	*0,221 (-7%)*
pircs2n04-thunv3	**0,232 (+14%)**	*0,221 (-7%)*
thunv1-thunv3	**0,228 (+12%)**	*0,235 (-1%)*
pircs2n01-thunv3	**0,226 (+11%)**	**0,24 (+1%)**
pircs2n02-thunv3	**0,226 (+11%)**	**0,24 (+1%)**
thunv3-thunv4	**0,216 (+6%)**	*0,226 (-5%)*
Ciir02tfkl-thunv3	**0,213 (+4%)**	**0,24 (+1%)**
Ciir02tfnew-thunv3	**0,213 (+4%)**	**0,24 (+1%)**
2002		

Systèmes fusionnés	Précision	Mesure F
thuirnv0315	(0,597)	(0,619)
thuirnv0315-uiowa03nov02	**0,65 (+9%)**	*0,568 (-8%)*
thuirnv0315-uiowa03nov01	**0,639 (+7%)**	*0,593 (-4%)*
thuirnv0313-thuirnv0315	**0,636 (+7%)**	*0,552 (-11%)*
thuirnv0314-thuirnv0315	**0,635 (+6%)**	*0,549 (-11%)*
thuirnv0312-thuirnv0315	**0,632 (+6%)**	*0,559 (-10%)*
thuirnv0311-thuirnv0315	**0,622 (+4%)**	*0,588 (-5%)*
meijihilf13-thuirnv0315	**0,618 (+4%)**	*0,592 (-4%)*
meijihilf14-thuirnv0315	**0,618 (+4%)**	*0,592 (-4%)*
isidscm203-thuirnv0315	**0,615 (+3%)**	*0,595 (-4%)*
2003		

Tableau 4 : Performance des meilleurs systèmes fusionnés par intersection avec le meilleur système

Le tableau 4 montre en particulier que, en 2002, la fusion pircs2n01-thunv3 améliore la précision de 11% par rapport au système thunv3 (de 0,204 à 0,226). En outre, l'augmentation correspond à 40% par rapport à pircs2n01. De la même façon, en 2003, la fusion isidscm203-thuirnv0315 améliore de 3% la précision par rapport à thuirnv0315 seul (0,615 contre 0,597). L'amélioration de la précision est de 15% par rapport à isidscm203 seul (0,615 contre 0,534).

De façon générale, la fusion par intersection améliore la précision des deux systèmes de façon importante ; même si les pourcentages d'augmentation sont moins importants que dans le cas du rappel.

6. Expérimentations complémentaires

Dans les premières expérimentations que nous avons reportées à la section 4, nous avons choisi les meilleurs systèmes uniquement sur la base de leur performance en termes de mesure F. D'autre part, la fusion a été réalisée en considérant le meilleur système, que nous avons fusionné aux autres. Les expérimentations que nous reportons ici ne sont pas centrées sur les performances du meilleur système. Les 10 systèmes qui sont sélectionnés sont combinés deux à deux, et le résultat de la fusion par union et par intersection est analysé, indépendamment du rang initial des systèmes.

6.1. Choix des « meilleurs » systèmes à fusionner

Si l'on observe les performances des meilleurs systèmes en considérant les différentes mesures de performance pour 2002 et 2003, on remarque que plusieurs versions de différents systèmes obtiennent les meilleures performances en 2002 et en 2003 (tableau 5). Par exemple, trois systèmes et leurs versions obtiennent les 10 plus grandes valeurs de précision en 2003 (il s'agit des versions des systèmes NLPPR03n et clr03n1, ainsi que le système ccsummeoqr). Ainsi, la sélection des meilleurs systèmes est plus proche d'une sélection dans un cadre réel : si les systèmes étaient disponibles, il serait possible de les utiliser sur différentes collections d'apprentissage pour décider des meilleurs systèmes en moyenne. Le fait d'en considérer plusieurs nous affranchi d'une certaine dépendance aux collections. Pour chaque année et chaque mesure, les 10 systèmes sélectionnés sont utilisés pour réaliser la fusion. Nous avons adopté deux stratégies différentes pour la sélection des 10 meilleurs systèmes. Dans la première stratégie (stratégie1), nous utilisons la mesure F comme base de sélection. Dans la stratégie2, les 10 meilleurs systèmes sont sélectionnés en fonction de la mesure visée. Ainsi, les 10 systèmes qui obtiennent le meilleur rappel sont sélectionnés lors de la fusion par union ; ceux qui obtiennent la meilleure précision sont sélectionnés pour la fusion par intersection. La fusion est réalisée en combinant 2 à 2 les résultats obtenus par les systèmes sélectionnés. Par exemple le premier système est fusionné avec les 9 autres systèmes, le deuxième avec 8 systèmes, et ainsi de suite pour l'ensemble des 10 systèmes.

2002					
Systèmes	Rappel	Systèmes	Précision	Systèmes	mesureF
nttcslabnvr2	0,597	thunv2	0,236	thunv3	0,237
UIowa02Nov4	0,574	thunv1	0,229	thunv1	0,236
CIIR02tfkl	0,556	thunv4	0,216	thunv2	0,236
CIIR02tfnew	0,556	thunv3	0,204	thunv4	0,226
2003					
ISIALL03	0,999	NLPR03n1w3	0,774	THUIRnv0315	0,619
MeijiHilF13	0,84	NLPR03n1w2	0,761	ISIDSCm203	0,597
MeijiHilF14	0,84	NLPR03n1f2	0,751	UIowa03Nov01	0,594
ISIDSCm203	0,832	NLPR03n1f1	0,726	THUIRnv0311	0,593
THUIRnv0315	0,792	clr03n1d	0,718	MeijiHilF13	0,589

Tableau 5 : Performance des systèmes en 2002 et 2003, ordonnés en fonction des performances (extrait)

6.2. Comparaison des stratégies pour l'union (rappel)

La figure 6 compare pour chacune des collections utilisées les résultats obtenus en termes de rappel en utilisant les 2 stratégies que nous proposons. Nous retenons pour chaque stratégie les 45 meilleures valeurs de rappel après fusion (2 à 2 en utilisant les 10 meilleurs systèmes) que nous comparons avec le classement des 45 meilleurs systèmes uniques pour le rappel. Dans cette figure, Rappel_i_2002 pour 2002 et Rappel_i_2003 pour 2003 correspond au rappel de ces 45

systèmes avant fusion. Dans la figure 6, les courbes sont décroissantes car les valeurs de rappel sont ordonnées de manière décroissante. Les valeurs en abscisses correspondent au classement de la valeur de rappel considérée par rapport aux autres valeurs de rappel.

Il faut noter que toutes les courbes ne doivent pas être toutes comparées de la même façon. Les courbes notées (1) et (2), en référence à la stratégie utilisée peuvent être directement comparées. Concernant la courbe notée (i), en référence aux systèmes initiaux utilisés de façon non fusionnée, le lecteur portera d'abord son attention sur les 10 premières valeurs, correspondant aux 10 meilleurs systèmes.

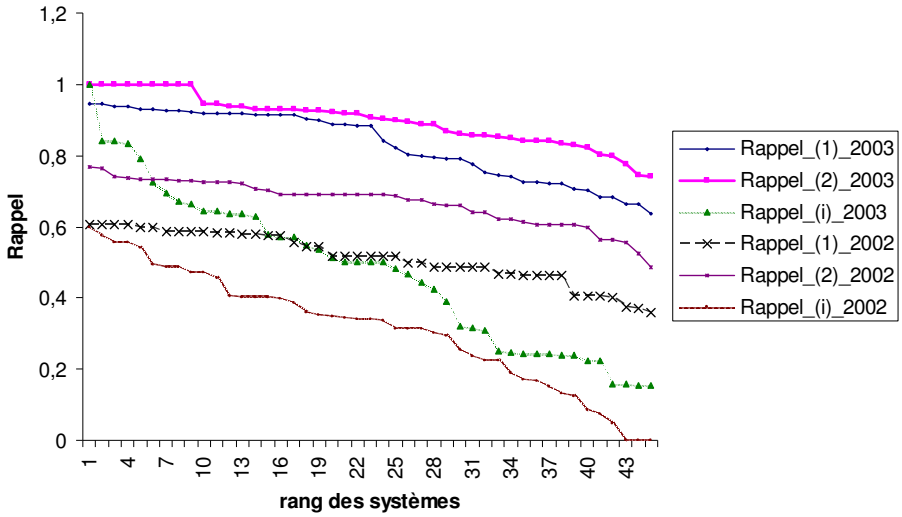

Figure 6 : Comparaison des mesures de rappel obtenues par chaque stratégie de fusion pour la fusion par union

La figure 6 montre que les deux stratégies proposées améliorent les performances initiales des systèmes. Les 10 premiers résultats obtenus avec la stratégie2 en 2003 obtiennent un rappel égal à 1.

Le tableau 6 indique la valeur de l'amélioration moyenne obtenue lors des fusions. Dans ce tableau, pour la stratégie 1 (sélection des meilleurs systèmes basée sur la mesure F), nous comparons la moyenne sur l'ensemble des combinaisons 2 à 2 des systèmes (45 couples de systèmes formés par la fusion des 10 systèmes sélectionnés par la stratégie 1) avec les performances du meilleur système (valeur du rappel du système étant le meilleur par rapport à la mesure F) et avec la moyenne du rappel obtenu par les systèmes utilisés isolément. La moyenne des systèmes utilisés isolément revient à calculer le rappel moyen obtenu par les 10 systèmes sélectionnés par la stratégie1. Dans ce même tableau, les performances de la stratégie2, qui sélectionne les meilleurs systèmes en termes de rappel sont comparées d'une part avec celles du meilleur système (sélectionné de la même façon) et avec celles obtenues en moyenne par les systèmes sélectionnés pris séparément. Les pourcentages indiquent l'amélioration obtenue par la moyenne des fusion 2 à 2, comparativement à meilleur système simple et par rapport à la moyenne des systèmes pris séparément.

		2002	2003
Stratégie 1	Meilleur système simple	0,404	0,792
	Moyenne systèmes simples	0,430	0,729
	Moyenne fusion 2 à 2	0,513 (27%-19,2%)	0,830 (4,9%-13,9%)
Stratégie 2	Meilleur système simple	0,597	0,999
	Moyenne systèmes simples	0,523	0,770
	Moyenne fusion 2 à 2	0,668 (11,9%-27,8%)	0,900 (0,1%-16,8%)

Tableau 6: Valeurs de rappel moyen pour la fusion par union des systèmes 2 à 2

Dans le tableau 6, pour la stratégie1, on remarque que le rappel moyen obtenu par les systèmes simples (0,4305) est supérieur au rappel du meilleur système sélectionné par la stratégie1 (0,404). Cela s'explique par le fait que le système détecté comme étant le meilleur avec la stratégie1, par rapport à la mesure F (Thunv3) est classé en 12ème position par rapport au rappel des autres systèmes. Dans ce cas, en appliquant la fusion par union sur les systèmes sélectionnés avec la stratégie1, on obtient une amélioration du rappel moyen d'environ 19% par rapport au rappel moyen des systèmes simple. En 2003, le rappel moyen des systèmes simples est inférieur au rappel du meilleur système avec la stratégie1 (ce système est classé en 5ème position par rapport au rappel des autres systèmes).

D'autre part, le tableau 6 montre bien que quelque soit l'année et la stratégie de fusion, en moyenne, les fusions améliorent les résultats par rapport à la moyenne des systèmes utilisés séparément. Par exemple, en utilisant la stratégie 1, pour l'année 2002, alors qu'en moyenne les meilleurs systèmes obtiennent 0,430 ; la fusion permet d'obtenir 0,531, soit une augmentation d'environ 19%. Bien évidemment, la stratégie 2 permet d'obtenir globalement de meilleurs résultats que la stratégie 1 (puisque les meilleurs systèmes sont choisis pour leur maximum de performance par rapport au rappel, mesure étudiée). L'augmentation relative est cependant plus importante. Par exemple, toujours pour l'année 2002, en utilisant la stratégie 2, alors qu'en moyenne les meilleurs systèmes obtiennent 0,523 ; la fusion permet d'obtenir 0,668, soit une augmentation d'environ 28% environ au lieu de 19%. De la même façon, pour l'année 2003, l'augmentation par rapport à la moyenne est de 14% pour la stratégie 1 alors qu'elle est de 17% environ pour la stratégie 2.

Pour la stratégie2, le meilleur système obtient un rappel supérieur au rappel moyen des systèmes simples. On constate alors en 2003 que le rappel moyen à l'issue de la fusion des 10 meilleurs systèmes est inférieur au rappel du meilleur système. La conclusion que l'on peut tirer est que les 9 autres meilleurs systèmes retrouvent tous des sous ensembles de l'ensemble des documents pertinents que le meilleur système restitue (le deuxième meilleur système obtient un rappel de 0,84 contre 0,999 pour le meilleur système).

6.3. Comparaison des stratégies pour l'intersection (précision)

Dans la figure 7, les résultats obtenus montrent une faible différence entre les performances de la stratégie1 et de la stratégie2 en 2002 pour les 10 meilleurs résultats. Cette faible différence s'explique par le fait que 80% des systèmes sélectionnés à travers leur valeur de précision et ceux sélectionnés grâce à leur valeur de F-mesure sont identiques (cf. tableau 5). De plus, l'intersection montre un accord entre les systèmes sur les documents retrouvés.

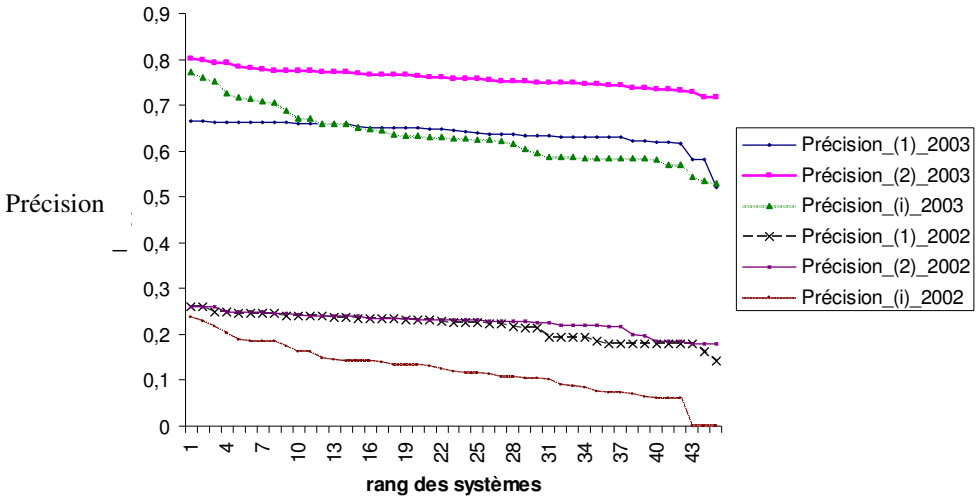

Figure 7 : Comparaison des mesures de précision obtenues par chaque stratégie de fusion pour la fusion par intersection

Le tableau 7 contient le même type de comparaisons que le tableau 6, mais en termes de précision.

		2002	2003
Stratégie 1	Meilleur système simple	0,204	0,597
	Moyenne systèmes simples	0,1857	0,5951
	Moyenne fusion 2 à 2	0,2169 (+16,8%)	0,6396 (+7,5%)
Stratégie 2	Meilleur système simple	0,236	0,774
	Moyenne systèmes simples	0,1959	0,7218
	Moyenne fusion 2 à 2	0,2268 (+15,77%)	0,7598 (+5,27%)

Tableau 7: Valeurs de précision moyenne pour la fusion par intersection des systèmes 2 à 2

Le tableau 7 montre qu'en moyenne la fusion deux à deux améliore les résultats. L'augmentation est plus conséquente en 2002 quelle que soit la stratégie : cela s'explique en partie par les faibles résultats de départ comparativement à 2003. La précision est augmentée de presque 17% avec la stratégie 1 en 2002 entre la moyenne des systèmes simples et la moyenne des fusions deux à deux. Cette augmentation est d'un peu plus de 7% en 2003. Concernant la stratégie 2, l'augmentation est de presque 16% en 2002 et de 5% en 2003. Ce résultat est intéressant dans la mesure où il n'implique pas d'avoir détecté « Le » meilleur système au préalable à la fusion ni au préalable à une recherche.

7. Discussions et conclusion

Les résultats des expérimentations que nous avons présentées dans cet article montrent quels sont les impacts des fusions par union et par intersection. Nous avons montré qu'il était possible, par cette simple stratégie d'améliorer fortement le rappel et donc de constituer un ensemble important de documents liés à un sujet donné. Ainsi, en 2002, en fusionnant le meilleur système en termes de mesure F avec un des 10 autres meilleurs systèmes, le rappel est augmenté de 48%. Lorsqu'un ensemble de systèmes est utilisé pour la fusion par union, pour l'année 2002, le rappel est augmenté d'environ 19% (moyenne des fusions par rapport à la moyenne des systèmes simples) pour la stratégie 1. En utilisant la stratégie 2, la fusion permet d'obtenir une augmentation d'environ 28%. De la même façon, pour l'année 2003, l'augmentation par rapport à la moyenne est de 14% pour la stratégie 1 alors qu'elle est de 17% environ pour la stratégie 2. Par la fusion par intersection, la précision est augmentée de presque 17% avec la stratégie 1 en 2002. Cette augmentation est d'un peu plus de 7% en 2003. Concernant la stratégie 2, l'augmentation est de presque 16% en 2002 et de 5% en 2003.

La mesure des améliorations ainsi obtenue permet de retenir ce type de stratégie dans le cadre de constitution de corpus. La constitution de corpus à de nombreux domaines d'application, que ce soit dans la veille scientifique (Dousset et Mothe 2004) ou dans la construction automatisée d'ontologie de domaine (Hernandez et al., 2006).

Parallèlement, nous avons montré qu'il était aussi possible d'améliorer fortement la précision. En 2002, la fusion du meilleur système avec un des meilleurs systèmes en termes de mesure F permet d'augmenter la précision de 15 % (9% en 2003).

Cette étude pourrait être complétée selon différents axes :

- en premier lieu, nous nous sommes intéressés à la fusion 2 à 2. L'étude pourrait être approfondie en fusionnant plus de deux systèmes,
- cette étude s'appuie sur une connaissance *a priori* des performances des systèmes (fusion des meilleurs systèmes). Ce choix est justifié par le fait que nous souhaitions mesurer les gains relatif à la fusion de systèmes performants. Il était donc nécessaire de les identifier. La généralisation de l'étude serait intéressante ; elle nécessiterait alors de disposer d'un certain nombre de systèmes que l'on puisse faire fonctionner pour un échantillon plus important de requêtes. Ce n'est malheureusement pas le cas ; aucun programme d'évaluation ne met à disposition les outils.
- Cette étude fait abstraction des caractéristiques des systèmes de recherche d'information pour se concentrer sur leurs résultats globaux. Il serait intéressant d'étudier les différents paramètres de chacun des systèmes afin d'analyser s'il existe une corrélation entre leur complémentarité en termes de documents retrouvés et leurs différences en termes de modèles ou techniques utilisées.

- l'objectif de cette étude était de s'intéresser d'une part au rappel, d'autre part à la précision ; certaines tâches devant privilégier l'une ou l'autre de ces mesures. Nous souhaitons poursuivre cette étude en nous focalisant plutôt sur une mesure globale (mesure F ou précision moyenne [mean average precision]). Dans ce cas, nous souhaitons étudier une combinaison moins systématique que l'intersection et l'union telle que nous les avons appliquées. Il s'agirait de définir des critères permettant de décider *a priori* quelle stratégie (fusion par intersection ou par union) devrait être utilisée en fonction de la requête en cours de traitement. Cette stratégie pourrait s'appuyer sur les taux de chevauchement des ensembles de documents retrouvés.

Références

Beitzel S.M., Jensen E.C., Chowdhury A., Grossman D., Frieder O., and Goharian N. (2004). *Fusion of effective retrieval strategies in the same information retrieval system*. In Journal of the American Society Information Science Technologies, 55(10), 859–868.

Dousset B., Mothe J., (2004). *Mining document contents in order to analyse a scientific domain*. In RC33 Sixth International Conference on Social Science Methodology, Amsterdam, Barbara Budrich Publishers , (support électronique).

Farah M., Vanderpooten D., (2007). *An outranking approach for rank aggregation*, In International ACM SIGIR Conference on Research and Development in Information Retrieval, 591–598.

Fox E.A, Shaw J.A., (1994). *Combination of multiple searches*, In Text Retrieval Conference (TREC-2), NIST special publication, 243–252.

Hernandez N., Chrisment C., Hubert C., Mothe J., (2006). *Mise à jour d'une ontologie de domaine à partir de l'analyse de nouveaux documents du domaine pour l'indexation de documents*, In Information - Interaction - Intelligence, Cépaduès Editions, Numéro spécial Textes et ressources terminologiques et/ou ontologiques : évolution et maintenance, Vol. Hors-série, 53–83.

Kantor P. B., Ng Kwong B., (2000). *Predicting the effectiveness of Naïve data fusion on the basis of system characteristics*, In Journal of the American Society for Information Science archive, 51(13), 1177 – 1189.

Kompaore D., Mothe J., Lemoing, E., (2006). *Fusion de systèmes pour la recherche de passages dans les textes*, Conférence francophone en Recherche d'Information et Applications (CORIA 2006), 295–300.

Lee, J., (1997). *Analysis of multiple evidence combination*, In International ACM SIGIR Conference on Research and Development in Information Retrieval, 267–276.

Lillis D., Toolan F, Peng L., Collier R., Dunnion J., (2006). *Probability-based fusion of information retrieval result sets*, In International ACM SIGIR Conference on Research and Development on Information Retrieval, 139–146.

Soboroff I., Nicholas C., et Cahan P., (2001). *Ranking retrieval systems without relevance judgements*, In Proceedings of 24th Annual International ACM SIGIR Conference, 66–73.

Spoerri A., (2007). *Examining the Authority and Ranking Effects as the result list depth used in data fusion is varied*, In Information Processing and Management: an International Journal, 43(4), 1044-1058.

Tukey J.W., (1977). *Exploratory data analysis*. EDA, Reading, MA, (Addison Wesley).

Vogt C.C., Cottrell G.W., (1998). *Predicting the performance of linearly combined IR systems*, In International ACM SIGIR Conference on Research and Development in Information Retrieval, 190–196.

Voorhees E.M., Gupta N.K., et Johnson-Laird B., (1994). *The collection fusion problem*. In 3rd Annual Text Retrieval Conférence (TREC-3), NIST.

Wu S., McClean S., (2006). *Performance prediction of data fusion for information retrieval*, In Information Processing and Management: an International Journal, 42(4), 899-915.