Guillaume Cabanac, Gilles Hubert, Mohand Boughanem, Claude Chrisment. Impact du « biais des ex aequo » dans les évaluations de Recherche d'Information. Dans : Document numérique, Hermès, Vol. 14, 2011 (à paraître).

Impact du « biais des *ex aequo* » dans les évaluations de Recherche d'Information

Guillaume Cabanac — Gilles Hubert
Mohand Boughanem — Claude Chrisment

Université de Toulouse, IRIT UMR 5505 CNRS
118 route de Narbonne, F-31062 Toulouse cedex 9
{cabanac, hubert, boughanem, chrisment}@irit.fr

RÉSUMÉ. *Cet article considère la problématique de l'évaluation en Recherche d'Information, en particulier dans le cadre de* TREC *avec le programme* trec_eval. *Nous montrons que les systèmes de RI ne sont pas uniquement évalués en fonction de la pertinence des documents qu'ils restituent. En effet, dans le cas de documents* ex aequo *(trouvés avec le même score) leur nom est utilisé pour les départager. Nous assimilons cette façon de départager les* ex aequo *à un biais expérimental qui influence les scores attribués aux systèmes, et argumentons en faveur d'une stratégie pour les départager plus équitablement. L'étude de 22 éditions de* TREC *révèle une différence significative entre la stratégie conventionnelle et inéquitable de* trec_eval *et les stratégies équitables proposées. Ces résultats expérimentaux suggèrent l'intégration des stratégies proposées dans* trec_eval *afin d'encourager la réalisation d'expérimentations plus équitables.*

ABSTRACT. *We consider Information Retrieval evaluation in the* TREC *framework with the* trec_eval *program. It appears that IR systems obtain scores regarding not only the relevance of retrieved documents, but also according to document names in case of ties, i.e., documents retrieved with a same score. We consider this tie-breaking strategy as an uncontrolled parameter influencing measure scores, and argue the case for fairer tie-breaking strategies. A study of 22* TREC *editions reveals significant difference between the conventional unfair* trec_eval *strategy and the fairer strategies that we propose. This experimental result advocates integrating these fairer strategies into* trec_eval *for conducting fairer experiments.*

MOTS-CLÉS : *Recherche d'information, évaluation, expérimentation, biais expérimental.*

KEYWORDS: *Information retrieval, measurement, experiment, uncontrolled parameter.*

Ces travaux ont été en partie réalisés dans le cadre du programme QUAERO, financé par OSEO, agence française pour l'innovation.

1. Introduction

La Recherche d'Information (RI) est un domaine caractérisé par une longue tradition d'expérimentation initiée dès les années 1960 (Robertson, 2008). Depuis 1992, les campagnes d'évaluation de RI TREC offrent aux universitaires et industriels l'opportunité de mesurer l'efficacité[1] de leurs systèmes et d'en discuter les aspects théoriques et pratiques (Harman, 1993; Voorhees *et al.*, 2005). Dans ce contexte, les résultats d'évaluation des systèmes de RI (SRI), c'est-à-dire des moteurs de recherche, sont calculés par le programme trec_eval (NIST, n.d.). Ce savoir-faire a été capitalisé par de nombreuses autres initiatives d'évaluation de la RI qui exploitent également ce programme, comme de nombreuses tâches des campagnes NTCIR (Kando *et al.*, 1999) et CLEF (Peters *et al.*, 2001).

Or, pour toute expérimentation scientifique, une règle fondamentale impose d'identifier les n paramètres en jeu, puis de tous les fixer excepté celui dont on désire mesurer l'impact sur l'artefact considéré. Il est crucial de s'assurer qu'un seul paramètre pourra varier (par conséquent que les $n - 1$ autres paramètres sont fixés) sans quoi les conclusions que l'on tirerait pourraient être ambiguës, deux paramètres ou plus ayant varié en même temps pendant l'expérimentation. Dans la lignée des analyses portant sur la méthodologie d'évaluation en RI telles que (Voorhees, 1998; Zobel, 1998), nous avons identifié un biais expérimental dans TREC au travers de trec_eval : les scores attribués aux SRI (valeurs des mesures) ne dépendent pas seulement des documents qu'ils restituent mais aussi du nom de ces derniers en cas d'*ex aequo*. C'est un problème important car des SRI « chanceux » (resp. « malchanceux ») peuvent obtenir de meilleurs (resp. pires) scores qu'ils ne mériteraient dans un cadre d'évaluation non biaisé.

Afin d'évaluer l'impact de ce biais sur les évaluations de RI, cet article est organisé comme suit. La section 2 décrit comment les SRI sont communément évalués dans les campagnes de type TREC. La section 3 détaille la limite que nous avons identifiée dans trec_eval : les documents restitués avec un score de pertinence identique sont départagés et réordonnés en fonction de leur nom. Cette stratégie introduit un biais expérimental qui n'est pas souhaitable. La section 4 propose deux nouvelles stratégies visant à éliminer l'impact du biais expérimental identifié. La section 5 compare la stratégie actuelle de trec_eval avec nos propositions sur 22 éditions des campagnes *adhoc*, *routing*, *filtering* et *web* de TREC qui se sont déroulées de 1993 à 2004 et mesure la significativité de la différence observée. La section 6 discute ces analyses et leurs limites. Enfin, la section 7 présente un panorama des recherches liées à la validation de la méthodologie d'évaluation en RI, avant de conclure cet article en donnant un aperçu des perspectives à ce travail de recherche.

---

1. En anglais, *effectiveness* décrit la capacité des systèmes à restituer à l'utilisateur des documents pertinents, alors que le terme *efficiency* fait référence aux performances du moteur de recherche (temps de réponse, etc.) sans considérer la pertinence des résultats. Dans cet article, nous employons « efficacité » pour désigner *effectiveness*.

2. Évaluation de l'efficacité des systèmes de RI

Cette section introduit les concepts d'évaluation en RI qui sont manipulés dans cet article. Cette présentation succincte n'a pas vocation à l'exhaustivité et pourra être complétée par les travaux de Buckley *et al.* (2005) qui détaillent davantage l'évaluation de la RI dans le cadre des campagnes TREC, grâce au programme trec_eval.

2.1. *Concepts fondamentaux pour l'évaluation de la RI*

La campagne d'évaluation TREC propose chaque année au moins une tâche de RI. Une tâche est généralement constituée d'au moins 50 *topics* qui représentent autant de besoins en information exprimés par un individu. Chaque *topic* est au minimum caractérisé par son identifiant (qid), un titre (title), une description ainsi qu'une narration décrivant l'information que l'individu recherche et qu'il considérerait pertinente. Par exemple, le *topic* qid = 54 est intitulé *« Contrats de lancement de satellites »* et la narration associée est *« Un document pertinent mentionnera la signature d'un contrat ou d'un accord préliminaire, ou la tentative d'une réservation visant le lancement d'un satellite commercial »*.[2]

Pour un participant, prendre part à une tâche nécessite de fournir aux organisateurs de l'évaluation au moins un *run* : la liste des documents restitués pour chaque *topic* traité, classés par pertinence décroissante. Le tableau 1(a) illustre un extrait de fichier *run* tel que présenté dans la documentation de trec_eval (NIST, n.d.), où chaque ligne est composée de six champs.

(a) Fichier *run*.

qid	iter	docno	rank	sim	run_id
030	Q0	ZF08-870	0	4238	prise1

(b) Fichier *qrels*.

qid	iter	docno	rel
030	Q0	ZF08-870	1

Tableau 1. *Champs des fichiers* qrels *et* run *et exemples de ligne valide.*

Nous ne détaillerons que les champs utilisés par trec_eval pour calculer les valeurs des mesures, les autres champs étant ignorés (NIST, n.d.). Le champ qid est l'identifiant du *topic*, docno est l'identifiant du document restitué et sim représente le score de similarité associé, c'est-à-dire la valeur que le SRI a calculée entre le document docno et la requête qid.

2. "Satellite Launch Contracts. A relevant document will mention the signing of a contract or preliminary agreement, or the making of a tentative reservation, to launch a commercial satellite."

2.2. *Stratégie conventionnelle pour départager les documents* ex aequo *d'un* run

Afin d'évaluer l'efficacité des SRI, leurs *runs* sont comparés avec une « vérité terrain » : la perception humaine de la pertinence. Comme une collection TREC contient généralement entre 800 000 et un million de documents (Voorhees, 2007), il s'avère impossible de juger leur pertinence pour chaque *topic*. C'est pourquoi TREC met en œuvre la technique du *pooling* : pour chaque *topic* t, un *pool* de documents est constitué à partir des 100 premiers documents restitués par chacun des systèmes participant à la campagne d'évaluation, les doublons sont supprimés (opération d'union ensembliste). L'hypothèse est que le nombre et la diversité des SRI permettra de trouver de nombreux documents pertinents. Enfin, un individu appelé « assesseur » examine chaque document du *pool* afin d'identifier s'il répond ou pas au besoin d'information spécifié dans le *topic* t. Le document est alors qualifié de pertinent ou de non-pertinent.

Suite à cette opération manuelle de jugement, les documents pertinents pour chaque *topic* sont connus. Ils constituent les « jugements de pertinence » et sont regroupés dans un fichier *qrels* (*"query relevance judgments"*). Le tableau 1(b) présente la structure d'un tel fichier : qid est un identifiant de *topic*, iter est ignoré, docno est un identifiant de document et rel représente la pertinence du document docno par rapport au *topic* qid. La valeur $0 <$ rel < 128 dénote un document pertinent. Les autres valeurs, en particulier rel = 0, dénotent un document non-pertinent.

À partir d'un *run* et des *qrels*, le programme trec_eval calcule les valeurs des mesures d'évaluation, par exemple la MAP (cf. section 2.3). Pour ce faire, il pré-traite en premier lieu le fichier *run*. Le champ rank étant ignoré, il classe les documents comme suit : *« des rangs internes [à* trec_eval*] sont calculés en classant [les documents] selon le champ* sim*, les* ex aequo *étant départagés de façon déterministe (en utilisant le champ* docno*). »*[3] (NIST, n.d.). Par ailleurs, Buckley *et al.* (2005, p. 55) commentent cette stratégie en soulignant à quel point il est important de départager les *ex aequo* :

> « Pour TREC-1 [...] Le système [participant] assignait également une valeur rank à chaque document, toutefois trec_eval ignorait délibérément ce rang. À la place, trec_eval calculait son propre classement des 200 premiers documents[4] à partir des valeurs RSV[5] pour s'assurer que les documents *ex aequo* soient départagés de façon identique et indépendamment du système (l'ordre des documents *ex aequo*, bien qu'arbitraire, était alors homogène quel que soit le *run*). Départager les *ex aequo* équitablement revêtait une grande importance à l'époque car de nombreux systèmes produisaient un grand nombre d'*ex aequo* – les modèles de RI booléen et à

3. "internally ranks are assigned by sorting by the sim field with ties broken deterministicly (using docno)."
4. À partir de TREC-2 les 1 000 premiers sont considérés (Buckley *et al.*, 2005, p. 58).
5. *Retrieval Status Value*, faisant référence au champ sim dans le tableau 1(a).

niveau de coordination pouvaient produire des centaines de documents avec le même RSV » [6] (Buckley *et al.*, 2005, p. 55).

La section suivante décrit le calcul des valeurs des mesures – scores des SRI reflétant leur efficacité – à partir des *runs* et des *qrels*.

2.3. *Mesures fondamentales pour l'évaluation de la RI*

Suite au réordonnancement d'un *run*, trec_eval calcule plusieurs mesures en fonction des *qrels*. Pour un système s donné, nous considérons dans cet article les mesures ci-dessous, ayant l'intervalle réel $[0; 1]$ comme domaine de définition. Le lecteur intéressé pourra se référer à (Manning *et al.*, 2008, ch. 8) pour davantage de détails.

Reciprocal Rank $RR(s,t) = {}^{1}/_{Rang(t,d)}$ est l'inverse du rang du premier document pertinent d restitué pour le *topic* t.

Precision $P(s,t,d) = \frac{card\left(\left\{d' | \left(Rang(t,d') \leqslant Rang(t,d)\right) \wedge Pert(t,d')\right\}\right)}{Rang(t,d)}$ est la précision de s lorsque la liste des documents restitués pour t est considérée jusqu'à d. Cette valeur dépend du nombre de documents pertinents d' classés jusqu'à d dans la liste des résultats, c'est-à-dire avec un rang inférieur ou égal. La fonction booléenne $Pert(t,d')$ retourne vrai lorsque d' est pertinent pour t, faux dans le cas contraire (information présente dans le fichier *qrels*).

Average Precision $AP(s,t) = \frac{\sum_{d \in Run(s,t) | Pert(t,d)} P(s,t,d)}{NbPert(t)}$ est la précision moyenne des documents pertinents restitués $Run(s,t)$ pour t. Cette valeur dépend des précisions des documents restitués $P(s,t,d)$ ainsi que du nombre de documents pertinents $NbPert(t)$ pour t. Notons que $AP(s,t)$ n'est pas la moyenne arithmétique des précisions de la liste des résultats.

Mean Average Precision $MAP(s) = \frac{1}{|T|} \sum_{t \in T} AP(s,t)$ est la moyenne arithmétique des précisions moyennes, calculée à partir de l'ensemble T des *topics* à traiter dans la tâche.

La section suivante expose la problématique liée à la façon dont trec_eval départage les documents *ex aequo*. Nous montrons que les choix réalisés dans trec_eval constituent un biais dans les évaluations, que nous nommons « biais des *ex aequo* ».

---

6. "For TREC-1 [...] Each document was also assigned a rank by the system, but this rank was deliberately ignored by trec_eval. Instead, trec_eval produced its own ranking of the top two hundred documents based on the RSV values to ensure consistent system-independent tie breaking among documents that a system considered equally likely to be relevant (the ordering of documents with tied RSV values was arbitrary yet consistent across runs). Breaking ties in an equitable fashion was an important feature at the time since many systems had large number of ties—Boolean and coordination-level retrieval models could produce hundreds of documents with the same RSV." (Buckley *et al.*, 2005, p. 55)

3. Le biais des *ex aequo* influençant les résultats d'évaluations en RI

Considérons dans la figure 1 (partie gauche) l'extrait d'un *run* comprenant uniquement les trois premiers documents restitués pour un *topic* t donné (qid = 031). Supposons que $NbPert(t) = 5$, dont le document **WSJ5** (en gras). Comme trec_eval ignore le champ rank, il réordonne le *run* selon le qid croissant, les sim décroissantes, puis selon les docno décroissants pour départager les *ex aequo* (par la suite, « asc » désigne un tri croissant et « desc » un tri décroissant). La liste de documents alors obtenue est présentée en partie droite de la figure 1, où le document pertinent **WSJ5** se voit assigner le rang numéro 1. Notons que le rang inverse vaut $RR(s,t) = 1$, la précision à **WSJ5** vaut $P(s,t,\textbf{WSJ5}) = 1$ et $AP(s,t) = 1/5$.

qid	docno	sim	rank
031	LA12	0,8	1
031	**WSJ5**	0,8	2
031	FT8	0,5	3

→

qid	docno	sim	$RR(s,t)$	$P(s,t,d)$	$AP(s,t)$
031	**WSJ5**	0,8		1	
031	LA12	0,8	1	$1/2$	$1/5$
031	FT8	0,5		$1/3$	

Figure 1. *Réordonnancement effectué par* trec_eval *et évaluation d'un* run*.*

À présent, sans effectuer aucun changement au contenu des documents manipulés dans cet exemple, supposons que le document pertinent **WSJ5** ait été nommé **AP8**. La figure 2 illustre le *run* soumis ainsi que le résultat du processus de réordonnancement de trec_eval : le document pertinent **AP8** est initialement en position 2 (position de **WSJ5**). Suite au réordonnancement, LA12 obtient la première place en considérant un tri par docno décroissant, restant devant **AP8**. On notera alors que le rang inverse, la précision à ce document et la précision moyenne ont été divisés par deux.[7]

qid	docno	sim	rank
031	LA12	0,8	1
031	**AP8**	0,8	2
031	FT8	0,5	3

→

qid	docno	sim	$RR(s,t)$	$P(s,t,d)$	$AP(s,t)$
031	LA12	0,8		0	
031	**AP8**	0,8	$1/2$	$1/2$	$1/10$
031	FT8	0,5		$1/3$	

Figure 2. *Influence du nommage des documents sur les mesures RR, P et AP.*

Cet exemple minimal suffit à illustrer la problématique considérée dans cet article : les valeurs des mesures obtenues par un SRI ne dépendent pas uniquement de sa capacité à restituer des documents pertinents car le nom des documents rentre également en compte pour départager les *ex aequo*. Considérer le champ docno à cet effet sous-entend alors que la collection *Wall Street Journal* (documents WSJ*) est dans l'absolu, quel que soit le *topic*, plus pertinente que la collection *Associated Press* (documents AP*), ce qui est évidemment faux. Cette pratique introduit un biais expérimental qui

7. Une démonstration de ce phénomène, illustré avec des copies d'écran trec_eval, est présentée sur http://www.irit.fr/~Guillaume.Cabanac/tie-breaking-bias.

rend injuste les comparaisons dans les deux cas ci-dessous où nous considérons AP, pour autant la discussion est généralisable à d'autres mesures :

1) pour la *comparaison entre systèmes* où l'on considère les $AP(s_1, t)$ et $AP(s_2, t)$ obtenues par les systèmes s_1 et s_2 pour un *topic* t donné. Cette comparaison est inéquitable parce que les valeurs de cette mesure peuvent différer alors que les deux systèmes ont restitué la même liste de résultats $[P_{0.8}, N_{0.8}, N_{0.5}]$ où P_x (resp. N_x) représente un document pertinent (resp. non-pertinent) restitué avec une similarité sim $= x$. Par exemple, cette situation se produit si l'on assimile le *run* de la figure 1 au SRI s_1, et le *run* de la figure 2 au SRI s_2. En effet, $AP(s_1, t) = 1/1 \cdot 1/5 = 1/5$ alors que $AP(s_2, t) = 1/2 \cdot 1/5 = 1/10$, correspondant à une différence inéquitable et non souhaitable de 200 %.

2) pour la *comparaison entre* topics où l'on considère les $AP(s, t_1)$ et $AP(s, t_2)$ obtenues par un même système s pour deux *topics* t_1 et t_2 distincts. Une telle comparaison est notamment réalisée dans la tâche *robust* de TREC (Voorhees, 2004) pour distinguer les requêtes faciles des difficiles. Elle est inéquitable parce que le processus de réordonnancement de trec_eval a pu profiter au système s pour le *topic* t_1 (en réordonnant les documents pertinents *ex aequo* plus haut qu'initialement) tout en l'ayant défavorisé pour t_2 (en réordonnant les documents pertinents *ex aequo* plus bas qu'initialement). Par conséquent, le concepteur du SRI pourrait entreprendre une analyse approfondie visant à comprendre pourquoi son système est moins performant sur t_2 que sur t_1 alors que la différence de scores pourrait uniquement résulter de malchance, faisant que les documents pertinents ont été réordonnés plus bas dans la liste des résultats qu'initialement, et ce uniquement à cause de leur nom. Imaginons que les documents pertinents proviennent tous de la collection AP, ils seront alors pénalisés car mis en bas de la liste de résultats en cas de réordonnancement par docno décroissant.

Départager les documents *ex aequo* comme le fait trec_eval actuellement introduit un biais qui affecte les résultats d'évaluations en RI. Afin de pallier cette problématique, la section suivante propose des stratégies de réordonnancement plus équitables.

4. Stratégies « réaliste » et « optimiste » pour départager les *ex aequo*

La stratégie actuelle permettant de départager les *ex aequo* (qid asc, sim desc, docno desc) introduit un biais expérimental car elle repose sur le champ docno pour réordonner les documents caractérisés par une même valeur de similarité sim. Cette section présente deux stratégies originales visant à départager les *ex aequo* sans pour autant être affectées par le biais discuté dans cet article. À cet effet, la figure 3 illustre « *run* $=\!\!\bowtie_{\text{qid, docno}}$ *qrels* », c'est-à-dire la jointure externe gauche entre les relations (construites à partir des fichiers) *run* et *qrels* portant sur les champs qid et docno (seule une partie des attributs est représentée). Cet opérateur de l'algèbre relationnelle conserve les tuples de la relation *run*, la valeur du champ rel étant ajoutée si le document est présent dans la relation *qrels*. C'est le cas pour les documents issus du *pool* qui ont été jugés. Dans le cas contraire, pour les documents non présents dans le fichier *qrels* pour le *topic* considéré, la valeur rel $= 0$ est produite.

(a) *run* $\rtimes\!\!\!\!\bowtie_{\text{qid, docno}}$ *qrels*

qid	docno	sim	rank	rel
8	**CT5**	0,9	1	1
8	AP5	0,7	2	0
8	WSJ9	0,7	3	0
8	**AP8**	0,7	4	1
8	FT12	0,6	5	0

(b) R. réaliste

qid	docno	sim	rel
8	**CT5**	0,9	1
8	WSJ9	0,7	0
8	AP5	0,7	0
8	**AP8**	0,7	1
8	FT12	0,6	0

(c) R. conventionnel (TREC)

qid	docno	sim	rel
8	**CT5**	0,9	1
8	WSJ9	0,7	0
8	**AP8**	0,7	1
8	AP5	0,7	0
8	FT12	0,6	0

(d) R. optimiste

qid	docno	sim	rel
8	**CT5**	0,9	1
8	**AP8**	0,7	1
8	WSJ9	0,7	0
8	AP5	0,7	0
8	FT12	0,6	0

Figure 3. *Trois stratégies de réordonnancement pour « run* $\rtimes\!\!\!\!\bowtie_{\text{qid, docno}}$ *qrels ».*

La figure 3 représente aussi la stratégie conventionnelle implantée dans trec_eval pour départager les *ex aequo*, ainsi que les deux stratégies que nous proposons :

1) le *réordonnancement réaliste* postule qu'au sein d'un groupe de documents *ex aequo* les non-pertinents devraient être ordonnés avant les pertinents car le SRI n'a pas été capable de les distinguer (en y affectant des scores sim différents). L'expression de tri correspondant à cette spécification est « qid asc, sim desc, rel asc, docno desc ».

$$\text{Exemple : } [P_x, N_x, P_x] \xrightarrow[\text{qid asc, sim desc, rel asc, docno desc}]{\text{réordonnancement } \textit{réaliste}} [N_x, P_x, P_x]$$

2) le *réordonnancement optimiste* postule qu'au sein d'un groupe de documents *ex aequo* les documents pertinents devraient être ordonnés avant les non-pertinents car le système peut les représenter de façon uniforme, dans un cluster par ex. L'expression de tri correspondant à cette spécification est « qid asc, sim desc, rel desc, docno desc ».

$$\text{Exemple : } [P_x, N_x, P_x] \xrightarrow[\text{qid asc, sim desc, rel desc, docno desc}]{\text{réordonnancement } \textit{optimiste}} [P_x, P_x, N_x]$$

En fonction de la stratégie de réordonnancement sélectionnée – réaliste, conventionnelle ou optimiste – la valeur de la mesure considérée diffère éventuellement. Dans cet article, nous désignons par M_S la mesure M calculée avec la stratégie de réordonnancement $S \in \{R, C, O\}$. Par exemple MAP_O correspond à la mesure MAP appliquée sur la base d'un réordonnancement optimiste. Notons l'ordre total $M_R \leqslant M_C \leqslant M_O$ entre les différentes valeurs de mesures.

Cette section a illustré comment le biais expérimental des *ex aequo* affecte potentiellement les scores obtenus par les SRI en terme de valeurs de mesures. Dans le but de promouvoir des évaluations plus équitables, nous avons proposé les alternatives de stratégies de réordonnancement réaliste et optimiste. Afin d'évaluer l'impact de la

stratégie de réordonnancement sur les résultats des évaluations, nous détaillons dans la section suivante une analyse réalisée à partir des *runs* soumis à TREC.

5. Impact du biais expérimental des *ex aequo* sur l'évaluation des SRI

Nous avons étudié l'impact du biais des *ex aequo* sur les résultats de 4 tâches de TREC : *adhoc* (1993–1999), *routing* (1993–1997), *filtering* (limitée à sa sous-tâche *routing*, 1998–2002) et *web* (2000–2004). Ces 22 éditions comprennent 1 360 *runs* de longueur moyenne 50 196 lignes. L'ensemble représente 3 Go de données issues du site de TREC[8], analysées comme suit. La section 5.1 évalue à quel point les *runs* sont concernés par le biais des *ex aequo* en rapportant la proportion des documents *ex aequo* au sein des *runs*. Puis, la section 5.2 détaille les différences observées entre les valeurs de mesures calculées avec nos propositions de stratégies équitables *versus* avec la stratégie réordonnancement conventionnelle implantée dans trec_eval.

5.1. *Proportion des documents* ex aequo *dans 4 tâches de* TREC

Par la suite, nous appelons « liste-résultat » le sous-ensemble d'un *run* concernant un *topic* t soumis pour une $édition$ donnée. Une liste-résultat sera notée $\texttt{runid}_{\mathsf{qid}}^{édition}$. Dans la mesure où des différences de résultats peuvent apparaître lorsqu'une liste-résultat contient des *ex aequo*, cette section évalue la proportion d'un tel phénomène sur les données précédemment présentées.

Le tableau 2 présente des statistiques associées à chacune des tâches : l'édition considérée et le nombre de *runs* soumis (détaillé par année et agrégé au niveau tâche). Deux autres indicateurs représentent ★le pourcentage d'*ex aequo* dans les listes-résultats ainsi que ☆le nombre d'*ex aequo* par similarité (sim). Les statistiques concernant le minimum (min), la moyenne (moy), le maximum (max) et l'écart-type (é.-t.) de ces données sont également rapportées. Par exemple, la liste-résultat illustrée dans la figure 3 contient ★$3/5 = 60\,\%$ de documents *ex aequo* et ☆est caractérisée par $(1+3+1)/3 = 1{,}7$ documents *ex aequo* par sim en moyenne.

Les tâches considérées ont eu lieu à au moins cinq reprises. Une participation croissante caractérise la tâche initiale *adhoc* alors qu'elle est variable pour les autres tâches avec néanmoins 32 *runs* ou plus par édition. De façon globale, les SRI participant aux premières éditions de TREC *adhoc* ont soumis davantage de résultats avec des *ex aequo* qu'à l'occasion des éditions les plus récentes. Cette observation est en accord avec celle de Buckley *et al.* (2005, p. 55) citée en section 2.2.

L'analyse globale des 22 éditions considérées révèle que 89,6 % des runs soumis contiennent des documents *ex aequo*. Une analyse complémentaire, synthétisée par la figure 4, a porté sur la répartition des *ex aequo* au sein des runs. Elle montre que

8. http://trec.nist.gov/results.html

tâche	édition	nb *runs* soumis	★% d'*ex aequo* par liste-résultat				☆nombre d'*ex aequo* par sim			
			min	moy	max	é.-t.	min	moy	max	é.-t.
adhoc	1993	36	0,0	30,3	100,0	36,0	2,2	4,4	28,0	4,2
	1994	40	0,0	28,4	100,0	35,9	1,9	9,5	37,3	11,2
	1995	39	0,0	29,2	99,9	32,8	1,0	2,8	26,2	4,2
	1996	82	0,0	24,1	100,0	32,3	2,0	4,1	35,1	4,7
	1997	79	0,0	24,7	100,0	34,7	1,8	4,5	25,8	5,1
	1998	103	0,0	19,0	100,0	27,4	1,0	2,5	33,8	4,4
	1999	129	0,0	15,6	100,0	24,6	1,5	3,7	22,9	4,4
	Moyenne des 508 *runs* →		0,0	24,5	100,0	32,0	1,6	4,5	29,9	5,5
filtering	1998	47	0,0	26,8	100,0	40,8	41,0	42,0	51,8	2,2
	1999	55	0,0	7,5	100,0	23,8	2,1	2,1	2,7	0,1
	2000	53	0,0	21,1	100,0	38,1	15,3	22,3	37,1	10,0
	2001	18	0,0	25,6	100,0	30,3	19,8	33,3	69,6	17,0
	2002	17	0,0	34,6	100,0	37,2	2,5	23,3	97,9	33,2
	Moyenne des 190 *runs* →		0,0	23,1	100,0	34,0	16,1	24,6	51,8	12,5
routing	1993	32	0,0	32,9	100,0	39,9	1,1	4,1	38,2	6,0
	1994	34	0,0	31,0	100,0	37,6	2,3	5,5	30,9	5,9
	1995	27	0,0	24,9	99,2	27,4	1,0	1,5	14,7	1,4
	1996	26	0,0	21,3	100,0	24,5	1,4	7,2	40,0	10,7
	1997	34	0,0	27,4	100,0	33,7	6,7	13,0	54,3	10,9
	Moyenne des 153 *runs* →		0,0	27,5	99,8	32,6	2,5	6,3	35,6	7,0
web	2000	104	0,0	29,3	100,0	34,3	2,9	9,3	79,6	16,6
	2001	96	0,0	32,0	100,0	31,9	25,8	27,8	63,8	5,7
	2002	71	0,0	25,8	100,0	33,5	1,0	3,6	44,7	6,3
	2003	164	0,0	18,8	100,0	27,8	1,4	2,3	12,0	1,8
	2004	74	0,0	24,9	100,0	34,4	1,5	4,3	39,6	6,2
	Moyenne des 509 *runs* →		0,0	26,2	100,0	32,4	6,5	9,5	47,9	7,3
Moyenne globale des 1 360 *runs* →			0,0	25,2	100,0	32,7	6,2	10,6	40,3	7,8

Tableau 2. *Proportions des* ex aequo *observées dans 4 tâches de* TREC.

la répartition des *ex aequo* est homogène avec une fréquence légèrement supérieure autour du milieu des listes-résultats (deuxième et troisième quarts).

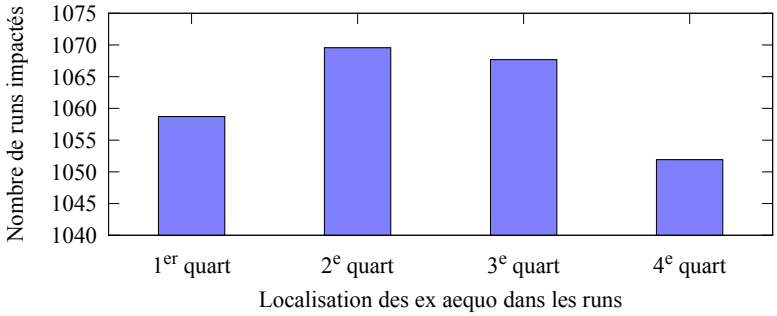

Figure 4. *Nombre de runs contenant des* ex aequo *en fonction de leur rang, regroupés en quarts.*

L'analyse au niveau tâche révèle qu'une liste-résultat contient 25,2 % de documents *ex aequo*, cette proportion étant très variable comme l'indique l'écart-type de 32,7 % en moyenne. La valeur d'écart-type est stable quelle que soit la tâche. De plus, des listes-résultats sans *ex aequo* ($\min^{\star} = 0{,}0$) ont été soumises chaque an-

née. À l'inverse, des listes-résultats uniquement composées d'*ex aequo* ont également été observées chaque année (il y en a 1 338 sur les 4 tâches). Ce phénomène est par exemple illustré par $\texttt{ibmge2}_{291}^{1996}$ dans la tâche *adhoc*, qui est un exemple de non-discrimination : tous les documents de cette liste-résultat possèdent le même score $\mathsf{sim} = -126{,}000000$. De telles listes-résultats sont les plus à même d'obtenir des résultats de mesures tout à fait différents selon la stratégie de réordonnancement utilisée.

Au niveau *run*, en considérant les documents groupés par sim, on observe une grande variabilité : certaines listes-résultats n'ont aucun *ex aequo* ($\min^{\star} = 1{,}0$ ce qui correspond à $\min^{\bigstar} = 0{,}0$) alors que d'autres ont une moyenne allant jusqu'à 97,9 documents par sim identique. La taille moyenne d'un groupe de sim identiques est de 10,6 documents, impliquant qu'un document classé en position $r + 11$ avec une stratégie réaliste peut être reclassé en position r avec une autre stratégie de réordonnancement lorsque le SRI est chanceux (il a gagné 11 places injustement). Généralisant cette observation, signalons que plus un groupe de similarité rassemble un grand nombre de documents, plus la perte ou le gain injuste de position sera important.

Nous avons également analysé le rang moyen des groupes de documents *ex aequo*. Dans le tableau 3, nous constatons le même phénomène pour toutes les tâches. Il existe toujours des groupes de similarité en tête de liste-résultat ($\min = 1$), tout comme certains sont en queue de liste résultat ($\max = 999$). En moyenne, le rang d'un groupe d'*ex aequo* est situé à la moitié de la liste-résultat. L'écart-type d'approximativement 280 positions indique que les groupes d'*ex aequo* sont répartis uniformément du premier au dernier rang. Il n'y a donc pas de position où les *ex aequo* se concentrent.

tâche	min	moy	max	é.-t.
adhoc	1	557,60	999	276,95
filtering	1	541,47	999	284,35
routing	1	525,96	999	285,02
web	1	490,30	999	296,53

Tableau 3. *Rangs minimum, moyen, maximum des groupes de documents* ex aequo. *L'écart-type de ces rangs est également présenté.*

Cette section a montré que les listes-résultats sont susceptibles de contenir de nombreux documents *ex aequo*. Par conséquent, les évaluations de RI peuvent être affectées par le biais que nous avons identifié. L'importance graduelle de cet impact sur le résultat des mesures est évaluée dans la section suivante.

5.2. *Différences des résultats d'évaluation en fonction de la stratégie retenue*

Après avoir estimé la proportion des documents *ex aequo* dans une liste-résultat, nous étudions dans cette section l'impact du biais des *ex aequo* sur les valeurs des mesures présentées en section 2.3. Cette étude est focalisée sur la différence entre M_R et M_C (réaliste *versus* conventionnel) pour montrer à quel point le hasard a injustement

augmenté le score M_R que le SRI aurait obtenu dans un contexte expérimental non biaisé. Notons que dans le pire des cas, l'impact maximal du biais est obtenu avec M_O car les mesures suivent un ordre total $M_R \leqslant M_C \leqslant M_O$, cf. section 4.

Pour chaque mesure, nous présentons en premier lieu les systèmes qui ont le plus bénéficié du biais des *ex aequo* en listant les 5 différences les plus importantes entre la stratégie inéquitable M_C et l'équitable M_R. Puis, nous généralisons ces résultats en rapportant la valeur de significativité des tests statistiques calculée pour la différence $M_C - M_R$ au niveau tâche, en considérant l'ensemble des *runs* associés. Les p-valeurs de significativité sont calculées avec le test t de Student pairé (la différence est calculée entre les paires de valeurs M_C et M_R) et unilatéral (car $M_C \geqslant M_R$). Bien que nécessitant théoriquement une distribution normale des données, Hull (1993) précise que ce test est en pratique robuste aux violations de cette condition. Par ailleurs, Sanderson *et al.* (2005) montrent que ce test est bien plus fiable que d'autres, tel que le test des rangs signés de Wilcoxon. Concrètement, lorsque $p < \alpha$ avec $\alpha = 0{,}05$ la différence entre les deux échantillons testés est qualifiée de statistiquement significative (Hull, 1993). Plus la valeur p est petite, plus la différence est significative. Enfin, les tests de corrélation sont effectués grâce au coefficient de corrélation r de Pearson.

5.2.1. *Impact du biais des* ex aequo *sur la mesure Reciprocal Rank*

Le tableau 4 montre l'impact de la stratégie de réordonnancement choisie sur le rang du premier document pertinent. Il présente les valeurs tronquées à 4 positions décimales de l'inverse du rang RR_x (uniquement pour l'affichage, les calculs étant réalisés en valeur exacte). En complément, nous rapportons les positions des rangs ($1/RR_x$) car cela nous semble plus intelligible. Le tableau 4 est trié par différence $\delta_{RC} = 1/RR_R - 1/RR_C$ décroissante, permettant de se focaliser sur les systèmes les plus chanceux, qui bénéficient le plus du biais des *ex aequo*. Les tests statistiques rapportés dans le tableau 5 indiquent une différence statistiquement significative entre RR_C et RR_R. Par conséquent le premier document pertinent est classé plus haut dans la liste-résultat (différence significative) avec une stratégie conventionnelle qu'avec une stratégie réaliste, sans que le moteur n'ait été amélioré pour autant. Malgré cette différence avérée, il est intéressant de constater une importante corrélation ($\geqslant 99\,\%$) entre les valeurs des mesures obtenues avec les deux stratégies. Le cas de la tâche *filtering* diffère, avec une corrélation plus faible (89 %). En résumé, les valeurs de RR_C et RR_R sont corrélées (augmentation faible) mais suffisamment distinctes pour que leur différence soit statistiquement significative.

5.2.2. *Impact du biais des* ex aequo *sur la mesure Average Precision*

Le tableau 6 montre les cinq SRI les plus affectés pour la mesure AP, pour chaque tâche et stratégie de réordonnancement. De plus, les gains entre les stratégies pairées sont calculés. On s'intéressera en particulier au gain$_{CR}$ entre AP_C et AP_R qui représente le gain injustement obtenu par les SRI bénéficiant du biais des *ex aequo* lors du réordonnancement conventionnel de trec_eval. Ce dernier est par exemple de 406 %

	liste-résultat	RR_R	RR_C	RR_O	$1/RR_R$	$1/RR_C$	$1/RR_O$	δ_{RC}
adhoc	$\texttt{padre2}_{195}^{1994}$	0,0011	0,0667	0,0769	946	15	13	931
	$\texttt{anu5aut1}_{297}^{1996}$	0,0010	0,0149	1,0000	992	67	1	925
	$\texttt{anu5aut2}_{297}^{1996}$	0,0010	0,0149	1,0000	992	67	1	925
	$\texttt{ibmgd2}_{291}^{1996}$	0,0010	0,0133	1,0000	998	75	1	923
	$\texttt{padre1}_{187}^{1994}$	0,0011	0,0556	0,5000	924	18	2	906
filtering	$\texttt{antrpohsu00}_{32}^{2000}$	0,0000	0,5000	1,0000	988	2	1	986
	$\texttt{antrpnohsu00}_{62}^{2000}$	0,0000	0,0909	1,0000	988	11	1	977
	$\texttt{antrpohsu00}_{62}^{2000}$	0,0000	0,0909	1,0000	988	11	1	977
	$\texttt{antrpohsu00}_{51}^{2000}$	0,0000	0,0476	1,0000	992	21	1	971
	$\texttt{antrpohsu00}_{39}^{2000}$	0,0000	0,0714	1,0000	980	14	1	966
routing	$\texttt{cir6roul}_{118}^{1997}$	0,0010	0,1429	1,0000	970	7	1	963
	$\texttt{cir6roul}_{161}^{1997}$	0,0010	0,0250	0,1429	998	40	7	958
	$\texttt{virtue3}_{228}^{1997}$	0,0011	0,2000	0,5000	949	5	2	944
	$\texttt{cir6roul}_{11}^{1997}$	0,0011	0,3333	1,0000	942	3	1	939
	$\texttt{cir6roul}_{108}^{1997}$	0,0011	0,2500	1,0000	925	4	1	921
web	$\texttt{irtLnut}_{516}^{2001}$	0,0010	1,0000	1,0000	993	1	1	992
	$\texttt{ictweb10nfl}_{525}^{2001}$	0,0010	0,1667	1,0000	992	6	1	986
	$\texttt{ictweb10nf}_{525}^{2001}$	0,0010	0,1667	1,0000	992	6	1	986
	$\texttt{ictweb10nfl}_{502}^{2001}$	0,0010	0,1667	1,0000	990	6	1	984
	$\texttt{ictweb10nf}_{502}^{2001}$	0,0010	0,1667	1,0000	990	6	1	984

Tableau 4. *Top 5 des différences entre rangs conventionnel $1/RR_C$ et réaliste $1/RR_R$.*

pour $\texttt{cir6roul}_{194}^{1997}$ qui mérite $AP_R = 0{,}0262$ avec une stratégie équitable alors qu'il obtient injustement $AP_C = 0{,}1325$ avec la stratégie conventionnelle.

Les tests statistiques rapportés dans le tableau 5 soulignent une différence significative entre AP_C et AP_R quelle que soit la tâche. Néanmoins, cette différence est faible en pourcentage, expliquant la valeur très élevée du coefficient de corrélation.

5.2.3. *Impact du biais des* ex aequo *sur la mesure Mean Average Precision.*

Le tableau 7 montre les cinq systèmes les plus affectés pour la mesure MAP, pour chaque tâche et stratégie de réordonnancement. On notera que les gains$_{CR}$ sont plus faibles que pour la mesure AP. Ceci est dû au fait que plusieurs valeurs d'AP sont prises en compte par la moyenne pour constituer la MAP. Or certaines listes-résultats ne sont pas affectées par le biais des *ex aequo*. De fait, l'influence des AP affectées sur la MAP est limitée, contrebalancée, par ces AP non affectées. Malgré cet effet de lissage imputable à l'opération de moyenne arithmétique, on remarque néanmoins

tâche	RR_C *versus* RR_R		AP_C *versus* AP_R		MAP_C *versus* MAP_R	
	δ_{RC} (%)	corr. r	δ_{RC} (%)	corr. r	δ_{RC} (%)	corr. r
adhoc	0,60*	0,99	0,37*	1,00	0,37*	1,00
filtering	9,39*	0,89	3,14*	0,99	3,12*	0,99
routing	1,14*	0,99	0,57*	1,00	0,58*	1,00
web	0,55*	1,00	0,40*	1,00	0,45*	1,00

Tableau 5. *Corrélation et significativité des différences entre les mesures M_C et M_R. L'astérisque* indique que le test t de Student est significatif avec $p < 0{,}001$.*

	liste-résultat	AP_R	AP_C	AR_O	gain$_{OR}$ %	gain$_{CR}$ %	gain$_{CO}$ %
adhoc	$\texttt{ibmgd2}^{1996}_{291}$	0,0000	0,0001	0,0074	49 867	318	11 858
	$\texttt{issah1}^{1995}_{246}$	0,0001	0,0003	0,0018	2 718	311	585
	$\texttt{harris1}^{1997}_{327}$	0,0139	0,0556	0,0556	300	300	0
	$\texttt{ETHAB0}^{1998}_{394}$	0,0002	0,0006	0,0062	3 388	268	848
	$\texttt{padre1}^{1994}_{181}$	0,0051	0,0169	0,2143	4 066	229	1 167
filtering	$\texttt{IAHKaf12}^{1998}_{13}$	0,0005	0,0116	0,0116	2 200	2 200	0
	$\texttt{IAHKaf32}^{1998}_{13}$	0,0005	0,0116	0,0116	2 200	2 200	0
	$\texttt{IAHKaf12}^{1998}_{39}$	0,0029	0,0625	0,2500	8 400	2 025	300
	$\texttt{IAHKaf32}^{1998}_{39}$	0,0029	0,0625	0,2500	8 400	2 025	300
	$\texttt{TNOAF102}^{1998}_{30}$	0,0263	0,5000	1,0000	3 700	1 800	100
routing	$\texttt{cir6rou1}^{1997}_{161}$	0,0000	0,0008	0,0060	11 995	1 435	688
	$\texttt{cir6rou1}^{1997}_{194}$	0,0262	0,1325	0,2626	902	406	98
	$\texttt{erliR1}^{1996}_{77}$	0,0311	0,1358	0,5714	1 736	336	321
	$\texttt{erliR1}^{1996}_{125}$	0,0070	0,0290	0,1795	2 462	313	520
	$\texttt{virtue3}^{1997}_{228}$	0,0065	0,0239	0,3852	5 863	271	1 509
web	$\texttt{ICTWebTD12A}^{2003}_{15}$	0,0064	0,2541	0,2544	3 861	3 856	0
	$\texttt{irtLnut}^{2001}_{516}$	0,0012	0,0355	0,2667	22 070	2 853	651
	$\texttt{iswt}^{2000}_{490}$	0,0000	0,0004	0,0007	4 248	2 173	91
	$\texttt{ictweb10nf}^{2001}_{502}$	0,0008	0,0154	0,1358	16 511	1 779	784
	$\texttt{ictweb10nfl}^{2001}_{502}$	0,0008	0,0154	0,1358	16 511	1 779	784

Tableau 6. *Top 5 des gains entre AP conventionnelle (AP_C) et réaliste (AP_R).*

des gains injustifiés. Par exemple $\texttt{padre1}^{1994}$ a gagné 37 % de MAP en bénéficiant du biais expérimental. Ainsi, sans aucune contribution propre, il a été gratifié d'une $MAP_C = 0{,}1448$ alors qu'il ne mérite que $MAP_R = 0{,}1060$ dans un cadre non biaisé. S'il avait été encore plus chanceux il aurait même pu obtenir indûment jusqu'à $MAP_O = 0{,}2967$. Le tableau 5 indique que MAP_C et MAP_R sont corrélées mais significativement différentes. En complément, le coefficient τ de Kendall indique que les rangs des SRI calculés à partir de la MAP ne diffèrent pas significativement quelle que soit la tâche ou la stratégie de réordonnancement, la différence de valeur de mesure n'étant pas assez importante pour affecter le classement des SRI.

Par ailleurs, nous avons étudié l'influence de la stratégie de réordonnancement (MAP_R *vs* MAP_C) sur la significativité des différences entre les SRI appariés, pour chaque édition. Les conclusions du test montrent que jusqu'à 9 % de conclusions sont erronées, correspondant aux deux cas suivants. Dans le premier cas, le test t a conclu à une différence significative avec la stratégie conventionnelle tandis que cette différence n'est pas significative avec la stratégie réaliste. Dans le deuxième cas, le test t n'a pas conclu à une différence significative avec la stratégie conventionnelle alors qu'elle l'est avec la stratégie réaliste.

Nous avons également analysé la différence en terme de rang de SRI, en calculant les rangs MAP_R *vs* MAP_C. Le rang de 23 % des SRI est différent, soulignant ainsi l'influence de la stratégie sur le classement des SRI.

Une analyse approfondie consiste à déterminer la position de ces SRI qui changent de rang (du fait d'un nombre élevé de documents *ex aequo* restitués en début de liste-résultat). Une hypothèse suppose que les plus mauvais SRI sont les plus concernés.

	Result-list	MAP_R	MAP_C	MAP_O	gain$_{OR}$ %	gain$_{CR}$ %	gain$_{CO}$ %
adhoc	padre1[1994]	0,1060	0,1448	0,2967	180	37	105
	UB99SW[1999]	0,0454	0,0550	0,0650	43	21	18
	harris1[1997]	0,0680	0,0821	0,0895	32	21	9
	padre2[1994]	0,1524	0,1752	0,2237	47	15	28
	topic3[1994]	0,0249	0,0286	0,0364	46	15	27
filtering	IAHKaf12[1998]	0,0045	0,0396	0,0558	1 140	779	41
	IAHKaf32[1998]	0,0045	0,0396	0,0558	1 140	779	41
	TNOAF103[1998]	0,0144	0,0371	0,0899	524	158	142
	TNOAF102[1998]	0,0134	0,0290	0,1089	712	116	276
	sigmaTrec7F1[1998]	0,0357	0,0702	0,3592	906	97	412
routing	cir6roul[1997]	0,0545	0,0792	0,2306	323	45	191
	erliR1[1996]	0,1060	0,1412	0,2507	137	33	78
	topic1[1994]	0,2062	0,2243	0,2543	23	9	13
	topic2[1994]	0,2550	0,2774	0,3177	25	9	15
	brkly14[1996]	0,2402	0,2601	0,2766	15	8	6
web	ictweb10nf[2001]	0,0210	0,0464	0,4726	2 150	121	919
	ictweb10nfl[2001]	0,0210	0,0463	0,4660	2 119	120	907
	irtLnut[2001]	0,0102	0,0221	0,2343	2 202	117	960
	iswtdn[2000]	0,0317	0,0412	0,0760	140	30	84
	iswtd[2000]	0,0251	0,0325	0,0721	187	29	122

Tableau 7. *Top 5 des gains entre MAP conventionnelle (MAP_C) et réaliste (MAP_R).*

Nous avons donc éliminé les 25 % des SRI les moins bien classés. Nous constatons alors que 17 % des SRI restent affectés. Cette observation contredit l'hypothèse que la plupart des *ex aequo* sont restitués par les plus mauvais SRI. En particulier, nous pouvons remarquer l'example de uwmt6a0[1997] qui a été classé premier alors qu'il contient 57 % de documents *ex aequo*.

Cette section a montré que les mesures $M \in \{RR, AP, MAP\}$ dans leur version réaliste (M_R) sont statistiquement différentes de la mesure conventionnelle M_C. Ainsi, nous avons observé une différence notable entre la stratégie de réordonnancement équitable que nous proposons et la stratégie inéquitable implantée dans trec_eval. La section suivante discute les implications de cette observation.

6. Discussion : biais des *ex aequo* et « phénomène de bourrage »

Après avoir montré que le biais des *ex aequo* peut avoir un impact sur les valeurs des mesures, nous l'avons observé et quantifié sur des données relatives à différentes tâches et différentes éditions de TREC. Il s'avère qu'en fonction de la stratégie de réordonnancement retenue, les scores attribués aux listes-résultats et *runs* peuvent être statistiquement différents. C'est à nos yeux un problème important qui complique la comparaison objective des SRI, certains ayant pu bénéficier du biais des *ex aequo*. Afin d'encourager des évaluations de RI plus équitables, il nous paraît alors opportun d'intégrer au programme trec_eval la possibilité de spécifier la stratégie de réordonnancement choisie : réaliste, conventionnelle (comportement actuel) ou optimiste. De ce fait, les campagnes TREC mais aussi NTCIR, CLEF et toutes celles se basant également sur trec_eval pourront bénéficier de cet apport.

En complément au biais des *ex aequo*, nous avons identifié un « phénomène de bourrage » pratiqué par certains SRI. Dans le cas de TREC, une liste-résultat est constituée au maximum de 1 000 documents. Or, nous avons remarqué que 10,5 % systèmes trouvent moins de 1 000 documents en réponse à un *topic* et « bourrent » la liste-résultat avec des documents associés à une similarité sim = 0. Ce phénomène est conceptuellement intriguant : pourquoi un SRI restituerait-il un document qui n'a rien à voir avec la requête ? Une réponse rationnelle peut-être la suivante : parmi ces documents de bourrage, certains peuvent s'avérer être pertinents (appartenant au *pool* et jugés pertinents) et contribuer au score du SRI, même faiblement. Or, avec la stratégie de réordonnancement de trec_eval, des documents restitués avec une sim = 0 peuvent finalement remonter en haut de la partie « bourrage » dans la liste-résultat. De fait, ils contribuent davantage que s'ils avaient été classés en bas de la liste avec la stratégie réaliste que nous préconisons. Il apparaît donc nécessaire de décourager le bourrage voué à accroître les valeurs des mesures artificiellement et injustement, ce qui représente un argument supplémentaire en faveur de la stratégie de réordonnancement réaliste que nous proposons. Une analyse complémentaire montre qu'aucun des six premiers *runs* de toutes les éditions étudiées ne présente ce phénomène. En d'autres termes, les meilleurs SRI ne bourrent pas.

7. Travaux connexes concernant l'évaluation de la RI

La littérature présente des travaux relatifs à la validation de la méthodologie d'évaluation en RI, notamment au travers de TREC. Les questionnements sont principalement liés à quatre aspects :

1) fiabilité du *pooling*. Le fait de ne juger que les documents issus du *pool*, c'est-à-dire ceux soumis par les SRI des participants a été analysé par Zobel (1998) qui démontre la fiabilité de cette technique. Par ailleurs, Sanderson *et al.* (2004) ont étudié la constitution éventuellement manuelle de collections de test sans avoir recours au *pooling*, montrant une qualité obtenue similaire aux collections de test TREC. Plus récemment, Buckley *et al.* (2007) ont argumenté en faveur de l'adaptation de la profondeur du *pool* en fonction de la taille des collections pour éviter d'oublier un trop grand nombre de documents pertinents.

2) fiabilité des *qrels*. La qualité des jugements de pertinence a été étudiée dans (Voorhees, 1998) où sont observées les différences d'appréciation selon les assesseurs, ces dernières ne compromettant pas globalement la pertinence des évaluations de TREC. Al-Maskari *et al.* (2008) identifient un désaccord entre assesseurs TREC « habituels » et assesseurs non-TREC, soulignant la relativité de la notion de pertinence.

3) fiabilité du nombre de *topics* par édition. Buckley *et al.* (2000) montrent la nécessité de proposer au moins 25 *topics* par édition pour lisser les phénomènes de performance locale : un score élevé sur un *topic* contrasté par des faiblesses sur d'autres. Le standard de TREC à 50 *topics* est préférable, étant caractérisé par un taux d'erreur acceptable de 2 % à 3 % pour la plupart des mesures. Certaines mesures, telles

que P@10 ne seraient alors valides qu'avec bien plus de 50 *topics*. Par la suite, des taux d'erreurs ont été formalisés et calculés dans (Voorhees *et al.*, 2002), permettant d'extrapoler le taux d'erreur en fonction du nombre de *topics* retenus. Enfin, Voorhees (2009) souligne des défaillances dans les conclusions des tests de significativité et recommande la validation des propositions de recherche sur plusieurs collections de test, même en présence de différences en termes de score relativement élevées ($> 10\,\%$).

4) fiabilité des mesures utilisées. Buckley *et al.* (2000) étudient la fiabilité des mesures, montrant par exemple que P@30 est caractérisée par un taux d'erreur deux fois plus élevé qu'AP. Par ailleurs, allant à l'encontre d'études précédentes, Sakai (2008) traite du « biais du système » : un SRI n'ayant pas participé au *pool* peut être surestimé ou sous-estimé selon le type de mesure utilisé. Des travaux tel que (Moffat *et al.*, 2008) proposent de nouvelles mesures plus à même de condenser la satisfaction présumée des individus. Enfin, des initiatives complémentaires visent à identifier un sous-ensemble minimal de mesures non redondantes (Baccini *et al.*, 2010) parmi le vaste panel des 85 valeurs calculées par trec_eval (Voorhees, 2007).

Notre contribution relative aux biais des *ex aequo* s'inscrit dans ce dernier aspect. Le problème de l'évaluation de SRI restituant des documents *ex aequo* par des mesures courantes n'étant pas conçues pour gérer les *ex aequo* (telles que la précision, le rappel, $F1$, AP, RR, $NDCG$) a été identifié dans la littérature (Raghavan *et al.*, 1989; McSherry *et al.*, 2008). Une solution proposée par Raghavan *et al.* (1989) repose sur la mesure Precall en tant qu'extension de la précision à divers degrés de rappel, prenant en compte les groupes de documents *ex aequo*. Par ailleurs, McSherry *et al.* (2008) ont étendu les six mesures citées précédemment en calculant la moyenne des valeurs obtenues par toutes les permutations possibles des documents dans tous les groupes d'*ex aequo*. Ces deux approches permettent la comparaison déterministe de SRI.

Pour résoudre ce même problème, nous avons opté pour une approche différente qui n'impose pas la conception de nouvelles mesures. Cette approche est basée sur des stratégies de réordonnancement appliquées aux *runs* des SRI. La stratégie actuelle, que nous avons qualifiée de conventionnelle, a été implémentée dans TREC depuis son origine. En plus d'être déterministes, les stratégies réaliste et optimiste que nous proposons permettent de mesurer le gain (resp. la perte) d'efficacité d'un système qui ordonnerait correctement (resp. incorrectement) les documents *ex aequo*. La différence entre les bornes de la mesure étudiée ($M_O - M_R$) peut être interprétée comme une limite du SRI face à la prise en compte des documents *ex aequo*. Cet indicateur met donc en lumière le potentiel d'amélioration d'un SRI que l'on pourrait atteindre en départageant les *ex aequo*.

À notre connaissance, l'impact du biais des *ex aequo* dans les évaluations de RI (notamment dans les campagnes TREC) n'a pas fait l'objet d'étude quantitative jusqu'à présent.

8. Conclusion et perspectives

Cet article a considéré la thématique de l'expérimentation en RI, notamment grâce au programme trec_eval qui est mis en œuvre dans les campagnes internationales telles que TREC, NTCIR et CLEF pour calculer les valeurs des mesures pour les systèmes participant, telle que la MAP. Nous avons montré que les valeurs des mesures dépendent de deux facteurs : *i*) la pertinence des documents restitués pour la requête traitée, et *ii*) les noms des documents qui sont exploités pour départager les documents *ex aequo*, restitués avec la même similarité. Ce dernier facteur constitue un biais expérimental influençant les scores de façon non souhaitable. En effet, le hasard peut bénéficier à un SRI lorsque des documents pertinents sont réordonnés plus haut qu'initialement dans la liste-résultat, uniquement grâce à leurs noms.

Afin de corriger cette stratégie conventionnelle biaisée implantée dans trec_eval, nous avons proposé deux stratégies plus équitables nommées « réaliste » et « optimiste ». L'étude de 22 éditions des tâches *adhoc*, *routing*, *filtering* et *web* de TREC a révélé des différences statistiquement significatives entre la stratégie équitable réaliste que nous promouvons et la stratégie inéquitable conventionnelle pour les mesures RR, AP et MAP. Toutefois, les valeurs des mesures ne sont pas assez affectées pour modifier le classement des SRI calculé à partir de la MAP. Nous suggérons l'intégration des trois stratégies dans trec_eval, permettant ainsi à tout expérimentateur de choisir le comportement qu'il escompte, ce qui rendrait possible et pourrait même encourager des expérimentations plus équitables. Cela permettrait également d'identifier les « améliorations » rapportées dans la littérature alors qu'imputables au biais des *ex aequo*, étant uniquement dues au hasard du nommage des documents.

Plusieurs perspectives s'offrent à ces travaux. À court terme, il s'agit de confirmer les conclusions des tests statistiques utilisés grâce à des tests plus sophistiqués tels que *bootstrap* et *randomization*, comme recommandé dans (Savoy, 1997; Smucker *et al.*, 2007). À moyen terme, l'étude du phénomène de « bourrage » discuté en section 6 permettra d'identifier la part des scores obtenue par des effets de bord imputables à la bonne connaissance du cadre d'évaluation et non à la qualité intrinsèque des SRI. Par ailleurs, nous pourrons évaluer l'impact du biais des *ex aequo* dans d'autres cadres d'évaluations de la RI, tels que NTCIR et CLEF. Le système DIRECT (Di Nunzio *et al.*, 2005) offrant un accès aux données des participants à CLEF pourra être employé à cet effet. Enfin, à plus long terme, le biais des *ex aequo* a pu fausser l'apprentissage des systèmes reposant sur l'approche de « *learning to rank* » (Joachims *et al.*, 2007). Il conviendrait de vérifier si l'utilisation de stratégies de réordonnancement non biaisées conduisent à un meilleur apprentissage. Par ailleurs, nous envisageons la définition de nouvelles mesures qui ne nécessiteraient pas le pré-traitement des *runs* actuellement réalisé, prenant en compte différemment les documents *ex aequo* et les autres documents que le SRI s'est efforcé à discriminer.

9. Bibliographie

Al-Maskari A., Sanderson M., Clough P., "Relevance Judgments between TREC and Non-TREC Assessors", *SIGIR'08: Proceedings of the 31st annual international ACM SIGIR conference*, ACM, New York, NY, USA, p. 683–684, 2008.

Baccini A., Déjean S., Kompaore N. D., Mothe J., "Analyse des critères d'évaluation des systèmes de recherche d'information", *Technique et Science Informatiques*, vol. 29, p. 289–308, 2010.

Buckley C., Dimmick D., Soboroff I., Voorhees E. M., "Bias and the limits of pooling for large collections", *Inf. Retr.*, vol. 10, n° 6, p. 491–508, 2007.

Buckley C., Voorhees E. M., "Evaluating Evaluation Measure Stability", *SIGIR'00: Proceedings of the 23rd international ACM SIGIR conference*, ACM, New York, NY, USA, p. 33–40, 2000.

Buckley C., Voorhees E. M., "Retrieval System Evaluation", *in* Voorhees *et al.* (2005), chapter 3, p. 53–75, 2005.

Di Nunzio G. M., Ferro N., "DIRECT: A System for Evaluating Information Access Components of Digital Libraries", *ECDL'05: Proceedings of the 9th European Conference on Digital Libraries*, vol. 3652 of *LNCS*, Springer, p. 483–484, 2005.

Harman D. K. (ed.), *TREC-1: Proceedings of the First Text REtrieval Conference*, NIST, Gaithersburg, MD, USA, February, 1993.

Hull D., "Using Statistical Testing in the Evaluation of Retrieval Experiments", *SIGIR'93: Proceedings of the 16th annual international ACM SIGIR conference*, ACM Press, New York, NY, USA, p. 329–338, 1993.

Joachims T., Li H., Liu T.-Y., Zhai C., "Learning to Rank for Information Retrieval (LR4IR 2007)", *SIGIR Forum*, vol. 41, n° 2, p. 58–62, 2007.

Kando N., Kuriyama K., Nozue T., Eguchi K., Kato H., Hidaka S., "Overview of IR Tasks at the First NTCIR Workshop", *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, NACSIS, p. 11–44, 1999.

Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, July, 2008.

McSherry F., Najork M., "Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores", *ECIR'08: Proceedings of the 30th European Conference on IR Research*, vol. 4956 of *LNCS*, Springer, p. 414–421, 2008.

Moffat A., Zobel J., "Rank-biased precision for measurement of retrieval effectiveness", *ACM Trans. Inf. Syst.*, vol. 27, n° 1, p. 1–27, 2008.

NIST, "README file for trec_eval 8.1", n.d., `http://trec.nist.gov/trec_eval`.

Peters C., Braschler M., "European Research Letter – Cross-Language System Evaluation: the CLEF Campaigns", *J. Am. Soc. Inf. Sci. Technol.*, vol. 52, n° 12, p. 1067–1072, 2001.

Raghavan V., Bollmann P., Jung G. S., "A critical investigation of recall and precision as measures of retrieval system performance", *ACM Trans. Inf. Syst.*, vol. 7, n° 3, p. 205–229, 1989.

Robertson S., "On the history of evaluation in IR", *J. Inf. Sci.*, vol. 34, n° 4, p. 439–456, 2008.

Sakai T., “Comparing Metrics across TREC and NTCIR: The Robustness to System Bias”, *CIKM’08: Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ACM, New York, NY, USA, p. 581–590, 2008.

Sanderson M., Joho H., “Forming Test Collections with No System Pooling”, *SIGIR’04: Proceedings of the 27th annual international ACM SIGIR conference*, ACM, New York, NY, USA, p. 33–40, 2004.

Sanderson M., Zobel J., “Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability”, *SIGIR’05: Proceedings of the 28th annual international ACM SIGIR conference*, ACM, New York, NY, USA, p. 162–169, 2005.

Savoy J., “Statistical inference in retrieval effectiveness evaluation”, *Inf. Process. Manage.*, vol. 33, n° 4, p. 495–512, 1997.

Smucker M. D., Allan J., Carterette B., “A Comparison of Statistical Significance Tests for Information Retrieval Evaluation”, *CIKM’07: Proceedings of the 16th ACM conference on Conference on information and knowledge management*, ACM, New York, NY, USA, p. 623–632, 2007.

Voorhees E. M., “Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness”, *SIGIR’98: Proceedings of the 21st annual international ACM SIGIR conference*, ACM, New York, NY, USA, p. 315–323, 1998.

Voorhees E. M., “Overview of the TREC 2004 Robust Track”, *in* E. M. Voorhees, L. P. Buckland (eds), *TREC’04: Proceedings of the 13th Text REtrieval Conference*, NIST, Gaithersburg, MD, USA, 2004.

Voorhees E. M., “TREC: Continuing information retrieval’s tradition of experimentation”, *Commun. ACM*, vol. 50, n° 11, p. 51–54, 2007.

Voorhees E. M., “Topic Set Size Redux”, *SIGIR’09: Proceedings of the 32nd annual international ACM SIGIR conference*, ACM, New York, NY, USA, p. 806–807, 2009.

Voorhees E. M., Buckley C., “The Effect of Topic set Size on Retrieval Experiment Error”, *SIGIR’02: Proceedings of the 25th annual international ACM SIGIR conference*, ACM, New York, NY, USA, p. 316–323, 2002.

Voorhees E. M., Harman D. K., *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, Cambridge, MA, USA, 2005.

Zobel J., “How Reliable are the Results of large-scale Information Retrieval Experiments?”, *SIGIR’98: Proceedings of the 21st annual international ACM SIGIR conference*, ACM, New York, NY, USA, p. 307–314, 1998.