Adapter les moteurs de recherche aux besoins en information

Prise en compte de la difficulté du besoin

Anthony Bigot

Institut de Recherche en Informatique de Toulouse,
Université Paul Sabatier,
118 Route de Narbonne,
F-31062 TOULOUSE CEDEX 9

anthony.bigot@irit.fr
Thème : Évaluation des Systèmes d'Information

RÉSUMÉ. *Les campagnes telles que Text REtrieval Conference (TREC) offrent un cadre qui permet d'évaluer des systèmes de recherche d'information (RI). L'évaluation utilise des mesures qui se basent sur une moyenne des résultats obtenus pour un ensemble de besoins en information: les succès et échecs sur chacun des besoins sont masqués. Cet article propose une analyse qui prend en compte la diversité des résultats et qui a pour but de sélectionner le système de RI le plus apte à traiter les besoins selon leur difficulté. Deux méthodes de sélection des meilleurs systèmes de RI sont proposées. La première méthode considère la moyenne des mesures de performance des systèmes et la seconde considère leurs rangs selon cette moyenne. Une expérimentation est conduite afin de sélectionner les meilleurs systèmes selon la difficulté des besoins en information. L'évaluation montre un gain maximal en robustesse de 68% et de 10% sur les performances.*

ABSTRACT. *Evaluation campaigns such as Text REtrieval Conference (TREC) offer a framework to evaluate information retrieval (IR) systems. Generally, these evaluation results are avereged over information needs so successes and failures on specific needs are hidden. This paper analysis aims at treating differently information needs through their difficulty. Two selection methods of the bests IR systems are considered. The first method is based on the average performance measure of systems ; the second is based on their ranks according to this average. An experiment is driven in order to select the best systems according to the information need difficulty. Evaluation results show a maximum gain of 68% in reliability and of 10% on performances.*

MOTS-CLÉS : *évaluation en recherche d'information, rang des systèmes, difficulté des besoins, classifications*

KEYWORDS: *IR evaluation, system ranks, information need difficulty, classifications*

1. Introduction

Les campagnes d'évaluation telles que Text REtrieval Conference (TREC)[1] ont permis de grandes avancées dans le monde de la recherche d'information (RI). Ces campagnes sont basées sur le modèle d'évaluation de Cranfield (Cleverdon *et al.,* 1966). Selon ce modèle, un système à évaluer recherche les documents à restituer à l'utilisateur pour un ensemble de besoins d'information et ce à partir d'une collection figée de documents ; la liste des documents pertinents étant connue pour chaque besoin. En comparant la liste de documents restituée par le système avec la liste de documents attendue pour un besoin donné, il est possible de calculer des mesures de performances. Ce cadre peut être utilisé pour comparer les systèmes entre eux sur le même ensemble de données. Plusieurs études ont montré la variabilité des résultats des systèmes (Harman *et al.,* 2009). Ainsi, un système S_1 peut être très performant sur un besoin A mais échouer sur un besoin B et un système S_2 peut avoir des performances inverses. Notre travail vise à privilégier le système S_1 pour traiter le besoin A et le système S_2 pour traiter le besoin B. La variabilité des résultats est issue de la variabilité qui existe au sein des systèmes de RI d'une part, et de celle qui existe parmi les besoins en information d'autre part.

Nous nous intéressons aux combinaisons de résultats des systèmes de RI différents afin de tirer profit de leur variabilité. Dans la littérature, différents travaux se sont intéressés à combiner les résultats provenant de différents systèmes. Lors de la campagne TREC3 ont eu lieu les premières expérimentations sur la fusion des systèmes de recherche d'information (Fox *et al.,* 1994). Dans cet article, l'auteur analyse une série de méthodes de fusion des systèmes de RI, dont le célèbre CombMNZ, et montre une amélioration significative des résultats par rapport aux performances des systèmes initiaux. L'étude de ces méthodes de fusion montre que l'utilisation de rangs au lieu de la similarité produit de meilleurs résultats dans la combinaison des résultats lorsque les systèmes initiaux ont des rangs très différents (Lee, 1997). De nombreux auteurs se sont employés à la mise au point de nouveaux algorithmes de fusion des systèmes de RI existants. Par exemple, Lillis (Lillis *et al.,* 2006) étudie les résultats de différents algorithmes de RI et propose un modèle de fusion basé sur les probabilités (probFuse). L'auteur compare ces résultats à ceux obtenus par CombMNZ et montre une amélioration de la MAP allant de 15% à 50% selon la taille de la collection utilisée durant la phase d'apprentissage.

Des analyses ont également porté sur les différents paramètres intervenant dans la combinaison comme le nombre initial de systèmes, leurs performances et leur diversité (Liu *et al.,* 2012). Cette étude considère les rangs obtenus par les systèmes et montre que la performance de la combinaison n'augmente pas forcément avec le nombre de systèmes utilisés. L'auteur montre également que les performances sont améliorées lorsque les systèmes initiaux sont relativement performants et qu'ils sont suffisamment différents selon le critère RSC ("Rank-score characteristics" en anglais)

1. http://trec.nist.gov/

Une seconde source de variabilité dans les performances des systèmes de RI provient de la variabilité des besoins en information et du contexte dans lequel ils sont formulés. La littérature montre de nombreuses analyses menées pour catégoriser les besoins en information. Mothe (Mothe *et al.,* 2005) analyse 16 traits linguistiques afin de prédire la difficulté d'un besoin en information. Les auteurs montrent que 3 de ces traits ont un impact significatif sur les mesures de rappel et/ou de précision. Carmel (Carmel *et al.,* 2010) s'intéresse à la prédiction de la difficulté d'un besoin en information nouvellement formulé et donc non connu au préalable.

La variabilité des besoins en information et du contexte dans lesquels ils sont formulés a conduit la RI à évoluer vers une RI adaptative (Joho *et al.,* 2008). Les premiers travaux dans ce domaine s'intéressent aux modifications à apporter au processus de RI selon les informations tirées du contexte dans lequel il se déroule (Brown *et al.,* 2001). D'autres travaux se focalisent sur la formulation du contexte sous forme d'un besoin en information afin d'améliorer la pertinence des résultats restitués par les systèmes de RI(Menegon *et al.,* 2009). Cependant, la mise en place d'un cadre d'évaluation de tels algorithmes comportent de nombreuses difficultés (Mizzaro *et al.,* 2008).

En s'appuyant sur l'hypothèse d'une variabilité dans la manière de traiter les besoins en information, Kompaore (Kompaoré *et al.,* 2007) étudie le meilleur choix possible de systèmes de RI selon des critères linguistiques du besoin en information. Dans ce contexte, Bigot (Bigot *et al.,* 2011) propose une méthode d'apprentissage du meilleur système de recherche d'information par classe de difficulté des besoins. Les auteurs montrent une amélioration significative de la mesure MAP de 10% pour les besoins classés faciles et difficiles sur la collection de documents test. Le gain en performance augmente à 24% sur les besoins classés moyennement difficiles à traiter.

Dans l'analyse que nous proposons ici, nous émettons l'hypothèse que des besoins différents doivent être traités par des systèmes de RI différents afin de tirer avantage de la variabilité de ces derniers comme dans (Bigot *et al.,* 2011). Nous supposons que les meilleurs systèmes pour traiter les besoins les plus faciles sont différents des systèmes plus aptes à traiter les besoins les plus difficiles. Nous cherchons à sélectionner les meilleurs systèmes selon des classes de difficulté des besoins ; ces classes sont définies par les performances des systèmes de RI. Contrairement aux travaux proposé dans (Bigot *et al.,* 2011), ce papier utilise une combinaison de plusieurs mesures d'évaluation des systèmes et les sélectionne selon deux aspects différents : leur performance et leur robustesse.

Le papier est structuré comme suit : la section 2 présente le contexte dans lequel les systèmes de RI sont évalués et explique la construction du jeu de données utilisé pour l'analyse. Dans la section 3, nous présentons la méthodologie de sélection du meilleur système. La section 4 propose un groupement des besoins en information selon leur difficulté et définit une implémentation pratique des méthodes proposées en section 3. La section 5 analyse l'expérimentation et propose une évaluation des méthodes de sélection. Enfin, la section 6 conclut et propose des pistes d'analyse sur le sujet.

2. Contexte et données

2.1. *Text Retrieval Conference*

Dans ce papier, nous cherchons à choisir les meilleurs systèmes de RI pour traiter les besoins en informations. Pour cela, nous avons besoin d'étudier les performances d'un nombre suffisamment grand de systèmes de RI. Cranfield (Cleverdon *et al.,* 1966) propose un cadre d'évaluation qui permet d'étudier les performances de systèmes de RI : les différents systèmes traitent des besoins en information prédéfinis sur une collection fixée de documents. Selon ce même cadre d'évaluation des systèmes, TREC propose chaque année des challenges à accomplir pour les systèmes de RI. Nous avons choisi d'utiliser la collection de la tâche ad-hoc issue de TREC. Pour cette tâche, TREC fournit une collection composée de documents, de besoins en information et de jugements de pertinence pour chacun des besoins (le "qrel").

La collection que nous avons utilisée est issue de TREC adhoc 8 et est composée de 528155 documents ; chaque système participant doit produire une liste ordonnée de 1000 documents pour chaque besoin en information. Les besoins en information sont composés d'un titre, d'une description et d'une partie narrative. La requête est une représentation interne que chaque système de RI définit librement à partir du besoin en information traité. Les jugements de pertinence permettent d'évaluer les performances des systèmes.

Ainsi chaque participant utilise un ou plusieurs systèmes de RI pour rechercher les documents de la collection susceptibles de répondre au besoin traité. Le résultat de cette recherche, appelée "run" en anglais et que nous appelons "exécution", est une liste ordonnée des documents restituée pour chaque besoin en information.
Afin de calculer les performances d'un système, chaque exécution est confrontée à la liste des documents pertinents attendus par TREC (le "qrel"). TREC fournit un outil d'évaluation (*trec_eval*) qui permet de calculer de nombreuses mesures de performances pour chaque couple (système ; besoin en information).

2.2. *Structure des données*

Dans ces travaux, nous nous appuyons sur les résultats obtenus par les participants à la tâche (adhoc TREC 8). Parmi les 130 mesures de performances, calculées pour chacun des 129 systèmes participants cette année-là, pour chacun des 50 besoins en information de la collection, nous sélectionnons les 118 mesures qui prennent leurs valeurs dans [0 ; 1]. Pour mener les analyses, nous avons besoin d'un format des données adapté à l'utilisation de méthodes mathématiques. Nous structurons donc les données sous forme de matrice. La matrice des données initiales est un cube de données à trois dimensions.

La première dimension est constituée des 50 besoins en information ; la seconde dimension représente les 129 systèmes et la troisième dimension correspond aux 118

mesures de performances. Chaque cellule du cube contient la valeur de la mesure de performance pour l'exécution d'un système de RI sur un besoin en information.

Baccini (Baccini *et al.,* 2011) montre que les 130 mesures de performances issues de trec_eval sont redondantes et les classe en 6 groupes. Les auteurs montrent également qu'un sous-ensemble composé d'une mesure représentative de chaque groupe est suffisant. Selon ces critères, nous retenons les 6 mesures suivantes pour construire le jeu de données dit "réduit" de notre analyse : la "mean average precision" (MAP) ; la précision à 30 (P30) ; la précision à 100 (P100) ; le rappel exact (exact_recall) ; la préférence binaire (bpref_retall) ; le rappel à 30 (R30). Baccini et al. (2011) ont montré que la corrélation entre les mesures moyennes des systèmes de RI en utilisant les données complètes et les mesures moyennes en utilisant les données réduites est supérieure à 0.99, ce qui confirme le choix de ces 6 mesures.

3. Méthodes de classement des systèmes de recherche d'information

Dans cette section, nous proposons deux méthodes de sélection des meilleurs systèmes de RI. Ces méthodes sont générales et ne dépendent ni des besoins en information, ni des systèmes de RI considérés. Les résultats des méthodes sont ensuite comparés.

Calcul des rangs. Nous considérons le calcul à besoin en information fixé. Pour ce besoin, les systèmes de RI sont ordonnés par ordre décroissant de la mesure considérée. Le système ayant la plus forte valeur se retrouve en tête de liste et le rang 1 lui est attribué. Le système suivant de la liste est le second meilleur système et le rang 2 lui est attribué et ainsi de suite jusqu'au dernier système qui obtient le rang maximum. Si plusieurs systèmes obtiennent la même valeur de mesure, ils sont ex-æquo et obtiennent un rang équivalent à la moyenne des rangs attribués s'il n'y avait pas d'ex-æquo. L'algorithme d'attribution des rangs reprend son cours normal après le dernier ex-æquo.

Calcul de la mesure moyenne d'un système pour un besoin donné. Pour chacune des méthodes de sélection, nous commençons par créer la matrice des mesures moyennes. Cette matrice est composée de *s* lignes (pour les systèmes) et de *m* colonnes (pour les mesures). Chaque cellule de la matrice est la performance moyenne $\bar{M}^{(S)}$ de chaque système *S* pour chaque mesure *M* donnée par :

$$\bar{M}^{(S)} = \frac{1}{b} \sum_{B=1}^{b} m_B^{(S)}$$

où $m_B^{(S)}$ est la valeur de la mesure attribuée pour la performance du système *S* sur le besoin *B*.

Dans la suite de cette section, $x_{i,j}$ désignera la valeur de la moyenne de la mesure *j* pour le système *i*. De même, r_i désignera le rang du système *i*.

Mesure moyenne et rangs de la mesure moyenne. La mesure moyenne "MeMo" d'un système consiste à calculer la moyenne générale de chaque système sur la matrice des mesures moyennes (figure 1). Ensuite, les systèmes de RI sont ordonnés selon cette valeur pour obtenir les rangs de la mesure moyenne MeMo.

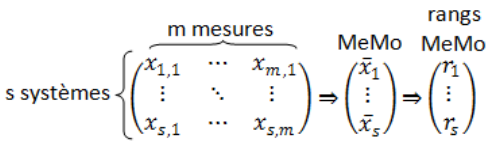

Figure 1. *Calculs de la mesure moyenne et rangs de la mesure moyenne des systèmes*

Les données réduites et la méthode de classement ci-dessus seront utilisées dans la suite de ce papier afin de déterminer quels sont les meilleurs systèmes selon la difficulté des besoins à traiter. Nous opposerons les systèmes obtenus par les rangs MeMo avec les systèmes obtenus en travaillant directement avec la MeMo.

4. Définition des meilleurs systèmes de recherche d'information par groupe de besoins en information

Les systèmes de RI ne traitent pas les besoins en information de la même manière : là où certains systèmes réussissent, d'autres échouent, et inversement (Harman *et al.,* 2009). Nous faisons l'hypothèse que les systèmes de RI sont plus ou moins performants selon la difficulté des besoins traités.

4.1. *Regroupement des besoins en information*

Afin de déterminer différentes classes de besoins en information selon un critère de difficulté et afin de déterminer si elles nécessitent d'être traitées par des systèmes de RI différents, nous procédons à un regroupement des besoins en classes.

À partir des données présentées en 2.2, il est possible de calculer une mesure moyenne représentative de la performance du système selon l'ensemble de ces performances sur un besoin donné. Ensuite, pour chaque besoin, le rang MeMo de chaque système est calculé comme le montre la figure 2.

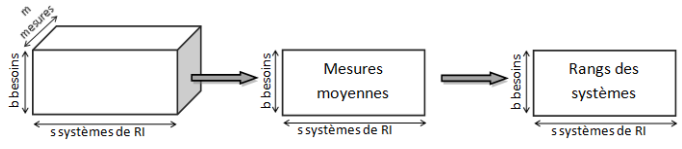

Figure 2. *Transformation des mesures en rangs de la mesure moyenne*

Pour grouper les besoins en information, une classification ascendante hiérarchique (CAH) est appliquée aux rangs MeMo. La CAH considère dans un premier temps chaque besoin en information comme un groupe à part entière puis assemble les deux groupes qui correspondent au critère choisi jusqu'à obtention d'une unique classe (Lebart *et al.,* 2006). Pour cette étude, le critère de Ward (Ward, 1963) est utilisé ; il minimise le gain en inertie entre les rangs d'un même groupe à chaque étape.

L'équation de décomposition de l'inertie nous indique que plus les besoins à l'intérieur des groupes sont similaires (i.e. plus l'inertie à l'intérieur de chaque groupe est petite), plus les groupes sont différents les uns des autres.

Le choix de 4 classes de besoins est raisonnable selon le critère usuel de rupture dans les éboulis de l'inertie (Lebart *et al.,* 2006). Ici, le nombre de classes obtenues par la CAH est dépendant d'un choix subjectif que nous avons fait. Afin de diminuer le biais introduit par ce choix, une méthode de ré-allocation dynamique non-supervisée est appliquée sur les classes obtenues. Les centres de gravité des classes ont été utilisés comme point de départ de l'algorithme des K-moyennes (Jain *et al.,* 1999). Cette méthode consiste à agréger chaque élément un à un au groupe dont la moyenne est la plus proche. La figure 3 donne une signification aux groupes de besoins en information créés. Pour chaque besoin en information, les MeMo des différents systèmes sont agrégées de manière à ce que :

– 25% des valeurs soient sous la boite (premier quartile : Q1) ;

– 50% des valeurs soient sous le trait gras (second quartile : médiane) ;

– 25% des valeurs soient au dessus de la boite (troisième quartile : Q3).

Les points isolés correspondent à des valeurs situées à une distance de la boite supérieure à une fois et demi l'écart inter-quartile (Q3-Q1).

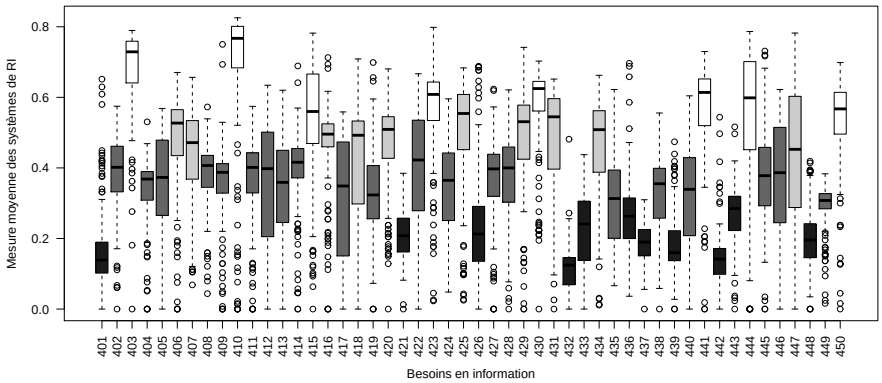

Figure 3. *Caractérisation des classes de besoins en information*

Sur cette figure, chaque groupe de besoins est représenté par un niveau de gris. Le groupe de besoins en information pour lesquels 75% des systèmes ont leur MeMo

supérieure à 0,5 est donc qualifié de groupe facile à traiter (en blanc). De la même manière, le groupe de besoins pour lesquels les systèmes obtiennent un Q1 supérieur à 0,3 et un Q3 inférieur à 0,6 est qualifié de groupe moyen (en gris clair) ; le troisième groupe est qualifié de groupe difficile à traiter (en gris foncé) avec une médiane des MeMo des systèmes comprise entre 0,3 et 0,4. Enfin, le groupe de besoins pour lesquels 75% des systèmes obtiennent une MeMo inférieure à 0,3 est qualifié de groupe très difficile (en noir).

4.2. *Performance et robustesse des systèmes de recherche d'information*

La première question qui se pose est quelle définition donner au meilleur système pour un groupe de besoins donné. Nous avons retenu deux approches en fonction de l'objectif visé. Le meilleur système peut présenter le meilleur classement possible pour la grande majorité des besoins en information, et avoir un très mauvais classement pour quelques uns. Il s'agirait d'un système de RI qui donne les meilleurs résultats mais pas en toutes circonstances. Nous appelons ce système de RI le système le plus *performant*.

Alternativement, le meilleur système peut être un système avec un bon classement (pas nécessairement le meilleur) pour l'ensemble des besoins en information considérés. Il s'agit d'un système robuste en permanence mais ne donnant pas les meilleurs résultats possibles. Ce système de RI est qualifié de système le plus *robuste*. Idéalement, pour un groupe de besoins en information donné, le meilleur système de RI aura les deux propriétés susmentionnées : les meilleurs performances et cela pour l'ensemble des besoins en information. Dans les faits, un tel système n'existe pas et les meilleurs systèmes de nos données connaissent des failles pour au moins un besoin. C'est pourquoi dans la suite de l'analyse, nous essayerons de trouver le système le plus performant d'une part, et le système le plus robuste d'autre part ; et ce pour chacun des groupes de besoins introduits en 4.1.

5. Analyse

5.1. *Détection des meilleurs systèmes par groupe de besoins*

Nous définissons deux règles qui correspondent aux données que nous traitons afin de déterminer les candidats aux deux types de *meilleurs* systèmes pour chacun des groupes de besoins considérés :

– candidats au système le plus *performant* : ces systèmes sont parmi les dix premiers pour au moins un besoin du groupe. Ces systèmes doivent également obtenir un rang inférieur à 20 pour au moins la moitié des besoins (i.e. la médiane de leurs rangs doit être inférieure à 20) ; ainsi nous sommes certains que les systèmes de RI candidats sont très performants pour au moins la moitié des besoins du groupe.

– candidats au système le plus *robuste* : ces systèmes ont un classement maximum inférieur à 80. De tels systèmes seront donc dans les deux premiers meilleurs tiers des systèmes de RI. La valeur 80 est fixée arbitrairement car suffisamment restrictive : en fixant ce seuil à un rang inférieur, aucun système de RI ne serait candidat au système le plus robuste pour certains groupes de besoins. Il est clair que les règles présentées ici peuvent mener à plusieurs systèmes candidats concourants pour être le plus performant et/ou le plus robuste. Nous verrons par la suite que le choix d'un système unique parmi les candidats est parfois possible et parfois non.

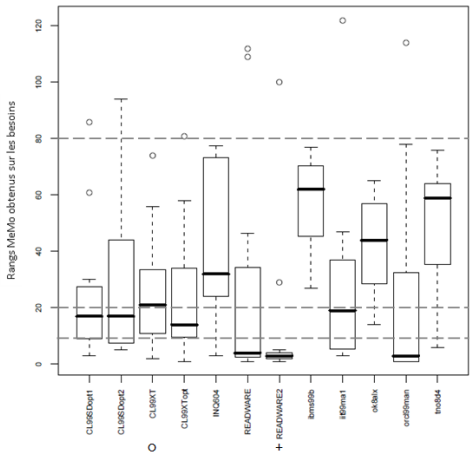

Figure 4. *Rangs MeMo : candidats aux meilleurs systèmes - groupe très difficile*

La figure 4 présente les 12 systèmes candidats pour le groupe des besoins que nous avons définis comme très difficiles en section 4.1. Un système candidat est un système qui correspond à au moins une des définitions présentées en début de cette section. Les différents seuils à respecter pour qu'un système soit candidat sont représentés par les lignes horizontales en pointillés. Le signe "+" (respectivement "o") sous le nom des systèmes indique le(s) système(s) finalement retenus comme le(s) plus performant(s) (resp. robuste(s)).

D'après cette figure, sept systèmes de notre échantillon sont candidats au système de RI le plus performant. Cependant, le système READWARE2 (signe "+") est parmi les 10 meilleurs systèmes de RI pour tous les besoins classés très difficiles sauf deux. De plus, aucun autre candidat au système le plus performant pour ce groupe de besoins n'a de telles performances. READWARE2 est donc désigné comme le système le plus performant pour les besoins très difficiles.

Ici, cinq systèmes de RI peuvent correspondre au système le plus robuste car ils n'ont que des rangs inférieurs à 80 : CL99XT, INQ604, ibms99b, ok8alx et tno8d4. Cependant, CL99XT a 75% de ses rangs (Q3) inférieur à 40 ce qui signifie qu'il est classé 75% du temps dans le premier tiers des systèmes pour les besoins très difficiles.

En conséquence, le système CL99XT est choisi comme système le plus robuste pour ce groupe de besoins.
Nous pouvons procéder à des analyses similaires pour les différents groupes de besoins obtenus en 4.1. Le tableau 1 présente les résultats de ces analyses. Ce tableau

Groupes de besoins	**Système le plus**	
	perfomant	**robuste**
Faciles	MITSLStd	MITSLStd
Moyens	CL99XT	READWARE2
Difficiles	READWARE2	ibms99a
Très Difficiles	READWARE2	CL99XT

Tableau 1. *Résultats de la sélection des rangs MeMo par groupe de besoins*

nous montre que la méthode de sélection basée sur les rangs MeMo permet de choisir le meilleur système à utiliser selon deux critères. Le premier critère est la difficulté du besoin en information à traiter. Le second concerne les attentes de celui qui exprime ce besoin : attend-il les meilleurs résultats en acceptant quelques échecs ? ou préfère-t'il obtenir des résultats un peu moins bons mais qui le restent toujours ?

Cependant, pour un groupe de besoins fixé, le choix des "meilleurs" systèmes de RI n'est pas toujours trivial et nécessitent une étude plus approfondie sur les rangs des systèmes candidats. Nous supposons que cela est dû aux ex-æquos possibles lors de la transformation des MeMo en rangs. Nous procédons donc à une étude similaire des meilleurs systèmes de RI par groupe de besoins en information, cette fois basée directement sur les MeMo.

5.2. *Détection des meilleurs systèmes par groupe de besoins par la mesure moyenne*

Les critères de sélection du système le plus performant et du système le plus robuste sont analogues à celles présentées précédemment. L'étude est adaptée à la mesure moyenne. Dans l'étude précédente, une valeur de rang élevée était mauvaise puisque cela signifie que le système était classé parmi les moins bons. Désormais nous travaillons sur des mesures de performance : une valeur élevée signifie que le système fait partie des meilleurs.

La figure 5 présente les systèmes de RI candidats selon les MeMo pour le groupe de besoins très difficiles.

Les médianes respectives du système "READWARE2" et du système "orcl99man" sont nettement supérieures aux médianes des autres systèmes. La position de la médiane de "READWARE2" proche du troisième quartile montre une forte densité des MeMo obtenues dans l'intervalle [0,45 ; 0,50]. Cela nous pousse à le choisir comme unique système le plus performant pour ce groupe. Étant donné que les moins bonnes valeurs de MeMo de READWARE2 sont supérieures aux MeMo de tous les autres

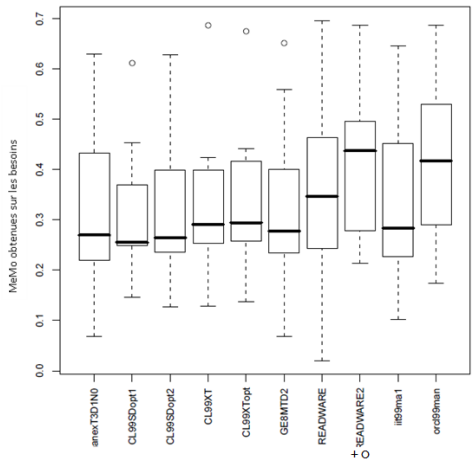

Figure 5. *MeMo : candidats aux meilleurs systèmes - groupe très difficile*

candidats, le système "READWARE2" est également choisi ici comme système le plus robuste. Le tableau 2 présente l'ensemble des systèmes retenus par groupe de besoins en information selon les MeMo.

Groupes de besoins	**Système(s) retenu(s) comme le(s) plus performant(s)**	**Système(s) retenu(s) comme le(s) plus robuste(s)**
Faciles	—	—
Moyens	ii99mal	Flab8x, ok8amxc, CL99SDopt1
Difficiles	READWARE2	ibms99a
Très Difficiles	READWARE2	READWARE2

Tableau 2. *Résultats de la sélection des MeMo par groupe de besoins*

Pour les groupes de besoins faciles et moyens, le choix d'un unique meilleur système est une tâche difficile voire impossible. En effet, ces groupes de besoins sont définis tels que tous les systèmes réussissent à bien les traiter : les performances obtenues par les systèmes sur le groupe facile sont très bonnes et donc très peu discriminantes pour choisir les "meilleurs" systèmes. Ceci n'est pas un problème, car comme tous les systèmes sont performants sur ces besoins, améliorer les résultats est une tâche compliquée. De plus, améliorer un processus déjà très robuste et performant n'est pas nécessaire.

5.3. *Comparaison des systèmes retenus par les deux approches*

Le tableau 3 montre que les deux méthodes de sélection produisent des résultats similaires pour les besoins en information les plus difficiles (groupes difficiles et très difficiles) et de plus larges variations dans les choix pour les besoins moins difficiles

(classés moyens et faciles). Nous remarquons également que la méthode de sélection basée sur les rangs a tendance à gommer les faibles écarts entre différents systèmes. Cela présente l'avantage de proposer un choix unique de "meilleur" système lorsque plusieurs obtiennent des performances proches les unes des autres. Les besoins les

Groupes de besoins	**Système le plus performant**		**Système le plus robuste**	
	Rangs MeMo	MeMo	Rangs MeMo	MeMo
Faciles	MITSLStd	—	MITSLStd	—
Moyens	CL99XT	ii99mal	READWARE2	Flab8x ok8amxc CL99SDopt1
Difficiles	READWARE2	READWARE2	ibms99a	ibms99a
Très Difficiles	READWARE2	READWARE2	CL99XT	READWARE2

Tableau 3. *Comparaison des méthodes de sélection par groupe de besoins*

plus difficiles sont aussi les plus discriminants pour les systèmes, c'est à dire qu'ils permettent de repérer plus facilement les meilleurs systèmes de RI et donc de les sélectionner. Nous avons vu que la méthode de sélection des systèmes basée sur le rang MeMo permet toujours de faire le choix d'un unique système au contraire de la sélection directement basée sur les MeMo. Finalement, la méthode de sélection basée sur les rangs MeMo donnent des résultats similaires à la sélection basée sur les MeMo pour les besoins les plus difficiles. Elle permet aussi d'opérer une sélection unique pour les besoins les moins difficiles donc la sélection sur les rangs MeMo devrait être préférée. Afin de valider ces observations, nous procédons à l'évaluation des deux méthodes de sélection.

5.4. *Évaluation des méthodes de sélection*

Ici, nous cherchons à déterminer si la robustesse (resp. la performance) des méthodes de sélection est meilleure que la robustesse (resp. la performance) des systèmes considérés initialement. En d'autres termes, nous cherchons à vérifier si les méthodes de sélection peuvent apporter une plus-value à un ensemble de systèmes de RI existants.

Sélection des systèmes de référence. Un système de référence doit être sélectionné pour tester les méthodes selon l'approche désirée : la performance ou la robustesse. Pour tester la performance, le système qui a obtenu la meilleure mesure moyenne sur l'ensemble des besoins est sélectionné. Ce système est "orcl99man" avec une mesure moyenne générale de 0.5220. Cette moyenne générale est utilisée comme référence pour évaluer la performance des méthodes de sélection sur l'ensemble des besoins. Pour affiner l'évaluation, nous procédons également à des tests sur chacun des groupes de besoins. La référence alors utilisée est la moyenne des mesures moyennes obtenues par "orcl99man" sur les besoins du groupe étudié.
Pour tester la robustesse, nous pourrions nous contenter d'utiliser le système ayant ob-

tenu la variance la plus faible. Le risque est d'obtenir une référence certes très stables mais avec des performances faibles. Pour palier à cela, nous sélectionnons les deux meilleurs tiers des systèmes sur l'ensemble des besoins considérés ; parmi les systèmes restants, celui à plus faible variance est choisi comme référence.

Création des méta-exécutions à partir des sélections. La procédure d'évaluation est identique pour les deux méthodes. Pour chaque groupe de besoins, le système sélectionné pour ce groupe est utilisé. La moyenne des mesures de performance du système est calculée pour chaque besoin du groupe. Si pour un groupe de besoin, plusieurs systèmes sont retenus, la moyenne de leurs mesures moyennes est utilisée. En concaténant les résultats obtenus pour chacun des groupes, nous obtenons les mesures moyennes d'une méta-exécution de tous les besoins. Une méta-exécution est utilisée pour les tests de performance et une seconde est utilisée pour les tests de robustesse.

Tests statistiques. La procédure de test est identique pour chaque groupe de besoins et pour l'ensemble des besoins. Pour tester si la différence des performances entre le système de référence et la méta-exécution est significative, nous calculons le t-test (ou test de Student) au seuil de 5%. Pour cela, nous comparons les mesures moyennes obtenues par le système référence avec les mesures moyennes de la méta-exécution sur les besoins considérés. Le t-test unilatéral est employé afin de déterminer si la performance de la méta-exécution est significativement supérieure (resp. significativement inférieure) au point de comparaison lorsque celle-ci est supérieure (resp. inférieure) à la référence. Pour tester si la robustesse de la méta-exécution est significativement différente de la robustesse de la référence, nous procédons au test de Fisher. Ce test nous permet d'établir si la variance des mesures moyennes obtenues par la méta-exécution est significativement plus petite ou plus grande que la variance des mesures moyennes du système de référence. Un t-test est également réalisé sur les moyennes pour comparer les performances de ces exécutions.

Résultats de l'évaluation. Les tableaux suivants présentent les résultats obtenus par les méthodes de sélection pour le système le plus performant (tableau 4) et pour le système le plus robuste (tableau 5). Dans ces tableaux, "(++)" (resp. "(- -)") indique que la sélection est significativement supérieure (resp. inférieure) à la référence selon le test employé ; "(+)" et "(-)" indiquent un écart non significatif ; "(=)" indique qu'en valeur absolue l'écart est inférieur à 1%.

Groupes de besoins	**MeMo**		**Rangs MeMo**	
	Référence	Sélection	Référence	Sélection
Faciles	—	—	0.6792	(+) 0.6974
Moyens	0.6110	(+) 0.6253	0.6110	(=) 0.6082
Difficiles	0.4723	(+) 0.5195	0.4723	(+) 0.5195
Très Difficiles	0.4215	(=) 0.4207	0.4215	(=) 0.4207
Tous	0.4920	(+) 0.5188	0.5220	(+) 0.5440

Tableau 4. *Évaluation de la sélection du système le plus performant - test sur la moyenne*

Dans le tableau 4, nous observons que hormis pour les besoins les plus difficiles, la sélection basée sur la MeMo obtient des performances supérieures à celles de la référence. La sélection sur les rangs MeMo est également meilleure pour les besoins faciles et difficiles et obtient des résultats très proches de la référence pour les deux autres groupes de besoins. Le meilleur gain est obtenu pour le groupe de besoins difficiles avec une amélioration de 10,0% par rapport à la référence.

Groupes de besoins	**MeMo**				**Rangs MeMo**			
	Référence		Sélection		Référence		Sélection	
	Moyenne	Variance	Moyenne	Variance	Moyenne	Variance	Moyenne	Variance
Faciles	—	—	—	—	0,6066	0,00703	(++) 0,6974	(-) 0,00439
Moyens	0,4805	0,00431	(++) 0,5831	(- -) 0,00138	0,4805	0,00431	(++) 0,6298	(+) 0,00498
Difficiles	0,3992	0,00553	(++) 0,4513	(-) 0,00368	0,3992	0,00553	(++) 0,4513	(-) 0,00368
Très Difficiles	0,2306	0,00470	(++) 0,4207	(++) 0,02749	0,2306	0,00470	(++) 0,3321	(++) 0,02134
Tous	0,3744	0,01341	(++) 0,4752	(-) 0,01273	0,4116	0,01962	(++) 0,4789	(+) 0,02987

Tableau 5. *Évaluation de la sélection du système le plus robuste - tests sur la variance et la moyenne*

Dans le tableau 5, pour les besoins de difficulté moyenne, nous observons que la variance de la sélection MeMo est significativement inférieure à celle de la référence ; on en déduit que la sélection est significativement plus robuste que la référence pour ces besoins avec un gain de 68% en robustesse avec des performances 21% plus élevées que celles de la référence. Au contraire, sur les besoins très difficiles, la référence est significativement plus robuste. Cependant, la sélection MeMo obtient des performances significativement plus élevées quelque soit le groupe de besoin considéré.

La sélection sur les rangs MeMo permet d'obtenir une meilleure robustesse lorsque les besoins faciles et difficiles sont traités. Là encore, la robustesse est significativement moins bonne que celle de la référence pour les besoins très difficile malgré des performances supérieures.

Lorsque tous les besoins sont considérés, les deux méthodes de sélection des systèmes selon la robustesse permettent d'obtenir des meilleurs performances que la référence. La sélection MeMo améliore la robustesse de 5% en conservant des performances de 27% supérieures à la référence. La sélection sur les rangs MeMo quant à elle dégrade la robustesse de moitié en conservant des performances 16% meilleures.

6. Conclusion et travaux à venir

Nous avons vu deux méthodes de sélection des systèmes. Une méthode consiste à calculer une unique mesure pour quantifier la performance moyenne de chaque système. La seconde méthode introduit les rangs des systèmes selon leur performance moyenne. Nous avons montré qu'avec ces méthodes il est possible de choisir le meilleur système selon un critère de robustesse ou de performance pour différents groupes de difficulté des besoins en information. La sélection des meilleurs systèmes est compliquée pour les besoins classés faciles ou moyens : ceux-ci sont très peu discriminants pour les systèmes de RI car par définition, les systèmes réussissent tous à

traiter ces besoins avec succès. Pour les groupes de besoins difficiles et très difficiles, le choix d'un unique meilleur système basé sur la performance ou sur la robustesse est plus aisé. L'évaluation montre que, à partir d'un ensemble de systèmes de RI donné, les méthodes de sélection analysées dans ce papier sont capables d'améliorer les résultats obtenus par les systèmes de manière individuelle. L'évaluation montre que l'on peut atteindre une amélioration de 10% des performances si celle-ci est privilégiée. Il est également montré que la robustesse peut être améliorée significativement jusqu'à 68% de gain tout en conservant de bonnes performances.

Pour les deux méthodes, un choix final des systèmes est opéré à la main et introduit un biais subjectif dans la sélection. Une future analyse devrait proposer une façon d'automatiser la détection des seuils fixés arbitrairement dans ce papier et analyser les différences de performances selon les valeurs de seuil retenues. L'emploi de méthodes statistiques telle que les skylines, les arbres aléatoires ou les analyses factorielles pourraient être étudiées pour détecter les valeurs des seuils à employer.

Dans ce papier, nous considérons des classes de besoins en information définies selon la difficulté que les systèmes de recherche d'information ont à les traiter. Lorsqu'une nouvelle requête arrive, il faut déterminer quel système doit la traiter. Il existe des travaux qui prédisent la difficulté, par exemple à partir de traits linguistiques (Carmel *et al.,* 2010).

Les travaux futurs devraient proposer une méthode d'apprentissage en commençant par entrainer les méthodes de sélection sur un ensemble connu de besoins en information. Ensuite, à partir de la difficulté prédite d'un besoin en information nouveau, au sens où il n'a pas été utilisé lors de l'apprentissage, le système sélectionné pour le groupe de difficulté concerné devra être utiliser. Avec un tel procédé, l'efficacité et les gains apportés par les méthodes de sélections pourront être vérifiés.

7. Remerciements

Nous tenons à remercier l'ANR pour son soutien financier à ces recherches au travers du projet ANR-2010-CORD-001-01 CAAS (contextual analysis and adaptive search, http ://www.irit.fr/CAAS/) ainsi que la fédération CNRS FREMIT FR3424 (http ://www.irit.fr/FREMIT/).

8. Bibliographie

Baccini A., Déjean S., Mothe J., Lafage L., « How many performance measures to evaluate Information Retrieval Systems ? », , vol. 30, n° 3, p. 693-713, 2011.

Bigot A., Chrisment C., Dkaki T., Hubert G., Mothe J., « Fusing different information retrieval systems according to query-topics : a study based on correlation in information retrieval systems and TREC topics », *Information Retrieval Journal*, vol. 14, n° 6, p. 617-648, 2011.

Brown P. J., Jones G. J. F., « Context-aware Retrieval : Exploring a New Environment for Information Retrieval and Information Filtering », *Personal and Ubiquitous Computing*, vol. 5, n° 4, p. 253-263, 2001.

Carmel D., Yom-Tov E., *Estimating the Query Difficulty for Information Retrieval*, Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool (ed.), 2010.

Cleverdon C., Mills J., Keen M., *Factors Determining the Performance of Indexing Systems*, vol. 1, ASLIB Cranfield Research Project, 1966.

Fox E. A., Shaw J. A., « Combination of multiple searches », *TREC-2, Proceedings of the Second Text REtrieval Conference*, D. Harman (ed.), p. 243-249, 1994.

Harman D., Buckley C., « Overview of the Reliable Information Access Workshop », *Information Retrieval*, vol. 12, n° 6, p. 615-641, 2009.

Jain A. K., Murty M. N., Flynn P. J., « Data Clustering : A Review », *CSURV : Computing Surveys*, vol. 31, n° 3, p. 264-323, 1999.

Joho H., Urban J., Villa R., Jose J. M., van Rijsbergen C. J., « AIR 2006 : First International Workshop on Adaptive Information Retrieval », *SIGIR Forum : Special Interest Group on Information Retrieval Forum*, vol. 42, n° 1, p. 63-66, 2008.

Kompaoré D., Mothe J., Baccini A., Déjean S., « Prédiction du SRI à utiliser en fonction des critères linguistiques de la requête », *COnference en Recherche d'Information et Applications*, Université de Saint-Étienne, p. 239-254, 2007.

Lebart L., Piron M., Morineau A., *Statistique exploratoire multidimensionnelle : visualisations et inférences en fouille de données.*, Dunod, 2006.

Lee J.-H., « Analyses of Multiple Evidence Combination », *SIGIR : Special Interest Group on Information Retrieval*, ACM, p. 267-276, 1997.

Lillis D., Toolan F., Mur A., Peng L., Collier R. W., Dunnion J., « Probability-based fusion of information retrieval result sets », *Artificial Intelligence Review*, vol. 25, n° 1-2, p. 179-191, 2006.

Liu H., Wu Z., Hsu D. F., « Combination of Multiple Retrieval Systems Using Rank-Score Function and Cognitive Diversity », *AINA : Advanced Information Networking and Applications*, IEEE, p. 167-174, 2012.

Menegon D., Mizzaro S., Nazzi E., Vassena L., « Evaluating Mobile Proactive Context-Aware Retrieval : An Incremental Benchmark », *ICTIR : International Conference on the Theory of Information Retrieval*, vol. 5766 of *Lecture Notes in Computer Science*, Springer, p. 362-365, 2009.

Mizzaro S., Nazzi E., Vassena L., « Retrieval of context-aware applications on mobile devices : how to evaluate ? », *IIiX*, vol. 348 of *ACM International Conference Proceeding Series*, ACM, p. 65-71, 2008.

Mothe J., Tanguy L., « Linguistic features to predict query difficulty - a case study on previous TREC campaigns », *ACM Conference on research and Development in Information Retrieval, SIGIR : Special Interest Group on Information Retrieval, Predicting query difficulty - methods and applications workshop*, p. 7-10, 2005.

Ward J. H., « Hierarchical grouping to optimize an objective function », *Journal of the American Statistical Association*, vol. 58, p. 236-244, 1963.