MOR : Mesure orientée rappel pour les systèmes de recherche d'information

Bissan Audeh, Philippe Beaune, Michel Beigbeder

MOR : Mesure orientée rappel pour les systèmes de recherche d'information

Bissan Audeh, Philippe Beaune, Michel Beigbeder

Institut Henri FAYOL, École des Mines de Saint-Étienne
Institut Mines-Télécom
158 Cours Fauriel, CS62362, 42023 Saint-Étienne, France
{audeh,beaune,mbeig}@emse.fr

RÉSUMÉ. *La majorité des métriques d'évaluation en recherche d'information donnent plus d'importance à la précision qu'au rappel, ce qui est cohérent avec certains contextes, comme la recherche sur internet, où les utilisateurs regardent uniquement les premiers documents apparaissant en début de la liste de résultats. Ces métriques d'évaluation ne sont pas capables de prédire à quel point l'utilisateur sera satisfait du système s'il cherche à trouver le maximum de documents pertinents, comme dans le cas de la recherche de brevets ou de la recherche dans un contexte médical. Dans cet article, nous définissons les besoins principaux pour évaluer et comparer des systèmes de recherche d'information dans un contexte orienté rappel. Nous montrons que les mesures précédemment proposées dans la littérature ne répondent pas à ces besoins, ce qui nous amène à proposer la mesure* MOR *qui permet de comparer correctement le rappel des systèmes de recherche d'information.*

ABSTRACT. *Most evaluation metrics in Information Retrieval (IR) give more weight to precision, which is consistent within several contexts, such as web retrieval, where users check only few documents appearing at the beginning of the result lists. These metrics are not able to measure the satisfaction of the user in contexts where the IR system is supposed to find the maximum of relevant documents, which is the case of patent or medical retrieval. In this article, we define the basic elements that should be considered to evaluate and compare IR systems correctly in a recall-oriented context. We demonstrate that previous measures do not consider the totality of these elements. We propose MOR which is a recall oriented measure that allows the evaluation and the comparison of information retrieval systems in a recall context.*

MOTS-CLÉS : *recherche d'information, métriques d'évaluation, rappel, rappel normalisé.*

KEYWORDS: *information retrieval, evaluation metrics, recall, normalized recall.*

1. Introduction

L'évaluation dans le domaine de la recherche d'information est un problème compliqué. Cela vient du fait que la pertinence est une notion subjective : le même document peut être jugé diversement par deux utilisateurs, ou dans deux conditions différentes. Ainsi, rendre les machines capables d'évaluer l'efficacité des systèmes de recherche d'information n'est pas facile. Cependant, cela n'a pas empêché les chercheurs de proposer des mesures d'évaluation qui permettent de comparer les systèmes de recherche d'information. La plupart de ces métriques se basent sur les deux notions historiques : la précision [1] et le rappel [2] (Cleverdon *et al.*, 1966 ; Jones, 1981), où dans le cas idéal, on trouve tous les documents pertinents au début de la liste de résultats. Alors que la précision est une notion facile à évaluer, car elle ne demande aucune information sur le nombre total de documents pertinents dans la collection, le rappel est une notion qui pose plusieurs problèmes. Il est d'ailleurs impossible de l'évaluer dans les cas réels où l'utilisateur ne connaît pas à l'avance le nombre de documents pertinents correspondant à sa requête.

Ce que nous proposons dans cet article est une nouvelle mesure qui permet d'évaluer la capacité d'un système à trouver le plus de documents pertinents sans pour autant négliger leur ordre d'apparition. Nous commençons en section 2 par définir le contexte dans lequel nous effectuons notre travail et les besoins que l'on cherche à satisfaire. Puis, en section 3 nous analysons les mesures existantes dans l'état de l'art pour l'évaluation du rappel. Dans la section 4 nous proposons la *mesure orientée rappel* (MOR) qui répond aux contraintes que nous avons définies, et qui ne sont pas remplies de façon satisfaisante par les mesures précédentes. Les caractéristiques et l'évaluation de cette mesure sont proposées dans les sections 5 et 6 respectivement. Nous terminons par un résumé dans la section 7.

2. Définition du contexte et des besoins

Pour pouvoir étudier le rappel, nous nous intéressons aux situations où l'utilisateur souhaite trouver le maximum de documents pertinents, comme dans les cas de recherche de brevets ou de dossiers médicaux. La priorité de l'utilisateur dans ces contextes est le rappel. Pour cette raison, il est prêt en général à évaluer plusieurs pages de résultats pour sa requête, contrairement aux contextes de recherche dirigée précision comme la recherche sur le Web. Ainsi, nous considérons qu'une mesure orientée rappel doit satisfaire les deux besoins suivants :

– **B1** : Pouvoir favoriser un système qui restitue plus de documents pertinents.

1. Pour une requête q, la précision est le pourcentage de documents pertinents trouvés par rapport au nombre total de documents trouvés par le système.

2. Le rappel est le pourcentage de documents pertinents trouvés par rapport au nombre total de documents pertinents dans la collection pour une requête q.

– **B2** : Pour deux systèmes qui rendent le même nombre de documents pertinents, pouvoir favoriser celui qui les donne plus tôt que l'autre dans sa liste de résultats.

Les sections suivantes abordent les mesures d'évaluation de l'état de l'art en considérant ce contexte. Pour mieux illustrer le comportement des différentes mesures, nous montrons dans le tableau 1 une extension de l'exemple proposé par (Magdy, Jones, 2010) : étant donnés cinq systèmes de recherche d'information évalués par rapport aux réponses rendues pour une seule requête q qui rend 100 documents, en sachant que le nombre de documents pertinents dans la collection pour la requête q est 4. Les systèmes sont ordonnés d'une façon décroissante par rapport à un un contexte de rappel qui exige les besoins *B1* et *B2*, donnant la priorité au Rappel puis au rang du dernier document pertinent retourné : le système 1 est le meilleur alors que le système 5 est le moins performant. Les mesures d'évaluation présentées dans ce tableau seront expliquées dans la section suivante. Le tableau 1 sera utilisé pour montrer le comportement de chaque mesure et pour justifier notre proposition.

Tableau 1. Les différentes mesures d'évaluation orientées rappel pour cinq exemples

Système	Rang des doc. pert.	AP	Rappel	F_1	F'_1	F'_4
système 1	$\{1, 2, 3, 4\}$	1	1	$0,077$	1	1
système 2	$\{50, 51, 53, 54\}$	$0,047$	1	$0,077$	$0,092$	$0,459$
système 3	$\{1, 98, 99, 100\}$	$0,273$	1	$0,077$	$0,429$	$0,864$
système 4	$\{1, 54\}$	$0,259$	$0,500$	$0,341$	$0,341$	$0,474$
système 5	$\{1\}$	$0,250$	$0,250$	$0,019$	$0,250$	$0,250$

3. Évaluation du rappel en recherche d'information

Comment évaluer le rappel ? Quand on affirme avoir amélioré le rappel d'un système, quelle mesure faut-il utiliser pour justifier cette amélioration ? Malgré la simplicité et la clarté de ces questions, la réponse n'est pas simple. Contrairement à la précision, l'utilisation de la notion pure du rappel, n'est pas une mesure correcte, car le fait de rendre toute la collection de documents va garantir d'obtenir le rappel maximal. De plus, cette mesure seule n'est pas capable de distinguer deux systèmes qui fournissent le même nombre de documents pertinents dans des positions différentes de la liste des résultats. Dans les campagnes d'évaluation comme TREC et CLEF, les mesures officielles d'évaluation pour les tâches orientées rappel (tels que les tâches médicales ou légales) sont très nombreuses. Nous présentons dans les sections suivantes les mesures d'évaluation les plus utilisées par ces campagnes en focalisant sur leur efficacité par rapport aux besoins définis dans la section 2.

3.1. Mesures d'évaluation traditionnelles

Dans la littérature de la recherche d'information, la mesure la plus utilisée est la MAP (*Mean Average Precision*)(Sanderson, 2010) qui est calculée par l'équation 1.

$$MAP = \frac{\sum_q AP_q}{Q} \tag{1}$$

où Q est le nombre de requêtes jugées et AP_q (*Average Precision*) est la précision moyenne pour une requête q qu'on obtient à partir de l'équation 2 (Harman, 1994) .

$$AP_q = \sum_{x<|Ren(q)|} \frac{Per(d_x, q) \cdot P_{@x}(q)}{n_q} \tag{2}$$

où $Ren(q)$ est la liste de documents rendus pour la requête q, n_q est le nombre de documents pertinents dans la collection pour la requête q, $P_{@x}(q)$ est la précision à la position x de la liste de résultats, et $Per(d_x, q)$ est la valeur booléenne de la pertinence entre la requête q et le document d à la position x, elle prend la valeur 1 si le document est pertinent, 0 sinon.

Le tableau 1 nous permet de voir clairement que la MAP, qui est aussi le AP [3] dans notre exemple d'une seule requête, n'est pas la mesure adaptée à notre contexte. Par exemple, selon la MAP, le système 5 est cinq fois meilleur que le système 2 même si ce dernier a réussi à rendre tous les documents pertinents.

Pour mieux prendre en considération le rappel, on peut utiliser la mesure F_β (Rijsbergen, 1979) qui est une extension de la moyenne harmonique de la précision (P) et du rappel (R) comme présentée par l'équation 3. Le paramètre β dans cette équation permet d'équilibrer l'importance du rappel par rapport à celui de la précision.

$$F_\beta = \frac{(1+\beta^2)P \cdot R}{\beta^2 \cdot P + R} \tag{3}$$

Dans le tableau 1, on peut constater que la mesure F_1 (F_β avec β=1) ne distingue pas la performance des systèmes 1, 2 et 3. Elle n'est donc pas capable d'ordonner les systèmes par rapport aux rangs des documents trouvés quand le nombre de documents pertinents trouvés est identique. Pour résoudre ce problème, une extension de cette mesure utilisant la précision moyenne au lieu de la précision pure peut être utilisée comme dans l'équation 4.

$$F'_\beta = \frac{(1+\beta^2)AP \cdot R}{\beta^2 \cdot AP + R} \tag{4}$$

En observant le tableau 1, on peut constater que l'utilisation de *Average Precision* AP au lieu de P a résolu le besoin B2, c'est-à-dire que F' est capable de distinguer les systèmes par rapport au rang des documents pertinents trouvés. Par contre, la mesure F' a largement favorisé les systèmes qui ont trouvé des documents tôt dans la liste de résultats par rapport aux autres qui ont détecté plus de documents pertinents. Par exemple, le système 4 qui ne rend que deux documents pertinents obtient des scores F'_1 et F'_4 plus élevés que ceux obtenus par le système 2 qui a rapporté tous les documents pertinents. Cela s'est produit, car le premier document pertinent rendu pour

3. Nous utilisons la notation AP au lieu de AP_q, et n au lieu de n_q dans le reste de ce papier, dès lors que nous ne considérons qu'une seule requête.

le système 4 est en première position, alors qu'il n'est qu'en cinquantième position pour le système 2. La mesure F' n'est donc pas capable de répondre au besoin B1 même en donnant au rappel un poids quatre fois plus élevé qu'à la précision dans l'équation 4.

D'autres mesures d'évaluation existent dans le domaine de la recherche d'information. La plupart de ces métriques mesurent, comme la MAP, la capacité d'un système à retrouver des documents pertinents tôt dans la liste. Parmi ces alternatives, on peut nommer le $GMAP$ (*Geometric Mean Average Precision*) (Robertson, 2006) qui est la moyenne géométrique des AP, et le MRR (*Mean Reciprocal Rank*) (Voorhees, Tice, 1999) qui est utilisé lorsque l'utilisateur n'a besoin que d'un seul document, il s'agit de l'inverse du rang du premier document pertinent trouvé. Par ailleurs, certaines mesures considèrent le degré de pertinence des documents trouvés, comme la mesure $NDCG$ (*Normalized Discounted Cumulative Gain*) (Järvelin, Kekäläinen, 2002). $NDCG$ récompense ou pénalise le score d'évaluation en prenant en compte le niveau de pertinence d'un document et le rang auquel il se trouve dans la liste des résultats. Quand on ne dispose pas de jugements de pertinence gradués, cette mesure a une corrélation très forte avec la précision moyenne (Sakai, 2007), elle n'est donc pas considérée comme une mesure d'évaluation du rappel. Une autre mesure qui peut paraître adaptée à une recherche orientée rappel est la précision au rang R (R-Prec)(Harman, 1994), où R est le nombre de documents pertinents (*Relevant*) dans la collection pour une requête. Cette mesure reste une variante de la précision pure à un rang précis dans la liste de résultats, donc elle n'est pas adaptée à un contexte de rappel. Par exemple, avec R qui vaut 4 dans notre exemple précédent, le système 2 qui a rendu tous les documents pertinents va obtenir un score R-Prec de 0, alors que le R-Prec du système 5 vaut 0,25 bien que le système 5 n'ait rendu qu'un seul document pertinent.

3.2. Mesures basées sur le rappel normalisé

Nous avons mentionné, jusqu'à présent, les mesures traditionnelles en recherche d'information les plus susceptibles d'être adaptées à un contexte orienté rappel. Nous avons illustré, en utilisant des exemples simples, que ces mesures ne sont pas capables de correspondre au jugement humain dans notre contexte. Pour cette raison, plusieurs travaux ont essayé récemment d'étudier ce problème (Zobel *et al.*, 2009 ; Magdy, Jones, 2010 ; Magdy, 2012 ; Webber, 2010). On trouve dans la littérature certains papiers remettant en question la notion du rappel elle-même, comme le papier de (Zobel *et al.*, 2009) qui critique l'ambiguïté de l'utilisation du rappel à cause de son lien avec plusieurs aspects, comme la « totalité », c'est-à-dire la capacité de trouver la totalité de documents pertinents, et la « persistance » qui indique la volonté de l'utilisateur d'examiner des documents aux rangs avancés dans la liste de résultats. Nous soulignons que nous nous intéressons à un contexte de totalité, où la priorité de l'utilisateur est de trouver, au mieux, tous les documents pertinents pour sa requête, car cette signification correspond le mieux à la définition originale du rappel, et c'est l'évaluation du rappel dans ce sens qui est la plus problématique.

D'autres études proposent de nouvelles mesures pour évaluer le rappel dans le contexte qui nous intéresse. Nous nous intéressons particulièrement au travail de Magdy et Jones (Magdy, Jones, 2010) qui présente la mesure $PRES$. Dans leur étude, les auteurs proposent de reprendre le rappel normalisé ($RNorm$) (Rijsbergen, 1979 ; Rocchio, 1964) et de l'adapter pour qu'il soit utilisable avec les collections de documents de grande taille. Les deux mesures $RNorm$ et $PRES$ se basent sur l'idée de définir un meilleur et un pire scénario pour un processus de recherche (figure 1). Dans le meilleur scénario (S1), tous les documents pertinents sont trouvés au début de

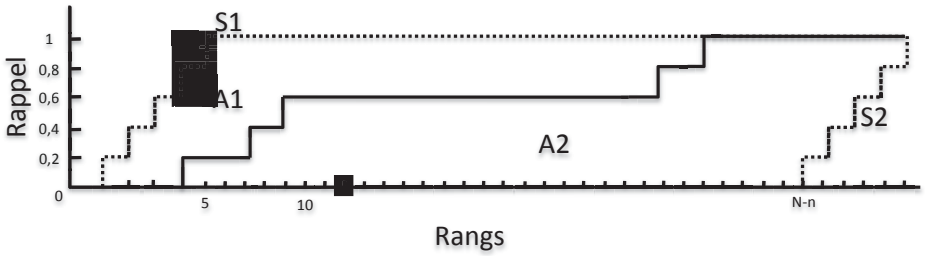

Figure 1. Meilleur et pire scénarios de la mesure $RNorm$

la liste de résultats, dans le pire scénario (S2) ils sont rendus à la fin de la liste à partir de la position $N - n$, où N est le nombre de documents dans la collection et n est le nombre de documents pertinents. Comme présentée par la figure 1, la courbe créée par un cas de recherche réel divise la surface entre S1 et S2 en deux parties : A1 et A2. Ainsi, l'idée du rappel normalisé est de calculer le pourcentage de la surface A2 par rapport à la surface totale entre S1 et S2, ce qui se résume par l'équation 5.

$$RNorm = \frac{A2}{A1 + A2} \tag{5}$$

Alors que cette mesure semble un bon candidat pour le contexte de recherche d'information dirigée rappel, son défaut majeur est que l'on a besoin de juger toute la collection de documents par rapport à une requête, ce qui est impossible avec la taille des collections d'aujourd'hui. Pour cette raison, la mesure $PRES$ (Magdy, Jones, 2010) prend en compte un seuil N_{max} qui représente le nombre maximum de documents que l'utilisateur a l'intention de juger au lieu de la taille totale de la collection N (figure 2).

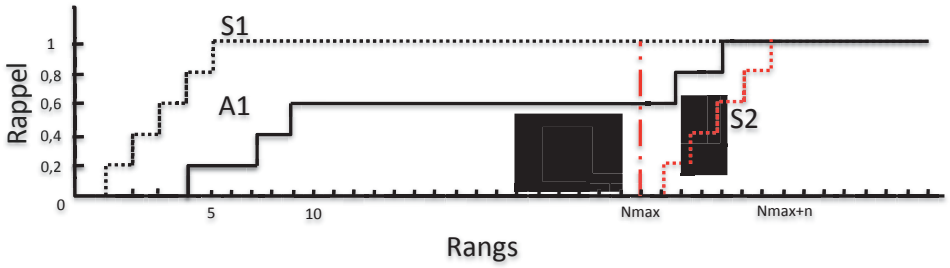

Figure 2. Meilleur et pire scénarios de la mesure $PRES$

L'introduction de ce seuil N_{max} ne modifie pas la définition du meilleur scénario pour $RNorm$, alors que le pire scénario se produit lorsque le système trouve tous les

documents pertinents après la position N_{max}, ce qui signifie que l'utilisateur ne les verra pas. La mesure $PRES$ est exprimée par l'équation suivante :

$$PRES = 1 - \frac{\frac{\sum r_i}{n} - \frac{n+1}{2}}{N_{max}} \tag{6}$$

où n est le nombre de documents pertinents dans la collection, $\frac{\sum r_i}{n}$ est la moyenne des rangs des documents pertinents trouvés avant le rang N_{max} si tous les documents pertinents ont été trouvés avant ce rang, sinon $\sum r_i$ est calculé par l'équation 7 :

$$\sum r_i = \left(\sum_{1<i<nR} r_i \right) + nR(N_{max} + n) - \frac{nR(nR-1)}{2} \tag{7}$$

où R est le rappel à N_{max} [4].

En ajoutant la mesure $PRES$ à notre tableau démonstratif (cf. tableau 2), on

Tableau 2. La mesure $PRES$ comparée aux mesures basées sur AP.

Système	Rang de docs. perts	AP	$Rappel$	F_1	F'_4	$PRES$
système 1	$\{1, 2, 3, 4\}$	1	1	$0,077$	1	1
système 2	$\{50, 51, 53, 54\}$	$0,047$	1	$0,077$	$0,091$	$0,500$
système 3	$\{1, 98, 99, 100\}$	$0,273$	1	$0,077$	$0,429$	$0,280$
système 4	$\{1, 54\}$	$0,259$	$0,500$	$0,341$	$0,341$	$0,370$
système 5	$\{1\}$	$0,250$	$0,250$	$0,019$	$0,250$	$0,250$

constate que la mesure $PRES$ a été capable de corriger les défauts des mesures traditionnelles précédentes. Contrairement à la mesure F', le score $PRES$ ne donne pas plus d'importance aux documents repérés tôt dans la liste par rapport au nombre de documents pertinents trouvés, c'est pour cette raison que le système 4 n'est plus privilégié par rapport au système 2 malgré le document pertinent qu'il a détecté en début de liste. Par contre, $PRES$ pénalise le fait de trouver des documents pertinents à la fin de la liste tronquée à N_{max}. Par exemple, $PRES$ estime que le système 3 est moins bon que le système 4, car le premier a repéré des documents à la fin de la liste, et cela a baissé son score bien qu'il ait trouvé tous les documents pertinents. Ce comportement peut être expliqué par la figure 3 où on constate que la surface créée par un système (A) peut être inférieure à celle créée par un autre système (B) qui trouve moins de documents pertinents.

$PRES$ est une mesure conçue pour la recherche de brevets. Elle est censée prendre en compte d'un côté le rappel et de l'autre côté l'effort que l'utilisateur va fournir pour obtenir le maximum de résultats. Les auteurs voient ce comportement de $PRES$ comme un outil permettant de mesurer l'effort que l'utilisateur doit fournir pendant le processus de recherche : tant que l'utilisateur trouve des documents loin dans la liste,

4. Ce qui revient à dire que nR est le nombre de documents pertinents rendus avant N_{max}.

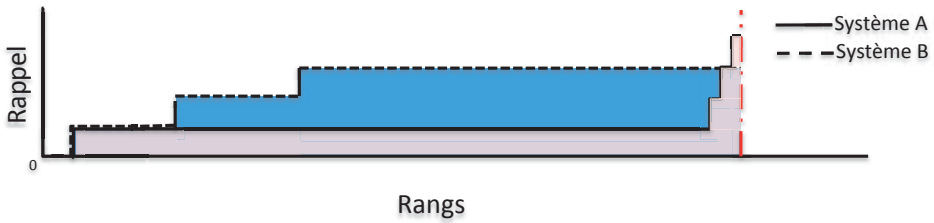

Figure 3. La surface créée par les performances des systèmes A et B

il va continuer de vérifier plus de documents, ce qui signifie plus d'efforts, induisant donc un score faible.

À notre avis, l'hypothèse qui suppose que l'utilisateur préfère trouver moins de documents pertinents plutôt que d'en détecter plus loin dans la liste de résultats est une hypothèse discutable dans un contexte de recherche dirigée rappel. L'effort de l'utilisateur qui est l'élément justifiant le comportement de $PRES$ est une notion subjective dépendant fortement de l'utilisateur lui-même, mais aussi du contexte de la recherche. Dans une recherche de haute persistance (Zobel *et al.*, 2009), surtout avec un système rendant peu de résultats, l'utilisateur est généralement prêt à examiner plusieurs pages de résultats indépendamment des rangs des documents pertinents trouvés, car dans la réalité, il ne connaît pas par avance le nombre total de documents pertinents pour sa requête. De plus, le N_{max} est considéré comme le nombre de documents que l'utilisateur souhaite juger. Dans un cas de recherche de faible persistance, l'utilisateur peut choisir une petite valeur de N_{max} pour exprimer explicitement le nombre de documents qu'il souhaite voir.

4. MOR : une mesure orientée rappel

Dans les sections précédentes, nous avons évoqué le problème de l'évaluation du rappel. En définissant les besoins à remplir lors d'une évaluation dirigée rappel, et à l'aide d'exemples simples, nous avons constaté que la majorité des mesures de l'état de l'art ne remplissent pas ces besoins, car elles sont fortement influencées par les documents pertinents trouvés en début de liste ou alors pénalisent les documents repérés à la fin de la liste, ce qui, dans les deux cas perturbe leur capacité de détecter les systèmes qui ont un meilleur rappel.

Pour définir une nouvelle mesure, il faut avoir une vision claire du comportement qu'elle doit posséder. Afin d'être le plus proche possible d'un jugement humain sur la préférence d'un système par rapport à un autre dans un contexte dirigé rappel, nous cherchons à construire une mesure d'évaluation qui répond aux contraintes bien définies en fonction de cet objectif. Nous commençons par définir les paramètres qui nous permettent de formaliser ces contraintes et d'introduire notre mesure.

4.1. Définition des paramètres

Nous avons constaté dans l'état de l'art que la courbe rappel/rang, proposée à l'origine pour le rappel normalisé (Rijsbergen, 1979 ; Rocchio, 1964) et utilisée par (Magdy, Jones, 2010), est une bonne base pour une mesure orientée rappel. Nous reprenons donc cette courbe dans la figure 4 en remplaçant le rappel par le nombre de documents pertinents pour faciliter le développement de notre mesure.

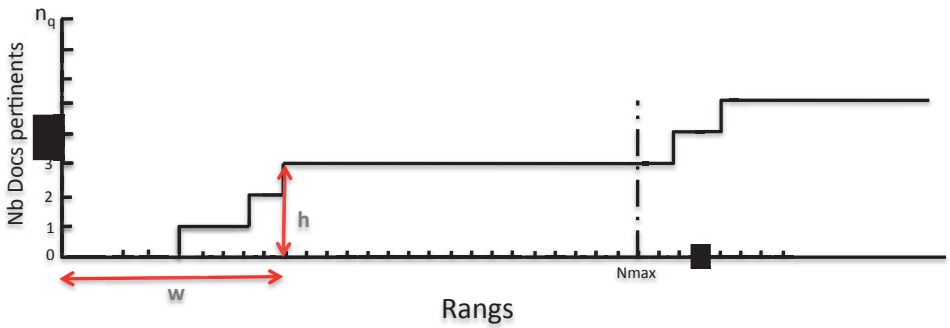

Figure 4. Les notions de hauteur et de largeur selon la mesure MOR

Cette courbe a l'avantage de donner une vision claire de la propagation des documents pertinents dans la liste de résultats, ce qui permet de bien visualiser le rappel et la précision. Pour bien contrôler ces deux notions dans notre mesure, nous définissons les paramètres suivants :

– La hauteur (h) : est le nombre de documents pertinents trouvés avant le rang N_{max}.

– La largeur (w) : est le rang du dernier document pertinent avant N_{max}.

– AP : est la précision moyenne de la liste tronquée à N_{max}.

Pour favoriser un système qui pour le même rappel trouve les documents plus tôt dans la liste, w seul ne suffit pas, car pour le même w, deux systèmes peuvent trouver des documents pertinents à des rangs bien différents. Par ailleurs, l'utilisation de la précision moyenne seule (ou de la précision totale) nous ramène au problème d'être fortement influencé par les documents pertinents trouvés en début de liste. Pour ces raisons, nous considérons qu'une combinaison de w et de AP est indispensable pour juger la précision, alors que pour le rappel, nous estimons qu'il est représenté par le nombre de documents pertinents trouvés, soit h.

Ainsi, la nouvelle mesure doit être une fonction dépendant des trois paramètres h, w, AP (figure 4) et prenant une valeur réelle dans le domaine $[0, 1]$, comme définie par l'équation 8.

$$f(h, w, AP) : \{0, .., \min(n, N_{max})\} \times \{0, .., N_{max}\} \times [0, 1] \rightarrow [0, 1] \qquad (8)$$

Cette fonction doit prendre en considération qu'un système de recherche d'informations cherche à maximiser h et AP et à minimiser w [5], et également tenir compte des contraintes formelles que nous allons définir par la suite.

4.2. Les contraintes formelles

Notre hypothèse est qu'une mesure d'évaluation adaptée à un contexte dirigé rappel doit remplir les contraintes citées dans les sections suivantes.

4.2.1. La priorité au rappel

Dans la section 2 (page 38), nous avons défini deux besoins qu'une mesure d'évaluation doit satisfaire pour évaluer le rappel des systèmes de recherche d'information : donner la priorité au rappel (B1) et ne considérer la précision qu'en cas de rappel équivalent (B2). Nous reprenons ces deux besoins pour formaliser les contraintes de la priorité à l'aide des paramètres h, w et AP. Avec ces trois éléments, nous distinguons trois niveaux de priorité, et ainsi les trois contraintes présentées dans le tableau 3. Dans ce tableau, nous précisons que la priorité est le rappel (h), et que

Tableau 3. Les contraintes de priorité de MOR entre deux triplets (h,w,AP) et (h',w',AP')

Contraintes	Définition
$c1$	$h > h' \Rightarrow f(h, w, AP) > f(h', w', AP')$
$c2$	$h = h', w < w' \Rightarrow f(h, w, AP) > f(h', w', AP')$
$c3$	$h = h', w = w', AP > AP' \Rightarrow f(h, w, AP) > f(h', w', AP')$

pour un h donné, la priorité est ensuite w. AP ne doit influencer la mesure qu'en cas d'impossibilité de juger un système par rapport à un autre selon h et w.

4.2.2. Meilleurs et pires scénarios

À chaque niveau de priorité, la nouvelle mesure doit être capable de détecter le meilleur et le pire scénario. Nous présentons ces scénarios dans le tableau 4. A priori,

Tableau 4. Les meilleurs et les pires scénarios de MOR

Contrainte	Niveau	Meilleur scénario	Pire scénario
$c4$	$Principal$	h=n	h=0
$c5$	$Pourun$h$donn$	w=h	w=N_{max}
$c6$	$Pourun$h$etun$w$donns$	$AP = AP1$	$AP = AP0$

concernant le rappel, le meilleur scénario est de trouver tous les documents pertinents ($h = n$) et le pire est de n'en découvrir aucun ($h = 0$). Pour un h donné, l'idéal est

5. Noter que par définition on a $w \geq h$.

d'obtenir les documents pertinents le plus tôt possible. Dans ce cas, le rang du dernier document pertinent est égal au nombre de documents pertinents trouvés ($w = h$), alors qu'au pire, ce rang se situe à N_{max} ($w = N_{max}$). Pour un h et un w donnés, le meilleur et le pire scénario dépendent de AP. Par contre, rien ne peut garantir qu'AP prendra respectivement la valeur 1 ou 0 pour ces deux scénarios. Pour cela, nous supposons que ces valeurs minimum et maximum sont $AP1$ et $AP0$ dont nous parlons dans la section suivante.

4.3. Construction de la mesure

Pour garantir la priorité de h par rapport à w et à AP (contrainte c1), plusieurs alternatives peuvent être utilisées. Vu que h est un entier positif et strictement inférieur à N, la méthode la plus simple est de considérer la somme de h avec une fonction de w et AP qui prend ses valeurs dans $[0, 1]$, ce que nous présentons dans l'équation 9.

$$f(h, w, AP) = \begin{cases} \frac{h+g_h(w,AP)}{\min(n,N_{max})+1} & \text{si } h > 0 \\ 0 & \text{si } h = 0 \end{cases} \tag{9}$$

où $g_h(w, AP)$ est cette fonction de w et AP pour un h donné, nous normalisons pour que la fonction finale ait une valeur entre 0 et 1 conformément à l'équation 8. La seule exception pour laquelle cette méthode ne peut pas garantir la priorité de h est quand h vaut 0, ce qui correspond à ne trouver aucun document pertinent. Ainsi, dans cette situation nous forçons la fonction finale à avoir la valeur 0, d'où la deuxième ligne de l'équation 9. Cette pratique, en plus de la normalisation, garantira les valeurs 1 et 0 pour le meilleur et le pire scénario respectivement, et satisfera ainsi la contrainte c4.
Nous suivons la même stratégie de contrôle de priorité pour définir la fonction $g_h(w, AP)$ comme cela est précisé par l'équation 10 :

$$g_h(w, AP) = \frac{(N_{max} - w) + g_{hw}(AP)}{(N_{max} - h) + 1} \tag{10}$$

C'est-à-dire, pour garantir la contrainte c2 (w est prioritaire à AP), nous considérons la somme de w qui est un entier positif avec une fonction de AP. Par contre, nous utilisons $(N_{max} - w)$ au lieu de w pour inverser l'effet de w, car plus sa valeur est petite plus le système évalué doit être récompensé. Ainsi, la fonction $g_h(w, AP)$ est définie par l'équation 10 qui est également normalisée pour avoir des valeurs dans $[0, 1]$ (contrainte c5). Il est important de rappeler que la précision moyenne (AP) considère la liste tronquée au rang N_{max}. Ainsi, la prise en compte de la précision moyenne (contrainte c3) passe par la fonction $g_{h,w}(AP)$. Le but de l'utilisation de cette fonction au lieu de l'emploi simple d'AP (qui est déjà dans le domaine $[0, 1]$) est de satisfaire la contrainte c6 : l'utilisation d'AP directement dans $g_h(w, AP)$ va perturber les résultats du meilleur et du pire scénario. Pour garantir l'obtention des bonnes valeurs dans ces cas, nous définissons la fonction $g_{h,w}(AP)$ par l'équation 11.

$$g_{h,w}(AP) = \begin{cases} \frac{AP - AP0}{AP1 - AP0} & \text{si } AP0 \neq AP1 \\ AP & \text{si } AP0 = AP1 \end{cases} \tag{11}$$

où $AP0$ et $AP1$ sont respectivement le meilleur et le pire scénario pour un h et un w donnés (figures 5 et 6). Notons que, quand h est égale à w, le meilleur et le pire scénario sont identiques. Pour cela, la fonction $g_{h,w}(AP)$ prendra la valeur de AP dans ce cas particulier. Par ailleurs, pour calculer $AP0$ et $AP1$, nous utilisons l'équation 12 que nous obtenons en appliquant l'équation 2 pour le meilleur et le pire scénario.

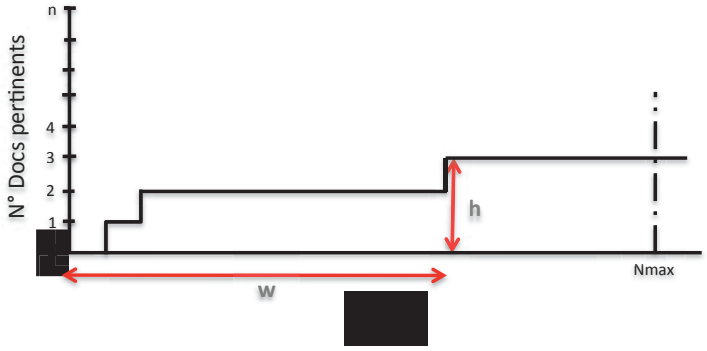

Figure 5. Meilleur scénario pour un h et un w donnés

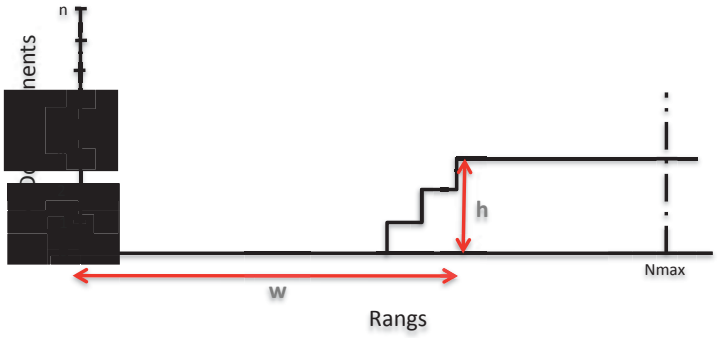

Figure 6. Pire scénario pour un h et un w donnés

$$AP0 = \frac{1}{n} \cdot \sum_{i=1}^{h} \frac{i}{w - h + i} \qquad AP1 = \frac{1}{n} \cdot (h - 1 + \frac{h}{w}) \tag{12}$$

4.4. Équation finale

En considérant les équations 9, 10 et 11, nous obtenons la définition finale de la mesure MOR

$$MOR = \begin{cases} \dfrac{h(N_{max} - h + 1) + N_{max} - w + \frac{AP - AP0}{AP1 - AP0}}{(\min(n, N_{max}) + 1)(N_{max} - h + 1)} & \text{si } h > 0 \text{ et } w > h \\ \dfrac{h(N_{max} - h + 1) + N_{max} - w + AP}{(\min(n, N_{max}) + 1)(N_{max} - h + 1)} & \text{si } h > 0 \text{ et } w = h \\ 0 & \text{si } h = 0 \end{cases}$$

où $AP0$ et $AP1$ sont définies par l'équation 12.

5. Caractéristiques de *MOR*

Avant d'évaluer *MOR*, nous citons dans les sections suivantes certaines de ses caractéristiques, notamment son rôle dans un contexte dirigé rappel, son lien avec le tri à plusieurs clés et enfin son comportement dans le cas d'un seul document pertinent. L'analyse de ces caractéristiques permet de mieux comprendre son comportement dans des contextes différents.

5.1. Le rappel en priorité

Contrairement à la mesure de rappel ordinaire avec laquelle il est très facile d'obtenir le score maximal en rendant toute la collection, il n'existe pas de méthode triviale pour obtenir le score maximal de 1 avec MOR. De plus, cette mesure ne dépend pas de la répartition de la surface, qui représente la performance du système dans la courbe rappel/rangs, par rapport à un meilleur et un pire scénario comme c'est le cas de $PRES$, même si ces deux scénarios sont bien distingués par MOR (cf. tableau. 4). Entre ces deux scénarios, MOR va ordonner les systèmes selon le nombre de documents pertinents trouvés puis selon leur capacité à rendre les documents pertinents plus tôt dans la liste.

Pour faire la démonstration, nous reprenons les exemples de la section 2 pour le cas des cinq systèmes de recherche d'information qu'on souhaite classer par rapport à leur rappel pour une seul requête. Nous rappelons que ces systèmes rendent 100 documents comme résultat de recherche et que la collection contient quatre documents pertinents. Nous présentons à nouveau dans le tableau 5 les valeurs des mesures AP, Rappel, F_1, F'_4, $PRES$ et nous ajoutons la mesure MOR. Nous constatons facilement de ce tableau que MOR est la seule mesure qui classe les systèmes d'une façon identique

Tableau 5. MOR comparé à AP, Rappel, F'_4 et $PRES$

Système	Rang de docs.perts	AP	$Rappel$	F'_4	$PRES$	MOR
système 1	$\{1, 2, 3, 4\}$	1	1	1	1	1
système 2	$\{50, 51, 53, 54\}$	$0,047$	1	$0,092$	$0,500$	$0,895$
système 3	$\{1, 98, 99, 100\}$	$0,273$	1	$0,429$	$0,280$	$0,801$
système 4	$\{1, 54\}$	$0,259$	$0,500$	$0,341$	$0,370$	$0,495$
système 5	$\{1\}$	$0,250$	$0,250$	$0,250$	$0,250$	$0,398$

aux préférences humaines [6]. Le système 5 est le système le moins performant selon MOR malgré le fait qu'il a trouvé un document pertinent au premier rang de la liste de résultats. De plus, contrairement à $PRES$, MOR a été capable de mieux placer le système 3, qui a trouvé tous les documents pertinents, par rapport au système 4 qui n'en a trouvé que la moitié.

Nous constatons qu'il est difficile pour un humain de donner une préférence entre les systèmes 2 et le système 3 qui ont trouvé le même nombre de documents pertinents dans les positions présentées dans le tableau 5. Si, ces documents ont tous été trouvés en début de la liste de résultats du système 2, par exemple, et tous à la fin de la liste du système 3 un utilisateur pourrait plus facilement décider que le système 2 est mieux, ce qui n'est pas le cas dans notre exemple. Nous remarquons que les valeurs de MOR pour ces systèmes (système 2 et système 3) sont très proches (0,895, 0,801) par rapport aux valeurs données par $PRES$ (0,5, 0,28), ce qui nous permet de considérer que MOR représente bien l'hésitation d'un humain pour donner une préférence dans ce cas.

5.2. MOR et le tri à plusieurs clés

Il faut noter que pour une requête donnée, il est possible de trier directement les systèmes de RI par rapport au rappel, puis relativement aux w et AP, ce qui donnera le même résultat qu'en les triant par rapport à leur score MOR. Par contre, le fait d'avoir une mesure numérique capable de reproduire ce tri nous permettra de pouvoir moyenner cette mesure sur l'ensemble des requêtes évaluées afin de pouvoir comparer les systèmes.

5.3. Cas d'un seul document pertinent

Dans le cas particulier où il n'existe qu'un seul document pertinent pour la requête, comme dans le contexte de question/réponse, la mesure souvent utilisée est l'inverse du rang du document pertinent trouvé (MRR). Dans ce cas, l'AP dans la formule de MOR deviendra le MRR, car nous ne pouvons trouver qu'un seul document

6. Nous rappelons que les systèmes sont présenté dans le tableau 5 dans un ordre qui correspond à l'avis d'un utilisateur dans un contexte de rappel.

pertinent. La figure 7 représente le comportement de MOR comparé à celui de $PRES$ et MRR dans le cas d'un seul document pertinent. De cette figure, nous pouvons

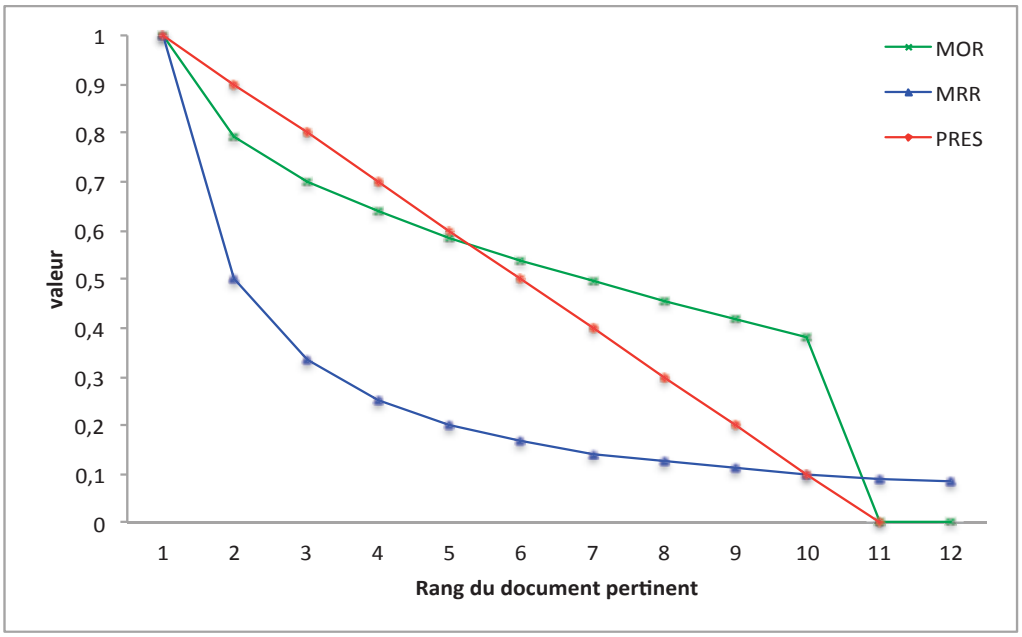

Figure 7. La mesure MOR comparée à MRR et $PRES$ dans le cas d'un seul document pertinent pour $N_{max} = 10$

voir comment l'utilisation du rappel à N_{max} (la valeur h) dans l'équation de MOR donnera la valeur 0 à la mesure MOR si le document pertinent a été trouvé après N_{max}. Ce comportement est cohérent avec le fait que l'utilisateur ne va jamais voir les documents aux rangs supérieurs à N_{max}, c'est aussi la même idée avec $PRES$ pour ce pire scénario. Comme nous pouvons le constater dans la figure 7, MOR pénalisera moins les systèmes qui trouvent le document pertinent vers la fin de la liste. Notre mesure est conçue pour un contexte où le rappel à N_{max} a le plus d'importance, alors que dans le cas d'un seul document pertinent le rappel n'est pas gradué (1 ou 0), ce qui signifie que MOR n'est pas adapté à ce cas. L'utilisateur, dans ce contexte, arrêtera la recherche après avoir trouvé le document pertinent, cela correspond plus à une mesure de comportement linéaire, dans ce cas comme $PRES$.

6. Analyse expérimentale

L'évaluation d'une mesure d'évaluation n'est pas simple. Logiquement, il faut évaluer à quel point la mesure satisfait les besoins pour lesquels elle a été construite. Dans notre cas, ces besoins ont été formalisés par des contraintes lors de la génération de la mesure, ce qui nous permet de garantir que MOR satisfait bien ces besoins, c'est-à-dire que pour évaluer un système de recherche d'information, MOR donnera la priorité au rappel, puis au positionnement des documents pertinents dans la liste de résultats (w et AP).

Néanmoins, il est intéressant de statistiquement étudier le comportement de la mesure sur un ensemble de cas. Ce que nous faisons dans cette section est de voir

la ressemblance entre MOR et le rappel. Une ressemblance très forte signifie que la mesure n'est pas intéressante car elle n'apportera pas une valeur ajoutée par rapport au rappel. Un écart très important entre les comportements des deux mesures signifie que MOR n'est pas une mesure dirigée rappel. Pour cela, nous vérifions que MOR soit plus proche du rappel que du MAP sans qu'elle soit une duplication du rappel.

6.1. Description de l'expérience

Nous nous intéressons aux comparaisons du MOR avec le rappel, la MAP et $PRES$. Nous réalisons deux expériences sur un ensemble de 88 runs [7] du track médical de TREC2012. La première expérience est de calculer la corrélation entre MOR et les autres mesures en utilisant le tau de Kendall (1938), et la deuxième est de comparer les décisions de ces mesures par rapport au run le plus performant et le plus mauvais.

6.2. Corrélation entre MOR et les autres mesures

Le tau de Kendall permet de mesurer le degré d'accord entre deux classements, où une valeur égale à 1 signifie un accord parfait. Pour cela, pour mesurer la corrélation entre deux mesures, nous calculons le tau de Kendall sur les deux listes ordonnées des systèmes que ces deux mesures ont évaluées. Les résultats de ces comparaisons sont dans le tableau 6. Nous constatons de ce tableau une valeur cohérente avec l'objectif

Tableau 6. La corrélation selon le tau de Kendall pour les différentes paires de mesures

Mesures	τ
$MOR \leftrightarrow Rappel$	$0,619$
$MOR \leftrightarrow PRES$	$0,205$
$MOR \leftrightarrow MAP$	$0,093$
$PRES \leftrightarrow Rappel$	$0,160$
$PRES \leftrightarrow MAP$	$0,075$
$Rappel \leftrightarrow MAP$	$0,123$

de la mesure MOR. Sans être complètement identique au rappel, MOR reste plus proche du rappel que de la précision. Cela indique que l'utilisation de w et AP a un effet important sur la mesure sans la biaiser vers la précision. Ce qui est intéressant dans le tableau 6 est que la correlation de la mesure Rappel avec MOR ($\tau = 0,619$) est plus élevée qu'avec $PRES$ ($\tau = 0,160$).

7. Un *run* dans la campagne d'évaluation TREC est la liste de résultats obtenue par un système pour un ensemble de requêtes. Les informations importantes dans cette liste sont, pour chaque requête, les documents pertinents et leurs scores de pertinence selon ce système.

6.3. Effet sur le classement

Nous présentons dans le tableau 7 les avis des trois mesures (MOR, rappel, MAP et $PRES$) sur le meilleur et pire systèmes parmi les 88 runs évalués. Nous constatons

Tableau 7. Le meilleur et le pire système selon les mesures d'évaluation

Mesure	meilleur (score)	pire (score)
MOR	$83(0,819)$	$49(0,018)$
$Rappel$	$83(0,826)$	$49(0,012)$
$PRES$	$84(0,734)$	$49(0,011)$
MAP	$30(0,461)$	$49(0,002)$

que toutes les mesures ont choisi le même système comme le moins performant, mais ils ne sont pas d'accord sur le système qui a la meilleure performance. Là aussi, on voit la cohérence entre le rappel et MOR sur les décisions des meilleur et pire systèmes. Alors que ce tableau indique que MOR et le rappel ont les mêmes préférences concernant les meilleurs et les pires systèmes, l'expérience précédente confirme que les choix des deux mesures ne sont pas toujours identiques.

7. Conclusion et perspectives

Nous avons présenté dans cet article la mesure d'évaluation MOR qui est adaptée à l'évaluation de systèmes de recherche d'informations dans un contexte de rappel. En utilisant des contraintes formelles inspirées par les besoins de ce contexte, nous avons construit la fonction finale étape par étape. Par conséquent, la mesure proposée correspond entièrement à ces besoins. Bien qu'il soit difficile de mesurer l'efficacité d'une mesure d'évaluation, nous avons présenté deux expériences qui permettent de confirmer que notre mesure correspond à l'objectif d'être orientée rappel sans pour autant être identique au rappel.

Nous pensons que *MOR* peut améliorer la qualité d'évaluation des tâches orientées rappel dans les campagnes d'évaluation. Pour confirmer cette idée, il est intéressant de comparer les jugements de la mesure avec les jugements humains dans un contexte rappel. On peut, par exemple, présenter les résultats de recherche de plusieurs systèmes en indiquant aux utilisateurs le placement des documents pertinents dans ces listes (pour éviter le biais de jugement de pertinence), et demander à ces utilisateurs de classer les systèmes par préférence. Les classements des utilisateurs pourront ainsi être utilisés dans nos deux expériences pour voir la relation entre les mesures automatiques et un classement des systèmes d'un point de vue humain.

Bibliographie

Cleverdon C. W., Mills J., Keen E. M. (1966). *Factors Determining the Performance of Indexing Systems.* (Aslib Cran éd.). Cranfield, College of Aeronautics.

Harman D. (1994). Overview of the second text retrieval conference (trec-2). In *Proceedings of the workshop on human language technology*, p. 351–357.

Järvelin K., Kekäläinen J. (2002). Cumulated gain-based evaluation of ir techniques. *Transactions on Information Systems (TOIS)*, vol. 20, p. 422–446.

Jones S. (1981). *The Cranfield tests*. London, Butterworths.

Kendall M. (1938). A new measure of rank correlation. *Biometrika*, vol. 30, p. 81–93.

Magdy W. (2012). *Toward Higher Effectiveness for Recall- Oriented Information Retrieval : A Patent Retrieval Case Study Walid Magdy*. Thèse de doctorat non publiée, Dublin City University.

Magdy W., Jones G. J. F. (2010). PRES : A Score Metric for Evaluating Recall-Oriented Information Retrieval Applications. In *Sigir*. ACM.

Rijsbergen V. (1979). *Information Retrieval*. Butterworths, Butterworth-Heinemann 1979.

Robertson S. (2006). On gmap: and other transformations. In *Proceedings of the 15th acm international conference on information and knowledge management*, p. 78–83. ACM.

Rocchio J. (1964). Performance indices for document retrieval systems. *Information storage and retrieval*.

Sakai T. (2007). On the reliability of information retrieval metrics based on graded relevance. *Information Processing & Management*, vol. 43, n° 2, p. 531–548.

Sanderson M. (2010). *Test collection based evaluation of information retrieval systems* (vol. 13). Now Publishers Inc.

Voorhees E. M., Tice D. M. (1999). The trec-8 question answering track evaluation. In *Trec*, vol. 1999, p. 82.

Webber W. E. (2010). *Measurement in Information Retrieval Evaluation*. Thèse de doctorat non publiée, University of Melbourne.

Zobel J., Moffat A., Park L. A. (2009). Against recall: is it persistence, cardinality, density, coverage, or totality? In *Acm sigir forum*, vol. 43, p. 3–8.