Analyse des critères d'évaluation des systèmes de recherche d'information

Alain Baccini * — Sébastien Déjean *— Désiré Kompaoré ** — Josiane Mothe**

* *Institut de Mathématiques de Toulouse, UMR 5219 - Université de Toulouse et CNRS*

{baccini/sebastien.dejean }@math.univ-toulouse.fr

** *Institut de Recherche en Informatique de Toulouse, UMR 5505 - Institut Universitaire de Formation des Maîtres - Université de Toulouse et CNRS,*

{kompaore/mothe}@irit.fr

Université Paul Sabatier
118, route de Narbonne, F-31062 Toulouse cedex 9

RÉSUMÉ. *Le modèle d'évaluation utilisé en recherche d'information implique une collection de documents sur laquelle les recherches sont effectuées, un ensemble de requêtes de test et la liste des documents de la collection pertinents pour chacune des requêtes. Ce modèle inclut également des mesures d'évaluation permettant de contrôler l'impact, sur la performance de la recherche, de la modification de certains paramètres d'un système. Trec-eval permet de calculer un grand nombre de mesures, certaines étant plus couramment utilisées comme la précision moyenne ou les courbes de rappel-précision. Le choix de l'ensemble minimal des mesures de performance nécessaires pour comparer deux systèmes a motivé l'étude qui est présentée dans cet article. Nous étudions la corrélation de 27 mesures parmi les plus utilisées dans la littérature. Nous montrons que 7 mesures sont suffisantes pour représenter les 27 mesures étudiées : ircl_prn.80, MAP, ircl_pr.20, recip_rank, P15, précision exacte et Rappel exact.*

ABSTRACT. *Evaluating information retrieval implies a document collection on which search is carried out, a set of test queries and the lists of the relevant documents for each query. This evaluation framework also includes evaluation measures making it possible to control the impact of search parameters on the performance. Trec-eval calculates a large number of measures, some being used more widely, like the mean average precision or recall-precision curves. The aim of this paper is to choose the minimal set of measures necessary to compare different information retrieval systems. In this paper, we present the study we carried out on relationships between 27 measures among the most used in the literature. We show that a set of 7 measures is enough to represent 27 studied measures: ircl_prn.80, MAP, ircl_pr.20, recip_rank, P15, exact_precision, and exact_recall.*

MOTS-CLÉS : recherche d'information, mesures de performance, évaluation, analyse statistique de données.

KEYWORDS: information retrieval, evaluation measures, evaluation, statistical data analysis.

1. Introduction

L'évaluation des systèmes de recherche d'information (SRI) est au cœur des problématiques de cette discipline. Le modèle d'évaluation généralement utilisé aujourd'hui est basé sur celui développé dans le projet Cranfield (Cleverdon *et al.*, 1966). Ce modèle implique une collection de documents sur laquelle les recherches sont effectuées, un ensemble de requêtes de test et la liste des documents de la collection pertinents pour chacune des requêtes. Ce modèle inclut également des mesures de performance associées au silence et au bruit documentaire, notions bien connues des documentalistes : le rappel et la précision (voir les définitions dans le tableau 3). Ce modèle permet de contrôler l'impact, sur la performance de la recherche, de la modification de certains paramètres d'un système, avec un coût raisonnable par rapport à une évaluation impliquant des utilisateurs à chaque étape (Robertson, 1981), (Voorhees, 2002).

Même si certains cadres d'évaluation sont proposés pour mieux prendre en compte l'utilisateur et l'acquisition d'information (Hersh *et al.*, 1994), (Järvelin & Kekäläinen, 2000), il n'en demeure pas moins que le modèle Cranfield reste l'approche dominante lorsqu'il s'agit d'évaluer les SRI (Borlund, 2003). Il faut toutefois noter qu'au cours des années, le modèle s'est affiné. Un premier aspect concerne la confiance que l'on peut avoir dans la comparaison des performances obtenues par différents systèmes via les tests statistiques (Hull, 1993). Un second aspect, qui nous intéresse plus particulièrement dans cet article, concerne le nombre de mesures d'évaluation maintenant utilisées. Le programme d'évaluation trec_eval, utilisé en particulier dans le programme TREC (Text REtrieval Conference : trec.nist.gov), permet de calculer un ensemble de 135 mesures. Certaines ont été introduite récemment comme par exemple la mesure Bpref (Buckley & Voorhees, 2004), comme pour compléter un ensemble de mesures déjà conséquentes. Il faut également noter que certaines sont plus couramment utilisées lorsqu'il s'agit de choisir des mesures de comparaison entre plusieurs SRI. Parmi ces dernières, nous pouvons citer la moyenne de la précision moyenne [*Mean Average Precision*] et les courbes de rappel-précision utilisées pour les comparaisons globales (Voorhees, 2007). En effet, il est difficile d'imaginer de comparer deux systèmes, ou bien l'impact de la variation d'un paramètre d'un même système, sur l'ensemble des mesures existantes. Il est donc important de choisir l'ensemble minimal des mesures nécessaires pour une comparaison complète de deux systèmes. Ce choix crucial a motivé les travaux que nous avons menés sur un ensemble de 27 mesures parmi les plus utilisées dans la littérature et que trec_eval permet de considérer. Dans cet article, nous nous intéressons donc à l'étude des corrélations existant entre les mesures utilisées dans l'évaluation des SRI. Dans un contexte plus large, nous souhaitons dégager les mesures les plus pertinentes pour permettre la comparaison de différents systèmes ou de différentes versions d'un même système.

Les travaux de la littérature relatifs à cet aspect sont présentés dans la section 2 de cet article. Nous présentons les données utilisées dans ce travail dans la section

3 : il s'agit des résultats obtenus par les participants au programme d'évaluation TREC (tâche recherche adhoc) par rapport à un ensemble de mesures d'évaluation. Nous étudions quel ensemble minimal de mesures il est judicieux de choisir lorsque l'on souhaite comparer des systèmes. Cette analyse repose sur des méthodes statistiques présentées dans la section 4. Les résultats de l'analyse sont présentés dans la section 5 et y sont discutés. Finalement, la section 6 présente une synthèse des résultats et des discussions et la section 7 conclut nos travaux et en donne quelques perspectives.

2. Travaux connexes

Peu de travaux ont été publiés concernant la corrélation des mesures d'évaluation en RI.

(Tague-Sutcliffe & Blustein, 1995) ont montré que les mesures R-précision et précision moyenne sont très fortement corrélées. L'étude, qui portait sur TREC3, a été confirmée depuis, comme dans (Voorhees & Harman, 1999) par exemple. (Aslam *et al.*, 2005) approfondissent l'étude de ces deux mesures et montrent, au moyen de l'analyse des données de TREC8, que cette forte corrélation est probablement liée au fait que les deux mesures approximent géométriquement la surface qui est située en dessous de la courbe rappel précision. (Buckley et Voorhees, 2005) ont mesuré la corrélation qui existe entre 7 mesures via la mesure de Kendall sur les données de TREC7. Ils ont montré qu'il y avait une corrélation d'au moins 0.6 entre chaque paire de mesure et que la corrélation la plus forte est celle entre la R-Précision et la MAP.

(Ishioka, 2003) analyse les relations qui existent entre les mesures d'évaluation que sont la mesure F, le point d'équilibre (point auquel le rappel est égal à la précision) et la précision aux 11 points de rappel, en se basant sur des tables de contingence 2x2. De son côté, (Egghe, 2008) étudie la corrélation entre la précision, le rappel, la mesure F (qui mesure à quelle vitesse la précision diminue lorsque le rappel augmente) et le silence. Tous deux montrent que l'évolution de la précision en fonction du rappel suit une fonction décroissante concave alors que l'évolution du rappel en fonction du *fallout* (pourcentage de documents non pertinents retrouvés ; ainsi, fallout + précision =1) est une fonction croissante concave.

(Sakai, 2007) s'intéresse à la comparaison de mesures basées sur la pertinence booléenne des documents et de mesures basées sur une pertinence graduée ; il montre que le rappel basé sur une pertinence graduée est fortement corrélé avec la précision moyenne. Enfin, (Melucci, 2007) étudie plus spécifiquement le cas de la corrélation entre les listes ordonnées de documents [*ranking*].

Les études de la littérature s'intéressent *a priori* à un ensemble très réduit de mesures. Face à ce constat, nous avons opté pour une approche plus globale visant à

sélectionner, après étude des redondances, un ensemble minimal permettant de comparer deux SRI.

3. Données

3.1. Exécutions issues de TREC

TREC est un programme d'évaluation qui, depuis TREC3 en 1994, comprend un certain nombre de tâches, dont la tâche *adhoc*. Cette tâche correspond à une recherche classique d'un utilisateur qui soumet une requête au système et attend en réponse un ensemble de documents pertinents par rapport à cette requête. Nous nous sommes intéressés plus spécifiquement à cette tâche qui a été conduite pendant 7 années consécutives. Nous avons retenu 5 années, préférant écarter les deux premières années qui correspondaient à la mise en place de la campagne TREC.

Chaque année, 50 besoins d'information [*topics*] sont proposés aux participants à la tâche. Chacun d'eux soumet une ou plusieurs exécutions. Une exécution [*run*] correspond à la liste des documents retrouvés pour chacun des 50 besoins pour un système et ses paramètres associés.

Ainsi, un même participant concourt généralement avec un seul système dont il modifie les paramètres pour soumettre plusieurs exécutions.

Le tableau 1 indique pour chaque collection TREC utilisée (tâche *adhoc* de TREC3 à TREC7) les identifiants des besoins d'information et le nombre d'exécutions soumises officiellement (*#E*).

Besoins	151-200	201-250	251-300	301-350	351-400
#E	40	34	80	79	103

Tableau 1. *Caractéristiques des collections TREC.*

Ainsi, lorsque les 5 années sont considérées, les 336 exécutions (40+34+80+79+103) pour 50 requêtes correspondent à un maximum de 16 800 ensembles de valeurs de mesure de performance (en réalité, il y en a 336 supplémentaires car nous avons également considéré, pour chaque exécution, les mesures moyennes sur l'ensemble des 50 requêtes). En fait, nous verrons par la suite que nous ne disposons que de 15 772 ensembles, la différence s'expliquant par des données manquantes.

3.2. Evaluation

Les fondements de l'évaluation en RI ont été développés dans le cadre du projet Cranfield (Cleverdon *et al.*, 1966) . Une collection de tests comprend un ensemble

de documents, un ensemble de requêtes et la liste des documents de la collection pertinents pour chaque requête. L'évaluation repose alors sur la comparaison de la liste de documents retrouvés par un système et celle des documents pertinents. Cleverdon propose deux mesures permettant d'évaluer l'efficacité d'une recherche : le rappel et la précision. Notons que la collection Cranfield 2 comprend seulement 1400 documents ; cette taille permet de constituer manuellement la liste exhaustive des documents pertinents pour chaque requête. Dans les programmes d'évaluation tels que TREC ou NTCIR[1], les collections de documents contiennent près d'un million de documents, ce qui rend impossible le jugement exhaustif de pertinence. Ainsi, dans le cas de grandes collections, les jugements de pertinence sont effectués à partir de l'ensemble des documents retrouvés par au moins un des systèmes participant ; ainsi certains documents de la collection sont non jugés. Les mesures d'évaluation se sont enrichies au fil des ans, la plupart se basant sur l'hypothèse que les documents non jugés sont non pertinents (Sakai & Kando, 2008).

Dans le cas de TREC, les valeurs des mesures de performance des systèmes sont calculées par le programme trec_eval[2] (Buckley, 1991). Ce programme calcule 135 mesures dans la version 8.1 que nous avons utilisée. Parmi ces mesures, nous avons retenu les 27 les plus communément utilisées dans la littérature. Pour chaque mesure, le résultat de chaque requête d'une exécution est évalué individuellement, la moyenne (équipondérée sur l'ensemble des requêtes) permettant d'obtenir des mesures globales pour chaque exécution. Ce sont ces mesures moyennes que nous avons analysées dans cet article. Notons que les documents pertinents sont supposés connus, les documents non jugés étant considérés comme non pertinents et diminuant potentiellement le rappel et la précision.

Le tableau 2 indique, pour chaque mesure, son nom dans les graphiques présentés plus loin, ce qu'elle représente et sa définition. Les paramètres de ces mesures ont été définis par les programmes internationaux d'évaluation et sont utilisés à l'identique par l'ensemble des campagnes.

Mesure	Description	Définition
rappel	Rappel exact par rapport à l'ensemble des documents retrouvés. Le rappel mesure la capacité du système à restituer l'ensemble des documents pertinents (en lien avec le	$R = \frac{Nb\ de\ doc\ pertinents\ retrouvés}{Nb\ de\ doc\ pertinents}$ Nombre total de documents retrouvés limité aux 1000 premiers (valeur fixée par les programmes d'évaluation

[1] http://ntcir.nii.ac.jp/

[2] http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/trec_eval_video/README version du 24 juillet 2006, consultée en décembre 2007.

	silence documentaire).	utilisés).
precision	Précision moyenne non interpolée par rapport à l'ensemble des documents pertinents. Mesure la capacité du système à ne restituer que des documents pertinents (en lien avec le bruit documentaire). Comme les listes de résultats qui sont évaluées sont limitées à 1000 documents, précision exacte et P1000 (voir plus bas) sont proches.	$P = \frac{Nb\ de\ doc\ pertinents\ retrouvés}{Nb\ de\ doc\ retrouvés}$ Nombre total de documents retrouvés, limité aux 1000 premiers (valeur fixée par les programmes d'évaluation utilisés).
ircl_prn.xx par exemple ircl_prn.0.00 pour [*Interpolated Recall - Precision Averages*] au point de rappel 0.00	Mesures de précision à 11 différents niveaux de rappel. La précision et le rappel ayant tendance à varier en sens inverse, il s'agit là d'un moyen de combiner les deux mesures. L'ensemble des 11 mesures est utilisé pour créer des courbes rappel/ précision.	$P(R_j) = \max_{R_j < R \leq R_{j+1}} P(R)$ P(R) correspond à la précision au point de rappel R. Rj correspond au point de rappel Rj avec j ∈ {0, 0.1, ….}. Les niveaux de rappel retenus sont 0, 0.10, …0.9 ; 1. Le niveau de rappel 0.1 correspond au point théorique où 10% des documents pertinents ont été retrouvés. Le point de rappel 0.10 n'existe pas forcément pour certaines listes de documents retrouvés. Les mesures sont donc interpolées. L'interpolation consiste à considérer que la précision au point de rappel 0.10 correspond au maximum de la précision pour tous les points de rappel ≥ 0.10.
Pn	Précision à différents niveaux de coupe. Elle permet en particulier de s'intéresser à la haute précision, lorsque peu de documents sont	Pn = Précision après que n documents aient été retrouvés (qu'ils soient pertinents ou non). avec n ∈ {5, 10, 15, 20, 30, 100, 200, 500, 1000}.

	restitués (ceci est à rapprocher de l'intérêt des internautes à la seule première page de résultats par exemple). Cette mesure peut être qualifiée de locale puisqu'elle s'intéresse à des points spécifiques. Cette mesure ne prend pas en compte le fait que certaines requêtes comportent peu de documents pertinents (voir R-Précision) ; la précision diminuant alors plus vite que n augmente.	Si moins de n documents sont retrouvés, les documents manquants sont considérés comme non pertinents.
R-Précision	Variante de la Pn où n est le nombre de documents pertinents.	Précision après que R documents ont été retrouvés, où R est le nombre de documents pertinents pour la requête considérée. Cette mesure a été introduite dans TREC2 pour limiter l'influence du nombre de documents pertinents : ce nombre varie en fonction des requêtes.
recip_rank	Correspond à une mesure alternative de calcul de la précision. Si l'utilisateur analyse la liste des documents, cette mesure permet d'évaluer le nombre de documents qu'il faut considérer avant de retrouver le premier document pertinent.	Rang du premier document pertinent.
Bpref	Préférence binaire. Contrairement aux autres mesures, bpref se focalise sur les documents réellement jugés.	Nombre de fois que des documents jugés non pertinents sont retrouvés avant un document pertinent. Cette mesure a été introduite en 2004 afin de réduire l'effet du jugement de pertinence qui n'est réalisé que sur certains documents.

F-measure	Mesure qui combine le rappel et la précision. En effet, rappel et précision ayant tendance à varier en sens inverse, il s'agit d'une mesure alternative aux mesures ircl_prn.xx relativement aux taux de rappel et précision exacts.	Mesure F = 2xRP/(R+P) où R (resp. P.) est le rappel (resp. précision) exact (resp. exacte).
MAP [*mean average precision*]	Précision moyenne. Cette mesure peut être qualifiée de globale puisqu'elle combine différents points de mesure. Elle est moins sensible au nombre de documents pertinents que les mesures Pn.	Moyenne des précisions obtenues chaque fois qu'un document pertinent est retrouvé. Cette mesure a été introduite dans TREC2 pour sa capacité à résumer les mesures de précision aux 11 points de rappel.

Tableau 2. *Présentation des 27 mesures utilisées.*

4. Outils statistiques

4.1. Démarche

Le point d'entrée des méthodes d'analyse statistique correspond généralement à une matrice de données. D'un point de vue statistique, nous disposons d'un tableau comportant en lignes l'ensemble des exécutions considérées (les *runs*) et en colonnes les différentes mesures de performance retenues. En l'occurrence, le tableau analysé ici comporte 15772 lignes (les exécutions définies en 3.1) et 27 colonnes (les mesures de performance définies en 3.2).

L'objet est d'analyser les colonnes (les mesures de performance) afin d'étudier leur redondance. Pour cela, nous avons mis en œuvre trois méthodes statistiques exploratoires :

- analyse de la matrice des corrélations, méthode la plus simple pour appréhender la structure générale des mesures de performance ;
- analyse en composantes principales, méthode qui, à partir des mêmes fondements, fournit des indicateurs et des graphiques détaillant mieux la structure des corrélations entre colonnes ;
- classification des mesures de performance, dont l'objet est de définir un petit nombre de classes, chacune regroupant des mesures redondantes.

4.2. Matrice des corrélations

Il s'agit de la matrice carrée d'ordre 27, symétrique, donnant au croisement de chaque ligne et de chaque colonne le coefficient de corrélation linéaire entre la mesure de performance figurant dans cette ligne et celle figurant dans cette colonne. Ce coefficient est encore appelé coefficient de Pearson (en particulier dans les logiciels anglo-saxons) ou, parfois, coefficient de Bravais-Pearson.

Rappelons que, si l'on note Yj et Yk les deux mesures de performance (les deux variables) considérées, leur coefficient de corrélation s'écrit sous la forme :

$$r_{jk} = \frac{c_{jk}}{s_j s_k}$$

où c_{jk} désigne la covariance des deux variables et s_j et s_k désignent leurs écarts-types respectifs.

Ce coefficient est nécessairement compris entre -1 et +1, son signe indiquant le sens de la corrélation (en moyenne, les deux variables varient dans le même sens lorsque le signe est positif et en sens opposés lorsque le signe est négatif) et sa valeur absolue l'intensité (l'importance) de cette corrélation : plus cette valeur absolue est proche de 1, plus la corrélation est forte.

Malheureusement, dès que le nombre de variables est un peu élevé (de l'ordre de quelques dizaines), la matrice des corrélations est assez délicate à lire et à interpréter. C'est pour cette raison que l'on trouve de plus en plus souvent, aujourd'hui, ces matrices présentées sous forme graphique comme nous le verrons au paragraphe 5.1.

4.3. Analyse en Composantes Principales (ACP)

Méthode très utilisée en statistique multidimensionnelle, l'ACP permet d'obtenir une représentation graphique du nuage des points associé à un tableau de données, et cela dans un espace de dimension réduite. La méthode est conçue de telle sorte que l'information expliquée (c'est-à-dire l'inertie, généralisation multidimensionnelle de la variance) dans cet espace de dimension réduite soit la plus grande possible. Cette méthode est basée sur la recherche des axes principaux d'un nuage de points : voir, par exemple, (Lebart *et al.*, 2006).

De façon générale, le tableau de départ est une matrice de dimension n x p, où n est le nombre d'individus (ou d'unités observées, ici les exécutions) et p le nombre de variables (ici les mesures de performance). Les données sont relatives à des variables quantitatives, homogènes ou non. À partir de cette matrice, des représentations graphiques des individus peuvent être réalisées relativement aux variables prises deux-à-deux. Ces représentations sont très lourdes (il y en a p(p-1)/2) et ne donnent qu'une vue partielle des corrélations entre variables. Au contraire, l'ACP permet de prendre en compte la globalité de ces corrélations. Il

s'agit de déterminer les axes principaux d'inertie, c'est-à-dire les axes maximisant l'inertie du nuage des points projetés. Ces nouveaux axes correspondent à des combinaisons linéaires des variables initiales et sont déterminés par la diagonalisation (recherche des valeurs propres et des vecteurs propres) soit de la matrice des variances-covariances lorsque les variables sont homogènes, soit de la matrice des corrélations lorsqu'elles sont hétérogènes. La représentation graphique des individus (les lignes de la matrice initiale) permet d'observer la forme du nuage de points et d'apprécier les distances entre deux points, donc entre les deux entités qu'ils représentent. De même, une représentation graphique des variables (les colonnes de la matrice initiale) permet de visualiser les corrélations entre ces variables. Une représentation selon les deux ou trois premiers axes principaux suffit, en général, à visualiser l'essentiel de l'information, notamment des corrélations entre variables.

4.4. Classification

Deux sortes de méthodes de classification sont couramment utilisées en statistique : la classification ascendante hiérarchique (CAH) et les algorithmes d'agrégation autour de centres mobiles (nuées dynamiques, k-means...). Lorsque cela est possible (nombre d'objets à classer ne dépassant pas quelques milliers), il est en général recommandé d'enchaîner les deux méthodes en commençant par la CAH (Lebart *et al.*, 2006). La classification peut être réalisée indifféremment sur les lignes ou sur les colonnes de la matrice initiale. Dans certaines applications, il peut même être intéressant de réaliser les deux classifications.

Dans le contexte qui est le nôtre, ce sont les mesures de performance que nous avons classées, autrement dit les colonnes de la matrice des données.

L'idée de la CAH est la suivante : au départ de l'algorithme, chaque variable considérée constitue une classe à elle toute seule (on dispose donc de p classes). À partir des individus observés sur l'ensemble des variables, il est possible de définir une distance entre tout couple de variables (par exemple la distance euclidienne classique de R^n). On commence alors par regrouper les deux variables les plus proches au sens de cette distance : on ne dispose plus ainsi que de p-1 classes. On continue ensuite, de proche en proche, jusqu'à l'obtention d'une classe unique. Cela nécessite toutefois de définir au préalable la distance entre deux classes, ce qui peut se faire de différentes façons, par exemple la distance entre les points moyens de ces classes. On notera toutefois qu'il est souvent préconisé, par les statisticiens, d'utiliser la méthode dite de Ward, qui utilise un critère lié à l'inertie et qui est donc cohérent avec l'ACP. La classification ainsi réalisée génère un arbre que l'on peut couper à différents niveaux pour obtenir des classes plus ou moins nombreuses.

Une fois le nombre de classes choisi, il est courant de mettre en œuvre un algorithme de type agrégation autour de centres mobiles. Partant de k centres donnés a priori, une étape de l'algorithme affecte chaque individu au centre le plus proche et recalcule ensuite les nouveaux centres (par exemple en prenant le barycentre de

chaque classe). L'algorithme est terminé lorsqu'aucun individu ne change de classe entre deux étapes successives, ce qui se produit en général très vite (entre 2 et 5 itérations). Lorsque l'on enchaîne CAH et agrégation autour de centres mobiles, les centres initiaux de cette dernière sont les barycentres des classes obtenues par la CAH. Ceci a pour effet de stabiliser les résultats de la CAH en réaffectant éventuellement des points à la frontière entre deux classes.

5. Résultats

5.1. Matrice des corrélations

Afin de faciliter l'interprétation de la matrice des corrélations (comprenant 27 x 27 = 729 valeurs, dont 351 intrinsèques), nous l'avons représentée sous la forme d'une image pouvant coder les valeurs numériques entre –1 et 1 selon un nuancier de couleurs ou de niveaux de gris. Dans la figure 1, le nuancier de gris commence à la valeur minimale –0.08 (blanc) et termine à la valeur maximale 1 (noir). Nous avons de plus procédé à une réorganisation optimale de l'image produite afin de mettre plus clairement en évidence les zones de fortes corrélations (Caraux & Pinloche, 2005).

En complément de la figure 1, les principales caractéristiques de l'ensemble des corrélations sont données dans le tableau 3. On remarque la quasi absence de valeurs négatives (minimum -0.08 et 5 valeurs négatives sur 351 ; les valeurs autour de 0 correspondent au fait que ircl_prn.1 est non corrélée avec P100, P200, P500, P1000 et Fmesure), ce qui montre qu'il n'y a pas de mesures contradictoires. Par ailleurs, plus des trois quarts des coefficients de corrélation (zones noires, ou foncées) sont très significatifs car supérieurs à 0.8 (calcul effectué sur plus de 15000 observations). Cela signifie que toutes les mesures de performance évaluent globalement les systèmes dans le même sens : un « bon » système l'est quelle que soit la mesure de performance utilisée.

L'avantage de la réorganisation optimale de l'image de la matrice des corrélations est la mise en évidence très explicite de groupes de mesures. Chaque groupe se caractérise par une information très redondante traduite par des carrés noirs (corrélations proches de 1).

Notons une corrélation importante des mesures *ircl_prnxx* entre elles. Ce résultat était évidement attendu puisque, par exemple, ircl_prn0.4 est lié à *ircl_prn0.3* : les 40% des documents pertinents retrouvés contenant nécessairement les premiers 30%. On s'attend à ce que ces précisions diminuent continuellement au fur et à mesure que le pourcentage de documents pertinents retrouvés augmente. De la même façon, il n'est pas étonnant que les précisions à différents niveaux de coupe (*Pn*) soient fortement corrélées. Logiquement, les valeurs de précision diminuent progressivement lorsque n (nombre de documents définissant la coupe) augmente.

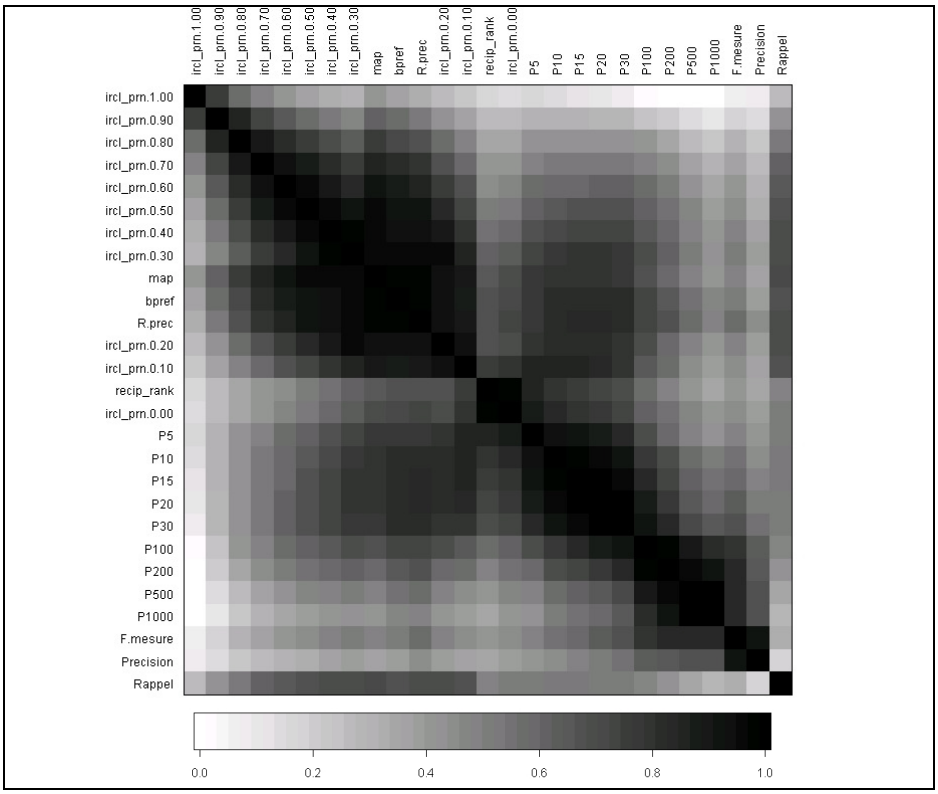

Figure 1. *Image réorganisée de la matrice des corrélations.*

Min.	Premier Quartile	Médiane	Moyenne	Troisième Quartile	Max.
-0,08	0,41	0,61	0,59	0,80	1

Tableau 3. *Caractéristiques numériques de l'ensemble des corrélations.*

Un autre aspect intéressant concerne les corrélations entre la *MAP* et ces deux ensembles de mesures. La *MAP* est plus fortement corrélée aux précisions lorsque 30 à 50% des documents pertinents sont retrouvés, bien moins lorsque plus de documents pertinents sont retrouvés. Ainsi lorsque 90% des documents sont retrouvés, la corrélation entre la *MAP* et la précision est seulement de 0.58 alors que lorsque 30% des documents sont retrouvés, elle est de 0.95. La figure 2 (a) montre bien que la *MAP* est redondante avec la précision aux niveaux de rappel de 0.2 à 0.6.

Ainsi, il est intéressant, lors d'une étude comparative de résultats, de présenter la *MAP* mais également la précision lorsqu'un faible pourcentage de documents pertinents est retrouvé et lorsqu'un fort pourcentage de documents est retrouvé. Cela reste vrai alors même que les valeurs de *MAP* varient de façon importante (valeur minimum 0, maximum 1, médiane 0.14, premier quartile 0.03, troisième quartile 0.31).

Les corrélations entre les *Pn* et la *MAP* sont assez élevées mais diminuent nettement à partir de n égal à 100 (figure 2 (b)).

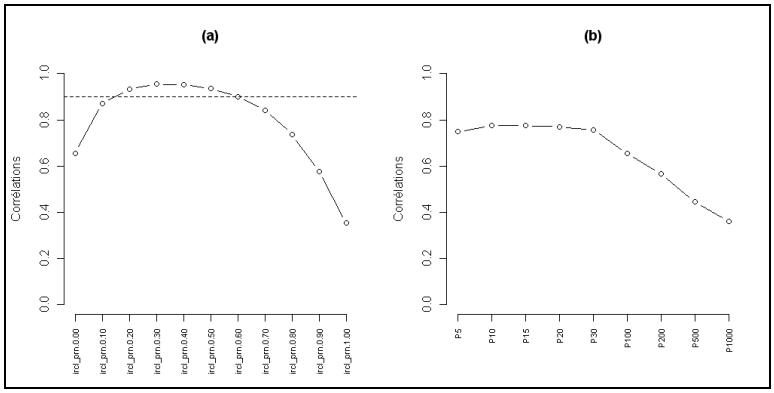

Figure 2. Corrélation entre *MAP* et les *ircl_prnxx* à gauche (la ligne horizontale indique une corrélation de 0.9) ; entre *MAP* et les *Pn* à droite.

À noter également le caractère particulier de l'indice *Rappel* qui constitue un groupe à lui tout seul car ne présentant de corrélation très élevée avec aucune autre mesure.

5.2. Analyse en Composantes Principales

L'ACP apporte un autre point de vue sur les relations entre les indices de performance et confirme la redondance d'information. Bien qu'elle permette de représenter conjointement variables et individus d'un tableau de données, nous nous focalisons ici sur la représentation des variables (figure 3).

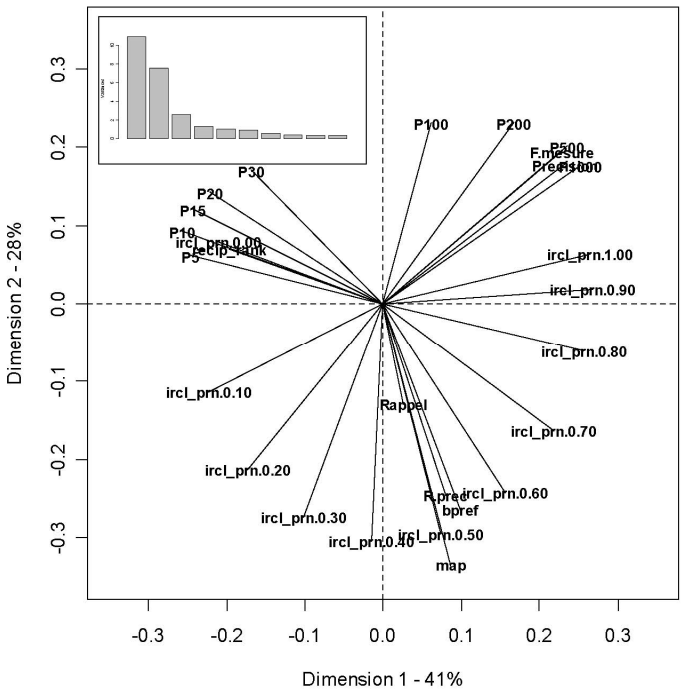

Figure 3. *Représentation des variables sur le premier plan principal et éboulis des valeurs propres (coin en haut, à gauche).*

Dans la figure 3, l'éboulis des valeurs propres, traduisant la part d'inertie expliquée par chaque composante principale (41% et 28% pour les deux premières), justifie une représentation en deux dimensions. Cependant, un examen des dimensions supérieures (dont le graphique n'est pas présenté ici) permet de mettre en évidence les variables *ircl_prn.0.00* et *recip_rank* sur l'axe 3 et la variable *Rappel* sur l'axe 4. Lors de l'interprétation, la représentation sous forme de traits permet de mettre en évidence les corrélations entre les variables, le cosinus de l'angle entre deux traits étant assimilable à une corrélation entre les variables qu'elles représentent.

On note sur l'axe 1 l'opposition entre, à droite, *ircl_prn.0.80*, *ircl_prn.0.90* et *ircl_prn.1.00* et, à gauche, *P5*, *P10*, *P15* et *P20*, deux ensembles de variables dont on peut noter la faible corrélation sur la figure 1 (valeurs inférieures à 0.4). Leur présence sur l'axe 1 montre leur meilleure capacité à dissocier les exécutions. Ainsi, il serait plus intéressant de présenter les comparaisons des résultats de différents systèmes sur la base, par exemple, de leur haute précision plutôt que via la *MAP*.

Sur l'axe 2 s'opposent, en bas, *MAP, ircl_prn.0.30*, *ircl_prn.0.40* et *ircl_prn.0.50* et, en haut, *P100* et, dans une moindre mesure, *P200*. Dans ce cas, les deux ensembles de variables sont un peu plus corrélés que précédemment tout en restant relativement faible (de l'ordre de 0.65).

Le premier plan permet de distinguer également des groupes d'indices de performance dont le caractère « progressif » de certains (les *ircl_prn.x.xx* et les *Pn*) est illustré par un faisceau ordonné. Cette structure très cohérente nous aidera à choisir des représentants des classes issues de la classification.

5.3. Classifications

Après la mise en œuvre de méthodes exploratoires mettant en évidence de nettes redondances entre mesures de performance, il est naturel d'envisager une classification pour limiter le nombre de mesures.

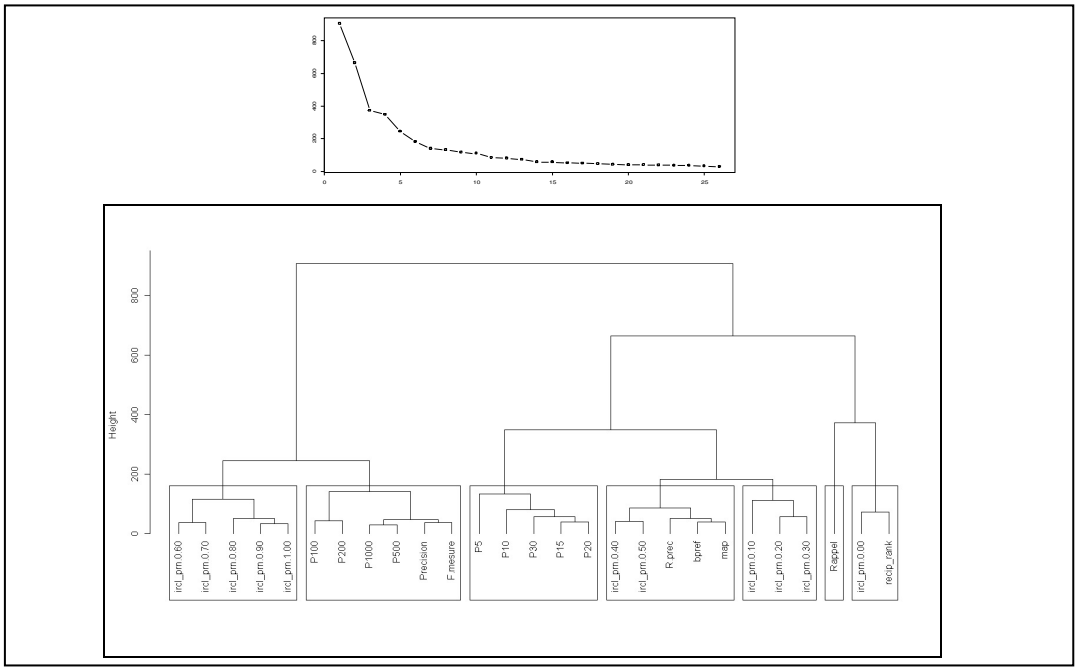

Figure 4. *Graphique de la hauteur des noeuds. Dendrogramme représentant la classification hiérarchique ascendante des indices de performance et mise en évidence de 7 groupes.*

Le choix du nombre de classes se fait à partir de l'arbre et du graphique associé de la hauteur des nœuds (figure 4). L'échelle verticale de l'arbre représentant la distance entre les groupes, un niveau de coupe pertinent est caractérisé par un écart important entre les hauteurs de deux nœuds successifs. Ceci se traduit, sur le graphique de la hauteur des nœuds, par un point pour lequel il y a une forte pente à gauche et une faible pente à droite. Dans ces conditions, selon le degré de finesse souhaité, on peut ici retenir 3, 7 ou 11 classes. Le choix de 7 classes nous a paru le plus approprié (ni trop grossier, ni trop fin).

La mise en œuvre postérieure d'un algorithme de type k-means entraîne deux changements par rapport à la classification hiérarchique de la figure 4 : cela concerne les variables *ircl_prn0.30* et *ircl_prn0.60*. Ces deux mesures basculent dans le groupe contenant *ircl_prn0.40*, *ircl_prn0.50*, *R.prec*, *bpref* et *MAP* (cf description des groupes ci-dessous). La position de ces deux variables sur le premier plan principal (figure 3) illustre leur caractère frontalier entre les groupes G4 et G6 pour *ircl_prn.0.30* et entre les groupes G4 et G7 pour *ircl_prn.0.60*.

Finalement, les 7 groupes retenus sont les suivants :

G1. *Rappel* ;
G2. {*Précision, F.mesure*, *P100*, *P200*, *P500*, *P1000*} *;*
G3. {*P5, P10, P15, P20, P30*} ;
G4. {*ircl_prn.0.30, ircl_prn.0.40, ircl_prn.0.50, ircl_prn.0.60, bpref, R.prec, MAP*} ;
G5. {*ircl_prn.0.00*, *recip_rank*} ;
G6. {*ircl_prn.0.10*, *ircl_prn.0.20*} ;
G7. {*ircl_prn.0.70*, *ircl_prn.0.80*, *ircl_prn.0.90*, *ircl_prn.1.00*}.

6. Synthèse

Nous proposons ici un retour sur les différentes vues de la matrice des corrélations (image et ACP) en les combinant avec l'information apportée par la classification. Ainsi dans la figure 5 à droite, les éléments ont été entourés en fonction de la classe à laquelle ils appartiennent. À gauche, ils ont été selon le même critère séparés par des lignes blanches.

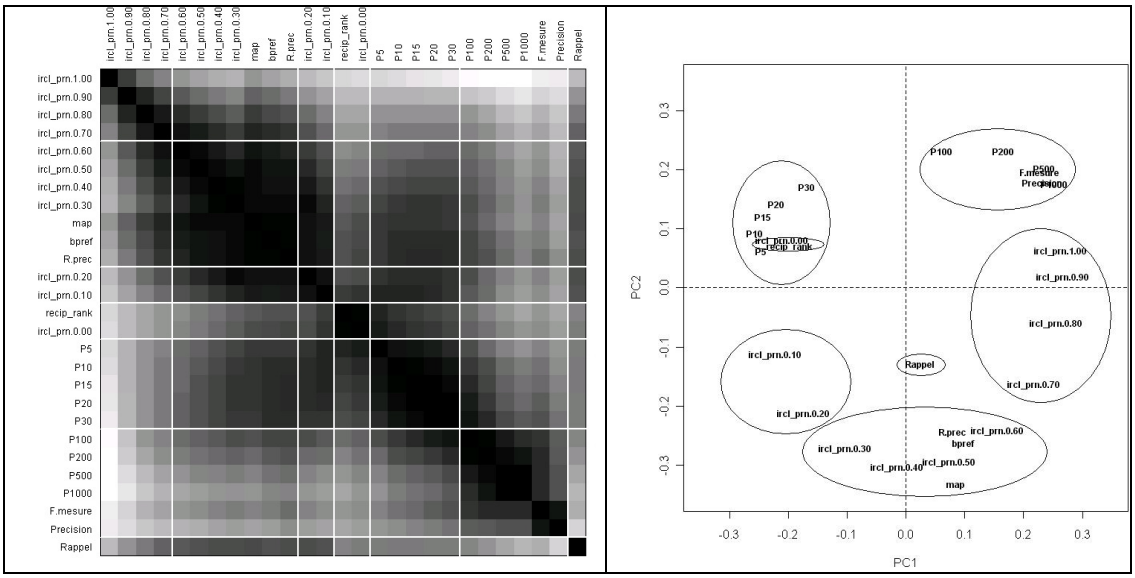

Figure 5. *Représentation des groupes issus de la classification (CAH + k-means) sur la matrice des corrélations réorganisée ainsi que sur le premier plan de l'ACP.*

La cohérence des sept groupes obtenus par classification apparaît très clairement sur les graphiques de la figure 5. Sur l'image réorganisée, les mesures appartenant au même groupe apparaissent côte à côte de façon systématique. Sur le premier plan de l'ACP, cinq groupes se caractérisent par des faisceaux très clairement distincts, les deux autres, {*ircl_prn.0.00*, *recip_rank*} et {*Rappel*}, ne sont pas très bien représentés sur ce plan et renvoient respectivement aux axes 3 et 4.

Ces sept groupes étant clairement identifiés, nous nous attachons ici à choisir un représentant unique pour chacun d'eux et commentons ces résultats d'un point de vue de ce que représentent les différents critères.

- ***ircl_prn.0.80*** : indicateur central (sur l'ACP) dans le groupe {*ircl_prn.0.70, ircl_prn.0.80, ircl_prn.0.90, ircl_prn.1.00*}. D'un point de vue de la RI, ce critère représente la précision lorsque la majorité des documents pertinents sont retrouvés (70 à 100%). Il s'agit donc d'une mesure qui permettra d'évaluer les systèmes lorsque la recherche doit être orientée vers un haut niveau de rappel. Cela est le cas dans le cadre de collecte d'information pour des activités de veille technologique par exemple.
- ***MAP*** : pour le groupe {*ircl_prn.0.30, ircl_prn.0.40, ircl_prn.0.50, ircl_prn.0.60, bpref, R.prec, map*}. Sa position centrale dans le groupe ainsi que sa plus grande variabilité (traduite par la longueur du vecteur en ACP) en font un bon représentant. D'un point de vue RI, la *MAP* est parmi les mesures les plus souvent utilisées pour rapporter les résultats d'un SRI. Elle a donc aussi, de ce point de vue là, un rôle central. Notons que l'obtention de ce groupe de mesures montre que la mesure *bpref* n'apporte pas d'éléments nouveaux pour l'évaluation des SRI, alors même que son objectif était de supprimer le biais lié au fait que les documents non jugés sont considérés comme non pertinents.
- ***ircl_prn.0.10*** ou ***ircl_prn.0.20*** : dans ce groupe, il n'est pas possible de trancher entre les deux sur des critères statistiques. Les deux mesures sont également sémantiquement très proches : il s'agit de la précision lorsque peu de documents pertinents sont retrouvés (rappel très faible). Nous retenons arbitrairement *ircl_prn.0.20* (comme « complément » de *ircl_prn.0.80* retenu plus haut). D'un point de vue de la RI, même si ces deux mesures forment un groupe particulier, l'intérêt d'utiliser ces mesures est moindre.
- ***ircl_prn.0.00*** ou ***recip_rank*** : comme précédemment, il n'est pas possible de trancher entre les deux sur des critères statistiques, y compris en regardant la dimension 3 de l'ACP. D'un point de vue RI, on préfèrera la mesure *recip_rank* (premier document pertinent) que le point de départ de la courbe rappel/précision.
- ***P15*** : nous retenons cette mesure pour sa position centrale dans le groupe {*P5, P10, P15, P20, P30*}. D'un point de vue de la RI, le choix alternatif de P10 aurait pu se justifier par le fait que certains moteurs de recherche sur le web affichent par défaut les 10 premiers résultats. Par ailleurs, il est intéressant de noter que l'ensemble des précisions à différents niveaux de coupe (Pn) peuvent être résumées à deux groupes (celui-ci et le suivant).

- ***Précision*** : en tant que représentant du groupe {*Précision, F.mesure*, *P100*, *P200*, *P500*, *P1000*}. C'est l'indice vers lequel les mesures intermédiaires (Pn) convergent et, par ailleurs, la définition du critère F.mesure comme moyenne harmonique de *Rappel* et *Précision* ne nous semble pas pertinente dans ce contexte.
- ***Rappel*** : le caractère particulier de cette mesure de performance a été mis en évidence de façon systématique par les trois méthodes.

7. Conclusions et perspectives

Dans cet article, nous nous sommes intéressés à l'étude approfondie des corrélations existant entre les mesures de performances utilisées pour évaluer les SRI, en particulier dans les campagnes internationales. Nous nous sommes focalisés sur les données issues de TREC *adhoc*, de TREC3 à TREC7. Nous avons ainsi obtenu un jeu de plus de 15 000 ensembles de mesures, ce qui, du point de vue statistique, permet de proposer des conclusions significatives. Nous n'avons retenu pour cette étude que 27 des 135 mesures utilisées dans le programme trec_eval (les plus courantes). Les différentes méthodes statistiques utilisées ont permis de montrer que la liste des 27 indices de performance considérés pouvait être réduite à un ensemble minimal de 7 mesures, sans perte d'information : *ircl_prn.80, map, ircl_pr.20, recip_rank, P15, précision exacte* et *rappel exact*.

Dans cette première étude, nous avons considéré les mesures qui, du point de vue du domaine de la recherche d'information, nous semblaient les plus importantes. Nos travaux actuels visent à élargir l'ensemble des mesures de performances choisies en considérant les 135 mesures disponibles dans le programme trec_eval.

Cette première étude s'est attachée à étudier l'ensemble de la tâche TREC *adhoc*. Les conclusions permettent donc de dire que, de façon générale, lorsqu'il s'agit de comparer des systèmes répondant à la tâche *adhoc*, 7 mesures suffisent à caractériser complètement les différences. Les collections utilisées chaque année varient. Il serait donc intéressant de savoir si les 7 classes sont identiques lorsque l'on se cantonne à une collection spécifique (une année de TREC *adhoc*).

De même, notre étude s'est limitée à une seule tâche. Nous souhaitons étudier l'effet de la tâche sur les corrélations entre mesures. Nous tenterons ainsi de répondre aux questions : les corrélations entre mesures sont-elles dépendantes de la tâche ? Sont-elles dépendantes de la collection utilisée ? Quels sont les groupes stables ?

Remerciements

Nous tenons à remercier le programme pluri-formations de Recherche en Mathématiques et Informatique de Toulouse (FREMIT) pour son soutien au projet ACRIC (Analyse Canonique et Recherche d'Information Contextuelle) dans le cadre duquel cette étude a été menée.

Références

Aslam, J.A., Yilmaz, E., Pavlu V., « A Geometric Interpretation of R-precision and its Correlation with Average Precision », *Proceedings of the international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, 2005, p. 664-671.

Borlund, P., « The IIR evaluation model: a framework for evaluation of interactive information retrieval systems », *Information Research*, vol. 8, n°3, paper no. 152. [Available at informationr.net/ir/8-3], Springer, 2003.

Buckley, C., « Trec_eval, [Available at trec.nist.gov/trec_eval], 1991.

Buckley C., Voorhees, E.M. « Retrieval system evaluation » dans *Voorhees E.M. et Harman D.K., « TREC : experiment and evaluation in information retrieval »*, Cambridge, Mass. MIT Press, 2005, pp. 53-75 .

Buckley, C., Voorhees, E. M., « Retrieval evaluation with incomplete information », *Proceedings of the international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, 2004, pp 25-32.

Caraux, G., Pinloche, S., « Permutmatrix: A Graphical Environment to Arrange Gene Expression Profiles in Optimal Linear Order », *Bioinformatics*, vol. 21, 2005, p. 1280-1281.

Cleverdon, C. W., Mills, J., Keen, E. M., « Factors determining the performance of indexing systems », *Cranfield, UK: Aslib Cranfield Research Project*, College of Aeronautics (Volume 1: Design; Volume 2: Results), 1966.

Egghe, L., « The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations », *Information Processing & Management*, vol. 44, n°2, 2008, p. 856-876.

Hersh, W.R., Elliot, D.L., Hickam, D.H., Wolf, S.L., Molnar, A., Leichtenstien, C., « Towards new measures of information retrieval evaluation », *Proceedings of the annual symposium on Computer Application in Medical Care*, 1994, p. 895-899.

Hull, D., « Using statistical testing in the evaluation of retrieval experiments », *Proceedings of the international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, 1993, p. 329-338.

Ishioka, T., « Evaluation of criteria for information retrieval », *Web Intelligence*, 2003. Proceedings, IEEE/WIC International Conference, 2003, p. 425-431.

Järvelin, K., Kekäläinen, J., « IR evaluation methods for retrieving highly relevant documents », *Proceedings of the international ACM SIGIR Conference on Research and Development of Information Retrieval*, ACM Press, 2000, p. 41-48.

Lebart, L., Piron, M., Morineau, A., « Statistiques exploratoire multidimensionnelle : Visualisations et inférences en fouille de données », 2006 , 4ème édition, Dunod.

Melucci, M., « On rank correlation in information retrieval evaluation », *ACM SIGIR Forum*, vol. 41, n°1, ACM Press, 2007, p. 18-33.

Robertson, S.E., « The methodology of information retrieval experiment », dans Sparck Jones, K. ed., *Information retrieval experiments*. London: Butterworths, 1981, p. 9-31.

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

Sakai, T. et Kando, N., « On information retrieval metrics designed for evaluation with incomplete relevance assessments », *Information Retrieval Journal*, Springer, vol. 11, 2008, p. 447–470.

Tague-Sutcliffe, J. et Blustein, J., « A statistical analysis of the TREC-3 data », Proceedings of the Third Text Retrieval Conference (TREC-3), NIST Special Publication, 1995, p. 385-398.

Voorhees, E.M. et Harman, D.K., « Overview of the seventh text retrieval conference (TREC-7) », *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, NIST Special Publication, 1999, p. 1-23.

Voorhees, E.M., « The Philosophy of Information Retrieval Evaluation », Lecture Notes in Computer Science, Volume 2406/2002, ISSN 0302-9743, Springer, Berlin / Heidelberg, 2002.

Voorhees, E.M., « Overview of the TREC 2006 », *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, NIST Special Publication:SP 500-272, 2007, p. 1-16.