Une méthode pour l'évaluation automatique de la "difficulté" d'une requête

Jens Grivolla

Laboratoire Informatique d'Avignon – Université d'Avignon / CNRS
339 ch. des Meinajariès, Agroparc BP 1228
84911 Avignon Cedex 9, France
Tél : +33 (0) 4 90 84 35 30 Fax : +33 (0) 4 90 84 35 01
jens.grivolla@univ-avignon.fr

RÉSUMÉ. *Dans des applications de recherche documentaire, il est souvent utile d'avoir une mesure de confiance dans l'ensemble de document trouvés, afin de pouvoir proposer un traitement spécifique (automatique ou interactif) des requêtes particulièrement difficiles, ou encore simplement avertir l'utilisateur de la faible fiabilité de l'information proposée. Pour cela, nous avons analysé différents indicateurs potentiels de performance de recherche par rapport aux requêtes données. Cet article se concentre sur les scores utilisés par différent types de systèmes de recherche documentaire pour le classement relatif des documents, et leur utilisation comme estimateurs absolus de performance. Nous allons montrer que, malgré une forte variation entre les scores fournis par différents systèmes, un grand nombre de ceux-ci peut effectivement être utilisé pour prédire la précision globale d'un ensemble de documents trouvés pour une requête.*

ABSTRACT. *In document retrieval it is often useful to have a measure of confidence in the retrieved document set in order to allow for specific treatment of particularly difficult queries, or simply alert the user to the low reliability of the information offered by the system. We have analyzed a number of potential retrieval performance indicators. This article concentrates on the scores used by different types of document retrieval systems for the relative ranking of documents for a given query, and their use as estimators of absolute retrieval performance. We will show that, despite a great variability among the scores used by different systems, an important number of these can effectively be used in order to predict the overall precision of a set of retrieved documents for a query.*

MOTS-CLÉS : *recherche documentaire, évaluation, classification automatique, difficulté des requêtes*

KEYWORDS: *document retrieval, evaluation, automatic classification, query difficulty*

1. Prédire la performance de recherche

Au-delà de fournir à l'utilisateur d'un système de recherche documentaire la meilleure liste possible de documents pour sa requête, il est souvent nécessaire de pouvoir estimer la qualité des résultats.

Avoir une bonne estimation de la performance en termes de pertinence pour une requête spécifique peut permettre de proposer à l'utilisateur des stratégies adaptées pour améliorer la qualité des résultats pour des requêtes pour lesquelles on obtient initialement une précision faible. Ces stratégies peuvent en particulier inclure de l'interaction avec l'utilisateur ou l'application de techniques coûteuses en temps de calcul ou autres ressources qu'on préférerait éviter quand elles ne sont pas nécessaires.

L'analyse de la difficulté d'une requête peut ainsi être intégrée dans le processus de recherche, qui se décompose alors en recherche initiale, analyse, et éventuellement une nouvelle phase de recherche adaptée spécifiquement à la requête. L'estimation de difficulté peut aussi se placer plus simplement à la fin du processus et fournir des informations supplémentaires à l'utilisateur sans intervenir directement dans la recherche d'information.

À ce jour, on trouve peu de travaux approfondis sur cette question, et typiquement avec un succès très modéré ([LOU 00, ROR 99, BAN 98]). Une des tentatives ayant eu le plus de succès semble être [CRO 02], utilisant un score décrivant la clareté (*clarity*), c.à.d. l'absence d'ambiguité d'une requête. Une approche différente est décrite dans [SUL 01], basée sur la similarité entre requêtes.

Il semblerait qu'aucun paramètre individuel ne puisse permettre de faire des prédictions précises. Cependant, certaines mesures peuvent contribuer à prendre des décisions fondées sur la performance de recherche attendue. Nous verrons dans cet article en particulier des mesures basées sur les scores fournis par différents systèmes de recherche documentaire.

Après une introduction au cadre expérimental de nos travaux et des méthodes d'évaluation utilisées, nous présenterons les mesures sur lesquelles nous nous sommes basés, avant de donner nos résultats expérimentaux et quelques conclusions.

2. L'environnement expérimental

2.1. *Collection de documents, requêtes et évaluations*

Nos travaux ont été conduits sur la base de données fournies par le NIST (*National Institute for Standards and Technology*), obtenues des campagnes annuelles TREC (*Text REtrieval Conference*[1]), et qui contiennent une grande collection de documents (environ 500 000) ainsi qu'un ensemble de requêtes (50 par an pour l'épreuve *ad-hoc*).

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

TREC propose différents contextes d'expérimentation. Nous avons choisi le contexte le plus "général", dans lequel les requêtes sont formulées de manière *adhoc*.

Les requêtes sont disponibles sous différentes formes, pouvant être composées d'un *titre* (quelques mots clés), d'un *descriptif* (une ou deux phrases) et d'un *narratif* (explication détaillée de la thématique recherchée).

Nous avons utilisé les soumissions de différentes équipes ayant participé à la campagne TREC 8 (voir [VOO 99]). Tous les résultats présentés ici ont été obtenus utilisant les requêtes de longueur moyenne (titre et descriptif) de la piste *adhoc*.

2.2. *Mesurer la performance d'un système de recherche documentaire*

Il existe différentes mesures qui permettent de juger de la performance d'une recherche documentaire.

Nous avons choisi la précision moyenne, car il s'agit d'une mesure utilisée communément pour un environnement applicatif générique, combinant différents aspects de qualité, comme la précision, la couverture, et l'ordre des documents trouvés. Elle est définie de la façon suivante :

Soit D l'ensemble des documents dans la collection, $R_q \in D$ l'ensemble des documents pertinents pour une requête q. $Rank_q(d)$ est la position du document d dans la liste ordonnée retournée pour la requête q par le système de recherche documentaire. Soit $P_q(n)$ la précision du résultat de la recherche pour la requête q sur les n premiers documents dans la liste.

La précision moyenne $AP(q)$ est alors :

$$AP(q) = \frac{\sum_{d \in R_q} P_q(Rank_q(d))}{|R_q|}$$

Nous allons dans la suite appeler "difficiles" les requêtes pour lesquelles un système de recherche documentaire renvoie un ensemble de documents avec une précision moyenne faible. De même, dans le cas d'une forte précision moyenne, nous considéreront la requête concernée comme facile.

Selon le contexte, la définition de difficulté peut être définie de manière plus générale, mais nous nous contenterons pour les résultats présentés ici de cette définition liée directement au système de recherche documentaire utilisé.

3. La qualité d'un prédicteur de performance

Dans un premier temps, il est nécessaire de quantifier la qualité d'un prédicteur de la précision moyenne obtenue pour une requête.

De manière générale, il est relativement difficile de quantifier la qualité d'un score x en tant que prédicteur d'un attribut y (dans notre cas de la précision moyenne obtenue pour une requête). Il y a une variété de mesures, allant de la simple corrélation linéaire à des méthodes statistiques complexes.

Nous avons utilisé deux mesures :

– l'impureté (ou surtout la réduction de l'impureté en séparant les données en classes) qui reflète assez intuitivement les résultats qu'on pourrait attendre en utilisant des algorithmes de classification et décision automatique.

– la corrélation de rangs de Spearman qui est un moyen reconnu de quantifier les dépendances entre différents attributs

3.1. *L'impureté*

Afin de trouver une mesure relativement simple et intuitive, nous avons choisi de calculer l'impureté des classes obtenues en séparant les examples selon le score choisi. L'impureté est définie comme la somme des variances des précisions moyennes dans chaque classe, pondérée par le nombre d'exemples :

$$Impurity = \frac{\sum_{c \in C} |c| \cdot \sigma_{AP_c}}{|Q|}$$

avec Q l'ensemble des exemples/requêtes, C l'ensemble des sous-ensembles disjoints $c \subset Q$, tel que $\bigcup c \in C = Q$, et σ_{AP_c} la variance des précisions moyennes des éléments de c.

Afin de construire les classes nous avons décidé de procéder par succession de séparations binaires optimales. L'optimalité de cette séparation est définie comme l'obtention de la plus faible impureté. La séparation consiste à déterminer une seuil sur le score x résultant en deux classes : $x < limite$ et $x \geq limite$.

Chaque classe est ensuite à nouveau séparée jusqu'à obtention du nombre désiré de classes. À cause du faible nombre d'exemples (50 requêtes) pour chaque système, nous nous sommes contentés de quatre classes.

Il est clair que cette approche ne garantit pas l'optimalité du résultat, mais la détermination de la séparation optimale en deux classes est triviale, alors qu'une séparation n-classes optimale nécessite des algorithmes beaucoup plus lourds. Ceci ne nous semblait pas nécessaire dans ce contexte.

3.2. *Le coefficient de corrélation de Spearman*

Afin de comparer nos résultats à ceux publiés dans [CRO 02], nous avons décidé de calculer également le coefficient de corrélation de rangs de Spearman ρ. Il s'agit là d'une mesure de corrélation reconnue et couramment utilisée.

Nous avons également déterminé la *P-value*, c.à.d. la probabilité que la distribution des données soit celle que l'on constate dans l'hypothèse d'une absence de corrélation. Cette *P-value* est basée sur le score de corrélation et le nombre d'exemples (de requêtes) utilisés pour le calculer : plus cette valeur est basse, plus on peut être certain qu'il existe réellement une corrélation.

4. Les scores utilisés

Afin de prédire ou estimer la difficulté d'une requête nous recherchons des mesures qui attribuent à chaque requête une valeur numérique (le score x) afin de prédire sans évaluation humaine la précision moyenne qu'obtiendra un système de recherche documentaire pour cette requête.

Nous avons analysé différents prédicteurs potentiels de performance de recherche, allant de caractéristiques simples comme la longueur des requêtes, l'ambiguïté des sens des mots, la synonymie ou spécificité des termes contenus dans la requête à des mesures plus sophistiquées, comme l'entropie ou la similarité cosinus moyenne de l'ensemble des documents classés en tête.

En particulier, nous avons trouvé une corrélation significative (bien que loin d'être parfaite) entre les scores utilisés pour classer les document trouvés et la précision moyenne obtenue pour une requête donnée.

Nous avons pu confirmer cette constatation pour un bon nombre de systèmes répandus, fondés sur des paradigmes différents.

Pour une requête donnée, la plupart des systèmes de recherche documentaire classent les documents grâce un score dérivé plus ou moins directement, soit :

– de la similarité entre documents et requêtes (p.ex. Smart [SAL 89])

– de la probabilité bayesienne $P(d|q)$, typiquement basée sur des probabilité d'occurrences de termes (p.ex. Okapi [ROB 96])

– des réseaux d'inférence bayesienne (p.ex. InQuery [CAL 92]).

Ces scores sont appropriés pour le classement relatif des documents en réponse à une requête, mais ils ne sont pas supposés donner une mesure absolue de la pertinence.

Les scores sont généralement disponibles dans les fichiers de résultats soumis aux évaluations TREC par les participants. On trouve ainsi pour chaque requête la liste des 1000 documents ayants obtenu les meilleurs scores, avec les scores associés ayant servi à classer les documents.

Nous avons utilisé dans nos expériences différentes transformations et normalisations de ces scores. Sur la base des scores attribués par le système de recherche aux documents trouvé, nous avons ainsi calculé des valeurs pour les requêtes. Une requête peut par exemple être représentée par la moyenne des scores des N premiers documents, le score du n^{ieme} document, ou encore le rapport entre les scores du 1^{er} et du n^{ieme} document.

Tous les résultats présentés dans cet article ont été calculés utilisant le score moyen sur les 20 premiers documents pour chaque requête.

5. Résultats expérimentaux

Nous notons que sur la base des scores de ranking de différents systèmes de RD, nous obtenons des mesures fortement liées à la précision moyenne obtenue pour chaque requête. Nous pouvons utiliser ces mesures pour séparer les requêtes en différentes classes qui diffèrent fortement quand à la précision moyenne.

La figure 1 montre les classes trouvées pour les système Okapi (ok8amxc) avec leurs limites supérieures et inférieures, la figure 2 fait de même pour le système AT&T.

Les bornes supérieures et inférieures que nous trouvons peuvent être très utiles pour décider de la manière de traiter une requête ou les options à présenter à un utilisateur. Pour le système Okapi, aucune requête avec un score en dessous de 3.15 n'obtient une précision moyenne supérieure à 0.28, et aucune requête avec un score en dessous de 3.7 n'atteint une précision moyenne au-delà de 0.59. On peut donc clairement supposer qu'une requête ayant un faible score produira des mauvais résultats de recherche. Elle devra être traitée en conséquence.

Des seuils correspondants peuvent être déterminés pour le système AT&T et d'autres systèmes. Ils sont représentés par des carrés dans les figures correspondantes.

Par contre, le score ne semble pas être du tout corrélé avec la précision moyenne obtenue pour une requête pour certains des systèmes. C'est en particulier le cas pour certains des systèmes manuels, peut-être en raison de l'utilisation d'un langage de requête formel et d'un classement des documents basé initialement sur une recherche booléenne. La figure 3 en montre un example. Certains systèmes automatiques affichent un comportement semblable (voir figure 4). Dans ces cas il n'est pas possible de déduire des seuils utiles sur la performance de recherche attendue.

Ceci démontre que l'estimation de pertinence en termes absolus et le classement relatif des documents pour une requête sont en fait des objectifs très distincts. La forte corrélation trouvée pour les scores utilisés par certains des systèmes nous paraît d'autant plus intéressante.

Nous pensons que l'analyse des différences entre les scores fortement corrélés avec la performance absolue et ceux qui ne le sont pas peut apporter des connaissances

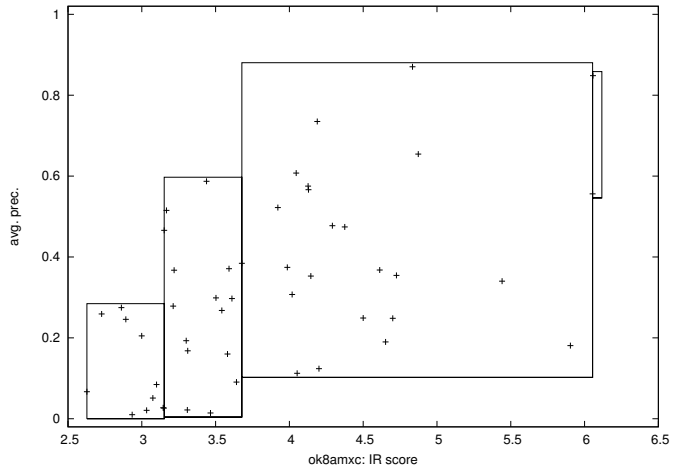

Figure 1. *ok8amxc : séparation en 4 classes*

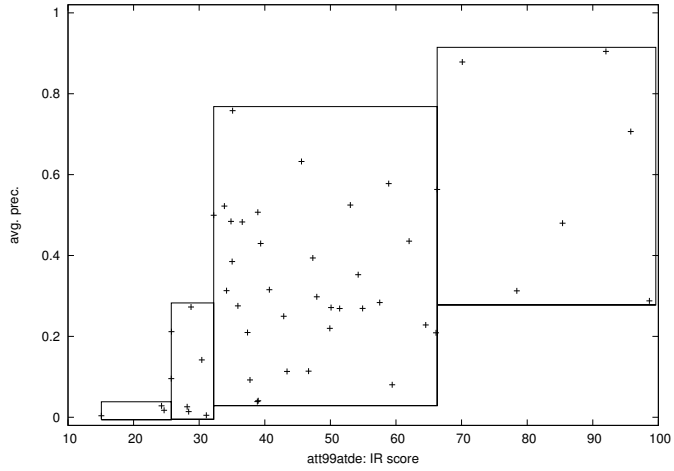

Figure 2. *att99atde : séparation en 4 classes*

intéressantes. Celles-ci pourraient être utiles pour divers aspects de la recherche documentaire.

Bien que la *corrélation de rangs* et le quotient $\frac{impurete}{variance}$ montrent un comportement similaire en général, nous notons qu'en particulier les scores *InQuery* ont une forte *corrélation de rangs* alors que l'impureté n'est pas significativement plus basse que pour d'autres systèmes. Il n'est pas entièrement clair comment cela se traduit

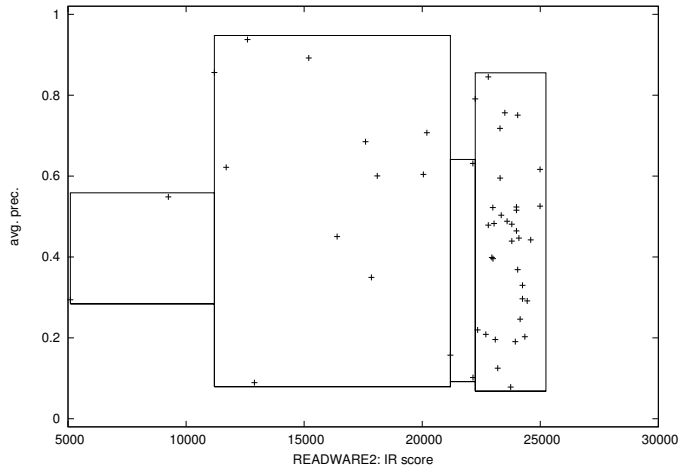

Figure 3. *READWARE2 : séparation en 4 classes*

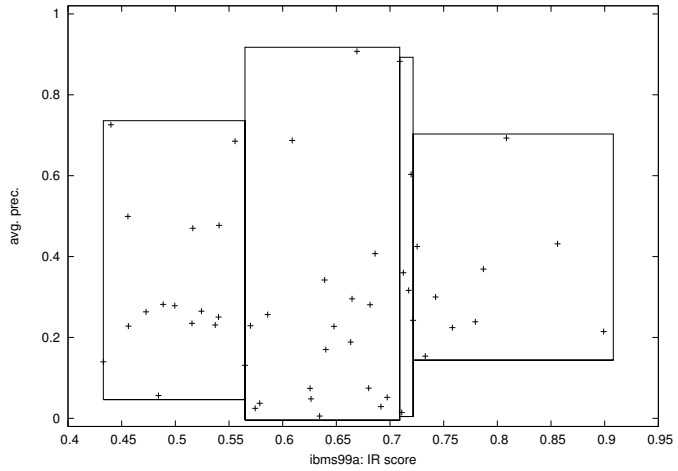

Figure 4. *ibms99a : séparation en 4 classes*

dans une réelle situation d'application en essayant de prendre des décisions basées sur l'estimation de difficulté d'une requête.

Le tableau 1 montre la moyenne (sur toutes les requêtes) des précisions moyennes, la variance globale, l'impureté et le quotient de l'impureté par rapport à la variance, ainsi que la corrélations de rangs et la valeur P pour certains des participants à TREC. Les noms des systèmes sont ceux utilisés lors des soumissions des résultats aux campagnes TREC.

	Name	Variance	Impurity	*Ratio*	Rho	*P-Value*	M. Av. Pr.
automatic	INQ603	0.0433	0.0193	0.4457	0.7129	$3.26e-08$	0.2658
	tno8d3	0.05	0.0301	0.602	0.5695	$2.24e-05$	0.2921
	MITSLStd	0.054	0.0313	0.5796	0.5490	$4.88e-05$	0.2978
	Sab8A4	0.041	0.0215	0.5243	0.5264	$1.09e-04$	0.2607
	ok8amxc	0.0475	0.0276	0.5810	0.5183	$1.43e-04$	0.3168
	att99atde	0.0509	0.0289	0.5677	0.4648	$7.64e-04$	0.3165
	Flab8atd2	0.0508	0.0352	0.6929	0.3272	$2.07e-02$	0.2929
	pir9Attd	0.0581	0.0451	0.7762	0.1331	0.3553	0.3206
	ibms99a	0.0484	0.0386	0.7975	0.0653	0.6509	0.3005
	fub99td	0.0488	0.0402	0.8237	-0.0238	0.8691	0.3063
manual	READWARE2	0.0492	0.0409	0.8313	-0.2392	0.0943	0.4692
	8manexT3D1N0	0.0419	0.0318	0.7589	0.1405	0.3291	0.3346
	iit99ma1	0.0671	0.0577	0.8599	0.1088	0.4504	0.4103
	CL99XTopt	0.0487	0.0427	0.8767	0.1001	0.4877	0.3765

Tableau 1. *impureté, corrélation de rangs et moyenne des précisions moyenne*

Une bonne performance de classification (c.à.d. la capacité à différencier par exemple entre les requêtes “faciles” et “difficiles”) se distingue généralement par une faible variance intra-classe (faible impureté). Plus les classes de requêtes sont homogènes quand à leur difficulté, plus la variance de la précision moyenne est faible. Le quotient $\frac{impurete}{variance}$ normalise cela par la variance d’origine de la précision moyenne.

En résumé, nous trouvons typiquement une réduction considérable de l’impureté par rapport à la variance globale de la précision moyenne.

6. Conclusions

Il semble évident que certains des scores ont un grand potentiel pour aider à estimer de manière automatique la qualité des résultats de recherche pour une requête donnée. Le fait que ce n’est pas le cas pour tous les scores des différents systèmes montre que le rapport entre les valeurs attribuées aux documents et la qualité du résultat n’est pas trivial.

Malheureusement nous n’avons pas encore pu discerner de différence fondamentale entre les scores reflétant réellement la pertinence probable et d’autres scores tout aussi efficaces pour le classement relatif des documents (à part quelques cas avec une normalisation à posteriori évidente qui rend le score inutilisable pour notre application).

Les résultats obtenus en termes de réduction d’impureté sont meilleurs que ceux de la plupart des autres mesures que nous avons traitées à ce jour. Un grand nombre de scores montrent également une *corrélation de rangs* avec la précision moyenne beaucoup plus forte que par exemple le score de *clarté (clarity)* de [CRO 02].

Nous pensons que l’utilisation de scores appropriés issus des systèmes de recherche peut aider considérablement à estimer la qualité du résultat avant de le présen-

ter à l'utilisateur. En conjonction avec d'autres techniques ceci pourrait nous permettre de traiter les sujets difficiles d'une manière mieux adaptée et ainsi augmenter la satisfaction de l'utilisateur.

Selon l'environnement applicatif, plutôt que de déterminer des limites supérieures et inférieures de la performance attendue, il peut être utile de définir des seuils sur la précision moyenne et d'optimiser la classification de manière à maximiser la probabilité que les requêtes de chaque classe sont par exemple "faciles" ou "difficilies".

Il est intéressant de noter qu'il n'y a pas de lien apparent entre la possibilité de prédire la performance de recherche sur la base des scores utilisés par les systèmes et les performances en tant que système de recherche documentaire. Ainsi le score utilisé par *InQuery* (INQ603) est fortement lié à la qualité du résultat, ce qui veut dire qu'il s'agit en quelque sorte d'une mesure absolue de pertinence (probable) des documents correspondants. Par contre, par exemple le système *IBM* (ibms99a), en utilisant un score qui ne donne aucune indication absolue de la qualité du résultat, obtient globalement de meilleures performances.

7. Perspectives

Une comparaison plus détaillée des scores utilisés par les différents systèmes pourrait être utile dans l'objectif d'une meilleure compréhension de la manière dont ils reflètent réellement la probabilité de pertinence des documents. Ceci pourrait amener au développement de normalisations inter-requêtes des scores pour les rendre plus utiles en tant qu'indicateurs absolus de qualité du résultat.

Nous avons commencé à valider ces résultats sur différentes collections de documents et requêtes en utilisant notre propre système de recherche documentaire. Un autre aspect de nos expérimentations en cours est l'exploration des effets de différents paramètres du système sur la qualité en tant que prédicteur de performance.

Nous travaillons également sur l'intégration de multiples estimateurs en utilisant différents algorithmes de classification automatique [GRI 04]. En combinant les mesures présentées ici avec d'autres indicateurs nous arrivons à des résultats très satisfaisants pour une classification binaire "facile/difficile". La classification est utilisée dans un processus de décision qui nous permet d'avoir une paramétrisation dépendante de la requête du système de recherche et de contrôler l'utilisation de différentes techniques d'enrichissement. Un objectif particulier est le développement de méthodes interactives sur la base de l'analyse des résultats d'une recherche initiale.

8. Bibliographie

[BAN 98] BANKS D., OVER P., ZHANG N.-F., « Blind Men and Elephants : Six Approaches to TREC data », *Information Retrieval*, , 1998.

[CAL 92] CALLAN J. P., CROFT W. B., HARDING S. M., « The INQUERY Retrieval System », *Proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications*, 1992, p. 78–83.

[CRO 02] CRONEN-TOWNSEND S., ZHOU Y., CROFT W. B., « Predicting query performance », *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, 2002, p. 299–306.

[GRI 04] GRIVOLLA J., « Méthodes statistiques et apprentissage automatique pour l'évaluation de requêtes en recherche documentaire », *actes de Récital*, Fès, Maroc, 2004.

[LOU 00] DE LOUPY C., BELLOT P., « Evaluation of document retrieval systems and query difficulty », *Acte de "Using Evaluation within HLT Programs : Results and Trends"*, Athènes, Grèce, 2000, p. 31-38, http ://www.limsi.fr/TLP/CLASS/ClassD43.pdf.

[ROB 96] ROBERTSON S., WALKER S., JONES S., HANCOCK-BEAULIEU M., GATFORD M., « Okapi at TREC-3 », *Overview of the Third Text REtrieval Conference (TREC-3)*, 1996.

[ROR 99] RORVIG M., « Retrieval Performance and Visual Dispersion of Query Sets », *TREC-8 Proceedings*, 1999, http ://trec.nist.gov/pubs/trec8/t8_proceedings.html.

[SAL 89] SALTON G., *Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer*, Addison–Wesley, 1989.

[SUL 01] SULLIVAN T., « Locating question difficulty through explorations in question space », *Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries*, ACM Press, 2001, p. 251–252.

[VOO 99] VOORHEES E. M., HARMAN D., « Overview of the Eighth Text REtrieval Conference », *TREC-8 Proceedings*, 1999.