Comment mesurer la qualité d'un système de classement de résultats ?

Par Guillaume et Sylvain Peyronnet

Domaine :	Recherche	Référencement
Niveau :	Pour tous	Avancé

Tout moteur de recherche a besoin de classer ses résultats. Cette notion de mesure de la pertinence par rapport à une requête donnée a fait l'objet de nombreux travaux scientifiques depuis des années. Voici quelques pistes suivies par les chercheurs et quelques modèles utilisés, et quels sont les moyens utilisés pour évaluer ces méthodes ayant pour objectif d'atteindre le Graal du search : la meilleure pertinence possible...

Pour commencer cet article, nous allons revenir à la base la plus évidente : l'objectif d'un moteur de recherche est de déterminer parmi un très grand ensemble de documents (textuels d'abord, mais aussi plus évolués comme par exemple des images ou des vidéos) ceux qui sont pertinents par rapport à un besoin informationnel. La notion même de besoin informationnel, ainsi que celle de pertinence, sont très difficile à capturer puisqu'elle est dépendante de chaque personne. Par ailleurs, le canal de communication entre le moteur et l'utilisateur est imparfait et parcimonieux (les requêtes sont courtes, et l'utilisateur peut se tromper).

En conséquence, les moteurs doivent « apprendre » ce qui est pertinent et ce qui ne l'est pas. Dans cet article nous ne parlerons pas de comment se fait cet apprentissage, mais de la façon dont le moteur va savoir si il est dans l'erreur ou dans le vrai, c'est-à-dire l'évaluation de la qualité de l'algorithme qui détermine la pertinence. Nous avons déjà évoqué le problème de l'évaluation dans notre article du mois dernier, mais plutôt et principalement sous l'angle de l'approche dynamique : on extrapole la satisfaction des utilisateurs à l'aide d'un monitoring de son comportement. Ici, on verra plutôt quelles sont les mesures chiffrées que l'on peut utiliser pour noter les classements (et implicitement les algorithmes de classement) de manière objective, dans l'idée de nourrir des algorithmes d'apprentissage qui produiront de meilleurs classements.

La notion de modèle de recherche

Un modèle de recherche est une famille d'algorithmes qui va, pour un besoin informationnel particulier (souvent exprimé par une requête) renvoyer un sous-ensemble de l'ensemble des documents connus (le corpus, ou encore l'index du moteur de recherche).

Les documents renvoyés seront appelés les documents pertinents, ceux qui ne sont pas renvoyés seront les documents non pertinents. Dans l'idéal, un modèle de recherche est correct et renvoie toujours des documents pour lesquels cette pertinence algorithmique correspond à la pertinence sémantique : les documents répondent au besoin informationnel.

On le sait bien, l'idéal n'arrive jamais dans un contexte de documents produits et

demandés par des êtres humains, et il va être nécessaire de mettre en place des mécanismes d'évaluation des classements pour mieux mesurer l'efficacité des différents modèles de recherche, et de leurs implémentations.

Il existe trois grands types de modèles de recherche. Les plus simples sont **booléens** *(voir l'article de Salton et al. [1])*. Les modèles booléens considèrent que les documents sont des ensembles de termes. Les requêtes vont être des formules propositionnelles (des mots reliés par les opérateurs ET, OU, NON) mises au regard des termes. Ainsi la requête « jaguar ET voiture » se verra associée comme résultat pertinent les pages qui contiennent les termes « jaguar » et « voiture ». Historiquement, ces modèles ont été les premiers mis au point et ont été utilisés pendant plusieurs dizaines d'années (jusqu'aux années 90, ils étaient prédominants dans les outils commerciaux). Les modèles **vectoriels** *(voir [2])* sont des modèles qui sont plus riches que les modèles booléens car ils permettent de renvoyer une liste ordonnée de documents pertinents. Ils fournissent donc une notion de pertinence partielle ou graduée, c'est-à-dire une graduation de l'adéquation des documents à une requête donnée. L'avantage de ces modèles est qu'ils permettent de ne fournir qu'un petit nombre de documents à un utilisateur (les « plus pertinents » au lieu des « pertinents »).

Enfin, le troisième type de modèle est **probabiliste** et à pour objectif de quantifier la probabilité qu'un document soit pertinent pour une requête *(voir [3])*. Parmi ces modèles, on trouve par exemple le modèle de référence Okapi BM25 *[4]*.

Nous allons maintenant rentrer dans le vif du sujet, et aborder les mesures d'évaluations.

Pourquoi et comment évaluer ?

L'évaluation va être un moyen pour le moteur de recherche de déterminer l'efficacité de son modèle de recherche, afin de le paramétrer pour augmenter la qualité des résultats qu'il fournit.

Une évaluation est un jugement de la pertinence réelle de résultats fournis par l'algorithme dans un contexte d'utilisation réelle. C'est-à-dire que l'on va regarder, pour chaque requête de test, les documents qui sont renvoyés par le moteur. Ces documents sont notés (par exemple par un score 0/1 – pertinent ou pas pertinent – ou par un score allant de 0 à 5 – correspondant à une graduation dans la pertinence) et les notes obtenues sont agrégées au sein de mesures globales qui permettent de déterminer l'efficacité globale de l'algorithme. Dans cet article, on part donc du principe qu'il y a des humains qui sont capables de noter chaque page comme pertinente ou non pour une requête donnée.

Si on résume : on va tester le moteur sur un certain nombre de requêtes, puis on va utiliser le score de chaque requête pour obtenir une mesure de qualité qu'on va espérer être universelle. Une notion clé qui apparaît donc est celle du corpus de test (ou *dataset*) qui est un ensemble de requêtes qui vont être suffisament représentatives pour qu'on puisse comparer différents algorithmes avec les résultats de recherche sur ces requêtes.

Pour des raisons statistiques, ce nombre ne peut pas être inférieur à une trentaine de requêtes car sinon les tests statistiques usuels de type *student* ou *chi2* ne peuvent pas être utilisés. Nous rappelons ici que ces tests servent à déterminer si le résultat d'une expérimentation est du au hasard ou s'il est reproductible. Par ailleurs, le

nombre de requêtes ne peut pas être trop important, car l'expérimentation est une notation par des êtres humains, et si il y a trop de requêtes, le coût de l'expérience et sa durée sont trop importants.
Il existe de nombreux corpus de test qui sont disponibles sans contraintes pour la recherche, et nous allons par la suite en évoquer quelques uns qui peuvent être utiles pour les lecteurs qui souhaiteraient développer leurs propres outils.
La banque de corpus de test la plus connue est sans conteste celle de la conférence TREC *(Text REtrieval Conference) [5]*. Cette conférence existe depuis 1992 et a rendue public des corpus de test pour de très nombreux aspects de la recherche d'information.
Une autre collection de requêtes est disponible dans le corpus CRANFIELD (évoqué dans notre article du mois dernier). Cette dernière est cependant très petite (225 requêtes) et trop ancienne pour être vraiment utile. Un corpus intéressant car centré sur les langages des pays européens est celui de l'initiative CLEF (associée à la conférence éponyme) *[6]*. On trouve même des corpus d'images dans cette initiative, qui promeut tous les aspects de la recherche d'information.
Il existe encore bien d'autres corpus, mais avec TREC et CLEF, on peut réaliser l'évaluation de la plupart des systèmes que l'on peut imaginer, sans aucun problèmes.
Nous allons maintenant voir comment on mesure l'efficacité des méthodes de calcul de la pertinence.

Evaluation d'ensembles non ordonnés

En tant qu'utilisateur de moteur de recherche de type Google, nous n'avons pas l'habitude des systèmes qui fournissent des résultats non ordonnés. Ils sont pourtant les plus intuitifs et correspondent à l'idée de base de la recherche d'information, qui est sensée renvoyer quelques documents pertinents, sans les classer entre eux.
Les deux mesures canoniques pour l'évaluation d'ensembles non ordonnés sont celles que l'on retrouve dans de nombreux domaines expérimentaux : la précision et le rappel (precision et recall en anglais). Ces deux mesures sont basées sur l'observation du résultat pour chaque document renvoyé. Ce résultat peut être « OK, le document est pertinent » ou « pas OK, le document n'est pas pertinent ».
On a donc des mesures qui vont se baser sur le comptage des documents renvoyés par l'algorithme et qui ne sont pas pertinents (des faux positifs) et sur le comptage des documents qui ne sont pas renvoyés mais qui auraient du l'être.
Les formules définissant le rappel et la précision sont indiquées sur la figure 1.
La figure 2 donne une représentation intuitive de ces deux quantités.

$$\text{rappel} = \frac{\text{nombre de documents pertinents renvoyés par l'algorithme}}{\text{nombre de documents pertinents}}$$

$$\text{précision} = \frac{\text{nombre de documents pertinents renvoyés par l'algorithme}}{\text{nombre de documents renvoyés}}$$

Fig.1.

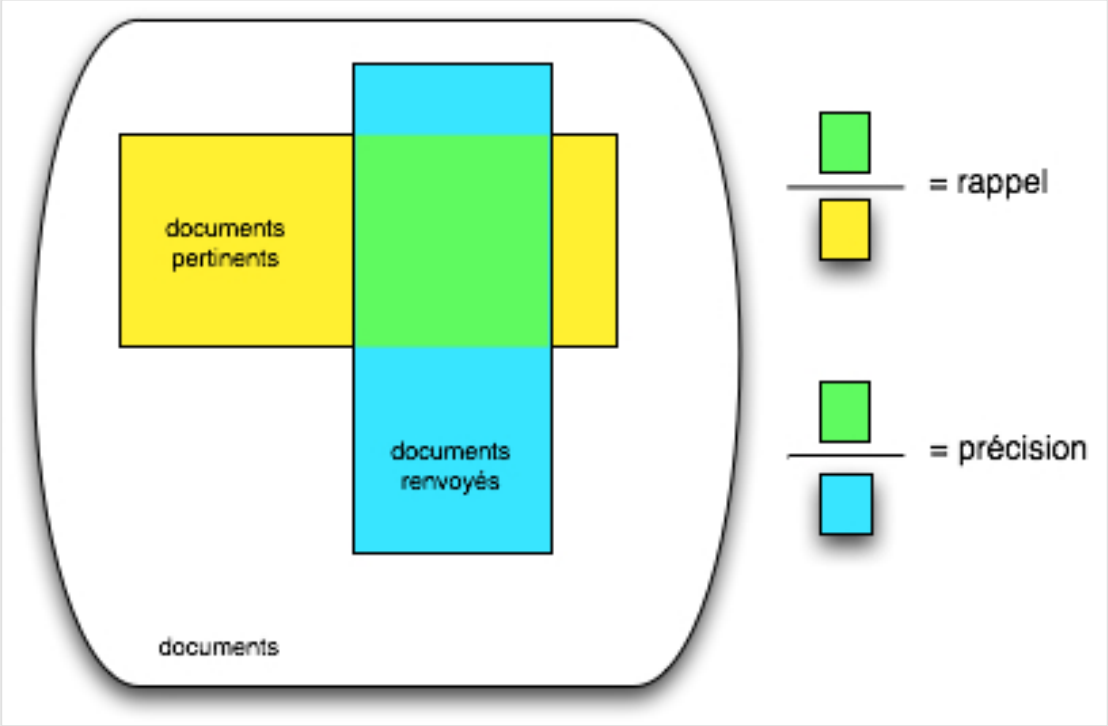

Fig.2.

Lors d'une évaluation, ces deux mesures sont calculées pour chaque requête, et ensuite on fait la moyenne pour l'ensemble des requêtes. On obtient donc deux scores pour caractériser un système de classement. Un algorithme qui renvoie tous les documents a un rappel de 1 et une faible précision tandis qu'un algorithme qui renvoie peu de documents, mais tous réellement pertinents, aura une forte précision, mais un faible rappel. Ce dernier cas est typiquement ce que l'on attend d'un moteur de recherche.

Souvent, plutôt que d'utiliser ces deux mesures séparément, on utilise une troisième mesure, qui agrège la précision et le rappel. Il s'agit de la mesure-F *(voir [7])*. Il s'agit en fait d'une famille de mesures, qui sont définies ainsi (images prises sur wikipedia) :

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}$$

Le choix d'un paramètre beta de 1 donne autant de poids à la précision et au rappel. En faisant varier ce paramètre, on privilégie l'une ou l'autre des mesures, notamment, si on fixe un beta inférieur à 1, on donne plus de poids à la précision, ce qui permet de mesurer ce que l'on attend d'un moteur de recherche usuel.
Enfin, il existe une dernière mesure agrégé, la **mesure-G**, qui dépend du produit de la précision et du rappel, mais que l'on utilise moins.

Evaluation d'ensembles ordonnés

Les mesures que nous avons vues précédemment permettent des évaluations intéressantes sur des ensembles non ordonnés. Le problème est que le but d'un moteur de recherche est de renvoyer des résultats dans un ordre de pertinence décroissante. L'ordre d'apparition des résultats est donc important et doit être valorisé par la mesure choisie. Il n'est pas question de donner autant de poids à un document pertinent renvoyé en 10eme position qu'à un document pertinent placé en première position.
Voici quelques mesures qui sont utilisées pour le cas des algorithmes renvoyant des ensembles ordonnés.

Courbe précision versus rappel

Supposons que le moteur de recherche renvoie 10 résultats. Pour prendre en compte l'ordre, nous allons analyser 10 ensembles non ordonnés construits à partir de l'ensemble ordonné des 10 résultats. Imaginons que le moteur renvoit les pages dans l'ordre P1, P2, P3, …, P10.

- Le premier ensemble non ordonné (E1) contient P1.
- Le deuxième (E2) contient P1 et P2.
- Le troisième (E3) contient P1, P2 et P3.
- Et ainsi de suite jusque E10 qui contient les 10 pages.

On va ensuite calculer la précision et le rappel pour chaque ensemble. Une fois ce calcul effectué, on place dans un espace à 10 dimensions les points correspondant à chaque ensemble en matière de précision et de rappel, et on prend le top (courbe en rouge de la figure 3).
Cette mesure, qui est très graphique, est la plus simple à comprendre parmi les mesures pour l'évaluation des ensembles ordonnés, mais elle a aussi un sens : la courbe en rouge est une assez bonne représentation du pourcentage de

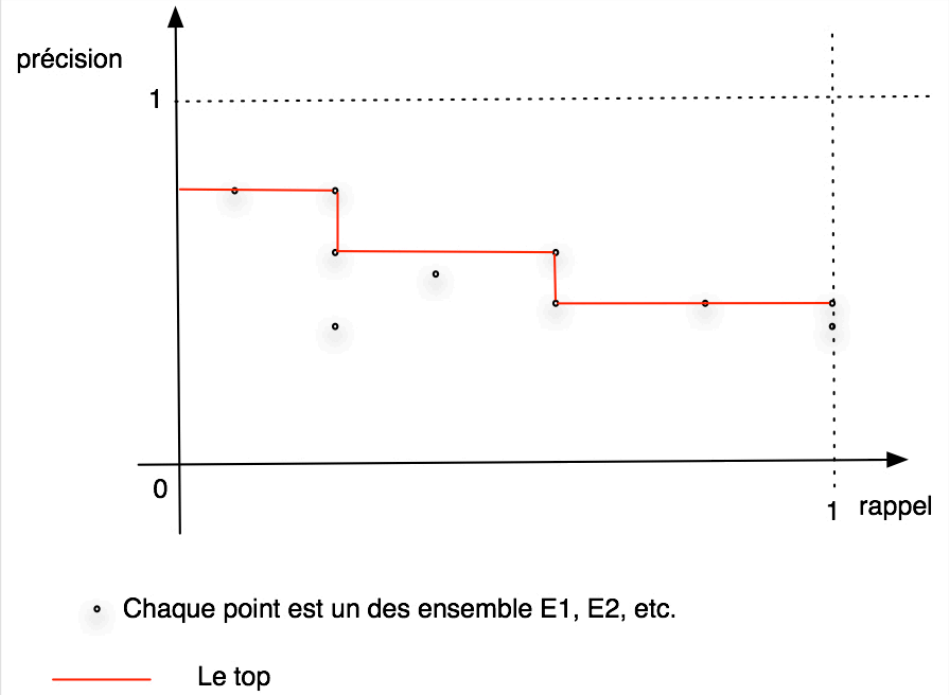

Fig.3.

documents pertinents qu'un utilisateur peut obtenir avec une valeur donnée du rappel. Par exemple, le point le plus à droite de la courbe de la figure signifie que si je veux que tous les documents pertinents soient fournis par le système, j'aurai environ la moitié de mes résultats qui seront des documents parasites non pertinents.

Courbe ROC

La courbe ROC *(Receiver Operating Characteristics en anglais) [8]* est une mesure générique que l'on retrouve dans tous les domaines scientifiques pour évaluer les mécanismes de classification. Concernant les algorithmes de classement pour les moteurs de recherche, la courbe ROC va indiquer la capacité de l'algorithme à présenter les documents pertinents avant (dans l'ordre de lecture) les documents non pertinents. L'idée est donc intuitive : on va favoriser les méthodes qui présentent les résultats pertinents le plus haut possible dans le classement.

Plus concrètement, on va partir d'une liste ordonnée de documents (construite par l'algorithme) et on va estimer le rappel à chaque rang dans la liste. Cela veut dire qu'on regarde la proportion de documents pertinents renvoyés par l'algorithme à chaque rang. En même temps, on va regarder le taux de documents non pertinents ordonnés avant ce rang. Cela va nous donner un couple de points que l'on peut placer dans une espace à deux dimensions. Le couple (0,2 ; 0,5) pour le rang 100 signifie que le moteur a renvoyé 20% des documents pertinents dans les 100 premiers du classement, mais qu'en même temps il a placé dans ces 100 premiers 40% des documents non pertinents.

Parce qu'une figure est toujours plus simple à comprendre, voici un exemple d'une telle courbe sur la figure 4.

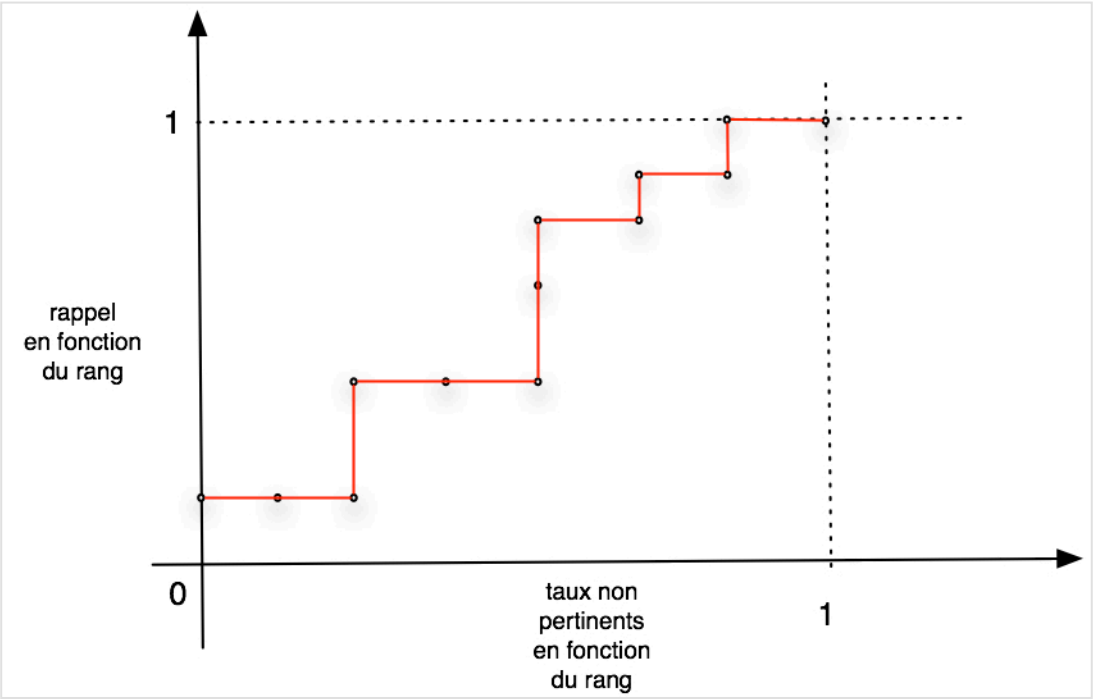

Fig.4.

La courbe ROC est la courbe en rouge. La mesure d'efficacité se fait en regardant l'aire sous la courbe. Plus cette aire est importante, plus le système qui est évalué est capable de présenter les résultats pertinents avant les résultats non pertinents. Pour un moteur de recherche, cela veut dire que les meilleures réponses à une requête sont avant les mauvaises.

Autres mesures

Il existe de très nombreuses autres mesures qui sont définies de manière plus mathématiques, ce qui explique pourquoi nous n'allons pas rentrer dans les détails ici mais simplement les évoquer et donner les intuitions associées.

Une mesure qui est très souvent utilisée est la **précision moyenne**. Nous avons évoqué dans la précédente lettre la notion de précision à 10 (p@10). La précision moyenne d'un algorithme de classement pour une requête donnée est la moyenne des valeurs de précision pour les documents pertinents pour la requête en question, dans la liste ordonnée des réponses. Si on a un corpus de requêtes complet, on va faire ce calcul de précision moyenne pour chaque requête, puis on va ensuite faire la moyenne de ces précisions moyennes. Cette moyenne des moyennes a un nom : le MAP (*Mean Average Precision* en anglais) du système. La mesure MAP est la mesure la plus utilisée dans les évaluations en recherche d'information, car elle est très stable et capable de séparer finement les algorithmes selon leur efficacité. Les lecteurs voulant en savoir plus trouveront des informations dans les références [9] et [10] à la fin de cet article.

Enfin, une dernière mesure que nous souhaitons aborder est le gain cumulatif réduit (DCG pour *Discounted Cumulative Gain*). Sa définition est particulièrement complexe et nous ne dirons rien de plus que le fait qu'il s'agit d'une mesure qui est capable de prendre en compte des jugements gradués. En effet, toutes les mesures précédentes utilisent un jugement de pertinence totale au niveau du couple document – requête. On peut bien sûr demander à des humains de noter de façon abrupte, mais naturellement les évaluateurs ont des réponses plus fines : chaque document peut être beaucoup, un peu ou pas du tout pertinent. Pour agréger des avis plus fins, il faut une mesure spécifique telle que la mesure DCG. Plus de détails sur cette mesure se trouvent dans la référence [11].

Conclusion

Dans cette article, nous avons évoqué les différentes mesures qu'on peut utiliser pour comparer l'efficacité des algorithmes de classement. Toutes ces mesures agrègent une notation humaine de la pertinence d'une page donnée pour une requête donnée en un score qui permet de qualifier la qualité des algorithmes eux-mêmes. Il s'agit d'un élément important et indispensable pour les moteurs de recherche, sans lequel il serait impossible de paramétrer finement les systèmes de classement.

Références

[1] Salton, G., Fox, E. A., & Wu, H. (1983). *Extended Boolean information retrieval. Communications of the ACM*, 26(11), 1022-1036.
http://www.ecommons.cornell.edu/bitstream/1813/6351/1/82-511.pdf

[2] Salton, G., Wong, A., & Yang, C. S. (1975). *A vector space model for automatic indexing. Communications of the ACM*, 18(11), 613-620.

https://wwwold.cs.umd.edu/class/fall2009/cmsc828r/PAPERS/VSM_salton-2.pdf

[3] Robertson, S. E., & Jones, K. S. (1976). *Relevance weighting of search terms*. Journal of the American Society for Information science, 27(3), 129-146. http://www.soi.city.ac.uk/~ser/papers/RSJ76.pdf

[4] Whissell, J. S., & Clarke, C. L. (2011). *Improving document clustering using Okapi BM25 feature weighting*. Information Retrieval, 14(5), 466-487.

[5] Les corpus de TREC. http://trec.nist.gov/

[6] Site de l'initiative CLEF http://www.clef-initiative.eu/

[7] Mesure-F https://en.wikipedia.org/wiki/F1_score

[8] Fawcett, T. (2006). *An introduction to ROC analysis. Pattern recognition letters*, 27(8), 861-874. http://people.inf.elte.hu/kiss/13dwhdm/roc.pdf

[9] Sanderson, M., & Zobel, J. (2005, August). *Information retrieval system evaluation: effort, sensitivity, and reliability*. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 162-169). ACM. http://ww2.cs.mu.oz.au/~jz/fulltext/sigir05.pdf

[10] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). New York: ACM press.

[11] Järvelin, K., & Kekäläinen, J. (2002). *Cumulated gain-based evaluation of IR techniques*. ACM Transactions on Information Systems (TOIS), 20(4), 422-446. http://www.sis.uta.fi/infim/julkaisut/fire/KJJK-nDCG.pdf

Guillaume Peyronnet *est gérant de Nalrem Médias.* **Sylvain Peyronnet** *est co-fondateur et responsable des ix-labs, un laboratoire de recherche privé. Ensemble, ils font des formations, pour en savoir plus : http://www.peyronnet.eu/blog/*