L'évaluation des SERP par le monitoring du comportement des utilisateurs

Retour au sommaire de la lettre

Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

Un point stratégique indispensable pour un moteur de recherche consiste à mesurer la pertinence de son algorithme au travers de la satisfaction (ou non) de ses utilisateurs. Pour cela, plusieurs méthodes sont utilisées, soit en faisant inetrvenir des experts dédiés à cette tâche (quality raters), soit en mesurant diretcement le comportement des visiteurs du site. Voici une petite revue d'effectif des différents tests le plus souvent mis en place à l'heure actuelle...

Dans cet article, nous allons aborder un sujet important dans la mise en place des moteurs de recherche : **l'évaluation humaine de la qualité des classements mis en place algorithmiquement**. Comme nous allons le voir, il existe de nombreuses méthodes pour évaluer cette qualité, mais pour de multiples raisons, il est important de prendre en compte l'avis des utilisateurs réels du moteur (qui peut être différent de l'avis des utilisateurs experts que sont par exemple les *quality raters* professionnels). Prendre en compte les avis des utilisateurs lors de leur usage standard du moteur de recherche, c'est faire ce que l'on appelle de l'évaluation dynamique ou en-ligne (*online evaluation* en anglais).

Nous allons tout d'abord définir précisément quel est la question à laquelle répondent les méthodes d'évaluation, quelles soient statiques ou dynamiques. Nous verrons ensuite ce qu'est l'évaluation statique ou non connectée (*offline evaluation* en anglais) . Enfin nous présenterons les principales méthodes d'évaluation dynamique.

Il est important de noter que, même si le grand public en entend rarement parler, cette problématique de l'évaluation des SERP (page de résultats des moteurs de recherche) sest au cœur de tous les moteurs de recherche, car c'est pour eux le seul moyen de s'assurer que l'algorithmique mis en place fournit réellement des résultats qui sont considérés comme pertinents par les utilisateurs.

Une partie du contenu de cet article est largement inspiré d'un tutoriel de Filip Radlinski et Katja Hofmann à la conférence SIGIR 2011 (liens et sources disponibles à la fin de cet article).

A quoi sert l'évaluation des SERP ?

Une méthodologie d'évaluation de la qualité des SERP va essentiellement servir à répondre à une question très simple : étant donnés deux algorithmes de classement différents, quel est celui qui est va fournir les meilleurs résultats, c'est-à-dire fournir les sites web considérés comme les plus pertinents par l'utilisateur ?

De prime abord, on peut imaginer qu'il suffit de présenter deux SERP différentes (chacune obtenue à l'aide d'un algorithme particulier) et de demander à un utilisateur celle qu'il préfère pour régler le problème. Mais cela n'est bien sûr pas si simple : tout d'abord chaque classement dépend certes d'un algorithme, mais il dépend aussi de la requête et de l'utilisateur. Pour répondre de manière correcte, il va donc falloir considérer de nombreuses requêtes, et de nombreux utilisateurs. Par ailleurs, on peut imaginer des méthodes beaucoup plus robustes que de demander l'avis sur deux SERP totalement indépendantes. On va ainsi créer des mécanismes d'entrelacement : on ne présente qu'un seul classement, avec des résultats provenants de plusieurs algorithmes différents, et après plusieurs observations, on sera capable de dire quel est le meilleur algorithme. Procéder ainsi paraît plus compliqué, mais donne de meilleurs résultats car on ne « pollue » pas le comportement de l'utilisateur en l'observant dans son interaction avec un seul classement.

Un bon mécanisme d'évaluation doit répondre à plusieurs critères pour être mis en place dans le workflow d'un moteur de recherche. Il doit être **pratique** : s'il nécessite des ressources trop conséquentes, alors le mettre en place n'est pas faisable car cela coûtera plus que cela ne rapportera. Il doit également être **correct** : si l'évaluation prédit qu'un classement est

meilleur qu'un autre, cela doit être aussi l'avis global de tous les utilisateurs du moteur. Et enfin, il doit être **efficace** : l'interférence entre l'évaluation et la production doit être la plus faible possible (c'est-à-dire que la plupart des utilisateurs doivent voir le meilleur classement, pas le moins bon).

Evaluation « offline »

Le principe de l'évaluation offline est très simple : on va demander à un groupe d'experts ou d'utilisateurs choisis d'évaluer explicitement le système de classement. Pour cela, on va présenter des classements pour un certain nombres de requêtes qu'on estime représentatives des demandes de l'ensemble des utilisateurs du moteur, et on va demander à chaque expert de noter ces classements.

Il existe de nombreuses métriques pour noter les classements. La littérature sur le PageRank et ses variantes fait par exemple un grand usage de la « précision à 10 » que l'on note généralement **p@10**. La p@10 est maximale quand un expert estime que les 10 premiers résultats d'un classement méritent d'être sur la première page (donc d'être dans les 10). Cette mesure est assez faible, car par exemple, elle ne distingue pas la première position de la quatrième, seulement le fait d'être dans les 10 premiers ou dans le reste. Nous renvoyons le lecteur intéressé à la référence [1] pour en savoir plus sur les métriques utilisées en recherche d'information.

Cette méthode d'évaluation offline est encore à l'heure actuelle celle qui est utilisée majoritairement, car il est plus simple pour un moteur de recherche d'encadrer des quality raters que de monitorer le comportement des utilisateurs (avec le problème de robustesse que cela entraine, c'est-à-dire le fait que des utilisateurs assez nombreux pourrait essayer de manipuler les SERP). Comme de nombreux « outils » de la recherche d'information, c'est dans les années 60 que cette technique a été mise en place, par Cleverdon à l'Université de Cranfield (on parle d'ailleurs de l'approche Cranfield, voir [2]).

Ce type d'approche implique plusieurs hypothèses qui ne sont pas nécessairement valides. Tout d'abord, elle part du principe que les experts ont raison, c'est-à-dire qu'ils comprennent les besoins des utilisateurs réels et sont capables de décider si des résultats correspondent à ces besoins. Ensuite, il faut que les requêtes analysées soient représentatives, ce qui n'a rien d'évident, sauf à en choisir des dizaines de milliers. Enfin, et c'est le principal écueil, il faut que la métrique utilisée (comme la p@10 par exemple) soit en adéquation avec la notion de qualité réelle (celle qui est « dans le cerveau » de l'utilisateur). Lorsque ces hypothèses ne sont pas satisfaites, il n'y a aucune garantie que l'évaluation soit correcte, ce qui est bien évidemment contre-productif pour le moteur.

Enfin, pour finir d'enterrer l'évaluation offline, on pourra mentionner le fait qu'elle coûte très cher, puisqu'il faut, pour la réaliser, entretenir des équipes de quality raters comprenant des centaines d'experts.

C'est pour cela que récemment plusieurs chercheurs ont proposés des méthodes d' « évaluation online », dont nous allons parler maintenant.

Evaluation « online »

Le principe de l'évaluation online est très simple : on fournit à des utilisateurs réels différentes versions des SERP, et on regarde comment ces utilisateurs agissent (les clicks, le temps passé, etc.). A partir des données d'interaction, on doit déterminer quelle est la qualité perçue pour le classement fourni.

Le lecteur intéressé trouvera dans la section « références » quelques articles récents sur le sujet, notamment écrits par Katja Hofmann ou Filip Radlinski, qui sont des experts du sujet et ont mis au point des méthodes très élaborées pour l'évaluation online.

Comme pour son pendant « offline », l'évaluation online fait des hypothèses spécifiques, mais qui sont raisonnables. Ainsi, il est généralement admis que les utilisateurs d'un moteur de recherche ont un but (ils ne cliquent pas sur les liens dans les SERP uniquement pour le plaisir

du clic) et ne tentent pas systématiquement de manipuler le moteur. Il y a donc une grande sincérité dans le processus d'évaluation online : on observe directement le comportement de l'utilisateur, et donc sa satisfaction (ou insatisfaction).

Le problème qui va alors se poser est celui de « l'extraction » de la satisfaction. Sachant que lors d'une évaluation online on ne peut pas interagir avec l'utilisateur, comment va t-on deviner si il est content ou mécontent ? Il va falloir mettre en place un processus de choix basé sur les données disponibles, qui sont essentiellement : les URL présentées, les requêtes tapées et les clics réalisés. Plus précisément, on a accès à la séquence d'utilisation, c'est-à-dire aux URL visitées, au temps de retour sur le moteur, aux requêtes formulées et reformulées. Si on le souhaite (mais c'est plus coûteux), on peut également mettre en place un mécanisme de tracking de la souris pour voir les endroits de clic et de survol.
Ensuite, on va se poser la question de savoir ce que l'on veut exactement :

- Est-ce que je souhaite connaître la valeur de chaque URL classée en telle ou telle position pour une requête donnée, ou bien je ne m'intéresse qu'à la méthode de ranking dans son intégralité, et donc je peux me satisfaire d'un ranking moyen si il est meilleur que les autres ?

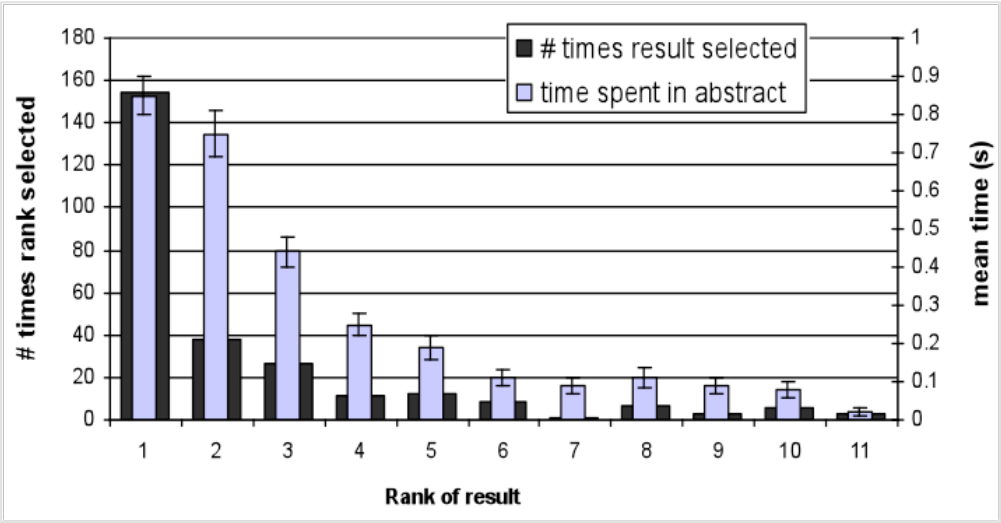
- Est-ce que je veux une évaluation absolue (une note pour chaque document ou algorithme) ou relative (est-ce que ce site-ci est meilleur que celui-là) ?

Selon les choix que l'on fait, ce que l'on va observer est différent, et est résumé dans ce tableau (non exhaustif car nous avons mentionné uniquement ici des métriques simples) :

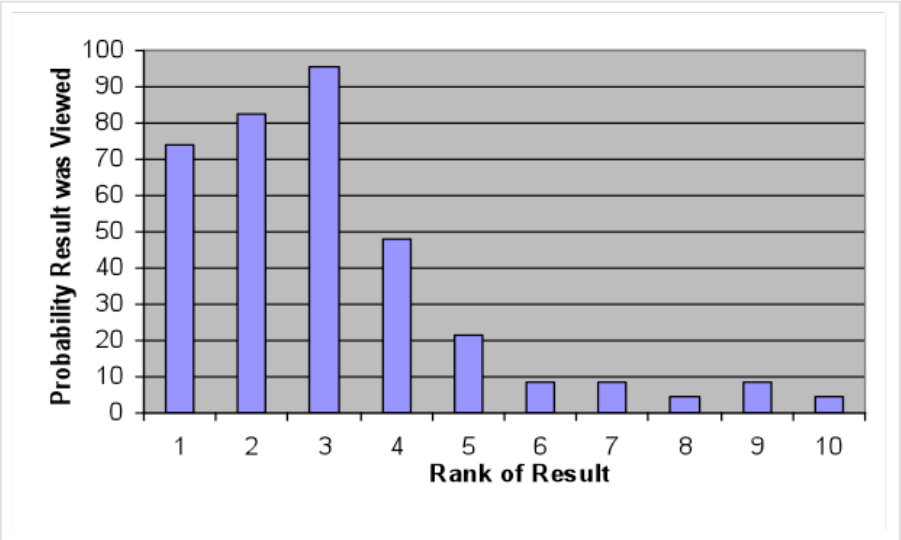
	Evaluation absolue	**Evaluation relative**
Valeur de chaque URL	Taux de click Modèle de la cascade	Click-skip
Méthode de ranking	Abandon Temps pour cliquer Taux de reformulation Clicks par requêtes Temps depuis le dernier click Etc.	entrelacement

Nous allons maintenant discuter de ces différents critères.

- **Valeur de chaque URL (absolue)**. Le critère du taux de clic consiste à compter le nombre de clics sur chaque résultat pour déterminer la valeur de l'URL correspondante. Il s'agit d'un critère difficile à appréhender pour le moteur car il est sujet à de nombreux biais liés au moteur lui même. En effet, la position d'un résultat a une influence sur le taux de clic, indépendamment de sa qualité. On trouve par exemple dans l'article [3] le graphique suivant (il s'agit de l'évaluation d'un moteur de recherche d'articles).



Cette figure montre que le premier résultat est cliqué quasiment 4 fois plus souvent que le second, et que la différence entre le deuxième et le troisième est faible. Si un moteur veut utiliser ce critère, il doit donc prendre en compte ce biais.
Il existe d'autres biais : le contexte et l'apparence des résultats modifient également le taux de clic. Une figure édifiante est la suivante :



On y voit que les utilisateurs d'un moteur de recherche ne regardent pas les résultats au delà du quatrième avant de cliquer sur un des trois premiers !

Pour pallier ces problèmes, on va « mélanger » les résultats dans les classements proposés. Ainsi, une même URL sera tantôt en première position, puis septième, puis deuxième, etc. Ceci permettra de corriger les biais « positionnels ». C'est ainsi que le taux de click moyen deviendra un bon estimateur de la qualité absolue d'un résultat. Le modèle de la cascade est un moyen de réaliser cette correction (nous n'en dirons pas plus sur ce modèle ici).

- **Méthode de ranking (absolue).** Il existe ici plusieurs critères qui ont le même but : déterminer si une méthode de classement est meilleure qu'une autre. Ces critères sont donc globaux et comportementaux : taux d'abandon, taux de reformulation, nombres de requêtes dans une session de recherche, nombre de clics par requête (le fait d'essayer plusieurs résultats pouvant indiquer une insatisfaction), temps entre deux clics, pourcentage de documents non testés, etc. Pour s'assurer que le monitoriing de ces critères fonctionne, il faut garantir certaines propriétés statistiques (avoir des

pools d'utilisateurs représentatifs) et surtout il ne faut pas avoir besoin de comprendre pourquoi un résultat plait ou ne plait pas. On aura ici une indication de la satisfaction, mais pas de sa raison (on ne peut pas vraiment savoir, notamment, si l'utilisateur est insatisfait à cause de la présence d'une URL précise).

Le tableau suivant précise comment on doit conclure selon l'évolution de chaque critère.

Critère	Quand la qualité diminue, le critère...
Taux d'abandon (proportion de requête sans clic)	augmente
Taux de reformulation	augmente
Nombre de requêtes par session de 30 minutes	augmente
Clics par requêtes	diminue
Nombre de secondes avant le premier clic	augmente
Nombre de secondes avant le dernier clic	diminue

Il existe de nombreux autres critères, que l'on peut retrouver dans les différents articles proposés en référence.

- **Valeur de chaque URL (relative).** Le critère principal utilisé ici pour évaluer les qualités relatives entre deux résultats est le « skip ». Un skip est le fait de passer un résultat et de cliquer sur un suivant. Par exemple, si en position 1 on a l'URL toto.com et en position 2 tata.com et qu'on clique directement sur le résultat tata.com, c'est probablement qu'il est de meilleure qualité que toto.com. Le problème de cet indicateur est qu'il ne permet *a priori* que de dire que le 2eme est meilleur que le 1er, car il arrive que l'on clique sur le premier seulement par sa position et non par sa qualité. Pour éviter ce phénomène, on va aléatoirement inverser l'ordre des résultats, pour pouvoir avoir des comparaisons dans les deux sens et donc avoir une conclusion fiable sur la différence de qualité. Bien sûr, le processus devient assez long puisqu'il va falloir construire de nombreuses variantes de chaque classement pour avoir une conclusion sur toutes les URL. On peut alors se demander pourquoi on va mettre cela en place alors que l'évaluation absolue est plus facile à réaliser ? La réponse est tout simplement que l'évaluation relative est plus correcte, elle correspond plus à la réalité de la satisfaction utilisateur (les humains arrivent mieux à comparer qu'à noter, voir [8] par exemple).

- **Méthode de ranking (relative).** Ici on est dans un cadre très balisé : on doit présenter deux ranking aux utilisateurs, mais en en montrant un seul. En effet, l'idée est que l'on veut décider si la méthode A est mieux ou moins bien que la méthode B. Quand on fait une comparaison avec des quality raters, on peut leur montrer deux classements (l'un produit par A, l'autre par B). Mais dans un contexte de monitoring de l'utilisateur, cela n'est pas possible. On doit donc ruser, et pour cela on va créer des classements en entrelaçant des résultats issus de A et de B. Ainsi, selon la perception de chacun de ces classements uniques par des utilisateurs, on pourra reconstruire une préférence pour A ou B. Bien sûr, la réalisation est plus complexe que ce que l'on vient juste de mentionner, et nous renvoyons le lecteur à un post de notre blog [9] pour en savoir (un tout petit peu) plus.

Il faut savoir que l'entrelacement est une méthode qui est fiable et raisonnable (avec environ 10 requêtes on peut différencier assez justement deux méthodes de ranking), mais qu'en revanche, elle fournit des résultats difficilement réutilisables et pas utiles au delà de savoir qu'une méthode de ranking « marche » mieux qu'une autre (pas de connaissances des causes de la différence).

Conclusion

Dans cet article, nous vous avons évoqué les principales méthodes d'évaluation online. Elles ne sont pas encore utilisées de manière systématique pour faire le paramétrage des moteurs de recherche, mais les recherches actuelles commencent à porter leurs fruits. Il y a fort à parier que ces méthodes soient donc de plus en plus utilisées.

Lorsque ce sera le cas, il sera alors temps pour les référenceurs de réfléchir aux meilleurs moyens d'en tirer parti, pour « aider » le moteur à fournir les « bons » classements à ses utilisateurs.

Références

[1] Page Wikipedia en anglais sur la recherche d'information
https://en.wikipedia.org/wiki/Information_retrieval - Performance_and_correctness_measures

[2] Cleverdon, C. (1967, December). The Cranfield tests on index language devices. In Aslib proceedings (Vol. 19, No. 6, pp. 173-194). MCB UP Ltd.

[3] Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, *25*(2), 7.
http://cs303.stanford.edu/papers/joachims_etal_07a.pdf

[4] Radlinski, F., Kurup, M., & Joachims, T. (2008, October). How does clickthrough data reflect retrieval quality?. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 43-52). ACM.
http://clair.si.umich.edu/~radev/767w10/papers/Week09/search_log_mining/Radlinski08.pdf

[5] Joachims, T., & Radlinski, F. (2007). Search engines that learn from implicit feedback. *IEEE Computer*, *40*(8), 34-40.
http://luci.ics.uci.edu/predeployment/websiteContent/weAreLuci/biographies/faculty/djp3/LocalCopy/04292009.pdf

[6] Hawking, D., Craswell, N., Bailey, P., & Griffihs, K. (2001). Measuring search engine quality. Information Retrieval, 4(1), 33-59.
http://comminfo.rutgers.edu/~muresan/IR/Docs/Articles/irHawking2001.pdf

[7] Hofmann, K., Whiteson, S., & de Rijke, M. (2011, October). A probabilistic method for inferring preferences from clicks. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 249-258). ACM.
http://staff.science.uva.nl/~whiteson/pubs/hofmanncikm11.pdf?origin=publication_detail

[8] Carterette, B., & Bennett, P. N. (2008, July). Evaluation measures for preference judgments. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 685-686). ACM.
http://maroo.cs.umass.edu/pdf/IR-665.pdf

[9] Article de notre blog :
http://www.peyronnet.eu/blog/comment-evaluer-automatiquement-la-qualite-des-serps/

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/