Comment le comportement de l'utilisateur permet d'améliorer les résultats de recherche

Par Guillaume et Sylvain Peyronnet

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

On sait depuis longtemps que les moteurs de recherche actuels, comme Google, prennent en compte le comportement de l'internaute lorsqu'il consulte ses SERP (pages de résultats). La notion de « pogosticking », ou action de cliquer sur un lien avant d'éventuellement revenir sur le moteur, est souvent évoqué à ce niveau. Mais ce serait oublier que cette prise en compte comportementale est certainement bien plus complexe que cela et prend en considération bien d'autres critères...

Pour cet article, nous revenons encore une fois sur un point très important pour les moteurs de recherche, et qui reste encore trop méconnu des référenceurs : celui de la prise en compte du comportement de l'utilisateur pour calibrer les résultats fournis pour chaque requête.

Une question « serpent de mer » revient assez souvent dans la communauté SEO : « est-il possible de manipuler le classement en fabriquant du faux comportement positif au niveau des SERP », c'est-à-dire en cliquant abusivement sur ses propres sites, dans l'espoir de faire croire au moteur que ces sites sont particulièrement populaires et pertinents ? La réponse est assez simple : il est très difficile de modifier une SERP ainsi, même si le constat de base est tout à fait exact : les moteurs de recherche modernes prennent en compte le comportement de l'utilisateur pour améliorer les résultats.

Dans cet article nous allons voir que les critères utilisés sont bien plus complexes que le simple taux de clics sur un résultat particulier.

Classement et prise en compte implicite du comportement utilisateur

Nous allons commencer par présenter une approche proposée par Microsoft il y a presque une dizaine d'année (voir l'article [1]). L'idée de cette approche est de créer un modèle du comportement de l'utilisateur, non pas pour évaluer la qualité du classement proposé (ce que nous avons expliqué dans un article précédent portant sur le *monitoring online*), mais plutôt pour fournir un critère de classement supplémentaire à ceux déjà utilisés (tf.idf, pagerank, etc.).

Ces travaux s'inscrivent dans un courant de pensée initié en 2002 par l'article de T. Joachims [2] : la rétropropagation implicite de pertinence (*implicit relevance feedback*). Comme le nom l'indique, il s'agit d'extraire depuis les actions de l'utilisateur des informations permettant d'inférer la qualité perçue du classement proposé, et de la modifier en conséquence.

L'avantage premier de cette approche est qu'elle « passe à l'échelle » : comme elle est basée sur le comportement des utilisateurs réels, il n'y a pas besoin d'entretenir de coûteuses équipes de *quality raters* pour faire tourner la machinerie ! Un autre avantage est que le volume de données est très important naturellement, et plus on a de données plus il y a de chance qu'on réussisse à comprendre quels sont les buts des utilisateurs. Enfin, un tel mécanisme est complémentaire aux autres métriques, il n'y a donc pas à faire mieux que le PageRank ou une autre métrique, il s'agit au contraire de rajouter un moyen supplémentaire pour améliorer les classements.

On se heurte bien sûr à des challenges à la mise en place d'un mécanisme raisonnable et efficace qui utilise les comportements des utilisateurs : il faut tout d'abord bien comprendre ce que l'on veut observer, puis il faut déterminer comment faire une observation « contrôlée », et enfin il faut mettre en place un mécanisme de généralisation : on observe le comportement de quelques utilisateurs sur quelques requêtes, et on extrapole à tous les utilisateurs et toutes les requêtes. Pas si simple...

Modéliser le comportement de l'utilisateur

On va donc tenter de modéliser le comportement de l'utilisateur par des caractéristiques comportementales. Ces caractéristiques peuvent être de deux types : les premières sont des valeurs observées, les deuxièmes sont des critères « distributionnels ».

Une valeur observée est très simple à définir : c'est littéralement la valeur numérique mesurée associée à un comportement. Par exemple, on peut mesurer le temps qu'il a fallu à un internaute pour cliquer sur une des pages proposées par un moteur lorsqu'il tape une requête. De la même façon, on peut mesurer le pourcentage de personnes qui cliquent sur le quatrième résultat pour la requête « machine à pain ». Les valeurs observées sont naturellement les critères les plus simples à obtenir, et sont les plus intuitifs.

Un critère distributionnel est plus difficile à appréhender, mais amène finalement beaucoup plus d'informations. L'idée pour ce type de critère est de quantifier la déviation par rapport à un comportement attendu (au sens statistique du terme). Par exemple, si on a mesuré que pour la requête « belle loutre », l'utilisateur moyen clique sur le premier résultat après 3,7 secondes, alors une page cliquée après 4 secondes en moyenne sur la même requête avec la même position à une valeur de -0,3 pour le critère distributionnel associé.

Le deuxième point important est bien sûr le choix des caractéristiques, que nous verrons plus tard, mais dont il nous faut dire maintenant qu'elles sont étroitement associées au comportement de l'utilisateur, et que ce comportement peut se découper en trois phases :

1. **Présentation**. Entre le moment où l'utilisateur tape sa requête et le moment où il va cliquer sur l'un des résultats proposés par le moteur, se déroule le temps de présentation. Certaines caractéristiques permettent de modéliser ce que l'utilisateur voit avant de cliquer.
2. **Action**. L'action typique de l'utilisateur est de cliquer sur l'un des résultats. On

peut mesurer cette action en terme de fréquence, mais aussi de temporalité.

3. Browsing. Une fois que l'utilisateur a cliqué sur un lien, il va browser la page ciblé par le lien. On peut également définir des caractéristiques comportementales sur ce moment post-clic.

Voyons maintenant une sélection des caractéristiques les plus utilisées par les moteurs, avec une courte description à chaque fois :

Phase de présentation :

- *Position*. Il s'agit de la position de la page en question dans le classement proposé par le moteur (la SERP).
- *Recouvrement titre vs requête*. Combien y a-t-il de mots en commun entre la requête et le Title de la page ?
- *Recouvrement snippet vs requête*. Combien y a-t-il de mots en commun entre la requête et le snippet de la page ?
- *Longueur*. Quelle est la longueur de la requête en nombre de mots ?
- *Recouvrement avec le futur*. Quel est le nombre de mots en commun entre la requête actuelle et la prochaine requête que va taper l'utilisateur ? Cette caractéristique est très importante puisqu'elle va donner une information de reformulation : si la valeur est grande, cela veut dire que le même besoin informationnel est reformulé, et donc que l'utilisateur n'est pas satisfait de ce qu'il a vu.

Phase d'action :

- *Délibération*. Il s'agit du temps écoulé entre le moment où l'utilisateur tape la requête et le moment où il va cliquer sur un des résultats.
- *Fréquence*. La fraction des clics (sur un ensemble d'utilisateurs) qui portent vers une page spécifique pour une requête donnée.
- *Déviation*. Il s'agit d'un critère distributionnel qui quantifie la différence par rapport au taux de clic moyen.
- *Clic dessus/dessous*. Est-ce que le résultat du dessus ou du dessous à été cliqué ?

Phase de browsing :

- *Dwell time*. Il s'agit du temps passé par l'utilisateur sur la page. Assez difficile à quantifier (sauf à monitorer totalement le comportement de l'utilisateur, ce que certains moteurs peuvent faire *via* des informations par un navigateur propriétaire, ou par un surf en mode connecté).
- *Déviation au dwell time*. Critère distributionnel qui mesure l'écart par rapport au temps passé sur la page moyen. Cela permet de mesurer la qualité : si le temps passé est plus grand, cela veut sans doute dire que la page cible est de grande qualité.
- *Re-clic*. Si l'utilisateur continue son surf à partir de la page qu'il a cliqué, cela signifie que celle-ci lui fournit des informations intéressantes qu'il veut continuer à explorer.

Il existe bien sûr de très nombreuses autres caractéristiques, mais celles mentionnées ici sont intuitives et suffisent à comprendre l'idée.

De l'observation à la prédiction

Une fois qu'un modèle est défini, il est possible de récolter des données en observant le comportement des internautes lors de leur interaction avec le moteur. Il se pose alors la question de ce que l'on peut faire avec ces données. Il existe essentiellement deux possibilités.

La première consiste à utiliser les informations comportementales pour créer un classement indépendant de celui que les algorithmes du moteur vont créer par ailleurs, puis de substituer (ou fusionner avec) ce classement indépendant à celui déjà existant.

L'autre possibilité est d'utiliser un score lié au comportement utilisateur comme l'un des (nombreux) critères de classement : on obtiendra ainsi un modèle utilisateur qui agira au même titre que le PageRank, TrustRank, la pertinence, etc. Cette dernière approche est la plus efficace mais également la plus difficile à mettre en place puisqu'il s'agit de modifier en profondeur le fonctionnement du moteur de recherche.

On notera que pour réussir à prédire ce qu'il va se passer globalement à partir de données observées (au niveau d'un moteur, les données observées vont représenter une minorité du volume de requête), il faut utiliser des techniques issues du machine learning, comme celles présentées dans l'article [3]. On notera d'ailleurs que les deux premiers auteurs de cet article (Burges et Shaked) ont déposé un brevet portant sur cette méthode [4].

Résultats expérimentaux

Nous allons maintenant voir si les SERP sont améliorées par l'utilisation de ces informations de comportement utilisateur. Les auteurs de l'article [1] ont utilisé plusieurs métriques, dont la précision à k (p@k). Pour rappel, La p@k est maximale quand un expert estime que les k premiers résultats d'un classement méritent d'être à cette place.

Les auteurs de cette recherche ont testé l'incorporation d'information comportementale dans plusieurs contextes. Cette information provient de l'observation de 8 semaines de production sur un moteur majeur, avec l'enregistrement de millions de requêtes uniques et des interactions associées.

Deux grandes expériences ont été effectuées : tout d'abord le test face à un moteur qui n'utiliserait que le contenu des pages pour faire le classement (avec l'algorithme BM25), puis ensuite le test face à un moteur réel, avec un très grand nombres de critères.

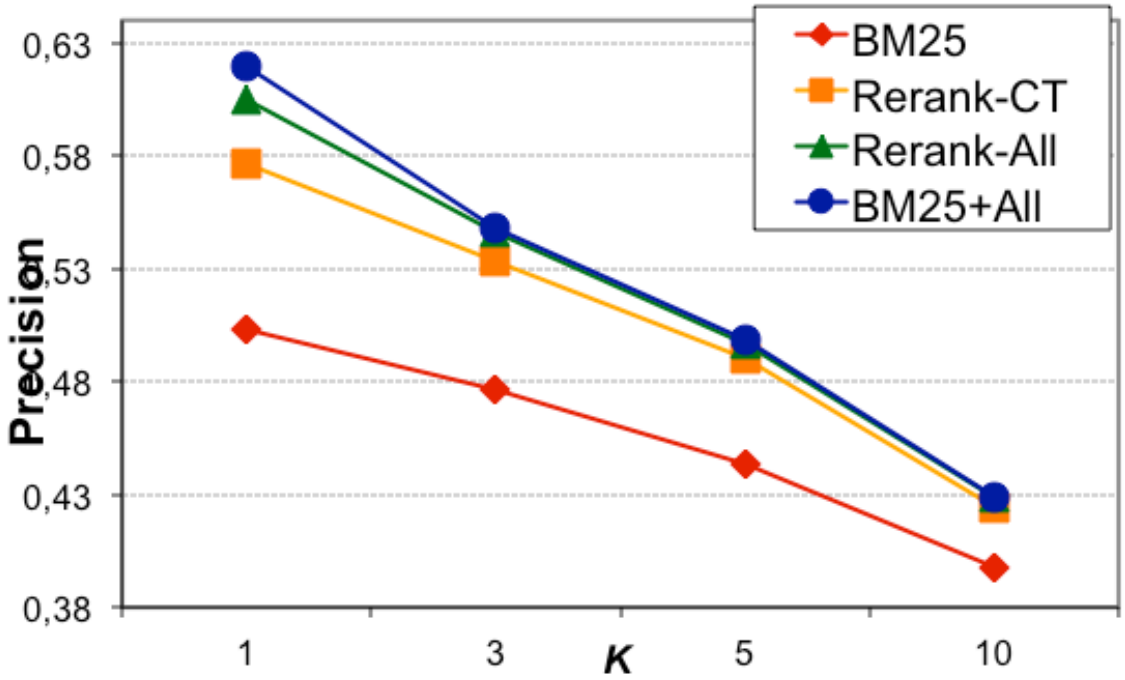

Fig.1. Expérimentation face à BM 25

1er test – face à BM25

La figure 1 montre ce qui est obtenu. Dans cette figure, BM25 est l'étalon de comparaison, Rerank-CT est le classement obtenu avec l'information liée à la phase d'action uniquement, Rerank-all est le classement comprenant les comportements des trois phases, et enfin BM25+ALL est l'utilisation conjointe de BM25 et de l'information comportementale. Il est important de notant que Rerank-XX est un classement obtenu sans même regarder le contenu des pages (uniquement le comportement des utilisateurs) !
On voit très clairement que l'utilisation d'information comportementale améliore dans tous les cas la qualité perçue. On voit aussi que les trois classements avec information comportementale sont très proches les uns des autres.

2eme test – face à un moteur complet

Sur la figure 2, on voit une plus grande différenciation. Le moteur classique est RN, Rerank-All est toujours le classement comprenant les comportements des trois phases, et enfin RN+All est un nouveau moteur qui utilise comme critère supplémentaire de classement le modèle utilisateur en trois phases.

Très clairement, la qualité est la meilleure pour ce nouveau moteur avec une amélioration autour de 15% pour la p@10. Il est donc intéressant d'incorporer cette information au niveau du moteur.

Cependant, une autre information est particulièrement importante, indiquée sur la figure 3.

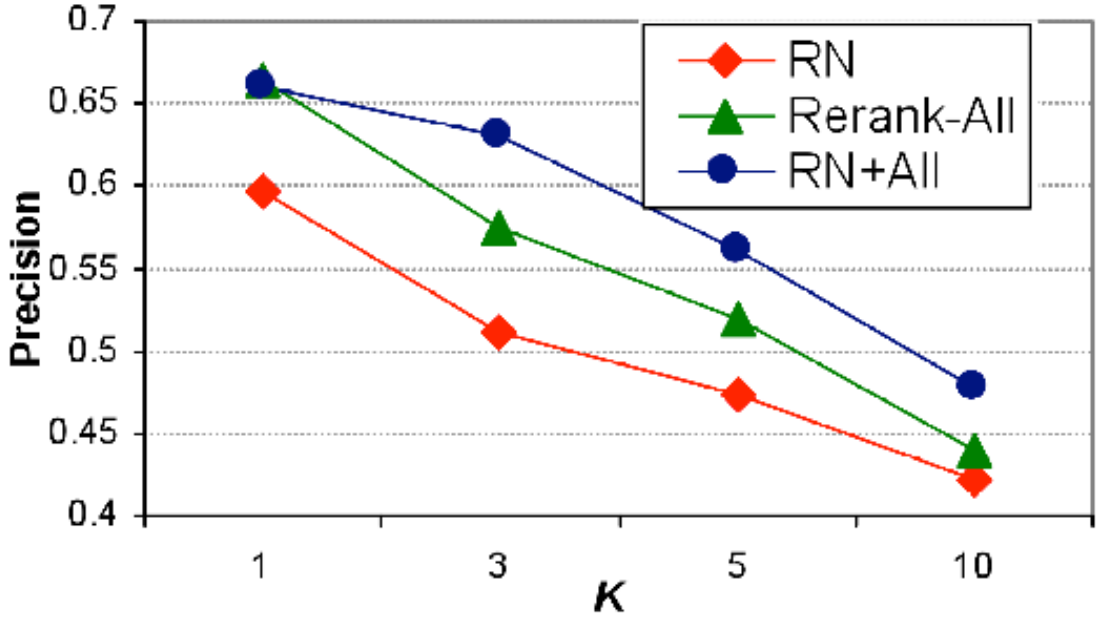

Fig.2. Expérimentation face à un moteur complet

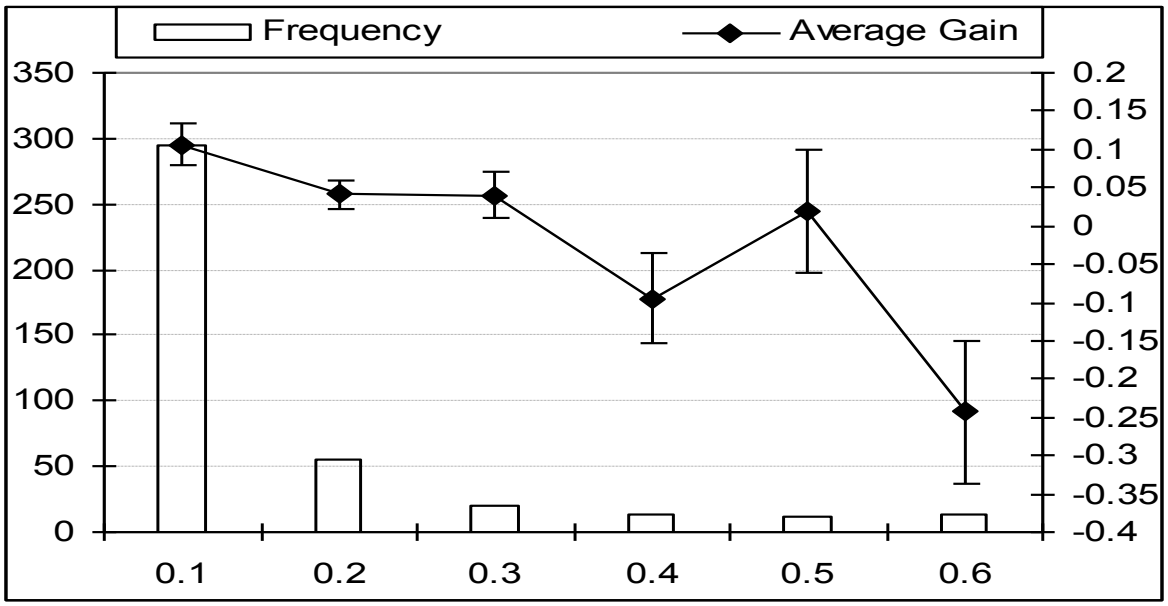

Fig.3. Qualité moyenne (MAP) des requêtes

Sur ce graphique on voit en abscisses la qualité moyenne (métrique MAP) pour un certain nombre de requêtes (ordonnées de gauche). Par exemple, on trouve dans cette expérience presque 300 requêtes avec une MAP de 0,1 (valeur indiquant une mauvaise qualité) alors qu'il y en a environ 15 avec une MAP de 0,6 (bonne qualité). La courbe en noir indique le gain moyen obtenu en utilisant l'information comportementale. Et ce que l'on voit est édifiant : utiliser le modèle d'interaction améliore fortement les résultats sur les requêtes qui étaient problématiques, mais dégrade la qualité sur les requêtes qui étaient bien traitées.

Ce résultat est particulièrement important car il montre qu'il n'existe pas de corrélation entre les résultats standards du moteur et ceux réalisés par un méthode

utilisant l'information comportementale. On peut donc raisonnablement espérer qu'avec un bon choix du ratio d'utilisation de cette information, on aboutira à une bonne amélioration des SERP. Pour cela, il faudra cependant rajouter une brique algorithmique pour prédire quelles sont les requêtes bien traitées, et quelles sont les autres.

Aller plus loin : les nouvelles métriques

Les résultats que nous avons vu jusqu'ici datent de 2006, et depuis, de nombreuses autres recherches ont été menées, qui pour la plupart confirment ce qui a été dit plus haut, en ajoutant également de nouveaux critères comportementaux.

Ainsi, Dupret et Lalmas rajoutent aux critères connus le taux d'abandon, c'est-à-dire les sessions de recherche qui n'ont donné lieu à aucun clic dans les SERP. Il ajoute également une notion plus complexe : le temps d'absence. Il s'agit d'une mesure très bruitée qui capte le temps qui s'est écoulé entre deux visites à un site web. La mesure n'est pas simple à comprendre : si le temps d'absence est court, il s'agit d'un point positif (le service plait) sauf si on détecte un retour associé à une reformulation (*« je reviens pour reformuler, parce que je ne suis pas content de ce que l'on m'a donné »*), si il est long, il s'agit d'un point négatif. Mais il existe un bruit naturel : le temps d'absence peut être long car l'utilisateur est parti faire autre chose… Pour réussir à prendre en compte cette nouvelle métrique, il faut l'associer à des indicateurs statistiques robustes, ce qui complique la tâche du moteur.

Conclusion

Nous avons vu dans cet article que prendre en compte l'information comportementale liée à l'interaction entre l'utilisateur et le moteur permet d'améliorer les SERP. On a vu également qu'un acteur comme Microsoft à breveté des algorithmes pour cela… A bon entendeur…

Références

[1] Agichtein, Eugene, Eric Brill, and Susan Dumais. "*Improving web search ranking by incorporating user behavior information.*" Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006.
http://www.msr-waypoint.com/en-us/um/people/sdumais/SIGIR2006-fp345-Ranking-agichtein.pdf

[2] Joachims, Thorsten. "*Optimizing search engines using clickthrough data.*" Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.
http://ebook-rush.googlecode.com/svn/trunk/Click-Model/optimizing_search_engines_using_clickthrough_data.pdf

[3] Burges, Chris, et al. "*Learning to rank using gradient descent.*" Proceedings of the 22nd international conference on Machine learning. ACM, 2005.
http://machinelearning.wustl.edu/mlpapers/paper_files/icml2005_BurgesSRLDHH05.pdf

[4] http://www.google.com/patents/US7689520

[5] Dupret, Georges, and Mounia Lalmas. "*Absence time and user engagement: evaluating ranking functions.*" Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013. http://www.dcs.gla.ac.uk/~mounia/Papers/wsdm2013.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/*