Quelques projets de recherche et développement chez Qwant

Par Guillaume et Sylvain Peyronnet

Domaine :	Recherche	Référencement
Niveau :	Pour tous	Avancé

Guillaume et Sylvain Peyronnet font partie de l'équipe qui travaille sur les projets de recherche et développement du moteur de recherche Qwant. Dans cet article, ils nous dévoilent quelques-uns de leurs travaux pour faire en sorte qu'un moteur ayant comme ligne de conduite de ne pas utiliser les données privées de ses utilisateurs soit aussi pertinent qu'un Google qui puise abondamment dans ce type de data. Pertinence, lutte contre le spam, réseaux sociaux et filtre familial sont au rendez-vous. Entre autres sujets de R&D mélant algorithmique et machine learning...

L'article que nous écrivons ce mois-ci est un peu particulier. En effet, nous allons vous évoquer une partie des problématiques de recherche qui sont abordées chez Qwant, le moteur de recherche européen. Il s'agira donc d'un article tourné en mode « première personne » puisque nous sommes largement impliqués dans les opérations de recherche et développement du moteur. Cette implication concerne de notre coté plusieurs personnes, dont principalement Guillaume et Sylvain Peyronnet et Thomas Largillier. Il y a ensuite plusieurs ingénieurs de R&D dans les équipes de développement du moteur (la quasi totalité de ces équipes se trouvent dans les bureaux Niçois de Qwant). Par ailleurs, quelques personnes basées sur Paris contribuent à la recherche en *machine learning* pour l'amélioration des résultats. Enfin, l'un d'entre nous (Sylvain) est « chief scientist » du moteur.

La R&D pour un moteur de recherche, qu'est ce que c'est ?

Au delà des domaines en eux-mêmes (sécurité, machine learning, UX, infrastructure, bases de données, etc.), la recherche au sein d'un moteur (de recherche) a quelques caractéristiques génériques. Tout, d'abord, cette recherche est **appliquée**. Même si il est tout à fait possible d'aborder des sujets complexes et sur un terme assez long, la plupart des sujets étudiés correspondent à des problèmes réels, et l'objectif est donc de trouver des solutions pratiques dans un délai assez court. Cela peut sembler être une porte ouverte largement enfoncée, mais cela a en fait une incidence très forte sur les aspects théoriques : la plupart des algorithmes mis au point font partie d'une vision approchée et on-line (on-line signifie ici qu'on a le droit de lire une seule fois les données pour prendre une décision).

Ensuite, il existe trois grands types de sujets de recherches :

1. Les sujets qui ont pour enjeu la création de nouveaux algorithmes basés sur de nouveaux principes. L'exemple typique est la sortie du PageRank en 1999 (voir l'article [1]) : un nouvel algorithme, qui propose une solution complètement différente pour résoudre un problème bien connu. Il s'agit d'une **rupture algorithmique**.

2. Les sujets qui ont pour but de réaliser des **améliorations drastiques de tâches déjà effectuées de manière efficace par d'autres algorithmes**. Chez Qwant, nous avons par exemple mis au point un algorithme qui fait du calcul de popularité des pages 20 à 30 fois plus vite que le PageRank. L'impact principal de la création de ce type d'algorithme est de passer la barrière du coût : on peut obtenir une qualité de résultats similaires aux concurrents, mais avec une infrastructure matérielle moins coûteuse.

3. Les **sujets exploratoires**. Pour ces derniers, l'objectif est de créer des nouveaux services, basés sur de nouveaux algorithmes. On trouve actuellement par exemple beaucoup de travaux sur l'analyse des réseaux sociaux, qui restent exploratoires car les services commercialisés dans le domaine sont assez pauvres (Klout, Klear, etc.).

Chez Qwant, les trois types de recherche coexistent, et même si il y a forcément une nécessité à avancer le plus vite possible vers des algorithmes efficaces mis en production, c'est la coexistence des différents rythmes de recherche qui permet d'avoir des bonnes idées, et des bons résultats. La R&D n'est pas une tâche de développement toute bête, le cycle de réalisation qui va de l'idée nouvelle jusqu'à la mise en production d'un algorithme prend de 6 mois à 1 an au minimum, parfois beaucoup plus.

Les algorithmes qui sont décrits dans cet article sont à divers stades de ce cycle : certains sont au niveau de la mise en production, d'autres sont évidemment encore embryonnaires (existence d'un prototype par exemple).

Est-ce qu'il y a une spécificité Qwant en matière de R&D ?

Pour faire court, la réponse est OUI. Les fondateurs de Qwant ont fait un choix idéologique : le respect de la vie privée, qui a un impact extrêmement important sur l'algorithmique du moteur. La situation est très simple, sur le Web, la plupart des algorithmes qui ont été conçus jusqu'ici partent du principe que l'on sait tout sur les utilisateurs. Qwant se place en opposition : on ne sait rien sur chaque utilisateur en particulier (on ne sait même pas qui il est, ni si un visiteur au moment M était déjà visiteur auparavant).

Cela ne signifie pas qu'on ne peut rien faire, car contrairement à ce que l'on voudrait nous faire croire, on peut faire la plupart des tâches avec des algorithmes qui respectent la vie

privée. Simplement, ces algorithmes doivent être parfois pensé différemment, ce qui induit de nouveaux résultats théoriques, et ensuite une réalisation pratique sans faille.

Quelques exemples de sujets abordés…

Nous allons maintenant voir quelques-unes des problématiques qui sont abordées par l'équipe R&D de Qwant.

Trouver les pages de qualité

La priorité pour un moteur est d'abord de trouver les pages de qualité. Dans cette objectif, il faut travailler sur le dyptique popularité-pertinence. Pour la totalité des moteurs modernes, la popularité vient d'un modèle de surfeur aléatoire (ce qui définit la notion de PageRank par exemple). Qwant ne fait pas exception à la règle, il faut bien un algorithme de calcul de popularité pour réussir à fournir un bon classement. Nous avons développé notre propre algorithme, nommé **Iceberg**, qui classe les pages par importance, en utilisant les propriétés structurelles du graphe du web. Cet algorithme est extrêmement rapide : un gain de performance de 20% est atteint par notre prototype.

Par ailleurs, nous travaillons à la mise en place de filtres déterminant la qualité des pages web. Il s'agit principalement d'algorithmes de classification supervisé, qui analysent certains critères objectifs des pages web et qui ont appris les avis de *quality raters* pour prédire le niveau de qualité d'un site web. Ici il s'agit d'une recherche balisée, dont l'enjeu est principalement d'obtenir la meilleure qualité et efficacité. Nous avons déjà évoqué dans cette lettre l'étude webspam que nous avons finalisée il y a peu. La création d'un tel filtre suit la même méthodologie, c'est la base de critères et le *quality rating* qui va être différent pour la passage à l'échelle d'un moteur de recherche.

Enfin, la combinaison de tous les critères et le choix des meilleures pages se fait via des algorithmes dits « top-k ». Qwant a en interne des spécialistes de ce type de méthode, dont on peut trouver le principe dans l'article scientifique [8].

La lutte contre le webspam

Les lecteurs de la lettre d'Abondance son bien placés pour le savoir : il y a des acteurs du web qui « luttent » contre le moteur, avec des techniques plus ou moins agressives.

Dans le milieu du SEO, on trouve aux extrêmes des personnes qui vont générer automatiquement du contenu et des liens pour améliorer les classements de leurs sites web. Détecter les manipulations de ces personnes est une nécessité pour un moteur de recherche.

Il existe deux familles d'algorithmes pour cela : les algorithmes de détection (des filtres dans le vocabulaire SEO, comme Penguin et Panda, par exemple) et les algorithmes de déclassement (des « updates » du noyau du moteur). Comme leur nom l'indique, les algorithmes de détection détectent les « mauvaises » pratiques pour les pénaliser, tandis que le déclassement consiste à annuler l'effet de ces pratiques, sans les détecter.

Nous travaillions sur la lutte contre le webspam bien avant l'existence de Qwant (voir les articles [5], [6] et [7]) et ce travail se poursuit au sein des équipes du moteur.

Filtrer les contenus pour adultes

Nos équipes ont beaucoup travaillé aux filtrages de pages web pour créer Qwant Junior. Il existe de nombreux mécanismes actifs pour réaliser un moteur pour les enfants, et parmi ceux-ci, un filtre pour détecter les contenus pornographiques.

Nous sommes très fiers du résultat obtenu, et un article de recherche (référence [3]) décrivant le filtre a été publié à la conférence internationale SEXI 2016 (la bien nommée…). Ce filtre consiste en une forêt de classifieurs qui étudient le contenu textuel de chaque page. Les résultats sont impressionants : nous avons une erreur globale de 2.15% et les seuls sites qui passent à travers le filtre sont détectables autrement car il s'agit principalement de sites qui n'ont pas de contenu textuel.

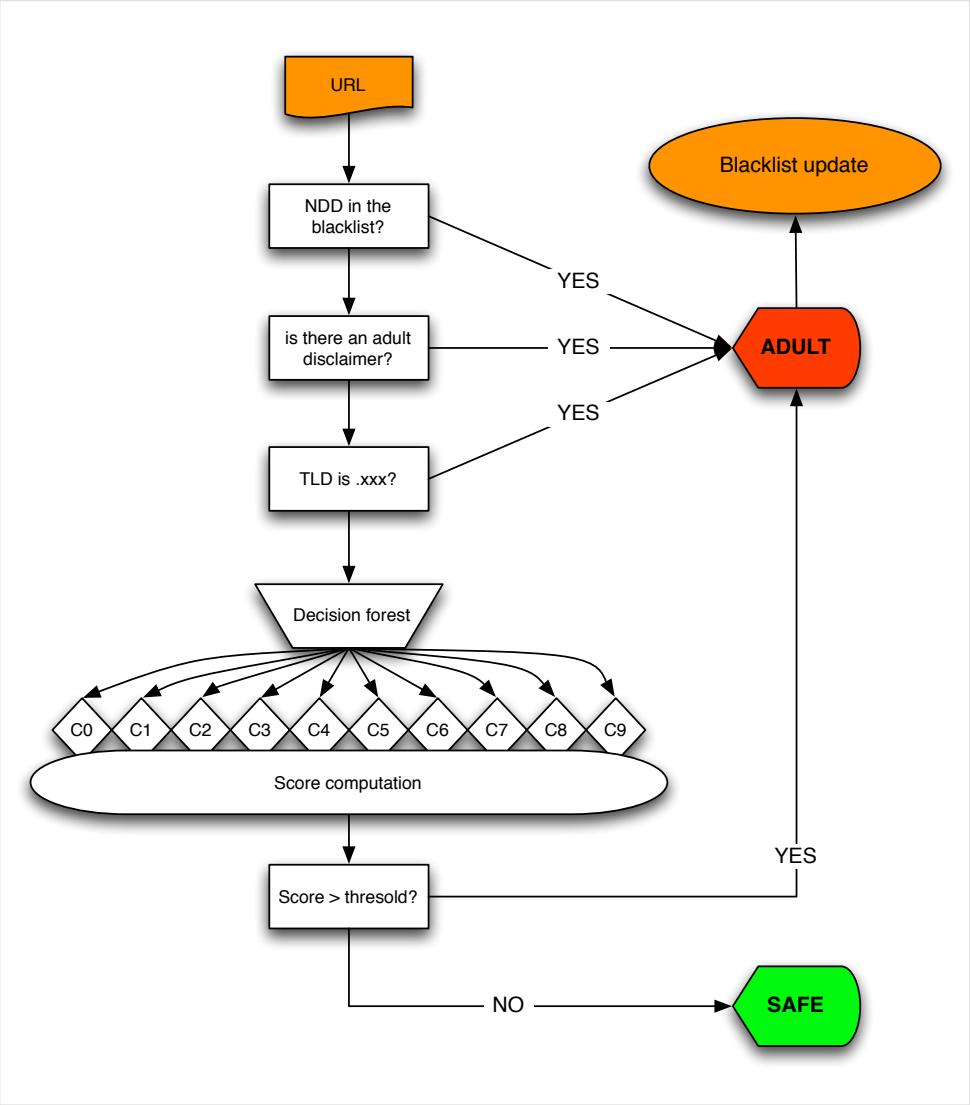

Fig 1 : principe du filtre anti-pornographie

La Figure 1 présente le schéma de fonctionnement du filtre. Chaque losange C0, C1, etc. correspond à un mécanisme de classification supervisé indépendant des autres. La décision finale est donc une agrégation des décisions prises par chaque algorithme. Certains des classifieurs sont très simples et se contentent de regarder 2 critères (par exemple le nombre d'images sur la page et la présence de noms de stars du porno) tandis que d'autres sont complexes et analysent plus d'une dizaine de critères différents. Le principal enseignement de cette recherche au-delà de la réalisation pratique, est que l'on peut filtrer la plupart des sites pornographiques sans analyser le contenu graphique (images et vidéos). Cette conclusion a d'ailleurs été une surprise pour nos confrères lors de la conférence SEXI.

Analyser les réseaux sociaux

Plus haut dans cet article nous avons mentionné que certaines recherches sont exploratoires. C'est le cas par exemple de nos travaux sur les réseaux sociaux. Nous nous intéressons particulièrement à deux sujets : la quantification du niveau d'influence d'une personne sur les réseaux sociaux, et la détection de communautés de personnes partageant des centres d'intérêt communs. Sur ce deuxième problème, nous avons déjà obtenus de beaux résultats théoriques, dont deux ont été publiés dans des conférences internationales (voir les articles [2] et [4]) et un troisième va bientôt l'être également.

Nous avons ainsi mis en place un algorithme qui détecte les communautés grâce à des parcours aléatoires du graphe des connexions entre personnes. Les résultats sont bons, mais pas encore assez à notre goût, et nous souhaitons donc continuer à explorer de nouvelles pistes avant de mettre en production quoi que ce soit sur le sujet. Sur le sujet de l'influence, c'est un peu la même chose, pour l'instant on travaille, mais nous ne faisons pas mieux que la compétition, il n'y a donc aucune raison d'en parler plus que cela.

Et bien d'autres choses encore…

Les équipes travaillent sur bien d'autres sujets. C'est ainsi qu'un petite *task force* s'est constituée à cheval sur les équipes de Paris et de Nice pour le travail sur la partie Actualités. En effet, les actualités sont un domaine très spécifique : les articles sont courts, ils contiennent beaucoup d'entités nommées (noms de personnes, lieux, entreprises, etc.). Le contenu a une « date de péremption » et est très largement en duplication entre les différents sites de presse. Il faut donc des algorithmes très différents, ne serait-ce que parce qu'on doit analyser très vite ce type d'article. Nous avons déjà des algorithmes de clustering (mise en thématiques) très efficaces qui ont été crées par l'équipe de Nice, mais le travail ne s'arrête pas là, car on peut toujours faire mieux.

D'autres choses encore ? Oui bien sûr, mais laissons planer un peu de mystère… Et nous aurons certainement l'occasion de vous en parler bientôt...

Conclusion

Comme les plus grands moteurs, Qwant s'est doté d'une équipe de recherche et développement, certes de dimension encore modeste, mais capable de contribuer à l'état de l'art dans le domaine de la recherche d'information « adversarial » (c'est comme cela que l'on appelle la recherche d'information au sein d'un index qui peut être manipulé par des acteurs externes). Les contributions se situent principalement en algorithmique et en utilisation du machine learning, et nous espérons bien que cela ne s'arrêtera pas de sitôt, avec des challenges de plus en plus excitants !

Références

[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: bringing order to the web.*

[2] Creusefond, J., Largillier, T., & Peyronnet, S. (2016). *On the evaluation potential of quality functions in community detection for different contexts.* In Advances in Network Science (pp. 111-125). Springer International Publishing.
http://arxiv.org/pdf/1510.01714.pdf

[3] Largillier, T., Peyronnet, G., & Peyronnet, S. (2015). *Efficient filtering of adult content using textual information.* arXiv preprint arXiv:1512.00198.
http://arxiv.org/pdf/1512.00198.pdf

[4] Creusefond, J., Largillier, T., & Peyronnet, S. (2015, August). *Finding compact communities in large graphs.* In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (pp. 1457-1464). ACM.
http://arxiv.org/pdf/1410.2105.pdf

[5] Largillier, T., Peyronnet, G., & Peyronnet, S. (2014, March). *Rocovo: robust communal publication scheme.* In Proceedings of the 29th Annual ACM Symposium on Applied Computing (pp. 579-586). ACM.
https://hal.archives-ouvertes.fr/hal-01081279/document

[6] Largillier, T., & Peyronnet, S. (2012). *Webspam demotion: Low complexity node aggregation methods.* Neurocomputing, 76(1), 105-113.
http://www.berbiqui.org/thomas/neucom.pdf

[7] Largillier, T., & Peyronnet, S. (2011). *Detecting Webspam Beneficiaries Using Information Collected by the Random Surfer.* International Journal of Organizational and Collective Intelligence (IJOCI), 2(2), 36-48.
http://www.berbiqui.org/thomas/ijoci0203-03.pdf

[8] Badr, M., & Vodislav, D. (2015). *Exact and Approximate Generic Multi-criteria Top-k Query Processing.* In Transactions on Large-Scale Data-and Knowledge-Centered Systems XVIII (pp. 53-79). Springer Berlin Heidelberg.

[9] https://www.qwantjunior.com/?l=fr

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/*