## Spam de contenu : comment le reconnaître ? Comment le détecter ?

Retour au sommaire de la lettre

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

*De nombreuses méthodologies sont utilisées par les moteurs de recherche actuels pour détecter le spam dans leur index et produire ainsi des résultats de la meilleure qualité possible. Parmi celles-ci, la notion de "classifieur" est fondamentale. Cet article détaille une étude et un algorithme détaillés il y a quelques années à cette fin, ainsi qu'un projet actuellement en cours et qui permettra d'avoir, d'ici quelques mois, plus d'informations précises à ce sujet, ainsi que des outils pour définir le niveau de qualité d'une page web. Passionnant...*

Nous allons aborder aujourd'hui la notion de spam de contenu, ce que l'on appelle souvent "webspam" dans la littérature scientifique. Sans aucune surprise, la détection du webspam est un enjeu important pour les moteurs de recherche. L'objectif de cet article est de définir ce qu'est le spam de contenu et de donner les clés permettant de comprendre comment un moteur va le détecter en utilisant des outils de classification.

Nous décrirons ensuite l'étude du projet que nous réalisons actuellement, et qui a pour objectif de mettre en évidence les critères constitutifs du spam de contenu, pour ensuite créer un outil de qualification de contenu.

## Qu'est-ce que le spam de contenu ?

La formule, amusante mais convenue, nous le dit : « *webspam, you know it when you see it* ». Dans le folklore scientifique, il n'y a pas donc pas de vraie définition formelle du spam de contenu. Au final, on s'aperçoit généralement que le spam de contenu se traduit par une page web de faible qualité, dont le texte est réalisé au kilomètre (automatiquement ou à la main) et qui est "moche" (template de faible qualité graphique). Mais ceci n'est pas une réelle définition, et pour le chercheur, c'est la qualification par des quality raters qui fera foi : une page est du spam si elle a été notée « spam » par des êtres humains.

Et là apparaît un petit drame, car les humains ne sont pas toujours d'accord entre eux. Bien sûr, pour les pages contenant du texte généré par des méthodes automatiques de base, tout le monde est d'accord sur la qualité du contenu. Mais il existe de nombreux cas très *borderline*, et souvent, seuls les initiés qui connaissent les astuces y verront du spam, tandis que l'internaute lambda n'y trouvera rien à redire. C'est par exemple le cas des sites qui vont faire une copie de Wikipedia en y rajoutant quelques publicités.

Bref, la détermination de ce qui est du spam de contenu est réalisé par le monitoring du comportement humain. Ce monitoring peut être explicite ou implicite. Explicite quand on demande à des "quality raters" de noter des pages web, ou implicite quand le moteur va regarder le comportement de ses utilisateurs (retour "post click", temps passé sur une page, etc.).

## Que faire quand un ensemble de pages ont été qualifiées ?

Avoir des pages qualifiées c'est bien, mais ce n'est pas un aboutissement acceptable pour un moteur de recherche. En effet, il est impossible de faire qualifier toutes les pages du web. Pour pallier ce problème, le moteur n'a d'autre choix que d'extrapoler la qualification de tout son index à partir de la qualification d'un petit ensemble de pages web.

Est-ce qu'il est possible, et facile, d'extrapoler ? La réponse est deux fois « oui ». C'est possible grâce aux techniques mises au point en apprentissage (le fameux *Machine Learning* qu'on retrouve à toutes les sauces dans le discours de Google, Facebook et consorts). Et c'est raisonnablement facile en utilisant des techniques issues de la classification supervisée, comme par exemple l'algorithme C4.5 de Ross Quinlan.

Qu'est ce que la classification supervisée ? Le problème de la classification est celui de faire la correspondance entre des objets en cours d'analyse et des catégories dans lesquelles on cherche à ranger ces objets. Un programme qui réalise cette identification s'appelle un "classifieur". On parle de classification supervisée quand la méthode de classification est initialisée avec des données qualifiées.

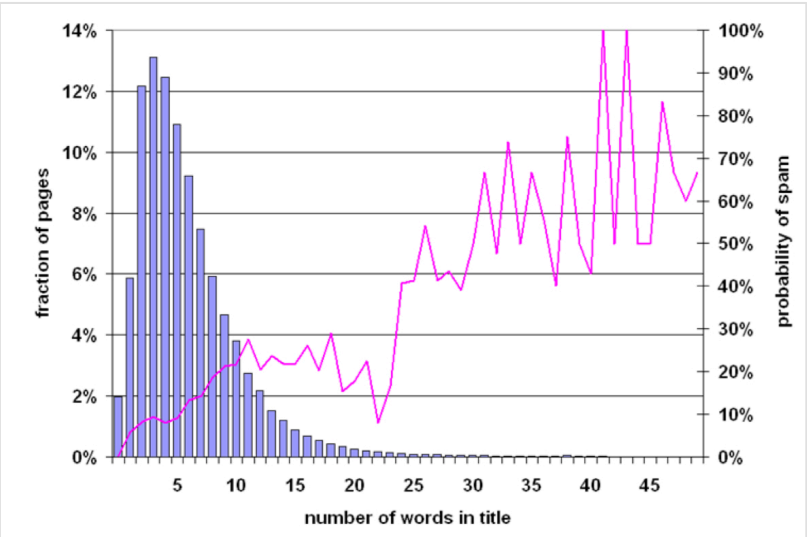
Pour revenir à notre question de base : une fois un ensemble de pages qualifié, on va calculer un certain nombres de statistiques sur des critères choisis de la page. Puis, grâce à l'algorithme C4.5 (ou un autre plus évolué), on va créer un classifieur. Dans notre cas, ce classifieur sera un arbre de décision. C'est-à-dire un ensemble de règles permettant de confronter dynamiquement une page en cours d'analyse avec certains critères (pas forcément tous, l'enjeu étant de diminuer le nombre moyen de critères utilisés pour chaque page), pour déterminer si une page est du spam ou du contenu de qualité.

Une fois ceci réalisé, on s'assurera que les résultats obtenus sont raisonnables, et notamment que l'on n'a pas trop de faux positifs. Un faux positif étant une page qui a été considérée comme du spam par le classifieur, alors qu'en fait elle n'en est pas.

## *Le cas d'école : l'étude de Ntoulas, Najork, Manasse et Fetterly*

Nous allons maintenant vous parler d'une étude scientifique réalisée il y a plusieurs années (en 2006) et publiée (en partie) dans l'article de Ntoulas et al. (voir les sources à la fin de cet article). Cette étude est ancienne, et son seul intérêt de nos jours est qu'il s'agit d'un bel exemple méthodologique.

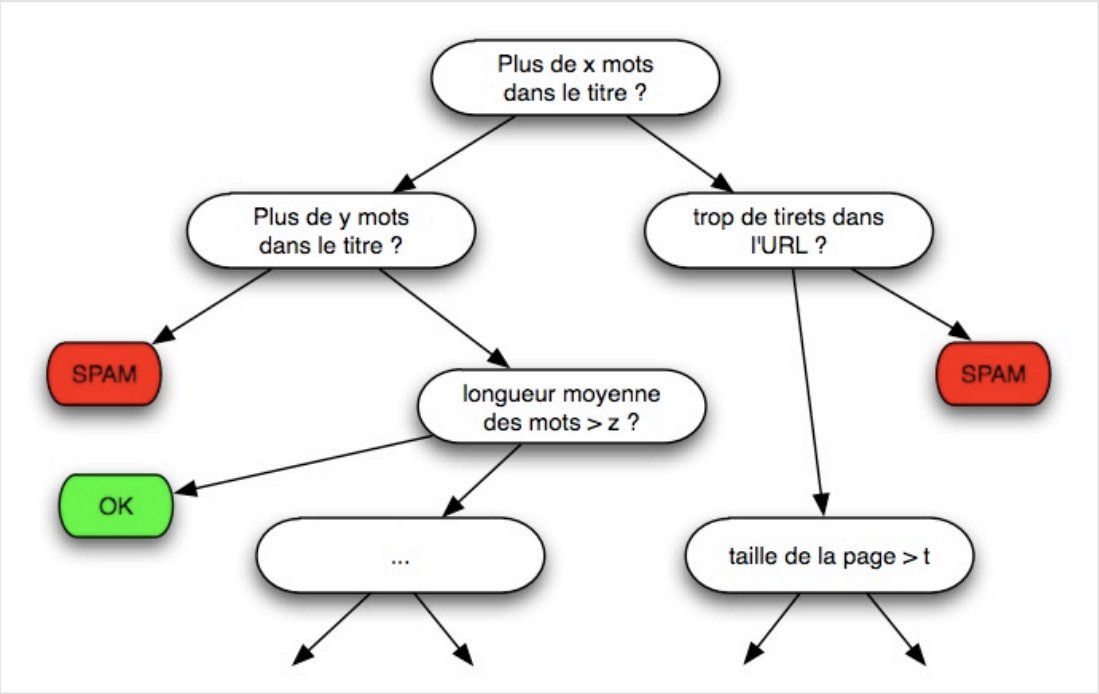
Les auteurs de cette étude ont utilisé un dataset de 18 000 pages qualifiées par des *quality raters*. Ils ont ensuite extrait les caractéristiques du dataset qui permettent selon eux de déterminer ce qui caractérise le spam et qui est détectable algorithmiquement. Il faut noter qu'ils ne se sont intéressés qu'à des critères *in text*. Par exemple, imaginons que l'on pense que le nombre de mots dans le <title> est discriminant pour le spam. Ce que l'on va faire pour tester cette hypothèse est de tracer la courbe suivante (issue de l'article de Ntoulas et al.) :



Cette courbe donne la proportion de pages qui contient un certain nombre de mots dans le <title>, corrélé (courbe violette) par la proportion de spam qui correspond. On voit en regardant cette figure que les pages qui ont un grand nombre de mots dans <title> ont une forte probabilité d'être du spam (par exemple probabilité 83% avec un <title> de 45 mots, alors que la probabilité d'être du spam est de 1% pour 5 mots dans le <title>).

Ntoulas et ses collègues ont ainsi étudié un grand nombre de critères, parmi lesquels :
- Taille de la page : quand un humain écrit, il fait une page ni trop courte, ni trop longue.
- Nombre de points, tirets et autres séparateurs dans l'URL : l'idée est qu'une page de spam a une URL générée automatiquement et qui contient le maximum de mots clés, avec des séparateurs entre eux.
- Longueur du nom de domaine : un long domaine est forcément louche car généralement on choisit un acronyme ou une courte suite de mots.
- Nombre de mots dans le titre : comme son nom l'indique.
- Fraction du texte qui est du texte d'ancre de lien : si tous les mots de la page sont des liens, on a sans doute une page de spam qui fait abusivement des liens.
- Longueur moyenne des mots : dans un texte standard, les mots apparaissent avec certaines probabilités, par conséquent les mots ont une longueur moyenne particulière. Un texte artificiellement généré est souvent déviant car la présence de certains mots est renforcée.
- Rapport de compression, rapport contenu/contenant : quelle est la taille de la page compressée par rapport à sa version normale ? Quelle est la proportion de code par rapport au réel contenu ?
- Vraisemblance d'indépendance : est ce que le texte de la page pourrait apparaître avec grande ou petite probabilité si il était généré par des singes tapant au hasard sur une machine à écrire ?

Il existe bien sûr de très nombreux autres critères, mais notre but ici n'est pas d'en faire une liste exhaustive. Une fois les critères mis au point, le travail est quasiment fini puisqu'il va suffire de passer les données obtenus dans l'algorithme C4.5 pour obtenir un arbre de décision efficace. Voici un exemple (totalement fantaisiste) d'arbre de décision (issu de notre blog) :




Avec le classifieur obtenu, Ntoulas et ses coauteurs annoncent un taux de reconnaissance du spam supérieur à 85%, et un taux de faux positif de 1,4%. 1,4%, cela a l'air d'être très petit, mais c'est à remettre en perspective de la taille de l'index d'un moteur de

recherche. Si l'index contient 20 000 milliards de pages de qualité, alors 280 milliards d'entre elles seront considérées (à tort !) comme du spam.

Et chèr(e) lecteur/lectrice, il faut bien se rendre compte que Panda et Penguin sont deux filtres construits avec ce type de méthode ! Attention, nous avons écrit « avec ce type » car bien sûr, vu la quantité de données en jeu, Google doit utiliser des algorithmes plus efficaces pour générer un ou des arbres de décision. Et c'est incroyable, car dans les labos de Google, une équipe de chercheurs a écrit un article intitulé « *PLANET: Massively Parallel Learning of Tree Ensembles with MapReduce* » dont le premier auteur s'appelle Biswanath Panda !

## 2006, c'est vieux ! Et le futur ?

Vous en avez peut-être déjà entendu parler, mais nous sommes en cours de réalisation d'une étude similaire, à plus grande échelle et avec des critères réactualisés.

Cette étude sera intéressante pour les chercheurs en algorithmique du Web. En effet, une fois le jeu de données et les critères connus, on peut travailler sur des nouveaux algorithmes pour améliorer le domaine de la recherche d'information sur le web. Elle sera aussi intéressante, de manière évidente, pour les moteurs. Mais elle sera également une grande source d'information pour les référenceurs : avec les critères, vous savez ce que vous devez faire (enfin surtout, ne pas faire) pour ne pas être filtré. Avec le classifieur vous avez un outil qui vous dit si vous êtes dans la zone verte, orange ou rouge pour un moteur qui se préoccupe de la qualité perçue.

Par ailleurs, nous souhaitons obtenir des résultats qui vont au-delà du spam. En effet, la qualité d'une page web peut se noter de manière plus fine qu'au travers de la dualité spam/pas spam. Nous avons choisi d'avoir un niveau de qualité intermédiaire : "spam" versus "haute qualité" versus "faible qualité".

Enfin, nous espérons voir des nouvelles choses, car le web a beaucoup changé depuis 2006. Les webmasters ont de nouvelles manières de penser les sites web. Les filous ont des techniques nouvelles pour générer du contenu. Et enfin, les réseaux sociaux ont largement changé la donne au niveau de la publication des contenus sur le web.

## Nos objectifs et notre méthodologie

Nos objectifs, annoncés dès le début, restent toujours les mêmes :

- Un jeu de données constitué d'environ 300 000 pages en langue anglaise, toutes taguées selon leur qualité (spam, haute qualité, faible qualité).

- Un jeu de données constitué d'environ 150 000 pages en langue française, toutes taguées selon leur qualité (spam, haute qualité, faible qualité).

- Deux rapports analysant ces jeux de données pour expliciter les critères constitutifs du spam, des contenus de haute qualité et des contenus de faible qualité.

- Deux classifieurs (un par langue) qui utilisent les critères pour décider automatiquement du niveau de qualité d'une page web fournie en entrée.

Notre méthodologie est similaire à celle de Ntoulas et al. Notre équipe (Jérôme Darbon, Thomas Largillier, Guillaume Peyronnet et Sylvain Peyronnet) est en cours de finalisation d'un crawl de plusieurs centaines de millions de pages sur le web pour créer une immense base de données d'URL. Le crawl a été (étonnamment) ce qui a été le plus complexe à mettre en place. Toutes les briques pour la suite ont été faciles à réaliser, mais le crawl a été une source constante de problèmes (matériels, logiciels, mais aussi conceptuels).

Une fois ce crawl terminé, nous tirerons uniformément des pages dans la base d'URL pour obtenir les deux *datasets* (anglais et français). Ensuite, nous utiliserons (à l'été 2014) le système déjà en place pour faire taguer par des êtres humains toutes les pages. On parle de 450 000 pages, qui doivent toutes être taguées entre 3 et 5 fois pour garantir les bonnes propriétés du résultat. Cela fait autour de 2 millions de tags à obtenir.

Ensuite, nous analyserons les *datasets*. Pour cela, il faudra lister tous les critères candidats, puis analyser toutes les pages au regard de chaque critère. En plus de tous les critères usuels, nous avons rassemblé des nouveaux critères candidats, avec l'aide de plusieurs acteurs des « méthodes agressives de référencement ». On trouve parmi ces nouveaux critères : le ratio entre nofollow et dofollow sur les liens sortants de chaque page, la position des liens sur la page (plus de liens dans le footer que plein texte par exemple), les signaux d'auteurs et les partages sociaux. On trouve aussi des choses plus ésotériques comme par exemple le résidu post-lemmatisation, qui caractérise la richesse de la langue du texte, ou encore l'analyse du TTFB (*Time To First Byte*, souvent assimilé à la vitesse d'affichage) de la page.

Une fois tous les critères analysés, la dernière tâche est de créer un classifieur avec les résultats obtenus. Cela demande une bonne puissance de calcul, du type de ce que fournit une grosse station de calcul. Encore une fois, nous irons plus loin que Ntoulas et ses collègues, notamment en testant plusieurs méthodes de classification supervisée. Notamment, nous utiliserons des méthodes qui pondèrent le coût des critères pour générer des arbres de décision « économiques ». En effet, il est préférable de tester d'abord les critères les moins coûteux à étudier, pour garder la puissance de calcul pour les pages les plus ambigües.

## Conclusion

Il n'existe pas vraiment de conclusion, car le meilleur est à venir, avec les résultats de cette étude, dont nous dévoilerons une partie (mais pas tout) ici même ;)

Sources

Quinlan, J. R. C4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.
https://www.google.fr/search?hl=fr&tbo=p&tbm=bks&q=isbn:1558602380

Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar. F*oundations of Machine Learning*, The MIT Press, 2012.
https://www.google.fr/search?hl=fr&tbo=p&tbm=bks&q=isbn:026201825X

Alexandros Ntoulas, Marc Najork, Mark Manasse, Dennis Fetterly: *Detecting spam web pages through content analysis*. WWW 2006: 83-92. 2006.
http://www2006.org/programme/item.php?id=3052

Page de Biswanath Panda :
http://research.google.com/pubs/author38399.html

Biswanath Panda, Joshua S. Herbach, Sugato Basu, Roberto J. Bayardo. PLANET: *Massively Parallel Learning of Tree Ensembles with MapReduce*. Proceedings of the 35th International Conference on Very Large Data Bases (VLDB-2009).
https://32mech.googlecode.com/files/vldb2009.pdf

***Sylvain Peyronnet**, Professeur des Universités à l'Université de Caen Basse-Normandie (http://sylvain.berbiqui.org/) et **Guillaume Peyronnet**, gérant de Nalrem Médias (http://www.gpeyronnet.fr/nalrem-medias.html). Ensemble, ils font des formations (http://www.peyronnet.eu/blog/masterclass-moteurs-seo/) et essaient de battre les loutres à la pêche à la truite.*