Comment les moteurs de recherche détectent-ils le content spinning ?

Retour au sommaire de la lettre

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

Le content spinning, déjà évoqué dans nos colonnes, est une méthode utilisée par de nombreux référenceurs pour obtenir des textes pertinents, vraisemblables et originaux pour leur site web. Mais il peut s'agir de méthodes dangereuses si elles sont détectées par les moteurs de recherche (Google a d'ailleurs clairement indiqué dans le passé que le spinning était interdit). Mais comment font ces moteurs pour détecter le content spinning ? Retour sur les travaux de trois chercheurs américains à ce sujet...

Chers lecteurs et lectrices de la lettre R&R, vous savez très certainement déjà ce qu'est le spin, car plusieurs articles vous en ont parlé par le passé (voir articles de la lettre d'avril et mai 2013). Nous n'allons donc pas vous faire l'affront de vous le (ré)expliquer en détail. Cependant, pour s'assurer que nous parlons très exactement de la même chose, nous allons dans un premier temps définir le vocabulaire que nous allons utiliser dans cet article.

Pour de nombreuses tâches liées à nos métiers de référenceurs ou webmaster, nous avons besoin de disposer de contenus textuels. Que ce soit pour faire de l'acquisition de liens et ainsi apporter de l'information contextuelle, ou tout simplement agrandir facilement la taille d'un site web, ou bien encore pour créer de la valeur ajoutée sur des pages à faible intérêt (listing, catégories), on a toujours besoin de contenus, parfois même de plus en plus de contenus.

Disposer de nombreux textes est donc un atout non négligeable dans une optique de référencement. Mais pas n'importe quels textes ! En effet, ceux-ci doivent pouvoir répondre à quelques exigences fondamentales :
- **être pertinents** pour les requêtes ciblées (n'oublions pas que l'enjeu final est d'obtenir un bon positionnement) ;
- **être vraisemblables** (sinon on parle de spam) ;
- **être originaux** (les textes dupliqués ou quasi-dupliqués peuvent être détectés et donc pénalisés, cf. lettre R&R de décembre 2013).

Pour obtenir ces textes de façon peu onéreuse et rapide, trois options sont généralement possibles :
- **Acheter du contenu offshore**. La qualité est variable et selon le pays où sont sous-traités les textes, on peut avoir des soucis de ciblages (certains termes ont des fréquences d'usage différents, certaines expressions n'existent pas ou au contraire sont sur-utilisées).
- **Copier le contenu d'un site tiers**. Nous l'avons vu, ce n'est pas une option si l'on souhaite éviter d'être repéré par les filtres de détection des contenus dupliqués ou quasi-dupliqués.
- **Utiliser des méthodes de réécriture de contenu**. Le *content spinning* est une technique de réécriture de contenu qui vise à remplacer des mots d'un texte ou à modifier la structure de ce texte (ou les deux à la fois) pour créer un texte différent du premier, mais qui porte la même information.

A l'heure actuelle, il n'existe pas d'algorithme, et par conséquence pas d'outil logiciel, capable de créer à partir d'un texte libre - écrit en langage naturel - ce que l'on appelle un ***masterspin***. Un masterspin, c'est une expression qui permet de générer de nombreux textes différents, qu'on appelle des ***spuns***, *via* un outil de génération automatique (ce que l'on désigne sous le terme de ***spinner***).

Un masterspin ressemble beaucoup à une expression régulière. On peut y décrire des alternatives et des substitutions à l'aide d'opérateurs spécifiques. Pour cela on utilise les

accolades { et }, ainsi que le *pipe* |. Par exemple, l'expression suivante est un masterspin qui peut générer 4 phrases différentes :

La {loutre | pie} est {joyeuse | voleuse}.

En effet, ce masterspin définit les phrases suivantes : *la loutre est joyeuse, la loutre est voleuse, la pie est joyeuse et la pie est voleuse*. On peut construire des masterspins complexes en imbriquant les opérateurs. Ces masterspins permettent alors la génération de très nombreux textes, avec une variabilité très grande entre les textes.

On voit de suite que la qualité du processus de content spinning est directement dépendante de la qualité du masterspin qui a été mis en place pour la génération. Ecrire un bon masterspin est devenu un vrai métier, et ceux qui ne savent pas le faire et ne souhaitent pas faire appel à un professionnel pour cela n'obtiennent jamais de bon résultats !

Afin d'aider les plus allergiques aux masterspins, certains outils de content spinning proposent de prendre en entrée des textes en langage naturel et d'opérer de simples substitutions des mots par des synonymes pour ainsi créer des textes capables de passer aux travers des outils de détection de contenu dupliqué.

Même si ce type de fonctionnalité ne donne que des résultats médiocres, il devient alors possible à l'amateur le plus complet d'accéder à une méthode de génération automatique de contenu.

Naturellement, plus les techniques de content spinning sont utilisées, plus il devient indispensable pour les moteurs de recherche d'être capable de détecter leur utilisation. Zhang, Wang et Voelker *(voir sources à la fin de l'article)* annoncent ainsi que sur environ 430 000 pages de sites de type wiki qu'ils ont étudié, 52% contenaient des spuns.

Sur le plan de la détection, il n'y a pas eu beaucoup d'avancées ces dernières années. La plupart du temps, il n'est de toute façon pas nécessaire de déterminer si un texte est un spun pour pénaliser la page qui le contient. En effet, dans la grande majorité des cas, les pages sur lesquelles sont mis en place les spuns sont de toute façon de très faible qualité, et sont donc pénalisées pour d'autres raisons (profil de liens abusif, site ayant toutes les caractéristiques du spam, etc.).

Cependant, le problème de la détection des spuns sur le web est devenu un sujet de choix pour certaines équipes de recherche. D'abord pour la beauté du sport, car le sujet en lui même est un vrai challenge algorithmique. Mais aussi car, au delà de la détection des spuns, se profile le problème de leur source.

Dans ce cas, c'est le moteur de recherche qui y a intérêt. S'il sait déterminer si un texte est un spun ou un véritable texte, s'il est capable de regrouper les spuns générés à partir d'un même spinner ou d'un même masterspin, voire même par un même référenceur, il peut alors mettre en place des contre-mesures efficace pour pénaliser des réseaux, des outils et des personnes.

L'article scientifique de Zhang, Wang et Voelker est le premier, à notre connaissance, à s'intéresser aux problèmes de la détection de spuns. Même s'il ne vise que la détection des spuns obtenus par la seule substitution de synonymes, il a le mérite de montrer que la détection est possible pour les méthodes de spinning les plus basiques. Bien sûr, les techniques évoluées passent encore largement au travers des mailles du filet, ce qui fait que les bons spinneurs ont encore de beaux jours devant eux !

Dans la suite de cet article, nous allons vous exposer les travaux de Zhang, Wang et Voelker, ce qui vous permettra de bien comprendre comment on peut détecter du spin simple. Tous les trois sont en poste à l'Université de Californie à San Diego (UCSD).

Une cible : The Best Spinner

L'approche des trois chercheurs vise avant tout à détecter le contenu généré avec *The Best Spinner* (un logiciel du commerce vendu à l'adresse http://thebestspinner.com et qu'on notera TBS à partir de maintenant). TBS permet de générer, à partir d'un texte qu'on lui fourni, des centaines de variantes de ce texte, toutes différentes mais toutes globalement similaires sémantiquement.En pratique, il est considéré que TBS prend en entrée un texte et le transforme en masterspin en repérant des blocs de mots qui peuvent être remplacés par des groupes de mots de sens similaires (synonymes, formes conjuguées, etc.).

Pour comprendre le fonctionnement de TBS, les chercheurs ont mis en place du *reverse engineering* pur et dur. Plus précisément, ils ont découvert que TBS télécharge le dictionnaire des synonymes et le sauve localement dans un format crypté de manière très basique. Ils ont ensuite désassemblé le binaire pour comprendre le mécanisme de connexion au serveur qui fournit le dictionnaire, ce qui leur a permis d'en obtenir une version lisible. La version téléchargée en août 2013 contient 750 114 synonymes groupés sur 92 386 lignes. Chaque ligne contient un mot, et les synonymes de ce mot. Il faut noter que le dictionnaire n'est pas transitif : si le mot XXX a pour synonyme YYY, et que ZZZ a pour synonyme XXX, cela ne veut pas dire que TBS va considérer que ZZZ et YYY sont synonymes.

Des méthodes qui ne fonctionnent pas

Dans un premier temps, il est intéressant de voir quelles sont les méthodes qui ne permettent absolument pas de détecter les spuns, même lorsque la stratégie de génération est la plus basique qui soit (remplacement des mots par des synonymes). La première méthode sans résultat est, étonnamment, l'algorithme des *shingles* que nous avons abordé dans la précédente lettre R&R. Toutes les expériences menées sur ce sujet montrent que les textes issus du spin ont les mêmes distances de Jaccard entre eux que des textes qui ne sont pas issus du spin. Cela prouve une chose : le spin, qui est utilisé pour passer à travers les méthodes de détection de contenu dupliqué, remplit parfaitement son rôle d'origine.

La deuxième méthode qui ne fonctionne pas consiste à analyser la structure même des phrases, en faisant ce que l'on appelle l'étiquetage morpho-syntaxique du texte. Sous un nom qui paraît ésotérique se cache une technique simple : on va associer aux mots les informations grammaticales qui leurs sont associés, comme le genre, le nombre, le fait d'être un adjectif, un verbe, etc. (notons que l'étiquetage morpho-syntaxique est parfois compliqué, comme l'illustre la phrase « l'homme ferme la porte » : le terme « ferme » est-il un un adjectif ou un nom commun ? S'agit-il d'un homme, qualifié de ferme, qui porte une femme, ou bien un homme qui fait l'action de ferme la porte ?).

On pourrait croire que cette technique va permettre de différencier les spuns car lorsque l'on spinne un masterspin, il va y avoir des mots qui vont être remplacés par d'autres avec un étiquetage morpho-syntaxique différent (car TBS ne fait pas cette opération d'étiquetage). Même si ceci est exact, cela ne permet absolument pas de différencier nettement les textes légitimes des spuns comme le montre les expériences réalisées par l'équipe de UCSD.

La question est donc toujours la même : comment faire pour détecter efficacement les spuns produit par TBS ?

Une méthode qui fonctionne, mais qui est spécifique à un outil

L'idée de la méthode efficace proposée par Zhang, Wang et Voelker est basée sur la connaissance du dictionnaire de synonymes dont nous vous avons parlé plus haut. Une fois que l'on connaît les mots et motifs qui sont susceptibles d'être remplacés et/ou de remplacer des parties du texte, on peut classer tous les mots des textes que l'on analyse dans deux catégories. La première est celles des **mutables**, il s'agit des mots qui sont

dans le dictionnaire utilisé par TBS. La seconde est celle des **immutables**. Les immutables sont les mots qui ne sont pas modifiés par l'outil de spinning. N'étant pas modifiés, ils sont des entités des textes idéales pour différencier les spuns des contenus légitimes.

D'un point de vue algorithmique, la méthode devient alors très simple. Elle sera similaire à la méthode des shingles mais sans shingles. A leur place, on va utiliser les immutables comme représentants des textes en cours d'analyse.

Pour résumer la méthode, voici comment procéder pour comparer deux textes et déterminer si il s'agit de deux spuns issus d'un même masterspin :
1. On calcule les ensembles des immutables de chaque texte. On note A l'ensemble des immutables de l'un, et B l'ensemble de ceux de l'autre.
2. On calcule l'intersection et l'union de A et B.
3. La similarité de spin entre A et B est donné par le coefficient de Jaccard de A et B pour les immutables, c'est-à-dire par la quantité fournie par l'opération :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Pour être tout à fait correct, la méthode employée par les trois chercheurs n'est pas réellement un calcul de coefficient de Jaccard, notamment parce que l'approche est gloutonne. Qu'est-ce que cela veut dire ? Que chaque texte est traité mot par mot, et donc que le calcul des immutables peut être en partie incorrect. Par exemple, si le texte est « *la loutre agile fait la sieste* » et que la réécriture se fait de deux manières possibles avec dans l'une « *la loutre* » qui devient « *la petite bête* » et dans l'autre « *la loutre agile* » devient « *le leste animal* », alors la méthode de détection ne verra que la première et considérera « *agile* » comme un immutable vu que « *loutre* » sera supprimé car présent dans le dictionnaire.

Cependant, malgré ce petit bémol, la technique donne selon les auteurs de très bons résultats. Le score obtenu dans les expériences présentées dans l'article par la méthode des immutables varie de 75% à 90% selon les cas, là où la méthode des shingles va de 20% à 30%, et celle de l'étiquetage morpho-syntaxique va de 20% à 40%.

Les principaux défauts de la méthode sont assez clairs. Nous avons déjà évoqué le fait qu'il ne s'agit pas d'un vrai coefficient de Jaccard. Cela pourrait être gênant dans certains cas pathologiques de textes contenant uniquement des mots qui sont dans des groupes de mots spinnables. Le mot « *loutre* » dans l'exemple précédent est l'illustration de ce type de mot. Mais l'autre défaut, qui est de taille, est qu'il faut connaître le dictionnaire utilisé par l'outil pour appliquer la méthode.

Dans le cas présenté ici, TBS ne protège pas bien son dictionnaire, ce qui fait qu'on peut facilement calculer les ensembles d'immutables. Mais rien ne permet de dire que cela sera ainsi dans le futur. Par ailleurs, TBS peut mettre en place une contre-mesure en modifiant substantiellement et souvent son dictionnaire. Les auteurs ont d'ailleurs étudié l'évolution dans le temps du dictionnaire et ont montré qu'entre deux versions du dictionnaire, 94% de son contenu était resté inchangé, ce qui permettait de conserver des résultats de détection très corrects.

On remarquera que dans le cas d'outils online, le dictionnaire devient inaccessible totalement, et le seul moyen de retrouver l'information pour la détection est de faire de l'attaque active sur l'outil. C'est-à-dire qu'il faut l'utiliser pour générer des spuns de textes bien choisis, pour découvrir le contenu du dictionnaire de synonymes avec un processus d'essai/erreur. La détection devient alors de fait particulièrement coûteuse.

Conclusion

L'article de Zhang, Wang et Voelker présente bien d'autres aspects de la détection, notamment pour l'accélérer quand il faut analyser des milliers de textes, choses que nous n'aborderons pas ici afin de ne pas complexifier notre propos.

Ce qu'il faut retenir : le spin de base est assez facile à détecter quand on détient le dictionnaire servant à faire la génération de textes. Ce dictionnaire est donc le nerf de la guerre et nous encourageons les référenceurs à créer le leur ! Nous encourageons aussi les créateurs d'outils à faire du spin plus évolué, et à protéger leurs bases en fournissant le service de manière déportée, online.

Sources

DSpin: Detecting Automatically Spun Content on the Web :
http://www.cs.ucsd.edu/~voelker/pubs/dspin-ndss14.pdf
Qing Zhang (http://cseweb.ucsd.edu/~q5zhang/), David Wang (http://cseweb.ucsd.edu/users/dywang/)

Proceedings of the Network and Distributed System Security Symposium (NDSS) :
http://www.internetsociety.org/events/ndss-symposium-2014
Geoffrey M. Voelker (http://cseweb.ucsd.edu/~voelker/), San Diego, CA, February 2014)

The Best Spinner :
http://thebestspinner.com

***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.*