CONTENU DUPLIQUÉ : TYPES, ALGORITHMES ET MÉTHODES D'OPTIMISATION (1ÈRE PARTIE)

Posted on 15 juin 2022

Le contenu dupliqué sur Internet est un problème aussi vieux comme le Web lui-même. Une facilité absolue de copie (voire de pillage) de contenu propre à l'espace web multipliée par des constellations de solutions techniques non-optimisées comme les paramètres de tracking ou les erreurs humaines engendre des milliards de pages doublons à côté des pages déjà existantes. Ceci en fait une des tâches prioritaires à gérer par les moteurs de recherche. Et comme d'habitude, ce que veut Google se répercute inévitablement sur le travail des responsables SEO. Dans cet article en deux parties nous allons passer en revue les différents types de contenus dupliqués, les algorithmes de détection et les particularités de traitement du contenu dupliqué par Google, les méthodes et outils permettant de l'identifier et bien sûr de le corriger.

Qu'est-ce que le contenu dupliqué ?

Commençons par la définition du contenu dupliqué et pour cela reprenons l'explication officielle de Google :

*« Par **contenu en double**, on entend généralement des blocs de contenu importants, appartenant à un même domaine ou répartis sur plusieurs domaines, qui sont identiques dans la même langue ou sensiblement similaires. Dans la plupart des cas, ces contenus ne sont pas trompeurs à l'origine. »*

En se basant sur cette définition, nous pouvons facilement élaborer quelques typologies de contenu dupliqué.

En fonction du lieu d'apparition du contenu en double, on peut avoir :

- Duplications internes (la page dupliquée se trouve au sein du même site).
- Duplications externes (la page dupliquée se trouve sur un autre site, un autre nom de domaine).

En fonction du taux de similitude, on distingue :

- Duplications complètes (« exact duplicate »).
- Duplications partielles (« near duplicate »).

En fonction de la nature des duplications :

- Duplications volontaires et trompeuses.
- Duplications involontaires ou accidentelles.

A ces trois types de duplications, on peut ajouter une 4ème :

- Duplications techniques.
- Duplications sémantiques (pages qui utilisent des mots et tournures différentes, mais finalement parlent au fond exactement de la même chose sans valeur ajoutée).

Selon le type de duplication, la gravité, la réaction et les méthodes de correction ne seront pas les mêmes. C'est ce que nous allons voir plus tard dans cet article.

Comment Google identifie-t-il le contenu dupliqué ?

Du côté des moteurs de recherche, la comparaison de documents web dans l'objectif d'en identifier les doublons est toujours une affaire de compromis entre précision et ressources machine consommées.

Beaucoup d'algorithmes qui sont à notre disposition et que nous pouvons utiliser sans aucun problème pour nos projets, s'avèrent très vite inefficaces à l'échelle du Web quand il faut effectuer la comparaison avec des millions, voire des milliards de pages web.

Pour identifier si un site contient du contenu dupliqué, Google utilise plusieurs niveaux, méthodes et algorithmes d'analyse.

La suite de cet article est réservée aux abonnés.

Vous êtes abonné(e) ? Tapez vos identifiant ci-dessous pour accéder à l'article complet :
(ou Abonnez-vous en cliquant ici)

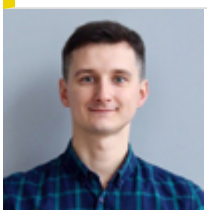
Nom d'utilisateur ou adresse e-mail

Mot de passe

Se souvenir de moi

Connexion

Mot de passe oublié ?



Alexis Rylko, directeur technique SEO chez iProspect (https://www.iprospect.com/ & https://alekseo.com/)