

Référencement Web

1.Introduction

Le référencement naturel

- Premiers moteurs de recherche :
 - Yahoo : un des premiers annuaires du Web
- 1990 : nombre croissant de sites Web
 - Besoin d'outils de recherche
- Vers 1994 : les annuaires
 - Yahoo : sites des catégories / sous-catégories
- Moteur : Webcrawler, excite, Lycos, Infoseek, ..., Altavista
- Les annuaires longtemps utilisés
- Tous disparues
- Le dernier : DMOZ : 2017

Le référencement naturel

- Début 90
 - Peu de sites, peu de pages
 - Algorithmes simples
 - Le nombre de fois de requêtes – apparition dans la page
 - Introduction du « keyword stuffing »
- Google
 - IA
 - Apprentissage automatique
 - Réseaux de neurones

Le référencement naturel

- Produits dans la grande distribution
- Site indexé / référencé : BD du moteur de recherche
- Ne montrer que les pages les plus intéressantes
- Position en fonction des mots clés les plus intéressants
- Est ce que le positionnement a porté ses fruits ?

Comment fonctionne le moteur de recherche ?

- Semble simple mais très complexe en réalité
- Maîtriser le fonctionnement du moteur de recherche pour bien référencer et optimiser votre site
- Le moteur de recherche parcourt le site et indexe les pages
- Constitution d'une base de données index
- Traitement de requêtes et interrogation de l'index
- Restitution des résultats (SERP)
 - Liens organiques (naturels)
 - Liens commerciaux

Technologies des moteurs de recherche

- Principaux : Google et Microsoft
- France : Qwant (Bing) et Orange/voilà
- Technologies de recherche par les portails de recherche francophones (2018)

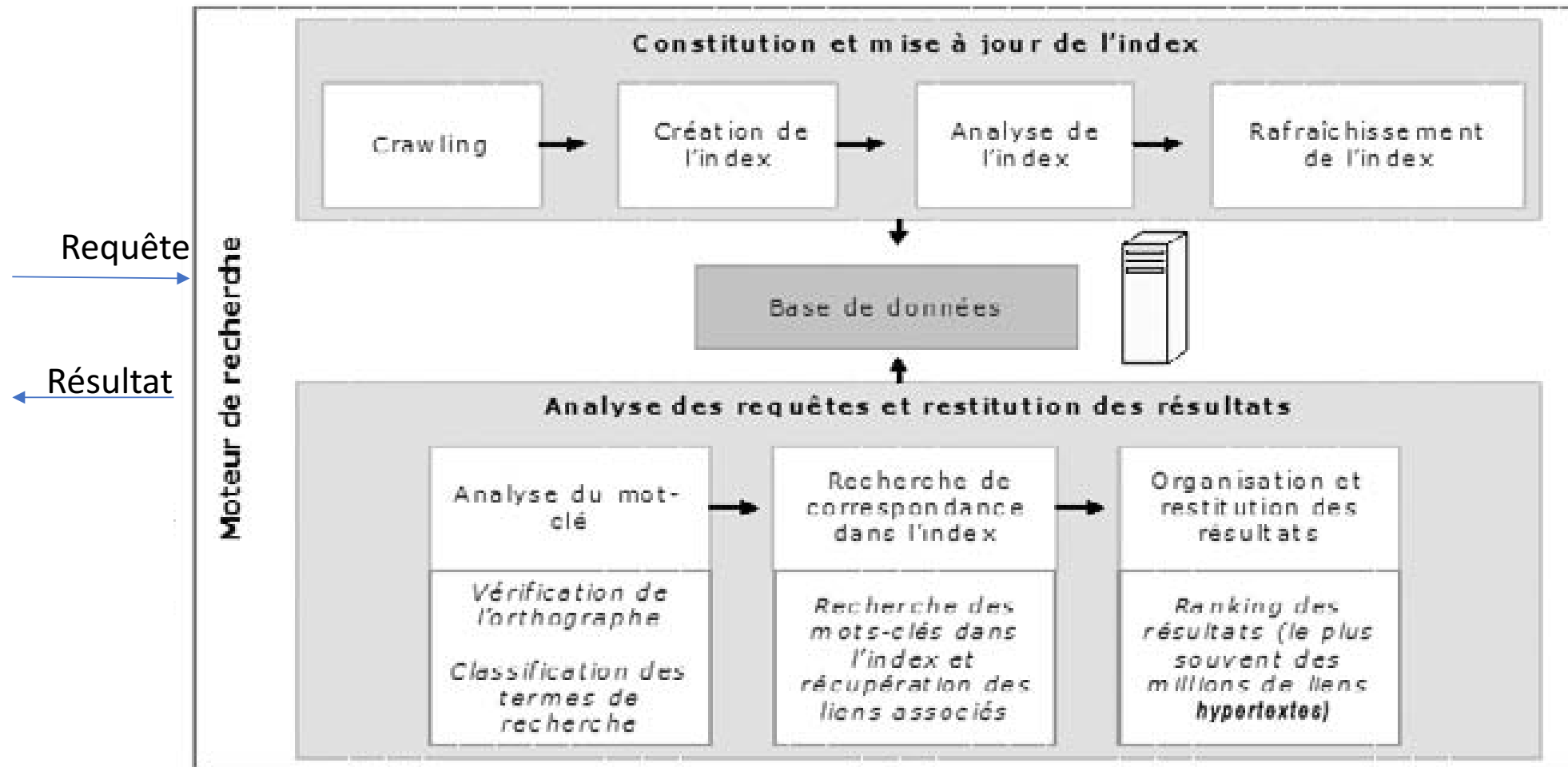
Technologies de recherche	Google	Yahoo	Bing	Orange
Google	X			
Yahoo			X	
Bing			X	
Orange				X
Qwant			X	
Free	X			
SFR/Alice	X			
Bouygues	X			

Technologies des moteurs de recherche

Technologies de recherche par les portails de recherche anglophones
(2018)

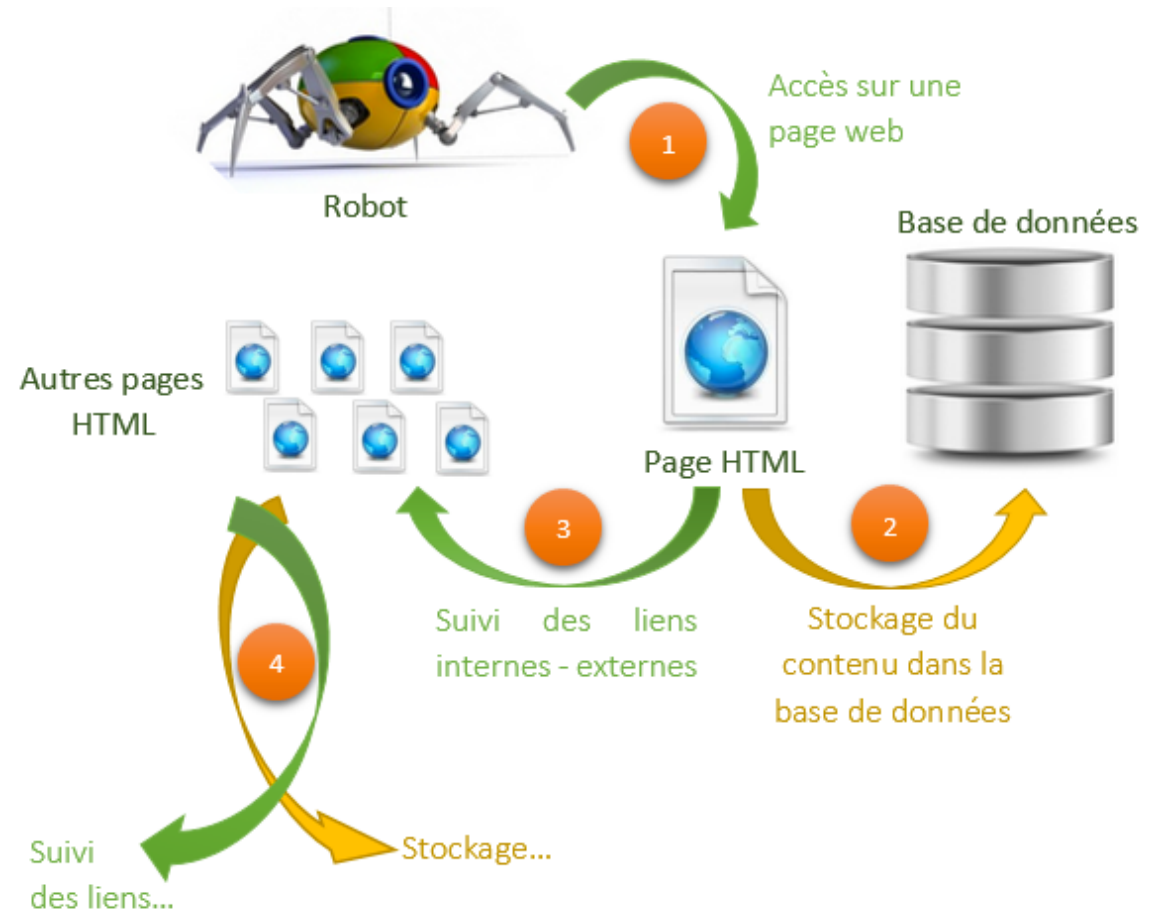
Technologies de recherche	Google	Yahoo	Bing
Google	X		
Yahoo			X
Bing			X
MSN			X
Facebook			X
AOL	X		

Fonctionnement des moteurs de recherche



Les spiders

- Robots, crawlers, agents ou bots
- Visitent en permanence les pages Web (les liens)
- Prennent en compte les modifications
 - Lit les pages web et stockent le contenu sur le DD du moteur
 - Détecter les liens et identifier de nouvelles pages
- Googlebot / Bingbot
- Robots.txt et balise meta « robots » utilisés par les webmasters



Les spiders

- Vérifie s'il connaît la page, si la version est plus récente
 - Oui : supprime l'ancienne et sauvegarde la nouvelle et l'index est mis à jour
- Date de dernière modification
- La taille de la page
- Taux de modification du code html
- Les zones modifiées (ex : lien)
- Le spider doit détecter une vraie modification

Les spiders

- Période « Google Dance » : Mise à jour des index mensuelle
- Aujourd'hui : en continu
 - Visitent plus les pages qui changent souvent le contenu (ex : site d'actualité)
 - Moins les pages statiques
- Affichage de google de la date et l'heure d'indexation
- Suivi des liens par les spiders :
 - Indexation de pages orphelines
 - Liens non pris en compte par les moteurs de recherche
- Fichiers logs : historique de connexions et spiders
- Outils : RobotStats, Watussi, Botify, OnCrawl

Budget Crawl et indexation

- Ressources allouées au crawl du site Web :
 - Les performances du serveur
 - La profondeur dans l'arborescence
 - La qualité du contenu
 - Le pageRank
- Indexation
 - Limite de la capacité d'indexation
 - < 100 liens
 - Aujourd'hui : pas de limite

Indexation

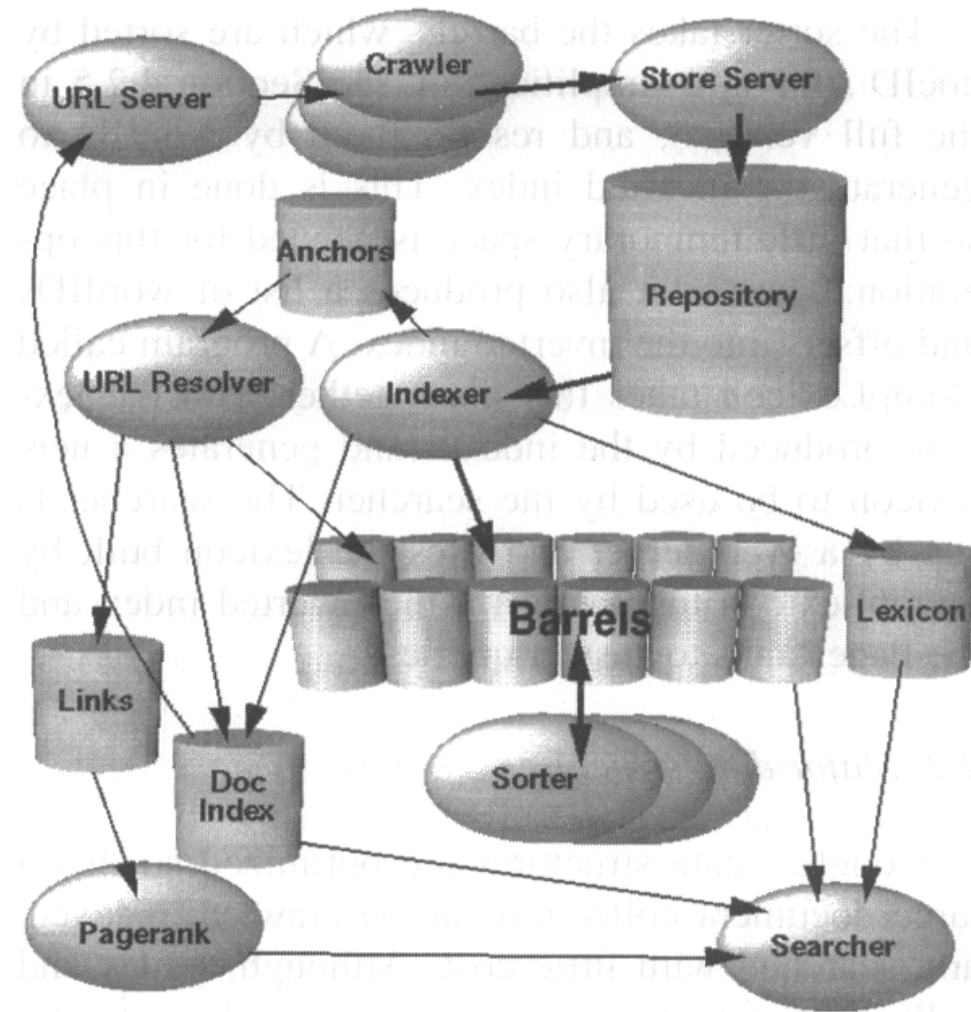
- Index principal (corpus)
- Index inversé (termes clés)
- Traitement d'une requête dans le moteur de recherche
- 2017 : Google crawle 20 milliards de sites Web par jour

Système de ranking

- Algorithmes pour le classement
- Pertinence
 - Position du mot dans la page
 - Poids
 - Similarité avec la requête
 - Distance entre les mots
- Popularité
- Comportement utilisateur
- Clustering

Fonctionnement de Google

- Serveur d'url.
- explorateur (crawler).
- Serveur de stockage.
- Indexeur.
- Solveur d'url.
- chercheur(searcher).
- Pagerank.
- Repository.
- Lexique.
- Ancre
- lexique



Fonctionnement de Google

- Le serveur d'URL : envoie aux crawlers les adresses qui doivent être visitées
- Serveur de stockage: – Compression des pages – Sauvegarde dans le repository – Nouvelle page associée à un docID.
- Indexeur: – Lecture dans la repository. – Décompression des documents. – Parsing des pages. – Conversion des pages en « hit list » (liste d'occurences). – Distribution des « hits » dans les « barrels » de façon partiellement triée. – Parsing des url et envoie des informations dans les ancrs. – Création du lexique.
- Gestionnaire d'ancres : stocke certaines informations créées par l'indexeur (les liens hypertextes et les ancrs)

Fonctionnement de Google

- Le gestionnaire de liens : contient des paires de docID
- PageRank récupère les informations de cette BD et calcule l'indice de popularité
- Le trieur : docID -> WordID. Création de l'index inversé
- Comparaison des mots du trieur avec ceux du lexique. Ajout de nouveaux mots
- Interface recherche : répond aux requêtes utilisateurs. Elle utilise le lexique, l'index inversé, les adresses URLs associées et le page Rank.

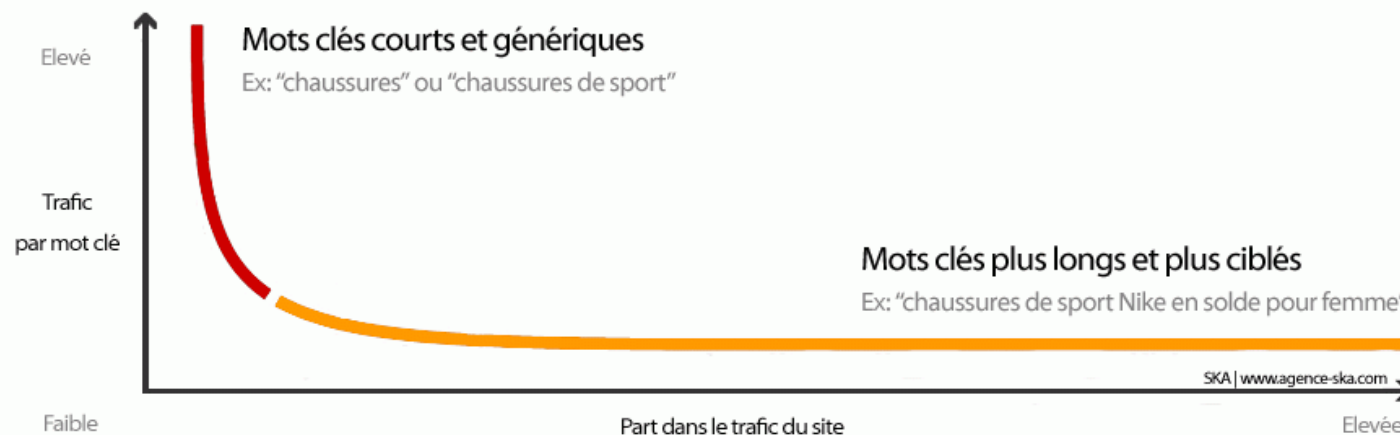
2. Au début du référencement

Méthode

1. Choix des mots clés
2. Choix des moteurs à prendre en compte
3. Création/ modification des pages en fonction de 1 et 2
4. Vérification de la prise en compte des pages par le moteur de recherche
5. Positionnement
6. Amélioration du résultat

Choix des mots clés

- Choisir les bons mots clés
- Etape très importante
 - L'intérêt : souvent saisis par les utilisateurs
 - La faisabilité : positionnement rapide
- Long tail : Les requêtes longues sont d'actualité



Trouver les mots clés

- L'intuition
- Les bases de données : SEMRush, Wordtracker, Keyword Discovery : termes souvent demandés
- Mots clés à partir des termes de départ
- Les générateurs de mots clés
- Demander à son entourage
- Utilisation des moteurs de recherche
- Related searches
- Consultation des balises meta des sites concurrents
- Faute de frappe
- Autocomplétion
 - Correction des requêtes
 - Les requêtes composés

Fautes de frappe

- Identifier les fautes de frappe
 - Générateurs de mots clés
 - L'analyse des mots clés sur le site
 - Google suggest
 - Demander à l'entourage
 - Outil : générateur de faute de frappe
- Référencer son site
 - Insérer des pages
 - URL et attributs alt
 - Mettre des fautes dans ses pages
 - Ne pas cacher dans les pages

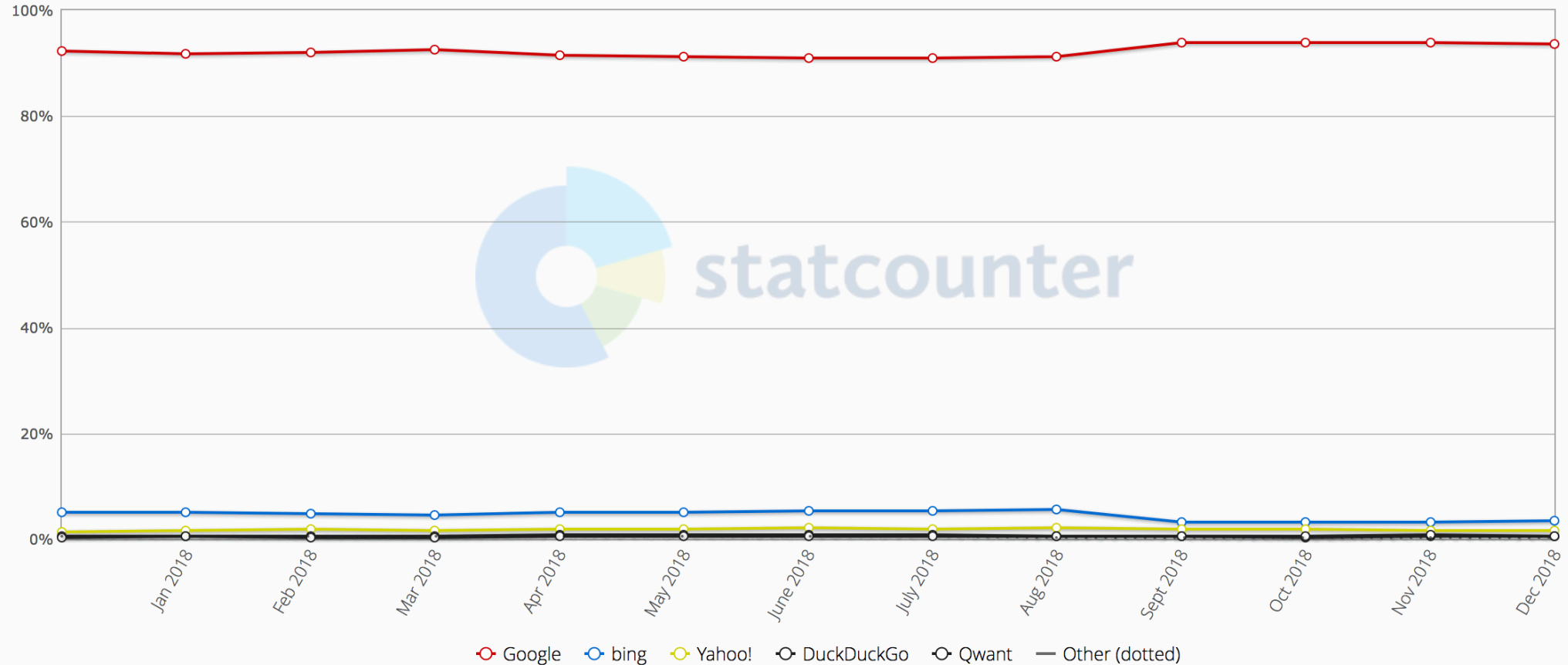
Choix des mots clés

- Une liste de 10 mots clés
- KeywordPlanner : Google
- Etudiez la liste des mots proposés par l'outil
- Classer par nombre de recherches
- Supprimer les expressions qui ne vous intéressent pas
- Etablir des fourchettes de notes
- Prendre en compte le potentiel et la faisabilité

Référencement et moteurs de recherche

Search Engine Market Share France

Dec 2017 - Dec 2018

[Edit Chart Data](#)

3. Référencement

« In page »

Critères in page

- Comprendre le fonctionnement des spiders
- Comment google voit le site (cache)
- Utiliser le navigateur Lynx (que le texte)

Critères in page

- Balise title
 - Critère in page très important
 - Une balise <title> pour toutes les pages
 - Un sujet précis par page
- Balise hn
 - Pour attribuer un poids aux pages Web pour une requête
 - Zones stratégiques
 - Menu de navigation pour comprendre la structure du site (dans la page d'accueil)
 - Pas dans le footer/header/nav (dans une page contenu)
 - Choisir les mots clés pour une balise hn
 - On peut l'utiliser pour une image, alt sera pris en compte : **ne pas abuser !**

Critères in page

- La mise en gras
 - Balise
 - Le mettre dans le css ne sera pas pris en compte par le moteur de recherche
- Les liens internes
 - Pour la page
 - Pour la page pointée
- Les balises meta
 - Dublin core n'a pas d'intérêt
 - Meta description : reprise par le moteur de recherche sinon un extrait de la page (doit être assez long et cohérent)
 - Meta keywords : que le moteur bing semble l'utiliser
 - New_keywords : Pris en compte par Google actualité depuis 2012

Critères in page

- Attributs alt et title des images et des liens
 - Alt de img par google pas bing
 - Title ni ar google, ni par bing
- Nom de domaine
 - Avoir son propre nom de domaine
 - En 2012 google a placé un filtre de nettoyage «EMD » Exact Match Domain
 - Quel nom de domaine utiliser ?
 - Quel hébergeur ?
 - Google prend en compte les distances des adresses IP : liens entre les documents
 - L'ancienneté du domaine est importante

Critères in page

- Noms composés ?
 - Tiret plutôt que coller les mots
- Nom de domaine ?
 - Mots clés ou nom de l'entreprise
- Communiquer sur un seul nom de domaine
- Référencement international :
 - Plusieurs solutions : un nom de domaine par pays, un sous-domaine par pays, un répertoire par pays, pages multilingues (à éviter)
- URL : des termes claires

Contenu textuel

- Réfléchir moteur de recherche
- La qualité du contenu est importante
- Le texte visible est analysé par le moteur de recherche
- Les occurrences des mots clés : comment doser ?
- 1 page = 1 requête
- La casse n'est pas importante
- Ordre et distance dans le texte
- Un seul thème par page
- Plusieurs langues ?
- Localisation du mot ?
- Avoir une requête principal et un champ lexical
 - Les positionner dans les <hn>
 - Balise : pas dans le <hn> : peut être pénalisé par les moteurs

Contenu textuel

- Mettre des liens connexes (ex : amazon)
- Emplacement des liens : pas loin du texte. Evitez les lien (suite, voir plus...)
- Pas de limite de nombre de liens : il faut que cela reste lisible
- Mettre des liens externes
- Un titre pas page
- Nom du site au début de la page d'accueil et à la fin des pages internes

Critères off page

- Les liens sont très importants pour les moteurs de recherche
 - Les robots suivent les liens pour indexer
 - L'attribut href doit être 100% valide
 - L'ancre sert de thème dans la page (cohérence pour un bon positionnement)
 - Le Google Bombing : multiplier sur les sites sources la requête
 - Pages avec un PR important
 - Eviter images et JS
 - Insérer l'attribut alt dans img
 - Eviter les liens dans les formulaires
 - Prise en compte des liens sortants ?

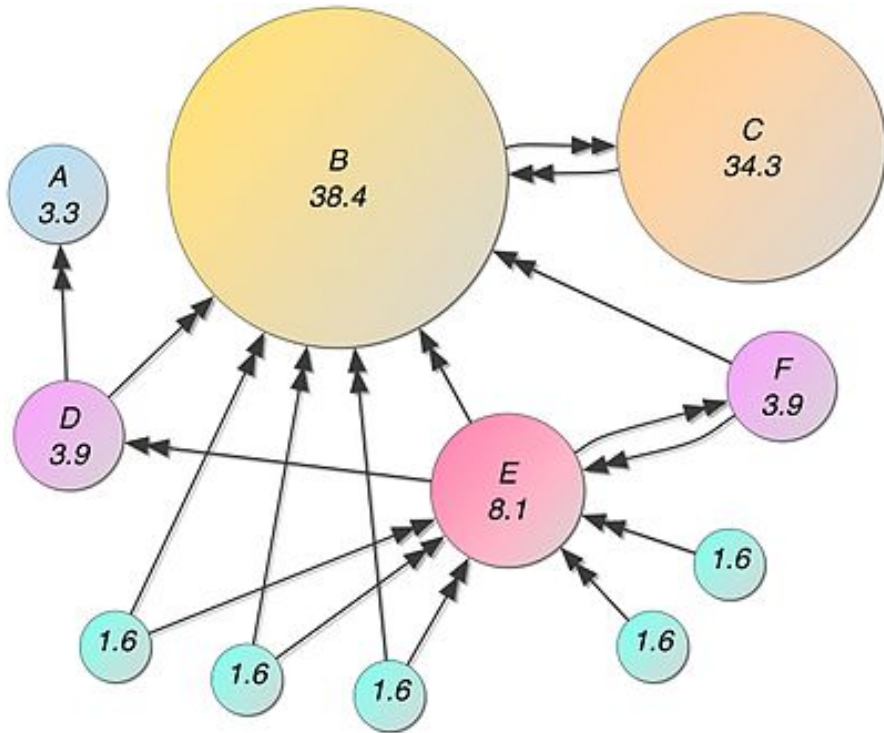
Cocon sémantique

- Stratégie de création de site web et d'arborescence qui s'articule autour de 4 grandes étapes
- Nouveau concept (2015) imaginé par Laurent Bourrely
 1. La définition des requêtes visées, classées par grandes familles et besoins de l'internaute.
 2. Création d'une mindmap
 3. La création de liens internes entre pages mères, filles et sœurs.
 4. L'optimisation de chacune des pages, au niveau des balises et de la contextualisation/sémantisation des liens.

PageRank

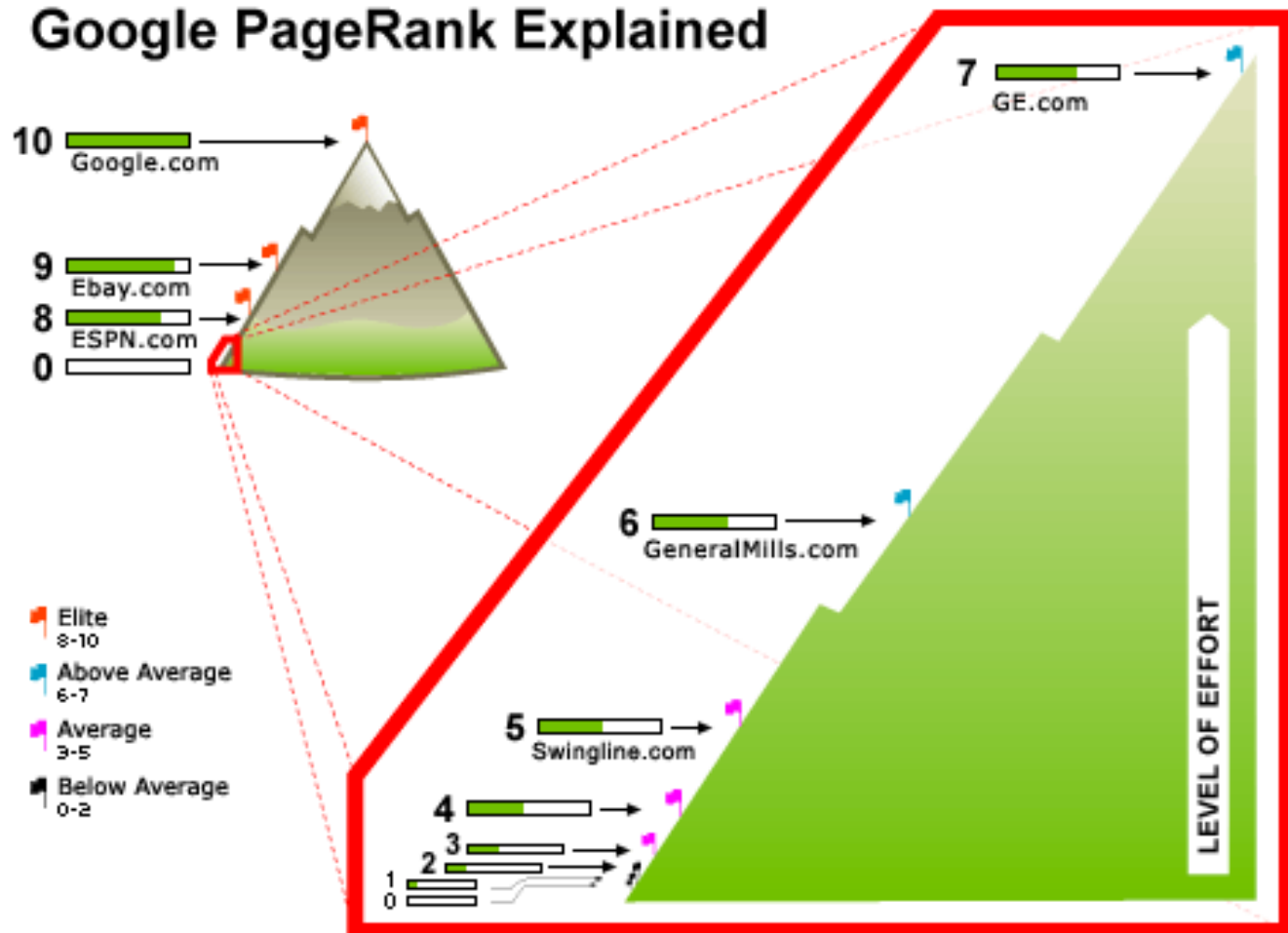
- Les moteurs populaires utilisent l'indice de popularité
- Google : PageRank
- La popularité de la page et des pages qui pointent vers elle
- Le nombre de liens dans la page pointante
- Un poids plus faible pour les liens internes ?
- PR d'une page pas d'un site
- Seuls les liens pointant vers la page
- Toolbar supprimé en Avril 2016
- Le PR ne suffit pas. Un des 200 critères utilisés par google : que la popularité de la page

Calcul du page Rank



- $PR(A) = (1-d) + d (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$
- $PR(A)$ Le site est représenté par A.
- $PR(Ti)$ à $PR(Tn)$ est le PageRank des pages Ti à Tn qui pointent sur le site A.
- $C(Ti)$ à $C(Tn)$ est le nombre de liens sur les sites Ti à Tn .

Google PageRank Explained

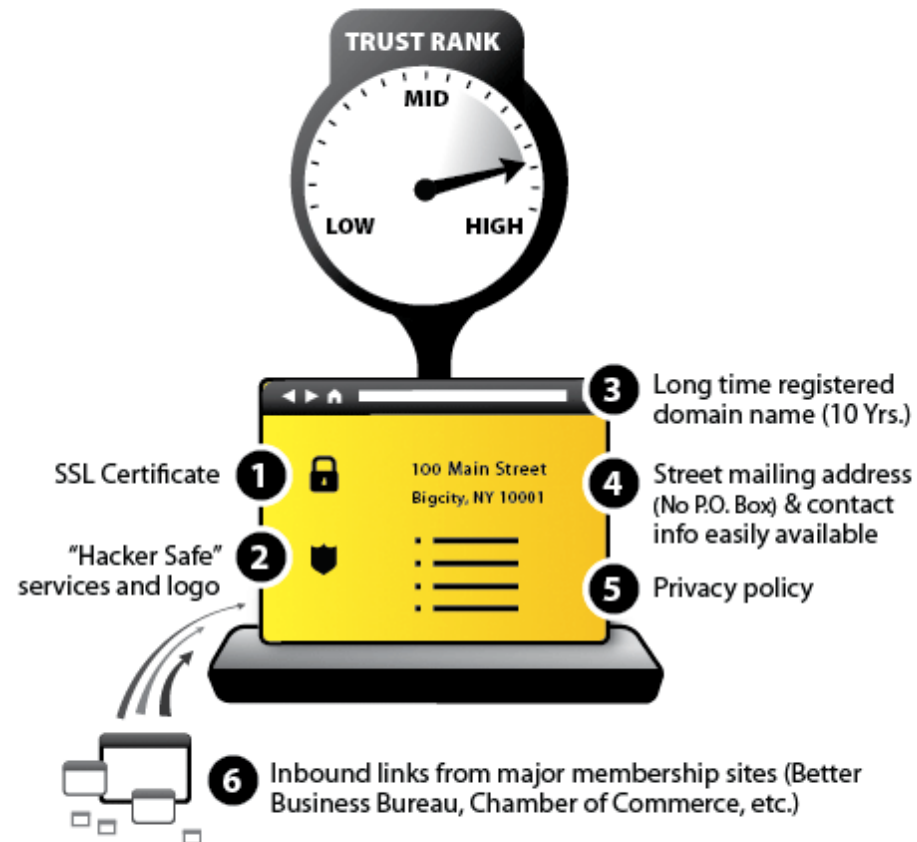


Améliorer son indice de popularité

- Cibler des sites populaires
- Evitez de trop échanger de liens : tous les liens ne sont pas égaux devant Google (liens dans les footer des pages)
- Des liens triangulaires plutôt que réciproques
- La qualité plutôt que la quantité
- Liens payants : formulaire pour les webmaster
- Créer une charte de liens
- Suivre les liens
- L'attribut nofollow

TrustRank (indice de confiance)

Factors That Increase TrustRank



Référencement multimédia

Référencement de l'image

- Le nom de l'image
- Le format de l'image : gif, jpeg, png
- Le texte alternatif : description sans accents
 - Google ne prend pas en compte longdesc (pas valide en html5)
- Texte du lien
- Texte autour de la page
- Texte de la page

Référencement de vidéo

- Comment les moteurs trouvent les vidéos
 - Spider classique
 - L'internaute charge directement la vidéo
- Optimisation des vidéos
 - Le nom du fichier
 - Les métadonnées
 - Les caractéristiques techniques
 - Les tags
 - La réputation
 - Le texte autour de la vidéo

Fichier pdf et word

Tableau 6-1 Champs pris en compte par Google et Bing pour des fichiers PDF

	Google	Bing
Contenu textuel	OUI	OUI
Métadonnée Titre (Title)	OUI	NON
Métadonnée Sujet (Subject)	NON	NON
Métadonnée Auteur (Author)	NON	NON
Métadonnée Mots-clés (Keywords)	NON	NON
URL	OUI	OUI

Tableau 6-2 Champs pris en compte par Google et Bing pour des fichiers Word

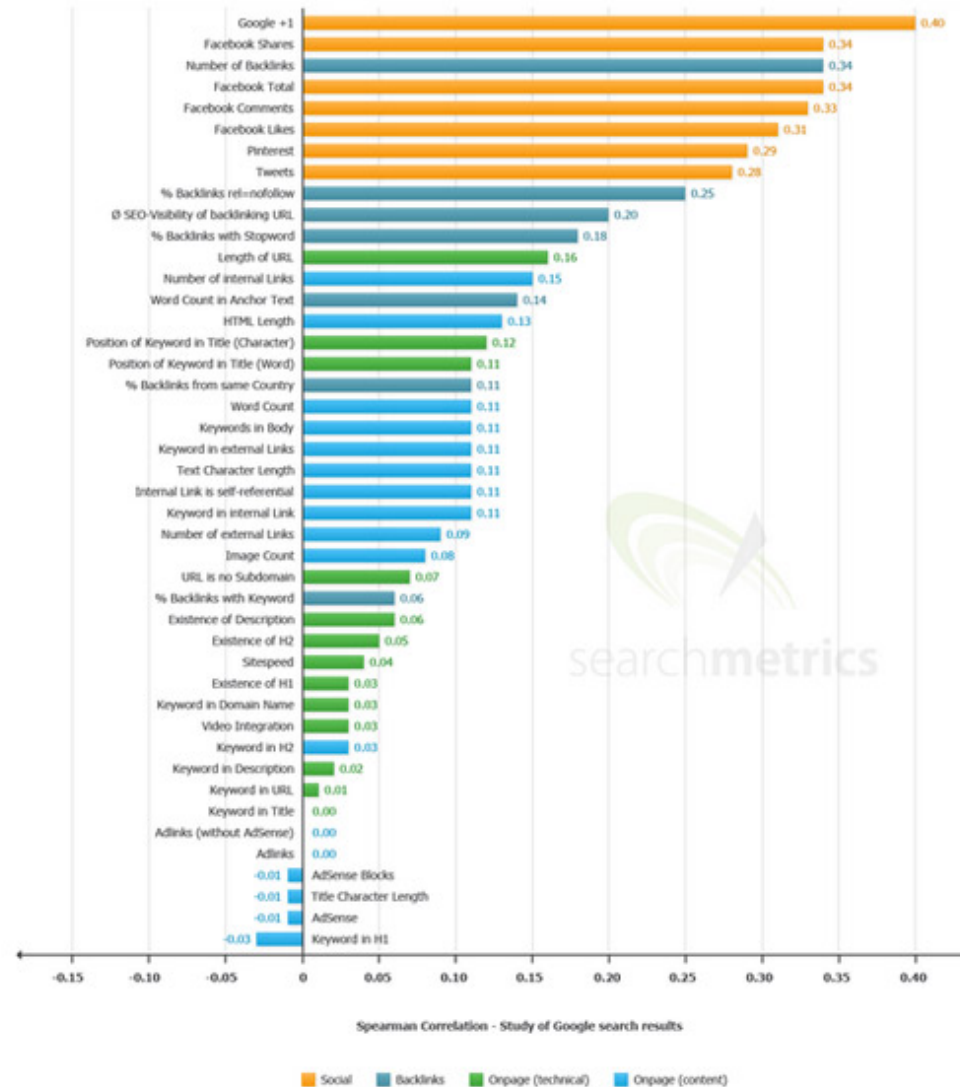
	Google	Bing
Contenu textuel	OUI	OUI
Métadonnée Titre (Title)	OUI	NON*
Métadonnée Sujet (Subject)	NON	NON
Métadonnée Manager	NON	NON
Métadonnée Auteur (Author)	NON	NON
Métadonnée Compagnie	NON	NON
Métadonnée Category	NON	NON
Métadonnée Mots-clés (Keywords)	NON	NON
Métadonnée Commentaires (Comments)	NON	NON
URL	OUI	OUI

SMO

Social Media Optimisation

Réseaux sociaux pour le référencement

- Comment profiter des réseaux sociaux pour le référencement ?
- Inciter les visiteurs et internautes à diffuser l'information
- Twitter, Facebook et Google+ : réseaux sociaux « stars »
- Ils ont des liens nofollow : invisibles pour les moteurs de recherche



Réseaux sociaux pour le référencement

- Les études se basent sur la corrélation et non la causalité
- Les réseaux sociaux comme une source d'analyse et de détection de pages
- J'aime, followers... ne sont pas pris en compte
- SEO/SMO complémentaires

Visibilité dans les moteurs de
recherche

Plus de visibilité

- Pour une page déjà positionnée :
 - Confiance aux auteurs (jusqu'en 2014 affichage de l'auteur)
 - Rich snippets : RDFa, microdata, JSON-LD (Web sémantique)
 - Schema.org
 - Knowledge Graph : combinaison de schema.org et de profile Google+

Spams et pénalités

A ne pas faire

- Le cloaking : autre contenu
- Le content spinning : du contenu sans le dupliquer
- Le scraping : récupération de contenu

Pénalités

- Sandbox : quarantaine
- Blacklist
- Déclassement
 - Un formulaire de « reconsidération »

Ne pas être référencé

- Balise <meta>
- Pour *empêcher la plupart des moteurs de recherche* :
 - <meta name="robots" content="noindex">
- Pour n'empêcher *que les robots d'exploration Google*
 - <meta name="googlebot" content="noindex">
- **pages confidentielles**
- **pages en maintenance**
- **pages en double**
- **pages qui n'ont pas d'importance réelle**

General Ranking Factors

Rank Correlations Top20

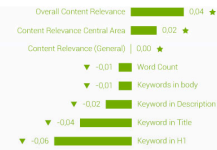
Google.com

TRENDS TO 2016

▲ Up
▶ Same
▼ Down
★ New Feature
★ New Calculation



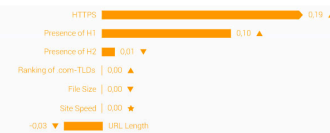
CONTENT



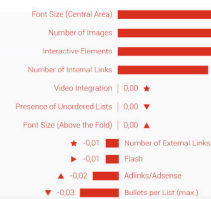
USER SIGNALS



TECHNICAL



USER EXPERIENCE



SOCIAL SIGNALS



BACKLINKS



Je

Download the complete whitepaper at searchmetrics.com/ranking-factors



Références

- Réussir son référencement web - Edition 2018-2019: Stratégies et techniques SEO. Olivier Andrieu