F.BOBET
P.CALVET
T.CUBEL
T.LARGILLIER
G.PEYRONNET
S.PEYRONNET (DIR)
G.PITEL
B.SAMANCHE

babbar

LA PUISSANCE DU LIEN

RÉFÉRENCER SON SITE WEB GRÂCE AUX LIENS

La puissance du lien

PUBLIÉ PAR BABBAR - 72 RUE DE LA RÉPUBLIQUE - 76140 LE PETIT-QUEVILLY

ISBN : 978-2-493567-03-1

DÉPÔT LÉGAL : SEPTEMBRE 2022

Table des matières

Préambule

Après un premier volet *Contenu web, paroles d'experts*, condensé de conseils d'experts liés au contenu et au référencement web, l'équipe de Babbar est ravie de présenter ce nouveau livre.

Si on vous parle de popularité, cocon SEO et netlinking, vous faites le... lien ? Une fois encore, les auteurs de ces articles sont des experts du web. Ils apportent une vision complète sur les sujets liés à la popularité sur le web et sur la puissance des liens.

Toujours dans l'objectif que chacun d'entre vous puisse en tirer le plus profit, leurs témoignages sont techniques et précis mais vulgarisés.

Sylvain Peyronnet, CEO de Babbar, chercheur en algorithmique et spécialiste des moteurs de recherche présentera, dans son chapitre « A propos du Pagerank », tout ce qu'il y a à savoir sur le fonctionnement de cet algorithme et sa relation au surfeur aléatoire.

Guillaume Peyronnet, co-fondateur de Babbar est l'un des pionniers du référencement web en France. Dans « Le Cocon SEO, une méthodologie pour maximiser la popularité », Guillaume expliquera à la fois le fonctionnement d'un Cocon SEO mais surtout, partagera son expérience quant à la mise en place d'un cocon.

Frédérik Bobet, président de l'agence Trikaya Communication, co-créateur de Yourtext.guru est un passionné de SEO. Il est l'auteur du chapitre « Les

enjeux du maillage interne » dans lequel il développera les éléments indispensables pour mettre en place un maillage interne efficace.

« Netlinking : ce qu'il faut savoir », un titre de chapitre évocateur et qui est, parfois, sujet de débat. Pierre Calvet, SEO et Customer support chez Babbar, apportera des précisions sur cette pratique et nous expliquera comment réaliser une stratégie de netlinking.

Comment reconnaître un bon lien d'un mauvais lien ? C'est le sujet du chapitre « Détection des fermes de liens » présenté par Thomas Largillier. Thomas, co-fondateur de Babbar, chercheur en informatique spécialisé dans l'algorithmique liée au web, développe les notions de fermes de liens et de structures optimales.

Thomas Cubel, consultant, formateur dans le domaine du SEO et spécialiste des enjeux on-site nous parlera dans son chapitre, de liens. Il est question de backlinks bien entendu, mais aussi de lien social. Thomas s'intéresse à l'importance de trouver un équilibre entre ces différents liens afin d'assurer la réussite d'une marque.

Le chapitre « Intelligence artificielle : le futur de la rédaction web » est présenté par Bob Samanche, éditeur de site web. L'auteur met en évidence le rôle de l'IA dans la rédaction web, nous parlera des outils d'aide permettant d'améliorer la qualité du contenu rédigé. Enfin, Bob partagera son expérience sur les différentes étapes à suivre pour rédiger un article pour le web.

Enfin, le dernier chapitre « Comment fonctionne un crawler web ? » s'intéresse, d'un point de vue technique, à la création de notre outil, Babbar. Guillaume Pitel, CTO de Babbar et expert en machine learning, donne une définition précise d'un moteur de crawling et nous raconte, comment en contournant les difficultés qui sont liées au crawl à grande échelle, Babbar a vu le jour.

Bonne lecture.

1. A propos du Pagerank

Sylvain Peyronnet est chercheur en algorithmique et spécialiste des moteurs de recherche. Il est CEO et co-fondateur de Babbar, qui exploite les outils SEO `yourtext.guru` et `babbar.tech`. Depuis plus de 20 ans il développe des algorithmes pour l'aide à la prise de décision dans un contexte de grande volumétrie. En 2016, il a reçu le SEMY award de la personnalité search de l'année en marge du SMX de Paris.

1.1 Introduction

Dans ce chapitre je vais vous exposer les principaux éléments à savoir et comprendre, lorsque l'on est SEO, concernant l'algorithme du PageRank et sa relation au surfeur aléatoire. Nous précisons au lecteur que nous distinguons deux concepts : le PageRank, qui est l'algorithme de calcul, et le pagerank, qui la quantité mathématique calculée par l'algorithme en question.

Avant de rentrer dans les détails, quelques considérations générales et historiques sont bienvenues. Le web est constitué de pages qui sont reliées entre elles par des liens hypertextes. L'invention d'*hypertext* a nettement précédé celle du web puisqu'on retrouve dès 1945, dans un article de Vannevar Bush [1], l'idée de la navigation entre éléments de connaissances

1. Bush, V. (1945). As we may think. The atlantic monthly, 176(1), 101-108.

via des liens. A partir de 1945, la route qui va mener à la création du web va être longue.

Cette route se poursuit avec de nombreux jalons, on notera par exemple 1958, année de l'invention du modem, ou encore 1961 pour les premiers travaux théoriques sur les communications par paquets [2]. En 1965, Ted Nelson introduit le terme *hypertext* [3].

La technologie va ensuite rapidement progresser, puisque dès 1973 on assiste à la mise au point d'ethernet et du protocole IP. Très rapidement (1975), Arpanet va être opérationnel. Tout est alors en place pour qu'émerge un "objet technologique" permettant de naviguer dans l'information. Ce sera le World Wide Web (WWW ou web), officiellement actif à partir du 13 mars 1989. A l'origine de ce projet on trouve Tim Berners-Lee, qui est rapidement rejoint par Robert Cailliau.

D'un point de vue mathématique, le web est un objet très simple : il s'agit d'un graphe orienté, dont les noeuds sont les pages web, et dont les arcs sont les liens hypertextes entre ces pages. Ce qui rend le web complexe à manipuler et comprendre, c'est avant tout sa structure particulière. Cette structure est en forme de noeud papillon. La figure 1.1, mise en avant pour la première fois par Andrei Broder et ses collègues [4], présente cette structure.

On voit tout d'abord que le web contient une unique composante fortement connexe géante. Le terme vous est peut-être étranger, mais cela veut dire que pour une très grosse partie du web, il existe toujours un chemin de liens entre deux pages prises au hasard dans cette partie. Ce n'est pas très surprenant. En effet, les moteurs de recherche, les annuaires qui référencent de très nombreux sites web, ainsi que les plus gros sites gouvernementaux et commerciaux forment naturellement une composante fortement connexe géante. Par ailleurs, cette composante fortement connexe géante est unique avec grande probabilité. En effet, s'il y en avait deux, il suffit d'un seul lien de l'une vers l'autre et réciproquement pour les fusionner, ce qui arrive naturellement au regard de la taille de telles composantes.

---

2. Kleinrock, L. (1961). Information flow in large communication nets. RLE Quarterly Progress Report, 1.

3. Nelson, T. H. (1965, August). Complex information processing : a file structure for the complex, the changing and the indeterminate. In Proceedings of the 1965 20th national conference (pp. 84-100).

4. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., ... & Wiener, J. (2000). Graph structure in the web. Computer networks, 33(1-6), 309-320.

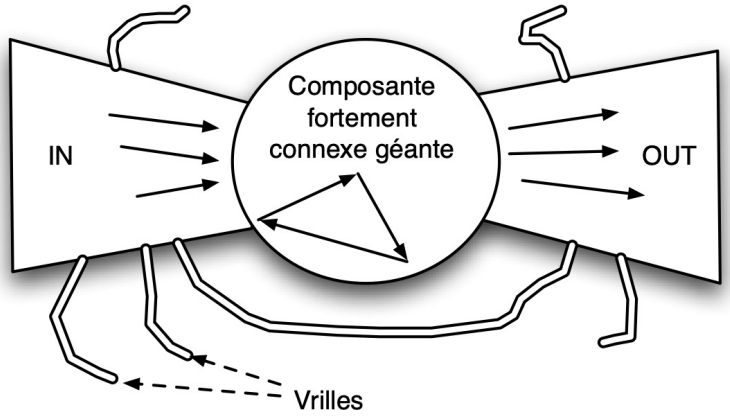


FIGURE 1.1 – Le web est un noeud papillon.

En plus de cette composante fortement connexe géante, on retrouve dans cette structure globale plusieurs composants intéressants. On remarque par exemple l'existence de deux autres grands ensembles de pages web. L'ensemble *IN* correspond à des pages qui forment des chemins vers la composante géante, tandis que l'ensemble *OUT* correspond à des pages qui forment des chemins partant de cette même composante géante. D'autres ensembles sont présents. Les vrilles sont constituées de pages sortant de *IN* ou se dirigeant vers *OUT*. Enfin, des composantes totalement déconnectées du reste du web existent, elles sont d'ailleurs difficiles à détecter puisqu'il faut les connaître explicitement pour les trouver. Cette structure globale est bien sûr mouvante : les frontières de chaque ensemble changent au fur et à mesure que les liens entre les pages évoluent.

Un autre point important concernant le graphe du web est sa taille. On parle de millions de milliards de pages, une toute petite partie étant connue des moteurs de recherche. En 2008, Google annonçait ainsi avoir connaissance de mille milliards de pages web uniques [5].

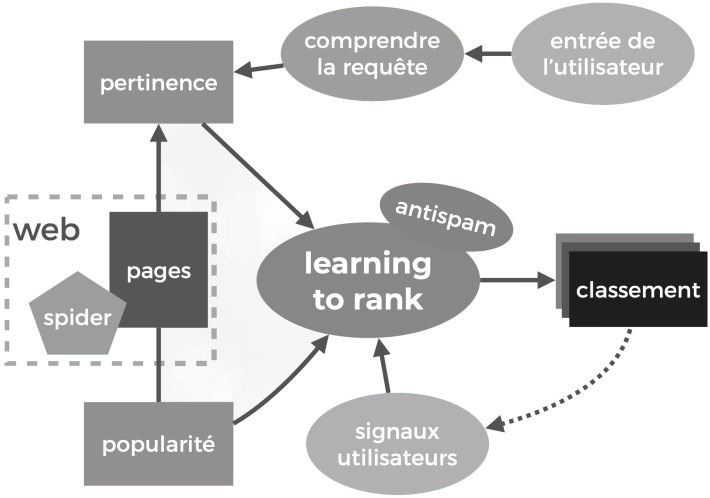


FIGURE 1.2 – Principe de fonctionnement d'un moteur de recherche.

1.2 Principe d'un moteur de recherche

Puisque j'ai mentionné Google, principal moteur de recherche à l'heure actuelle, je vais pouvoir maintenant faire une habile transition vers le sujet des moteurs.

Le principe d'un moteur de recherche est très simple et est résumé dans la figure 1.2, que vous avez probablement déjà vue de nombreuse fois. La première opération du moteur est de créer un index composé de pages web. Pour cela, il *crawle* le web en suivant les liens des pages qu'il rencontre. Ce crawl va fournir au moteur deux types d'information : une information sémantique via le contenu des pages, et une information structurelle via les liens entre les pages.

L'information liée au contenu va être analysée en utilisant des algorithmes issus du traitement du langage naturel et de la recherche d'information. Le sujet des contenus ayant été abordé dans le précédent livre édité par Babbar [6], nous ne l'aborderons pas ici.

Ce qui nous intéresse va être l'utilisation qu'il est possible de faire de

5. `http://googleblog.blogspot.fr/2008/07/we-knew-web-was-big.html`.
6. `https://landing.babbar.tech/fr/formulaire-livre-contenu-web`

l'information structurelle. L'idée forte, présentée par Brin et Page [7] en 1998, est qu'un lien d'une page vers une autre correspond à un vote de la source pour la cible. Ainsi, plus une page reçoit de liens plus elle est considérée comme populaire, et plus elle doit être mise en évidence dans les résultats de recherche. L'algorithme du PageRank formalise cette idée, et surtout la légitime en mettant en correspondance une quantité mathématique simple et le comportement d'un modèle comportemental de l'internaute (le *surfeur aléatoire*).
Dans la suite de ce chapitre, je vais présenter la notion de pagerank, ses évolutions, et quelques points intéressants au sujet de la mesure de popularité sur le web.

1.3 Le PageRank et la popularité

Les moteurs de recherche stockent de manière différenciée l'information structurelle (les liens) et l'information de contenu (les textes des pages). L'information structurelle va être utilisée principalement pour analyser l'importance des pages. En effet, lors de la construction des SERPs pour une requête donnée, le moteur va prendre les pages les plus importantes parmi les plus pertinentes pour la requête et les renvoyer à l'utilisateur.
L'analyse de l'importance des pages peut se faire de différentes manières, mais l'algorithme le plus connu pour calculer le classement des pages en fonction de leur popularité, c'est le fameux PageRank. Cet algorithme qui fit la réussite de Google est au final conceptuellement très simple.
Dans le folklore SEO, on dit du pagerank que c'est une mesure d'autorité des pages web. Cette affirmation est fausse. En réalité le pagerank est une mesure de popularité des pages. On va cependant finir par confondre les deux notions, ce qui est de toute façon le résultat du moteur : en mettant en avant des pages populaires au sens de ses algorithmes, il les transforme en pages d'autorité. Vous lirez souvent aussi que « le PageRank considère qu'un lien vers un site est un vote pour ce site ». Ceci est plutôt correct, comme l'explication qui suit va vous permettre de le comprendre.
La question que va se poser tout moteur de recherche est celle de trouver un signal qui permet d'arbitrer entre plusieurs pages d'égales pertinences pour une requête donnée. En effet, si je suis en mesure de donner plusieurs sources pertinentes, quelle sera celle qui sera préférée par mes utilisateurs ?

7. Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking : Bringing order to the web. Stanford InfoLab.

L'expérience montre que la page préférée sera toujours une page très populaire, les utilisateurs faisant plus confiance à des sources connues qu'à des sources inconnues.

Si cette question parait simple, y répondre techniquement sans monitorer le comportement de tous les internautes est en réalité assez complexe. Le PageRank est un algorithme qui va approcher la notion de popularité, qui est une notion liée au comportement des internautes, de manière formelle et quantifiable par un calcul.

1.3.1 La véritable définition du PageRank

Vous avez sans doute tous déjà vu la formule du calcul du pagerank (qui est un peu plus loin dans ce chapitre), elle parait à la fois simple et compliquée, et au final elle est généralement assez mal comprise. En effet, elle donne le sentiment que le pagerank dépend uniquement des liens vers la page que l'on considère, alors qu'en pratique le pagerank dépend plus du nombre de chemins sur le web qui passent par la page en question.

En réalité, l'idée derrière le PageRank est intuitivement très simple. Quand on y réfléchit, déterminer qu'un site web, et plus précisément une page spécifique d'un site web, ou n'importe quelle page du web, est populaire, c'est quelque chose de très logique et évident : une page est populaire si les internautes s'y attroupent ! Ce n'est pas une surprise, une page qui est visitée est une page qui attire, qui est populaire.

En 1998, Larry Page a tout compris : Si on ne peut pas connaître la popularité réelle d'une page web en comptant les internautes, peut-être peut-on théoriser la façon dont ces derniers parcourent le web pour prédire où ils vont majoritairement. Le modèle théorique qui va permettre de résoudre ce problème, c'est le modèle du visiteur virtuel, appelé "surfeur aléatoire". Cet internaute "modèle" a un comportement de voyageur passant de pages en pages au fil des liens rencontrés.

Au départ, on place un surfeur sur une page web, choisie au hasard, puis on considère qu'il liste tous les liens sortants présents sur cette page, et en choisit un, au hasard, qu'il va suivre, pour arriver sur une autre page. Parfois, le surfeur va souhaiter s'intéresser à quelque chose d'autre, qui n'est pas en lien avec cette page. Il va repartir d'une page web tirée au hasard sur le web. On dit alors qu'il se "téléporte" puisque même sans la présence d'un lien, un chemin qui le guiderait, il va ailleurs, sur le web.

Ce que l'on va appeler pagerank d'une page web, c'est la probabilité qu'un surfeur aléatoire, qui parcourt le web inlassablement de la manière

vue ci-dessus, se trouve sur cette page donnée à un moment donné.

Pourquoi est-ce une mesure de popularité ? Imaginons maintenant que le surfeur aléatoire n'est plus tout seul sur le web. Ils sont maintenant des milliards à se promener de manière aléatoire sur le web. Si une page à une grande probabilité, pour chaque surfeur aléatoire, qu'il s'y trouve à un moment donné, cela veut dire qu'il seront en permanence très nombreux sur la page, qui est donc populaire. Au contraire si la probabilité est petite, alors ils seront très peu à être sur la page.

Petit pagerank est donc synonyme de petite popularité et grand pagerank de grande popularité, c'est au final très simple !

1.3.2 Un peu de mathématiques

Dans le cadre d'une pratique standard du SEO, connaître l'intuition derrière le pagerank est assez souvent suffisant. Mais pour les curieux (oui, je sais que vous l'êtes), voici quelques éléments de mathématiques.

Techniquement, calculer la probabilité que le surfeur aléatoire soit sur telle ou telle page lors de sa « promenade » passe par le calcul de la distribution de la chaîne de Markov associée au comportement du dit surfeur sur le graphe du web. Ceci parait fort compliqué, et à vrai dire ce n'est pas si simple que cela, mais une formule très simple permet d'exprimer ce calcul :

$$\pi_i = \frac{1-c}{n} + \sum_{j \to i} \frac{\pi_j}{d^+(j)}$$

Dans cette formule, π_i est le pagerank de la page i, c est la probabilité pour le surfeur aléatoire de continuer à suivre les liens, $j \to i$ signifie que la page j fait un lien vers la page i (il y a un arc de j vers i dans le graphe), et enfin $d^+(j)$ est le degré sortant de la page j, c'est-à-dire le nombre de liens externes de la page. En outre le graphe contient $|G| = n$ pages. Pour un moteur de recherche, le graphe dont on parle est toujours la partie du web qui est dans l'index du moteur.

Cette formule du pagerank a quelque chose de notable, qu'il faut bien prendre en compte pour l'appliquer efficacement : c'est une formule itérative, c'est à dire qu'on l'applique plusieurs fois de suite, jusqu'au moment où les résultats se stabilisent. Le nombre d'itérations à faire avant la stabilisation est dicté par la valeur de la constante c, qui vaut 85% dans l'article de Brin et Page.

1.3.3 Quel impact pour le référencement web ?

A partir de cette définition on peut en déduire quelques conséquences particulièrement intéressantes pour le référencement web, c'est à dire pour maximiser la popularité d'une page.

La première conséquence, souvent vue comme contre-intuitive, est qu'on peut gagner du pagerank en faisant un lien sortant.

C'est une conséquence très peu intuitive pour les SEOs, qui est pourtant une évidence au regard du fonctionnement du surfeur aléatoire. En effet, la popularité d'une page c'est tout simplement la fréquence de passage du surfeur aléatoire. La bonne pratique pour maximiser cette popularité est donc de créer des chemins avec des liens, autour de sa page, qui vont faire venir et revenir le surfeur aléatoire le plus souvent.

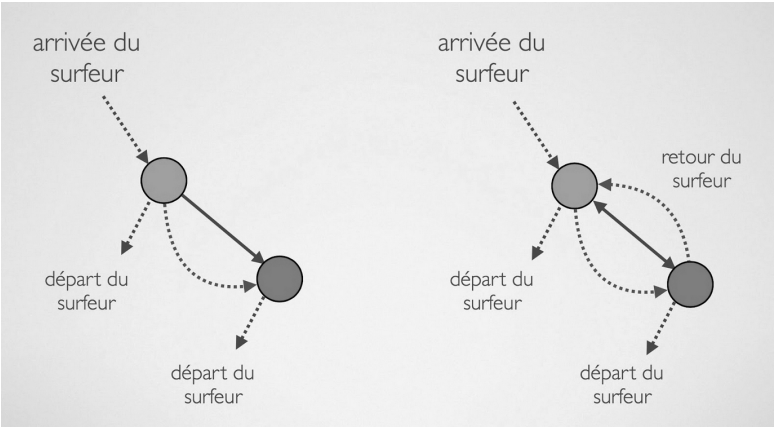


FIGURE 1.3 – Augmenter son pagerank avec un lien sortant.

La figure 1.3 illustre bien ce phénomène. A gauche de l'image, le surfeur arrive sur la page du haut et a deux comportements possibles seulement : soit il se téléporte pour aller ailleurs sur le web. Soit il suit le lien et va sur la page du bas, puis il ne peut que se téléporter. Le surfeur passe une seule fois sur la page du bas.

Sur le schéma de droite de la même figure, le surfeur aléatoire, une fois arrivé sur la page du bas, possède une autre option : revenir sur la page du haut. Ce simple lien en plus, qui crée une boucle entre les deux pages, permet de faire « cycler » le surfeur aléatoire. Concrètement cela

va augmenter sa fréquence de passage, le seul moment où il partira sera lorsqu'il se téléportera (avec probabilité 15% donc).

En pratique, lors d'opérations de netlinking, il faut toujours penser en termes de boucles en plus de penser simplement en nombre de liens. En effet, faire des liens vers l'extérieur peut dans certains cas augmenter fortement votre pagerank, et donc votre popularité au sens du moteur.

Une seconde conséquence est que l'on peut gagner du pagerank grâce à un lien entre deux pages sur des sites qui ne sont ni l'un ni l'autre celui qu'on veut pousser.

Vous avez sans doute deviné l'idée : si une page A me fait un lien, et que je fais un lien vers une page B, alors un lien entre B et A ramène le surfeur aléatoire sur une route qui mène à ma page, amplifiant ainsi la probabilité qu'il soit chez moi, donc mon pagerank.

C'est une façon discrète de faire de l'amélioration de pagerank puisqu'on ne fait pas pointer en direct des liens vers nos pages. Comment soupçonner le site du bas de la figure d'essayer de manipuler sa popularité ? Cependant, l'amplification n'est pas aussi forte qu'en cas de liens réciproques directs.

1.3.4 Faut-il pratiquer un netlinking différent ?

Il y a au final très peu de choses à savoir pour faire de l'amélioration algorithmique :

- ⊙ Les liens réciproques, ça fonctionne mais c'est facile à détecter du côté de Google. Il faut en faire mais avec parcimonie.
- ⊙ Les boucles autour d'une page. C'est une très bonne pratique, qui prend tout son sens dans le cadre d'un réseau de sites par exemple.
- ⊙ Les liens classiques, directement vers la page. À chaque fois qu'on fait un lien classique on amène de la popularité. L'enjeu est de ne pas la perdre et de bien la dispatcher dans le site, pour cela le maillage interne est crucial.

Si vous devez faire impérativement des liens vers des tiers (fournisseurs ou partenaires par exemple), évitez de faire ces liens sur toutes les pages. Un exemple concret : si vous êtes un e-commerce et que vous vendez un produit de la marque X, ne faites pas un lien vers le site de X sur chaque fiche produit. Faites plutôt un lien vers une page interne à votre site, dédiée à la marque X, avec un lien vers le site du fournisseur, et des liens vers d'autres pages internes à votre site. Ainsi vous évitez l'hémorragie de popularité au profit de votre fournisseur.

1.4 Au-delà du surfeur aléatoire de 1998

Le modèle décrit ci-dessus est très simple, et il a permis à Google de prendre l'avantage sur ses concurrents à partir des années 2000. Mais on se doute bien que depuis, le moteur a affiné sa modélisation pour qu'elle colle mieux à la réalité.

1.4.1 Le surfeur est raisonnable

Un modèle de l'internaute qui suppose que ce dernier clique au hasard sur les liens n'a effectivement pas beaucoup de sens. Quand vous êtes sur un site web, vos clics vont vers des pages qui peuvent vous intéresser, pour lesquelles vous avez une curiosité légitime. Ces pages sont souvent liées, sur les pages, depuis les endroits que vous êtes en train de lire, car l'objectif de tout webmaster est de vous garder sur son site.
Le moteur va donc naturellement faire varier la probabilité de chaque lien d'être choisi par le surfeur aléatoire. Cette variation va dépendre de plusieurs caractéristiques des liens et des pages. Certains liens auront moins de valeur car internes, d'autres parce qu'ils sont cachés dans un footer, etc.
En 2012 Google a déposé un brevet sur ce sujet, qu'on appelle le brevet du surfeur raisonnable [8]. Ce brevet porte le nom de Jeff Dean, qui est maintenant l'un des responsables de l'IA chez Google. Le brevet liste plusieurs critères de pondération. Je ne vais pas me lancer dans une description exhaustive car cela n'aurait pas de sens (on parle d'un brevet de 2012), mais on y trouve des éléments de position dans la page, mais aussi de mise en évidence typographique, voire même de sémantique (sur l'ancre par exemple).
Les conséquences de la mise en place du surfeur raisonnable au niveau d'un moteur sont nombreuses, mais **en termes de bonnes pratiques pour le SEO, voilà ce qu'il faut retenir** :

- ⊙ Un lien en footer a assez peu d'intérêt, sauf si votre site est vraiment gigantesque.
- ⊙ Un lien important doit apparaître comme important : il doit être en plein contenu et bien visible.
- ⊙ Si vous voulez faire de la structuration interne, mettez au maximum les liens au dessus de la ligne de flottaison, et c'est valable même pour les menus.

8. Ranking documents based on user behavior and/or feature data. Brevet Google.

⊙ Lors d'opération d'achats de liens, assurez vous que les liens soient contextualisés et plein texte, sinon ils perdent vraiment de leur valeur.

N'oubliez pas que le but de Google est de mettre en avant ce qu'attend l'être humain. Toute opération qui met en avant un élément aux yeux des humains devrait en théorie être vue positivement par le moteur.

1.4.2 Le surfeur est thématique

Un autre aspect du comportement naturel des humains qui n'est pas pris en compte par le pagerank de 1998 est l'aspect sémantique du comportement que l'on a sur le web. En effet, on parcourt les pages du web via les liens, dans le but de s'informer sur un sujet choisi. Il y a donc un biais important lors du choix de la prochaine page et donc du clic que l'on va faire : la plupart du temps il va avoir pour cible une page qui parle du sujet qui nous intéresse.
Parmi toutes les approches algorithmiques qui ont été envisagées à partir de 2000 pour créer un surfeur aléatoire sensible à la sémantique, c'est le pagerank thématique de Taher Haveliwala qu'il faut retenir [9].
Dans le contexte du pagerank thématique de Haveliwala, chaque page a une popularité différenciée par thématique. Ainsi une même page peut être populaire pour, par exemple, la thématique sport, mais pas pour la thématique people. Le surfeur aléatoire sensible à la thématique a des comportements plus riches que le surfeur aléatoire classique. Lorsqu'il est sur une page donnée, il peut soit se téléporter soit suivre des liens (jusqu'ici rien de nouveau). S'il se téléporte, ce sera vers une page tirée au hasard, mais en s'intéressant spécifiquement à une des thématiques. S'il suit un lien, il peut soit continuer à s'intéresser à la même thématique ou au contraire changer de centre d'intérêt. Comme chaque page possède des scores thématiques on sait quelles sont les thématiques portées par la page, et dans quelle mesure elles sont importantes ou pas au sein du contenu de la page.
Cette ventilation de la popularité thématique pour chaque page est stockée dans un vecteur qui a autant de composantes qu'il y a de thématiques considérées dans l'index du moteur.
Lorsqu'un utilisateur du moteur tape une requête, une attribution thématique est faite pour la requête, ce qui permet de prendre en compte la

9. Haveliwala, T. H. (2005). Context-sensitive Web search. Stanford University.

popularité des pages de l'index pour les thématiques qui sont celles de la requête, et seulement celles là.
Bien entendu, cet algorithme, comme quasiment tous ceux des moteurs, a changé lors du passage aux méthodes d'embeddings vectoriels pour la prise en compte de la sémantique. Mais si la technique est différente, le principe reste exactement le même.
En tant que SEO, cet algorithme est probablement l'un des plus structurants pour vous. En effet, il a une conséquence directe, c'est que **les liens doivent être thématisés pour permettre le transfert de popularité**. Une autre conséquence de cet algorithme est la notion de cocon sémantique [10]. Je ne parlerais pas plus de ce sujet ici, puisqu'il est abordé dans l'article de Guillaume Peyronnet un peu plus loin dans ce livre.

1.5 Le calcul moderne du PR

Le pagerank est maintenant une notion ancienne, que ce soit pour les moteurs de recherche ou pour les SEOs. Il y a eu de nombreuses avancées, mais plus de rupture extraordinaire dans la prise en compte des liens depuis maintenant plusieurs années. C'est surtout sur l'angle technique que la plupart des travaux de recherche ont permis des améliorations, un article récent de Stergios Stergiou [11] explique, par exemple, comment on calcule le pagerank sur un index de 100 milliards de pages.
Pour faire un micro instant promo, c'est typiquement le type de sujet de R&D sur lequel l'équipe de Babbar travaille à l'année.
Pour des graphes de plusieurs centaines de milliards de pages, ce n'est plus réellement possible de voir le pagerank comme la solution d'une seule équation. Pour faire le calcul, plusieurs approches sont possibles. Pour certaines, le graphe du web va être distribué sur plusieurs machines de calcul. Ensuite le calcul va consister à "simuler" localement le pagerank, mais parfois il faudra que le surfeur aléatoire "passe" d'une machine à l'autre. Comme fera-t-il ? Il doit accéder à des données qui sont sur d'autres machines, et pour faciliter l'opération, de savants calculs sont faits pour ne pas utiliser des index énormes pour faire passer le surfeur aléatoire. D'autres approches (c'est la nôtre par exemple) vont calculer diverses métriques, dont le pagerank, lors du crawl. L'avantage d'une

10. Notion mise en avant par Laurent Bourrelly sous ce nom, mais que l'on retrouve sous d'autres noms chez d'autres acteurs du SEO.

11. Stergiou, S. (2020, April). Scaling PageRank to 100 billion pages. In Proceedings of The Web Conference 2020 (pp. 2761-2767).

telle méthode est qu'il n'y a pas besoin de décorréler stockage du graphe et calcul des métriques. En contrepartie les métriques calculées ne sont correctes qu'une fois que plusieurs milliards de pages ont été crawlées.
Dans tous les cas, calculer le pagerank ou des métriques similaires est devenu une tâche de grande complexité en terme d'infrastructure et de volumétrie de données à considérer.

1.6 Les outils SEO et le PR et ses dérivés

Pour finir ce chapitre, je vais aborder le sujet des outils SEO et des métriques structurelles qu'ils calculent.

1.6.1 Les outils ne sont pas Google

Il y a de très nombreux outils qui proposent des métriques liées à la structure du graphe du web (ce qu'en SEO on simplifie en « analyse des backlinks »), et un peu moins qui sont les sources de données de tous les autres. Parmi ces derniers on va trouver des outils comme Ahrefs, Semrush, Majestic, Moz, Sistrix, `Babbar.tech` (c'est nous !) et sans doute quelques autres encore. Chacun communique sur les chiffres de son index, par exemple au moment de la sortie de ce livre Babbar connait autour de 1500 milliards d'URLs.
Ce chiffre parait énorme, mais au regard de Google qui annonce connaître 130 000 milliards d'URLs [12], c'est très peu. Est-ce que cette différence est un problème ? En réalité assez peu. Si une page est dans l'index de Google mais pas dans celui d'un outil, il n'y a tout simplement pas de données sur la page dans l'outil, c'est évident. Mais si la page est dans l'index, la qualité de la mesure des métriques est bonne vu les volumes de données stockées par les opérateurs des outils. Je ne vais pas me lancer dans un cours de statistiques, mais avec des volumes en milliers de milliards les calculs sont proches de la réalité avec très grande probabilité.
Le seul point qui peut s'avérer problématique est celui de l'agrégation des données : il n'existe pas de notion algorithmique de popularité d'un site, on reconstruit cette information en agrégeant la popularité des pages du site connues par l'outil. Si un outil n'a pas dans son index une grosse proportion d'un site, la mesure peut avoir une imprécision assez grande.

12. `https://www.seroundtable.com/google-130-trillion-pages-22985.html`

C'est la stratégie de crawl de l'outil qui peut pallier ce problème, chacun ayant sans doute sa recette secrète pour cela.

1.6.2 Les métriques ne sont pas des signaux Google

C'est un point qu'il ne faut jamais oublier, les métriques sont des indicateurs qui sont là pour aider à comprendre la perception algorithmique qu'on peut avoir d'un site web, et sauf grosse surprise il n'y a absolument aucune raison que Google possède un signal exactement équivalent. Souvent parce qu'il va capter des phénomènes de manière plus fine, ou au contraire plus large mais en croisant l'information avec un autre signal. En revanche, il ne faut pas oublier également que les signaux de classement sont également des métriques algorithmiques, qui ont souvent pour vocation de capter les mêmes « primitives ».

En conséquence, l'utilisation des métriques des outils est plutôt une bonne pratique (pour peu que la métrique soit de qualité, ce qui est un autre problème, que je n'aborderais pas ici). Mais il ne faut pas oublier qu'on doit utiliser la métrique pour comprendre, décider et agir, mais que bouger une métrique n'est pas le but car ce n'est pas parce qu'une métrique bouge que le moteur va être plus enclin à mieux positionner un site.

Pour bien faire comprendre cela, je vais faire une analogie : une métrique c'est comme une note sur un contrôle à l'école. Si je modifie ma note en effaçant ce qu'à inscrit le professeur et en remplaçant par une meilleure note, mon niveau ne s'améliore pas. Si ma note augmente parce que j'ai plus travaillé, alors mon niveau augmente. L'augmentation artificielle de la métrique n'a aucun impact sur la réalité de mon niveau, en revanche, l'augmentation suite à une action indirecte (plus de travail) indique bien une augmentation de mon niveau.

Ensuite, chaque outil a ses métriques. A vous de choisir l'outil qui vous offre les meilleures métriques pour la prise de décision, celles qui collent le mieux aux algorithmes modernes des moteurs de recherche me paraissent les plus utiles.

1.7 Conclusion

A vrai dire il n'y a pas tellement de conclusion à avoir en fin d'un tel chapitre. Le pagerank est au final une notion assez simple, qui colle au comportement des humains sur le web, ce qui est normal puisque c'est exactement son but. Quelques bonnes pratiques émergent naturellement,

celles que vous connaissez déjà si vous êtes un SEO aguerri.

2. Le Cocon SEO, une méthodologie pour maximiser la popularité

Guillaume Peyronnet a d'abord été éditeur de sites web monétisés grâce à la publicité. Après quelques années, il s'est mis au conseil et à l'audit SEO, puis à la formation via les fameuses masterclass SEO des frères Peyronnet. C'est un expert reconnu pour ses compétences en référencement naturel.
Il est l'un des co-créateurs de `yourtext.guru` et `babbar.tech`.

Parmi les techniques de référencement web à la mode – mais peut-on encore parler de mode quand celle-ci se maintient depuis des années maintenant – on entend souvent parler du cocon sémantique, des grappes sémantiques, des silos, ou, encore de cocon SEO. Les appellations sont diverses pour finalement évoquer une seule et même technique, appliquée avec une méthodologie différente.

2.1 Définition du cocon SEO

Cette technique, qui porte donc plusieurs noms, consiste à améliorer le référencement d'une page web – voire d'un ensemble de pages – en l'intégrant, par le biais de liens, à un groupe de pages (ledit « cocon »). Ce groupe a une particularité : les pages qui sont liées entre elles sont proches sémantiquement, deux à deux. C'est-à-dire qu'elles partagent la

même thématique et qu'elles se complètent d'un point de vue contenu.

2.2 Comment fonctionne le cocon SEO ?

L'idée derrière la technique du Cocon SEO est d'utiliser à notre profit notre connaissance des algorithmes de Google afin de se faire mieux voir. Pour être très précis, on tire profit de l'algorithme du Pagerank afin d'améliorer la popularité de notre cocon. C'est vraiment parce que l'on fait des liens entre les pages que la magie opère. Cependant, le Pagerank nous impose de respecter une proximité sémantique entre pages liées.
On sémantise donc notre groupe de pages afin que le passage d'une page à une autre soit fluide d'un point de vue thématique. Pour aller encore plus loin, on respecte toujours l'intelligence de l'internaute : des liens importants sont toujours des liens qui sont placés en évidence, généralement dans le contenu principal des pages, c'est-à-dire dans le texte, pas dans le menu de navigation ou dans une partie annexe de la page.

Toutes ces conditions, qui sont évidentes lorsque l'on connait le Pagerank (cf. l'excellent article de Sylvain Peyronnet dans cet ouvrage pour tout savoir sur le Pagerank), permettent alors :

- ⊙ D'améliorer le positionnement des pages du cocon grâce à l'optimisation de la popularité ;
- ⊙ De se positionner sur de la longue traîne : en se forçant à créer de nouvelles pages dont on améliore globalement le référencement, en ciblant des contenus annexes, on obtient de la présence supplémentaire dans le moteur de recherche. Si l'on finit par avoir un cocon qui traite tous les sujets d'une thématique, on peut être positionné sur toutes les requêtes reliées à cette thématique ;
- ⊙ De créer des opportunités d'acquisition de liens. En effet, grâce aux nombreux contenus dont on dispose, on a plus de possibilité d'aller chercher des liens depuis l'extérieur de notre cocon. Mieux encore : avant on avait une page pour laquelle on n'arrivait pas à trouver de liens. Maintenant on a des dizaines voire des centaines de pages qui sont dans la même thématique que notre page principale, mais traitant d'aspects variés. Imaginez : vous avez besoin de pousser une page qui vend une machine à café. Personne ne souhaite vous faire de liens. En mettant en place un cocon vous avez maintenant des pages qui permettent de répondre aux questions techniques comme

« comment détartrer ma machine à café à dosettes ». Et vous pouvez alors rêver que sur un forum quelqu'un pointe vers votre page. Un lien qui grâce au maillage de votre cocon se transforme en apport de popularité indirect pour votre page de vente. Bravo !

Faire un cocon SEO c'est multiplier indirectement les opportunités pour recevoir des liens !

2.3 Le cocon SEO est parfois une mauvaise idée

Il ne semble y avoir que des avantages à mettre en place un cocon, c'est pourquoi c'est une recommandation courante dans le cadre d'une stratégie SEO. Pourtant, il existe deux raisons pour lesquelles il faut aussi savoir parfois éviter de faire un cocon SEO :

- ⊙ Vous pensez encore avoir de nombreuses opportunités pour récupérer des liens depuis des sites tiers. Alors, foncez ! Si votre page est apte à gagner encore en popularité complètement légitime, il ne faut pas hésiter : ce sera généralement beaucoup moins cher que de monter un cocon complet. Et parfois c'est suffisant pour devenir premier sur toutes les requêtes que l'on vise. Alors pourquoi faire plus ? Qu'est-ce qui est mieux que premier ? Gardez en tête que le cocon amplifie la popularité : avant toute chose on doit faire de l'apport de popularité. Si on a un site tout neuf, faire un cocon aura peu d'impact : faire des liens sera plus essentiel. Au contraire, si on a un site déjà très populaire, il suffit d'ajouter des pages bien thématisées pour aussitôt augmenter la visibilité globale du site.
- ⊙ Vous avez un budget vraiment très réduit, ou pas de temps pour écrire vous-mêmes les nombreux contenus nécessaires pour faire un cocon… Vous risquez alors de faire un cocon très réduit, peu efficace. Et s'il y a bien quelque chose que l'on n'aime pas, c'est faire les choses inutilement. C'est difficile d'évaluer ce qu'est une petite taille pour un cocon car selon les requêtes souhaitées, le nombre de pages à faire pour obtenir une belle amplification varie. Mais il faut être potentiellement prêt à faire des centaines de nouvelles pages et donc de nouveaux contenus. A défaut, avec un budget réduit, acheter de la popularité peut être plus intéressant. Ou alors il faut être prêt à accepter qu'avec un budget moindre on ne sera pas premier.

2.4 Méthodologie de mise en place d'un cocon SEO

2.4.1 Le choix du sujet principal

Le sujet principal, c'est le sujet de la page que vous souhaitez rendre plus visible, cela peut être un sujet très général comme « cuisine » si vous avez un contenu très généraliste et la volonté de vous positionner sur des mots-clés assez larges (c'est généralement le cas quand on fait un cocon), mais on peut aussi avoir des sujets plus précis comme « couteau de cuisine » si votre page s'y prête.

2.4.2 Les sujets connexes

A partir du sujet principal, on cherche des sujets qui en découlent, que ce soit par association, élargissement ou rétrécissement. Par exemple, si le sujet principal est la « cuisine » (faire la cuisine), on pourrait trouver des sujets connexes comme :

- les ingrédients ;
- les recettes.

Et si l'on n'a pas beaucoup d'imagination, ou bien si l'on n'est simplement pas un grand connaisseur de la thématique, on risque de rencontrer des difficultés à trouver davantage de sujets connexes.
Alors, pour nous faciliter la tâche, on va utiliser le savoir de Google.

Pour cela, tapons « cuisine » dans le moteur de recherche, puis étudions de quoi parlent les résultats renvoyés :

- cuisine équipée ;
- cuisines pas chères ;
- cuisine sur mesure.

On dirait bien que le moteur de recherche a pris ma requête à la lettre ! J'ai demandé cuisine, mais je pensais « faire la cuisine ». C'est bien de m'en rendre compte maintenant car je pensais cibler la requête « cuisine », mais en réalité vu que je veux parler de « faire la cuisine », c'était une mauvaise requête pour le site : en accentuant l'effort SEO direct sur « cuisine », je risque d'obtenir des visiteurs qui souhaitent fabriquer leur cuisine, et pas du tout faire la cuisine.

Je décide donc d'utiliser « faire la cuisine » comme requête principale. Je trouve alors des sujets connexes bien différents chez Google :

- ⊙ définition de « faire la cuisine » ;
- ⊙ les choses à savoir faire en cuisine ;
- ⊙ comment faire la cuisine ;
- ⊙ cuisiner.

Pour trouver les sujets connexes j'ai regardé les titres qui apparaissent dans la SERP. Pour aller encore plus loin et plus vite, je vais m'intéresser aux pages qui apparaissent dans la SERP, et plus précisément aux mots importants de ces pages.

FIGURE 2.1 – Groupes de 2 mots pour la requête « faire la cuisine »

Pour cela, j'utilise `yourtext.guru` qui utilise la data du moteur de recherche pour faire des recommandations sur la façon d'écrire un contenu. Je demande une analyse sémantique pour la requête « faire la cuisine », puis en parcourant la liste des mots proches (par groupes de deux mots par exemple, comme dans la figure 2.1, mais on peut aussi utiliser les groupes de 1 ou 3 mots, voire les entités nommées), je trouve les sujets connexes suivants :

- ⊙ faire plaisir ;
- ⊙ quiche lorraine ;
- ⊙ cuisine française ;
- ⊙ pâte feuilletée ;
- ⊙ livre de recettes.

Je décide de m'arrêter sur ces sujets connexes directs, pour ensuite procéder par rebond pour trouver d'autres sujets connexes.

Je demande donc à `yourtext.guru` des recommandations pour chaque sujet connexe, en prenant soin de préciser la requête en cas d'ambiguïté

(par ex. « Faire Plaisir » n'a d'intérêt que si c'est un sujet connexe bien relié à la cuisine. Donc c'est plutôt « Faire plaisir en faisant la cuisine » qu'il faut utiliser). Voilà une liste de sujets connexes que je peux récupérer de cette façon en quelques minutes :

alimentation saine, aliments, appareil, appareil creme prise, appareil quiche, beurre, beurre detrempe, blanquette veau, boeuf, boeuf bourguignon, boeuf carottes, brisee, chef, classique cuisine francaise, coquilles saint jacques, couches, couches pate, creme, creme epaisse, creme lait, creme liquide, cuisine, cuisine francaise, cuisine francaise plats, cuisine francaise recettes, cuisine francaise traditionnelle, cuisine maison, cuisine plaisir, cuisine recette, cuisine saine, cuisine therapie, cuisiner, cuisson, culturel immateriel humanite, cyril lignac, detrempe, eau, ecrit, epaisse, estouffade boeuf carottes, etaler, etaler pate, excedent farine, faire cuisine, faire plaisir, farine, farine ble, farine plan travail, fast food, feuilletage, feuilletee, feuilletee maison, feuilletee pate, feuilletee recettes, feuilletee recettes pate, fondue savoyarde, four, four chaleur tournante, francais, francaise, francaise recettes, france, france cuisine francaise, fromage, galette rois, gastronomie, gastronomie francaise, histoire cuisine, histoire cuisine francaise, idee faire, idees, idees faire, ingredients, ingredients pate brisee, lait, lardons, lardons fumes, legumes, livre, lorraine, lorraine pate, lorraine quiche, maison, manger, minutes, minutes preparation, moule, moule tarte, moyen age, noix muscade, noix muscade rapee, nutrition, oeufs, oeufs creme, oeufs entiers jaunes, oeufs mimosa, pate, pate brisee, pate brisee farine, pate brisee feuilletee, pate feuilletee, pate feuilletee apero, pate feuilletee beurre, pate feuilletee classique, pate feuilletee creme, pate feuilletee expliquee, pate feuilletee ingredients, pate feuilletee maison, pate feuilletee recettes, pate plan travail, pate quiche, pates, pates feuilletees, patisserie, paton quart tour, patrimoine culturel, patrimoine culturel immateriel, pivoter quart tour, plaisir, plaisir cuisine, plaisir cuisiner, plaisir mangeant, plaisir manger, plan travail, plan travail etalez, plat, plats, plats typiques cuisine, plier pate, poissons fruits, pommes, pommes terre, poulet, preparation, prix litteraires, produits, puree pommes terre, quart tour, quart tour aiguilles, quiche, quiche lorraine, quiche lorraine creme, quiche lorraine fromage, quiche lorraine maison, quiche lorraine pate, quiche lorraine recette, quiche lorraine recettes, quiche lorraine traditionnelle, quiche tarte salee, quiches, recette, recette clafoutis cerises, recette fondue savoyarde, recette oeufs, recette pate, recette pate feuilletee, recette quiche, recette quiche lorraine, recette tarte, recettes, recettes cuisine, recettes cuisine francaise, recettes pate, recettes pate feuilletee, recettes quiche, recettes quiche lorraine, recettes quiches, repas, repas gastronomique francais, restaurant cuisine francaise, retirez excedent farine, rouleau, rouleau patisserie, sante, saveurs textures, sel, sel poivre, sel poivre muscade, specialites cuisine francaise, tarte, tour, tour aiguilles montre, tours, tours pate, traditionnelle, traditionnelle cuisine francaise, travail, travail etalez pate, typiques cuisine francaise, versez appareil quiche, viande, vin blanc.

Dans cette liste, on peut se rendre compte que certains sujets connexes sont trop vagues, il faudra alors les compléter pour bien comprendre de

quoi il s'agit (par ex. « pivoter quart tour » se réfère à « pâte feuilletée : pivoter quart tour »). On peut aussi vouloir faire du ménage car on ressent qu'il existe des doublons (exacts ou non).

2.4.3 Transformer les sujets connexes en intentions

L'approche classique

Maintenant que nous disposons des sujets connexes, on souhaite les rendre compréhensibles facilement par le moteur de recherche. On l'a déjà vu à l'étape précédente, parfois il a fallu ajouter un ou deux mots pour comprendre ce qu'un sujet désignait. Mais on veut être certain que le moteur de recherche lui aussi va bien comprendre ! Alors, nous allons transformer nos sujets connexes en intentions.

Il y a une façon simple de faire une belle intention, c'est d'exprimer le sujet connexe sous forme de question. Par exemple, « pâte feuilletée » devient « comment faire une pâte feuilletée », ou encore « pourquoi faire une pâte feuilletée ».
Pour que ce soit pratique à systématiser, on va utiliser une liste fixe d'intentions primaires, le *QQOQCCP* : Qui, Quoi, Où, Quand, Comment , Combien, Pourquoi.

En associant chaque sujet connexe à une intention on essaie de former des questions. C'est un exercice amusant, mais on va l'automatiser un peu plus en tapant ces associations dans Google (voir la figure 2.2).
Au fur et à mesure que l'on saisit la requête, on voit Google suggest proposer des recherches d'internautes qui sont porteuses d'intentions. Il n'y a plus qu'à se servir :
« qui pâte feuilletée » $\Longrightarrow$ qui a inventé la pâte feuilletée
« quoi pâte feuilletée » $\Longrightarrow$ quoi faire quoi pâte feuilletée
« où pâte feuilletée » $\Longrightarrow$ où acheter de la pâte feuilletée sans gluten
« quand pâte feuilletée » $\Longrightarrow$ quand utiliser une pâte feuilletée
« comment pâte feuilletée » $\Longrightarrow$ comment faire des croissants avec de la pâte feuilletée
« combien pâte feuilletée » $\Longrightarrow$ combien de temps se conserve une pâte feuilletée industrielle
« pourquoi pâte feuilletée » $\Longrightarrow$ pourquoi piquer la pâte feuilletée

On peut parfois avoir pour chaque couple d'intention primaire et de sujet

FIGURE 2.2 – Les intentions via Google

connexe plusieurs questions qui ressortent. En tout cas, pour une thématique sur laquelle Google renvoie de nombreuses suggestions, en quelques minutes on récupère des centaines de questions.
Si jamais la thématique est mal connue de Google, il ne faut pas hésiter à travailler en anglais, puis à traduire. A défaut, il sera nécessaire de trouver les questions en mettant à profit son mental, parfois cela peut être très long.

Si Google connait particulièrement bien un sujet connexe, il va faire apparaître des PAA (*People Also Ask*). Dans ce cas, le travail de recherche est bien plus rapide : on peut prendre les PAA pour les intégrer dans le cocon.

Aller encore plus vite

L'étape de transformation des sujets connexes en questions peut parfois être un peu délicate, ou longue. Alors on peut vouloir utiliser les outils d'exploration de `yourtext.guru` (questions, idées) ou le *semantic explorer* de `babbar.tech`, voire des outils d'IA (Intelligence Artificielle), comme SEO-TXL de `yourtext.guru` (voir la figure 2.3), pour trouver directement les questions/intentions et ainsi aller plus vite et avoir une diversité plus prononcée, ne s'arrêtant pas purement au SEO.

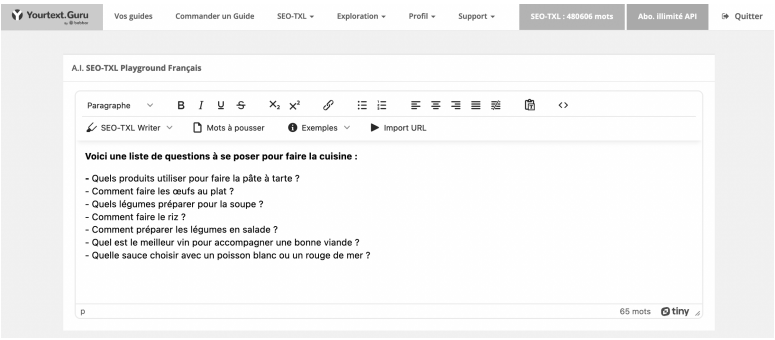


FIGURE 2.3 – utilisation de SEO-TXL pour trouver des questions

2.4.4 Organiser son cocon avec le maillage interne

Maintenant que l'on dispose de toutes les questions (les intentions) que l'on souhaite intégrer à notre cocon, la grande question est de savoir comment organiser le cocon. C'est-à-dire comment faire son maillage interne.

Pour cela, on va vouloir associer ensemble les intentions qui sont proches. C'est-à-dire celles pour lesquelles les contenus qui seront présents sur les pages les traitants seront proches.

Pour chaque question à insérer dans le cocon, je fabrique donc un guide `yourtext.guru`. Puis j'utilise la fonctionnalité de cocon afin d'afficher pour chaque question les 10 autres questions les plus proches sémantiquement.

La figure2.4 montre la relation sémantique de la SERP entre le guide central et les 10 guides avec les plus fortes relations sémantiques avec ce guide. L'idée maintenant va être de prendre le guide central pour en faire un contenu qui sera publié sur une page de mon cocon. En prenant alors les hauts scores de relation sémantique, je saurais vers quels autres types de contenus / pages pointer pour maximiser le transfert de pagerank thématique.

Ici, si je souhaite faire 3 liens (n'hésitez pas à varier le nombre de liens : si vous êtes sur un cocon au sein d'un seul site, vous pouvez y aller gaiement. Si le cocon est fabriqué entre plusieurs sites, soyez plus restreints dans vos ambitions) depuis « comment faire un livre de recettes de cuisine », je

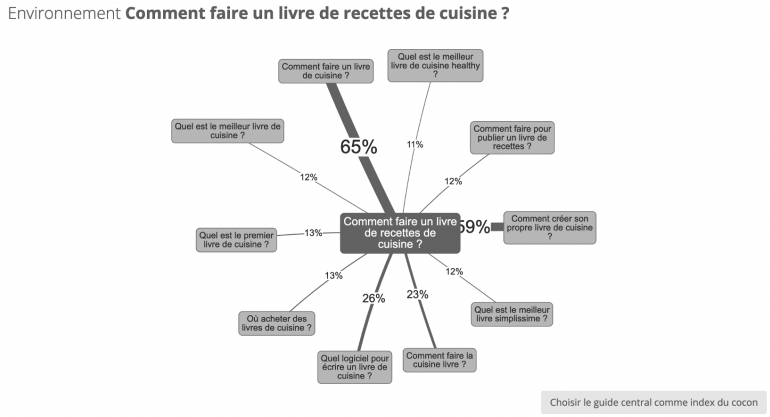


FIGURE 2.4 – Maillage via la fonctionnalité Cocon

pointerai vers « comment faire un livre de cuisine », « comment créer son propre livre de cuisine » et « quel logiciel pour écrire un livre de cuisine ».

C'est un excellent exemple pour aborder une problématique importante : ici les scores de similarité sémantique sont parfois très élevés (> 40-50%). Cela veut dire que les SERPs affichent des pages qui sont très semblables. Donc on doit pouvoir se permettre de dire qu'il s'agit de contenus tellement proches qu'on peut les traiter ensemble. Faire une seule page pour traiter « comment faire un livre de recettes de cuisine », « comment faire un livre de cuisine » et « comment créer son propre livre de cuisine » semble plus adéquat. En général un score > 50% amène à fusionner les contenus pour ne faire une seule page. Entre 40 et 50% on se pose humainement la question. En dessous, c'est généralement OK de conserver les contenus sur des pages différentes.

Finalement, « comment faire un livre de recette de cuisine » pointera donc vers « quel logiciel pour écrire un livre de cuisine », « comment faire la cuisine, livre » et « où acheter des livres de cuisine ».
Maintenant je peux généraliser ce maillage en passant en revue toutes les relations entre elles, j'obtiens alors le maillage global de mon cocon. Parfois les liens seront réciproques, parfois ils ne le seront pas.
Attention, ce maillage repose sur la similarité sémantique, il permet ainsi

d'optimiser le transfert du pagerank sémantique. C'est le but du cocon, alors c'est parfait. Mais il ne faut pas oublier l'internaute qui parcourt le cocon ! Alors, en dehors de ces liens faits directement au cœur de vos articles, il ne faut pas oublier d'avoir une navigation faite pour l'humain. Là où le moteur de recherche trouvera les liens hypertextes pleins de bon sens, l'internaute humain aura besoin d'un menu plus classique. N'hésitez donc pas à faire cohabiter les liens optimisés pour faire un cocon SEO avec une arborescence habituelle.

Si jamais pour une intention donnée les scores alentours sont très faibles (<10), on peut se demander si les relations sémantiques seront assez fortes entre intentions pour garantir une bonne optimisation. Il est recommandé alors d'agir, soit en supprimant le contenu du cocon, ou alors d'ajouter encore des intentions que l'on pense proches, puis de refaire le calcul du cocon pour le vérifier. Un lien efficace entre deux pages doit permettre une navigation qui ne change pas complètement le contexte.

2.4.5 Rédiger les contenus du Cocon SEO

Maintenant que vous êtes en possession de la liste des contenus à écrire, et des liens à placer sur chacune des pages de contenu. Il est temps de rédiger.

Lorsque nous avons calculé les relations sémantiques entre les intentions/questions, c'est bel et bien le contenu des guides qui a été pris en compte. Cela veut dire que pour avoir le meilleur transfert de pagerank thématique, il est indispensable de rédiger en optimisant vos contenus pour les guides reliés.

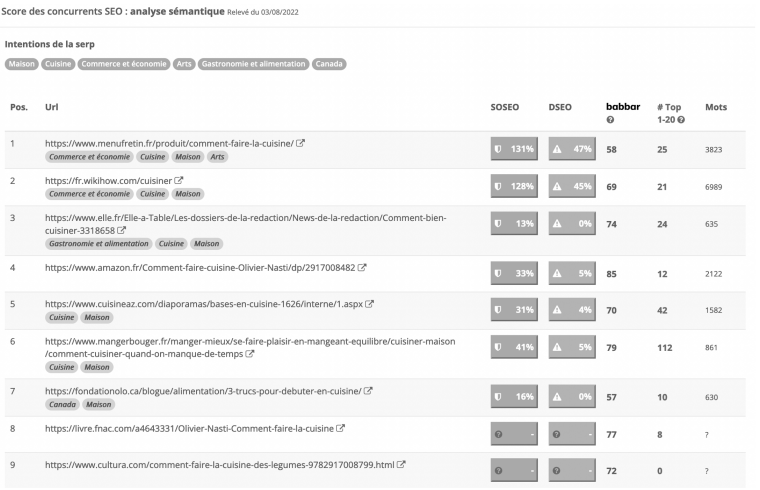
Si on fait un beau cocon mais que les guides ne sont pas écrits en suivant les bonnes pratiques d'optimisation sémantique, alors l'effet du cocon va être aussitôt réduit.

Pour bien écrire avec l'envie d'optimiser pour le référencement web, c'est facile : il faut utiliser les bons mots, ceux qui ont de l'importance pour le moteur de recherche. Pour cela on « imite » les premiers de la classe sur la page des résultats du moteur de recherche.

Pour bien faire les choses, nous utilisons `yourtext.guru` comme outil d'analyse sémantique. Pour comprendre son fonctionnement en détail, vous pouvez suivre les formations gratuites à `yourtext.guru` sur la Babbar Academy (`https://babbar.academy/`).

Dans `yourtext.guru`, on regarde les scores des premiers de la SERP,

puis on va chercher à s'aligner sur eux, en faisant un peu mieux. Il n'y a pas besoin de faire extrêmement mieux, un peu mieux suffit : le moteur de recherche étant une machine à classer, on veut être meilleur, mais pas forcément atteindre la suroptimisation.



Score des concurrents SEO : **analyse sémantique** Relevé du 03/08/2022

Intentions de la serp

Maison | Cuisine | Commerce et économie | Arts | Gastronomie et alimentation | Canada

Pos.	Url	SOSEO	DSEO	babbar	# Top 1-20	Mots
1	https://www.menufretin.fr/produit/comment-faire-la-cuisine/ — Commerce et économie, Cuisine, Maison, Arts	131%	47%	58	25	3823
2	https://fr.wikihow.com/cuisiner — Commerce et économie, Cuisine, Maison	128%	45%	69	21	6989
3	https://www.elle.fr/Elle-a-Table/Les-dossiers-de-la-redaction/News-de-la-redaction/Comment-bien-cuisiner-3318658 — Gastronomie et alimentation, Cuisine, Maison	13%	0%	74	24	635
4	https://www.amazon.fr/Comment-faire-cuisine-Olivier-Nasti/dp/2917008482	33%	5%	85	12	2122
5	https://www.cuisineaz.com/diaporamas/bases-en-cuisine-1626/interne/1.aspx — Cuisine, Maison	31%	4%	70	42	1582
6	https://www.mangerbouger.fr/manger-mieux/se-faire-plaisir-en-mangeant-equilibre/cuisiner-maison/comment-cuisiner-quand-on-manque-de-temps — Cuisine, Maison	41%	5%	79	112	861
7	https://fondationolo.ca/blogue/alimentation/3-trucs-pour-debuter-en-cuisine/ — Canada, Maison	16%	0%	57	10	630
8	https://livre.fnac.com/a4643331/Olivier-Nasti-Comment-faire-la-cuisine	-	-	77	8	?
9	https://www.cultura.com/comment-faire-la-cuisine-des-legumes-9782917008799.html	-	-	72	0	?

FIGURE 2.5 – Les scores de la SERP

Par exemple, pour faire un contenu sur « Comment faire la cuisine », la SERP (voir la figure 2.5) donne des scores sémantiques pour la concurrence SEO. Une bonne pratique passe-partout est :

- ⊙ avoir un score SOSEO un peu plus haut que le score moyen de SOSEO des trois premiers de la SERP ;
- ⊙ avoir un score DSEO plus bas que le score moyen des cinq premiers de la SERP. Il n'y a pas de limite basse.

Sur cet exemple, cela veut dire que l'on va viser un SOSEO un peu plus haut que 90 et un DSEO plus petit que 20. En atteignant ces objectifs, on aura un texte bien optimisé pour le SEO.

En dehors de cette optimisation pour le référencement web, il ne faut pas oublier de prévoir où seront placés les liens du Cocon SEO. En ayant d'avance une idée des liens que l'on doit faire, on peut écrire quelques mots, quelques phrases, voire un paragraphe complet pour introduire des

liens réellement contextuels. Qui a envie de voir un lien apparaître au sein d'un article comme un cheveu sur la soupe ?

En plus du contexte du paragraphe d'où le lien est fait, il faut prêter attention au texte d'ancrage. Ce texte est vu par le moteur de recherche comme étant une extension de contenu pour la page liée.

Par exemple, en faisant un lien sur le texte « cuisine » d'une page A vers une page B, le contenu de la page B devient plus fort pour le mot « cuisine » sans que l'optimisation sémantique de la page B ne devienne suroptimisée (c'est tellement efficace que Google a mis en place un filtre appelé Penguin il y a quelques années pour éviter les abus de linking entre sites).

On a donc tout intérêt à placer les liens sur quelques mots, bien adaptés, avec une forte adéquation sémantique avec les contenus de la page cible. C'est aussi là où l'on peut mettre le plus facilement l'expression exacte sur laquelle on souhaite positionner la page. Afin de rester naturel, on va généralement mettre un texte proche de la requête ciblée, mais un peu plus large, et on réservera les textes d'ancres exacts à quelques liens pointant vers les pages que l'on souhaite vraiment pousser.

2.4.6 Comment mettre en ligne les pages du cocon ?

Une fois que l'on est en possession de tous les contenus et tous les liens, la façon la plus facile pour publier le cocon SEO est de mettre en une fois tout le contenu en ligne, à disposition de Google.

De cette façon on est tranquille et on laisse le moteur de recherche découvrir les pages à son rythme. On peut même faire un peu d'acquisition de liens pour aider à la découverte et rapidement permettre d'obtenir des positions.

En pratique, quand on a un cocon de plusieurs centaines de pages, il faut parfois plusieurs mois pour en rédiger les contenus. Et on ne va pas attendre plusieurs mois avant de commencer à profiter de l'effet Cocon SEO : plus vite les contenus seront en ligne, mieux ce sera ! A la fois pour permettre de commencer à obtenir des positions de longue traine, de créer des opportunités d'acquisition de liens, et pour améliorer le référencement global du site sur lequel on travaille et plus particulièrement le nœud central de notre cocon.

Dans ce cas-là, où l'on ne peut publier le Cocon SEO qu'au compte-goutte, on va adopter une stratégie un peu différente :

- ⊙ Quand on met en ligne une page, on peut faire des liens uniquement vers des pages déjà en ligne ;
- ⊙ Si on publie une nouvelle page du cocon et que cela ouvre, maintenant, une possibilité de faire un lien depuis une page déjà publiée, alors on va :
  - ◇ Ne pas ajouter de lien depuis la page déjà existante, c'est trop artificiel ;
  - ◇ Ou ajouter le lien depuis la page déjà publiée mais alors on va ajouter ou modifier un paragraphe complet de l'article pour placer le lien. L'idée est alors de montrer que l'apparition du lien résulte d'une véritable modification de l'article, pas simplement de la création automatique du maillage interne.

L'avantage d'être dans une optique où l'on publie au fur et à mesure est qu'on peut se permettre d'étendre un Cocon SEO déjà en ligne, que ce soit pour amplifier son effet sur le positionnement ou pour introduire encore plus d'intentions de longue traîne.

2.5 Conclusion

Faire un cocon SEO, c'est avoir une stratégie éditoriale faite pour le SEO. En quelques heures, vous pouvez définir de nombreuses intentions à traiter pour améliorer le référencement de tout ou partie de votre site.
Si jamais vous avez un besoin éditorial qui sort du SEO (soyons réalistes, c'est presque toujours le cas), n'hésitez pas à intégrer les sujets de vos articles dans le calcul du cocon. Ces contenus ne seront peut-être pas les plus adaptés pour faires des liens internes qui soient les plus efficaces possibles, mais ce seront des contenus qui pourront tout de même être liés depuis des pages faites pour le référencement web ; elles pourront donc profiter du maillage interne.
De même, si vous n'avez aucune envie de faire un cocon SEO parce que publier autre chose que ce que votre stratégie éditoriale vous dicte est impossible, utilisez les outils de calcul de maillage du cocon SEO pour trouver le plus facilement possible les meilleurs liens à faire pour le SEO entre vos contenus bichonnés avec soin.
Tout peut devenir un cocon SEO. Et rien n'est parfait : en référencement web on tire le meilleur de ce que l'on peut faire en fonction de ses propres contraintes. Si on ne peut pas faire tout parfaitement, on le fait quand même, ce sera toujours plus efficace que de ne rien faire du tout.

3. Les enjeux du maillage interne

Frédérik Bobet est président de l'agence Trikaya Communication. Passionné de SEO et de technologies du web, il est l'un des co-créateurs de `yourtext.guru`, c'est d'ailleurs à son initiative que le projet avait été lancé. Il est également créateur de la chaine Youtube Trikaya Live Stream.

3.1 Introduction

Le maillage interne de votre site internet est crucial pour faciliter la compréhension des internautes et des robots de recherche. Présenter les bonnes informations aux endroits stratégiques de votre site permet de fluidifier la navigation de vos internautes, mais aussi de clarifier la compréhension des informations par les moteurs de recherche. Une bonne catégorisation permet de proposer des landing pages optimales qui visent des intentions précises et qui sont entourées d'éléments complémentaires ou connexes.
La circulation du pagerank interne a aussi son importance et un bon maillage interne permet d'optimiser concrètement l'interprétation de votre architecture par les robots de recherche.
Plusieurs éléments du maillage interne sont maintenant rentrés dans les conventions de développement, des éléments qu'on a tous l'habitude d'utiliser comme par exemple le menu principal, la navigation à facettes, la pagination, le fil d'Ariane, ou encore le pied de pages et bien d'autres

encore.

Je vous propose de découvrir dans ce chapitre comment on réalise un bon maillage interne.

3.2 Prérequis

Avant toute chose, il est indispensable de s'imprégner un maximum de la thématique sur laquelle on travaille et de la stratégie du site. On doit être en mesure de justifier chacune des décisions qui vont être prises : pas de place au hasard ici !

- ⊙ Quels sont les objectifs du site à court, moyen et long terme ?
- ⊙ Quelles sont les attentes de l'audience cible ?
- ⊙ Identifiez le plus précisément possible les principaux personas marketing.
- ⊙ Procédez à une recherche de mots clés exhaustive et priorisez-les afin de concevoir une véritable stratégie mots clés. Pour cela vous pouvez regarder la vidéo de la figure 3.1 [1].
- ⊙ Une fois en possession de votre stratégie mots clés bien travaillée, vous allez pouvoir passer à la conception d'une structure hiérarchique optimale en vous appuyant sur la data contenue dans votre stratégie mots clés et concevoir une structure *data driven* (figure 3.2 et vidéo associée [2]). profitez de ce moment pour :
  - ◇ réfléchir à la typologie de chaque contenu ;
  - ◇ établir un plan d'action à court terme : Quick win et longue traîne ;
  - ◇ établir un plan d'action à moyen terme : moyenne traîne et contenu moins vente directe ;
  - ◇ établir un plan d'action à long terme : courte traîne et comment devenir LE média.

  Souvent la réflexion sur la structure permet d'aider les décisions UX. N'hésitez pas à partager vos fichiers avec l'équipe d'UX designers, qui les accueillerons avec des étoiles dans les yeux.

1. `https://youtu.be/ehGQfQtGkNw`
2. `https://youtu.be/xusQA9bECl g`

FIGURE 3.1 – Trouver les bons mots-clés en SEO.

3.3 Définition d'un lien hypertexte

Je me permets de redonner ici la définition d'un lien hypertexte, car je vois quelques aberrations apparaître parfois.
Un lien est défini par une page source, une page de destination et une ancre, il est représenté en HTML par la balise `<a href="URL_target">Ancre</a>`.
Les liens `mailto, tel, #` ne sont pas considérés comme des liens et ne consomment donc pas de pagerank.

3.4 Le crawl

Le travail de crawl effectué par les robots de recherche est la base de découverte de votre site et de ses pages.
Le crawl consiste à découvrir l'intégralité de la face visible d'un site internet en découvrant de manière itérative l'ensemble des pages qui sont reliées à la structure publique du site.
Le robot arrive sur la page d'accueil et découvre les liens qui sont présents sur cette page, il visite ces nouvelles URLs qui présentent elles-mêmes de nouveaux liens vers de nouvelles URLs et ainsi de suite jusqu'à ne plus trouver de nouvelles URLs.
Comme vous pouvez le comprendre ce processus itératif suit un ordre bien précis et va nous permettre, pour chaque page, de lister l'intégralité

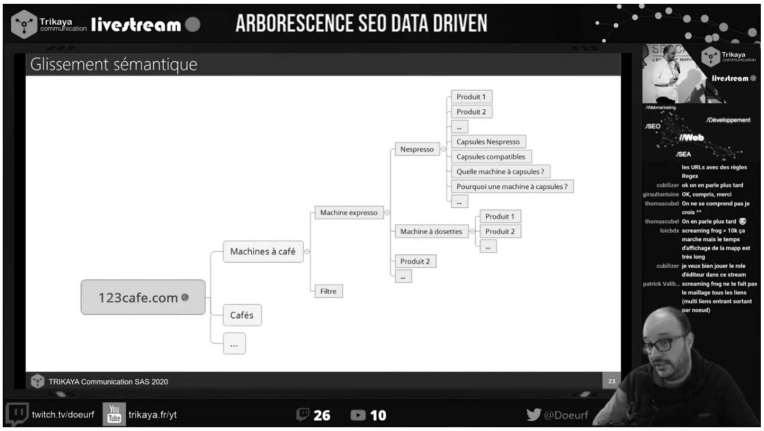


FIGURE 3.2 – Arborescence SEO *Data Driven*.

des pages qui lui font référence. On peut aussi conserver une information importante qui est le niveau de profondeur de chaque page.

Si vous voulez que votre site soit bien compris par les robots de recherche, il faut commencer par travailler les chemins de crawl vers chaque ressource proposée sur votre site web. Plus l'entourage d'une page sera pertinent, plus la page sera pertinente.

Une notion importante à bien comprendre est la notion de chemin le plus court, en effet même s'il existe plusieurs cheminements possibles pour atteindre une page, il y aura au moins un chemin le plus court. Cette notion comprend aussi le chemin le plus court onpage (un lien en haut d'une page sera découvert avant un lien en bas de page).

Il faut bien comprendre cette notion, car c'est cela qui permet de réaliser les représentations graphiques qu'on rencontre le plus souvent. Afin de ne pas alourdir inutilement la visu graph, on représente uniquement le chemin le plus court vers une ressource. C'est évident que si on tente de représenter l'intégralité des liens d'un site sur une même image, cela risque d'être illisible et inexploitable (voir la figure 3.3).

Ce chemin le plus court est celui qui embarque les informations les plus directes qui circulent jusqu'à la page.

Ce qu'il faut bien comprendre ici c'est qu'un lien vers une page n'est pas qu'une simple URL source avec une ancre, mais un lien est un contexte.

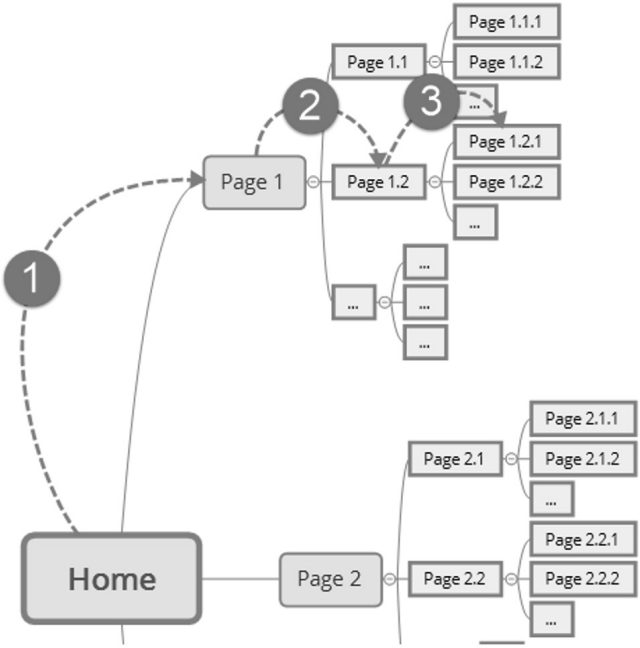


FIGURE 3.3 – Plus court chemin.

On comprend rapidement que si tous les contextes qui circulent vers une page sont précis et pertinents plus le contexte de la page sera renforcé et donc pertinent.

Tous les chemins vers une ressource donnée sont importants et il devient facile de comprendre qu'on peut aussi bien renforcer le contexte entourant une ressource que le parasiter.

C'est très important de comprendre cette notion de cheminements de navigation vers une ressource qui représente un contexte.

J'ai pour habitude d'utiliser un exemple simple, mais efficace. Imaginons que je vende des produits de sport de montagne et que depuis mon mega-menu j'ai directement un lien vers « fond » et « alpin », sorti de tout contexte on ne sait pas si je parle du fond de la vallée, du massif alpin, etc.

Si par contre je force le robot de recherche à passer par la catégorie « skis » avant de découvrir « fond » et « alpin » alors j'ai ajouté un contexte qui va permettre au robot de recherche de mieux comprendre de quoi il s'agit.
Lors du crawl avec un outil comme Screaming Frog vous pourrez vous apercevoir de problématiques comme des pièges à robots (spider trap) qui vous donneront l'impression que le crawl ne se terminera jamais, d'ailleurs c'est parfois le cas !
Enfin, pensez à vérifier le nombre de pages indexées par Google avec la commande `site:ndd.com` versus le nombre de pages HTML indexables crawlées par votre logiciel. Si vous constatez une très grande différence entre les 2 il sera nécessaire de se pencher sur l'existence de pages orphelines.

3.5 Le pagerank

Créé par Larry Page, d'où son nom Pagerank - et oui ce n'est pas le rang des pages, mais bien le nom du créateur – cet algorithme permet de calculer l'importance d'une page au sein d'un réseau de pages maillées par des liens hypertextes. La formule d'origine ci-dessous peut vous paraître obscure si vous n'êtes pas familier avec les mathématiques

$$PR(u) = \frac{(1-c)}{N} + c \cdot \sum_{v \to u} \frac{PR(v)}{\#liens(v)}$$

Pour faciliter la manipulation de cet objet on peut se permettre de faire un raccourci brutal : Plus une page a de liens qui pointent vers elle, plus elle est importante au sein du réseau de page.

3.5.1 Surfeur aléatoire

Le pagerank est un modèle mathématique qui a pour but de simuler le comportement d'un internaute qui visiterait des pages de manière aléatoire.
Imaginons qu'on a 100 watts de pagerank à distribuer depuis une page donnée, si cette page comporte 2 liens, chaque lien embarquera 50 watts chacun (autrement dit l'internaute aléatoire a une chance sur 2 de cliquer sur un lien ou l'autre), si la page propose 100 liens, chaque lien n'embarquera plus que 1 watt chacun (autrement dit l'internaute aléatoire a une chance sur 100 de cliquer sur un des liens).
Notez bien que le principe est le même avec des liens nofollow, si la page propose 99 liens nofollow et 1 lien « dofollow » alors celui-ci

n'embarquera qu'1 watt et les 99 autres watts seront perdus.

Conclusion : n'utilisez pas de liens nofollow !

Même chose si vous envoyez des liens « dofollow » vers des pages noindex.

3.5.2 Surfeur raisonnable

La notion de surfeur raisonnable est un peu plus fine que celle de surfeur aléatoire vue précédemment, puisqu'ici on va considérer que la place du lien dans la page a une importance. Ainsi un lien tout en haut à gauche de la page aura plus de chances d'être cliqué par un internaute lambda – enverra plus de jus - qu'un lien tout en bas à droite de la page (si on considère des pages dans une langue écrite de gauche à droite évidemment).
Ici en revanche, impossible de connaître le niveau de pondération appliqué par Google, nous savons juste qu'un lien très haut dans la page sera plus puissant qu'un lien très bas dans la page. Il est aussi souvent admis que les liens faisant partie de l'interface sont pondérés négativement par Google.
Tout au long de ce chapitre, nous raisonnerons plutôt en surfeur aléatoire qu'en surfeur raisonnable, moins facile à manipuler. Nous pourrons toutefois nous permettre de réfléchir en surfeur raisonnable à l'échelle *onpage* simplement en gardant à l'esprit qu'un lien plus haut est plus puissant qu'un lien plus bas dans la page.

3.6 Pagerank et niveau de profondeur

La plupart du temps le pagerank interne décroit en fonction du niveau de profondeur, plus une page est profonde moins elle reçoit de pagerank des autres pages. A contrario la page d'accueil est censée être la page qui reçoit le plus de pagerank puisqu'elle est censée être maillée depuis toutes les pages du site.
Les études à grande échelle fournies par les grands noms de l'analyse de logs comme Oncrawl, révèlent exactement ce phénomène, comme l'illustre la figure 3.4.
On voit clairement qu'à partir du niveau 4 le taux de crawl décroit significativement.
En effet, la plupart des moteurs de recherche priorisent leur travail de crawl par rapport au pagerank des pages. Ainsi plus une page va être profonde, moins elle a de pagerank, moins elle est crawlée.

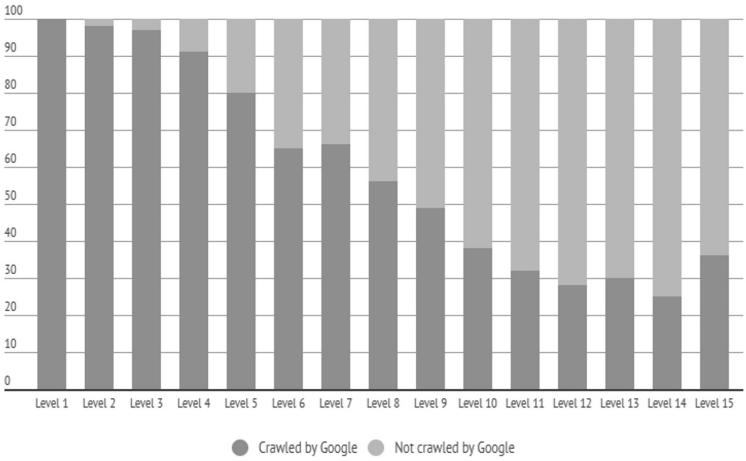


FIGURE 3.4 – Etude Oncrawl montrant le taux de crawl en fonction du niveau de profondeur.

Afin d'analyser finement la répartition des pages de votre site par niveaux de profondeur, vous pouvez utiliser Screaming Frog et vous rendre dans l'onglet Site Structure de la partie droite du logiciel. Vous pouvez voir le graphique qui représente la distribution des pages d'un site par niveau de profondeur (voir figure 3.5).

3.7 Les impacts du maillage interne

Le maillage interne a des impacts notables sur plusieurs facteurs importants pour les moteurs de recherche, il influence directement sur plusieurs aspects.

3.7.1 La circulation du pagerank

Jouer sur la valeur du pagerank de chacune des pages nous permet de montrer aux moteurs de recherche où sont les priorités du site. Nous nous attarderons à apporter beaucoup de pagerank sur les pages visant les requêtes les plus génériques / les plus concurrentielles et nous aurons besoin de moins de puissance sur des pages visant des requêtes de longue traîne peu concurrentielles.

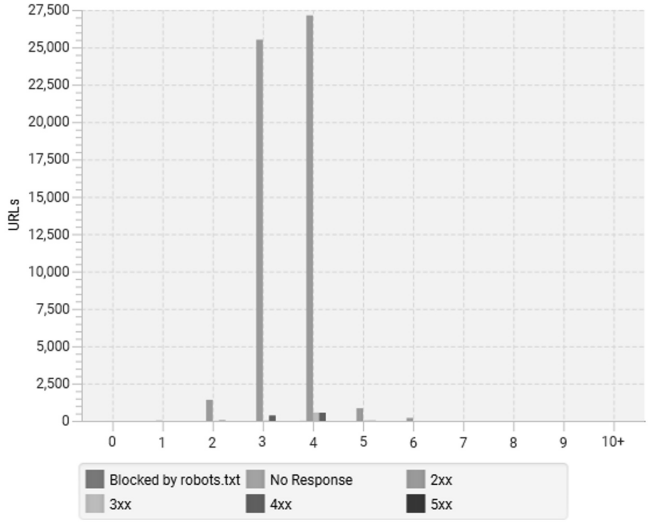


FIGURE 3.5 – Graphique Screaming Frog Niveaux de profondeur / Crawl Depth.

3.7.2 Le cloisonnement sémantique

Afin d'éviter tout parasitage contextuel entre les différents univers traités sur notre site, nous pouvons travailler notre maillage interne pour enfermer chaque thématique dans son propre silo sémantique. L'idée est de canaliser le cheminement du crawler pour qu'il reste « emprisonné » dans un univers une fois qu'il y est entré, et donc que le surfeur aléatoire ait beaucoup plus de chances d'aller cliquer sur un lien de la même thématique plutôt que lui laisser la possibilité d'aller sauter du coq à l'âne.

C'est exactement pour cela que l'utilisation d'un mega-menu est fortement déconseillée par les référenceurs tout comme l'implémentation d'un système de maillage automatique.

3.7.3 Les sitelinks

La détermination des sitelinks par Google est fortement influencée par la valeur du pagerank interne des pages. Il a tendance à afficher les pages représentatives de l'activité du site qui ont le plus fort pagerank interne.

Le comportement des utilisateurs vis-à-vis du site est aussi pris en compte et peut venir mettre son grain de sel dans cette sélection.

3.8 Les outils de conception et visualisation

3.8.1 Outils de mind mapping

Il existe pléthore d'outils de mind mapping - ou carte mentale - comme Mindmeister, Freemind, Maxmind, Gloomaps, Coggle, Xmind.
Personnellement j'affectionne tout particulièrement Xmind 8 pro (attention pas la version Xmind Zen) pour sa prise en main facile, sa gratuité et le design final des cartes mentales qu'il permet de concevoir.
Les outils de mind mapping sont très utiles pour concevoir votre structure en partant de zéro ou pour schématiser la structure de maillage interne d'un site existant.
Afin de faciliter la description du travail à réaliser sur la structure de maillage interne d'un site, je vous conseille de toujours faire une schématisation du maillage de chaque page type. Vous pourrez ainsi décrire précisément les améliorations à apporter sur chaque template de page qui constituent le site.
Par exemple, pour un site e-commerce, on peut identifier les types de pages suivantes :

- home ;
- catégorie ;
- produit ;
- marque.

Pour les e-commerces les plus avancés en termes de maillage interne, vous trouverez en plus les typologies de pages suivantes :

- catégorie + marque ;
- catégorie + facette ;
- catégorie + facette + marque.

Chaque type de page va proposer des liens vers des types de pages complémentaires, on trouvera ainsi assez logiquement pour un site de bricolage :

- un lien vers la marque depuis la page produit ;

- ⊙ des liens vers des produits similaires depuis la page produit ;
- ⊙ des liens vers catégorie + marque depuis la page catégorie (Perceuse sans fil Bosch) ;
- ⊙ un lien vers catégorie + marque depuis la marque ;
- ⊙ des liens vers les facettes depuis la catégorie (Perceuse sans fil 18V) ;
- ⊙ des liens vers catégorie + marque + facette depuis catégorie + marque ou catégorie + facette (Perceuse sans fil Bosch 18V).

On s'attardera donc à schématiser chaque typologie de page et les interactions qu'elle a avec les autres typologies. Cela permet également de travailler par template de page et donc de s'adresser plus facilement à l'équipe de développement.

3.8.2 Screaming Frog (payant)

Screaming Frog est la boîte à outils par excellence du référenceur aguerri ! Depuis quelques années ce crawler SEO propose la possibilité de visualiser le maillage interne de votre site.
Ce type de visualisation permet de détecter très rapidement des problématiques discrètes qui auraient été difficiles à identifier manuellement même au cours du travail de schématisation précédemment exposé.
Vous allez pouvoir identifier des erreurs de redirections 301 ou 302 internes, des erreurs de canonicals mal paramétrées, etc.
Ce type de visualisation vous permet aussi et surtout d'identifier les grands univers du site et s'ils sont bien cloisonnés.
Grâce à l'outil de visualisation de Screaming Frog vous allez aussi pouvoir faire varier la taille des nœuds en fonction de différentes métriques :

- ⊙ link score (pagerank) : Si vous voulez visualiser en fonction du pagerank il faudra au préalable lancer un crawl analysis / analyse de crawl depuis le menu dédié ;
- ⊙ niveau de profondeur ;
- ⊙ nombre de liens entrants ;
- ⊙ nombre de mots.

Vous pouvez aussi paramétrer pas mal de choses comme :

- ⊙ les couleurs de certains nœuds spécifiques ;

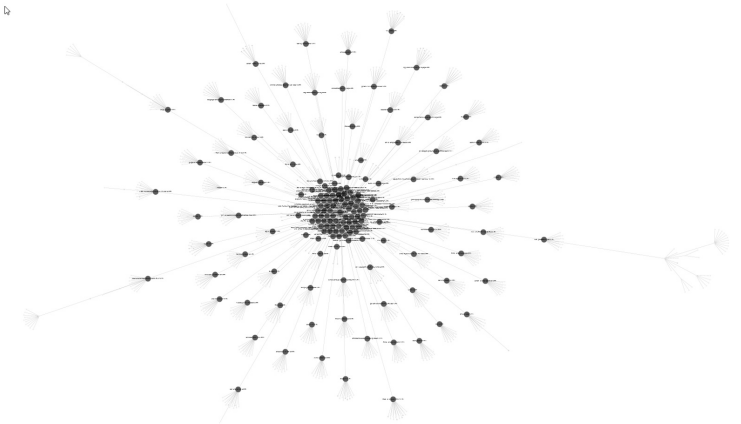


FIGURE 3.6 – Rendu visualisation Screaming Frog.

- ⊙ la taille des liens ;
- ⊙ la taille des nœuds ;
- ⊙ les niveaux de profondeur affichés ;
- ⊙ mettre en surbrillance des nœuds grâce à une expression régulière.

Cet outil compris dans l'abonnement annuel de Screaming Frog vous permettra ainsi de suivre régulièrement votre maillage interne et son évolution.

Les seuls inconvénients à mon sens sont les suivants :

- ⊙ peu présentable pour un client ou pour la direction ;
- ⊙ difficile de faire une comparaison entre 2 visualisations ;
- ⊙ représentations parfois assez brouillonnes avec des branches qui s'enroulent autour du centre.

3.8.3 Gephi (gratuit)

Outil open source développé à la base pour l'analyse scientifique des réseaux, il peut aussi s'avérer intéressant pour le référencement naturel. Cela dit il faudra de la patience et de l'acharnement pour prendre en main ce logiciel assez obscur aux premiers abords.

Crawl et exports

Afin de réaliser une visualisation sous forme de graph de votre site il faut commencer par réaliser un crawl avec un logiciel comme Screaming Frog ou Xenu. Pour l'exemple nous prendrons Screaming Frog.
Une fois votre crawl terminé vous procéderez à :

- ⊙ l'export pour l'ensemble de vos pages (Onglet Internal + type de pages HTML -> export) et enregistrez sous nodes.csv ;
- ⊙ l'export de l'ensemble de vos liens internes (Bulk export -> Links -> All Inlinks) et enregistrez sous edges.csv.

Traitement et nettoyage des fichiers exportés

Editez le fichier nodes.csv avec Excel et renommez la colonne Address en Id. Vous pouvez aussi conserver la colonne Title et la renommer en Label. Conservez les colonnes status code et inlinks.
Editez le fichier edges.csv et renommez la colonne destination en Target et la colonne Anchor en Label. Ne gardez que Source, Target et Label.
Faites bien attention que vos fichiers respectifs soient encodés en UTF-8.

Import dans Gephi

Une fois tous ces traitements effectués on va pouvoir passer à la phase d'import dans Gephi.

1. Créez un nouveau projet et enregistrez-le sur votre disque.
2. Allez dans laboratoire de données.
3. Cliquez sur importer feuille de calcul et sélectionnez votre fichier nodes.csv en sélectionnant table de nœuds dans la liste « importer en tant que ».
4. Une fois l'import des nœuds terminé, cliquez de nouveau sur importer une feuille de calcul, sélectionnez cette fois-ci votre fichier edges.csv et sélectionnez table de liens dans la liste « importer en tant que ».
5. Une fois vos imports terminés, rendez-vous dans vue d'ensemble.
6. Dans la sidebar spatialisation, sélectionnez Force Atlas 2 et configurez les éléments de spatialisation comme le montre la figure 3.7.
7. Réglez votre nombre de processus en fonction de votre processeur et cliquez sur exécuter.
8. Attendez que la visualisation se stabilise.

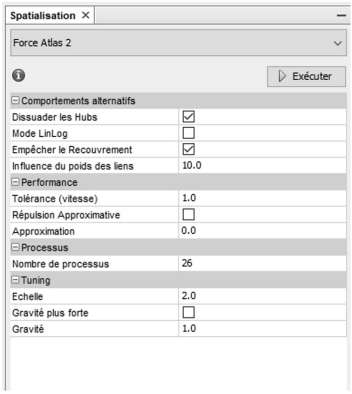


FIGURE 3.7 – Configuration de Gephi.

La figure 3.8 montre le même site que celui observé dans Screaming Frog précédemment.
Vous pouvez ensuite lancer une séquence de calcul du pagerank sur tout le graph depuis la boite à outil statistique située dans la sidebar de droite en cliquant sur le bouton exécuter en face de pagerank.
Une fois le calcul terminé, vous pourrez soit afficher cela directement dans le graph en dimensionnant les nœuds en fonction de la valeur du pagerank.
Vous pourrez aussi accéder à cette nouvelle donnée directement dans le laboratoire de données pour l'exploiter sous forme de données tabulaires, dans votre tableau de nœuds.
Voilà vous aurez compris que cela demande un gros travail et que ce n'est pas évident à exploiter, mais on peut faire des analyses très poussées grâce à Gephi et on peut croiser toutes les données que l'on souhaite.

3.8.4 Cocon.se (payant)

L'outil Cocon.se est un outil SAAS très facile à utiliser puisqu'il est doté d'un crawler interne, il suffit de lui donner l'URL de la page d'accueil de votre site, de configurer quelques paramètres et de lancer un crawl. Une fois le crawl terminé, vous avez directement accès à la visualisation du graph de votre site (voir la figure 3.9).
Hormis la facilité d'utilisation de cet outil, il a le mérite d'offrir des rendus

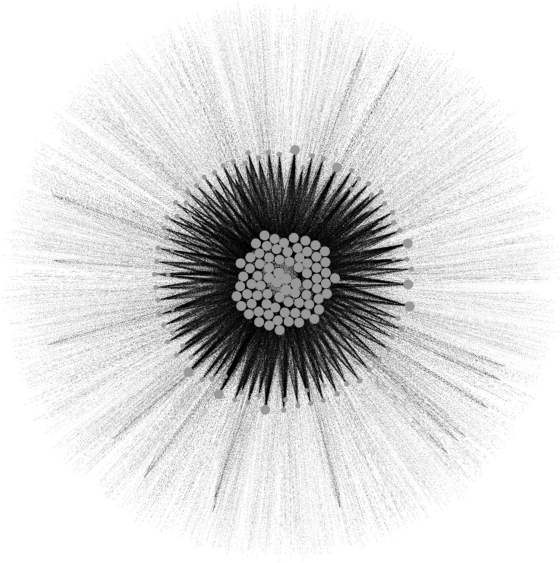


FIGURE 3.8 – Visualisation dans Gephi.

jolis et vendeurs !

De plus chaque nœud embarque 2 informations qui permettent de faciliter la lecture de ces documents :

— la couleur qui représente le niveau de profondeur de la page ;
— la taille du nœud qui représente la valeur du pagerank interne de la page.

Enfin c'est le seul outil que je connaisse qui permet de s'y retrouver quand on compare un graph avant modification et après modification de la structure (voir la figure 3.10).

Bien sûr il faut avoir pratiqué beaucoup de fois pour être à même d'exploiter les visualisations graph, mais une fois qu'on y a gouté on ne peut plus s'en passer ! Ces représentations permettent de gagner un temps précieux dans vos analyses, elles permettent d'avoir une vision macro et d'aller mettre le doigt sur des problèmes de maillage interne bien souvent difficiles à détecter à la main.

Pour moi l'outil le plus abouti du marché actuellement !

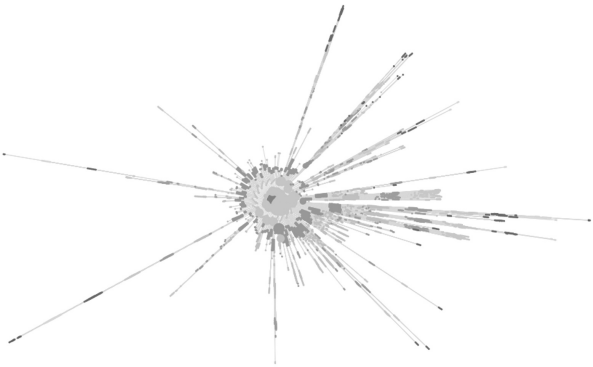


FIGURE 3.9 – Exemple de visualisation obtenue avec cocon.se.

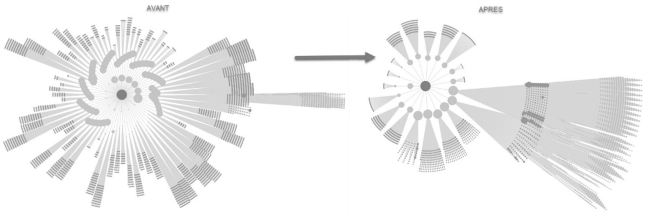


FIGURE 3.10 – Illustration avant / après modification de structure avec cocon.se.

3.8.5 Seolyzer.io (payant)

Seolyzer est un outil SAAS d'analyse de logs qui propose comme tout bon analyseur de logs un crawler interne.
Une fois votre 1er crawl réalisé avec l'outil, vous aurez accès à l'outil de visualisation embarqué.
Les visualisations graphiques sont très propres et exploitables.
Mais là où l'on va trouver la valeur ajoutée de cet outil pour l'analyse du maillage interne, c'est dans les analyses croisées.
Comme je vous le disais précédemment Seolyzer est un outil d'analyse de logs, et le fait de croiser les données issues des logs avec celles issues du crawler va nous apporter des informations cruciales pour détecter les

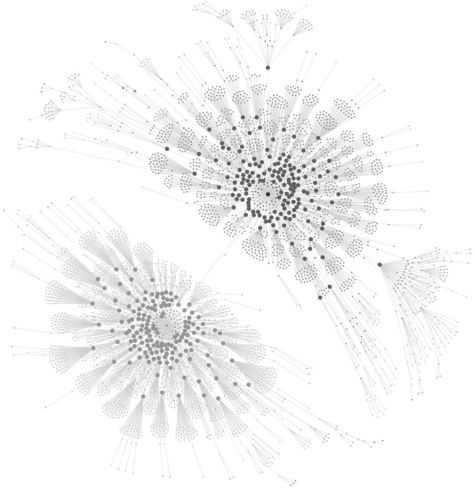


FIGURE 3.11 – Représentation graphique Seolyzer.

pages orphelines.
Les pages orphelines sont donc les pages qui ont été trouvées dans les données de logs serveur (qui sont connues et visitées par Google), mais qui n'ont pas été trouvées dans les données de crawl.
Les pages orphelines sont des pages qui n'ont plus aucun lien depuis la structure principale du site. Si vous avez bien suivi ce qui a été dit jusque là vous comprendrez rapidement qu'une page orpheline ne reçoit donc aucun pagerank interne et que cela peut poser de graves problèmes en termes de capacité de positionnement dans les moteurs de recherche.
En plus de ne pas recevoir de pagerank depuis la structure principale, les pages orphelines ne contribuent pas non plus à la circulation du pagerank.

3.9 Les bonnes pratiques du maillage interne

3.9.1 Navigation principale et navigations transversales

Si vous souhaitez travailler votre maillage interne, il faut bien comprendre les enjeux de votre marché et les habitudes de navigation des internautes cibles.

Votre navigation principale doit être déterminée pour remplir cet objectif principal : un internaute propulsé au beau milieu de votre site par un moteur de recherche doit pouvoir s'y retrouver sans mode d'emploi.
C'est ici le point de rencontre entre le marché, le référencement naturel et l'UX.
Déterminer votre navigation principale c'est déterminer vos verticales, pour un magasin de bricolage ce sera, entre autres :

- ⊙ outillage
- ⊙ quincaillerie
- ⊙ jardin.

Une fois que vous avez déterminé vos verticales principales, il sera simple de déterminer vos navigations transversales. C'est-à-dire des navigations qui ont des points communs avec différentes verticales. Le meilleur modèle pour comprendre cette notion de transversale est le modèle des marques. Une marque peut proposer des produits dans plusieurs de vos rayons.
Pour reprendre la thématique du bricolage, qui parle à tout le monde, les transversales seront :

- ⊙ marque ;
- ⊙ caractéristiques produit (facettes) ;
- ⊙ guides éditoriaux (jardinage vient en transversale de la verticale produit jardin).

Afin de capitaliser sur vos navigations transversales il faudra trouver des points de rencontre avec vos verticales principales dans le but de renforcer le contexte autour de vos verticales principales.
Il faudra évidemment prendre en compte l'intérêt SEO de ces points de rencontre avant d'investir dedans. Dans le e-commerce ces nœuds de rencontre seront :

- ⊙ catégorie + marque ;
- ⊙ catégorie + facette(s) ;
- ⊙ catégorie + marque + facette(s).

Evitez d'aller à l'inverse des grands acteurs du marché, cela vous amènera plus de soucis que d'avantages. Google conçoit ses clusters statistiques en compilant l'ensemble des informations d'une thématique et si les grands

acteurs ont tous l'habitude de ranger de la même manière ils auront un impact statistique très fort. Google aura l'habitude de trouver une catégorie entourée de telle et telle autre catégorie et il aura donc du mal à comprendre une navigation qui proposerait un contexte inhabituel. Chaque glissement sémantique deviendrait non naturel entre vos pages et la pertinence du contexte de vos pages serait mauvaise.

3.9.2 Niveaux de profondeur

Il faut éviter d'avoir un site trop profond de manière générale. Mais attention cela ne veut pas dire balancer toute sa hiérarchie de catégories dans un mega-menu. Si vous avez des raisons de créer un niveau de profondeur, n'hésitez pas.

Rangez naturellement les choses sans vous soucier du niveau de profondeur tant que vous maitrisez votre hiérarchie, que celle-ci est logique et qu'elle va bien du général au particulier.

On commencera à s'inquiéter des niveaux de profondeur excessifs quand on aura un volume de pages important au-delà du niveau 5 et surtout si le volume de pages se concentre au-delà du niveau 10.

La plupart du temps dans cette situation c'est que le site présente des spider trap (pièges à robot) :

- ⊙ Navigations à facettes totalement « open bar » permettant de générer un nombre quasi infini d'URLs.
- ⊙ Paginations très profondes avec trop peu d'éléments par page et/ou posant des problèmes de navigation (voir les suivants / voir les suivants…, système de pagination trop restreint, etc.).

3.9.3 Maillage de la home

C'est une convention depuis les origines du web, la home est maillée depuis toutes les pages du site (en général depuis le logo) et heureusement, car les moteurs de recherche auraient bien des problèmes à extraire l'ordre d'importance des données d'un site web.

Je tenais à le rappeler : n'oubliez pas de mailler la home depuis toutes les pages de votre site !

3.9.4 Maillage des catégories de premier niveau

Les catégories de plus haut niveau dans un site web pensé avec soin correspondent aux requêtes les plus génériques, donc très souvent les plus concurrentielles.
De plus ce sont souvent des points d'entrée vers vos différents univers.
Je conseille de les mailler depuis toutes les pages de votre site afin de leur donner le maximum de pagerank interne.

3.9.5 Mega-menu

Les mega-menus ont déjà fait couler beaucoup d'encre. Non seulement ils diluent fortement le pagerank interne que chaque page est capable de redistribuer, mais bien souvent ils créent aussi un parasitage sémantique important puisqu'ils créent des liens entre des pages qui n'ont aucune relation sémantique logique.
On ne peut toutefois pas leur enlever leurs avantages en termes d'UX, en particulier sur desktop.
Vous pouvez toutefois, si vous voulez continuer à travailler avec un mega-menu, utiliser la technique d'obfuscation dynamique de menu.

3.9.6 Maillage cocon sémantique

J'aborderai uniquement ici l'aspect technique d'implémentation de la brique de base du maillage cocon sémantique.
Le but est d'augmenter le pagerank interne d'une page dont on veut augmenter les capacités de ranking : la page mère.
On va créer des pages filles (si possible qui se positionneront sur un mot clé intéressant). La page mère propose un lien vers chacune de ses filles. On s'attachera à contextualiser ce lien en les intégrant dans un paragraphe dédié depuis la page mère afin que ces liens aient un contexte fort à leur gauche et à leur droite ; en d'autres termes un beau lien en plein milieu d'un paragraphe riche de mots pertinents.
Chaque page fille va renvoyer un lien bien contextualisé vers la page mère au sein d'un paragraphe de très haut niveau ou mieux encore depuis le 1er paragraphe (voir la figure 3.12).
Enfin chaque page fille va faire un lien vers chacune de ses sœurs (figure 3.13). Ce lien se présentera sous forme de liste à puce et ne sera pas contextualisé pour des raisons de pratique et de maintenabilité.
La figure 3.14 montre comment ceci est implémenté en HTML.

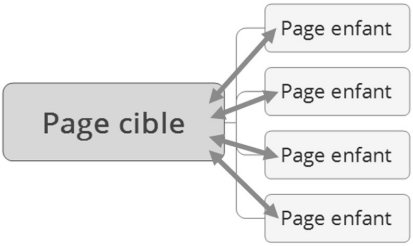


FIGURE 3.12 – Maillage cocon : lien mère-fille.

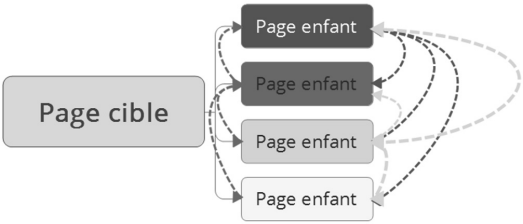


FIGURE 3.13 – Maillage cocon sœurs-soeurs.

3.9.7 Technique d'obfuscation de lien

On transforme ceci :

```
<a href="http://www.example.fr/contact/">CONTACTEZ
NOUS</a>
```

En cela :

```
<button onclick="location.href='http://www.example.fr
/contact/'">CONTACTEZ NOUS</button>
```

<h1>Titre optimisé ou avec co-occurrences</h1>

Avant de se lancer dans des travaux [lien avec ancre optimisée vers la page parent], il convient donc d'en envisager le bénéfice, qu'il soit financier ou en termes de préservation de l'environnement

Blablablabla.....

Blablablabla.....

<h2>Isolation du garage</h2>

... Afin de voir l'impact de ce pont thermique et les solutions pour y remédier rendez-vous sur la page isolation garage de notre site ...

<h2>Isolation intérieure</h2>

Il existe une technique d'isolation intérieure qui dans certains cas peut s'avérer être la seule qui puisse être faite. ...

<h2>Sujets similaires</h2>

Page sœur 1 (même niveau, même silo)

Page sœur 2 (même niveau, même silo)

FIGURE 3.14 – L'implémentation du maillage.

3.10 Conclusion

Un bon maillage interne permet de concentrer le pagerank interne sur les pages stratégiques du site.
Il permet d'ordonner les ressources en fonction de leur niveau d'importance et des enjeux concurrentiels.
Le cloisonnement sémantique devient optimal et permet ainsi au surfeur aléatoire de naviguer dans un univers fermé qui ne parle que de la même chose et qui aborde des sujets complémentaires avec des glissements sémantiques propres et pertinents.
Enfin un bon maillage interne permet une circulation optimale du pagerank interne, mais aussi du pagerank issu du travail de netlinking. Vous pourrez ainsi économiser beaucoup de budget de netlinking pour des performances de ranking équivalente !

4. Netlinking : ce qu'il faut savoir

Pierre Calvet est dans l'univers du SEO depuis la fin de ses études en 2013. Avec un parcours chez l'annonceur au national et à l'international ainsi qu'un passage en agence pour des grands comptes, Pierre est aujourd'hui consultant SEO et customer support pour `babbar.tech` et `yourtext.guru`. Il présente des cas d'utilisation des outils en vidéo, en fait la démonstration et répond sur le support.

4.1 Introduction

S'il y une prestation qui fait débat au sein des communautés SEOs quelles qu'elles soient, c'est bien le netlinking : certains ne veulent pas entendre parler de cette pratique, d'autres ne jurent que par elle et entre les deux se trouvent de nombreuses personnes qui ne savent pas vraiment sur quel pied danser.

4.1.1 Obtenir des liens, mais pourquoi ?

Obtenir des liens permet de travailler l'autorité, l'un des facteurs de ranking les plus importants [1], mais si vous lisez ce livre, vous vous en doutez déjà. Il ne vous manque que l'explication de la méthode de sélection des liens, et c'est ce que nous allons voir dans ce chapitre.

1. Les liens pointant vers votre site sont parmi les facteurs les plus importants pour votre ranking : `https://www.youtube.com/watch?v=l8VnZCcl9J4&t=1820s`

4.1.2 Les techniques d'obtention de liens

Avant d'ouvrir sur le contenu du chapitre, il faut préciser une chose : sélectionner les liens dont on va faire l'acquisition ne peut se faire que sur une partie des techniques d'obtention de liens, celle qui correspond à de l'outbound. Dans vos stratégies de netlinking, il faut également considérer les techniques qui se rapprochent de l'inbound.

Voici une liste non exhaustive des différentes options qui s'offrent à vous pour obtenir des liens :

- ⊙ Le linkbaiting (qui sert à pousser une personne ou une communauté de personnes à partager le contenu qu'on produit), et dont l'inconvénient majeur est la difficulté de maîtriser les sources de liens.
- ⊙ Les Relations Presse, qui ont le même inconvénient, compensé par la forte visibilité que procure un article de presse.
- ⊙ L'acquisition de liens (par du partenariat, généralement rémunéré en argent ou en service).
- ⊙ L'échange de liens ou les articles invités (le guestblogging, qui peut également comporter un complément financier).
- ⊙ D'autres techniques existent (créer un écosystème de blogs, par exemple, ou récupérer des sites expirés) mais elles nécessitent un investissement en temps et en ressources beaucoup plus élevé.

Nous nous concentrons dans ce chapitre que sur l'acquisition et l'échange de liens, car ce sont les techniques d'outbound les moins chronophages, les plus maîtrisées et donc avec un bon équilibre entre le travail à faire et les fruits à récolter.

4.1.3 Ouverture

Gardez donc à l'esprit que tout ce que nous abordons dans ce chapitre ne sera pas du goût de tous. Par exemple, les Britanniques et les Américains ne veulent généralement pas entendre parler d'achat ou de vente de liens (pour la plupart), ils ne jurent que par l'inbound et que le modèle français est particulièrement mature sur l'outbound. Même s'ils n'approuvent pas cette méthode, voyons quand même les bénéfices de celle-ci.

4.2 Comment choisir ses liens

4.2.1 Explication de la nécessité de cette étape

Savoir choisir ses liens, c'est un exercice d'équilibriste entre les métriques, le ROI, les mises à jour Google, et surtout c'est un travail que vous ne pouvez pas prendre à la légère : si vous ne faites pas attention, votre source de revenus risque fort d'en pâtir. Voilà pourquoi une méthode rigoureuse de sélection est indispensable.

4.2.2 Le rétroplanning pour anticiper les périodes cruciales

Avant de rentrer dans les détails, n'oubliez pas la règle suivante : anticipez ! Tout le travail que vous fournissez, dont l'achat de liens, aura un impact fixe après la fameuse durée des 3 mois (voir la section sur le Transition Rank). Cette période, adjointe d'un temps de préparation indispensable pour la conception du contenu et la publication, implique forcément un délai rallongé pour bénéficier du boost promis par le netlinking.
Par exemple, pour des sites de voyage, les objectifs de vente peuvent être différents d'un mois à l'autre, et la période de vente a en général lieu dans les 3 à 6 mois avant la période de consommation. On se retrouve parfois à faire publier des articles sur le ski à la fin du printemps, voire en été.
Pour un site e-commerce, la saisonnalité doit être prise en compte lorsque vous travaillez vos backlinks comme lorsque vous décidez de produire du contenu. Il vous faut donc anticiper le Transition Rank, la production du contenu, la validation éventuelle, et d'autres aléas potentiels.

4.2.3 Les approches stratégiques à considérer

Au niveau stratégique, tout dépend du contexte du site pour lequel vous tentez l'acquisition de liens :

- ⊙ Vous pouvez réaliser une approche quantitative pour pénétrer les pages de résultats de recherches (SERP) du marché ou essayer d'atteindre les mêmes niveaux que vos concurrents, surtout si vous êtes une jeune entreprise ou un jeune site.
  L'avantage de l'approche quantitative sera forcément la croissance rapide des sites référents. Les inconvénients, il n'est pas nécessaire d'être devin pour les identifier : si vous achetez beaucoup de liens publiés au même moment de sites peu qualitatifs, ça risque de se voir.
- ⊙ Vous pouvez réaliser une approche qualitative en dédiant une attention toute particulière aux sujets que vous souhaitez pousser.

L'avantage de l'approche qualitative, c'est que vous aurez peu de chances d'être détectés, ou en tout cas, les liens obtenus seront d'assez bonne qualité pour que ça ne vous pénalise pas. L'inconvénient est évident : vous allez passer énormément de temps et d'argent sur cette approche.

- ⊙ Vous pouvez dans tous les cas préférer une approche sécurisante pour éviter d'attirer l'attention en augmentant progressivement vos références chaque mois. L'avantage étant que vous rendez l'augmentation du nombre de références plutôt naturelle, l'inconvénient étant que ça risque de prendre plus de temps pour atteindre vos objectifs.
- ⊙ Vous pouvez aussi préférer une approche hybride, car chacune des approches précitées est compatible avec les autres. Il est d'ailleurs conseillé de diversifier les approches : à vous de piloter cette partie car en fonction de votre profil de liens existant, de votre objectif de profils de liens et de vos moyens, vous n'aurez pas les mêmes capacités.

Bien évidemment, si vous souhaitez réaliser de l'achat de liens de qualité, de façon sécurisante et en grande quantité, vous allez vous heurter à la limite de vos moyens : si le budget ne suit pas, vous n'avez pas les moyens de vos ambitions. Il faut dans ces cas-là revoir votre stratégie et adapter votre méthode mais cela se fera au détriment de vos délais de réalisation (et on le voit par la suite, les concurrents, eux, ne s'arrêtent pas).

4.2.4 A propos d'objectifs, comment les identifier ?

Il est indispensable que vous ayez une idée précise de ce que vous souhaitez atteindre : dans le cas général, vous souhaiterez faire au moins aussi bien que les concurrents, voire mieux sur les requêtes stratégiques que vous visez.

Pour ce faire, il y a une méthode très simple : regardez ce que font vos concurrents.

Lorsque vous regardez les pages de résultats de recherches (SERP) des mots clés que vous visez, vous allez pouvoir identifier, à travers l'analyse des pages présentes, plusieurs informations : le nombre de liens que ces pages ont obtenus, comment ces pages sont poussées sur leurs sites respectifs, le type de contenu attendu (ce qui peut vous permettre d'éviter d'être à côté de la plaque en proposant du texte quand le contenu attendu est une liste de produits) et si les pages ont été conçues pour répondre

spécifiquement à la requête ou non, etc.

Pour aller plus loin, vous pouvez vous faciliter la tâche sur l'identification des backlinks des pages et des métriques liées aux pages en utilisant des outils. Ça vous permet en général de situer votre page par rapport à la SERP et si une métrique semble corrélée au classement, il paraît évident qu'il faut travailler dessus pour améliorer son placement.

Alors que vous tentez de doper la production de liens vers votre site, il faut garder à l'esprit que vos concurrents font sans doute de même : leur profil de liens n'est pas statique, il est naturellement mouvant, comme le vôtre. Il vous faut conserver un historique pour identifier les tendances de gains de liens de vos concurrents et identifier le potentiel que vous avez pour dépasser les concurrents sur un objectif de délai que vous définissez, ou les acquisitions à réaliser pour dépasser vos concurrents. Encore une fois, les outils peuvent vous aider à identifier ces tendances.

Lorsque vous avez identifié ce qu'il vous faut travailler, vous avez plusieurs possibilités : soit vous essayez de vous calquer sur ce que le concurrent a réalisé et contactez les mêmes sites, soit vous essayez de faire mieux avec d'autres partenaires qui vous correspondent mieux. Dans tous les cas il faut s'assurer que votre approche ne soit pas trop visible (et donc adieu les règles de partenariat - balise *sponsored* par exemple - telles que Google les attend puisqu'elles sont équivalent à du nofollow).

4.2.5 Les facteurs de sécurité à ne pas oublier

Lorsque vous souhaitez pousser une page, ou plusieurs pages de votre site internet, vous devez prendre en compte le contenu de la page avant d'acheter du lien. Si votre contenu est bien conçu et correspond à ce qu'attend la SERP, les achats de liens doivent vous servir à légitimer la page telle qu'elle est. Vous ne devez pas faire de liens pour positionner une page qui n'est pas adaptée à la requête que vous visez.

Lorsque vous cherchez à obtenir un lien, il est impératif que le site que vous visez soit légitime pour parler du sujet qui va relier son article à votre page. Au-delà de ça, il est même indispensable que les deux pages associées par un lien parlent d'un sujet très proche, de préférence dans la même langue.

Si le site visé n'est pas expressément sur votre thématique, vous pouvez sans doute en obtenir des liens quand même, mais il vous faudra trouver un terrain d'entente entre la thématique du site source et votre thématique, et s'attendre à ce que le résultat ne soit pas aussi fort que vous pourriez

l'avoir avec un site plus proche.
Lorsque vous souhaitez obtenir un nouveau lien la question de l'ancre se pose également. Une ancre sur la requête visée peut être très intéressante, tout comme le contexte (donc le texte autour) du lien [2]. Mais l'ancre optimisée peut également poser un problème si vous obtenez trop de liens avec la même ancre : la limite officielle annoncée par Google est un peu floue à mais elle se situerait entre 4 et 10% des ancres obtenues. Essayez donc d'éviter de dépasser 4% pour une ancre optimisée, sauf s'il s'agit de votre marque.
Un autre élément à prendre en compte, c'est la quantité de sites référents : plus vous recevez de liens depuis des sources diverses plus c'est intéressant pour votre site… Obtenir plus d'un lien depuis un même site diminue l'intérêt de ces liens, et vous êtes également plus dépendant d'un même acteur.
Avec tout ça, comment appliquer une surcouche de sécurité à votre achat de liens ? Vous pouvez par exemple obtenir des liens optimisés vers vos pages stratégiques et noyer un peu le poisson en augmentant le nombre de liens non optimisés vers une page qui, normalement, accueille votre trafic marque : la page d'accueil ou une section de marque de votre site, ce faisant, vous augmentez le nombre de sites référents.
Enfin, ce qui est important de vérifier, c'est que les sites que vous visez soient des sites qui ne risquent pas de vous nuire : vérifiez l'empreinte des sites, la qualité de ces derniers, leur modèle économique potentiel, la propriété du domaine, l'hébergement, les informations légales, etc. Ce site a-t-il l'air d'être un site fiable à qui vous pouvez faire confiance ? En tant qu'utilisateur, qu'en pensez-vous ?

4.2.6 Les facteurs limitants auxquels vous allez vous heurter

Dans votre campagne de liens, tout ne fonctionnera pas comme prévu. En prestation, ou lorsque vous devez reporter à quelqu'un, un lien qui vous semblera parfait peut toujours être discuté par les décideurs (ceux qui ont la carte bancaire), parce qu'ils pensent que le site ne correspond pas à leurs valeurs, ou qu'ils ne comprennent pas l'intérêt d'obtenir un lien sur ces sites-là. La plupart du temps, les décideurs vont avoir des scrupules sur l'acquisition des liens car l'impact est difficile à percevoir pour quelqu'un dont le SEO n'est pas le métier (pour résoudre ça, je vous invite à aller

2. Les liens contextualisés sont très probablement un facteur de ranking : `https://www.searchenginejournal.com/ranking-factors/contextual-links/`

voir la partie sur le suivi de performances).
Un refus peut également venir du partenaire potentiel : si le partenaire considéré pense qu'il n'a aucun intérêt à produire un lien vers votre site, il y a des chances que vous ne puissiez pas obtenir de liens de sa part, à moins d'argumenter avec pédagogie et de longues discussions.
Un autre facteur limitant pourrait être le budget du lien car il n'y a pas de « grille tarifaire» canonique et chacun voit la valeur du lien comme il l'entend.
Plusieurs éléments sont à prendre en compte pour définir ce budget au mieux :

- La popularité du site et ses métriques
- Le maillage de ses articles
- Le trafic qu'obtient l'article
- Ce que vous attendez du lien (de l'autorité ? de la visibilité ? une validation par des médias ?)
- Les relais réseaux sociaux
- Le nombre d'acteurs du milieu
- Le nombre de vendeurs de liens du milieu
- L'ancienneté du site

4.2.7 Gagner du temps grâce aux outils

Parmi les tâches présentées, certaines sont quasiment impossibles à réaliser sans outil. Ils peuvent vous aider à mieux identifier les gains qu'une stratégie de netlinking permet d'obtenir.
Qu'il s'agisse du calcul et l'identification des métriques des pages d'un site (que vous pouvez retrouver sur Babbar) ou l'historique de ces métriques (également sur Babbar), l'avantage des outils est un gain de temps et un premier filtre dans la quantité faramineuse de pages qui composent le web.
Tous les outils ont un index différent et des méthodes de calculs et d'affichages différents. C'est pourquoi les métriques vont varier de l'une à l'autre, mais quelque soit l'outil que vous utilisez, vous pouvez bénéficier de perspectives intéressantes pour accélérer votre travail.
A vous de choisir l'outil qui correspond le mieux à vos besoins en fonction de la méthodologie utilisée pour le calcul des métriques ou des fonctionnalités présentées.
Comment passer les étapes citées dans les sous parties précédentes avec un outil comme Babbar ?
Identifier les objectifs, la première étape, est assez facile : vous vous

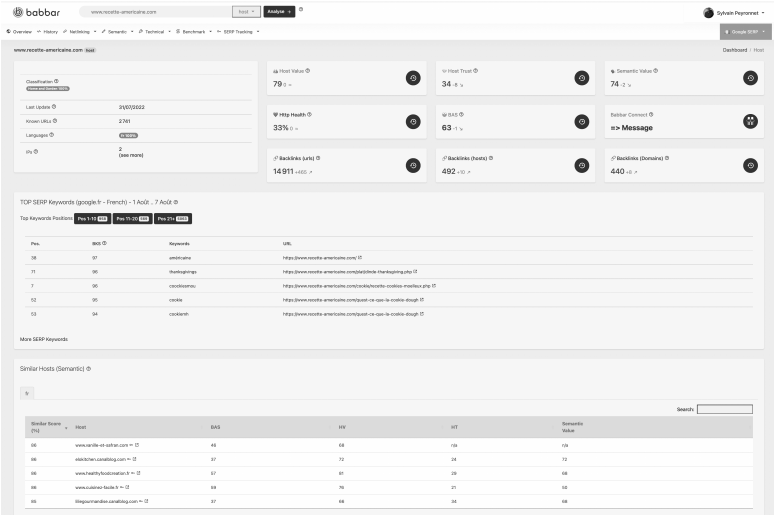


FIGURE 4.1 – Les métriques de Babbar

comparez aux concurrents et vous pouvez ainsi trouver les métriques dont ils disposent sur les urls présentes sur les requêtes que vous ciblez, et en passant par l'historique, vous identifierez également la tendance d'acquisition de backlinks de ces urls.
Vous pouvez ensuite passer à l'étape d'acquisition : trouvez ce que font les concurrents peut vous servir (avec les backlinks et la Force Induite), trouvez des sites prêts à faire du lien avec des pages proches sémantiquement de vos pages à optimiser (avec le Spotfinder et le filtre Babbar Connect).
Grâce aux métriques, vous pouvez faire le tri en faisant un ratio entre le prix demandé par le site et une métrique comme la Force Induite, mais aussi les équivalents du Trust Rank ou du PageRank [3], voire du PageRank thématique, selon l'approche qui vous intéresse. Et en comparant ces ratios, vous pouvez identifier les vendeurs de liens les plus intéressants pour vous, relativement à vos capacités financières.
Grâce aux outils, vous allez aussi pouvoir gérer les risques : que ce soit via les métriques et une limite que vous définissez en fonction du contexte que vous connaissez, ou via l'identification des IP des sites, ou selon les

3. Brevet du PageRank : Patent US6,285,999 B1 « Method for node ranking in a linked database » : `https://patents.google.com/patent/US6285999B1/en`

pourcentages des ancres fréquentes, ou enfin selon si l'outil obtient des positions ou non.

Un autre avantage des outils et non des moindres : les outils connaissent plus de pages que vous ne pourrez jamais en connaître. Ils ont donc une vision plus large et une compréhension plus vaste de ce qu'est le web. Cette compréhension est forcément pilotée par les interprétations proposées, et c'est donc sur cette interprétation que votre choix doit se faire : si vous êtes en phase avec l'une des interprétations, c'est avec l'outil correspondant que vous aurez les meilleurs résultats (même s'il faut le temps de prise en main de l'outil).

4.3 Suivre ses performances

Comme partout, lorsque vous faites quelque chose, vous ne le faites pas pour rien. Il faut donc suivre les résultats pour s'assurer qu'ils sont atteints. Ne pas le faire équivaudrait à jeter son argent par les fenêtres... Quelle que soit votre employeur ou votre client, le suivi des performances liées aux actions est un besoin pour eux : ils ne vivent pas vos actions au quotidien et ont besoin de comprendre rapidement comment ils se sont améliorés.

4.3.1 Les KPI indispensables

En SEO en particulier, la plupart des actions réalisées doivent être suivies : la complexité de la compréhension du métier pour les profanes impose de toujours comparer la version antérieure de la SERP à la nouvelle version obtenue à la suite des actions (et un petit délai de 3 mois). Cette nécessité a valu la création de la vue « historique des SERP » chez Babbar qu'on peut également retrouver sur d'autres outils, car elle est vraiment indispensable pour que le SEO puisse expliquer pourquoi on l'a payé et comment ça a marché. (Ou comment ça n'a pas marché, ce qui peut arriver).

De façon plus facile à suivre, le nombre d'impressions de vos pages, comparé aux périodes similaires précédentes, peut vous donner un indicateur que vous vous positionnez sur d'autres mots-clés.

Enfin, ce que la plupart recherche, le nombre de clics, qui, s'il augmente, traduit une amélioration de la captation des utilisateurs autour de la thématique abordée, que ce soit sur la requête ciblée ou sur des requêtes annexes : ce qui est important, c'est que vous obtenez plus de trafic.

4.3.2 Les KPI additionnels

D'autres métriques intéressantes, visibles dans les outils, peuvent être le nombre de mots clés sur lesquels l'URL optimisée est positionnée (si vous ne réussissez pas à vous améliorer pour la requête A, peut-être que les requêtes A', B, B', etc. vous apportent maintenant le trafic qualifié dont vous rêvez).
Si votre objectif est plus large qu'une seule URL, alors les métriques rentrent en lice : pour légitimer votre site, vous allez travailler le Trust Rank, pour faire croître la popularité de vos pages, le PageRank et le PageRank thématique (la semantic Value) vont vous servir, et alors vous pourrez vous comparer avec les URLs et les sites concurrents qui vous ont permis d'établir vos objectifs.
Vous pouvez également suivre les métriques simulées, comme la force induite, qui vous permet de vérifier que ce que vous espériez obtenir correspond bien à ce que vous avez acheté.
Enfin, *last but not least* comme disent les anglo-saxons, le nombre de liens et de sites faisant une référence à vos pages, ainsi que les proportions des ancres fréquentes, et tout ce qui sert à sécuriser votre stratégie de netlinking doit être suivi. Ce faisant vous assurez une surveillance nécessaire pour pousser les pages de votre site comme vous l'entendez, et limitez les excès.

4.3.3 La limite du transition rank

On entend souvent que le SEO prend du temps (on entend en général des périodes qui varient entre « 3 mois », « moins de 3 mois », « de 3 à 6 mois »). Tout cela est vrai (même si ce n'est pas expliqué du même point de vue à chaque fois).
Chaque action (une action n'est pas immédiate, elle prend toujours un minimum de temps entre la prise de décision et l'exécution/publication) est vue lorsqu'un GoogleBot (l'un des robots de Google) passe sur la page où a eu lieu l'action. Les nouvelles informations relatives à la page sont stockées, puis analysées (là encore ce n'est pas forcément immédiat). Passent ensuite les algorithmes de classification, qui identifient un changement qui semble jouer en la faveur de la page. Si ces changements sont comparables à du spam, la page va se voir attribuer une fonction de classement spéciale, parmi ce qu'on appelle le Transition Rank [4].

4. Le brevet « transition rank » : Patent US8,924,380 B1 « changing a rank of a document by applying a rank transition function » : `https://patents.google.com/patent/US`

Qu'est-ce qu'il se passe à ce moment-là ? Si vous optimisez votre texte ou que vous faites l'acquisition de liens, Google veut tester votre page : il la fait intégrer un chemin de classement entre l'ancienne position et la nouvelle position. Pour que votre page parcourt ce chemin, ça peut durer jusqu'à 3 mois. (Il n'est pas possible d'identifier ce chemin avec précision : il en existe plusieurs, sélectionnés en fonction du critère supposé attaqué par du spam et par une fonction aléatoire).
Une fois dans ce chemin de transition, si vous vous lancez dans une suroptimisation du facteur touché, vous allez repartir à zéro et recommencer un chemin, ce qui n'est jamais positif pour vous. Il faut donc parfois attendre 3 mois avant de voir les résultats et potentiellement les réajuster.
Avant la mise en production, vous aurez certainement eu à faire à une planification des actions, à de la rédaction, à de la sélection des acteurs, à de la validation : ajoutez-y les 3 mois du Transition Rank et on atteint facilement les 6 mois.

4.3.4 Préférez une stratégie long terme

Lorsque vous vous lancez dans l'acquisition de liens, que ce soit au bénéfice de votre entreprise ou pour le compte d'un client, votre action ne peut pas être ponctuelle. Si vous obtenez des liens, vous ne les obtenez pas en une fois : tout d'abord parce que ça correspondrait à une superbe synchronisation de vos actions avec les partenaires, mais aussi parce que ce n'est pas un signal normal pour un site web (dans la plupart des cas).
De plus, ne faire qu'une action sans poursuivre dans cette voie ne jouera pas en votre faveur : les pages de résultats évoluent tous les jours, le learning to rank change les signaux importants autant de fois, et vos concurrents ne s'arrêtent pas : lorsque vous vous arrêtez, vous risquez d'être dépassé.
Le SEO est une course sans arrêt où la lumière est mise sur les premiers, mais s'ils se reposent sur leurs lauriers, ils vont être rattrapés.

4.4 Les éléments à suivre en sus

4.4.1 Au-delà des performances de vos pages

Le suivi n'est pas fait uniquement pour la performance de vos pages. Il existe également pour s'assurer que tout ce que vous avez mis en place reste en place. En développement, il existe les tests de non-régression : la

méthode de ces tests est particulièrement utile pour suivre la vie de vos achats et vous assurer que le contrat est respecté par votre partenaire.

4.4.2 Codes de réponses

Suivre le code de réponse des pages avec lesquelles vous avez fait affaire est l'action la plus simple : aller voir les pages sources régulièrement, que ce soit via un crawler, via un code, ou à la main (si vous êtes patients), vous permet de suivre si l'url source est toujours présente et répond bien toujours en 200.

Si ce n'est pas le cas, il faut vérifier si le site est toujours actif. Si oui, vérifiez qu'il appartient toujours à la même personne et la recontacter pour lui demander de remettre le lien en place. Après tout, si vous avez payé pour un lien dont la durée de vie est censée être illimitée par exemple, retirer le lien est une rupture (attention, peut-être accidentelle) du contrat.

Ne sautez pas à la gorge de votre partenaire : il y a bien des raisons indépendantes de sa volonté qui peuvent expliquer ce retrait de lien (site HS, hébergement sans sauvegarde, etc.).

Si le site n'est plus actif ou s'il n'appartient plus au partenaire, c'est également plus délicat de faire valoir son droit : si le site n'existe plus, l'article est arrivé au terme de sa vie en même temps que le site et c'est la fin du contrat. Si le site a changé de propriétaire, le nouveau peut ne pas être au courant que les liens vendus font l'objet d'un contrat. (Surtout s'il a acheté un site expiré).

4.4.3 Présence du lien

L'article ou la page répond toujours en code 200 : ça ne suffit pas. Vous devez chercher si le lien existe dans la page et s'il est toujours en dofollow [5]. De la même façon que pour le suivi de la présence de l'article, ce relevé peut se faire via un scrapper, via un code, ou à la main.

Autre chose à surveiller également au niveau de la présence du lien : sa position dans la page. Si le lien est relégué sous les commentaires de l'article au bout de 6 mois, ça n'aurait plus le même impact qu'un lien en plein texte.

5. Les Nofollow ne sont pas comptés : `https://www.searchenginejournal.com/ranking-factors/nofollow-links/`

4.4.4 Présence sur le site et métriques de l'URL Source

La présence sur le site est plus compliquée à monitorer. Grâce aux outils, cependant, vous pouvez suivre l'évolution des métriques internes (chez Babbar, l'Internal Page Value correspond au PageRank Interne) et les mettre en relation avec la valeur de la page (la Page Value chez Babbar).

Bien évidemment, selon les sites (surtout les médias avec des millions de pages), vous allez rapidement vous rendre compte que la page est amenée à s'éloigner dans les limbes du site partenaire. Pour remédier à cette disparition, vous pouvez pousser la page source depuis d'autres sites, via d'autres acquisitions de liens. Cette méthode vous permettra d'améliorer l'autorité obtenue par votre page source sur un gros site (comme un média).

Via l'API Babbar, de simples appels réguliers suffisent à vérifier les métriques des pages sources : le code de réponse d'une part, l'Internal Page Value et la Page Value pour comparer le maillage de la page en interne et à l'externe, le Page Trust pour la proximité de la page avec un site de confiance, et la Force Induite pour suivre l'adéquation du couple source-cible suffisent pour le suivi des métriques.

Vous pouvez également suivre un autre indicateur de qualité de votre page référente : les positions obtenues en font des pages légitimes pour apparaître dans les résultats de recherche : ça ne peut qu'être un signal de qualité pour les pages qui bénéficient de liens de ces pages positionnées.

Bien entendu, si vous commencez à obtenir des liens vers les pages qui vous font des liens, il ne faut pas oublier de suivre ces acquisitions pour les mêmes raisons. En suivant les métriques et les liens obtenus par l'URL Source, vous pourrez également détecter et expliquer une amélioration ou une diminution de vos performances.

4.5 Conclusion

4.5.1 Le netlinking, ce n'est pas juste acheter des liens

Vous l'aurez sans doute remarqué, je ne parle que peu d'achat de liens, j'utilise surtout le terme « acquisition de liens » : on peut acquérir ou obtenir des liens autrement qu'en les achetant (même si c'est plus rapide). L'échange de liens est une pratique qui porte ses fruits également, mais elle est chronophage.

4.5.2 Les outils peuvent vous faire gagner un temps précieux

Vous pouvez sans doute presque tout faire sans outil, mais vous y passerez en général plus de temps que vous n'en avez à disposition, surtout si vous facturez au temps passé : vos clients s'attendent à une certaine efficience de votre part. C'est la raison première de l'existence d'outils : il s'agit d'un service que vous ne pouvez pas faire vous-même parce que vous manquez généralement de temps pour effectuer les tâches réalisées par ces outils qui vous facilitent l'aide à la décision. Et en SEO comme ailleurs, le temps est précieux (même avec un délai de 3 mois).

4.5.3 Ne jetez pas votre argent et suivez vos achats

Cette étape est essentielle et pourtant souvent ignorée : suivez un maximum les acquisitions de liens réalisées : de cette façon vous aurez des explications précises et un historique sur les améliorations et les baisses de performances des pages concernées.

4.5.4 Soyez toujours prêts à subir les mises à jour

Bien évidemment, en SEO on dépend énormément des mises à jour de Google. Il suffit d'une petite mise à jour sur la catégorie dont vous dépendez et vous pouvez perdre ce que vous avez déployé comme effort.
Il ne sert à rien de se voiler la face : c'est le jeu auquel les SEO ont tous accepté de jouer en se frottant aux algorithmes. Et comprendre les impacts de la mise à jour est une bonne façon de préparer l'adaptation de la stratégie pour regagner les positions perdues.

4.5.5 Ailleurs, le netlinking se fait autrement

Ce chapitre touche à sa fin, mais souvenez-vous : vous n'êtes pas seuls à faire du SEO et chacun a sa façon de faire. Si vous n'êtes pas d'accord avec une méthode ou les sensibilités des personnes qui sont sur un autre marché, il n'est pas nécessaire de décrier leurs méthodes : si elles ne fonctionnent pas, ça se verra rapidement chez eux. En revanche, s'ils continuent à exister, c'est qu'il y a des résultats et donc un marché pour ces pratiques.

5. Détection des fermes de liens

Thomas Largillier est docteur en informatique de l'Université Paris-Saclay. Il travaille sur les aspects théoriques du web depuis plus de dix ans et est auteur de plusieurs publications scientifiques portant sur la lutte contre le webspam. Actuellement en disponibilité d'un poste de maître de conférences à l'Université de Caen-Normandie, il est l'un des co-fondateurs de `Babbar.tech`.

5.1 Introduction

Avec l'avènement des moteurs de recherche et leurs algorithmes, tout le monde a cherché à optimiser son classement pour obtenir plus de visibilité. Au départ l'optimisation a surtout porté sur le texte des pages web. Avec l'arrivée de Google et de son algorithme du PageRank, les liens ont pris une importance capitale.

Afin d'optimiser son classement il ne suffit plus d'optimiser ses contenus mais il faut également avoir une vraie stratégie de « linking »

Aujourd'hui le faible coût de production d'une page web permet aux webmasters peu scrupuleux de monter autant de pages que nécessaire pour augmenter artificiellement leur popularité. Il s'agit d'un enjeu majeur pour les moteurs de recherche dont l'usage dépend de la confiance que leur accorde les utilisateurs. Il est donc primordial pour ces derniers de savoir repérer ces « fermes de liens » ou d'en annuler les effets.

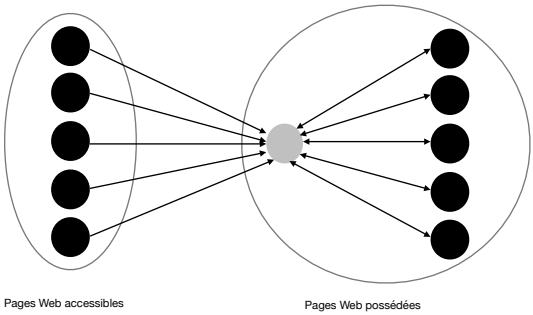


FIGURE 5.1 – Structure optimale pour booster la visibilité d'une seule page (en gris clair)

Les structures optimales sont connues depuis longtemps et ont été rendues inefficaces car trop facilement détectables. Il s'agit maintenant pour les tricheurs de « dégrader » suffisamment leur structure par rapport à la structure optimale tout en obtenant l'efficacité la plus grande.
Et aujourd'hui on va apprendre comment reconnaître un bon lien d'un mauvais lien.

5.2 Les structures optimales

Il existe plusieurs types de structures optimales en fonction de l'objectif visé, que l'on souhaite maximiser le pagerank d'une page ou d'un ensemble de pages. Comme défini dans l'article Web Spam Taxonomy de Gyöngyi *et al* [1] il existe 3 types de pages pour les webmasters sur le net, les pages qu'ils possèdent qu'ils peuvent modeler à leur guise, les pages accessibles sur lesquelles ils peuvent agir de manière très légère (commentaires, `etc.`) et les pages inaccessibles sur lesquelles ils n'ont aucun contrôle. Bien évidemment seuls les deux premiers types de pages sont intéressants et permettent de manipuler la popularité des pages.
La manière la plus simple de « booster » la popularité d'une page est

1. Gyongyi, Z., & Garcia-Molina, H. (2005). Web spam taxonomy. In First international workshop on adversarial information retrieval on the web (AIRWeb 2005).

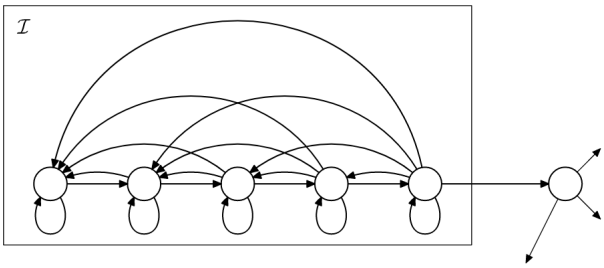


FIGURE 5.2 – Structure optimale pour booster la visibilité d'un ensemble de pages *I*

d'avoir le plus de liens entrants possibles. Le pagerank doit cependant circuler et donc les webmasters doivent renvoyer le pagerank des pages qu'ils reçoivent vers des pages qu'ils contrôlent pour le récupérer. L'objectif est de « capturer » le plus de pagerank possible et de ne jamais le renvoyer vers le reste du web.
Afin d'augmenter de manière optimale le pagerank d'une seule page il faut donc récupérer le plus de liens possibles depuis les pages accessibles et ensuite créer plusieurs pages faisant un unique lien réciproque vers la page cible comme illustré en Fig. 5.1. Bien évidemment ce genre de structures est assez peu naturel et très facilement identifiable par le moteur il est donc inutile d'espérer avoir un quelconque gain aujourd'hui en la reproduisant.
L'article [2] nous explique comment maximiser le pagerank d'un ensemble en modifiant le linking de ces dernières sans aller récupérer des liens dans les pages accessibles. L'article se concentre sur les pages possédées. De la même manière cette structure (représentée en Fig.5.2) bien qu'optimale est assez peu naturelle et non utilisable sur un site web.
Depuis les travaux d'Haveliwala [3] on sait que pour être utile un lien doit être catégorisé sinon il n'apportera pas assez de popularité. Cela ne représente qu'un léger frein pour les *spammers* qui doivent dorénavant décliner leurs fermes de liens sur plusieurs thématiques pour rester efficaces.

2. de Kerchove, C., Ninove, L., & Van Dooren, P. (2008). Maximizing PageRank via outlinks. Linear Algebra and its Applications, 429(5-6), 1254-1276.

3. Haveliwala, T. H. (2003). Topic-sensitive pagerank : A context-sensitive ranking algorithm for web search. IEEE transactions on knowledge and data engineering, 15(4), 784-796.

De nos jours l'idée derrière les fermes de liens est toujours de produire du pagerank artificiellement en créant des zones anormalement denses du web. C'est cela que les moteurs de recherche vont chercher à détecter ou à ne pas prendre en compte lors du calcul de popularité pour en annuler les effets.

5.3 Détecter les fermes de liens

La manière la plus intuitive de se débarrasser des fermes de liens consiste à les repérer pour ensuite ne plus les prendre en compte dans le calcul de la popularité des pages. Cette détection peut-être faite « offline » en parcourant l'index des pages et en repérant les maillages qui ne servent qu'à renforcer la position de certaines pages ou alors « online » lors du crawl pour éviter de prendre en compte les fermes de liens dans l'index et ne pas avoir à corriger son calcul de popularité a posteriori.
Les deux méthodes ne sont pas forcément exclusives. Il est possible d'utiliser une méthode « online » lors du crawl pour avoir un index le plus sain possible et d'utiliser des méthodes « offline » plus coûteuses sur son index pour évincer les fermes de liens qui seraient passées à travers les mailles du filet ou pour porter une une attention plus particulière sur les pages à forte popularité.
Plusieurs méthodes sur la connaissance de groupes de pages de « qualité » ou au contraire de pages de *spam*. Ces pages diffusent alors leur qualité à travers leurs liens. Les pages de qualité augmentent la qualité des pages vers lesquelles elles font des liens tandis que les pages de *spam* diminuent la qualité des pages vers lesquelles elles pointent.
Gyöngyi *et al*[4] ont proposé une méthode de détection du « webspam » basée sur l'estimation de la proportion du PageRank qui vient de pages de *spam* pour savoir si une page est elle-même du *spam*. Le problème de cette méthode est qu'elle nécessite une initialisation humaine pour étiqueter des pages comme étant honnêtes/de confiance ou l'inverse pour pouvoir lancer l'algorithme.
Nous avons développé une méthode de détection des pages bénéficiant de l'aide du *spam* pour améliorer leur popularité. Cette méthode est basée sur les vecteurs *ustat* présentés dans un article de Fischer *et al*[5].

4. Gyongyi, Z., Berkhin, P., Garcia-Molina, H., & Pedersen, J. (2005). Link spam detection based on mass estimation. Stanford.

5. Fischer, E., Magniez, F., & De Rougemont, M. (2010). Approximate satisfiability and equivalence. SIAM Journal on Computing, 39(6), 2251-2281.

Notre méthode est présentée en détail dans un article appelé *Detecting Webspam Beneficiaries Using Information Collected by the Random Surfer*[6].
Cette méthode consiste à identifier les structures permettant de « booster » la popularité des pages en regardant la « trace » que laisse le surfer aléatoire en se promenant dessus. Pour pouvoir comparer les traces entre elles il est important d'utiliser un langage commun et pour cela on va réétiqueter les pages web avec la distance qui les séparent du point de départ de la marche aléatoire. Bien sûr on ne va pas réétiqueter tout le graphe à chaque fois et l'on va se contenter d'un voisinage proche (distance de 2 ou 3 pages).
La comparaison entre les marches aléatoires se fait sur les différences entre les fréquences de changement de niveau de voisinage. Ainsi si les fréquences sont similaires il est probable que les structures sous-jacentes ont le même but même si elles ne sont pas identiques. Cela permet une robustesse de la méthode en ne recherchant pas un motif dans le graphe mais dans le comportement du surfer aléatoire. Peu importe les variations dans le schéma de liens mis en place si la ferme de liens manipule le surfer aléatoire de la même manière qu'un motif connu alors elle sera détectée.
Nous avons identifié plusieurs motifs utilisés par les *spammeurs* grâce à cette méthode. Les deux plus fréquents sont présentés en Fig. 5.3. Pour fonctionner cette méthode se repose effectivement sur une bibliothèque de motifs de *spam* qu'il faut construire. Pour valider notre méthode nous avons construit 14 motifs qui nous ont permis de valider la méthode.

5.4 Déclasser les fermes de liens

Une autre manière d'annuler l'effet des fermes de liens consiste à les déclasser naturellement. Cela consiste à avoir un calcul de popularité robuste qui ne prend naturellement pas en compte l'effet de « boost » escompté par les fermes de liens. Ces méthodes ne sont une fois de plus pas nécessairement à opposer aux méthodes de détection et fonctionnent très bien en complément de ces dernières.
Pour ce faire plusieurs algorithmes ont été proposés dans la littérature qui sont souvent des variations plus robustes du PageRank. Ainsi des

6. 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.

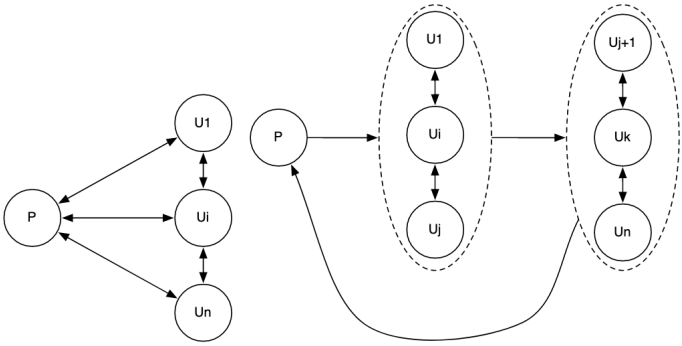


FIGURE 5.3 – Les deux motifs les plus fréquemment utilisés par les *spammeurs*

algorithmes comme le TrustRank [7], l'AntiTrustRank [8] propagent une information supplémentaire en plus de la popularité dans le graphe pour proposer un classement plus robuste. Le principal problème de ces méthodes étant qu'elles requièrent une initialisation humaine pour avoir un ensemble de « bonnes » pages pour pouvoir lancer l'algorithme.

Zhuang *et al* ont proposé [9] un modèle pour formaliser les algorithmes de déclassement et proposent deux nouveaux algorithmes. Les auteurs proposent un algorithme supervisé qui obtient des résultats bien meilleurs que les approches existantes mais souffrant du même problème d'initialisation et un algorithme non supervisé qui donne des résultats légèrement moins bons mais n'ayant aucun problème d'initialisation ce qui rend son utilisation plus confortable.

Nous avons développé une méthode pour calculer le PageRank de manière plus robuste aux fermes de liens. Cette méthode a été publiée dans un

7. Gyongyi, Z., Garcia-Molina, H., & Pedersen, J. (2004). Combating web spam with trustrank. In Proceedings of the 30th international conference on very large data bases (VLDB).

8. Krishnan, V., Raj, R. (2006, August). Web spam detection with anti-trust rank. In AIRWeb (Vol. 6, pp. 37-40).

9. Zhuang, X., Zhu, Y., Chang, C. C., Peng, Q., & Khurshid, F. (2017). A unified score propagation model for web spam demotion algorithm. Information Retrieval Journal, 20(6), 547-574.

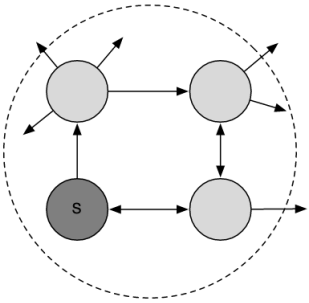


FIGURE 5.4 – Cette heuristique regroupe tous les noeuds faisant partie d'une boucle

article scientifique [10]. L'idée est de regrouper les noeuds du graphe qui font beaucoup de liens entre eux en *clusters* et de ne propager le PageRank qu'entre les *clusters* et non pas à l'intérieur. L'objectif est donc de proposer des méthodes de *clustering* qui vont regrouper les fermes de liens à l'intérieur d'un même *cluster* permettant d'annuler leur effet de « boost ». Les méthodes de *clustering* classiques sont très efficaces pour ce problème mais beaucoup trop coûteuses pour être appliquées à l'échelle du web. Il faut donc trouver des méthodes qui sont utilisables en pratique à cette échelle. Nous avons introduit plusieurs heuristiques de *clustering* pour réaliser cette tâche. Une heuristique est un algorithme qui donne un résultat satisfaisant, *i.e.* pas forcément optimal, en un temps raisonnable. En effet nous n'avons pas besoin de capturer l'entièreté d'une ferme de liens à l'intérieur d'un *cluster* pour en annuler les effets. Les fermes de liens étant des zones anormalement denses du web, le *clustering* permettra de regrouper une grande partie des noeuds au sein d'un *cluster* si les heuristiques sont conçues correctement.

Trois des heuristiques que nous avons proposées réussissent à annuler les effets des fermes de liens. La première (Fig. 5.4) consiste à regrouper ensemble tous les noeuds faisant partie d'une petite boucle (de taille 3 ou 4). L'intuition derrière cette méthode est que les *spammers* ont besoin que le surfeur aléatoire revienne fréquemment sur la page dont ils souhaitent augmenter la popularité. En pratique cela implique de calculer tous les

10. Largillier, T., & Peyronnet, S. (2012). Webspam demotion : Low complexity node aggregation methods. Neurocomputing, 76(1), 105-113.

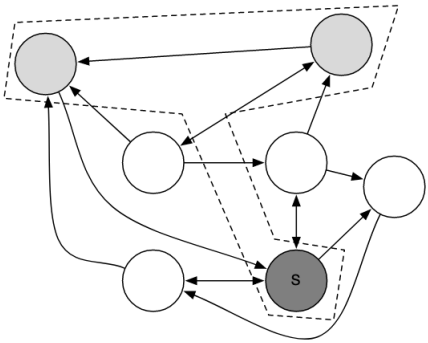


FIGURE 5.5 – Cette heuristique regroupe les arrivées fréquentes d'une marche aléatoire avec le point de départ

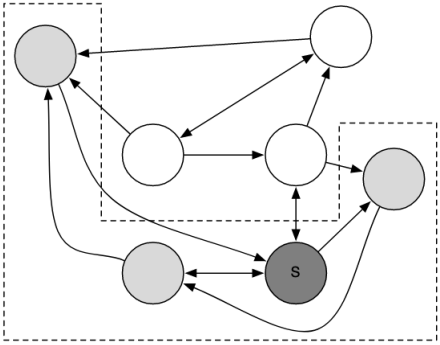


FIGURE 5.6 – Cette heuristique regroupe tous les noeuds faisant partie de marche aléatoire finissant sur le même noeud

chemins de taille k avec $k = 3$ ou $k = 4$ autour d'une page pour avoir toutes les boucles. Ceci est difficilement réalisable pour l'ensemble des noeuds du web et il faut donc choisir les noeuds sur lesquels appliquer un tel regroupement.

Les deux autres heuristiques sont très proches (Fig. 5.5 et Fig. 5.6) et sont basées sur des petites marches aléatoires. La première méthode consiste à regrouper le point de départ de la marche aléatoire avec les noeuds fréquemment atteints et la deuxième groupe également tous les noeuds sur le chemin qui permettent d'atteindre ces noeuds. En effet si peu importe comment on part d'un noeud on finit toujours au même endroit après quelques pas cela peut indiquer une volonté de manipuler le surfer aléatoire.

Ces méthodes peuvent être utilisées sur un sous-ensemble des pages du web que l'on soupçonne d'avoir recours à des méthodes malhonnêtes, par contre en les utilisant sur les pages qui ont un fort PageRank. Il est en effet inutile de passer du temps à identifier des tricheurs qui ne réussissent pas.

La meilleure méthode pour déclasser l'effet de « boost » des fermes de liens est la dernière puisqu'en plus elle augmente le PageRank des pages honnêtes. Cependant lorsque l'on soumet les résultats de chacune des heuristiques au test du χ^2 on voit que seule l'heuristique qui consiste à regrouper ensemble les pages faisant partie de petites boucles le passe. Il n'y a donc que pour cette méthode que nous avons une preuve statistique montrant que cette méthode traite différemment les pages de *spam* et les pages honnêtes.

5.5 Conclusion

Comme nous l'avons évoqué, le problème des fermes des liens est une priorité pour les moteurs de recherche, notamment pour garder la confiance des utilisateurs. Fort heureusement les moteurs de recherche ne sont pas démunis face à ces attaques et possèdent un arsenal de mesures et contre-mesures pour s'en occuper.

Les référenceurs mal intentionnés ne sont toutefois pas en reste et avec un coût quasi nul de production de contenu et de liens de nos jours il reste rentable pour eux d'avoir une stratégie agressive même avec un faible taux de réussite.

Les évolutions se font de moins en moins par à-coup de chaque côté mais plutôt en continu. Les forces en présence finiront sans doute par trouver

un équilibre, reste à savoir si celui-ci sera profitable aux utilisateurs qui se retrouvent pris entre deux feux.

6. Créer des liens : du pagerank au social, en passant par l'influence

Thomas Cubel est consultant et formateur dans le domaine du SEO. Spécialiste des enjeux on-site et fervent défenseur d'une vision holistique et engagée du référencement, il a été membre du Jury des Semy Awards Paris 2022. Thomas Cubel est également auteur de plusieurs centaines de ressources sur le sujet. Son site web : `https://www.thomascubel.com/`.

6.1 Introduction

Il existe plusieurs écoles pour créer des liens.
L'école des référenceurs, qui en grande majorité recherche le fameux pagerank, le « jus » qui fera positionner leurs pages dans les SERP [1].
L'école des spécialistes de la communication, qui a à cœur de faire passer le bon message, transmettre les valeurs d'une marque pour favoriser la bonne image par exemple.
L'école des spécialistes des réseaux sociaux, qui reste dans le dialogue, les interactions, le social marketing [2], qui aime la proximité entre la marque et le consommateur.

1. SERP : Search Engine Results Page. La page de résultats d'un moteur de recherche. `https://fr.wikipedia.org/wiki/Page_de_r%C3%A9sultats_d'un_moteur_de_recherche`
2. `https://fr.wikipedia.org/wiki/Marketing_social`

Une entreprise sérieuse, qui souhaite être leader de son marché, qui veut tenir dans le temps, se doit d'avoir un équilibre entre toutes ces « écoles » et tant d'autres à côté. Elles ont chacune leurs priorités, leurs visions, mais elles disposent toutes d'une partie de la vérité concernant la réussite sur le long terme d'une marque.

Dans ce chapitre, je vous propose de parler de la création de liens. Les backlinks dans le cadre du SEO, mais aussi, surtout, du lien social. Celui qui fera que vous tiendrez le marathon, qui fera votre réseau professionnel, qui vous emmènera à présenter dans une grande conférence « comment vous avez fait pour réussir et avec qui ».

6.2 Back to basic : le travail du référenceur

Pour performer avec un site web dans les pages de résultats des moteurs de recherche (SERP), nous devons respecter trois piliers du SEO : la technique, le contenu et la popularité (voir la figure 6.1).

La technique permet d'avoir un bon contenant et un bon outil de travail, compris et utilisable par les utilisateurs et les robots des moteurs de recherche.

Le contenu va ensuite s'imbriquer dans ce contenant technique pour apporter une offre qualitative, de la valeur ajoutée. C'est le moteur, intermédiaire entre l'offre et la demande, qui est en charge de mettre cette offre pertinente en face d'une demande exprimée par un utilisateur : la fameuse requête.

La popularité consiste enfin à populariser cet ensemble, via des actions externes, afin de montrer au monde la valeur proposée. Le but final étant d'obtenir une certaine légitimité et de ne pas être seul dans les méandres du web. Cela est possible en obtenant des backlinks par exemple (liens pointant vers une page / un site web cible), mais également en faisant vivre nos contenus, en créant de l'engagement ou des partenariats.

C'est sur ce dernier pilier que nous allons nous arrêter dans le cadre de ce chapitre.

Partons du principe que vous avez un site correctement travaillé au niveau technique, que les contenus ciblent des requêtes et intentions de recherche précises, que vous avez par exemple proposé des contenus optimisés avec `yourtext.guru`.

Comment favoriser une bonne popularité sur le long terme ? Comment s'en sortir en tant que marque ? Acheter des backlinks est-il suffisant ?

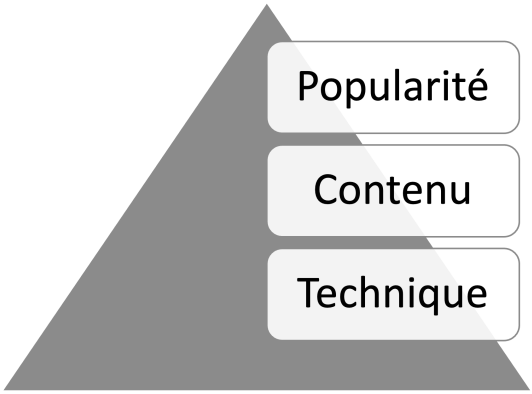


FIGURE 6.1 – Les trois piliers du SEO

Nous allons voir ensemble les limites de la recherche exclusive de pagerank, et pourquoi il vaut mieux aller vers un modèle beaucoup plus long terme, vertueux, alliant pagerank, social, influence marketing et plus encore.

6.3 Les référenceurs cherchent surtout du pagerank

Si vous étudiez les discussions entre référenceurs qui peuvent survenir sur des réseaux sociaux comme Twitter, Facebook, LinkedIn ou bien sur des groupes privés (Skype, Slack, Discord, etc.), vous remarquerez assez rapidement que ce sont les backlinks, le pagerank, les PBNs [3] qui sont au centre de toutes les discussions.
En effet, les backlinks faisant partie des tops priorités pour pouvoir améliorer sa visibilité sur les pages de résultats des moteurs de recherche, il est nécessaire d'en posséder une grande quantité pour percevoir des résultats intéressants.
Les référenceurs passent donc une large partie de leurs journées à demander les meilleurs spots, le meilleur service, de l'échange de liens ou le dernier bon plan. En tout cas en ce qui concerne les profils « netlinkeurs ». C'est d'autant plus vrai en France, puisque le marché français est un des marchés les plus mûrs sur ces sujets. Nous avons en effet des dizaines

3. PBN : *Private Blog Network*. Réseaux de blog privés.

(centaines ?) de marketplaces et services pour acheter des liens, des articles sponsorisés, des sites, des PBNs… Mais également du contenu, de l'influence, des articles de presse, des fans, des comptes de réseaux sociaux abandonnés, et plus encore.
C'est sans compter les innovations comme `Babbar.tech`, que je remercie d'ailleurs pour me permettre d'écrire ici, qui est la seule solution proposant le calcul d'un pagerank thématique et de métriques utiles et réalistes en 2022.
C'est important de le dire, car lorsque vous allez en Allemagne par exemple, vous payerez souvent très chers vos liens et vous pourrez même avoir des difficultés à trouver des marketplaces pour acheter du backlink. Cela n'existe tout simplement pas ou peu.
Pareil sur le marché anglophone, vous serez face à des services relativement lambda, proposant certes beaucoup de choix, mais du choix (ou plutôt une qualité) que nous voyons depuis plus de 10 ans, toujours avec les mêmes codes.
Les autres cultures sont en effet davantage tournées vers l'obtention de backlinks naturels, c'est-à-dire par le fruit d'un travail acharné et méritocratique. Aux Etats-Unis, on peut penser aux stratégies de content marketing [4] mêlant ebooks, études de cas, outils en ligne, blogging, infographies, vidéos et autres. Tout cela participe à créer une émulsion, du bouche à oreille, mais également du relais d'informations et donc des backlinks. Ça joue le jeu du web à fond. Un web utile, inclusif, qui partage.
Maintenant, faut-il pour autant blâmer l'achat de liens ou la culture intensive de pagerank ? Cela suffit-il à devenir leader dans sa thématique ? C'est ce que nous allons voir maintenant.

6.4 Le pagerank et la notion de leadership

Comme vous l'avez vu au fil de cet ouvrage, le Pagerank, bien qu'il puisse être appliqué à d'autres sujets d'études, est quand même connu pour être l'algorithme qui permet de connaître un certain score de popularité d'une page web.
Cela permet de faire ressortir les meilleures ressources présentes sur le web grâce à un système itératif et en complément d'une analyse de la

4. Content marketing : Marketing de Contenu. `https://fr.wikipedia.org/wiki/Marketing_de_contenu`

pertinence des contenus proposés pour une requête.

De ce fait, le pagerank est une notion surtout SEO qui s'arrête à la frontière de la SERP. Avoir beaucoup de backlinks vers son site n'augmentera pas de manière directe et significative votre nombre d'abonnés sur Twitter, YouTube, Instagram, TikTok, LinkedIn…

Il ne fera pas augmenter le chiffre d'affaires de votre entreprise non plus. Il ne travaillera que très peu votre image de marque et ce qui peut en découler.

Quand je dis cela, j'entends par là que beaucoup de backlinks achetés ici ou là, ou en faisant de l'échange, etc. sont très souvent faits rapidement, sans trop se soucier des contenus produits, de l'aspect marketing ou de la communication. On pose son lien, et la réflexion s'arrête là.

Cela n'est pas en soi un problème si on souhaite se contenter simplement de faire un travail de référencement lambda, très « mécanique », « pratico pratique » qui fera monter les pages dans la SERP. Mais cela peut poser problème si comme moi, vous souhaitez voir une marque grandir pour devenir à terme un *serious player*[5] plein d'ambition et indéboulonnable.

Le pagerank seul ne peut donc pas suffire à vous donner la place de leader, ou seulement dans de très rares exceptions et de manière assez temporaire.

6.5 Les risques à cultiver seulement le pagerank

Pour terminer cette partie sur le pagerank et après avoir expliqué son utilité, son importance et son utilisation dans le domaine du SEO, il me parait important de conclure sur les inconvénients de se focaliser seulement sur ce fameux « jus » pour bien comprendre le manque d'aspects sociaux.

Comme je viens de le dire, se focaliser seulement sur le pagerank pourra développer votre visibilité SEO si vous partez de zéro, pourra améliorer les classements de vos pages sur des requêtes plus ou moins concurrentielles. Cela pourra vous aider également à obtenir de belles métriques dans les outils, sur les marketplaces, et à certains endroits stratégiques pouvant déclencher des partenariats, de la vente de liens. Mais c'est à peu près tout.

Je pense qu'il est important de rappeler que tout cela est tout de même très centré sur la SERP et le relationnel entre SEO. En tout cas de manière directe.

5. Synonyme de leader, qui a de l'importance au sein d'un marché.

Si on va sur de l'impact indirect, on peut s'imaginer que le site bien conçu pourra multiplier son trafic, et donc son chiffre d'affaires (via des services, produits, vente de liens, affiliation, publicité, etc.). Cela peut permettre également d'obtenir toute sorte d'opportunités plus sérieuses comme des propositions de rachat, gros contrats, etc. Mais est-ce suffisant ?
La réponse à cette dernière question est sans doute « ça dépend ». Si on souhaite rester dans le spectre du SEO, à vouloir simplement faire ranker des pages et faire de l'argent pendant un temps, oui ça suffit.
Par contre, si l'entreprise veut que sa marque rayonne, perdure pendant des décennies, avec une certaine garantie d'être indéboulonnable, ou tout simplement pour sécuriser ses gains et ne pas mettre tous ses œufs dans le même panier, il faut aller plus loin.
Avez-vous connu Google Penguin version lancement orchestré ? Je pense que c'est un bon exemple.
Pendant 4 ans, ce filtre antispam a été déclenché manuellement pour pénaliser des sites qui ne répondaient pas aux exigences de Google. Le filtre attaque principalement les ancres et peut vous laisser dans les méandres de Google pendant des mois, voire des années. Vous deviez attendre le prochain lancement à chaque fois et faire le ménage entre temps (ou repartir de zéro).
Le problème avec ce filtre est qu'il a été considéré comme strict et injuste par de nombreux éditeurs et référenceurs. Il est aussi surtout considéré comme l'un des filtres les plus dévastateurs au niveau business avec Google Panda. Des choses dont on ne parle pas souvent d'ailleurs.
La réalité est là. Beaucoup d'éditeurs ont préféré arrêter ou switcher leurs activités pendant cette période. Beaucoup ont perdu énormément d'argent, d'employés, et d'amis. Beaucoup n'étaient pas référenceurs, mais des entreprises qui pensaient « bien faire », qui ne pensaient pas que « créer des simples liens » allait leur faire mettre la clé sous la porte.
Bien sûr, je parle de cas extrêmement difficiles pour faire passer un message (*spoiler alert* : ne mettez pas tous vos œufs dans le même panier et ne vous croyez pas invincible en SEO), mais il y a aussi des cas qui s'en sont mieux sortis. Il est intéressant de les étudier.
Par exemple, j'ai eu des cas qui avaient quand même été pénalisés (car il faut le dire, tout le monde faisait des liens pénalisables auparavant) et qui ont réussi quand même à faire du business grâce à leur branding, leurs réseaux sociaux, leurs systèmes d'affiliation, leur boutique pignon sur rue aussi.
L'inconvénient majeur donc de se focaliser seulement sur l'obtention de

pagerank en créant des backlinks lambda, c'est qu'on ne sait pas de quoi demain est fait. Tout peut très bien se passer, mais tout peut aussi vous mettre à terre. Un bon entrepreneur ne peut pas jouer avec ça, il doit anticiper et diviser les risques.
Si l'on comprend cela, on comprend tout de suite l'importance de développer du coup d'autres leviers comme la publicité, mais aussi l'emailing, les réseaux sociaux, les partenariats, l'affiliation, le content marketing, le mobile, le local et même ce qu'on appelle le marketing traditionnel.
Cela permet d'assurer la pérennité de l'activité et son leadership par la même occasion.

6.6 Faire du social : ingrédient de la réussite ?

Maintenant que nous avons parlé du pagerank, nous allons parler de ce que j'appelle les enjeux sociaux du web.
Lorsque vous recommandez à un ami un produit, un service ou une marque, vous faites du social. Quand vous envoyez un mail via une page contact, un MP sur Twitter ou un message dans un chat pour échanger avec une marque, vous faites du social. Quand vous relayez des informations sur les réseaux sociaux, vous faites du social.
Dès que vous avez ou que vous vous intéressez aux rapports entre individus, vous êtes dans le social.
Et une de mes convictions lorsqu'on s'occupe d'un site ou plus globalement d'une marque, c'est que sur le volet popularité du SEO, il faut, en plus du pagerank seul, du social.
En effet, je pense réellement que la bonne approche pour réussir et tenir sur le long terme, c'est de faire un hybride entre les enjeux SEO purs et ce que j'appelle le « social ». Cela veut dire faire de l'achat de liens, de la création de backlinks artificiels pour aller vite et répondre à des enjeux courts termes de positionnement. Mais aussi être dans la création d'actions sociales ou marketing favorisant les backlinks naturels.
Pour moi, cela est extrêmement important pour créer des bonnes relations, qui favoriseront une bonne image de marque, de l'échange, des ambassadeurs qui vont vendre notre marque à notre place, etc… Tout en allant vite sur les aspects SEO purs que sont le positionnement par exemple. Ce qui est intéressant aussi, c'est qu'ainsi on diversifie et enrichi beaucoup plus l'écosystème de ses liens. On a des liens hypertextes, on a des liens humains qui se créent, mais on a aussi moins de calculs dans toute cette

histoire. Et si je prends les liens hypertextes, c'est-à-dire les backlinks, on peut vraiment avoir un bon profil en général.
Là où l'achat d'un lien moyen, c'est un site avec de bonnes métriques, un peu de trafic, une apparence pas très glorieuse et des contenus vraiment basiques. Faire du social permettra d'avoir des liens follow, nofollow, sponsored, UGC, du petit site, du moyen site, du gros site, du contenu qui a du sens, de la cohérence, des liens qui ne sont pas sur catalogue, du relais sur les réseaux sociaux, dans des newsletters, voire même des livres pour certains.
Et comme je l'ai indiqué plus haut, cela participe à ne pas mettre tous ses œufs dans le même panier, et de contre balancer des possibles effets négatifs qui pourraient survenir comme lors d'une pénalité. Ces liens humains comptent, eux-aussi. Et beaucoup ont tendance à l'oublier.
Par exemple, j'ai récemment discuté avec un client qui m'expliquait que ce qui l'avait sauvé lors de sa chute en SEO il y a quelques années, c'était YouTube, l'emailing et les réseaux sociaux. Il passait du temps à faire des vidéos, à créer des séquences mails « humaines » et à répondre aux commentaires, faire interagir les gens sur ses réseaux. Il a juste eu à « activer » un peu plus la machine pour vendre lors de la chute. Les gens lui faisaient confiance.
Donc vraiment, si j'avais un bon conseil pour tous ceux qui ne pensent qu'à faire des backlinks « juste pour le pagerank » : ne restez pas dans votre coin, allez aux contacts des gens qui font votre marché. Ils vous le rendront. Donnez avant de demander et ça sera plus facile.
Et pour ceux qui pensent que les backlinks naturels, ça n'existe plus, sortez du web des référenceurs ! Car ça aussi, c'est un problème. Beaucoup de ressources remontent par des citations naturelles, comme celles que vous voyez dans les livres ! C'est le principe fondamental du web.

6.7 Faire du social avec le content marketing

Quand on commence avec son site web, sa marque ou bien lorsqu'on est resté trop longtemps dans son coin à faire des backlinks pour le pagerank seul, il faut déjà commencer par démarrer la machine sociale.
L'une des premières choses à instaurer, c'est ce qu'on appelle un « brise-glace ». C'est-à-dire ce qui va mettre à l'aise les gens, ce qui va augmenter la confiance envers vous, faire les premiers échanges, etc.
Vous ne pouvez pas vous pointer sur les réseaux, ou balancer des mails à tout va, et demander tout, tout de suite, sans instaurer le bon climat. Vous

risqueriez de vous « griller ».
Vous devez donc commencer par vous organiser et résoudre ce premier problème de « je dois m'intéresser aux gens, je dois les aider, leur donner des choses avant de recevoir ».
Et l'un des leviers qui peut le plus vous aider pour ça, c'est le content marketing.
Je vous recommande de suivre le fameux cycle du content marketing et de le coupler à mon processus en 4 étapes que j'utilise pour résoudre tous les problèmes : audit, stratégie, mise en œuvre et suivi (voir la figure 6.2).

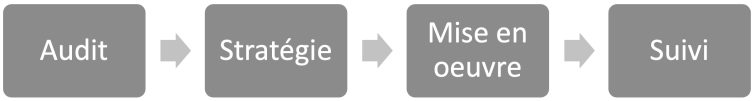


FIGURE 6.2 – Les quatres étapes

6.7.1 Audit

Premièrement, on commence par analyser la situation, on récupère des données sur les concurrents, les cibles, le marché et sur son propre projet, puis on réfléchit.

- ⊙ Qu'est ce qui intéresse les gens ? De quoi parlent-ils ?
- ⊙ Où vont-ils chercher de l'information et discuter ?
- ⊙ Qu'est ce qui génère des likes, des partages, des articles, des backlinks, etc. ?
- ⊙ Quels sont les réseaux sociaux les plus utilisés ?
- ⊙ Quels sont les formats et les supports les plus utilisés ?
- ⊙ Quelle est la fréquence de publication sur ces différents lieux ?
- ⊙ Où se trouve mon concurrent ? Que fais mon concurrent ?
- ⊙ Et moi ? Je suis où dans tout ça ?

Vraiment, essayez de voir grand, posez-vous des questions, ne vous limitez pas dans un premier temps. Dites-vous que vous avez des milliards d'euros, tout le personnel qu'il faut et regardez l'ensemble. Vous pouvez si vous le souhaitez utiliser la carte d'empathie ou les buyers personas à cette étape.

Vous devriez alors avoir une idée d'où aller, de comment parler, de quel sujet parler, à quelle fréquence et de quelle manière.
Ensuite, une fois que vous avez ces informations, vous pouvez vous poser la question du budget, des délais et de vos objectifs.
Vous souhaitez vendre plus ? Vous souhaitez travailler votre image ? Vous souhaitez attirer des investisseurs ? Faire des backlinks ? Sous 1 mois, 3 mois, 6 mois, 1 an ? Vous avez 5000€, 50 000€, 500 000€, 5 000 000€ ?
Décidez de ce qui correspond le plus à votre situation / profil tout en gardant à l'esprit qu'il va falloir quand même mobiliser des ressources pour intégrer le marché. Ensuite, vous pouvez passer à la stratégie.

6.7.2 Stratégie

Une fois que vous avez récolté toutes les données sur votre marché, vos concurrents et vous-même, que vous avez décidé de vos objectifs, budgets et délais, vous pouvez passer à la stratégie qui consiste à établir un rétro planning précis contenant les tâches à faire avec :

- un titre et une description des tâches ;
- le nom de la ou les personnes qui vont s'en occuper ;
- la date de publication ou lancement ;
- les ressources à engager (argent, sous-traitance, etc.) ;
- les ressources à créer ;
- où cela doit-il être posté ;
- les KPIs à atteindre et à mesurer.

Par exemple, si on souhaite des backlinks, car c'est tout de même l'objet de cet ouvrage, aller sur des infographies, des outils en lignes, des gros guides, des ressources téléchargeables vous permettra d'attirer du lien en nombre.
Si vous relayez ces ressources aux bons endroits comme sur des sites d'actualités, les réseaux sociaux plébiscités, que vous contactez les bonnes personnes, que vous faites même participer des influenceurs, ça fonctionne bien ! Et vous pouvez avoir 3-5 domaines référents minimum pour une infographie et plusieurs centaines pour un outil, jeu, etc.
Je trouve d'ailleurs qu'il faut davantage raisonner en valeur qu'en argent strict ici. Quelle est la valeur perçue pour la ou les personnes en face ? Comment je peux multiplier cette valeur avec un simple contenu, sujet ? Parfois, c'est une valeur inestimable si vous avez le bon timing, le bon

sujet, le bon format.

Par exemple, un des plus gros coups que j'ai vu dans ce domaine c'est la création d'une petite application en ligne pour savoir combien de temps une personne a perdu dans sa vie à jouer à un célèbre jeu de Riot Games : League of Legends. `Wol.gg`, c'est une page, un formulaire pour mettre son pseudo, un bouton et une API. Ce site et ses anciennes versions ont récupéré des milliers de domaines référents, des centaines de milliers de personnes connectées en simultané les premiers jours. Et ça tourne encore ! A ce niveau, on parle même plus de rentabilité, c'est de la chance. Tout cela, pour la simple et bonne raison que l'application attirait la curiosité, suscitait de l'engagement, du partage, du social... Et encore plus car l'interface et le partage initial avait été fait pour l'Asie, très friand de ce jeu vidéo populaire.

Pensez donc à ce qui pourrait créer de l'émulsion, un séisme, même temporaire, dans votre secteur. Vous pouvez utiliser `Babbar.tech` pour ça. Vous regardez pour quelques sites, les pages qui obtiennent le plus de backlinks. Vous tomberez souvent sur des guides fournis, des sujets qui se laisse partager, etc.

Et vous verrez que le content marketing, ce n'est pas juste du blogging comme je le vois souvent. Ça va bien plus loin, surtout hors de France.

6.7.3 Mise en oeuvre

Après la stratégie, c'est la mise en œuvre. On a le plan, il faut maintenant bâtir la stratégie. Ici, vous pouvez faire tout vous-même, mais vous vous rendrez assez vite compte qu'il faut plutôt sous-traiter et travailler avec les meilleurs des meilleurs. Ça ne coûte pas plus cher, car travailler avec du low-cost, c'est aussi souvent coûteux dans le temps de mon point de vue.

Je trouve même qu'il faut sortir de cette logique du moins cher. Quelqu'un qui a de bonnes idées, qui construit correctement les choses, qui prend le temps, ça se paie. Et franchement, on s'y retrouve par rapport aux équations simples du type 1 backlink = 200€. Le contenu, l'app ou autre, vous pouvez relancer plusieurs fois, recycler, le modifier, etc. Le backlink, pas vraiment.

En plus, vu les prix et l'engouement autour des plateformes, je suis sûr qu'on arrive à un moment où ça coûte même beaucoup moins cher de faire du content marketing et que c'est un meilleur investissement. C'est juste qu'il faut être formé et passer à l'action, ainsi que gérer une plus grosse charge mentale.

Entourez-vous donc de profils qui vous correspondent, trouvez ceux qui parlent le même langage que vous, informez-vous sur les tendances, faites de la veille, intéressez-vous aux autres, à ce qui est fait ou pas fait. Vous verrez, ensuite ça ira tout seul.
Pour l'organisation, vous pouvez vous tourner vers les célèbres Trello, Slack, Asana, Notion.so, mais aussi tous les outils pour surveiller les mentions et réseaux comme Brandmentions, Hootsuite, Publer et bien sûr Babbar.tech.
Optimisez vos processus. Comptez le temps moyen des tâches, essayez de faire le plus simple possible. La complexité peut parfois être utile, mais elle est dans 90% des cas pas toujours nécessaire.
De la même manière, attention aux idées reçues, aux biais de confirmation, etc. Beaucoup de personnes vont vous dire qu'une image plus jolie, elle convertira mieux etc. Pareil pour vos textes, vos boutons. Faites plutôt de l'A/B testing. Vous verrez que parfois (souvent), les images les plus moches, les vidéos pixélisées, avec un micro vraiment moyen sont les contenus qui cartonnent bizarrement le plus.

6.7.4 Suivi

Pour terminer, n'oubliez pas de suivre vos campagnes de content marketing, c'est-à-dire vérifier si le calendrier est respecté, que le travail est fait, mais aussi vos KPIs, le ROI de vos actions.
Vous avez la possibilité comme je le disais plus haut de recycler, de relancer la publication si ça ne décolle pas. Vous pouvez amplifier la résonance avec des outils comme la publicité, le relais par les influenceurs ou autre. C'est très important.
Par exemple, dans l'infographie de la figure 6.3 qui représente le cycle du content marketing, vous voyez bien qu'après la distribution des contenus, le sujet du suivi et de l'amplification sont nécessaires pour avoir de la résonance. Également, cela vous permet de retourner dans le cycle vertueux général avec notamment la phase de recherche statistiques.
En effet, les données concernant votre projet vont changer dans le temps. Il est donc nécessaire pour les campagnes futures de redéfinir les objectifs, les budgets, délais, mais également les types de contenus qui fonctionnent vraiment, ceux qui ne sont pas du tout porteurs, etc.
A terme, vous saurez ce qui fonctionne ou pas, ce qui coûte ou pas, ce qui est incertain ou pas. Vous pourrez même l'appliquer à d'autres sites ou secteurs. Cela vous permettra aussi de toujours remettre en question vos

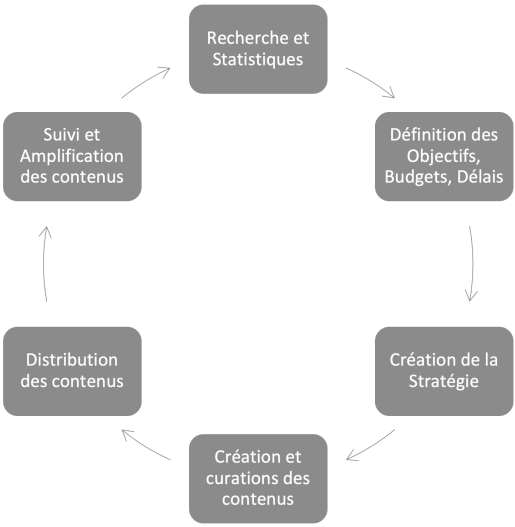


FIGURE 6.3 – Le cycle vertueux

stratégies par rapport à l'actualité, aux tendances, à vos enjeux en général, etc. Vous verrez ainsi ce qui rapporte le plus de backlinks dans le cadre du netlinking. En tout cas si c'est votre seul KPI.

6.8 Oui, mais ce n'est pas du netlinking, non ?

C'est une question que l'on me pose souvent en consulting. La réponse est « si pourtant, c'est juste qu'on a oublié les fondements du web et qu'on occulte une partie de celui-ci.»
En fait, je remarque (et je ne suis pas le seul), que la discipline du SEO est quand même assez souvent conceptualisée comme « une checklist à suivre, ni plus, ni moins ». Ça manque de cohérence, de sens dans les choses, d'une vue globale en synergie avec les autres métiers.
Personnellement, chacun voit comme il l'entend son métier de référenceur, mais je suis adepte d'une vision holistique des choses, une vision où chaque élément est en synergie avec un autre. Où le pilier technique fonctionne forcément avec le pilier contenu, et le pilier contenu avec le

pilier popularité et ainsi de suite.
Tout doit aller dans le même sens, sinon cela apporte des incohérences, de la perturbation, et au moindre changement majeur, l'instabilité est présente. Je souhaite avoir le contrôle sur un maximum de choses, tout en proposant un web utile. C'est pour cela que je travaille étroitement avec les équipes opérationnelles et les autres consultants. Je ne reste pas dans mon coin à faire du SEO, je m'ancre dans chaque projet à fond.
On est d'accord qu'il y a une part de contenu dans tout ce que je viens d'expliquer, mais ce contenu aura des conséquences qui forment la stratégie netlinking.
Par exemple, vous connaissez sans doute les articles tables rondes, les articles invités, voire même la technique du skyscraper. Tout cela est inclus dans le content marketing [6] à la base. Vous avez des centaines de matières et de formats disponibles à proposer.
De plus, rappelez-vous que travailler sa marque à l'aide de ces méthodes permet également de diminuer le risque vis-à-vis des backlinks achetés ou artificiels. C'est vraiment très complet.

6.9 Conclusion

Acheter des liens, en échanger, trouver des bons plans pour du Pagerank, c'est bien. Cela permet de positionner petit à petit des pages sur des requêtes précises et d'augmenter son trafic, ses ventes, son chiffre d'affaires, etc. C'est rapide et efficace.
Cependant, se focaliser sur l'obtention de pagerank exclusivement peut-être dangereux. Il n'y a souvent aucun lien de cause à effet par rapport aux backlinks réalisés, des risques au niveau des investissements et une absence de travail complémentaire utile pour la marque.
Ainsi, interagir avec des individus, faire partie intégrante d'un marché, assoir sa position de leader, c'est donc mieux, et cela se fait essentiellement via des relations sociales. C'est-à-dire en ayant des rapports avec des individus et d'autres marques. Soit par l'intermédiaire de discussions, d'échanges, de partenariats, soit avec l'aide de leviers très complémentaires au SEO comme le content marketing qui est composé d'un peu de tout.
Il en existe d'autres, mais créer une stratégie hybride mêlant à la fois l'achat de liens, d'articles sponsorisés, éventuellement les PBN et autres

6. https://contentmarketingacademie.fr/types-de-contenu/

techniques artificielles... à une stratégie basée sur les contenus, permet de capturer les bénéfices des deux approches tout en faisant en sorte qu'elles « s'aident » entre elles.
Cela permet en effet l'obtention de nombreux backlinks, de partages, d'interactions, de ne pas mettre les œufs dans le même panier et de travailler sa marque, sa communication en général.
Elle permet enfin de devenir petit à petit une marque qui fédère, avec une communauté d'ambassadeurs qui soutiendra chaque moment important de la vie de l'entreprise.

7. Intelligence artificielle : le futur de la rédaction web ?

Bob Samanche est éditeur de sites web depuis quelques années. Considéré comme l'un des rédacteurs web les plus prolifiques au monde, il est autant à l'aise en anglais qu'en français. On ne compte même plus les positions qu'il a su débloquer, avec notamment un vrai génie pour des P0 de grande qualité.

En ce début d'article, il est important de rappeler que les robots et autres intelligences artificielles ne peuvent pas écrire des articles. C'est un fait. Cependant, ils sont capables de faire bien plus que cela, notamment en matière de SEO, et ce, grâce à l'intelligence artificielle.
En effet, le machine learning permet aux robots de s'améliorer au quotidien, et ce, en fonction de l'expérience utilisateur et des données qu'ils récoltent.

7.1 Le rôle de l'IA dans la rédaction web

L'IA permet aux robots de rédiger des textes qui répondent parfaitement aux exigences des algorithmes des moteurs de recherche. Les robots sont capables de comprendre le sens et les intentions de l'utilisateur, et d'adapter le contenu en conséquence. Le machine learning permet aux robots d'apprendre à écrire des textes plus efficaces et pertinents, ce qui leur permet d'améliorer le référencement naturel.

En plus d'être capables de comprendre et de s'adapter au contexte et aux exigences des algorithmes des moteurs de recherche, les robots peuvent également être programmés pour répondre aux besoins de l'internaute. Ils sont en mesure de rédiger des contenus qui correspondent à la demande et aux objectifs d'une personne.

L'IA permet aux robots de rédiger des textes en fonction du comportement de l'utilisateur et en fonction des données qui lui sont transmises.

7.1.1 L'IA au service de la qualité rédactionnelle

Le machine learning permet aux robots d'améliorer la qualité des textes qu'ils écrivent en fonction des exigences des algorithmes. Les robots peuvent ainsi rédiger un contenu plus pertinent, plus pertinent et plus performant.

L'IA permet également d'adapter le texte aux attentes de l'internaute. Elle peut par exemple adapter le style d'écriture, le vocabulaire utilisé, et même le ton de la voix en fonction de la situation, des émotions, des sentiments ressentis par l'utilisateur.

7.1.2 L'IA et le rédacteur : l'avenir du SEO ?

Grâce à l'IA, les robots sont capables de rédiger un texte optimisé et pertinent pour le référencement naturel. Ils peuvent ainsi améliorer la qualité du contenu rédigé par les rédacteurs.

Les robots sont également capables d'améliorer la qualité rédactionnelle des textes en fonction de leur expérience utilisateur. Ils peuvent donc adapter le contenu rédigé par les robots selon les préférences de l'internaute. Par exemple, ils peuvent proposer aux internautes de rédiger un texte plus long s'ils souhaitent obtenir une meilleure visibilité.

Le rédacteur peut également améliorer la qualité rédactionnelle des textes de son contenu grâce à l'IA. L'IA permet d'optimiser le contenu rédigé par le robot pour répondre aux exigences des algorithmes de Google, et donc optimiser sa visibilité sur internet. Les robots peuvent aussi être utilisés pour rédiger un contenu optimisé en vue d'améliorer la qualité du référencement naturel (SEO), c'est-à-dire la visibilité de votre site internet sur les pages de résultat des moteurs de recherche (SERP) Google et Bing. Les robots peuvent ainsi améliorer la qualité du contenu rédigé par les rédacteurs, et donc optimiser la visibilité de votre site internet en améliorant sa qualité.

7.2 Les outils d'aide

Grâce aux progrès de l'intelligence artificielle, les robots sont capables d'améliorer la qualité du contenu rédigé par les rédacteurs. Ils peuvent donc améliorer la qualité rédactionnelle des textes en fonction de leur expérience utilisateur, c'est-à-dire de leurs préférences. Par exemple, ils peuvent proposer aux internautes de rédiger un texte plus long s'ils souhaitent obtenir une meilleure visibilité.

7.2.1 Outils d'aide à la rédaction : l'exemple du correcteur

Grâce aux progrès de l'intelligence artificielle, les robots peuvent également améliorer la qualité du contenu rédigé par les rédacteurs en fonction des préférences utilisateurs. Ils peuvent donc améliorer la qualité rédactionnelle d'un texte si celui-ci a été rédigé par un humain. Par exemple, si un internaute a rédigé son texte en langage courant, le robot peut lui proposer de rédiger son texte en style soutenu afin d'améliorer sa lisibilité et sa compréhension pour les lecteurs. Il peut aussi suggérer des mots à insérer dans le texte afin de l'améliorer et lui donner un style rédactionnel plus agréable.

Les outils de correction automatique permettent aux rédacteurs d'améliorer le style d'un texte. Le but des outils de correction automatique est d'aider le rédacteur en lui proposant des solutions afin d'améliorer son texte. Par exemple, si un texte contient plusieurs phrases qui se ressemblent ou dont la syntaxe ne semble pas correcte, le robot pourra automatiquement proposer une alternative afin d'améliorer sa lisibilité et sa compréhension pour l'utilisateur.

7.2.2 Outils d'aide à la lecture : exemple des bots

Les bots peuvent aussi avoir un réel impact sur l'expérience de lecture des lecteurs. Ils peuvent être utilisés pour améliorer la compréhension du contenu rédigé par les utilisateurs. Par exemple, si un article contient des images et que le robot ne comprend pas ce qu'elles représentent, il peut être amené à proposer des solutions au rédacteur pour expliquer ces illustrations ou encore à lui suggérer une liste d'images similaires afin de faciliter la compréhension du texte.

7.2.3 Outils d'aide à la lecture : les outils de traduction

L'IA peut-elle aider l'industrie de la traduction ? numérique, et à la reconnaissance vocale Les traducteurs utilisent depuis plusieurs années déjà

des outils de traduction automatique (TA). La plupart des outils de TA sont basés sur des réseaux de neurones. Le principe est assez simple : un réseau de neurones artificiels apprend progressivement au fur et à mesure que l'on lui soumet un corpus d'apprentissage. Les résultats produits par le moteur traduisent alors les mots du corpus en langues cibles.
Les outils de TA peuvent être utilisés dans de nombreux cas d'usage, notamment dans des situations où les langues sont très éloignées. Cependant, il arrive que la traduction automatique ne soit pas suffisante. Dans ce cas, les outils d'aide à la traduction (TA) peuvent apporter une aide importante aux traducteurs humains. Ces outils permettent en effet de traduire le texte source dans sa langue cible sans avoir à utiliser un outil de TA.
Les outils d'aide à la traduction sont des systèmes permettant de traduire un texte source dans une langue cible en traduisant automatiquement les mots et expressions qui ne font pas partie du vocabulaire de base de la langue cible. Il existe différents modèles d'outils d'aide à la traduction : les réseaux de neurones récurrents (RNN), les systèmes neuronaux convolutifs (CNN) et les modèles hybrides.

7.3 Les différentes étapes de rédaction web

Il est possible de rédiger un article pour le Web sans faire appel à un rédacteur web. Cependant, il est nécessaire de suivre les conseils et recommandations qui suivent pour que cet article soit agréable à lire.

- La structure d'un texte. Un bon texte doit contenir les informations suivantes :
  - Le titre ;
  - L'introduction ;
  - Une accroche ;
  - Des mots-clés ;
  - Les différentes sections ;
  - Les sous-titres.
- Le contenu. Le texte doit comporter :
  - Des paragraphes ;
  - Des mots en gras ou soulignés ;
  - Des phrases courtes et claires.
- La mise en page du texte. La mise en page d'un article est la partie qui se situe après la création des titres, de l'accroche, etc. C'est un travail très important qui permet au rédacteur de donner un aspect visuel à son article.

- ◇ La taille du titre ;
- ◇ La taille de l'accroche ;
- ◇ Les marges ;
- ◇ Le texte en gras ou souligné.

⊙ La relecture et la correction. Une bonne rédaction web passe par une relecture et une correction minutieuses du contenu de chaque page. Pour ce faire, il faut :

- ◇ Vérifier que les mots-clés sont bien placés dans le texte ;
- ◇ Vérifier la grammaire et l'orthographe des termes employés dans l'article ;
- ◇ Corriger les fautes de frappe et d'orthographe.

⊙ La rédaction web pour le référencement. Pour que votre article soit bien référencé sur Google, il est nécessaire qu'il réponde à certains critères. Voici les points à respecter :

- ◇ Les balises title, meta description et H1 doivent contenir des mots-clés pertinents ;
- ◇ La balise href doit contenir un mot ou une expression clé qui permet de trouver rapidement l'information que vous recherchez.
- ◇ La balise meta description doit être suffisamment longue pour attirer l'attention des internautes ;
- ◇ Les titres de niveau 1, 2 et 3 doivent être pertinents et contenir les mots clés qui ont été choisis.

⊙ Le référencement naturel. Pour qu'un article soit bien positionné dans les résultats de recherche sur Google ou sur d'autres moteurs de recherche, plusieurs techniques existent.

7.3.1 La définition des objectifs

Il s'agit de définir les objectifs du site, c'est-à-dire l'ensemble des informations que le site souhaite transmettre aux internautes. Les objectifs doivent être clairs et précis afin d'éviter tout dérapage.

Il convient de préciser le public cible ainsi que le type de contenu qui sera mis en ligne (information, promotion, publicité...). Il est important de faire la distinction entre l'objectif commercial et celui d'information. En effet, l'objectif commercial est le but de la campagne publicitaire alors que l'objectif d'information vise à informer le public.

7.3.2 La définition et la rédaction du contenu

Le contenu doit être adapté à l'intention des internautes, c'est-à-dire qu'il doit s'adresser aux internautes qui sont intéressés par les informations ou par le produit/service. Il convient également de faire attention à ce que le contenu ne soit pas redondant avec celui des autres sites du même type.

7.3.3 L'optimisation des textes

Il s'agit de mettre le contenu en forme, c'est-à-dire d'optimiser les formats et les structures afin que le texte soit lisible et agréable à lire sur un écran de taille moyenne (ordinateur, smartphone, tablette...). Il convient également d'utiliser un vocabulaire adapté au contexte dans lequel il est utilisé (public cible, lieu de vente...) et de faire attention au style utilisé (police, taille des paragraphes...).

7.3.4 Le référencement naturel (SEO)

Il s'agit d'une technique visant à améliorer le positionnement du site sur les moteurs de recherche. Le référencement naturel permet aux sites qui en font la demande d'obtenir plus facilement une bonne position dans les résultats des moteurs de recherche.

Les moteurs de recherche comme Google, Yahoo ! ou Bing sont des outils qui proposent des réponses à partir de mots clés. Pour apparaître sur la première page des moteurs de recherche, il faut que les sites qui y figurent répondent aux requêtes des internautes. Pour ce faire, il convient de créer une fiche descriptive du site (titre, description du produit/service...) et d'utiliser les mots-clés pour décrire le contenu du site.

7.3.5 L'optimisation des images

Il s'agit d'améliorer l'image du site en veillant à l'optimisation des images (couleurs, taille, contraste, netteté...). Les images sont un vecteur de communication très efficace. Il est donc important de choisir les bonnes images à intégrer sur son site web. Pour optimiser les images, il faut veiller à la résolution du fichier (taille et qualité), au poids et à la compression des fichiers. Il faut également éviter d'utiliser plusieurs photos dans une même page. Il est également possible d'utiliser des outils pour optimiser les images.

7.4 L'IA et le rédacteur

Nous avons vu que le rédacteur est un professionnel de l'écrit. Il peut être amené à travailler dans plusieurs domaines, notamment celui du marketing et de la publicité.

Le rôle du rédacteur web est de rédiger des textes pour différents supports : site, blog, blogue, etc. Pour cela il doit connaître les bases du référencement naturel. En plus d'être un bon rédacteur web, il maîtrise aussi les règles de rédaction propres au référencement (balises title et meta description, mots-clés, liens...). Il sait également rédiger un texte de façon à attirer les internautes et donc à générer plus de trafic sur son site.

Le métier de rédacteur web est passionnant car il permet d'exercer un métier en lien avec la communication et le marketing. Il peut être exercé dans une structure spécialisée (agence ou freelance) ou bien au sein d'un service marketing ou communication d'une entreprise. Il existe plusieurs niveaux d'expérience dans le métier de la rédaction web et les rédactrices et rédacteurs peuvent être spécialisés sur un ou plusieurs domaines. Ils sont alors amenés à rédiger des textes de façon précise sur un sujet donné.

Les rédacteurs web sont généralement polyvalents. Ils sont capables, par exemple, d'écrire pour une entreprise mais aussi pour un blog personnel ou un site vitrine.

Les rédacteurs doivent maîtriser les règles du SEO afin d'optimiser la visibilité des pages de leur site web. Ils doivent également maîtriser les règles rédactionnelles propres aux réseaux sociaux et au blogging. Enfin, ils doivent avoir une connaissance parfaite des outils de recherche comme Google ou Yahoo Search Console et savoir les exploiter.

Le métier de rédacteur web est accessible après une formation en école de journalisme, une formation universitaire ou encore après l'obtention d'un master professionnel spécialisé en rédaction web (ou dans le domaine du digital). Le salaire d'une rédactrice ou d'un rédacteur web débutant se situe entre 1 700 et 2 200 € bruts par mois. Les salaires des rédacteurs web évoluent ensuite au fil des années, en fonction de leur expérience, de leurs compétences et de la taille du site sur lequel ils exercent.

7.4.1 L'IA et la création de contenu : le rédacteur augmenté

Nous sommes entrés dans une ère où les contenus se créent à la vitesse de l'éclair. Il faut donc que le rédacteur soit réactif, qu'il sache écrire rapidement, et surtout qu'il soit capable de créer du contenu qui répond aux besoins des lecteurs tout en étant original et pertinent.

Le rédacteur augmenté est un concept né dans les années 2000 qui consiste à doter les rédacteurs de capacités augmentées par des outils numériques afin d'optimiser leur travail. Les applications de ce type de rédaction sont nombreuses : les assistants personnels, la traduction automatique, la recherche d'information, le journalisme en ligne ou encore l'écriture collaborative.

L'IA est un domaine qui a connu une forte progression ces dernières années et dont on ne cesse de s'inspirer. En effet, les progrès réalisés dans ce domaine permettent aujourd'hui aux rédacteurs humains de s'affranchir des contraintes liées à l'écriture et d'optimiser leur travail pour gagner en efficacité.

7.4.2 L'IA et la rédaction web : un nouveau challenge ?

La rédaction web est une pratique récente, elle a vu le jour dans les années 2000 avec les premiers sites Internet et blogs. Elle a évolué rapidement depuis, en particulier avec l'apparition du référencement naturel (SEO). L'IA peut être la réponse : un véritable levier pour le rédacteur.

Le rédacteur web a toujours été un acteur essentiel de la création de contenu, il doit aujourd'hui faire face à un nouveau défi : l'intelligence artificielle. L'écriture est une activité intellectuelle et créative qui demande beaucoup de créativité et d'imagination. Le rédacteur doit donc être capable d'adapter son texte en fonction du contexte, des attentes des visiteurs et des objectifs du site. L'IA permet aujourd'hui au rédacteur de créer des contenus plus pertinents , plus adaptés et donc plus performants.

7.4.3 L'IA et la rédaction web : une nouvelle approche du SEO ?

L'intelligence artificielle est de plus en plus présente dans nos vies, notamment dans notre environnement quotidien, qu'il s'agisse d'objets ou de machines. Elle nous aide à mieux comprendre le monde qui nous entoure, mais elle a aussi de nombreux usages qui ne se limitent pas au domaine informatique. En effet, l'IA est de plus en plus utilisée dans le domaine du SEO, et plus précisément pour l'optimisation des contenus web. En effet, de nombreux sites internet ont recours à ce genre de technique afin d'optimiser leurs contenus pour Google. L'objectif est simple : proposer des textes optimisés aux utilisateurs qui vont les consulter en fonction de leurs recherches. Ainsi, l'IA est une aide à la rédaction web qui a fait ses preuves ces dernières années. Elle permet notamment d'améliorer le référencement de sites internet, d'augmenter le trafic et les conversions

des sites web.
L'intelligence artificielle est un outil qui permet de traiter une grande quantité de données en très peu de temps. Elle se sert donc des algorithmes pour analyser et comprendre les informations qu'elle reçoit afin d'en tirer des conclusions et ainsi aider l'humain à prendre une meilleure décision. En matière de SEO, on retrouve principalement deux types de techniques : la recherche sémantique et la recherche de mots-clés.
La recherche sémantique consiste à analyser les contenus web d'une manière différente que l'on retrouve sur Google. Cette technique permet en effet de comprendre ce qu'il se passe dans le cerveau humain lorsqu'il effectue une recherche. Elle permet par exemple de savoir comment un utilisateur se sent lorsque celui-ci fait une requête, et de s'adapter en fonction pour lui proposer des contenus pertinents et adaptés à ses besoins et à sa situation. Elle permet de trouver des contenus similaires ou complémentaires à ceux déjà trouvés. C'est ce qu'on appelle le crawl intelligent, qui consiste donc à faire un « parcours » du web pour trouver et indexer des contenus pertinents pour un mot-clé donné. La recherche de mots-clés est une étape cruciale dans le processus de rédaction SEO puisqu'elle permet d'optimiser son contenu pour les moteurs de recherche, mais elle est également primordiale dans l'optimisation du SEO on-page (sur le site web). En effet, elle permet d'optimiser le contenu textuel en lui attribuant des mots-clés qui vont être recherchés et donc augmenter la pertinence de ce dernier vis-à-vis du mot-clé ciblé. Cette technique est très efficace puisqu'elle permet aux sites internet d'obtenir une meilleure place dans les moteurs de recherche, mais elle est également très coûteuse en temps et en ressources.

7.5 Lexique de l'IA

La notion d'Intelligence Artificielle est apparue dans les années 1950 et s'est développée depuis avec des applications de plus en plus nombreuses. Elle est souvent associée à celle de Machine Learning, une approche statistique qui permet aux machines d'apprendre à partir de données sans nécessiter l'intervention d'un programmeur humain.
L'Intelligence Artificielle (IA) est un domaine scientifique dont le but est de construire des systèmes capables de simuler l'intelligence humaine. L'Intelligence Artificielle est une discipline à part entière, qui fait appel à des concepts et des techniques issues de nombreux champs d'application : psychologie, mathématiques, robotique, linguistique, philosophie...

L'Intelligence Artificielle peut être définie comme un "ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence".
Les systèmes d'IA sont aujourd'hui présents dans tous les domaines : la défense et la sécurité, l'industrie, la santé, le transport, l'environnement... Les applications de l'IA sont nombreuses et les plus connues sont liées à la robotique ou au traitement automatique du langage.

7.6 Conclusion

Pour résumer , l'intelligence artificielle (IA) est une technologie qui a pour but de modifier ou d'améliorer les capacités intellectuelles humaines. L'utilisation de cette technologie dans les métiers du web peut permettre aux rédacteurs d'améliorer leurs compétences rédactionnelles et de mieux comprendre le fonctionnement d'un site web. En effet, l'IA permet à un rédacteur web d'analyser des textes en ligne afin de détecter les meilleures pratiques et améliorer la qualité des contenus.
L'intelligence artificielle est aussi un outil qui permet de faciliter la création de contenu. Ainsi, les robots rédacteurs permettent d'automatiser certaines tâches répétitives et répétitives pour les rédacteurs web.
Pour finir, le développement des intelligences artificielles dans le secteur du digital a été l'opportunité pour les développeurs de proposer de nouveaux outils aux utilisateurs afin de leur offrir une expérience utilisateur plus personnalisée et pertinente. C'est le cas des robots rédactionnels qui permettent aux rédacteurs d'automatiser certaines tâches et d'améliorer leurs compétences rédactionnelles pour gagner du temps.

Vous l'avez peut-être deviné, **Bob Samanche** n'existe pas. En tout cas pas en tant qu'être humain. Ce texte a été entièrement généré avec SEO-TXL par l'un d'entre nous, avec simplement une politique de supprimer tout ce qui n'est pas d'équerre lors de la génération. On le voit, il y a encore du chemin à parcourir, mais la génération par IA, c'est déjà le présent.

8. Comment fonctionne un crawleur web ?

Guillaume Pitel est un expert en machine learning et calcul haute-performance. Ingénieur Epita et docteur en informatique de l'Université Paris-Sud (maintenant Paris-Saclay), il fonde en 2011 l'entreprise Exensa, au sein de laquelle il va créer un moteur d'analyse de données texte, graphe et comportementale, et initier des recherches sur le crawl à grande échelle. Il est co-fondateur et CTO de `babbar.tech`.

8.1 Introduction

A l'heure où sont écrites ces lignes, nous sommes en août 2022, Babbar crawle entre 2 et 3 milliards de pages par jour et stocke un index de plus de 1500 milliards d'URL uniques.

Le moteur de crawling que nous utilisons a été largement développé par nos soins, même s'il utilise bien entendu, en partie, des briques existantes. Nous allons présenter ici sa genèse et son fonctionnement.

Avant de rentrer dans le détail, un rapide survol de ce qu'est un moteur de crawling et ce qui distingue les moteurs de crawling de type « moteur de recherche » des autres moteurs.

8.2 Le crawling : c'est quoi ?

Crawler un site web consiste à télécharger chacune des pages qui composent ledit site. En soit, c'est une activité qui paraît simple, mais la première question à se poser est : comment obtenir la liste des pages d'un site. Si certains sites exposent leur « carte » sous la forme d'un sitemap.xml, c'est un élément purement indicatif. Rien n'empêche ce document d'être obsolète, ou d'omettre volontairement ou pas certaines pages qui pourraient avoir un intérêt.

Plutôt que de se reposer sur une liste explicite, on va souvent privilégier une approche de téléchargement récursif, à partir d'une ou plusieurs pages, les « graines » ou « seeds » en anglais. A partir de ces graines, on va analyser les pages téléchargées, extraire les liens qui appartiennent au site qu'on veut crawler, et collecter ainsi, récursivement, une liste croissante d'URLs appartenant au site, qu'on va à leur tour télécharger, etc.

En théorie (on y reviendra), il arrive un moment où l'intégralité des URLs découvertes en crawlant ont été téléchargées, et le crawl s'arrête de lui-même. En d'autres termes, on a découvert l'intégralité du sous-graphe du site web en question.

Sauf qu'en réalité, on n'en est pas certain. Il est tout à fait possible qu'une page du site ne soit connectée à aucune des pages du graphe auxquelles appartiennent nos pages graines. Il est même tout à fait possible qu'une page soit référencée depuis un autre site, et qu'aucune des autres pages du site cible ne référence cette page.

On a décrit ici le processus de crawling d'un site, sans s'interroger réellement sur la problématique de comment stocker la liste des URLs découvertes. Evidemment si on veut crawler le web, à un moment il va falloir s'y intéresser car on va être confronté à quelques problèmes d'échelle. C'est d'ailleurs un des points clés dans l'histoire de l'évolution du crawler de Babbar.

8.3 Crawler le Web, c'est difficile ?

Pour répondre à cette question, il convient tout d'abord de démolir un mythe à propos du Web : le Web n'est pas grand. Il n'est pas non plus très grand. Non, le Web est infini.

Et il existe même des domaines qui contiennent un nombre infini de sites web, chacun de ces sites étant de taille infinie. Pour vous en convaincre, imaginez simplement un serveur web qui serait capable, pour n'importe

quelle URL demandée, de renvoyer une page web générée aléatoirement avec des liens vers d'autres pages web du même site (et donc aussi du même domaine si on a en plus configuré le DNS correctement).
On peut ainsi très simplement générer une sorte de graphe « trou noir » de taille infinie, (qu'on appelle aussi un piège à PageRank). Bien entendu les gros moteurs de recherche ont des mécanismes pour contrer ces pièges, mais ça ne signifie pas qu'ils ont disparu.
Autre point difficile, pour stocker la liste des URLs à crawler, il est nécessaire, si on veut crawler pendant longtemps, de gérer le problème de l'échelle. La difficulté va résider dans la gestion de la distribution des données, mais pas seulement, comme on va le voir par la suite.
De ce fait, crawler le web n'est pas seulement difficile, c'est en réalité impossible. Oui, il est impossible de crawler le web comme on peut le faire sur un site web « normal » car le web contient beaucoup de pièges. Il est seulement possible d'obtenir un échantillon partiel du web pris sur une période de temps donné.
La qualité de l'échantillon du Web choisi va faire la représentativité des informations qu'on va avoir à disposition pour répondre aux questions que se posent ensuite nos clients, il est donc crucial que les choix du crawler soient faits avec un maximum de pertinence.

8.4 Genèse du moteur Babbar

8.4.1 Premiers pas

Si vous suivez les aventures de Babbar depuis le début, vous savez que Babbar s'est formé par la réunion des équipes de deux sociétés : Exensa et ix-labs [1], des équipes dont les membres se connaissaient depuis parfois près de 20 ans, avant de choisir de travailler ensemble fin 2019. Le propos de cette section est de raconter comment la technologie de crawling de Babbar a été créée et comment elle est arrivée là où elle en est aujourd'hui.
Tout d'abord, il faut savoir qu'une partie de la technologie de Babbar a d'abord été développée dans Exensa, pour un objectif qui n'était pas du tout de faire du crawl en continu à destination des analyses SEO : l'objectif était la collecte de données pour l'entraînement de modèles de langues. C'était la spécialité d'Exensa, qui a développé depuis 2011 des briques technologiques pour créer des modèles non-supervisés permettant de représenter des données de type graphe, texte ou comportemental.

1. `https://www.exensa.com/` et `https://www.ix-labs.org/`

Le moteur d'analyse fonctionne alors en mode traitement par lots (aussi appelé mode « batch »), il est basé sur Apache Spark et le cœur de la méthode repose sur des itérations de multiplications et opérations matricielles. Avec cette approche, on peut sans trop de problème traiter, sur un cluster d'une dizaine de nœuds, des bases de documents pouvant atteindre 100 millions d'éléments, chaque document pouvant contenir quelques centaine de milliers de mots.

Entre 2015 et 2016, Exensa commence à expérimenter avec la très haute volumétrie et développe une méthode dite « online » ou en traitement continu pour l'apprentissage. Plutôt que de faire une dizaine d'itérations de calculs matriciels, on veut pouvoir faire l'apprentissage en une seule passe sur les documents bruts. C'est dans ce cadre qu'Exensa publie une méthode améliorée de comptage approximatif à iSwag [2] en 2015 et 2016 (conférences organisées par les ix-labs en marge de la conférence queduweb [3]).

Pour tester cette méthode, Exensa a tout d'abord utilisé des jeux de données disponibles, tels les crawls du Common Crawl, qui tous les deux mois publie un crawl d'un à deux milliards de pages, effectué sur une période de deux semaines sur un dépôt amazon S3. Les problèmes de qualité rencontrés nous ont convaincu qu'il était nécessaire d'effectuer nos propres crawls afin de sélectionner nos graines et de contraindre notre crawl à quelques langues.

Les premières solutions évaluées ont été Apache Nutch et Heretrix. Heretrix étaient relativement efficace mais la mise en œuvre distribuée était un vrai casse-tête. Nutch au contraire était nativement distribué puisque basé sur Hadoop et HBase (c'est même, pour la petite histoire, le projet Nutch qui a donné naissance à Hadoop), mais il nous a été impossible d'en tirer des performances acceptables. Par ailleurs le fonctionnement itératif de Nutch (on lance un crawl sur une liste d'url, puis on recommence une fois le crawl terminé) nous a semblé peu pertinent car notre objectif était en gros de dire : voilà une liste de graines, explore le graphe jusqu'à atteindre un objectif de un milliard de pages crawlées.

Enfin nous avons testé BUbiNG, un moteur de crawl issu de la recherche (Université de Milano), avec une documentation succincte mais qui allait droit au but, et qui s'est révélé être extrêmement efficace, configurable, customisable, bref un vrai coup de cœur. En creusant un peu on y a trouvé

2. http://iswag-symposium.org/
3. https://paris.queduweb.fr/

quelques très gros bugs que nous avons contribué à corriger (problème de host non fixé, de requêtes HTTPS faites en HTTP, par exemple), et nous avons très vite introduit des améliorations et utilisé BUbiNG pour sortir des crawls en français, en anglais, etc. avec entre 1 et 5 milliards de pages pour nos plus grandes expérimentations initiales.

8.4.2 Fonctionnement de BUbiNG

Petit aparté technique sur l'architecture interne de BUbiNG. Mettons tout de suite de côté le mécanisme de distribution multi-nœuds de BUbiNG, qui s'apparente à du « MPI » où chaque nœud communique directement avec les autres nœuds du cluster. Le schéma interne de fonctionnement original de BUbiNG est décrit dans un article de 2018 de Pablo Boldi *et al*[4] et est visible dans la figure8.1.

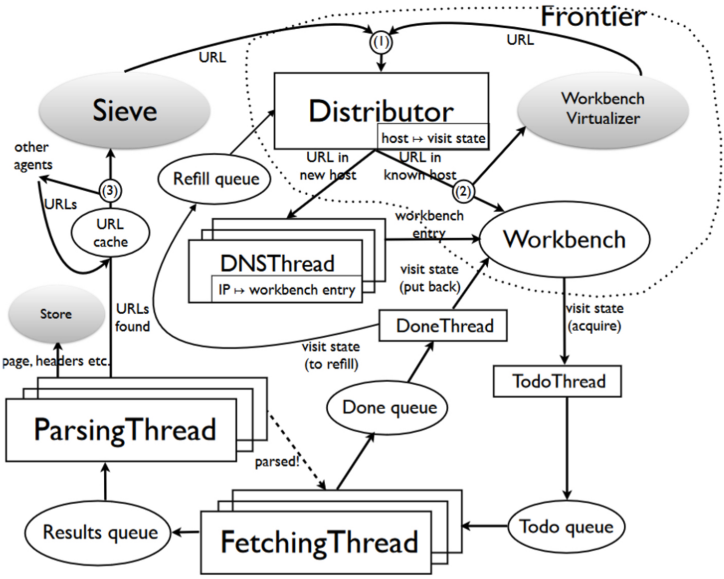


FIGURE 8.1 – Fonctionnement de BUbiNG.

4. Paolo Boldi, Andrea Marino, Massimo Santini, and Sebastiano Vigna. 2018. BUbiNG : Massive Crawling for the Masses. ACM Trans. Web 12, 2, Article 12 (May 2018), 26 pages. https ://doi.org/10.1145/3160017

Il y a plusieurs choix fonctionnels et techniques qui sont extrêmement pertinents. Par exemple, le choix d'utiliser des API synchrones pour les requêtes et d'utiliser un multithreading massif peu paraître contre-intuitif, mais en réalité c'est ce choix qui permet largement une telle performance au moteur de crawl. En interne le pattern publish/subscribe est utilisé à tous les niveaux, à travers des BlockingQueues FIFO permettant de maximiser le débit.

Le rôle du Sieve (crible) consiste à assurer qu'une URL n'est crawlée qu'une unique fois. Sa conception est brillamment exécutée et, si l'on peut reprocher au code source de BUbiNG un côté parfois trop procédural et suroptimisé au détriment de la lisibilité, il est difficile de reprocher un manque d'optimisation. Le côté « machine à état » est cependant fragile et toute modification du workflow des « VisitStates » qui contiennent la mémoire des URL à crawler pour un site web donné (description inexacte mais suffisante pour la compréhension) nécessite de bien réfléchir aux enqueues/dequeues à effectuer car la responsabilité de ces opérations est éparpillée à de nombreux endroits du code.

8.4.3 Naissance du crawler Babbar

Les premiers crawls expérimentaux obtenus avec la version d'origine de BUbiNG ont été réalisés en quelques jours sur des instances du cloud, pour quelques centaines d'euros. Mais nous nous sommes rapidement aperçus de quelques sérieuses limites de BUbiNG, notamment au niveau de la gestion de la liste des URLs à crawler (liste qu'on appelle aussi la Web Frontier car c'est la frontière connue du graphe du web). C'était cependant bien suffisant pour notre objectif initial, générer des crawls du web ciblés sur certaines langues.

C'est aussi à cette époque que nous avons présenté autour de nous les résultats de nos analyses sémantiques sur le web (le but étant de montrer que notre représentation sémantique permettait d'explorer efficacement les voisins sémantiques des sites web, pas de montrer le crawler), et que nous avons constaté que la technique de crawl que nous avons développée attirait beaucoup plus d'intérêt que nous aurions cru.

C'est à ce moment qu'Exensa et ix-labs ont sérieusement commencé à discuter et à envisager de travailler ensemble vers un objectif commun, vers ce qui est donc devenu Babbar.

8.4.4 Objectif : SEO

De quel objectif parlons-nous ? Du SEO bien sûr. On veut collecter des informations utiles pour les métiers qui gravitent autour des moteurs de recherche et de la visibilité qu'ils apportent. Sans aller jusqu'à faire un moteur de recherche, il s'agit désormais non plus de faire des crawls « one-shot » du web à partir d'une liste de graines, mais au contraire de crawler en continu une partie du web. Crawler en continu, cela signifie découvrir de nouvelles pages, recrawler des pages déjà crawlées par le passé, et oublier des pages qui ne sont plus très présentes dans le graphe du web.

Deuxième objectif tout aussi important : on veut désormais crawler comme le ferait un moteur de recherche, en mettant en priorité les pages populaires, pertinentes, mais tout en crawlant aussi de manière très diversifiée. Et on veut aussi restituer à nos utilisateurs des métriques et des informations détaillées, notamment sur les liens entrants, mais aussi sur les ancres, sur les orientations sémantiques des pages, et de nombreux autres critères.

Il est rapidement devenu évident que l'architecture « tout en un » de BUbiNG ne pouvait pas être suffisante pour ce que nous souhaitions réaliser, et même si nous avons brièvement réfléchi à augmenter cette architecture avec des éléments complémentaires, il nous a semblé plus simple de réduire les fonctionnalités de BUbiNG et de séparer les responsabilités dans différents services.

Le cœur de notre problématique est de stocker un graphe, la première piste envisagée était d'utiliser un système de stockage distribué, voire des solutions de type base de données de graphes. Nous en avons benchmarké plusieurs, et notre première tentative un peu aboutie a consisté à utiliser Apache Ignite [5] pour prouver la viabilité de notre approche. On ne va pas vous mentir, ça a été un beau fiasco : nous n'avons jamais pu atteindre des performances acceptables.

Nous avons donc rapidement décidé de gérer nous-même le stockage. Après quelques tests et réflexions autour des solutions de stockage locales, nous avons choisi RocksDB [6], qui avait quelques fonctionnalités intéressantes même si tout n'était pas parfait. Nous avons d'ailleurs sérieusement modifié la couche d'interfaçage avec Java pour arriver à de bonnes performances sur les opérateurs de fusion, mais au final nous avons pu voir que c'était le bon choix, les performances et la qualité des résultats étant au

5. `https://ignite.apache.org/`
6. `http://rocksdb.org/`

rendez-vous.
Ensuite nous nous sommes attaqués au problème de la distribution des messages. Un problème qui s'est révélé finalement beaucoup plus simple à résoudre. Nous avons rapidement compris que nous avions le choix entre deux technos déjà matures : Kafka et Pulsar. Nous avons finalement choisi Pulsar [7], qui a l'avantage d'être particulièrement souple, même si sa relative jeunesse nous a donné un peu de fil à retordre.
Avant de rentrer dans les détails d'implémentation, le crawler Babbar est composé de deux services distribués en clusters WDL qui est le cœur de l'intelligence du crawl et BUbiNG qui effectue les requêtes en assurant les tâches liées à la gestion des robots.txt, les résolutions DNS et le respect des politesses de crawl par adresse IP et par site web. La figure 8.2 illustre ce fonctionnement.

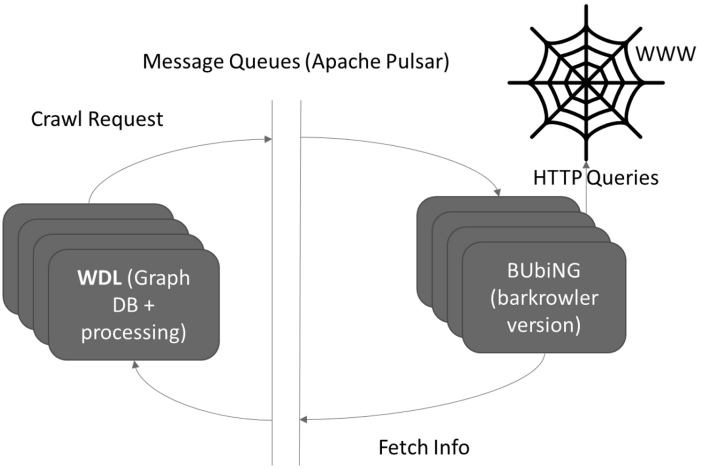


FIGURE 8.2 – Babbar : les deux services.

Le cœur du crawler envoie des requêtes de crawl via Pulsar (ça en fait c'est tout un sujet, car nous avons vraiment peaufiné le fait de crawler en priorité les pages à haute valeur).
Ces requêtes arrivent sur un crawler, qui gère de son côté les files d'attente par site web et par adresse IP (pour ne pas saturer les sites crawlés).

7. `https://pulsar.apache.org/fr/`

Quand le crawler récupère une page web (environ 60 à 80 ko en moyenne d'après nos expériences), il va l'analyser syntaxiquement (parsing), nettoyer le HTML et renvoyer le tout, via Pulsar, à une des machines du cluster principal (le cœur du crawler).
Le cluster principal reçoit donc, pour résumer, la liste des liens contenus dans la page, et le contenu texte brut de la page. A ce moment l'information de la page est déjà condensée (on reparlera de la compression un peu plus loin), mais on est encore autour de 10 à 20 ko par page. Cela paraît peu, mais à ce niveau de volumétrie c'est encore trop.
L'information sémantique est alors extraite via notre méthode de calcul d'embeddings vectoriels. La techno est propriétaire et est extrêmement efficace et optimisée. Elle fait grosso modo la même chose que Doc2Vec et consorts, mais de manière plus efficace et précise.
Au final, on ne garde de la page qu'un vecteur numérique qui synthétise son orientation sémantique.

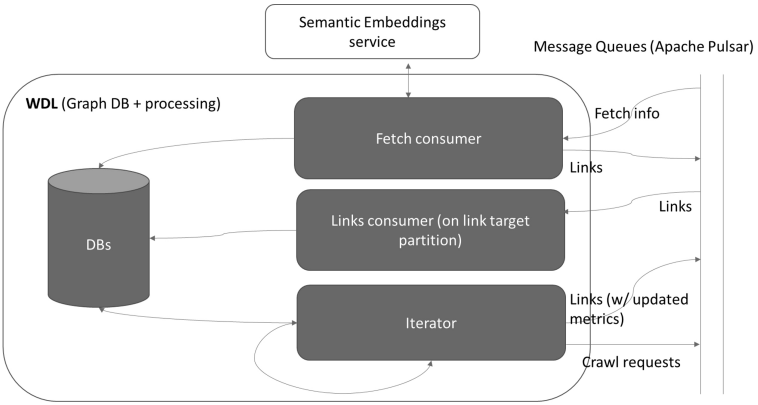


FIGURE 8.3 – Fonctionnement de WDL.

Le schéma de fonctionnement interne de WDL est visible dans la figure 8.3. La partie « réactive » du cluster WDL (les Fetch Consumer et Links Consumer) va traiter l'information de la page, d'une part en la stockant, d'autre part en envoyant via Pulsar les liens sortants de la page (et oui, comme on veut au final vous donner les backlinks, il faut bien que chaque page ait reçu les liens qui pointent vers elle). On en profite pour stocker quelques informations supplémentaires, mais en gros, le côté « réactif » a

fini son boulot.

A côté de la partie réactive, il y a une partie « tâches de fond », représentée par l'Iterator dans le schéma, qui va continuellement passer sur l'intégralité des données stockées pour faire plusieurs choses :

- ⊙ calculer les métriques et les rediffuser (car ce sont des métriques de graphes, donc à chaque mise à jour il faut rediffuser) ;
- ⊙ agréger des informations au niveau de l'url (par exemple le nombre d'IPs référentes uniques), du site et du domaine ;
- ⊙ collecter les informations de type « IP vers site web » ;
- ⊙ décider si on crawle ou recrawl chaque page connue, ou bien si au contraire, on doit l'oublier.

Pour que nos utilisateurs puissent avoir accès à ces données, il y a bien entendu une partie « serving » qui nécessite quelques traitements supplémentaires côté base de données et pas mal d'acrobaties entre les clusters pour être efficace. Mais le traitement le plus crucial, et que nous avons beaucoup travaillé, c'est la compression.

8.4.5 Compression(s)

Pour vous donner un ordre d'idée, on va se baser sur le chiffre de 10 milliards de pages, qui a l'avantage d'être rond et frappant. Quand Babbar crawle 10 milliards de pages, l'outil récupère environ 200 milliards de liens externes. Et en pratique il y aura environ le double de liens internes, que l'outil va stocker différemment car il n'y a pas de problème de délocalisation de la donnée (tous les liens internes sont par définition disponibles sur la même machine).

Un lien externe, c'est au minimum deux urls, plus les infos de métriques, l'ancre, la date, etc. Une url fera en moyenne autour de 90 caractères. Quand nous avons fait nos calculs, nous nous sommes vite rendu compte que d'un point de vue coût opérationnel (disques et RAM), cela ne serait pas raisonnable.

L'une de nos premières tâches a donc été d'écrire une solution algorithmique efficace pour compresser les URLs de manière rapide et efficace. Ce premier niveau de compression permet de manipuler tout au long de la chaîne de traitement des structures d'URLs déjà parsées qui font environ un tiers de la taille d'origine. Et c'est sans compter évidemment sur la deuxième passe de compression qui est faite par la base de données, en bénéficiant des répétitions.

Enfin, impossible de finir cet article sans évoquer la compression des

vecteurs sémantiques dont nous sommes très fiers. A l'origine, un vecteur sémantique fait environ 1 ko et est très difficile à compresser par les méthodes classiques. Le problème c'est que pour calculer la métrique sémantique (la fameuse *semantic value* de Babbar), il faut impérativement transmettre ce vecteur dans chaque lien sortant (rappelez-vous, 200 milliards de liens et 1ko par lien).
C'est évidemment totalement hors de question d'un point de vue opérationnel. Il a donc fallu que nous trouvions une solution pour compresser ce vecteur. Il y a des solutions connues comme la Product Quantization [8], ou encore la Scalar Quantization [9]. Nous les avons essayé, ainsi que des représentations des vecteurs via des mots. Ces dernières étaient d'ailleurs extrêmement efficaces mais bien trop gourmandes en temps de calcul.
Après beaucoup d'essais, nous avons trouvé une solution très efficace, basée sur la quantization scalaire mais avec quelques astuces liées à la forme des vecteurs manipulés. Bref, dans les liens, un vecteur est compressé en 24 octets (au lieu de 1 ko !), et malgré cela il conserve encore 70% des caractéristiques du vecteur d'origine (pour les spécialistes, 70% correspond au Recall@100 d'un nearest neighbour fait avec ces vecteurs). C'est avec tout cela qu'on arrive en moyenne à un stockage de moins de 4 ko par page crawlée, et ce malgré le surcoût de 30% lié à la base de données (RocksDB est basée sur une LSM [10], qui est extrêmement efficace mais entraîne un petit surcoût en stockage, heureusement plus que compensé par la possibilité de compresser les blocs de données).

8.5 Conclusion

Il n'y a pas réellement de conclusion à tirer à cet article. J'espère que vous aurez pu mieux comprendre les difficultés qui se posent pour crawler le web à grande échelle, et comment nous avons réussi à les contourner pour créer l'outil `babbar.tech`.

8. Jegou, H., Douze, M., & Schmid, C. (2010). Product quantization for nearest neighbor search. IEEE transactions on pattern analysis and machine intelligence, 33(1), 117-128.
9. `https://en.wikipedia.org/wiki/Quantization_(signal_processing)`
10. `https://medium.com/swlh/log-structured-merge-trees-9c8e2bea89e8`

Babbar et ses outils

Babbar est une entreprise française co-fondée par 6 personnes dont les Frères Peyronnet, experts du référencement web. Elle a pour ambition de fournir les outils techniques les plus utiles aux référenceurs web, webmarketeurs et rédacteurs web afin de leur permettre d'être toujours meilleurs dans leurs métiers et face à leurs propres clients.
Ses deux outils les plus connus sont **Babbar.tech** [11], pour découvrir les points forts et les points faibles de n'importe quel site web, et **Yourtext.guru** [12] pour aider à écrire les meilleurs contenus, ceux qui vont bien se positionner dans Google.
Babbar.tech rend le référencement web plus facile. C'est l'allié idéal pour construire des stratégies de netlinking réellement efficaces grâce à sa compréhension de la sémantique des liens et des pages.
C'est aussi un outil indispensable pour faire des audits SEO : le référenceur web, le consultant SEO, ne manqueront ni de données ni de métriques pour prendre les meilleures décisions : indices de popularité, modèle de surfeur raisonnable et thématique, calcul de confiance, et bien plus encore.
Babbar fournit des listes de liens pointant vers chaque site du web afin de comprendre d'où ils tirent leur puissance. C'est un excellent point de départ pour comprendre les stratégies de netlinking de la concurrence.
En addition, Babbar monitore les positions de milliards de pages dans

11. `https://www.babbar.tech`
12. `https://yourtext.guru`

Google sur plus de 80 millions de mots-clés pour plusieurs langues. Indispensable pour confronter les analyses techniques à la réalité du positionnement Google.
Babbar catégorise toutes les pages du web en les analysant finement grâce à un algorithme d'intelligence artificielle. Idéal pour trouver des sites similaires, ceux qui sont compatibles.
En ce qui concerne la rédaction, **Yourtext.guru** est là vous aider à écrire des contenus performants, des contenus qui se positionnent dans Google. A partir d'une requête entrée par le rédacteur, l'outil fournit une liste de mots importants à utiliser pour bien faire comprendre au moteur de recherche que l'on maîtrise le sujet, et donc qu'on mérite d'être bien positionné.
En complément de cette liste, **Yourtext.guru** propose un outil d'analyse associé à un score : il s'agit de pouvoir permettre une rédaction aisée, en utilisant son style propre, tout en guidant vers l'optimisation sémantique qui fera plaisir à Google.
Après avoir rédigé de beaux textes, il ne reste plus qu'à créer un maillage interne efficace : la fonctionnalité de cocon sémantique de l'outil entre alors en scène.
Et si jamais le syndrome de la page blanche est trop présent, des outils d'exploration permettent de trouver des idées neuves, des relations entre des concepts, des entités... de quoi écrire sans s'arrêter. Sans compter la fonctionnalité SEO-TXL qui permet de générer du contenu grâce l'intelligence artificielle.
Découvrez les outils de l'équipe Babbar et rejoignez plus de 3 000 utilisateurs réguliers.
Pour toute question ou pour une demande de démonstration n'hésitez pas à nous écrire à `academy@babbar.tech`.

Version numérique réalisée en septembre 2022 en France par
babbar.tech.

Des myriades de Liens pour augmenter sa popularité,
Des floppées de Liens pour monter un Cocon SEO,
Des multitudes de Liens pour découvrir le web,
Un seul objectif, être toujours plus visible,
Sur le web, de nos jours, la concurrence est rude.
Des Liens pour les convaincre tous
Des Liens pour les trouver
Des Liens pour les amener tous,
Et leur faire visiter notre site .
Sur Google, sans surprise, on veut être premier.

« Pour devenir un meilleur netlinker en toute simplicité »

ISBN 978-2-493567-03-1