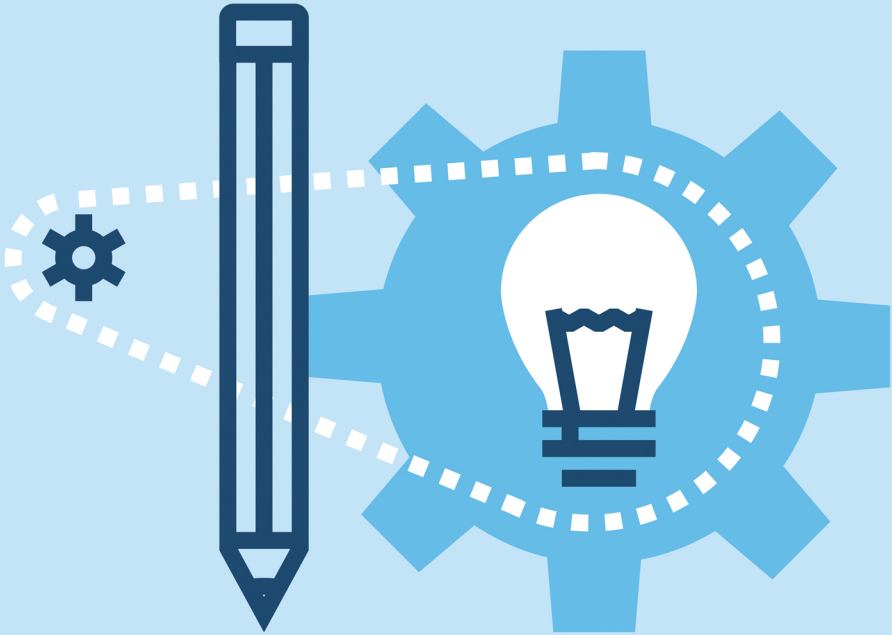


CONTENU WEB

Paroles d'experts

PRÉFACE D'OLIVIER ANDRIEU

AUTEURS AUTRICES

DIRECTION
S.PEYRONNET

S.BAY
M.CAIROU
C.GILLET
B.GUIRAUD
T.LARGILLIER
G.PEYRONNET
S.PEYRONNET
G.PITEL

babbar

Contenu web : paroles d'experts

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

ISBN : 978-2-493567-01-7

DÉPÔT LÉGAL : JUILLET 2022

Table des matières

Préface

Connaissez-vous la blague (méchante et injuste, certes) suivante : « Quand parle-t-on le plus souvent de Microsoft dans le domaine de la recherche d'information et du SEO ? » et sa réponse : « Quand on cite Bill Gates pour sa célèbre phrase : "content is King" » ?

Alors bien sûr, ce serait faire peu de cas du moteur Bing, qui s'est pourtant fortement amélioré en termes de pertinence depuis plusieurs années et qui dame même parfois le pion à Google en termes de fonctionnalités innovantes (le récent protocole IndexNow en est un exemple évident).

Mais cette petite blagounette à 3 sous nous montre bien que, depuis que les moteurs de recherche existent, cette notion de contenu est bien capitale et au centre de toutes les attentions pour tous les acteurs du secteur. Sans contenu, sans texte, point de salut. Ou alors, il ne reste plus que la triche pour tenter d'obtenir une certaine visibilité sur le web. Mais ceci est un autre débat...

Mais le contenu en lui-même ne se pense pas (ou plus) en termes de quantité uniquement. Le SEO, ça ne se fait pas au kilo ! Le temps de la « bouillie » est révolu, et la qualité est bien sur toutes les lèvres. Si les algorithmes actuels ont besoin d'un certain volume en entrée pour analyser un contenu, on peut penser que les systèmes vont s'améliorer au fur et à mesure et qu'à moyen terme, une brève pourra être aussi bien considérée en termes de pertinence qu'un long article de plus de mille mots. Et pourquoi pas qu'une image ou un enregistrement sonore !

En ces temps de Metaverse et de réalité augmentée et virtuelle, on peut

imaginer que la façon de présenter une idée, un concept ou une actualité va évoluer vers des formats différents et le plus souvent complémentaires. Mais ils auront tous la même finalité : nous informer sur des sujets aussi divers que variés.

Les plus anciens d'entre nous (dont je fais partie) se souviennent les galères qu'ils subissaient lorsqu'il fallait chercher, il y a 40 ans de cela, une information de type encyclopédique, qui plus est le plus rapidement possible. Pour ma part, la source se trouvait dans une suite de doctes ouvrages reliés de marque Larousse qui se trouvaient dans le cabinet médical de mon père. Malheureusement, pendant les heures de consultation, je n'y avais pas accès et devais souvent attendre plusieurs heures avant de consulter la signification de tel terme ou en savoir plus sur un pays, une ville ou un personnage historique. L'Internet a changé tout cela et a mis à portée de main - et plus encore avec l'avènement des smartphones - l'information au niveau mondial.

Celle-ci est désormais disponible par défaut. Le pré-carré de la connaissance est devenu une utopie. Le phénomène s'est inversé ou a tout du moins fortement évolué. Ce sont ceux qui digèreront, traiteront, diffuseront et rendront rapidement accessible les données disponibles qui gagneront. Et le challenge est bigrement plus exaltant...

C'est pour cela que le livre que vous avez entre les mains est intéressant, car il explore de nombreuses voies du contenu web, de sa création à sa mise en forme, de sa production à sa transmission. Et quoi de plus beau que de transmettre ?

Merci donc à Babbar pour ces quelques pages qui, je n'en doute pas un instant, sauront vous passionner. Puisque j'ai commencé cette préface en citant Bill Gates, terminons-la avec une autre citation du fondateur de Microsoft : « Dans le futur, les leaders seront ceux qui savent donner le pouvoir aux autres. » Et le pouvoir, c'est bien dans le contenu qu'il se trouve... La boucle est bouclée...

Olivier Andrieu, éditeur du site `Abondance.com`.

Préambule

Pour ce livre, l'équipe de Babbar s'est associée à des experts et expertes du référencement web pour vous offrir leur vision sur une collection de sujets liés au contenu et à son référencement. Les sujets abordés sont techniques et précis mais ne vous inquiétez pas les auteurs et autrices les ont vulgarisé pour que vous puissiez en tirer le maximum.

Guillaume Peyronnet, co-fondateur de Babbar et l'un des pionniers du référencement web en France, vous présentera dans son « Introduction au SEO » tout ce qu'il y a à savoir sur le fonctionnement d'un moteur. Fort de plus de 20 ans d'expérience, Guillaume vous présentera les piliers du SEO indispensables à connaître pour débuter dans le métier.

« Comment rédiger un bon contenu pour le web ? », un titre de chapitre très évocateur dans lequel Baptiste Guiraud, rédacteur web expert et passionné, vous partagera son expérience sur comment structurer, rédiger et optimiser ses contenus pour qu'ils plaisent autant aux moteurs de recherches qu'aux internautes.

Camille Gillet, autrice et storytelleuse nous parlera de l'importance de s'intéresser à l'identité de marque avant de se lancer dans la promotion de contenu et qu'il n'est jamais trop tôt pour s'intéresser à ces questions, la création de récit étant la colonne vertébrale de la communication professionnelle.

Le référencement web est avant tout là pour ramener du trafic sur son site et apporter de la visibilité mais cela n'est pas nécessairement suffisant

pour créer de la rétention d'utilisateurs. Marielle Cairou, spécialiste de la relation client et CSM chez Babbar nous explique, dans « *Inbound marketing* & SEO », comment combiner les deux approches pour obtenir un résultat optimal.

Syphaïwong Bay, consultante SEO et experte éditoriale est l'autrice du chapitre « Quand interagir avec Google sculpte notre vision du monde ». Elle aborde la place des moteurs de recherche, l'évolution de l'accès à l'information et l'influence du SEO sur la connaissance. Syphaïwong développe également comment, grâce au SEO, participer à la construction d'un Internet de la connaissance.

Dans le chapitre « A propos des *embeddings* », Guillaume Pitel, CTO de Babbar et expert en machine learning, nous explique pourquoi utiliser ces « plongements » qui transforment les contenus en vecteurs et en quoi cela nous aide à mieux comprendre les contenus web.

Sylvain Peyronnet, CEO de Babbar, chercheur en algorithmique et spécialiste des moteurs de recherche nous présente, dans le chapitre « A propos de la génération de contenu pour le web », un état de l'art des méthodes et outils de génération de contenu et quels sont les meilleurs usages pour les intégrer dans une stratégie SEO.

Enfin, le dernier chapitre « A propos de l'audit automatique de contenu web » se concentre sur les étapes techniques qui permettent d'analyser un contenu. Thomas Largillier, co-fondateur de Babbar et chercheur en informatique spécialisé dans l'algorithmique liée au web nous présente les points clés de l'analyse de contenu et comment celle-ci peut-être automatisée pour notamment étudier le positionnement d'un site sur une requête.

Bonne lecture.

1. Introduction générale au SEO

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`.

1.1 Introduction

Le référenceur ne craint rien. Il aime s'attaquer aux moteurs de recherche, se heurter aux algorithmes et aux milliers d'ingénieurs déployés par Google. Il adore comparer ses sites et ses pages aux concurrents du web. C'est peu dire qu'il ne faut pas avoir peur, ne pas avoir le syndrome de l'imposteur pour exercer ce métier. Et pourtant, le référenceur web a toujours une boule dans la gorge quand il doit répondre à la fatidique interrogation « Et toi, que fais-tu dans la vie ? ».

Définir le métier du référenceur et la discipline associée est loin d'être une mince affaire. Il existe presque autant de définitions que de pratiquants du référencement naturel.

Repartons des fondamentaux. Le web est un énorme espace où des milliards de milliards de pages web cohabitent. Quand on a besoin d'accéder

à une page donnée, il faut alors savoir comment s'y rendre. En connaissant son adresse, on y parvient vite, de la même façon qu'on arrive rapidement chez un ami dans une ville que l'on ne connaît pas, grâce à son adresse postale.

Mais imaginons que j'ai besoin de trouver un médecin, un plombier, un fleuriste, un boulanger, etc. Comme faire ? Errer au hasard en espérant tomber sur la bonne devanture ?

La richesse du web est à la fois son plus grand atout et sa malédiction : on peut tout y trouver… mais comment faire pour trouver efficacement ? Si je suis lâché à pied au coeur de New York, je rencontrerai bien des difficultés à parcourir toute la ville en quelques heures. Et quand je cherche une information sur le web, je veux trouver la réponse immédiatement. Il me faut une carte, un plan, un annuaire, quelque chose qui pointe les raccourcis.

C'est de ce constat que sont nés les annuaires de sites façon Yahoo, Dmoz, etc. Quand il y a trop de connaissances, de sites, de pages, il faut faire une sélection. On fournit ainsi un passe coupe-file à l'internaute. Mais ce n'est toujours pas assez face au phénoménal volume du web. Ainsi, les moteurs de recherche sont apparus. Maintenant, plus de listes à parcourir, une simple boite de saisie qui permet d'exprimer son besoin informationnel. « Acheter un vélo », « partir en voyage en Australie », « comment réparer ma machine à laver », etc. En quelques millisecondes, le moteur organisera sa base de données, réplication techniquement réfléchie du web, pour classer les pages et fournir une réponse sous forme d'une liste de pages. Le premier résultat est meilleur que le second, lui-même plus pertinent que le troisième, etc.

1.2 Le référencement web

On commence, à ce stade, à pouvoir parler de référencement web : les pages bien positionnées sont plus consultées par les internautes que les pages moins bien positionnées. On dira alors qu'un site bien référencé est un site dont les pages sont bien positionnées pour de nombreuses requêtes, mieux encore, pour des requêtes qui sont très demandées par les internautes (donc des requêtes à fort trafic).

C'est là qu'apparaît le métier du référenceur web : faire en sorte que le moteur de recherche classe mieux les pages de ses sites (ou les pages de ses clients) afin d'attirer du trafic. Le référenceur web a donc pour mission de donner de la visibilité à des pages web dans les moteurs de recherche.

On est donc sur un métier relié aux tâches du marketing : on fait connaitre et on persuade de venir chez soi.

S'arrêter là est trop réducteur. Avoir un apport de trafic sur un site n'est pas toujours intéressant. L'e-commerçant qui vend des livres et qui a un site très bien positionné uniquement sur des besoins informationnels reliés au détartrage de machine à café ne fera certainement pas de vente. Le référencement web c'est donc donner de la visibilité à un site dans les moteurs de recherche, mais une visibilité en adéquation avec les contenus et les services proposés par le site. Le référenceur web a donc un métier transversal : il doit à la fois améliorer les classements dans les moteurs de recherche, mais il doit en plus connaitre les sites dont il fait le positionnement, s'immerger dans les problématiques métiers de ses clients. On le devine déjà en filigrane, il aura souvent son mot à dire sur les textes du site.

Dans le monde entier, hormis en Chine et en Russie, Google est le moteur le plus utilisé, et généralement très loin devant tous les autres. En France, on parle de plus de 90% de part de marché. Dès lors, le métier du référenceur web consiste à travailler uniquement ou presque sur Google. Le référencement web, c'est lutter contre Google, ou travailler avec lui, selon le point de vue que l'on adopte.

Finalement, on arrive à une définition aisée : le référencement web c'est la part de visibilité qu'un site a dans Google pour son audience cible. On peut estimer qu'un bon référencement se quantifie par de bonnes positions pour des besoins informationnels correspondant à la cible du site, et par rapport au volume de trafic apporté.

Maintenant que l'on connait précisément l'enjeu théorique, on peut se demander comment les référenceurs web s'y prennent pour améliorer le référencement des sites web dont ils ont la charge.

Du point de vue du moteur de recherche, Google en l'occurrence, il n'y aurait pas de grande complexité : faire des sites utiles à l'utilisateur, agréables et rapides à consulter. C'est une doctrine qui semble un peu grossière, pourtant c'est en ligne directe de l'envie de Google : réussir à comprendre les sites de façon aussi fine qu'un humain. La firme de Mountain View met en place des algorithmes qui ont cette finalité, et voilà pourquoi la recommandation pour obtenir un bon positionnement est d'aller dans le même sens. Seuls les sites intéressants et travaillés devraient mériter les premières places.

Mais quand on se place du côté du webmaster, cette recommandation

fait sourire. Même avec toute la meilleure volonté du monde et l'envie de faire toujours tout pour l'internaute, on sait aussi qu'il y a de nombreux profils d'internautes... et satisfaire tout le monde est difficile. Et il y a des raisons plus pragmatiques, comme un budget plus faible que celui des concurrents, parce qu'on vient d'arriver sur le web. Est-ce qu'on devrait laisser toujours les mêmes en tête de tous les classements ? Ou bien ne vaut-il pas mieux mettre à profit les points forts des sites, les développer, réduire les problèmes faciles à corriger, et voir si les choses ne vont pas s'améliorer ?

Évidemment, et heureusement, en pratique, on peut largement influencer Google en faisant des modifications destinées à nous faire considérer de façon plus satisfaisante.

Le classement du moteur est fait par une agrégation de nombreux signaux, pondérés de façon distincte et dynamique. En clair, il y a de nombreuses zones d'ombres et l'on ne sait pas exactement ce qui a un impact et dans quelles mesures. Voilà pourquoi on croise le fameux « ça dépend » en réponse à de nombreuses questions sur le référencement web. C'est exact, ça dépend. Sur deux sites différents, la même action peut amener un gain différent. Prenons un exemple très simple comme réécrire tous les contenus d'un site pour les améliorer : pour un site qui aurait déjà de très bons contenus, cela peut n'avoir aucun impact. Pour un autre site qui aurait de mauvais contenus, cela peut permettre d'apparaître rapidement dans les premières pages de résultats de Google.

Devant tant d'incertitude, il est prudent de prendre un peu de recul pour observer le fonctionnement général d'un moteur de recherche.

1.3 Fonctionnement d'un moteur de recherche

Le moteur de recherche commence par parcourir le web. Grâce à un « spider », c'est-à-dire une brique logicielle qui parcourt le web et navigue de page en page jusqu'à en connaitre tous les contenus (en pratique, tout n'est jamais connu, mais le moteur n'a pas besoin de tout connaitre pour proposer des pages intéressantes, à partir d'un certain volume on ne risque pas de manquer de ressources). En parcourant le web, il suit les liens, ce qui lui permet d'estimer la popularité de chaque page (s'il y a beaucoup de liens dans le voisinage d'une page, c'est que la page est un nœud important sur le web). Il associe ainsi un score de popularité à chaque page. En parallèle, en parcourant les pages, il récupère les contenus. Il

détermine ainsi de quoi parle chaque page. Quand un internaute se saisit du moteur de recherche pour exprimer un besoin informationnel, il analyse la requête tapée et peut alors constituer une page de résultats :

— Il prend les pages qui répondent au besoin informationnel de l'internaute (chaque page acquiert un score de pertinence par rapport au besoin exprimé) ;
— Pour ces pages, il se souvient du score de popularité attribué ;
— Il agrège les deux indicateurs, avec un coefficient différent pour chaque signal, et classe ainsi des pages populaires et qui répondent au besoin informationnel.

Les résultats sont-ils bons ou non ? C'est l'humain qui va le déterminer. Selon la façon dont il interagit avec les résultats, l'humain va fournir une information sur la qualité perçue au moteur. Une fois qu'assez d'humains ont vu les mêmes résultats ou des résultats pour des requêtes très similaires, le moteur va pouvoir considérer la donnée récupérée comme valable. Si jamais les internautes sont contents, le moteur peut se féliciter. Si jamais les internautes ne sont pas heureux des résultats, le moteur va utiliser un algorithme de *machine learning*, le *learning to rank*, afin de modifier les coefficients utilisés pour le classement. La prochaine fois, un internaute n'aura plus tout à fait le même classement. C'est fantastique : le moteur s'améliore au fur et mesure qu'il est utilisé.

Finalement, une fois passées les analyses initiales, le moteur de recherche est surtout et avant tout une machine à classer des pages, à les comparer grâce à de nombreux signaux. C'est un point essentiel du référencement web : au départ, c'est un marathon, il faut être dans le peloton, faire le minimum vital pour être considéré par le moteur, ensuite sur la ligne d'arrivée, il faut être le meilleur sur les critères importants. C'est à retenir absolument : il faut trouver les critères importants pour les besoins informationnels que l'on souhaite travailler, et il faut les amener à des niveaux supérieurs à la concurrence.

Le schéma général de fonctionnement, visible dans la figure 1.1, illustre assez bien le consensus qui existe chez les référenceurs experts. Pour améliorer le référencement d'un site web, il y a trois piliers principaux et incontournables : le contenu (analyse des pages et des requêtes), la technique (le site est-il explorable de façon satisfaisante et les contenus sont-ils accessibles) et la popularité (le site est-il reconnu).

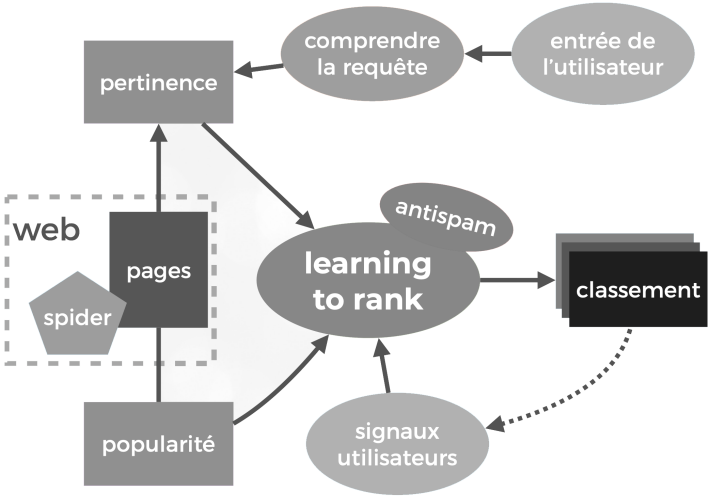


FIGURE 1.1 – Le schéma de principe d'un moteur de recherche

1.3.1 L'analyse des contenus

Le contenu, c'est ce que l'on trouve sur les pages du site : le design, les images, les sons, mais surtout les textes. Ce sont eux que le moteur de recherche sait le mieux analyser. Ce sont ces derniers qui sont lus par les internautes, c'est là où ils trouvent majoritairement les réponses à leur besoin informationnel. Le contenu est essentiel : sans contenu, une page est vide, elle ne porte aucune information, elle n'a aucun intérêt. Avec du bon contenu, une page est pertinente ; sans contenu ou avec du mauvais contenu, l'internaute qui la rencontre est déçu.

Le moteur de recherche souhaite donner de bons résultats à ses utilisateurs afin qu'ils soient satisfaits et qu'ils continuent à l'utiliser. La priorité est claire pour le moteur : il faut se focaliser sur l'intérêt, la pertinence des pages. Les contenus pauvres sont à écarter au plus vite. Pour le référenceur web, le bon contenu doit toujours primer, car c'est un prérequis pour faire un bon travail de référencement. Il demandera alors souvent à modifier, améliorer voire refondre les contenus des pages importantes du site à positionner. Ce sera souvent un chantier long et difficile, car un bon contenu n'est pas seulement un bon contenu pour l'humain. C'est un

contenu qui est à la fois intéressant pour l'humain et bon pour le moteur de recherche. Parfois un contenu vu comme parfait pas les équipes du site devra être retravaillé pour être vu comme parfait par Google. Diplomatie et pédagogie sont des maitres mots du référencement web.

En pratique, le référenceur utilise à son profit des outils tels que `yourtext.guru` afin de facilement extraire le vocabulaire important relié à un besoin informationnel. Il peut ainsi faire mieux que la concurrence en étant guidé.

Le contenu est un pilier à ne jamais négliger : il est déjà important à l'heure actuelle et il le devient de plus en plus avec les améliorations des algorithmes d'intelligence artificielle de Google. L'ambition du moteur est de toujours répondre correctement au besoin informationnel, et en comprenant de mieux en mieux le contenu il devient capable de s'affranchir davantage des autres critères de positionnement. La réponse directe aux questionnements des internautes, sans passer par des sites tiers, reste une motivation importante chez Google, ne l'oublions pas… Les récentes avancées en intelligence artificielle (IA) permettent de faire les questions et les réponses, de générer des textes, des images, de comprendre ce qui est affiché sur une photo, de détecter les textes générés par des modèles connus de génération de texte, pourtant très bien écrit. Il est plus que jamais temps de s'y intéresser.

1.3.2 La technique

Ensuite, la technique est un pilier relié à la façon dont est conçu le site. Le référenceur s'assure que le site n'a pas de problèmes qui empêcheraient le moteur de recherche de le lire correctement. On a beau avoir de beaux contenus, si jamais le moteur ne peut pas les consulter, ils ne serviront à rien. C'est une brique qui est souvent travaillée en collaboration avec les équipes techniques du site. Le référenceur web mettra son nez dans tous les recoins et ne remontera généralement que les choses à corriger. On parle souvent d'« erreurs » techniques, ce qui sera vu comme étant autant de blâmes remontés aux équipes techniques. Encore une fois, pédagogie et diplomatie seront des qualités.

Le pilier technique est généralement associé à des listes de critères à vérifier. Il y a un côté facile à détecter la grande majorité des problèmes (est-ce que les balises meta title sont remplies, le sont-elles de façon unique, etc.), tandis que la correction n'est pas toujours si évidente : le contexte du projet aidera souvent à comprendre pourquoi des anomalies

sont en ligne alors qu'elles sont connues.

En dehors des anomalies, le référenceur web proposera souvent des améliorations. Par exemple, Google fait la promotion depuis des dizaines d'années de la performance web afin d'obtenir le temps de chargement le plus réduit possible pour les pages web. C'est un serpent de mer qui revient sous plusieurs formes. Les communications récentes mettaient en avant les *core web vitals*, c'est-à-dire des métriques qui sont à cheval entre la vitesse de chargement et l'expérience utilisateur. C'est dans l'intérêt de Google que les sites web soient rapides à charger. Qui dit chargement efficace dit récupération des contenus plus rapides. C'est-à-dire qu'avec la même infrastructure technique on peut récupérer plus de pages web. Pour un moteur, c'est une économie substantielle. Dans ce cadre, le référenceur web indiquera de suivre les préconisations de Google afin de s'assurer d'être dans ses bonnes grâces. Il n'y a pas toujours de données qui permettent d'affirmer que les améliorations seront efficaces, mais c'est une bonne pratique générale de suivre les recommandations de Google à la lettre. C'est quand on s'en écarte qu'il faut pouvoir justifier des raisons de l'écart volontaire.

1.3.3 La popularité

Nous arrivons maintenant au dernier des trois piliers principaux, celui de la popularité. Au début de l'ère des moteurs de recherche, beaucoup de moteurs étaient capables de retourner des résultats corrects quand ils s'intéressaient à des pages de sites très bien identifiés et auxquels on pouvait faire confiance. Par exemple, faire un moteur de recherche à l'interne d'une entreprise est généralement aisé. Quand personne ne triche et que tout le monde joue avec les mêmes règles, le moteur doit trier, mais personne ne cherche à l'influencer. Quand les vannes s'ouvrent à l'ensemble du web et que des milliers de webmasters essaient d'être tous en même temps premiers sur les pages des résultats de recherche, le problème surgit. Tout le monde s'agite, fait un peu n'importe quoi, de façon terriblement efficace pour certains. Les résultats de recherche se modifient substantiellement : pendant que certains sites de référence ne connaissent pas le référencement web et ne font rien comme il faudrait le faire, par ignorance, de petits acteurs lancent les grandes manoeuvres, avec un fort succès. C'est l'aspect concurrentiel qui entraîne le « mauvais comportement », qui du point de vue du référenceur n'est que la nécessité d'obtenir de la visibilité (après tout, il n'y a que dix pages sur la page de

résultats de Google). Les annuaires de sites qui faisaient des sélections à la main n'avaient pas ce genre de problème. Pour eux, c'était le volume le seul souci.

Volontairement ou non, Larry Page, chez Google, trouve une réponse parfaite à ce défi. Il crée l'algorithme du *pagerank*, qui simule le comportement des internautes afin de savoir quelles sont les pages qu'ils aiment et donc les pages qui méritent d'être trouvées facilement dans le moteur de recherche.

Pour mieux visualiser l'algorithme du *pagerank*, imaginez une carte de France, zoomez autour de Paris et Lyon. Voyez comme les routes sont nombreuses autour de ces deux villes : il y a tellement de personnes qui habitent autour de ces villes qu'il faut beaucoup de routes pour leur permettre d'y accéder. Voyez les grands axes qui existent entre elles. Il y a beaucoup de trafic allant d'une ville à l'autre. Maintenant, remplacez les villes par des pages web, et les routes par des liens. Et bien sûr, ajoutez des humains qui empruntent les liens, qui consultent le web, sautent de pages en pages. À tout moment, vous pouvez faire une pause et voir où sont les humains. Plus ils sont nombreux à un endroit, plus cet endroit est populaire.

Grâce au *pagerank*, par le biais d'une simulation de l'internaute, on récupère la sélection humaine que faisaient les contributeurs aux annuaires de sites. En croisant cette donnée à la réponse apportée par les pages au besoin informationnel de l'internaute, Google peut proposer des résultats meilleurs que ses concurrents.

Depuis longtemps, des algorithmes comparables au *pagerank* sont implémentés dans tous les moteurs de recherche. Ce n'est donc plus tellement un avantage concurrentiel, mais plutôt un prérequis.

En référencement web, pour améliorer sa popularité, on fait levier sur l'algorithme du *pagerank*. Pour cela, on fait apparaître des routes vers les pages dont on souhaite améliorer la popularité. Ces routes, ce sont des liens que l'on récupère parce que quelqu'un sur le web pense que notre page est intéressante et pointe donc vers elle depuis sa propre page. Dans les faits, depuis quelques années, les liens naturels sont rares, surtout dans des domaines concurrentiels. Le référenceur web est souvent amené à faire de l'*inbound marketing* (voir le chapitre 4) ou de l'achat de liens pour pouvoir obtenir de nouvelles routes vers une page.

Cependant, le levier de popularité est tout à fait d'actualité. Il est même plutôt facile à actionner : les vendeurs de liens sont très nombreux, notamment en France, des plateformes spécialisées existent par dizaines.

Le plus difficile pour le référenceur web est de faire accepter un budget et de choisir les meilleurs liens disponibles selon ses objectifs.

Pour cela, le référenceur web utilise des outils d'analyse du web et des liens comme `babbar.tech`.

En faisant progresser chacun des trois piliers, le référenceur web parvient à améliorer le référencement web des sites dont il a la charge.

Le référencement web est une discipline à première vue simple : faire des liens, faire de bons contenus, s'assurer que le site est développé proprement. Derrière ces grandes formules, il y a beaucoup de travail d'artisan. Faire les choses proprement, avec qualité, pour que les pages produites ou améliorées soient vues comme autant de bijoux qu'on exposerait à Google.

2. Comment rédiger un bon contenu pour le web ?

Baptiste Guiraud est rédacteur web depuis plus de six ans. Il est passionné de web en général et est également éditeur de sites. Il lance actuellement une formation orientée rédaction web pour le SEO, en capitalisant sur son expérience et son expertise du sujet.

2.1 Introduction

L'écriture web est une discipline particulière. Elle demande une bonne maîtrise de la langue, une bonne compréhension de son lectorat et de ses attentes, ainsi que la capacité à délivrer la bonne information au bon moment.

Écrire pour le web ne demande donc pas simplement d'avoir une belle plume. Certes, bien rédiger, clairement et en tenant son lecteur en haleine est un gros avantage, mais ne suffit malheureusement pas pour rendre son contenu aussi visible qu'intéressant. Un ensemble de bonnes pratiques, tant d'un point de vue structurel qu'informationnel, seront attendues et nécessaires afin de permettre à une page de « performer » en ligne.

Il ne faut pas oublier que l'internaute ne sera pas seul juge de la qualité d'un contenu. Les algorithmes utilisés par les moteurs de recherche l'évalueront également et lui octroieront un classement « logique » selon leurs critères. En d'autres termes, si un contenu ne répond pas aux attentes

des algorithmes de classement, il n'obtiendra pas la visibilité escomptée, dans la mesure où il n'apparaîtra que très loin sur les pages de résultats.

Voici donc la dichotomie avec laquelle le rédacteur web doit composer au quotidien :

— Rédiger un contenu pertinent et de qualité, répondant aux attentes des internautes.
— Rédiger un contenu suffisamment travaillé pour qu'il obtienne les faveurs des moteurs de recherche.

Il n'y a pas de secret ni de mystère pour voir ses pages performer, simplement un ensemble de bonnes pratiques et d'astuces qui permettront de satisfaire les attentes des lecteurs, tout en bénéficiant d'un bon positionnement sur les moteurs de recherches.

Voici ces bonnes pratiques détaillées.

2.2 La structure du contenu et son importance

Un des premiers points qu'il est important d'avoir en tête, c'est la structure du contenu. Elle doit être à la fois claire, détaillée et logique.

La structure est aussi importante pour les moteurs de recherche que pour les internautes :

— En survolant les titres et sous-titres, le lecteur pourra se faire une idée de la teneur du contenu.
— Les moteurs de recherche vont évaluer ces titres et sous-titres pour s'assurer du contexte dans lequel est déployé le contenu.

Il est donc primordial d'accorder du temps à la construction du plan de chaque page, puis à la rédaction de ses titres et sous-titres qui jalonneront sa structure. Un exemple de structure de texte est présenté en figure 2.1.

Une bonne structure d'article facilite sa lecture et permet de travailler son contenu en détail, en contextualisant parfaitement chaque paragraphe.

Il faut structurer vos contenus à l'aide des balises Hn. Il n'est pas forcément utile de tout utiliser (H3/H4 maxi recommandé), mais il est important de doter chaque page d'une arborescence claire et facile à comprendre.

On reste logique : après un H2, soit on insère un H2, soit un H3 pour détailler. Mais on ne passe pas d'un H1 à un H3 par exemple, jamais ! On déroule une structure logique et ordonnée.

Apprendre à peindre (H1)
Introduction
Les différentes techniques de peinture (H2)
Paragraphe
Peindre au pinceau (H3)
Paragraphe
Peindre au couteau (H3)
Paragraphe
Les différents types de peinture (H2)
Paragraphe
La peinture acrylique (H3)
Paragraphe
La gouache (H3)
...

FIGURE 2.1 – Exemple de structure de texte

2.2.1 Titres et mots-clés

Il est vrai qu'insérer des mots-clés dans les titres et les sous-titres permet souvent de bénéficier d'un avantage par rapport aux sites concurrents qui ne le font pas.

D'ailleurs, plus le mot-clé principal est au début du titre, meilleur est le bénéfice. Attention, ce n'est pas une raison pour commencer tous vos articles de la même façon. Il faut savoir composer entre écriture originale et optimisation envisageable. Même si les titres du type « Apprendre à peindre : tout ce qu'il faut savoir » marchent bien, il y a moyen de trouver une formulation un peu plus originale.

« Cours de peinture : les astuces pour devenir le prochain Van Gogh » : Là, on est dans le « putaclic »… « Astuce » étant un mot souvent employé pour attirer les clics, on considère ces titres comme étant un peu trop vus et revus. Néanmoins, certains ne jurent que par ce genre d'intitulé.

« Découvrir une technique de peinture par semaine » : Ce titre est déjà un peu plus raisonnable et promet de détailler un peu le sujet. Même si « technique de peinture » est placé au milieu du titre, on est face à une formulation qui intéressera certainement les amateurs.

« Apprendre à réaliser les toiles de ses rêves » : Là, le titre est bien écrit,

mais le mot-clé est absent. Pour autant, si le reste de l'article est bien rédigé, le contenu peut parfaitement faire le job.

« Devenez un as du pinceau grâce à cette formation » : Là, on reste sur du « putaclic », mais l'optimisation est bien moins évidente. Est-ce que ce titre convaincra ? Je n'en suis pas sûr.

Vous comprendrez donc qu'il faut souvent jouer entre l'optimisation des titres et leur originalité.

En ce qui concerne les sous-titres, la mécanique est la même. Le seul point important est d'écrire un sous-titre clair, annonçant le contenu ou le ton du paragraphe qui suivra. Là encore, plus l'ensemble est cohérent du point de vue thématique, plus il sera facile de profiter de la structure du contenu pour y insérer des liens bien thématiques.

Le lecteur pourra également évaluer le contenu en survolant l'ensemble. Des sous-titres pertinents permettent de retenir l'attention des lecteurs rapides. Ils posent plus facilement leur regard sur un paragraphe dont l'intitulé leur parle. Il faut donc en profiter pour optimiser ce point.

Dans tous les cas, une bonne structure permet de faciliter la lecture. Donc on ne se gênera pas pour en profiter !

Si vous créez votre propre site, ajoutez un plugin qui génère une table des matières pour chaque page et chaque article. Le lecteur pourra apprécier la teneur des pages en un coup d'œil. A l'échelle de Google, c'est également un signal positif, dans la mesure où on lui mâche le travail pour saisir la structure du contenu. Et Google adore qu'on lui facilite la tâche !

2.2.2 Les paragraphes

Pour ce qui est de la rédaction des paragraphes, il est moins question de mécanique que de logique et d'expérience utilisateur.

Il n'y a pas de règle quant à la longueur d'un paragraphe. Tant que ce dernier traite bien le sujet prévu, et que les informations livrées sont pertinentes, pourquoi se limiter en longueur ? Simplement pour ne pas perdre le lecteur. Il faut garder à l'esprit que l'attention des internautes est relativement volatile. Ces derniers n'accordent leur attention qu'à condition de répondre à leurs attentes. De plus, la mécanique de lecture en ligne est différente de celle des supports papier. La taille des écrans en est la principale cause.

Concrètement, il faut faciliter la lecture de son contenu. Il ne faut pas hésiter à aérer sa rédaction et à faire en sorte que le rendu visuel ne soit

pas trop compact. Ce point participe à l'UX (l'expérience utilisateur). Le travail relatif à l'UX vise notamment à rendre la navigation aussi fluide et intuitive que possible. Aussi, la lecture d'un contenu peut s'avérer nettement plus simple si l'ergonomie est étudiée.

En ce qui concerne la longueur, il ne faut surtout pas se fier aux indicateurs de certains plugins (Yoast, Rankmath, etc.). Là où ces outils parlent de « 300 mots max par paragraphe », libre à vous d'adapter la longueur de vos parties en fonction du sujet et des informations que vous entendez délivrer.

Toutefois, entrecouper un très long paragraphe (plus de 500 mots par exemple) à l'aide d'un sous-titre peut être utile. Tant en termes de facilité de lecture que de structure du contenu, l'utilisation de sous-titres ou sous-parties permet d'alléger certains passages.

2.2.3 Utilisation du gras et de l'italique pour apporter des nuances au contenu

De même que la longueur des paragraphes et la disposition du contenu, le gras et l'italique permettent d'améliorer l'UX de l'internaute. Notez qu'il ne s'agit absolument pas de critères SEO, mais plutôt d'un moyen d'optimiser l'UX et de faciliter la lecture de certaines informations. Le but n'étant pas de surligner tout le texte, mais plutôt d'appuyer certains passages **pour attirer l'attention du lecteur**.

On va donc utiliser le gras avec modération et faire en sorte de le dédier à quelques points clés du contenu. Attention, ne surlignez pas chaque occurrence de mot-clé. Veillez simplement à mettre en valeur quelques points qui vous semblent importants dans le contenu rédigé.

L'italique a exactement la même utilité, à savoir, *mettre certains passages en valeur* ou en perspective. Une phrase en italique retient l'attention. De plus, en revenant à la ligne et en isolant un passage en italique du reste d'un paragraphe, on obtient un effet intéressant, permettant de sortir une phrase du contexte. On peut ainsi appuyer sur certaines notions importantes ou reformuler succinctement un paragraphe pour insister sur un point très précis. Ce n'est pas du SEO, c'est du bon sens et de l'UX.

2.2.4 Les listes à puces

La liste à puces est à considérer comme un outil. Certes, elle facilite la lecture de données précises, que l'on détaille facilement, mais elle permet également de s'attirer les faveurs de Google. Comme dit précédemment,

Google adore qu'on lui facilite la tâche. Or, une liste à puces est souvent une suite de faits ou de données facilement assimilables par le moteur de recherche, puis facilement traitée. D'ailleurs, de nombreuses positions 0 (P0) affichent une liste à puces. C'est la preuve que Google apprécie les données déjà mises en forme et structurées.

On profite d'une liste à puces pour utiliser du jargon métier, ou des mots importants pour le sujet à traiter. Toujours dans la veine de notre thématique « peinture », voici un exemple de liste à puces logique.

Tous les accessoires pour peintre sur toile correctement :
— Un chevalet.
— Une palette à peinture.
— Des pinceaux et des brosses (poils souples et poils raides).
— Un peu d'eau (pour diluer la peinture).
— Peinture de son choix (gouache, acrylique, à l'huile, à vous de voir).

On voit que l'on peut facilement utiliser du vocabulaire pertinent, sans devoir construire de vraies phrases. Dans certains cas, la liste à puces permet de mettre en avant des spécificités d'un produit par exemple, en n'incluant que des données chiffrées (poids, hauteur, largeur, etc.). De plus, les listes à puces facilitent l'assimilation de l'information par le lecteur. Il a sous les yeux les informations importantes concernant le sujet qui l'intéresse. En plus d'aérer et de structurer le contenu, la liste à puces permet donc de mettre en évidence des données précises et importantes.

Si on ne peut pas parler de vrai impact SEO, on peut néanmoins être sûr de plus facilement intéresser et retenir l'attention de l'internaute.

2.2.5 La méta description

Bien qu'elle n'ait pas de poids en SEO, la méta description reste utile, surtout pour l'internaute. La méta description apparaît sur la SERP, juste en dessous de la balise Title de la page affichée. Une bonne méta description, bien pensée, peut attirer l'internaute. Il faut qu'elle suscite l'intérêt et soit engageante. Si tel est le cas, il y a des chances pour qu'elle ait un impact positif sur le comportement des internautes qui voient votre site passer dans les pages de résultats.

Au-delà de résumer le contenu de votre page, faites en sorte d'éveiller la curiosité ou de susciter l'intérêt.

Exemple : *Arriver à peindre une toile n'est pas qu'une question de talent. Apprenez toutes les techniques utiles pour vous aussi vous adonner à la peinture simplement.*

La personne qui souhaite essayer de peindre voudra en savoir plus et sera tentée de visiter la page en question.

La méta description peut contenir jusqu'à près de 300 caractères, mais au-delà de 150 caractères, Google peut décider de n'en afficher qu'une partie, voire de la réécrire. On recommande donc de se limiter à environ 150-160 caractères. C'est un élément à ne pas sous-estimer et dont on soignera la formulation. En termes de trafic, la méta description peut apporter un véritable plus. Elle peut décider l'internaute à visiter une page. Elle contribue souvent à améliorer le CTR (*Click Through Rate*) d'une page.

2.2.6 L'importance de la structure du contenu

Pour clore le chapitre dédié à la structure du contenu, je tiens à insister sur l'importance qu'a l'architecture de chacun de vos textes. Au-delà du SEO, il y a une dimension UX à ne pas négliger. Certes, des textes bien faits et bien optimisés permettront certainement d'obtenir un meilleur classement, mais il ne faut jamais oublier que l'on écrit avant tout pour des humains.

Même dans le cadre de la rédaction de « textes SEO », qui serviront surtout à faire gagner des positions à une page, il est toujours rentable d'écrire correctement, sans simplement chercher à répéter un mot-clé ou une expression sur laquelle la page doit se positionner.

Les rares internautes qui liront ce texte de bas de page en garderont une impression mitigée, voire négative. Cela vaut vraiment la peine d'inclure quelques informations vraiment intéressantes à ce type de contenu. Même si les personnes qui le liront ne représenteront qu'une infime minorité du lectorat global, elles pourraient apprécier la petite information supplémentaire et passer à l'acte (achat, téléchargement, remplir un formulaire, etc.).

2.3 L'optimisation du contenu

Voilà une spécificité de l'écriture pour le web : l'optimisation sémantique. Grâce à un outil tel que `Yourtext.guru`, on va pouvoir mesurer l'effort fait par la concurrence en termes d'optimisation sémantique, et ainsi adapter son contenu pour lui permettre de performer. Il suffit d'écrire naturellement puis de tester son contenu dans l'outil pour se rendre compte que « l'écriture naturelle » ne permet généralement pas d'obtenir un score

intéressant.

En revanche, il faut garder en tête que ce n'est pas parce qu'on optimisera un texte en prêtant attention au moindre détail, que ce dernier sera forcément premier sur le mot-clé visé. En revanche, si ledit texte est publié sur un site en bonne santé et populaire, alors le résultat de l'optimisation devrait permettre d'obtenir un bon classement.

Si l'optimisation sémantique ne fait pas tout, elle contribue grandement à faire apparaître ses pages en bonnes positions dans les SERPs. De plus, elle permet de travailler la cohérence sémantique au sein d'un même site. En clair, pour performer en ligne, il est quasi impossible d'ignorer l'optimisation sémantique.

2.3.1 Comment optimiser un contenu ?

Avant tout, il faut avoir en tête que les pages soumises aux moteurs de recherche vont être évaluées. Des algorithmes dédiés à cette évaluation vont permettre d'octroyer un classement à une page, en prenant en compte de nombreux critères. Ces algorithmes se basent sur le contenu existant et indexé, sur ses résultats et ses performances, avant d'établir des critères attendus dans chaque thématique. Il faut donc que votre contenu réponde à certains points attendus par les moteurs de recherche.

2.3.2 Concrètement, comment cela se traduit-il ?

Il y a peu de chance pour que vous soyez le premier à traiter une thématique au travers de votre contenu. Cela signifie donc que d'autres pages sont déjà indexées et classées par Google sur la requête que vous visez.

Il « connaît » déjà la thématique. Il sait que tel ou tel mot sera certainement présent, qu'il sera à proximité de tel autre mot et qu'il reviendra certainement X fois. Il existe donc déjà un cadre qui servira de référence à Google pour juger de la pertinence et de la qualité de votre contenu. L'optimisation sémantique consiste à exploiter les critères que Google privilégie pour une thématique donnée. Il s'attend à une certaine densité de mots relatifs au métier. Pour optimiser un texte, il va ainsi falloir jouer avec la densité des lexies attendues. Ces lexies auront plus ou moins de poids.

Imaginons que nous souhaitions positionner une page sur le mot-clé « maçon ». Des mots comme « construction », « maison », « mur » ou « béton » vont forcément revenir dans la mesure où ils font partie du jargon

métier. Autrement dit, essayer de positionner une page sur le mot-clé « maçon » sera impossible sans l'utilisation de ces lexies. Qui plus est, il sera certainement nécessaire d'utiliser certains mots plusieurs fois, parfois en les associant à d'autres termes importants.

Comme expliqué dans la partie dédiée aux titres, l'insertion de mots importants dans les titres et sous-titres est une première étape. On contextualise ainsi plus facilement les paragraphes. Mais les paragraphes ne vont pas déroger à la règle ! On va donc faire en sorte d'utiliser le « jargon » propre au métier ou à la thématique.

Contrairement aux idées reçues, ce n'est pas la beauté du texte ni les formules littéraires employées qui permettront à un contenu d'être perçu comme meilleur par les moteurs de recherche. On constate même souvent le contraire en analysant de grandes quantités de pages. Plus un contenu est spécifique, plus il a de chances d'être bien classé. Les moteurs de recherche ne prêtent aucune importance aux figures de style ni aux beaux mots. Ils jugent un contenu sur la base de critères précis, d'ordre plus mécanique que littéraire. De fait, un langage très spécifique à un sujet permet d'éviter trop de « dilution » au niveau de l'information. Attention, on ne perd évidemment jamais de vue que la page s'adressera à des humains. Il faudra donc donner corps à un texte à la fois précis, agréable à lire, vraiment axé sur le sujet traité.

2.4 Yourtext.guru : présentation

Yourtext.guru est un outil d'assistance à l'écriture. Il permet de faire une lecture pertinente d'une SERP, tout en permettant de donner à son contenu « le poids » suffisant pour être visible et bien référencé. Si au départ cet outil représente une contrainte en matière d'écriture, il devient vite un assistant indispensable, qui permet d'améliorer l'analyse des pages web, ainsi que leur production. A force de l'utiliser, il devient difficile de s'en passer dans la mesure où l'outil fournit de précieuses informations sur la façon dont il faudra traiter chaque sujet souhaité.

2.4.1 La création d'un guide d'écriture

En renseignant une requête, un titre, un mot-clé ou une phrase, on peut générer un guide d'écriture correspondant. Ce guide va permettre d'analyser la concurrence (les 10 premiers résultats), ainsi que de connaître quels sont les mots importants attendus par Google. D'ailleurs, Yourte

xt.guru ne se contente pas de livrer une simple liste de mots-clés, il nous donne également des combinaisons de 2 et 3 mots qui compteront à l'heure de rédiger un contenu exhaustif.

L'autre partie du guide d'écriture comporte un outil d'analyse de son propre contenu. Ce dernier rend l'optimisation sémantique graphique. Une courbe se dessine au fur et à mesure de l'avancée de la rédaction et des vérifications de score. Cette courbe grimpe au fur et à mesure que les termes importants sont ajoutés au texte. Vous vous assurez d'inclure tous les mots importants, et vous pouvez également analyser le contenu de vos concurrents. Au-delà de la courbe, des scores sont attribués au contenu : Le SOSEO (optimisation sémantique) et le DSEO (suroptimisation).

L'un et l'autre sont à prendre en compte. Il suffit de jeter un œil en dessous du champ de rédaction, sur les 10 premiers résultats de la SERP, pour voir que ces scores sont importants. Ces scores reflètent les efforts faits par les concurrents. Plus le contenu est dense et optimisé, plus son score sera important.

Petite astuce : *focalisez-vous sur les 3 premiers résultats de la SERP que vous visez. Google ne les a pas mis en tête par hasard.*

2.4.2 Le score BABBAR et les positions des pages

Yourtext.guru offre désormais deux indicateurs supplémentaires à ceux concernant l'optimisation du contenu :

— Le score BABBAR du domaine où une page est publiée.

— La quantité de mots clés sur lesquels une page apparaît.

A eux deux, ces indicateurs nous renseignent sur la puissance d'un concurrent (la notoriété de son site obtenue à l'aide de liens) ainsi que sur les positions obtenues par son contenu. Cela permet d'évaluer l'effort à faire pour arriver à rivaliser avec son propre site.

Concrètement, si les concurrents qui arrivent en tête ont à la fois un domaine puissant et un contenu très dense et optimisé, obtenir un bon classement demandera beaucoup d'effort, tant en matière de rédaction, que de *netlinking*.

En revanche, une SERP où le score BABBAR de vos concurrents est plus bas que sur votre propre site et leur score d'optimisation moyen, il y a alors matière à se positionner facilement sur la requête en question.

« On ne peut améliorer que ce que l'on peut mesurer ». C'est exactement ce que permet Yourtext.guru : une lecture détaillée d'une SERP et du travail effectué par les concurrents.

2.4.3 L'analyse d'intention

Cette fonction permet de découvrir comment Google considère ou classe une requête. Selon l'intention rencontrée chez vos concurrents, vous pourrez adapter la nature et la production de votre contenu. Comme chaque indicateur, l'analyse d'intention permet d'affiner le travail en s'assurant de servir à Google une page au bon format, respectant ses attentes concernant la thématique traitée et ainsi susceptible d'être bien classée.

Le but est de s'assurer de choisir la bonne orientation et de ne pas produire une page qui dénoterait trop du top 10 de la requête visée.

A noter : *ce point rejoint celui de la structure du contenu. On s'applique à donner à Google des pages au format qu'il considère adéquat pour répondre à une requête. Dans ce cadre, on comprend que l'originalité et la spécificité d'une marque ou d'un site doivent s'appuyer sur des critères « communs » à toutes les pages concurrentes. Un contenu trop original, tant sur le format que sur sa teneur, a peu de chance de performer, à moins de devenir viral, indépendamment des performances SEO. En ce sens, l'écriture web est toujours sous contrainte.* Yourtext.guru *permet justement de tirer parti de ces contraintes, en créant un contenu ciselé pour se classer.*

2.4.4 Quelques astuces sur Yourtext.guru

Comme tout outil, Yourtext.guru ne fait que ce qu'on lui demande.

Première astuce Yourtext.guru **: N'hésitez pas à créer plusieurs guides avec des variations.**

Lorsque vous souhaitez rédiger une page, n'hésitez pas à générer un guide avec la requête stricte, puis au pluriel, puis sans mot de liaison.

Pourquoi ?

Simplement pour vous assurer que les 10 premiers résultats présentés sont toujours les mêmes. Si ce n'est pas le cas, il y a peut-être quelque chose à faire, notamment s'intéresser aux différences remontées sur ces SERP.

Deuxième astuce Yourtext.guru **: Internautes, les questions qu'ils se posent**

Ce n'est un secret pour personne, les PAA (*people also ask*) sont toujours intéressantes à traiter. Au lieu de créer un guide de rédaction dédié, ajoutez simplement ces PAA dans votre texte en y répondant. La création de sous-partie (H3 par ex.) répondant directement à l'une de ces questions peut représenter un vrai bénéfice pour votre contenu.

Troisième astuce `Yourtext.guru` : n'ayez pas peur du score DSEO

Le « D » induit la notion de danger, mais en réalité, il ne s'agit que d'un indicateur. Ne craignez pas la suroptimisation si vos concurrents s'y sont adonnés. En sachant que `Yourtext.guru` analyse une bonne partie de la SERP, il se peut que les 10 premiers aient un contenu bien plus dense que les autres.

Dans ce cas, le DSEO devient un enjeu intéressant à exploiter. A manipuler avec précaution toutefois.

2.5 Quelques astuces pour créer un contenu aussi pertinent que performant

Il n'y a pas de truc secret pour obtenir un bon contenu, simplement quelques petites astuces qui permettent de gagner en pertinence et en temps de recherche avant d'écrire.

2.5.1 Si vous maîtrisez le sujet que vous traitez

Il est parfois utile de prendre un peu de distance sur un thème que l'on connaît bien, simplement pour ne pas trop tomber dans les lieux communs.

Prendre de la distance se traduit notamment par visiter des forums spécialisés où les potentiels clients exposent leurs problèmes ou leurs craintes.

Ces informations sont très utiles. Lorsqu'on connaît les craintes et les frictions que rencontrent certains clients face à un produit, une solution ou quoi que ce soit que vous proposiez. On arrive plus facilement à jouer sur la corde sensible dans l'écriture. On rassure, on éclaircit, on explique de façon à rassurer, donc potentiellement convertir.

Les forums regorgent d'informations précieuses (encore faut-il trouver les bons).

2.5.2 Si vous ne maîtrisez pas la thématique traitée

Commencez par la SERP. En clair, épluchez les 10 premiers résultats et notez les points qui reviennent dans toutes les pages.

Pour étoffer vos connaissances sur le sujet traité, là encore, les forums sont utiles. Ils permettent de découvrir le jargon thématique, la façon qu'ont les internautes de parler de leur problématique. N'hésitez pas à chercher sur d'autres supports que sur les SERP. Youtube regorge de vidéos de passionnés qui donnent beaucoup d'informations utiles. Certains

sujets sont également traités sous forme de podcast. L'alimentation par exemple est une thématique sur laquelle on peut se renseigner de mille façons.

Ne vous contentez pas de chercher des informations seulement sur la SERP sur laquelle vous souhaitez positionner votre page. En revanche, commencez par cette SERP pour savoir quel type de contenu produire, quelles informations devront absolument y être et comment articuler le contenu. Pour le reste des informations, les autres sources vous permettront d'enrichir votre page.

2.5.3 N'hésitez pas à contacter un professionnel si vous le pouvez

Toujours dans l'optique de s'assurer d'utiliser le bon jargon ou de regarder le sujet sous le bon angle, il est parfois utile de contacter directement un professionnel et de le questionner. En expliquant clairement ce que vous recherchez comme information, vous serez probablement surpris de la quantité de gens disposés à vous accorder quelques minutes pour vous éclairer.

Si vous ne connaissez rien à la peinture en bâtiment par exemple, mais que vous êtes tenu d'écrire une page sur le sujet, après avoir lu en ligne, sur des sites et des forums, regardez quelques tutoriels sur Youtube et si quelques questions persistent, listez-les puis contactez un professionnel. Un appel de trois minutes vous donnera souvent plus d'informations que la lecture de beaucoup de sites web.

2.5.4 Ecrivez, puis dégraissez

Un autre point important concernant l'écriture, c'est la notion de *less is more*. Dès lors que vous terminez l'écriture d'un texte, repassez-le en entier, puis enlevez des mots. Enlevez autant de mots que possible, tout en conservant un texte agréable à lire. En enlevant des adjectifs, des transitions (parfois inutiles), des adverbes ou tout simplement en contractant quelques phrases, on obtient un contenu plus dense d'un point de vue sémantique. Sans altérer le score d'optimisation de `Yourtext.guru`, essayez de dégraisser votre texte.

Ce type de finition permet souvent d'obtenir de meilleurs résultats en termes de classement.

2.5.5 Construisez-vous des *buyers personas*

Un *buyer persona* est un « profil de client type ». Il reflète le lectorat ou la clientèle d'un site ou d'une marque.

En ayant votre *buyer persona* en tête, vous vous adresserez plus facilement à lui. Selon ses centres d'intérêts, ses motivations dans la vie et sa classe socio-professionnelle par exemple, vous aurez davantage d'accroches pour lui parler directement.

Que vous interveniez en tant que prestataire sur un projet, ou que vous lanciez votre propre site, essayez de faire un portrait-robot de votre futur lecteur. Votre écriture n'en sera que plus efficace. Elle touchera sa cible et déclenchera l'action (abonnement, achat, etc.).

2.6 Conclusion

L'écriture pour le web demande la maîtrise de pas mal de points clés pour arriver à produire un contenu à la fois plaisant, engageant voire vendeur et doté d'un bon potentiel de classement.

Si au départ tous les critères à prendre en compte sont des entraves, des contraintes, elles doivent devenir des guides ou des rails. Au fil de la pratique, des templates de pages sont déjà prêts dans votre tête. Le plan d'un article ou d'une page se dessine au fil de la documentation et la rédaction des paragraphes se fait simplement, assisté de `Yourtext.guru` pour s'assurer de coller aux attentes de Google.

En travaillant ainsi, vos contenus seront visibles, lus et même partagés. Il n'y a plus qu'à vous retrousser les manches !

3. Au commencement, la marque était le *storytelling*...

Camille Gillet est animée par une volonté de jouer les troubles fêtes irrévérencieux, en créant essentiellement du contenu à partager. Pour cela elle marie ses trois plaisirs que sont l'écriture, la pédagogie et le Web en orientant l'activité autour du *storytelling* (accompagnement et formation). Parmi ses réussites, on trouve deux livres, la création des LinkeFictions qui sont des micro-fictions satiriques autour de la *Startup Nation* et le podcast "Le héraut aux mille visages" sur le *storytelling*.

3.1 ... et le *storytelling* était la marque

Quand Han Solo se met à charger en hurlant dans les couloirs de l'étoile noire dans *Star Wars*, il n'a aucun plan en tête. La Princesse Leia soupire un « quel courage ! », le spectateur sourit devant l'audace, mais Luke Skywalker lève les yeux au ciel... Et pour cause ! Si certains prétendent que la meilleure défense serait l'attaque, personne n'a jamais dit que « foncer dans le tas » était productif. Pas même George Lucas qui fera faire demi-tour à son contrebandier adoré, poursuivi par un détachement de *Stormtroopers* armés jusqu'aux dents.

Utilité de son action : nulle. Prise de risques : maximale.

Créer une marque et monter un site associé en se ruant bille en tête sur des questions comme le SEO ou la production de contenus sans avoir une

stratégie bien ficelée est exactement la même chose. Connaître et maîtriser son identité de marque, formaliser son *storytelling*, sont des éléments clés pour s'assurer que les actions menées seront pertinentes... et donc rentables.

Nous allons voir pourquoi il faut investir dans ce temps de réflexion et comment le faire de façon efficace.

3.2 Identité de marque et *storytelling*

Pour nous comprendre, nous allons avoir besoin d'une base de vocabulaire commune et il est donc impératif de revenir sur les deux expressions maîtresses de cet article : que sont l'identité de marque et le *storytelling* ?

3.2.1 Comprendre l'identité de marque

En marketing, la *Brand Idea* (ou Identité de marque) est l'ensemble des éléments constitutifs de cette dernière. C'est-à-dire les éléments d'identité, de cibles et d'objectifs, d'histoire et de positionnement qui vont donner du sens à la marque. On y retrouve par exemple les valeurs, les *personas* [1], les choix de produits, les objectifs souhaités... mais également la ligne éditoriale, le *storytelling*, et jusqu'aux éléments de design.

Elle se distingue de la *Brand DNA* (ADN de marque) qui, elle, va cibler plus spécifiquement son positionnement de cœur.

Par exemple : chez Coca-Cola, on relèverait différents éléments d'identité comme les couleurs, l'utilisation d'imaginaires de l'enfance (père Noël, jeux vidéo, etc.), une segmentation plutôt jeune, etc. Mais, si l'on devait formaliser la *Brand DNA* de cette marque, cela serait plutôt « Nous souhaitons proposer un produit qui incarnerait les valeurs de partage et de bonheur » [2]. Bien entendu, l'ADN de marque compose l'identité de cette dernière.

3.2.2 Comprendre le *storytelling*

À l'instar de l'ADN de marque, le *storytelling* fait partie intégrante de l'identité de marque. Mais, contrairement à ce qu'on lit parfois ici ou là, il ne s'agit pas de « l'histoire de la marque ». « L'art du *storytelling* » n'est

1. Cible marketing définie par un ensemble de caractéristiques comme des valeurs démographiques ou des intérêts personnels.

2. À noter qu'un des slogans de Coca-Cola a été littéralement « Partage du bonheur ».

pas celui de raconter des histoires, quoi que semble vouloir nous faire dire la traduction littérale du terme.

Si, en théorie, le *storytelling* est bien le fait de créer des histoires autour d'une marque, en pratique, on utilisera l'expression « créer un métarécit » [3] qui traduit davantage la subtilité de la matière. Le *storytelling* n'est pas l'action de raconter des histoires, mais de vendre des histoires... Entreprises, entrepreneurs, politiques, associations [...] tout ce qui peut avoir un *branding* [4] peut avoir un *storytelling*.

Reprenons l'exemple de Coca-Cola : « Nous créons des produits qui offrent du bonheur à partager », c'est l'ADN de marque, et il y a eu tout un *storytelling* associé de toutes les façons possibles : slogan (« Open Happiness » [5]), publicité (Avec le spot « The Happiness Factory » [6]), et bien entendu, les supports de communication filent les éléments narratifs associés au moment.

Mais, un *storytelling* n'est pas figé pour autant ! L'autre arc narratif favori de la marque rouge est sur l'authenticité de son goût. Une sorte de « souvent copié, jamais égalé ». Et cela se traduit par des slogans comme : « Cette sensation s'appelle Coke » [7] ou encore « Seul un Coca-Cola fait l'effet d'un Coca-Cola » [8].

Le *storytelling* est donc un récit qui sera mis en scène par différents moyens, et non pas seulement un morceau de texte descriptif se voulant émotionnel.

Dans cet article, lorsque nous parlerons de *storytelling*, nous évoquerons tout le métarécit d'une marque.

3. Expression notamment utilisée par Sébastien Durand dans son livre « Le storytelling » aux éditions DUNOD.

4. L'image de marque est la notoriété auprès d'un public (consommateurs, électeurs, etc.).

5. Slogan créé en 2009 et traduit en France par « Ouvre un Coca-Cola, ouvre du bonheur ».

6. Une campagne faite en 2006 et primée par un Lion d'argent à Cannes et créée par l'agence Wieden+Kennedy. Cette publicité iconique montre ce qu'il se passe dans un distributeur après avoir inséré la pièce dans des décors magiques et très steampunk qui donnent l'impression que chaque bouteille est un événement heureux. Son succès sera tel que Coca fera par la suite « Happiness Factory 2 » puis, le troisième volet, et couplera cela à un petit film et un documentaire pour présenter les personnages.

7. Slogan français de 1987 qui donnera lieu à une question en 1993 de Monsieur José Balarello, alors député des Alpes Maritimes au ministre délégué à la santé de l'époque sur le danger supposé de ce slogan publicitaire. La question tourne autour du terme « Coke », qui est le nom déposé par la marque à l'origine, et qui serait selon notre législation de l'incitation à la toxicomanie (source : bases des questions du Sénat en 1993).

8. Slogan français de 2005.

3.3 Pourquoi faut-il prendre le temps de travailler son identité de marque et son storytelling ?

Une erreur que j'ai hélas souvent vue est celle de monter tout un site, de créer des campagnes sur les réseaux sociaux et sur le SEO sans être capable de répondre à quelques questions. Des questions qui sont finalement posées lorsque que, un jour, presque au hasard semble-t-il, la marque se retrouve à faire sa page « à propos ». Des questions tournant autour de la notion de valeurs, la notion d'UVP[9] ou encore tout simplement de *storytelling*. Et les raisons invoquées face à cette absence de travail sont fréquemment identiques :

— manque de temps ;
— manque d'argent ;
— manque d'intérêt pour l'exercice ;
— méconnaissance du champ d'action de ce dernier.

Pourquoi des entreprises comme Apple, Coca-Cola ou encore Innocent sont citées en exemples dès que l'on parle de marketing ? Parce que ce sont de grosses firmes ? Oui... mais surtout parce que ce sont des cas d'école parfaits à étudier pour comprendre certains mécanismes de la communication. Tout cela, parce que ce sont trois exemples d'entreprises avec une identité de marque très caractérisée. Ces entreprises ont pris grand soin de travailler ces éléments pour être littéralement remarquables, imprimer les esprits, pour prendre vie !

Pour comprendre toute l'importance de travailler son positionnement – identitaire et narratif – il suffit de puiser dans ces exemples, en particulier lorsque ces derniers doivent relever des défis.

Reprenons celui de Coca-Cola : En 1982, la marque sort le Coca-Cola *light*. Nous sommes en pleine époque des scandales sur la malbouffe, la société est pointée du doigt. La nourriture et le plaisir sont deux éléments associés, et Coca a un positionnement orienté bonheur, plaisir, récompense et joie. L'idée de devoir faire attention à sa ligne entre légèrement en contradiction avec ça, car c'est plutôt une action frustrante, mais Coca va jouer sur son produit en renommant presque en interne son Coca-Cola d'origine le « *Original Taste* ». Consciente que les goûts et donc expériences ne sont pas les mêmes, la marque transforme ce qui est un produit « *light* » en « autre goût ». Soit, non pas un produit pour répondre au scandale, mais bien « une nouvelle expérience Coca ». Le tout, pour rester cohérent avec son identité et son métarécit. Lorsque sortira le Coca-

9. *Unique Value Proposition* (Proposition de valeur).

Cola Zéro en 2005, le *storytelling* sera entièrement axé sur une expérience similaire, tout aussi riche que l'originale [10]. Depuis, la gamme s'est dotée de versions au goût citronné ou à la cerise, et la marque verbalise elle-même l'idée qu'il « existe un Coca-Cola pour tout le monde » [11].

Sans une identité forte et nommée, Coca-Cola aurait eu bien du mal à proposer autre chose qu'une boisson light. C'est cette précision qui permet aux campagnes de communication de pouvoir être percutantes et cohérentes avec la narration générale de l'entreprise.

Mais tout le monde n'est pas Coca-Cola et n'a pas forcément les moyens de la marque. Définir ses valeurs, savoir pourquoi son entreprise est née, savoir ce qu'on va communiquer est pourtant la base d'un lancement. La première chose que l'on fait quand on s'inscrit sur un réseau social ou un site de rencontres est bien de remplir sa bio et de mettre en avant des éléments pour « séduire », « engager », « rassembler », non ?

Malheureusement, il est fréquent que ce temps de réflexion soit jugé dispendieux ou impertinent, et le rattraper est plus coûteux sur le long terme lorsque vient la question de devoir créer des pages « à propos », ou simplement trouver une ligne distinctive de communication. Parce qu'au moment de créer des éléments narratifs, de préparer une action marketing ou simplement de répondre à des questions presse, ne pas connaître ses valeurs, ne pas savoir ce que l'on souhaite transmettre comme métarécit, c'est prendre le risque de créer quelque chose de creux et d'impersonnel qui ne représentera pas la marque.

Les entreprises comme Burger King ou KFC, quand elles se sont implantées en France, l'ont fait autour de campagnes fortes s'articulant sur une identité qui l'était tout autant !

En 2018, l'agence de marketing La/PAC gère le lancement français de KFC et propose un *teaser* mystérieux de 15 secondes, avec musique épique, personnage qui se prépare et une seule phrase « Il arrive » [12]. Le « il », fait référence au Colonel Sanders, figure emblématique de KFC. Bien que le personnage original ait réellement existé, ce dernier est mort depuis longtemps et son image a été retravaillée de sorte qu'elle devienne une icône marketing. Par ailleurs, le positionnement de KFC se fait autour

10. La promesse de Coca est que le produit d'origine, comme le Zéro, ont l'exact même goût, sans calories. Là, où le Coca-Light est plus léger en goût.

11. Cette *baseline* se retrouve fréquemment, notamment sur les sites officiels.

12. Pour plus d'informations sur cette campagne mêlant tweets, spots télévisés associés à la diffusion du film Expandable 2 sur TF1, voir l'article de La réclame à ce sujet : « KFC lance un improbable Colonel Sanders en France ».

d'une recette mystérieuse au nombre d'épices incroyable développée par le Colonel Sanders. Ces éléments sont repris dans toute la communication actuelle, mais figuraient déjà lors du lancement. Le *storytelling* d'un super Colonel qui vient nous « sauver » du poulet fade grâce à sa recette unique est rendu possible parce que formalisé en amont dans l'identité de marque. L'agence La/PAC n'a pas eu à imaginer de but en blanc quelles seraient les valeurs de KFC, au risque de trahir peut-être l'esprit d'origine.

Dans le cadre d'un site de petite ou moyenne entreprise (e-commerce ou site vitrine), il paraît malavisé de travailler ses pages, son référencement naturel et même son nom d'entreprise et logo sans avoir pensé tout d'abord ses valeurs, sa raison d'être et toute sa narration. Comment défendre la légitimité et comment porter la voix d'une entité inconsistante ? Plus important : comment travailler en équipes autour d'un projet s'il n'existe aucune base de référence ?

Ces questions de cibles, de valeurs, de message à transmettre sont des questions posées tout au long de la vie d'une marque :

— à la création de son nom ;
— dans le choix du logo ;
— dans le choix de son esthétique (charte graphique, *packaging* de produits, etc.) ;
— dans son choix sémantique (en plus des axes SEO) ;
— dans ses interventions sur les réseaux sociaux ;
— dans sa relation clientèle.

Et même dans la création du contenu le plus « insignifiant » pour un site Web. Pour une page simple d'explications autour d'une thématique, par exemple une page issue d'un « guide » [13], l'erreur souvent commise consiste à se focaliser uniquement sur la pertinence sémantique du point de vue du référencement naturel. Mais, même si ces pages sont à visée de référencement, elles donnent également une aura d'expertise à la marque. Leur fonction de « ressource » commande une certaine imprégnation et personnalisation de ces informations par l'entreprise. D'une société qui fait des burgers à une autre, les choix sémantiques, la conception de la page (son UX, son design) vont différer. McDonald's France ne parle pas de nutrition de la même manière que son concurrent Burger King. Les premiers démontrent par le design et le contenu qu'ils sont encore loin de chercher à séduire par la publicité leur audience déjà

13. Ici, on entendra ce terme du point de vue SEO, c'est-à-dire un ensemble de pages destiné à explorer entièrement une thématique sous le prétexte de la vulgarisation de cette dernière.

conquise, mais le choix même des informations et de leur format très « brut » raconte deux choses précises : ils souhaitent être sérieux et factuels, et, malheureusement, ils n'ont pas à cœur de convaincre sur cet aspect-là. De son côté, Burger King ne parle pas de nutrition, mais de « qualité ». Ce choix est cohérent avec le positionnement marque qui est sur un goût unique et un aspect artisanal (steaks grillés à la flamme). Là, où MacDo sait qu'il doit convaincre sur la question de la malbouffe, Burger King profite de ce genre de pages souvent délaissées pour, au contraire, placer ses valeurs et donc toute son identité.

Le résultat est sans appel : l'une semble plus engageante – et engagée que l'autre. La comparaison entre les deux marques peut aussi se faire sur les pages environnements, qui sont des pages destinées essentiellement à la presse. Et, là encore, la marque avec l'identité la plus définie produit les contenus les plus impactant et efficaces.

Mais, encore une fois, toutes les entreprises ne sont pas Burger King et toutes n'ont pas pour agence de communication Buzzman (Meetic avec LoveYourImperfections ou encore Delsey avec le film d'animation « What Matters is Inside » [14]). Mais, toutes les entreprises ayant besoin d'une identité et d'une voix, l'exercice de la définition/création de l'identité de marque et du *storytelling* est un incontournable pour la pérennité de l'entreprise.

Reste à pouvoir le pratiquer de la façon la plus adaptée au budget et aux équipes travaillant sur le projet.

3.4 Bien concevoir identité de marque et *storytelling*

Il n'y a pas de recette miracle permettant de faire l'impasse sur l'investissement en temps, en argent et en compétences, c'est évident. Pourtant, la comparaison entre Burger King et MacDonald's est intéressante, car il s'agit de deux mastodontes qui ne manquent pas de ressources et qui n'ont donc aucune excuse concernant la qualité de chacune de leurs actions de communication.

Concevoir une identité de marque et un *storytelling* est à la portée de toutes les structures pour peu qu'elles accordent à cet exercice toute

14. Un film d'animation qui utilise le *storytelling* pur pour vendre une valise connectée. On y suit un jeune homme qui vient de perdre son père faire une véritable chasse au trésor créée par son père avant de mourir, dans le but – semble-t-il, de lui faire découvrir le monde... et « ce qui importe vraiment »/« les reconnecter ». La narration de la valise « connectée » qui connecte par le voyage est un bijou qui a été largement salué par le public.

l'importance qu'il requiert.

Mais dans l'optique d'un budget et d'une équipe plus restreints, nous n'évoquerons que les indispensables à formaliser avant toute chose :

— les valeurs ;
— l'UVP ;
— le *Why* ;
— les personas ;
— le *storytelling*.

3.4.1 Définir son ADN de marque

Les valeurs, l'UVP et le *Why* forment l'ADN de marque. Chaque élément est relié aux autres et permet la compréhension ou la fabrication de ces derniers. Le *Why*, que nous n'avons pas défini jusqu'ici est la raison d'être de la marque. Une sorte de réponse existentielle à échelle marketing. À différencier impérativement de l'UVP qui est la rencontre entre les besoins des clients, le savoir-faire de la marque et la proposition de la concurrence.

Les valeurs forment le socle qui va justifier la raison d'être de la marque et qui va également donner ce « pourquoi elle et pas une autre » de l'UVP.

Nous l'avons vu : pour communiquer et mieux vendre, il faut savoir sur quels arguments on va se baser, quel lien on va créer avec sa cible, etc. Une entreprise de cosmétiques bio va avoir des valeurs en commun avec sa cible, valeurs que l'on ne retrouvera sans doute pas dans une industrie pétrochimique.

Pour les définir, il faut repartir du rêve d'origine, du « *why* primordial » de l'entreprise. Pourquoi cette idée ? Vous pouvez également explorer la thématique, la concurrence, écouter la cible pour entendre le champ sémantique – et donc distinguer les valeurs, entourant l'émotionnel et le lien avec les marques phares du secteur.

Exemple [15] : Un cabinet de naturopathie pratiquant également la médecine chinoise souhaite mettre en avant l'aspect humain et bienveillant. Le client énonce de but en blanc trois valeurs : « bienveillance, écoute, conseil ». En explorant l'univers de la thématique et en écoutant avec attention les éléments émotionnels qui filtrent dans les contenus concurrents ou dans les prises de paroles des consommateurs, on peut voir que la valeur

15. Cet exemple est tiré d'un cas rencontré avec une de mes étudiantes en master 2, qui devait créer l'image de marque d'une personne dans ce secteur.

à privilégier sera « l'humain » (ce qui permet d'englober les précédentes valeurs qui se trouvaient être redondantes), suivie de « naturel » et de « traditionnel ». Les deux dernières valeurs offrant la possibilité d'entériner l'idée d'une médecine « douce » et « traditionnelle », en opposition avec une médecine froide, sentant les produits chimiques.

En creusant les raisons de la création de l'entreprise, la personnalité de l'entrepreneur derrière, on peut aller au-delà et commencer, grâce à ces valeurs, à s'intéresser au *Why*.

Toujours le même exemple : Sa raison d'être était d'apporter une alternative humaine aux gens abandonnés par la médecine classique.

Il s'agit ici de la promesse qui sera faite en substance et dont le *storytelling* s'inspirera pour commencer à s'étoffer.

Enfin, l'UVP peut être formalisée, forte des valeurs sur lesquelles elle repose et d'une raison d'être tangible. Dans cet exemple, cela peut très bien être une alternative abordable et de proximité, une réponse aux déserts médicaux et à l'incompréhension des médecins.

3.4.2 Comprendre et identifier ses personas

Il existe plusieurs écoles en matière de création de personas et pour ma part, je prône non pas la data démographique, mais plutôt des méthodes comme la carte d'empathie (voir la figure 3.1).

FIGURE 3.1 – La carte d'empathie

Plus efficace que la data démographique qui se contente surtout d'offrir des éléments impersonnels et vidés d'émotions, la carte d'empathie offre tout un panel d'informations sur les mécanismes psychiques de la cible. À la manière d'une création de personnages pour une histoire, il est plus intéressant de comprendre ses motivations et ses freins, plutôt que de savoir la couleur de ses cheveux.

La carte d'empathie n'est pas nécessairement aussi complexe, la plupart du temps, dégager quelques *insights conso* [16] suffit.

L'intérêt étant de pouvoir dégager les valeurs importantes pour lui, identifier les freins et obstacles pouvant nourrir la narration, ainsi que les désirs sur lesquels appuyer émotionnellement pour créer de l'intérêt et de l'empathie.

Ce travail d'identification va être utile pour toutes les personnes intervenantes sur la marque – à l'exception peut-être des développeurs. La carte d'empathie peut être utilisée en référencement naturel pour comprendre l'intention de l'utilisateur et donc déployer un site qui puisse répondre à ses attentes, et, bien entendu, tous les créateurs de contenus auront besoin de ces éléments pour définir et comprendre le ton et la sémantique émotionnelle à employer. Ce qu'on appelle le *tone&manners* et que l'on retrouvera dans une identité de marque fouillée et formalisée, permettant de donner des éléments clés de cohérence d'exécution et narrative pour tous les prestataires.

3.4.3 Identifier et formaliser son *storytelling*

Il est dangereux de dire que l'on crée un *storytelling*. Cela reviendrait à dire qu'on fabrique un discours et tout un univers pour une personne. Ce n'est pas inexact, puisqu'entre la fiction qui nous dépeint des génies dans leur garage et la réalité qui se tisse sur un ensemble d'héritage et de diplômes prestigieux [17], il y a un gouffre que le *storytelling* – surtout à l'Américaine – franchit très régulièrement. Pour autant, lorsque l'on reprend les éléments cités au-dessus concernant l'identité de marque, on constate une chose : le *storytelling* est un métarécit issu naturellement de l'ADN de la marque. Il sera un diamant brut à tailler pour lui faire adopter la forme idéale pour son ouvrage.

Et si le *storytelling* n'est toujours pas l'histoire d'une marque, mais

16. Ensemble d'émotions et de raisons d'être/faire, ou de ne pas faire, clés chez son client type.

17. Jeff Bezos, Bill Gates...

bien la narration que cette dernière propose, l'histoire de la création de l'entreprise, de son ou ses fondateurs, de la création des produits ou encore de sa gestion des crises permet de nourrir ce *storytelling*.

Un des meilleurs exemples reste Red Bull : A l'origine simple boisson à base de taurine qui a vu une polémique sur ce composant jeter l'image d'une boisson dangereuse. Et, au lieu de faire preuve de pédagogie et de rassurer sa clientèle, la marque a délibérément choisi de rebondir sur cela en menant toute une campagne de *sponsoring*/distribution dans des soirées jeunes ou en club. Explorant la narration d'une boisson *underground*, voire interdite, Red Bull s'est rapidement imposée comme « le goût du risque », jusqu'à être aujourd'hui sponsor de sports extrêmes et de challenges hors du commun comme se jeter dans le vide depuis « l'espace » [18].

Red Bull n'a pas créé son *storytelling* de toutes pièces, la marque a prélevé les éléments narratifs naturels saillants dans son histoire et son parcours de création pour en faire un métarécit efficace.

Si on reprend l'exemple du cabinet de naturopathie, les *storytellings* possibles sont vastes ! Opposition entre deux visions de la médecine, narration de proximité et quasi-thérapeutique (au sens psychologique du terme), retour à une tradition naturelle ancienne (moyen-âge, antiquité, aborigène...), les inspirations ne manquent pas, les récits non plus.

Et pour mieux cerner le *storytelling* le plus cohérent, il faut repartir sur les bases de la conception du récit : l'obstacle. Le cœur de la narration – et donc du *storytelling*, est le conflit. C'est de la volonté de régler les problèmes et points de frictions que naissent toutes les histoires et toutes les entreprises. Enlevez la notion de conflits et obstacles à des oeuvres comme *le Seigneur des Anneaux* et vous obtenez l'histoire d'un jeune adulte qui hérite des affaires de son oncle qui a décidé de se mettre au vert parce qu'il ne supporte plus ses voisins et sa famille... Sans le problème de l'anneau maléfique, il n'y a pas d'appel à l'aventure et sans toutes les péripéties – et donc les obstacles rencontrés, le personnage n'évolue pas, sa route n'a aucun intérêt et le récit est vide de sens et d'émotions.

Ainsi, lorsque Red Bull est accusée d'être une boisson dangereuse, elle fait de cette problématique une opposition séduisante pour une cible qui adore se rebeller.

Coca-Cola qui est l'incarnation de la boisson sucrée sans intérêt nutritif

18. En 2012, la mission « Red Bull Stratos » voyait Felix Baumgartner faire le saut le plus haut du monde en chute libre depuis la stratosphère.

va capitaliser sur l'idée du plaisir si pur qu'il en devient un élixir de bonheur.

3.5 Une image chaotique ou maîtrisée ?

En définitive, une marque génère automatiquement son image de marque [19], et si cette dernière est issue d'éléments laissés en friche, elle pourra s'avérer néfaste sur le long terme.

Et si l'investissement peut effrayer les entreprises les plus modestes, il n'en reste pas moins que l'impérativité de l'exercice se posera à un moment de l'aventure. Reste donc le choix de subir entièrement le calendrier et la narration extérieurs ou bien de dessiner un plan solide et durable.

Ou en d'autres termes, le choix entre foncer avec son *wookie* de compagnie sur des *Stormtroopers* ou intervenir au moment opportun avec son vaisseau spatial, conformément à la stratégie officielle.

19. On fera la différence entre identité de marque et image de marque. La première étant ce que la marque veut être et est, la seconde ce que le public perçoit d'elle, parfois malgré elle.

4. *Inbound marketing* & SEO

Marielle Cairou a plus de sept ans d'expérience dans le marketing, la communication, l'organisation d'événements et la relation client. Elle a été chargée *inbound marketing* en agence, ses missions étaient la gestion et le suivi de portefeuilles clients (PME à grands comptes ; nationaux et internationaux). Marielle est actuellement la *customer success manager* de `babbar.tech`.

4.1 Introduction

Quand on rédige du contenu, on se pose différentes questions comme « mon contenu va-t-il être vu par beaucoup de gens, est-ce que ce contenu va avoir un bon positionnement ? » ou encore « est-ce que ce contenu va répondre aux problématiques de mon audience et les convertir en clients ? ».

À travers ces questions, on relève qu'il peut s'agir de différentes expertises comme le référencement web (SEO) et/ou de l'*inbound marketing* appelé aussi le marketing entrant. Comment ces deux expertises interviennent et pourquoi sont-elles indissociables ?

Pour mieux comprendre comment le SEO et l'*inbound marketing* sont liés, nous allons aborder l'*inbound marketing* et ses fondamentaux puis quelles stratégies sont à mettre en place pour combiner le SEO et l'*inbound marketing* dans votre entreprise.

4.1.1 Les origines de l'*inbound marketing*

Pour l'histoire, c'est Seth Godin [1], ancien vice-président de Yahoo, qui est à l'origine de l'*inbound marketing* grâce au concept de *permission marketing*.

Le marketing de la permission permet aux internautes de choisir les contenus qu'ils souhaitent et de donner la permission de recevoir des contenus de la marque en question. L'objectif final est de proposer du contenu qui réponde aux besoins de l'internaute, qu'il soit de plus en plus engageant, pour créer une relation de confiance et arriver à l'acte d'achat. Il s'oppose au marketing disruptif, qui repose généralement sur des méthodes qui forcent votre marque à se présenter devant des internautes qui ne s'y attendent pas et qui sont de plus en plus agacés de cet aspect intrusif dans leur quotidien.

En *inbound marketing*, le client est au cœur de toutes les actions marketing et commerciales. On parle de marketing pédagogique, permissif... L'objectif de l'*inbound marketing* est d'envoyer du contenu de qualité à la bonne personne au bon moment.

C'est en 2005, que l'*inbound marketing* se démocratise grâce à la création du logiciel Hubspot par Brian Halligan (CEO) et Darmesh Shah (CTO) aux Etats-Unis.

L'*inbound*, c'est une méthodologie marketing et commerciale structurée, dont l'objectif est de générer des contacts qualifiés puis d'envoyer les prospects chauds à la force de vente pour qu'ils achètent ; et ainsi développer l'activité de l'entreprise.

On oppose souvent le marketing traditionnel (dit *outbound*) à l'*inbound*, les approches sont certes différentes mais elles peuvent être complémentaires selon la stratégie et la vision de l'entreprise.

En effet l'*outbound* est centré davantage sur le marketing (la prospection *push* via le *phoning*, l'*emailing*, les publicités radio, TV etc..) alors que l'*inbound* est centré davantage sur le client, on parle de « *Customer centric* ». Les actions en *inbound marketing* peuvent se caractériser par du contenu de qualité, du SEO, des formulaires *optin*, du *nurturing* via la segmentation de votre base de contacts et la personnalisation de vos actions commerciales et marketing.

1. `https://fr.wikipedia.org/wiki/Seth_Godin`

4.2 Les fondamentaux de l'*inbound marketing*

On sait ce qu'est l'*inbound marketing* mais comment mettons-nous en place une stratégie en *inbound marketing* ?

En effet en *inbound marketing* on s'appuie d'abord sur des contenus experts et pédagogiques pour générer des contacts, puis on s'organise en interne entre le marketing et le commerce pour définir ce qu'est un contact qualifié, et comment on traite les contacts venus du site web grâce au SEO. Ensuite on s'équipe d'outils performants : plateforme *inbound*, CRM, outil d'aide à la rédaction, *netlinking*...

Pour mettre en place la méthodologie de l'*inbound marketing*, il y a 4 phases à retenir [2], présentées dans la figure 4.1.

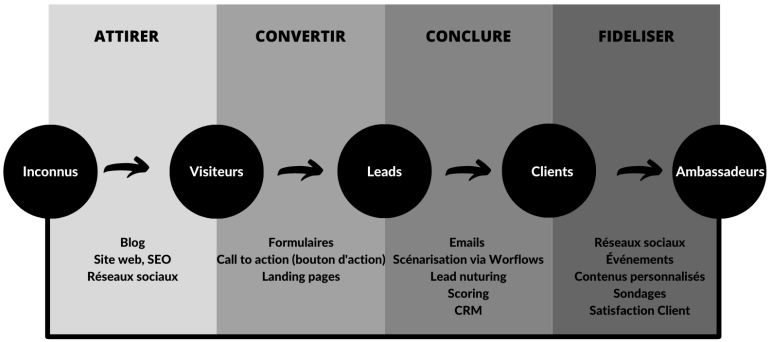


FIGURE 4.1 – Les phases de l'*inbound marketing*

4.2.1 1ère étape : ATTIRER

L'objectif de cette 1ère phase est d'attirer les internautes sur votre site internet pour qu'ils deviennent visiteurs. Ces visiteurs sont d'abord anonymes et sont attirés par vos contenus qui répondent à leurs problématiques.

On remarque que le SEO est l'un des points d'entrée, c'est d'ailleurs la prise en compte des fondamentaux du référencement web qui va permettre d'attirer les internautes. En effet, le comportement d'achat a changé :

2. TRUPHÈME, Stéphane. *L'Inbound Marketing-2e éd : Attirer, conquérir et enchanter le client à l'ère du digital*. Dunod, 2021.

85,5% des internautes français achètent sur l'internet et toutes les recherches sur internet passent essentiellement par Google [3].

C'est primordial de faire du SEO sur son site web pour être trouvé par les internautes et surtout pour les attirer. Pour cela, il faut d'abord analyser les mots et expressions clés qu'utilisent les cibles de votre entreprise lors de l'achat ou des requêtes d'informations avant l'achat. Une fois que vous avez ces informations, il vous suffit de vous positionner sur ces mots et expressions clés.

C'est pour cette raison que la rédaction de contenu a une part importante dans le marketing entrant et le SEO car chaque mot et expression ont une importance et une conséquence.

En *inbound marketing*, il y a 3 notions incontournables à cette phase pour attirer les prospects. Il faut :

— Comprendre ce qu'est une ***persona*** : c'est la cible idéale, l'objectif est de définir le parcours client et les contenus qui pourront intéresser la *persona* à chaque étape. Il faut raisonner par besoin client et non par secteur d'activité. Attention les *personas* ne sont pas des segmentations marketing, il faut penser à toutes les personnes qui auraient besoin de votre service.

— Créer le parcours d'achat : c'est ce qui va permettre de proposer des contenus selon les besoins du prospect et de l'engager progressivement. Il y a 3 étapes dans ce parcours d'achat : La phase de conscience, la phase de considération et la phase de décision.

1. **La phase de prise de conscience / *Awareness***. L'acheteur ressent les premiers signes de sa problématique et il s'intéresse aux différents contenus qui pourraient lui apporter des réponses. Par exemple, il dira « je veux faire des aquarelles mais je ne sais pas quel papier donc j'émets une requête « aquarelles + papier ».
2. **La phase de considération / *Consideration***. L'acheteur/ prospect a clairement défini sa problématique et il souhaite trouver une solution. Il va se dire alors « grâce à ce que j'ai lu et vu, j'ai compris qu'il me fallait du papier spécial aquarelle avec un grammage de 300 ».
3. **La phase de décision / *Decision***. L'acheteur sait comment résoudre sa situation et souhaite la meilleure solution. Il dira « j'ai choisi de prendre du papier aquarelle mais je dois choisir

3. `https://www.alioze.com/chiffres-web`

entre la marque Arches, Canson ou la marque du distributeur ».

— Mettre en place une stratégie éditoriale, ce qui permet de définir des campagnes de contenus qui ont plusieurs objectifs.
    1. Un meilleur référencement naturel grâce à la rédaction de contenus optimisés SEO. Pour cela vous pouvez utiliser l'outil sémantique `yourtext.guru` par exemple.
    2. Une meilleure conversion en proposant des contenus de qualité et qui répondent aux besoins de l'acheteur/prospect.

Maintenant que vous savez comment générer du trafic, l'étape clé de la stratégie *inbound marketing* est la conversion de vos visiteurs.

4.2.2 2ème étape : CONVERTIR

L'objectif est de faire rentrer les visiteurs en base et identifier les contacts qui sont de réels prospects pour engager la conversation. Pour cela, les formulaires sont indispensables car ils permettent de récupérer les coordonnées des visiteurs et de créer une relation de confiance avec votre marque. Plus le niveau d'engagement du contact est élevé, plus il devient un client potentiel.

Pour réussir la conversion, il faut que vous ayez des contenus sous différents formats, ex : inscription à une newsletter, une demande de démonstrations, une demande de contact, un téléchargement d'infographie ou ebook, etc. Ces contenus téléchargeables doivent être de qualité et pertinents dans leurs réponses aux problématiques de vos prospects car ils vont vous donner leurs données personnelles en échange.

Après avoir attiré et converti vos prospects en lead qualifié, il faut les transformer en client. En *inbound marketing*, on s'appuie également sur un tunnel de conversion qui décrit la maturité d'un prospect et les contenus que l'on peut lui soumettre dans le parcours d'achat. La figure 4.2 présente le schéma de ce tunnel de conversion.

Dans la figure, l'acronyme **MQL** signifie *Marketing Qualified Lead*. Il s'agit d'un contact qualifié par le marketing, aussi appelé prospect « tiède », via des contenus téléchargeables ou via le comportement du prospect sur le site via l'application d'un *scoring*. L'acronyme **SQL** signifie *Sales Qualified Lead*. Il s'agit d'un contact qualifié par le marketing et traité par le commerce, considéré comme prospect « chaud » : cette partie est très importante car elle marque l'alignement entre les équipes marketing et commerciales. Il faut que l'équipe marketing soit au courant des éléments du contact qui sont très importants pour le commerce, afin que celui-ci

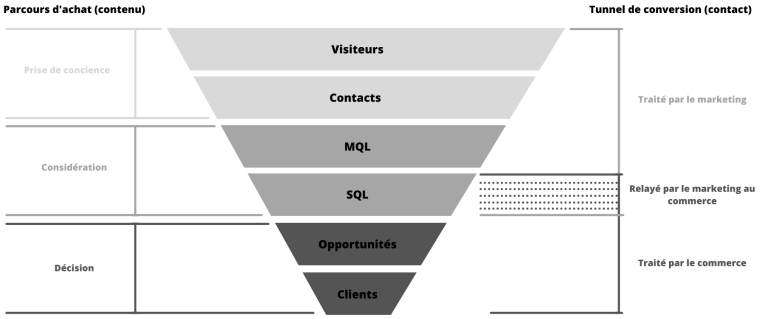


FIGURE 4.2 – Tunnel de conversion *inbound*

puisse prendre le relais et ne pas perdre de temps sur la prospection.

Plus les équipes marketing et commerciales travailleront main dans la main, plus les prospects chauds seront qualifiés et convertis. L'objectif est que l'équipe marketing reçoive les besoins de l'équipe commerciale sur les prospects qu'ils traitent pour leur demander des informations utiles via des contenus téléchargeables et les engager dans le tunnel de conversion. Cet alignement entre les équipes marketing et commerciales permet 3 choses :

— un gain de temps pour les équipes ;
— un prospect intéressé ;
— un gain d'argent.

Ces différentes étapes sont appelées *Lifecycle stages* : les étapes du cycle de vie d'un contact (*lead*). L'objectif est de faire passer le prospect d'une étape à une autre pour qu'il devienne client voire ambassadeur de votre marque.

4.2.3 3ème étape : CONCLURE

Une fois que les prospects sont dans votre base, l'objectif est alors de conclure la vente, de les transformer en client ! Pour cela, nous utiliserons deux techniques : le *scoring* et le *lead nurturing*.

On met alors en place des scénarios d'*emailing* et des méthodes de qualification, pour n'envoyer aux commerciaux que les prospects les plus « chauds » !

Nous devons pour cela définir les critères de qualification qui permettent de caractériser un contact à chaque étape. Nous n'enverrons alors

aux commerciaux que les contacts qui ont suffisamment de critères renseignés (ex : téléphone, email, code postal...).

En parallèle, Le ***scoring*** est un outil qui permet de donner des points aux contacts en fonction de leur comportement. Il permet donc de détecter des signaux faibles. Il permet aussi de prioriser le rappel d'un contact : si deux prospects sont suffisamment qualifiés, mais l'un des deux a un score plus élevé, on le contacte en priorité. Le score permet aussi de créer des *workflows* spécifiques : si le score augmente d'un coup, faut-il envoyer un e-mail en interne pour prévenir l'équipe commerciale ?

Le ***lead nurturing***, c'est le fait d'alimenter, de nourrir, d'engager les contacts. C'est un outil dont l'objectif est de faire passer un contact d'une étape à une autre du tunnel. Il s'appuie essentiellement sur de la scénarisation email et de l'envoi de contenus pertinents.

4.2.4 4ème étape : FIDÉLISER

Une fois la vente signée, l'objectif est de transformer le client et de le fidéliser grâce à des actions personnalisées. Dans cette phase, on va développer la relation client et assurer un suivi du client pour instaurer une relation de confiance. L'enjeu est triple car un client qui est écouté, considéré et satisfait de votre entreprise, est un client qui peut :

— renouveler son abonnement, ses commandes ;
— demander des services et/ou produits supplémentaires ;
— devenir un ambassadeur de votre marque et attirer de potentiels clients.

Il ne faut pas oublier que même si c'est devenu un client, il est nécessaire de l'informer et de lui proposer des contenus *premium* sur l'utilisation de vos services et produits ou autres problématiques qu'il rencontrera.

L'intention est de mieux le connaître pour lui proposer une meilleure expérience client. Pour cela différents indicateurs peuvent être mis en place pour mesurer la satisfaction de vos clients et ajuster vos actions en conséquence des résultats reçus. Parmi ces indicateurs, on trouve :

— Le ***Net Promoter Score***. Le NPS mesure la prédisposition d'un client à vous recommander sur une échelle de 1 à 10.
— Le **score de satisfaction client (CSAT)**. Le CSAT mesure la satisfaction des clients vis-à-vis de votre entreprise ou de l'usage de votre solution.
— Le ***Customer Effort Score***. Le (CES) permet de mesurer la satisfaction client suite à une réponse apportée. Cela permet d'avoir

une indication sur la qualité de la réponse et connaître les axes d'amélioration du support.

Le client est au centre de la stratégie d'*inbound marketing*, il est important de recevoir ses commentaires via ces sondages sur la qualité de réponses de votre service support, de vos services et de vos produits. Il est encore plus important de remonter ces informations à vos équipes pour agir et apporter des améliorations dans le but de satisfaire voire d'enchanter le client.

4.3 L'alliance de l'*inbound marketing* et du référencement web

Pour mieux comprendre ce qu'est l'*inbound marketing*, on va utiliser un terme de pêche. En effet, l'*inbound marketing* consiste plus à utiliser le bon type d'appât qu'à utiliser le filet le plus large. Il est basé sur l'idée qu'avec une approche personnalisée, on touche moins de personnes au total mais qu'on s'engage avec des personnes qui portent un réel intérêt à la marque et qui seront plus susceptibles de faire un achat ou d'utiliser vos services. De nombreuses stratégies classiques de référencement web reposent sur ce type d'approche : les métadonnées, les mots clés dans le contenu et la création de liens sont autant de moyens de signaler aux moteurs de recherche ce que votre site a à offrir.

Voici quelques-unes des stratégies de l'*inbound marketing* les plus courantes en matière de référencement web.

4.3.1 Les chatbots non intrusifs

La présence d'un chatbot non intrusif sur votre site est un très bon moyen d'obtenir un retour d'information immédiat. Il vous permet d'obtenir les pensées et les intentions d'un client pendant qu'il est encore sur votre site. Vous pouvez ensuite utiliser les informations obtenues par le chatbot pour adapter toutes vos autres stratégies en aval.

Des méthodes comme les chatbots permettent de guider les visiteurs de votre site web le long du tunnel de conversion. En interagissant ainsi avec les visiteurs, vous pouvez améliorer l'efficacité de vos autres stratégies, telles que l'optimisation du contenu et du taux de conversion.

4.3.2 Les réseaux sociaux

La présence sur les réseaux sociaux est l'une des stratégies de marketing les plus populaires dans les entreprises modernes, et ce pour une bonne raison. Elle permet une communication directe avec les clients actuels et potentiels sur une plateforme qu'ils utilisent déjà dans leur vie quotidienne. Parler à quelqu'un via Twitter, Facebook ou Instagram augmente considérablement les chances que cette personne se retrouve sur votre site.

Dans une stratégie d'*inbound marketing*, les réseaux sociaux ont également l'avantage d'améliorer potentiellement le référencement web. Une bonne présence sur les médias sociaux, basée sur un bon contenu qui est peut être partagé, signifie que votre site est plus susceptible d'obtenir de précieux liens retours et d'améliorer son référencement. Ces *backlinks* sont importants pour le classement car ils constituent le coeur du *pagerank* de Google. Ainsi, non seulement les médias sociaux améliorent le trafic vers votre site, mais ils démontrent également le lien entre l'*inbound marketing* et le référencement naturel.

Les réseaux sociaux peuvent contribuer à améliorer la visibilité de la marque dans votre secteur et à accroître l'association de noms pour vos produits ou services. Pensez également à la façon dont les réseaux sociaux peuvent renforcer votre stratégie en référencement et en *inbound marketing*. Bien que les interactions avec les réseaux sociaux ne soient pas un facteur de classement direct sur Google, elles peuvent donc avoir un avantage secondaire en améliorant la visibilité de votre marque et, comme nous l'avons mentionné, elles peuvent contribuer à la création de *backlinks*.

4.3.3 Le marketing de contenu

Le marketing de contenu et la stratégie de référencement web ne sont pas du tout séparés ! Ces deux stratégies de marketing courantes sont étroitement liées.

Une grande partie du marketing de contenu comprend la fabrication d'articles de blog, de livres blancs, d'ebooks, de vidéos, infographies etc. qui encouragent l'interaction des utilisateurs avec votre marque. Les billets de blogs, en particulier, sont un élément important de l'*inbound marketing*, car ils portent sur des sujets qui intéressent votre public cible et peuvent indirectement accroître la notoriété de votre marque ou service. Ils sont également très utiles pour le référencement web.

La clé de ces stratégies est de s'assurer que le contenu soit construit sur des sujets qui intéressent réellement votre public. Concentrez-vous sur l'objectif ultime de votre public et élaborez une stratégie de marketing de contenu qui touche le cœur de son objectif. Ne vous concentrez pas sur le contenu pour lui-même, demandez-vous si votre contenu est réellement utile. et s'il répond à ces problématiques. S'il l'est, le trafic viendra de lui-même.

Pour une stratégie *inbound* qui améliore le référencement, concentrez-vous sur la rédaction de contenu **E-A-T**[4] (*Expertise, Authority, Trustworthiness*). Lorsque Google a mis à jour l'algorithme *Medic Update* en 2018, il a souligné l'importance croissante du contenu légitime et digne de confiance. Les sites présentant les caractéristiques E-A-T ont un contenu qui est non seulement utile, mais qui est construit sur l'expertise, l'autorité et la réputation de confiance de votre marque dans votre secteur spécifique.

Google E-A-T oblige à tenir compte de certains signaux sur des sites qui peuvent améliorer le référencement. Des choses comme l'utilisation de mots-clés, les liens internes, la clarté du sujet, l'obtention d'avis, la mise à jour de vos contenus, etc.

Son objectif est d'apporter des contenus web de qualité surtout sur les sites **YMYL** (*Your Money Your Life*). Cela signifie que les sites qui se concentrent sur les achats pour le bien-être, la santé, la sécurité et la sûreté sont plus susceptibles d'être affectés. Si le contenu de votre site correspond à cette description, le contenu E-A-T sera la clé d'une bonne stratégie *inbound marketing* et de référencement web.

Comme vous pouvez le deviner, la qualité d'une page joue un rôle important dans son classement dans les résultats de recherche organiques de Google. Dans les directives, Google indique que les facteurs les plus importants utilisés pour déterminer la qualité globale d'une page web sont les suivants :

— L'objectif de la page (y a-t-il un objectif bénéfique ?).
— L'expertise, l'autorité et la fiabilité de la page.
— La qualité et la quantité du contenu principal.
— Les informations sur le site web ou le créateur du contenu principal.
— La réputation du site web ou de l'auteur du contenu principal.

Ainsi, plus une page fait preuve d'expertise, d'autorité et de fiabilité, plus elle devrait être bien classée. En plus du contenu E-A-T, l'algorithme de Google a également renforcé l'importance du contenu frais. En 2011,

4. `https://www.mariehaynes.com/resources/eat`

il a publié une mise à jour *Google Panda* qui mettait l'accent, entre autres, sur la « fraîcheur » du contenu et encourageait les spécialistes du marketing à réfléchir à un marketing SEO entrant qui fournisse aux chercheurs les informations les plus récentes et les plus pertinentes.

Si vous commencez à voir des pages plus anciennes fonctionner de moins en moins bien et perdre du trafic, demandez-vous si leur contenu est toujours pertinent et exact. Si ce n'est pas le cas, les republier avec des informations actualisées pourrait vous aider. Si le rafraîchissement de l'ancien contenu n'est pas une option, et qu'il a toujours une valeur d'entrée, alors envisagez de rediriger la page vers une page correspondante ou de diriger vos visiteurs vers une partie plus pertinente de votre site.

4.3.4 L'*inbound marketing* dans le référencement web

L'utilisation d'une stratégie de référencement web bien mise en œuvre garantit que les clients que vous avez ciblés ailleurs trouveront exactement ce qu'ils recherchent une fois sur votre site. En matière de référencement, l'*inbound marketing* consiste à cibler des mots-clés que les clients recherchent déjà et à optimiser les pages de votre site en fonction de ces termes. En ce sens, l'*inbound marketing* et le référencement sont indissociables.

Mais le véritable secret d'une campagne d'*inbound marketing* réussie réside dans la création de contenu. En générant continuellement du nouveau contenu produit dans le but réfléchi d'attirer votre public cible, vous attirerez des clients sur votre site. La qualité du contenu et une bonne utilisation des mots-clés sont les clés de l'amélioration des performances de votre référencement web et de votre *inbound marketing*. Ces stratégies peuvent vous permettre d'attirer des clients qui sont déjà intéressés par ce que vous vendez ou par le secteur d'activité de votre marque.

N'oubliez pas que vous ne voulez pas seulement avoir le plus de trafic possible sur votre site, vous voulez des prospects de qualité. Idéalement, ceux qui achèteront, donneront leur avis, parleront de vous à leurs amis et achèteront à nouveau. L'*inbound marketing* et le référencement web sont deux moyens (connectés) de s'adresser à ces types de clients et de développer votre entreprise en ligne.

Décortiquons un peu cela. Vous engagez vos clients en dehors de votre site et vous créez de nouvelles pages ciblées sur votre site. Il est donc à espérer que vos campagnes d'*inbound marketing* et de référencement web travaillent ensemble vers les mêmes objectifs. En gardant à l'esprit les

principes de base du référencement web lorsque vous créez de nouvelles pages, vous prenez déjà les mesures nécessaires pour vous assurer que le contenu contient les termes que vos clients recherchent, en utilisant les mots-clés exacts qu'ils tapent déjà dans les moteurs de recherche.

Le référencement naturel joue un rôle essentiel dans le succès d'une stratégie d'*inbound marketing*. La mise en œuvre d'une stratégie de référencement web solide augmente l'efficacité des deux autres piliers de l'*inbound marketing* : le marketing de contenu et le marketing des réseaux sociaux. En travaillant ensemble, ces stratégies génèrent du trafic ciblé vers votre site web. Toutes ces stratégies de contenu doivent être optimisées pour atteindre l'objectif commun de l'*inbound marketing* :

— Attirer les visiteurs.
— Convertir les visiteurs en prospects.
— Transformer les prospects en clients.
— Fidéliser les clients et les enchanter.

Depuis l'arrivée de l'*inbound marketing*, la création de contenu de qualité reste le critère le plus important pour que l'optimisation des moteurs de recherche soit efficace. Bien que les moteurs de recherche aient changé la façon dont ils présentent et classent les informations aux utilisateurs, ce changement a rendu les résultats beaucoup plus personnalisés et localisés pour chaque utilisateur individuel.

Investir dans la création de contenu original et de qualité est essentiel et n'est plus une option si vous voulez sérieusement faire de l'*inbound marketing*. Plutôt que de jouer avec les moteurs de recherche, fournissez-leur ce qu'ils veulent, rédigez d'une manière qui les aide à indexer et à classer votre contenu plus efficacement. Développez une stratégie de contenu SEO solide pour votre entreprise et produisez le contenu que votre public souhaite vraiment consulter.

4.4 Conclusion

Mettre de l'*inbound marketing* dans le référencement naturel signifie déterminer les besoins/problèmes de vos clients potentiels et créer du contenu pour engager les utilisateurs sur ces sujets. La clé d'un *inbound SEO* réussi est de s'engager avec votre client de manière humaine en utilisant les termes qu'il a déjà recherchés.

Pour ce faire, vous pouvez utiliser les pratiques de référencement web et de l'*inbound marketing* pour optimiser les pages de vos produits et/ou services existants, mais qu'en est-il des nouvelles pages ? L'une des

façons les plus simples d'ajouter continuellement du nouveau contenu à votre site est de créer un blog. Un blog est un excellent moyen de cibler les nouveaux mots-clés que vos clients tapent dans Google et les autres moteurs de recherche. Le contenu du blog doit être frais, pertinent et unique.

Vous ne devez pas vous contenter d'insérer un tas de mots-clés dans vos pages de produits et vos articles de blog, puis vous arrêter là. La lisibilité du contenu de chaque page est tout aussi importante que les mots-clés eux-mêmes. N'oubliez pas que vous essayez d'engager le dialogue avec votre futur client. Il est essentiel de garder à l'esprit que votre contenu doit répondre aux besoins des consommateurs pour réussir une stratégie de référencement web et d'*inbound marketing*.

5. Quand interagir avec Google sculpte notre vision du monde

Syphaïwong Bay est fondatrice de la société Assonance.
Consultante SEO et experte éditoriale, elle a obtenu en 2019 le prix SEMY du jeune espoir search de l'année en marge du SMX de Paris.

5.1 Introduction

Médiamétrie annonçait en avril 2021 que 85,7 % des Français de 2 ans et plus se connectent à Internet, pour environ 2 heures et 39 minutes de temps passé sur les réseaux numériques par chaque personne. Bien entendu il s'agit d'une tendance. Et comme pour estimer la consommation de raclette un soir d'hiver, nous savons qu'il y a toujours une personne qui en consomme énormément pendant que son voisin est intolérant au lactose. Pourtant, le premier chiffre est sans doute le plus évocateur. Car sans parler du temps passé sur le web, il y a bien 85,7 % de personnes en France qui vont sur Internet. Pour chercher une information sans doute : une recette de cuisine, la météo pour les prochaines vacances en bord de mer, ou la liste des effets secondaires recensés à la suite d'une vaccination contre un virus à diffusion mondiale.

Les moteurs de recherche sont bien ancrés dans le monde d'Internet. Sans forcément allumer son ordinateur, une commande vocale pour commander une pizza ou une recherche d'un restaurant à proximité avec son smartphone, font partie des nombreuses façons d'utiliser un moteur de recherche. Il y a quelques années, j'assistais à un débat avec quelques pontes du domaine numérique réunis autour de la question : le SEO est-il mort et enterré ? C'est un marronnier. Ma réponse fut à l'époque simple : tant qu'il y aura des moteurs de recherche, nos métiers sont saufs. Ils changeront, mais nous serons toujours là. Et finalement la mort du SEO n'est en ce sens pas prête d'arriver. En 2002, un groupe de chercheurs (Autret 2002) affirmait que le volume de publications mises en ligne en l'espace de 5 ans était supérieur à tout ce qui avait été édité jusque-là « depuis Gutenberg » [1]. Vingt ans plus tard, il est très facile de s'imaginer que cette proportion a vraiment augmenté. Et dans cet océan d'informations, les moteurs de recherche sont à la fois des outils et des guides. Des outils, car ils nous permettent d'indexer et de recenser un large éventail de données. Des guides, car les méthodes de classement nous orientent vers certaines publications plutôt que d'autres. Les résultats dorénavant enrichis par des images, vidéos, questions que d'autres se posent (*people also ask* ou « PAA ») permettent à chaque page de résultats naturels d'offrir une vision d'une thématique. On pourrait la croire exhaustive, pertinente, utile ou vraie. Elle l'est sûrement, grâce aux qualités des algorithmes de classement. C'est la raison pour laquelle il est important de réaliser que le SEO, tant dans la façon dont il fonctionne que la façon dont nous travaillons, exerce une influence directe sur la connaissance.

5.2 De la fracture numérique à la société de l'information faite de *threads* et de *hashtags*

Impossible de considérer le SEO comme étant une source de connaissance au niveau global sans parler de l'accès à Internet. Au début des années 2000, des paris furent lancés pour faire du web une ressource illimitée à la connaissance. L'ONU a d'ailleurs inscrit l'accès à Internet

1. Gérald Bronner, « Ce qu'Internet fait à la diffusion des croyances », Revue européenne des sciences sociales [En ligne], 49-1 | 2011, mis en ligne le 01 janvier 2015, consulté le 10 décembre 2020. URL : `http://journals.openedition.org/ress/805` ; DOI : `https://doi.org/10.4000/ress.805`

comme étant un droit fondamental en 2016 [2]. Bien que ce sujet soit encore un débat juridique en France pour que cela apparaisse dans la Constitution [3], les initiatives en faveur de la réduction de la fracture numérique vont en ce sens [4].

En 2010 par exemple, l'Allemagne lança une campagne intitulée « iD2010 Informationgesellschaft Deutschland » pour faire suite à la stratégie de l'Union européenne de « répondre aux défis de la société de la connaissance » [5]. En 3 ans seulement, la proportion des ménages ayant un accès à un Internet en haut débit est passée d'environ 30 % à 57 %. Ce chiffre paraît dérisoire, pourtant il faut imaginer les installations et infrastructures que cette croissance demande. Aujourd'hui si les cybercafés se font moins courants, des postes d'ordinateur sont encore disponibles en libre-service dans certaines bibliothèques et surtout spécifiquement pour permettre à certains publics d'accéder à Internet. C'est l'équivalent du poste de Minitel présent dans le bureau de poste sur lequel il était possible de se rendre afin de consulter ses résultats aux examens.

Notre utilisation des moteurs de recherche, inscrite dans le contexte de la lutte contre la fracture numérique pour favoriser l'accès aux savoirs diffusé sur le web, devient très concrètement un outil populaire d'accès à la connaissance. Les méthodes de classement se multiplient et les plateformes s'équipent de méthodes supplémentaires pour aider les moteurs. C'est le principe même de la folksonomie [6] (indexation personnelle), soit le classement par les utilisateurs, grâce à l'usage d'étiquettes (*tag, hashtag*). Autant de *patterns* qui nous permettent d'annoter un contenu en ligne pour qu'il soit trouvé plus facilement. Il s'agit du même schéma que lorsque l'on écrit le contenu d'un <TITLE>, n'est-ce pas pour être trouvé

2. « L'ONU déclare l'accessibilité à internet comme droit fondamental », Infopresse, Consulté le 05 Janvier 2021, URL : `https://www.infopresse.com/article/2016/7/5/l-onu-declare-l-accessibilite-a-internet-comme-droit-fondamental`

3. « Et si le droit à l'accès à Internet était inscrit dans la Constitution française ? », Numérama, Consulté le 05 janvier 2022, URL : `https://www.numerama.com/politique/755454-et-si-le-droit-a-lacces-a-internet-etait-inscrit-dans-la-constitution-francaise.html`

4. « Faire du droit d'accès à l'internet un droit fondamental et autonome », Propositions du 117e concret des notaires de France, consulté le 05 Janvier 2022, URL : `https://www.congresdesnotaires.fr/fr/les-congres/edition-2021/propositions/`

5. Solène Hazouard, « TIC : s'acheminer vers la société de la connaissance », Regards sur l'économie allemande [En ligne], 93 | octobre 2009, mis en ligne le 01 octobre 2011, consulté le 15 septembre 2020. URL : `http://journals.openedition.org/rea/3904`

6. Dominique Cardon, La Démocratie Internet. Promesses et Limites, coll. La République des idées, Seuil, 2010

et apparaître plus facilement selon une requête particulière ?

5.3 Collecter et classer les savoirs pour construire la mémoire collective

Fondée en 1537, la Bibliothèque nationale de France « a une mission de collecte, d'archivage et d'entretien (conservation, restauration), en particulier de tout ce qui se publie ou s'édite en France [...] » [7]. Cette mission est également remplie sur le web, les robots de la BnF collectent en effet les publications en ligne pour les archiver [8]. Un email est alors envoyé à l'éditeur pour le prévenir de l'attribution d'un numéro ISSN. Bien que l'on ne considère pas le travail de la BnF comme étant similaire à Google et son « Don't be evil » [9], les moteurs de recherche développés notamment dans le cadre du projet Gallica [10] suivent entièrement l'idée de donner accès à de l'information en utilisant les moteurs de recherche comme étant des outils facilitateurs. Tout récemment, le projet Gallicagram [11] par Benjamin Azoulay et Benoît de Courson (École normale supérieure Paris-Saclay) met à l'honneur les données issues de la presse grâce à un moteur et un graphe pour visualiser les termes issus d'archives de presse. Ce projet ne manque pas de faire écho au *Google Books Ngram Viewer* [12] qui utilise quant à lui les données issues des ouvrages présents dans le service Google Books.

7. « Bibliothèque nationale de France », Wikipedia FR, Consulté le 05 Janvier 2022, URL : `https://fr.wikipedia.org/wiki/Biblioth%C3%A8que_nationale_de_France`

8. « Capture de votre site web par le robot de la BnF », Bibliothèque nationale de France, Consulté le 05 Janvier 2022, URL : `https://www.bnf.fr/fr/capture-de-votre-site-web-par-le-robot-de-la-bnf`

9. « Google change de slogan et abandonne "Don't be evil" », Cnet, Consulté le 05 Janvier 2022, URL : `https://www.cnetfrance.fr/news/google-change-de-slogan-et-abandonne-don-t-be-evil-39868498.htm`

10. « À propos de Gallica », Bibliothèque nationale d'Information, Consulté le 05 Janvier 2022, URL : `https://gallica.bnf.fr/edit/und/a-propos`

11. Gallicagram, Consulté le 05 Janvier 2022, URL : `https://shiny.ens-paris-saclay.fr/app/gallicagram`

12. *Google Books Ngram Viewer*, Consulté le 05 Janvier 2022, URL : `https://books.google.com/ngrams`

5.4 Google, moteur de recherche dédié aux paroles de tous

Chercher dans Google, c'est chercher dans un espace public au sein duquel les éditeurs ne bénéficient pas tous des mêmes objectifs ou parcours. La facilité avec laquelle nous pouvons mettre en ligne et rendre visible une information, sur une échelle plus ou moins grande, pose la question de la légitimité.

Anecdote de producteur de contenus : il nous est demandé, à juste titre, d'être en mesure de présenter les sources des informations que nous citons. Dans quelques cas, pas si rares, il est complexe de retrouver la source initiale d'une donnée pourtant communiquée de façon très large. C'est notamment le cas pour des informations chiffrées que nous pouvons retrouver sur de nombreux sites différents. Cependant, quelques recherches révèlent alors assez rapidement que personne ne sait tout à fait qui a proposé cette assertion en premier. Dans ces conditions, il faut espérer trouver un éditeur (en ligne ou hors ligne) assez reconnu par ses pairs pour le poser en tant qu'expert.

La question de l'autorité arrive naturellement, son paradoxe également. Nous voudrions à la fois ne lire que des auteurs ou éditeurs justifiant d'une autorité reconnue tout en valorisant le fait que n'importe qui publiant des choses pertinentes (selon qui ?) mérite tout à fait notre attention.

De plus, bien que le web foisonne d'avis divers et variés, au moment où l'on est exposé à un contenu, notre attention lui serait a priori exclusivement dédiée. Ce serait simplement comme écouter un seul avis, se dire qu'il est valide et universel en partant du principe qu'il soit premier dans un moteur de recherche, puis ne pas avoir autant de temps que cela pour croiser les sources en tant que consommateur d'information. La dernière fois que j'ai cherché une recette de moelleux au chocolat en ligne, je n'ai pas comparé les différents dosages proposés durant des heures. Cela peut paraître désuet, pourtant on s'imagine facilement adopter une attitude similaire pour la recherche d'un conseil médical en tapant la requête « soigner piqûre moustique » dans Google. De la même manière, lorsque nous constatons l'importance des avis clients et le marché que cela représente [13], il est bien difficile de nier l'importance que peut prendre

13. « Tenez-vous compte des avis internet pour choisir un restaurant, un hôtel ou vos vacances ? », Oh Comptoir !, France Bleu, Replay du 13 Février 2020, URL : `https://www.francebleu.fr/emissions/oh-comptoir/champagne-ardenne/tenez-vous-compte-des-avis-internet-pour-choisir-restau-et-hotel`

toute publication en ligne.

5.5 Faut-il se méfier des résultats SEO ?

Les référenceurs sont à tort ou à raison connus pour être des manipulateurs des résultats des moteurs de recherche. La compétence de pouvoir publier, faire indexer et favoriser le classement de contenus est une force. Si elle est beaucoup utilisée pour des questions marchandes, il s'agit également d'un outil de valorisation du savoir. C'est la raison pour laquelle le SEO peut être utilisé dans un but de manipulation d'opinion. Rappelons-nous cet épisode où François Hollande alors en pleine campagne présidentielle eut son nom associé aux termes « incapable de gouverner » [14]. Même si ce type de phénomène ne dure pas dans le temps, ce fut assez efficace pour être encore cité ici 10 ans plus tard.

Encore un paradoxe du SEO, lorsque l'on pourrait croire que le moteur de recherche est une arme inégalée pour l'accès à la connaissance, il devient aussi une source de risques. Le classement, aussi pratique soit-il, n'en demeure pas moins un classement. Grâce à lui, nous n'avons plus besoin de naviguer après la première page de Google. Parfois même, nous nous économisons en évitant de *scroller*. C'est à ce moment qu'il faut se méfier non pas du web ou du moteur de recherche, mais sans doute de nous-mêmes. Le biais de confirmation et la portée qu'il peut représenter sont accentués par le web. Tout ce qui apparaît en première page influence un achat, une opinion ou même un savoir. Pour une idée ou une croyance, chaque publication peut être considérée comme étant une confirmation de son propre avis. Notre cerveau est ainsi fait qu'il croira plus facilement quelque chose qui va dans son propre sens. Ainsi, même si des avis contradictoires apparaissent, le biais de confirmation demeure un véritable obstacle. Outre la valorisation de publications malveillantes, les référenceurs ont constaté les risques d'un manque de pluralité dans les résultats de recherche avec le filtre Bigfoot [15] : un même site pouvait apparaître des dizaines de fois pour une même requête. Ce phénomène

14. « Incapable de gouverner : François Hollande victime d'un Google Bombing », Abondance.com, Consulté le 05 Janvier 2022, URL : `https://www.abondance.com/20120217-11270-incapable-de-gouverner-francois-hollande-victime-dun-google-bombing.html`

15. « Retours sur l'update Google du 15 août 2012 », Blog RESONEO, Consulté le 05 Janvier 2022, URL : `https://blog.resoneo.com/2012/08/retours-sur-lupdate-google-du-15-aout-2012/`

fut très étudié par les agences et professionnels du webmarketing. Mais imaginons l'existence d'un moteur de recherche majoritaire ou avec un quasi-monopole qui ne présenterait que des contenus édités et validés par les mêmes personnes ? Voici le début d'un potentiel récit dystopique qui pourrait mettre en scène la manipulation de l'opinion publique par les mots comme cela est narré dans un célèbre roman de George Orwell [16] au sujet d'une « novlangue ». Arrivé à ce point de l'argumentaire, ce devrait être le bon moment pour évoquer la responsabilité d'une entreprise privée aux intentions lucratives lorsqu'elle a une aussi grande influence sur la connaissance et l'opinion publique. Devrions-nous comptabiliser les apparitions dans Google des interventions des candidats à la présidence de la République ? À noter que si les publications en ligne ne sont pas concernées par le décompte des temps de parole durant les campagnes électorales (sauf demande du CSA aux éditeurs concernés), Internet est concerné par l'obligation du respect du « temps de réserve » la veille et le jour du scrutin [17].

5.6 Comment participer à l'internet de la connaissance grâce au SEO ?

Par facilité ou tout simplement par manque de temps ou de budget (est-ce la même chose ?), nous optimisons la plupart de nos contenus de façon à correspondre à ce que l'on attend de nous. Les outils dédiés au SEO se basent tous plus ou moins sur une logique de *reverse engineering* pour offrir aux robots des moteurs de recherche des contenus qu'ils vont *a priori* apprécier. Pour ce faire, nous nous basons sur des données de publications déjà existantes. Mais si vous utilisez toujours les mêmes feuilles de thé pour vos infusions, tout ce que vous produisez pourrait finir par devenir extrêmement fade.

Optimiser un contenu pour le web n'est pas incompatible avec la création de pertinence, d'actualité ou de nouveauté. Au contraire. N'oublions pas que la façon dont est rédigé un texte, bien que cela soit important, ne demeure pas moins un critère parmi tant d'autres. Bien d'autres publications vous parlerons des critères de positionnement en SEO.

16. George Orwell, 1984, paru en 1949

17. « Élection présidentielle 2022 : le rôle du CSA », Conseil Supérieur de l'Audiovisuel, Consulté le 05 Janvier 2022, URL : `https://www.csa.fr/Informer/Toutes-les-actualites/Actualites/Election-presidentielle-2022-le-role-du-CSA`

Profitons donc de la marge de manœuvre permise par la diversité des critères de *ranking* pour amener la voix de nouveaux experts sur le devant de la scène. Allons chercher des informations sorties des sentiers battus et développons des angles d'attaques originaux. Étendons nos champs de recherche en dehors de ce qui se positionne déjà en première page, pourquoi pas sur d'autres moteurs de recherche que Google et même dans des publications hors ligne, des conférences, des films, des documentaires.

Naturellement le lecteur devrait se poser ici la question du budget. Certes le fait de créer du contenu original n'est pas une entreprise bon marché. D'une part, il faut sans doute finir par accepter que la création de contenu n'est qu'un métier cérébral parmi d'autres qui demande du temps et des compétences, donc qui a priori devrait coûter un certain prix plus le temps passé et les niveaux d'exigence augmentent. D'autre part, cela ne signifie pas que pour sauver l'humanité, nous devrions faire que chacune de nos publications révolutionne les domaines dans lesquels elles évoluent. Certains contenus n'ont pas cet objectif. Mais aux entreprises qui souhaitent développer du *brand content* ou tout simplement créer de la valeur et faire-savoir qu'elle existe, la création de contenu intéressant devrait être un fondamental. Contentons-nous déjà de considérer que ce que nous publions sur le web comme étant assez intéressant pour en être satisfait et avoir envie que cela soit lu, vu ou écouté (selon le média utilisé). Pas uniquement dans un objectif marchand ou d'*inbound* numérique, mais tout simplement, car l'on considère ce que l'on a publié comme méritant réellement l'attention des personnes que nous visons.

Une autre piste serait celle de produire moins de contenus ? Le *slow content* propose de publier moins, mais de publier « mieux ». La similarité des contenus n'est pas une notion exclusive aux analyses de détection de duplication de contenus. Des centaines de pages web peuvent dire absolument la même chose sans pour autant alerter un algorithme de moteur de recherche. Autrement, le *content spinning* n'existerait pas. Et qu'en est-il de la génération de contenus par des intelligences artificielles ? Ces IA sont pour certaines tout à fait performantes et convaincantes dans leurs capacités à rédiger du contenu lisible et compréhensible en reprenant des jeux de données que nous leur proposons. Oui ces algorithmes créent du texte, mais elles ne créent pas nécessairement des idées. Gagnons du temps pour créer de la banalité avec l'intelligence artificielle, pourquoi pas, laissons le terrain de la singularité aux humains.

6. À propos des *embeddings*

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`.

6.1 Introduction

L'usage du terme anglais *embeddings* ou *vector embeddings* - ou plus spécifiquement sa première occurrence spécifiquement dédiée à la langue, les *word embeddings* pour désigner une méthode de représentation des mots et des documents - ou « plongement » en bon français, est relativement récent (2013-2014 si l'on en croit google trends, voir figure 6.1), et pourtant la notion qu'il dénote est assez ancienne. Dans le domaine du traitement automatique des langues, la problématique de la représentation est en effet étudiée depuis un certain nombre de décennies, et une rapide plongée historique permettra de comprendre la génèse des notions qui ont précédé l'émergence de ce terme, désormais utilisé largement pour désigner toute représentation vectorielle permettant de remplacer un symbole, une image, un son, de manière sémantiquement productive.

Notre tour d'horizon commence par une école de pensée qui a été

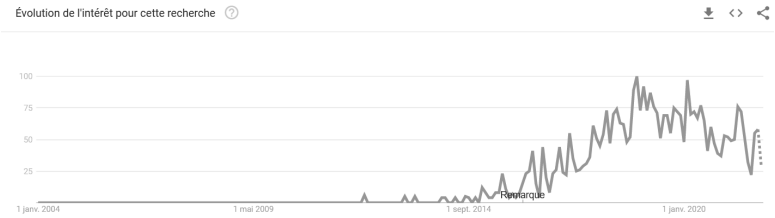


FIGURE 6.1 – L'intérêt pour les embeddings

déterminante pour l'évolution de la linguistique. Initiée par Bloomfield et Harris, la linguistique distributionnelle entend étudier les mots en conditions réelles, et donc à partir de corpus. Mais surtout, Harris et les membres de l'école distributionnelle posent les fondements de la représentation des mots.

6.2 Les mots et leurs ressemblances

> « If A and B have almost identical environments... we say that they are synonyms »
>
> Harris, 1954

> « You shall know a word by the company it keeps ! »
>
> Firth, 1957

Autrement dit, pour l'école distributionnelle, une représentation d'un mot se fait avec d'autres mots, mots qui ont été trouvés dans le même contexte (contexte qui peut être de taille variable, allant du voisinage immédiat jusqu'au document complet). On parle de **cooccurrence** entre deux mots lorsque deux mots se trouvent dans le même contexte.

On note aussi dans ces citations un des objectifs de la représentation : trouver des **synonymes**, des mots qui transportent le **même sens**.

6.3 La représentation des mots

A ce stade, une représentation possible d'un mot comme *choucroute*, d'après le premier paragraphe de sa page Wikipédia, ressemblerait à cette liste : `mets, chou, saumure, lacto-fermentation,etc.` Bien entendu, trouver un synonyme à « choucroute » ne semble pas évident, que ce soit à partir de cette liste ou par une autre méthode. L'existence de synonymes exacts est rare, par contre, on cherche souvent des mots qui ont des sens **proches**). Dans le cas de la choucroute, on peut imaginer, à tout le moins, trouver des termes proches dans les domaines de l'alimentation.

On voit aussi l'importance de la définition du contexte. L'un des premiers usages pour l'utilisation des mots étant la recherche d'information, le niveau document est le plus naturel pour le contexte.

6.4 Des statistiques pour représenter les mots

Les linguistes ont très tôt étudié les propriétés quantitatives des textes. Z. Harris précisera ainsi le fondement de l'approche distributionnelle en introduisant un élément clé : les **statistiques**.

> « Difference of meaning correlates with difference of distribution »
>
> Harris, 1970

Selon cette idée, on ne va donc plus simplement représenter un mot par la liste des mots avec lesquels il est en cooccurrence, mais par une liste de paires `(mot, nombre de cooccurrences)`.

La **proximité** entre deux mots ne va donc plus être simplement liée au nombre de mots en commun dans leur liste de cooccurrences, mais va prendre en compte la similarité de leurs distributions de fréquence.

6.5 Des mots aux vecteurs

Avec cette représentation, on est alors très proche de la notion communément comprise de vecteurs mathématiques. Il suffit de remplacer les mots par leur indice dans une liste de tout le vocabulaire connu, et on obtient, pour représenter un mot, un vecteur des fréquences de cooccurrences pour chacun des mots du vocabulaire. Selon le mot et la taille du contexte choisi pour les cooccurrences, le vecteur va être plus ou moins creux, c'est-à-dire qu'il va contenir plus ou moins de zéros.

Pour savoir si deux mots sont **proches**, il suffit de choisir l'une des nombreuses **distances** sur les espaces vectoriels : distances euclidiennes, distance cosinus, le choix est vaste pour mesurer une distance, et donc, une **similarité**. Ce dernier pas ne fut franchi que tardivement avec notamment HAL (*Hyperspace Analog to Language*), un modèle de similarité entre mots d'inspiration psychologique et utilisant des vecteurs de cooccurrences.

On pourrait penser être arrivé au bout du chemin de ce que permet l'approche distributionnelle statistique. Après tout, quelle autre information pourrait-on tirer d'un corpus que les nombres de cooccurrences entre mots selon différents contextes ? Pourtant à partir de la même information, il existe de nombreuses manières de l'exploiter et d'en tirer des résultats de qualité supérieure.

6.6 Régularisation statistique

En travaillant sur les vecteurs de fréquences de cooccurrence, on se rend rapidement compte d'un problème lié à la nature des langues humaines : certains mots sont extrêmement fréquents, alors que d'autres sont très rares. Le linguiste George Kingsley Zipf [1] a donné son nom à la loi qui exprime une relation empirique qu'il a, le premier, observée entre le rang d'un mot et sa fréquence : le mot le plus fréquent est environ dix fois plus fréquent que le dixième mot le plus fréquent, lui même dix fois plus fréquent que le centième mot le plus fréquent, et ainsi de suite. Cette très grande distorsion dans les fréquences implique un grand déséquilibre de représentation, et ce faisant, un impact disproportionné du bruit sur les mots les plus fréquents.

Ce problème est d'autant plus significatif que les mots les plus fréquents sont très peu porteurs de sens. Les mots les plus fréquents en français sont les suivants (en prenant la racine des mots) : le, de, un, être, et, à, il, avoir, ne. Les premiers mots « porteurs de sens » n'arrivent qu'autour de la cinquantième position.

A cause de cette distorsion, il est très facile pour deux mots sans rapport d'être considérés comme similaires si les distributions de leurs cooccurrences avec quelques uns de ces mots fréquents sont très similaires. Le poids écrasant des hautes fréquences rendant invisibles les plus faibles variations de fréquence sur les mots « riches de sens » .

1. `https://fr.wikipedia.org/wiki/George_Kingsley_Zipf`

Dans le domaine de la recherche d'information, la régularisation par la fréquence d'apparition dans un document a été proposée dès 1973 par Karen Spärck Jones, et a donné naissance à la mesure TF-IDF, consistant à moduler la fréquence d'apparition d'un mot dans un document (TF pour *Term Frequency*) par l'inverse de la fréquence d'apparition du mot par document (IDF, pour *Inverse Document Frequency*). Par exemple un terme fréquent apparaissant dans tous les documents aura un IDF neutre de 1, tandis qu'un terme rare apparaissant dans 10% des documents aura un IDF de 10. L'idée est simplement que plus un terme est spécifique à quelques documents, plus il est significatif.

Dans le domaine des associations de mots, la régularisation par *Pointwise Mutual Information* a été proposée dès 1990 par Church et Hanks. L'idée de ces derniers est de considérer les mots comme des événements a priori indépendants et d'extraire leur corrélation avec la formule de l'information mutuelle :

$$I = \log \frac{P(x,y)}{P(x) \times P(y)}.$$

Cette formule permet de calculer la dépendance statistique entre deux événements, en se basant sur le ratio entre la probabilité observée de cooccurrence des deux événements et le produit des probabilités d'occurrence des deux événements pris indépendamment. En effet si deux événements arrivent avec chacun une probabilité de 1/10, on suppose qu'ils vont arriver simultanément « naturellement » avec une probabilité de 1/100. Si on observe une probabilité significativement supérieure, alors les deux événements sont probablement corrélés, sinon, il n'existe pas de lien entre les événements.

Les vecteurs obtenus après ces transformations présentent des voisinages bien meilleurs que les vecteurs bruts.

6.7 Les *Embeddings* : densification et débruitage

Nous avons vu que l'origine de la représentation vectorielle des mots vient de l'approche distributionnelle de la linguistique, qui entend représenter les mots par leur voisinage observé en conditions réelles.

La régularisation des statistiques brutes permet de corriger un biais évident de la modélisation de la langue sous forme numérique, mais il en subsiste bien d'autres, notamment parce que la régularisation par information mutuelle repose sur l'idée que les mots se comportent comme des

événements, or cette analogie est largement réductrice. Par ailleurs, l'information mutuelle ne fonctionne qu'entre deux variables, or les dépendances croisées sont bien plus complexes dans la langue.

Intuitivement, des vecteurs tels que ceux qu'on obtient par l'approche distributionnelle semblent avoir **beaucoup trop de dimensions** : chaque mot du vocabulaire ne devrait pas être une dimension totalement indépendante des autres. Au contraire, on peut même penser qu'à la place de centaines de milliers de dimensions, quelques centaines de dimensions devraient suffire à exprimer la sémantique d'une langue.

Le propos de cet article n'est pas de rentrer dans des mathématiques compliquées, mais afin de bien comprendre la manière dont peuvent être obtenus les embeddings, et avant d'aborder le sujet des réseaux de neurones, il semble nécessaire d'évoquer rapidement des notions d'algèbre matricielle.

Avant tout, il faut réaliser que lorsqu'on a créé des vecteurs de représentation pour les mots, on a en réalité sous la main une matrice de cooccurrence régularisée, chaque cellule (x,y) de la matrice correspondant à l'information mutuelle entre le mot x et le mot y, ou bien zéro lorsqu'il n'y a aucune cooccurrence entre les mots.

Ensuite, il faut savoir que les matrices ont une dimensionnalité intrinsèque, appelée « rang ». Le rang est le nombre de dimensions qui restent après avoir réduit la matrice à une base orthogonale (en ligne ou en colonne). Le processus d'orthogonalisation est assez simple, et il est à la base de nombreuses opérations appelées décompositions. Après orthogonalisation, on se retrouve avec des vecteurs orthogonaux de normes variables. En triant ces vecteurs selon leur norme on s'aperçoit bien vite, sur une matrice de cooccurrence, que la distribution est loin d'être linéaire (voir figure 6.2). :

Évidemment, cette distribution change en fonction du corpus étudié, mais de manière générale, on constate que les normes décroissent très rapidement, jusqu'à atteindre des niveaux si faibles qu'on peut considérer qu'il s'agit plus de bruit que d'information utile.

C'est à partir de ce constat qu'ont été élaborées les premières méthodes permettant d'obtenir des représentations denses, courtes, et restreintes à quelques dimensions principales, à partir des vecteurs de cooccurrence. On peut citer l'ancêtre de ces méthodes : la LSA (*Latent Semantic Analysis*), aussi connue sous le nom de LSI [2] (lorsqu'elle est appliquée à la recherche

2. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990).

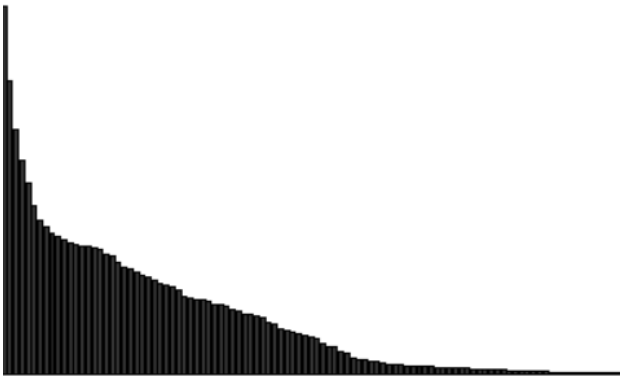


FIGURE 6.2 – La distribution n'est pas linéaire

de documents).

La LSA consiste simplement à appliquer une méthode algébrique de factorisation de matrice (la décomposition en valeurs singulière, ou SVD pour l'acronyme consacré) sur une matrice obtenue par une régularisation des vecteurs de comptes de cooccurrence mot/document. On obtient ainsi 2 matrices, une pour les documents, l'autre pour les mots, et il est possible d'exploiter chaque vecteur de document ou de mot pour calculer des similarités entre mots, entre document, ou entre mot et document.

Dès les années 1990, on parlait donc déjà de vecteurs de mots, d'algèbre sémantique, mais le terme *embeddings* n'était pas encore apparu. Les méthodes utilisées pour obtenir ces vecteurs étaient pour la plupart basées sur des méthodes de factorisation partielle de matrice, mais pas uniquement. On peut citer l'ICA (*Independant Component Analysis*), la pLSI (une LSI obtenue par des méthodes probabilistes), la NMF (*Non-negative Matrix Factorization*) ou encore la LDA (*Latent Dirichlet Allocation*) qui ajoute une non-linéarité lors de la factorisation.

Il a été montré dès 1997 que le choix du nombre de dimensions final était crucial pour la performance des modèles générés. Là où certains utilisaient avant tout la réduction de dimension à des fins d'optimisation, il a été démontré que la capacité des méthodes à séparer l'information pertinente du bruit permettait de manière significative d'améliorer la

Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391-407

qualité des explorations de voisinages sémantiques.

En 2013, deux méthodes, GloVe (***Glo**bal **Ve**ctors for Word Representation*) et Word2Vec ont particulièrement relancé la dynamique de recherche autour des représentations vectorielles de la sémantique. Les approches sont significativement différentes entre les deux méthodes, mais toutes deux s'attachent à traiter un ensemble de problèmes très proches, d'une part la tâche de similarité, qui consiste à mesurer la capacité d'un modèle sémantique à reproduire les jugements de similarité émis par des humains, et d'autre part la tâche d'analogie, consistant à faire deviner le mot D qui est à C ce que B est à A (par exemple, « roi » est à « reine » ce que « prince » est à « princesse »).

Word2Vec a très certainement été un tournant majeur dans les approches par représentations vectorielles, au point que pour certains, rien n'existait auparavant. C'est en tout cas une des méthodes qui, par sa simplicité, sa rapidité, et le fait qu'elle soit accompagnée d'un code clair et très performant, ainsi que par des datasets d'*embeddings* pré-appris, a déclenché un grand nombre de vocations. Les déclinaisons ont été nombreuses, pour calculer des embeddings de documents, de paragraphes, mais aussi pour des embeddings de graphes, pour de la recommandation, etc.

Arrêtons nous là pour les *embeddings* de texte, les modèles de langue avec BERT, GPT et autres ont beaucoup fait parler d'eux ces dernières années.

6.8 *Embeddings* à tous les étages

La représentation d'entité sous forme d'un vecteur n'est pas limitée au texte (on a déjà vu qu'on pouvait l'appliquer à différent niveaux : mot, phrase, document). Un des autres usages qui est apparu assez tôt a concerné les recommandations. Le cadre général de la recommandation est le suivant : un utilisateur interagit avec des objets (produits, films, livres, etc.) de manière implicite (en consultant, visionnant, achetant, etc.) ou explicite (en notant). On peut donc représenter l'information d'un système comme une matrice utilisateur/objet avec comme contenu des cases de la matrice soit une conversion numérique des actions implicites (par exemple le temps passé à consulter un livre ou à visionner un film) ou les valeurs brutes des notes données par l'utilisateur. Ne reste plus qu'à créer les embeddings correspondants aux utilisateurs et aux objets, on se retrouve dans la même situation que pour les mots et leurs cooccurrences.

L'histoire de la recommandation a été marquée par le Netflix Prize. En 2006, Netflix publiait un jeu de données assez inestimable contenant une partie des notes attribuées par leurs utilisateurs à un certain nombre de films. Le but du concours était de prédire les notes « cachées » qui avaient été volontairement enlevées du jeu de données. Le score était donné par une fonction d'erreur sur les prédictions. L'équipe victorieuse a combiné un certain nombre de méthodes pour remporter le prix, mais un des algorithmes les plus marquant se nomme SVD++ ou SVD Régularisée. Il s'agissait d'une version fortement modifiée de la SVD (en fait pas vraiment une SVD), par descente de gradient, avec une fonction d'erreur modifiée pour prendre en compte certaines spécificités des notes de la matrice, notamment la moyenne par utilisateur ou par film.

Au final, l'algorithme produisait bien une factorisation (non compatible avec les critères d'une SVD, mais qu'importe) avec deux matrices, une représentant les utilisateurs, et une représentant les films. Des *embeddings*, encore une fois. De nombreux autres algorithmes de factorisation ont été proposés pour faire de la recommandation, on retrouve d'ailleurs des méthodes utilisées pour le texte, comme la NMF, mais aussi des méthodes comme les *Factorization Machines*, dont l'objectif est d'intégrer automatiquement les croisement de variables, sensées démultiplier la dimensionnalité des données d'origine.

On l'a vu, dès qu'on peut représenter des données sous la forme d'une matrice, on peut la factoriser et donc en tirer des *embeddings*. Un type de structure de données qui se représente très bien sous forme de matrice est le graphe. En effet, on peut représenter un graphe sous la forme d'une matrice d'adjacence, où les noeuds sont les lignes et les colonnes, et la valeur à la case (i, j) vaut 1 s'il existe un arc entre les noeuds i et j, et zéro sinon. Il devient alors très simple de réduire la représentation d'un noeud à un vecteur numérique, et de trouver, par similarité, des noeuds ayant de fortes similarités du point de vue des connexions.

Les images quant à elles n'échappent pas à la règle, même si leur situation est quelque peu différente. En effet, une image est, de facto, représentée comme un vecteur numérique. Sauf que ce vecteur brut est inutilisable, d'abord car sa taille varie selon l'image considérée, ensuite car deux images visuellement très proches peuvent avoir des vecteurs très différents, il suffit pour cela d'appliquer quelques transformations simples. Cependant, il existe bien des méthodes pour compresser des images en conservant leur sémantique. La méthode la plus efficace est relativement récente, et consiste à exploiter les dernières couches intermédiaires d'un

réseau de neurones profond entraîné pour classifier les images. En effet ces couches intermédiaires, bien que construites dans un cadre supervisé, permettent d'extraire les caractéristiques clefs des images qui sont utilisées ensuite pour les classifier. Le pari est que ces caractéristiques soient suffisamment générales pour que le vecteur conserve bien la sémantique de l'image et non pas des détails discriminants mais non généralisables. En pratique, cette approche fonctionne très bien sur une grande classe d'images.

6.9 Des *embeddings* partout ?

Nous n'avons pas évoqué en détails les usages possibles des *embeddings*, mais on peut en citer rapidement trois.

— Le premier consiste à utiliser ces *embeddings* en entrées des systèmes d'apprentissage automatique non supervisés, on récupère ainsi une information déjà synthétisée et porteuse de sens.

— Le deuxième usage consiste à appliquer directement des méthodes de classification sur ces *embeddings* ou sur leur agrégation, on économise ainsi grandement sur le temps d'apprentissage et la taille du modèle.

— Enfin, le troisième usage, certainement le plus prolifique, consiste à utiliser les *embeddings* dans un index pour effectuer des recherches par similarité. On peut chercher des utilisateurs similaires, des textes similaires, des textes similaires à une requête, des utilisateurs « proches » d'un produit, etc. Le champs des index de recherche de vecteurs est très actif ces dernières années, tous les moteurs de recherche commencent à intégrer les distances vectorielles et les index approximatifs pour accélérer les temps d'accès.

En conclusion, oui, on va probablement trouver de plus en plus des *embeddings* un peu partout. Certains chercheurs et chercheuses se sont d'ailleurs penchés sur les informations qu'on peut faire « ressortir » de ces *embeddings*, en apparence opaques. Il a ainsi été découvert que dans les modèles de langue, notamment, il est assez facile de faire réapparaître les choses apprises. Ceci pose des questions sur le droit à l'oubli lorsque les modèles sont entraînés sur les données d'internet, mais aussi sur la capacité d'identifier des personnes à partir d'un *embedding* les représentant. Cette problématique est d'ailleurs au coeur d'une réflexion très actuelle sur l'identification, les cookies et les solutions comme le *federated learning*

of cohorts, qui consiste pour les grandes plateformes d'internet comme Google ou Facebook à fournir non plus des identifiants d'utilisateurs pour le ciblage publicitaire, mais des vecteurs représentant de manière bruitée une synthèse des goûts supposés de l'utilisateur, via des cohortes d'utilisateurs similaires.

7. À propos de la génération de contenu pour le web

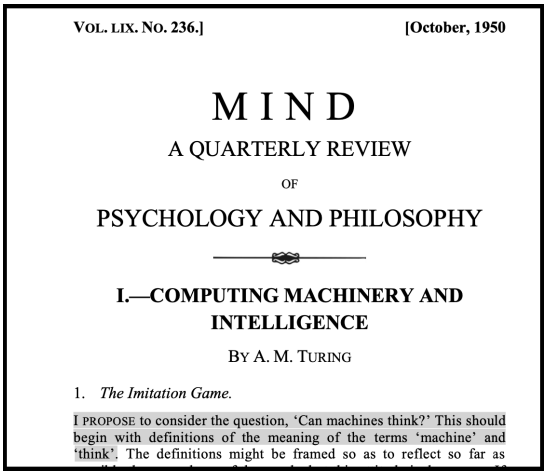
Sylvain Peyronnet est chercheur en algorithmique et spécialiste des moteurs de recherche. Il est CEO et co-fondateur de Babbar, qui exploite les outils SEO `yourtext.guru` et `babbar.tech`. Depuis plus de 20 ans il développe des algorithmes pour l'aide à la prise de décision dans un contexte de grande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

7.1 Introduction

La génération automatique de contenus textuels est un champ disciplinaire ancien qui vise à rendre intelligible par l'être humain des informations stockées « informatiquement ». Depuis déjà plusieurs années des éditeurs de sites web, mais aussi des prestataires spécialisés, utilisent des méthodes de génération automatique plus ou moins sophistiquées, pour par exemple générer des textes donnant des résultats électoraux, des comptes-rendus de matchs sportifs ou encore pour rédiger des fiches annuaires. Ces dernières années, l'émergence des méthodes basées sur les réseaux de neurones a transformé le secteur en profondeur, permettant à tous d'utiliser simplement des objets techniques normalement très complexes.

Les premières réflexions sur le sujet de la génération de contenu ne

sont pas récentes. C'est Alan Turing qui en parle le premier dans son article de 1950 [1]. La figure 7.1 met en évidence la première phrase de cet article, et on comprend en la lisant que finalement toute la question de la machine qui pense est reliée à la question de la machine qui est capable d'articuler un propos correctement et intelligiblement. C'est tout le coeur de ce que Turing a appelé *the imitation game*.



VOL. LIX. NO. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as

FIGURE 7.1 – *The imitation game*

Au delà de ces premières pensées théoriques, des travaux pratiques vont être initiés, en premier par Claude Shannon (le père de la théorie de l'information) qui va proposer, autour de la même époque, de générer des contenus textuels à l'aide de chaînes de Markov. Je ne vais pas rentrer dans la théorie, mais une chaîne de Markov c'est intuitivement un automate dont le passage d'un état à un autre est guidé par des probabilités.
En terme de génération de contenu, cela veut dire qu'on va choisir un premier mot au hasard parmi les mots qui dans la langue initie souvent un début de phrase, puis ensuite qu'on choisira le mot suivant selon les probabilités d'apparition de chaque mot de la langue quand le mot initial a été fixé. Par exemple, je vais choisir au hasard le premier mot « Le », puis ensuite je vais tirer au hasard parmi tous les mots qui peuvent suivre

1. A. M. TURING, I.—COMPUTING MACHINERY AND INTELLIGENCE, Mind, Volume LIX, Issue 236, October 1950, Pages 433–460.

« Le », par exemple le mot « chat », ce qui me donne le bout de phrase « Le chat », je continue et tire au hasard le mot « siamois » pour obtenir « Le chat siamois ». En continuant ainsi je vais générer des textes de plus en plus importants.

Bien entendu, cette approche naïve à ses faiblesses, notamment car elle va générer des phrases qui n'ont pas de sens. En générant de proche en proche comme ci-dessus je peux tout à fait générer la phrase « Le chat boit du whisky » qui est pourtant assez improbable.

Dans ce chapitre, nous allons évoquer la problématique particulière de la génération de contenu pour la publication sur le web. Il s'agit donc principalement de la question de la création de contenus perçus favorablement par les moteurs de recherche. Avant de parler de la génération à proprement parler, je vais commencer par évoquer les technologies modernes qui rendent possible cette génération, et je parlerai aussi de ce qui constitue un bon contenu au yeux d'un moteur.

7.2 Des technologies enfin matures industriellement

Toutes les méthodes de génération fonctionnent sur le même principe : elles « tissent » l'information en prédisant un ou plusieurs mots pour compléter un texte en cours d'écriture. L'idée pour cela est d'explorer un modèle de la langue, qui peut être construit à différents niveaux de granularité (au niveau des groupes de lettres, des mots, des phrases, voire plus).

Dès les années 90, des méthodes à base de réseaux de neurones ont permis de générer plus efficacement des textes. Tout d'abord apparaissent les RNNs (*Recurrent Neural Networks*) qui permettent de déterminer le mot le plus probable qui complète une suite de plusieurs mots. Les RNNs souffrent cependant d'un problème qu'on appelle la fuite de gradient. Je ne vais pas rentrer dans la technique, mais je vais faire une analogie : les RNNs vont se rappeler d'un certain nombres de mots lus dans le passé pour prédire le prochain mot, mais ils souffrent d'une forme d'amnésie et n'ont qu'une mémoire court terme. Si une phrase est trop longue, la prédiction sera faite seulement sur les derniers mots lus dans la phrase. En conséquence les RNNs ne peuvent pas facilement produire des phrases longues cohérentes. Pour palier ce problème les chercheurs vont concevoir les LSTMs (*Long Short-Term Memory*). Je vais encore une fois simplifier le propos, mais basiquement il s'agit d'une variante des réseaux de neurones qui va permettre de stocker provisoirement des informations

importantes pour le contexte, ce qui va améliorer la qualité de prédictions. Sous certaines conditions la mémoire est « effacée », pour s'adapter à un nouveau contexte.

Ce qui va tout changer c'est l'utilisation de ce que l'on appelle l'architecture encodeur-décodeur. L'idée de cette architecture est de mimer ce que l'on suppose être le comportement du cerveau qui lorsqu'il perçoit une information, l'encode pour la stocker, et qui pour s'en servir sort l'encodage de son stockage et décode l'information. En 2017 une équipe de chez Google va publier un article « *Attention is all you need*[2] » dans lequel va être présentée la notion de *transformer*. Un *transformer*, c'est une architecture de type encodeur-décodeur, avec une primitive d'attention, qui va avoir pour but de focaliser l'information utilisée sur des termes qui semblent importants au regard du contexte de la phrase.

Assez rapidement est alors sorti GPT-1, un premier modèle basé sur cette idée. Le modèle s'est montré plutôt pertinent, et capable de faire des tâches en zero-shot : résoudre des problèmes sans avoir vu d'exemples spécifiques du problème et de ses solutions. OpenAI a poursuivi ses travaux, et a sorti GPT-2, un modèle plus gros qui a permis de voir des premiers bons résultats émerger. Simultanément, d'autres modèles similaires sont sortis, notamment Megatron (NVidia) et Turing NLG (Microsoft).

Enfin, la réelle rupture d'un point de vue industriel, c'est GPT-3, qui en passant au-delà d'une taille critique, et avec plusieurs améliorations, montre actuellement des résultats très impressionnants pour la génération de textes. GPT-3, c'est le modèle qui est caché derrière la plupart des outils de génération comme `jarvis.ai` et bien d'autres. OpenAI prépare déjà GPT-4, et évoque même GPT-5. Les autres opérateurs ne sont pas en reste, on annonce par exemple un modèle commun Megatron-Turing NLG environ 3 fois plus gros que GPT-3, ou encore le modèle de AI21 Labs : jurassic-1 Jumbo.

Nous sommes maintenant en 2022, et la communauté des professionnels du web dispose de nombreux outils basés sur ces modèles de la langue. Ces outils permettent de rédiger du texte raisonnablement qualitatif, même si on se rend compte rapidement qu'à l'heure actuelle on ne peut pas laisser la machine rédiger toute seule si on souhaite avoir du texte totalement utile à l'humain.

Dans la prochaine section, nous allons voir ce que peuvent donner

2. Vaswani, Ashish, et al. *Attention is all you need.* Advances in neural information processing systems 30 (2017).

ces modèles, la plupart des exemples seront réalisés avec les modèles SEO-TXL qui sont proposés dans `yourtext.guru`, dans le cadre d'un partenariat entre Babbar et Lighton [3].

7.3 Générer du texte : la pratique

Pour bien comprendre comment l'utilisation de méthodes modernes de génération de contenu peuvent aider à obtenir des textes performants sur le web, il faut d'abord comprendre ce que le moteur de recherche attend en matière de contenu.

7.3.1 L'attente du moteur

Lorsque l'on prend la place du moteur de recherche, les choses sont plutôt simples : un bon contenu est un contenu qui va répondre au besoin informationnel exprimé par un utilisateur via une requête. Les référenceurs travaillent pour être les premiers sur leurs requêtes, et pour cela ils vont proposer des contenus qui vont maximiser les critères attendus par le moteur :

— Ces contenus sont pertinents, ils répondent à la requête. Même si dans l'absolu on pourrait écrire un livre complet sur la signification technique de la phrase précédente, pour un être humain c'est une notion intuitive dont on comprend assez bien ce qu'elle veut dire.

— Ces contenus passent un filtre de qualité minimale au niveau du moteur de recherche. Le point crucial est que la plupart de ces filtres sont basés sur des algorithmes de *machine learning* qui prennent en compte des statistiques sur les mots. Les méthodes de génération automatique de textes étant paramétrées par les statistiques attendues sur les textes générés, elles vont naturellement produire du contenu qui passe la barre en terme de qualité algorithmique attendue.

— Ces contenus sont optimisés pour le moteur de recherche. Ils sont construits en employant le vocabulaire spécifique attendu pour un contenu qui a vocation à bien se positionner pour la requête de l'utilisateur. L'optimisation est une notion indépendante à la fois de la pertinence et de la qualité. C'est essentiellement une propriété statistique des contenus, induite par le mécanisme de regroupement de pages au niveau des bases de données internes du moteur.

3. `https://lighton.ai`

Les outils d'optimisation sémantique comme `yourtext.guru` permettent de comprendre les propriétés statistiques de ces groupes et de guider les rédacteurs dans l'optimisation, c'est à dire dans la création de contenus qui sont les plus susceptibles d'être mis en avant, parmi ceux de l'index du moteur, pour une requête donnée.

Les référenceurs en font souvent un peu trop : le moteur veut mimer le comportement humain, et parfois l'humain est satisfait avec des contenus qui ne remplissent pas les 3 promesses ci-dessus. Cette manie d'en faire un peu trop est souvent ce qui va aboutir à des contre-performances, le moteur pouvant alors détecter la volonté SEO derrière des actions « trop belles pour être vraies ».

L'idée derrière les contenus générés automatiquement est de maximimiser de manière raisonnable chaque critère, sans devoir faire appel à une rédaction humaine (ou avec le moins d'interventions dans le cas d'une approche mixte).

7.3.2 Générer des textes, c'est facile

Sur ce sujet il n'y a pas besoin de grands discours. L'avantage d'une méthode automatique, c'est que comme son nom l'indique il n'y a pas besoin de grandes manipulations pour s'en servir. Il y a bien entendu selon les outils que l'on va utiliser plus ou moins de configuration à faire pour avoir un résultat optimal, mais en pratique sans rien faire on obtient déjà des résultats raisonnables.

Dans la suite je vais montrer deux exemples réalisés avec deux modèles différents. J'ai choisi de générer des textes en anglais, car c'est le langage sur lequel les résultats sont les plus impressionnants. La raison en est très simple : le volume de données d'apprentissage disponible pour la fabrication des modèles est plus important pour l'anglais que pour les autres langues.

On va tout d'abord utiliser le modèle de la langue GPT-3, créé par OpenAI. GPT-3 est actuellement considéré comme le modèle donnant les résultats les plus impressionnants, et il est d'ailleurs la base de nombreux outils en SaaS. Certains de ces outils sont dans une zone grise des conditions d'utilisation de GTP-3 [4]. Pour utiliser directement GPT-3, vous pouvez demander un accès via le formulaire accessible sur la homepage

4. voir la page `https://beta.openai.com/docs/use-case-guidelines/use-case-requirements-library`.

de OpenAI[5].

La figure 7.2 présente le résultat de la génération à partir de la phrase en anglais « *An otter is an animal* » (en français « Une loutre est un animal »). Le texte généré ne nécessite pas de grande explication : il est écrit dans un anglais correct, il est informatif et prolonge bien ce qui a été donné en entrée.

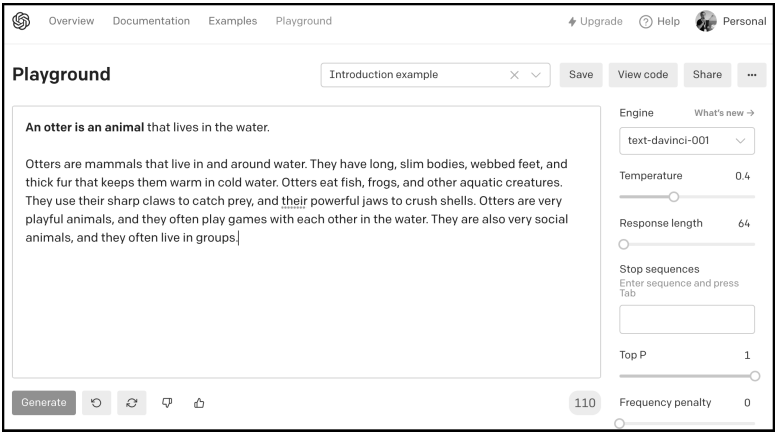


FIGURE 7.2 – Utilisation de GPT-3

Vous pouvez ensuite voir, dans la figure 7.3, le résultat de la génération à partir de la même phrase en anglais, mais pour le modèle SEO-TXL embarqué dans `yourtext.guru`. Là aussi le texte est correct et informatif. On remarque que le texte obtenu avec SEO-TXL est plus académique dans l'esprit, mais rien n'est extrapolable à partir d'un seul exemple.

Ce que l'on remarque en revanche, c'est qu'il y a un léger changement de contexte entre le premier paragraphe et le deuxième. Le premier paragraphe est descriptif de l'animal, tandis que le deuxième porte plus sur les spécificités des loutres d'Amérique du Nord. Bien entendu un tel micro-changement de contexte est tout à fait admissible dans un texte sur le sujet, mais il implique une orientation qui n'est pas forcément celle qu'aurait voulu un auteur humain. Si l'on veut « forcer » la main à l'algorithme, il faut en pratique interagir avec lui en écrivant des entames de paragraphe à la main, pour maintenir un niveau d'information suffisant sur le contexte

5. `https://openai.com/`

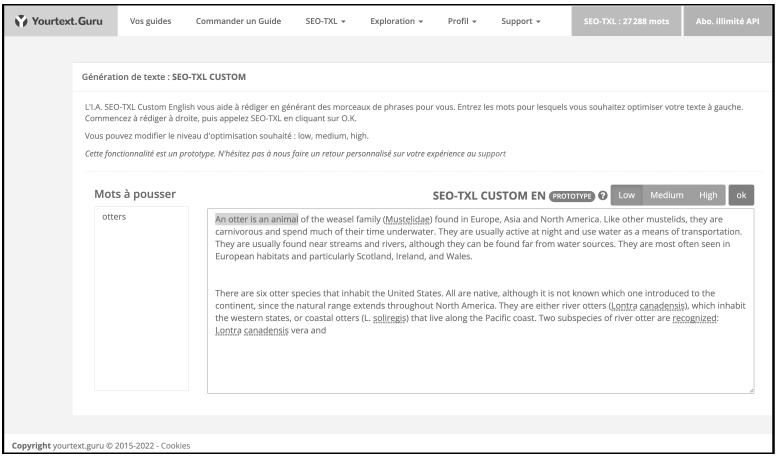


FIGURE 7.3 – Utilisation de SEO-TXL

souhaité.

Bien entendu, avec cet exemple on a l'impression d'avoir trouvé le graal de l'écriture. En pratique ce n'est pas vrai, si vous avez des sujets qui sont peu populaires, ou très techniques, ou basés sur des informations complexes, ou enfin basés sur des informations récentes (après 2020 par exemple dans le cas de GPT-3), alors vous serez sans doute assez déçus des textes générés, et vous devrez beaucoup intervenir pour maintenir un niveau de qualité acceptable.

Ensuite, dans l'optique d'un usage SEO, vous devrez faire une passe avec un outil sémantique pour vous assurer d'avoir un niveau d'optimisation cohérent avec les positions que vous visez au niveau de la SERP. Si vous utilisez SEO-TXL dans `yourtext.guru`, notez que vous pouvez générer du texte nativement optimisé.

7.4 Y a-t-il un risque à utiliser ces méthodes de génération ?

C'est une question complexe que celle de deviner si Google va pénaliser telle ou telle action que peut faire un webmaster ou un référenceur. Avant de se poser la question de la réalité de l'action, on peut déjà se poser celle de la possibilité de l'action.

A cette question, il est facile de répondre. Il est tout à fait possible de détecter des contenus qui sont générés automatiquement à l'aide d'un modèle de la langue. Il y a une littérature scientifique abondante sur le sujet de la détection, avec des résultats chiffrés indiquant des taux de réussite de plus de 80% dans les cas les plus complexes actuels [6].

Concernant le risque en lui-même, il est sans doute encore très mesuré. En effet, Google a d'abord pour but de mettre en avant des pages qui répondent à l'internaute. Si une page générée automatiquement répond idéalement à un besoin informationnel humain, et si cette page est appréciée par les humains qui la visitent via les moteurs de recherche, il y a fort à parier que pour Google l'« auteur » de la page ne soit pas une variable importante. Le moteur est pragmatique : si l'utilisateur est content, alors « il » est content aussi.

7.5 Conclusion

Aujourd'hui, il est déjà possible de réaliser des contenus de qualité avec les méthodes de génération automatique de contenu. C'est particulièrement le cas pour des contenus très normés, comme par exemple les fiches produits ou les résultats des élections. Avec un peu plus de travail, et notamment un va-et-vient constant entre la machine et l'humain, il est aussi possible de faire des contenus génériques de très bonne tenue. Enfin, pour des contenus complexes ou portant sur des évènements récents, l'humain reste bien meilleur que la machine.

Le sentiment actuel sur la maturité de ces technologies de génération est que nous sommes à un tournant. Personne ne sait de quoi demain sera fait, mais il est envisageable que les prochaines générations de modèles permettent d'automatiser encore plus la création des contenus sans grande valeur ajoutée, ne laissant à l'humain que la réalisation des briefs et le polissage des textes de qualité moyenne, et les textes haut de gamme pour ce qui est de la pure rédaction.

Dans ce cas cela se traduirait probablement par des changements massifs dans certains métiers du SEO. En effet, toutes les actions ne se caractérisant pas par la qualité des contenus produits et utilisés bénéficieront

6. C'est un sujet que j'ai abordé dans la lettre Réacteur d'Abondance (`https://www.reacteur.com/2022/01/sylvain-peyronnet.html`. J'y mets en avant les articles disponibles aux URLs `https://par.nsf.gov/servlets/purl/10212709` et `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8049133/`.

soudainement de coûts opérationnels plus bas, ce qui les généralisera dans un premier temps, puis sans doute en annulera l'intérêt sur le long terme.

8. À propos de l'audit automatique de contenu web

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`.

8.1 Introduction

L'analyse des contenus est un champ de recherche qui n'a pas attendu le web pour voir le jour. Néanmoins l'arrivée de ce dernier à changer les choses de par le nombre de documents disponibles et leur facilité d'accès. D'un point de vue académique, les étapes de l'analyse de contenus web ont été formalisées par Sally J. McMillan [1] :

1. Formuler une question
2. Sélectionner un échantillon
3. Définir un schéma de codage
4. Entraîner des codeurs à implémenter le schéma
5. Analyse et interprétation des données

1. McMillan, Sally J. (March 2000). "The Microscope and the Moving Target : The Challenge of Applying Content Analysis to the World Wide web". Journalism and Mass Communication Quarterly. 77 (1) : 80–98

La première étape consiste à savoir pourquoi on souhaite analyser des contenus. Différents acteurs vont se poser différentes questions. Les moteurs de recherche vont chercher à organiser les contenus pour pouvoir répondre au mieux aux besoins informationnels de leurs utilisateurs. Les acteurs du référencement web sont quant à eux plutôt intéressés par comprendre le positionnement d'un contenu.

L'étape de sélection d'échantillon va ici varier en fonction de la question que l'on se pose. Les moteurs de recherche vont chercher à être le plus exhaustif possible et il est dans leur intérêt de connaître le plus grand nombre de contenus. La sélection se fera donc plus sur la qualité des contenus, l'objectif étant de constituer un corpus le plus « propre » possible afin de ne pas stocker pour rien des contenus sans valeur. Si on cherche simplement à comprendre le positionnement d'une page par rapport à d'autres et bien le corpus pourra simplement être constitué des pages qui nous intéressent pour pouvoir les comparer.

Le schéma de codage consiste à définir comment nous allons représenter les informations en vue de les analyser. Il s'agira de définir une unité de codage, la plus petite entité que l'on va étudier. Cela peut varier énormément et cette unité de codage peut aller d'un simple mot à un site web entier en passant par un paragraphe ou une page web. La granularité sera une fois de plus définie par la question que l'on souhaite résoudre. Il peut sembler naturel pour des contenus web d'utiliser la page comme unité de codage mais on peut avoir des résultats plus fins si on arrive à segmenter correctement chaque page en éléments plus précis comme des paragraphes. Il se peut également que l'on souhaite se contenter d'information à l'échelle de sites entiers s'il est inutile de descendre dans le détail pour l'analyse que l'on souhaite faire. Il faut ensuite définir les métriques que l'on veut récupérer sur cette unité de codage ainsi que son contexte. Il est primordial lors de cette étape de s'assurer que les données que l'on collecte vont permettre de répondre à la question posée. Si on ne récupère pas les bonnes données alors il sera sans doute impossible de répondre ou alors on obtiendra une réponse erronée.

Une fois les données codées il ne reste plus qu'à faire l'analyse à proprement parler pour répondre à la question posée au départ. Cette étape est logiquement indépendante du fait que les données récupérées viennent du web. Il s'agit d'une analyse classique qui va dépendre de la question que l'on cherche à trancher et des données que l'on aura effectivement collectées. Les outils que l'on peut utiliser sont nombreux et le *machine learning* permet de traiter des questions complexes en traitant

un gros volume de données qui possèdent beaucoup de métriques. Il n'y a cependant pas d'outil recommandé qui sache répondre à toutes les questions et il appartient à la personne qui fait l'analyse de choisir les plus adaptés.

Dans le cadre de l'audit la question que l'on cherche à résoudre est de pouvoir expliquer le positionnement d'une page/d'un site et surtout de trouver des vecteurs d'amélioration de cette position. Pour cela il va falloir analyser le contenu du site ainsi que celui des concurrents pour pouvoir répondre à la question. En effet du point du vue du moteur un site n'est ni « bon » ni « mauvais » il est simplement meilleur ou moins bon qu'un autre. Lorsque l'on cherche à réaliser un classement cela n'a pas de sens d'étudier les contenus indépendamment les uns des autres. D'autant plus sur le web où les contenus sont « liés » entre eux.

8.2 Création de corpus

Puisque l'on connaît la question, « Comment expliquer le positionnement d'un site ? », il faut maintenant s'atteler à créer le corpus qui va nous permettre d'y répondre. Il est inutile d'analyser l'entièreté du web pour le faire on va donc pouvoir échantillonner et se concentrer sur un nombre restreint de sites. On va pouvoir analyser le site qui nous concerne et ses principaux concurrents. Trouver la liste des principaux concurrents d'un site web est une tâche aisément faite par un humain et très facile à automatiser. La difficulté dans la création du corpus va consister à être capable d'extraire correctement le texte des pages web pour pouvoir l'analyser par la suite.

Comprendre le contenu d'une page web implique tout d'abord de réussir à le récupérer. Il s'agit d'une tâche plutôt facile pour un être humain qui s'en acquitte d'un simple copier/coller. L'utilisateur peut simplement « voir » où se trouve le contenu important. Une dernière chose lui facilite grandement la tâche : le contenu important d'une page est défini comme celui qui est perçu comme important par un être humain.

Même dans le cadre d'un audit où l'on ne souhaite pas analyser l'entièreté du web, il est impossible pour un être humain de récupérer le contenu de milliers ou de dizaines de milliers de pages dans un temps raisonnable. Il faut donc automatiser cette tâche et la confier à des outils.

L'arrivée du HTML5 facilite grandement la récupération automatique de contenu important avec ses balises sémantiques `header`, `footer`, `nav`, `article`, etc. L'intérêt de ces balises n'est plus simplement de structurer

le contenu d'une page mais également de préciser le type du contenu qui s'y trouve. Cela permet de ne pas avoir à parcourir le DOM [2] en détail pour trouver le contenu le plus important.

Malheureusement HTML5 n'a pas encore été adopté par l'ensemble du web [3] et trop de pages n'utilisent pas ou pas totalement les balises sémantiques avec du contenu important se trouvant dans des `div`. Il faut donc être capable de traiter ces pages là pour avoir confiance en l'analyse que l'on fera par la suite.

Il existe plusieurs manières d'extraire le texte d'une page web. La plus simple consiste à simplement récupérer le corps de la page et d'enlever le balisage. Cette méthode est extrêmement rapide mais extrêmement peu précise. On récupère tout le contenu de la page avec les éléments de navigation et de mise en forme sans se concentrer sur l'essentiel.

Fort heureusement il est possible de faire mieux et plusieurs méthodes basées sur des heuristiques ont vu le jour. BoilerPipe [4] est une bibliothèque permettant l'élimination de contenu dit « boilerplate » (c'est-à-dire annexe) basée sur des métriques des blocs de texte [5] et plus récemment trafilatura [6] [7] est un outil qui peut être utilisé pour récupérer directement des contenus en ligne à partir d'une liste d'URLs.

Certaines méthodes quant à elles cherchent à apprendre la structure d'un site pour réussir à en extraire les contenus les plus importants [8]. En effet il est rare que sur un même site web la structure des pages soit hétérogène. Dès lors si un algorithme d'apprentissage réussit à identifier les zones importantes il pourra extraire le contenu d'un site web facilement. Ce type de méthode fonctionne très bien si l'on se concentre sur peu de sites différents, puisqu'il faut faire une passe d'apprentissage à chaque fois. Mais dans le cadre d'un audit où l'on souhaite comparer un site à quelques concurrents cela peut avoir un réel intérêt.

2. `https://fr.wikipedia.org/wiki/Document_Object_Model`
3. `https://w3techs.com/technologies/details/ml-html5`
4. `https://github.com/kohlschutter/boilerpipe`
5. Kohlschütter, Christian, Peter Fankhauser, and Wolfgang Nejdl. "Boilerplate detection using shallow text features." Proceedings of the third ACM international conference on web search and data mining. 2010.
6. Barbaresi, Adrien. "Trafilatura : A web Scraping Library and Command-Line Tool for Text Discovery and Extraction." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations. 2021.
7. `https://trafilatura.readthedocs.io/en/latest/`
8. Zhou, Ziyan, and Muntasir Mashuq. "web content extraction through machine learning." Standford University (2014) : 1-5.

web2Text [9] [10] utilise le *deep learning* pour réussir à faire la différence entre un bloc de contenu important et un bloc de contenu annexe. Une fois l'apprentissage passé le temps d'inférence est relativement court ce qui la rend utilisable en pratique.

Malgré l'efficacité de ces méthodes il faut rester prudent car le web change constamment et les différentes méthodes d'extraction automatique ne sont pas forcément stables dans le temps [11]. Il est donc important de se mettre à jour régulièrement pour être certain de toujours récupérer le contenu le plus important en attendant que les balises sémantiques soient utilisées par 100% des sites web.

Un autre point crucial pour la récupération de contenu concerne le contenu dynamique. En effet, quand on doit récupérer le contenu d'un grand nombre de pages le temps est souvent une ressource limitée. La plupart des outils que l'on peut utiliser en programmation n'interprètent pas le contenu dynamique d'une page. Cela peut s'avérer problématique si le contenu important d'une page est chargé dynamiquement. Dans certains cas il est donc important de pouvoir récupérer ce contenu dynamique.

La manière la plus simple de le faire automatiquement est d'utiliser un navigateur *headless* c'est à dire sans interface graphique. On va juste utiliser le moteur de rendu du navigateur pour avoir la version complètement interprétée de la page avant d'en extraire le contenu. Des outils comme PhantomJS [12] (abandonné en 2017) ou Selenium [13] permettent de faire tourner un navigateur sans interface dans son code pour interpréter complètement des URLs. Une initiative comme Browserless [14] permet également de faire tourner une instance de navigateur pour récupérer le contenu dynamique d'une page.

8.3 Codage de l'information

Ici, il est important de récupérer suffisamment d'informations pour répondre à la question que l'on se pose. Ai-je besoin de récupérer les

9. Vogels, Thijs, Octavian-Eugen Ganea, and Carsten Eickhoff. "web2text : Deep structured boilerplate removal." European Conference on Information Retrieval. Springer, Cham, 2018.

10. `https://github.com/dalab/web2text`

11. Weninger, Tim, et al. "web content extraction : a metaanalysis of its past and thoughts on its future." ACM SIGKDD Explorations Newsletter 17.2 (2016) : 17-23.

12. `https://github.com/ariya/phantomjs`

13. `https://www.selenium.dev/`

14. `https://www.browserless.io/`

liens ? Les menus ? Ai-je besoin de conserver la place des liens dans la page ? Toutes ces questions peuvent-être répondues par l'affirmative ou la négative en fonction de ce que l'on cherche à savoir. Un grand principe étant qui peut le plus peut le moins. Donc dans le doute et si cela est possible au niveau stockage il vaut mieux récupérer une information qui semble peu ou pas utile de prime abord, plutôt que de laisser sur le côté une information qui pourrait s'avérer capitale par la suite.

De nos jours l'immense majorité des informations est extraite automatiquement. Il n'est plus nécessaire d'entraîner des codeurs humains chargés de « transformer » des sites web en données utilisables pour l'analyse. Toutefois si l'on utilise des algorithmes d'apprentissage automatique qui ont besoin de données étiquetées par des humains il faudra être vigilant et bien penser à faire étiqueter chaque contenu par plusieurs évaluateurs pour éviter d'introduire des biais à ce moment là.

Il est également important dans le cadre d'un audit de récupérer des données qui sont actionnables. Si l'on parvient à expliquer le positionnement uniquement par des métriques sur lesquelles on n'a pas la main cela ne permettra pas de faire de recommandations pour améliorer ce dernier.

8.4 Analyse des contenus

Cette étape est évidemment entièrement déterminée par la question que l'on se pose en début de processus. Il existe tout un tas d'outils statistiques permettant d'analyser un contenu ou de le comparer à des contenus voisins.

Une des premières analyses que va faire un moteur de recherche quand il rencontre une page va être de s'interroger sur sa qualité pour déterminer si la page doit intégrer son index.

Une des premières analyses qui peut être faite est de comprendre de quoi parle le contenu. Il est inutile de le comprendre dans l'absolu mais il est nécessaire d'en avoir une représentation qui permette de comparer sémantiquement le contenu à d'autres que ce soit à la requête ou à d'autres pages.

Il est important de comprendre qu'il est impossible de considérer les contenus indépendamment les uns des autres. En effet que ce soit pour le moteur qui cherche à renvoyer les « meilleurs » contenus à l'utilisateur ou dans le cadre d'un audit où l'on cherche à comprendre un positionnement il est toujours affaire de classement et donc d'un ensemble de documents.

Une des manières les plus simples d'obtenir une distance entre deux contenus est d'utiliser le cosinus de Salton [15]. Pour cela il faut posséder une représentation vectorielle des contenus. La première manière de faire consiste à projeter les documents dans l'espace des termes du corpus avec la TF-IDF. Pour chaque document, le vecteur le représentant contient pour la composante associée à chaque terme du corpus sa fréquence d'apparition dans le document divisée par le nombre de documents dans lequel le terme apparaît. L'idée étant qu'un terme apparaissant beaucoup dans un document mais étant très peu présent dans le corpus est discriminant.

Ces méthodes vont évoluer jusqu'à des méthodes ou les vecteurs sont calculés grâce à des méthodes de « plongement lexical » (*word embedding*) qui permettent d'affiner grandement la proximité sémantique calculée. Pour en savoir plus sur ces outils vous pouvez lire le chapitre 6 écrit par Guillaume Pitel sur le sujet au sein de cet ouvrage.

Il faut également se concentrer sur l'organisation des contenus entre eux. La popularité étant un réel levier pour expliquer le positionnement il est important de comprendre d'où elle vient et comment elle est répartie au sein d'un site. Depuis les travaux de Taher H. Haveliwala [16] on sait que cette dernière est étroitement liée au contenu des pages. Il est donc impossible de dissocier les deux analyses.

8.5 Audit automatique

Maintenant que l'on a fait un tour d'horizon de ce qu'il faut faire pour analyser un contenu et donc réaliser un audit, est-il raisonnable de penser que tout ceci peut-être automatisé ? Il est évident que la collecte des données ne pose aucun problème à être faite de manière automatique. Un *scraper* n'a besoin que d'une URL de départ pour crawler un domaine en entier.

Extraire les données des pages peut également être fait de manière automatique. Un humain fera sans doute des choix *a priori* s'il s'aperçoit que collecter telle ou telle donnée n'est pas pertinent, mais un programme pourra lui faire ces choix-là *a posteriori* si la donnée s'avère inexploitable pour l'analyse, sans que cela ne pose le moindre problème.

15. `https://fr.wikipedia.org/wiki/Modèle_vectoriel`
16. Haveliwala, Taher H. "Topic-sensitive pagerank : A context-sensitive ranking algorithm for web search." IEEE transactions on knowledge and data engineering 15.4 (2003) : 784-796.

Faire des recommandations

Aujourd'hui cette partie est la plus compliquée à automatiser. En effet s'il y a un certain nombre de recommandations qui peuvent fonctionner sous forme de règles et sont donc faciles à automatiser (*e.g.* chaque page met > 8s à se charger → Améliorer la vitesse de chargement) ce n'est pas le cas de toutes les recommandations. Certaines viennent d'observations plus subtiles en combinant énormément d'informations très diverses.

Très souvent une recommandation est entraînée par le point de vue global de l'auditeur sur le site et la compétition. Il faut donc être capable d'intégrer toutes ces données et que la machine les « comprenne ». Il n'est pas si aisé d'utiliser directement de l'apprentissage automatique pour cette étape. En effet il faudrait déjà avoir un grand nombre d'audits avec toutes les données d'entrée et les recommandations faites.

Pour ces raisons il est peu probable qu'un audit complètement automatique atteigne un jour la finesse de l'analyse humaine. Néanmoins il est aisément imaginable d'avoir des recommandations automatiques qui soient de grandes directions à suivre. Mettre l'accent sur telle ou telle facette du référencement.

Un effort non négligeable est également à fournir sur le rendu de l'audit. S'il n'y a plus d'humain pour expliquer, il faut trouver des représentations suffisamment claires et explicites pour pour que la personne ayant demandé l'audit soit capable de comprendre et d'entreprendre les démarches pour améliorer son référencement.

8.6 Conclusion

Pour offrir un audit entièrement automatique il reste des leviers techniques assez lourds à lever mais il semble raisonnable d'imaginer que des solutions vont fleurir dans les années à venir.

Au delà de la simple réalisation d'un audit complètement automatisé est-il possible d'imaginer qu'un jour les recommandations pourront être réalisées automatiquement ? Les récents progrès en génération automatique de texte avec GPT-3 [17] ou Jarvis [18] permettront-ils à terme d'avoir du texte optimisé pour le référencement directement ? Sera-t-il possible d'optimiser son *linking* automatiquement de la même manière ? La partie technique du référencement sera probablement toujours hors d'atteinte

17. `https://openai.com/`
18. `https://www.jarvis.ai/`

des modifications automatiques. On imagine mal un responsable des infrastructures autoriser des modifications automatiques sur la configuration de ses équipements pour des raisons de sécurité.

L'analyse de contenu est un champ très formalisé que l'émergence de nouvelles techniques vient régulièrement bousculer. Les avancées technologiques permettent en permanence de traiter plus de documents avec des analyses de plus en plus fines. Les résultats deviennent plus précis et le champ des questions que l'on peut se poser ne cesse de s'agrandir. Sur un corpus de documents, dynamique comme le web, aucune réponse n'est figée et donc il faut penser à régulièrement mettre à jour ses connaissances et ses analyses.

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** [19], pour découvrir les points forts et les points faibles de n'importe quel site web, et **Yourtext.guru** [20] 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.

19. `https://www.babbar.tech`
20. `https://yourtext.guru`

En addition, Babbar monitore les positions de milliards de pages dans 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 juillet 2022 en France par babbar.tech.

Faire un bon contenu pour le web n'est pas anodin. Entre référencement web, storytelling, algorithmes et webmarketing, il faut prendre en compte de nombreux facteurs pour élaborer une stratégie de contenu.

Ce livre, proposé par Babbar.tech et yourtext.guru, est la **compilation de témoignages d'experts** permettant de mieux appréhender les difficultés que vous pourriez rencontrer dans le plan d'action de vos contenus web et référencement.

Prenez en main les concepts SEO et les clés de votre réussite en stratégie de contenus et SEO.

"Un condensé de conseils pratiques d'experts pour réussir votre stratégie de contenus web et votre référencement naturel"

babbar

ISBN 978-2-493567-01-7