1.fr, Un outil d'extraction des champs sémantiques d'une page web

Par Pierre-Yves Landanger

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

L'outil 1.fr permet d'analyser une page web et d'en extraire ses principaux champs sémantiques. Puis, par comparaison avec les résultats renvoyés par Google pour une requête donnée, il est possible de déterminer les champs sémantiques qui renforcent la compréhension de la page en fonction de la requête visée, mais aussi d'identifier les champs parasites négatifs. Explication du fonctionnement de l'outil et exemple d'utilisation...

Imaginez que vous ayez le pouvoir de comparer le contenu d'un article que vous avez rédigé à ce qu'attendent les moteurs de recherche. Ne serait-ce pas plus simple pour améliorer vos positions ? Vous allez découvrir dans cet article, une technique qui va dans ce sens au travers de la description d'un outil que nous avons développé et qui est disponible à l'adresse (on ne peut plus simple !) http://1.fr/.

L'idée de cet article est d'analyser les champs lexicaux des 100 premiers résultats de Google pour une requête donnée, puis de les comparer à ceux d'une page web dont vous avez soumis l'URL.

Comment ça marche ?

Reprenons quelques concepts pour commencer : un champ lexical est un groupe de mots qui partagent une relation de "sens".

Exemple avec un extrait du champ lexical pour le terme "SEO" : *référencement naturel, search engine optimization, moteur de recherche, optimisation, audit seo, moteurs, positionnement, votre site, optimiser, audit, outils seo, duplicate content, netlinking, balise, google adwords, maillage interne...*

La technique que nous utilisons pour notre outil démarre donc par une extraction des champs lexicaux du texte qu'on analyse. Pour y parvenir, on compare tous les mots du texte, deux à deux. S'il existe une relation entre les 2 mots, alors on les regroupe au sein d'un réseau, comme le montre la figure 1.

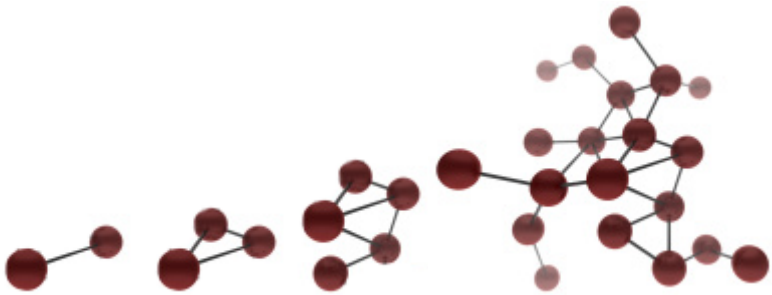

Fig.1. Création d'un réseau de mots ayant une relation entre eux.

Cependant, toutes les relations ne sont pas obligatoirement identiques. Exemple: "SEO" a :

- Une relation forte avec "netlinking" ou "duplicate content" ;
- Une relation moyenne avec "structure", "crawl" ;
- Une relation faible avec "entreprendre", "javascript", "automatiser" ;
- Une relation nulle avec "joint de culasse" ou "compote de pomme".

Autre exemple : la phrase *"Je suis une petite bête toute verte"* contient 3 relations sémantiques faibles avec le mot *"Grenouille"*. La phrase *"Batracien, je suis amphibien."* contient 2 relations sémantiques fortes avec le mot *"Grenouille"*.

Lorsque nous avons construit notre outil, nous avons extrait les champs lexicaux de plusieurs millions de pages web identifiées sur la Toile.

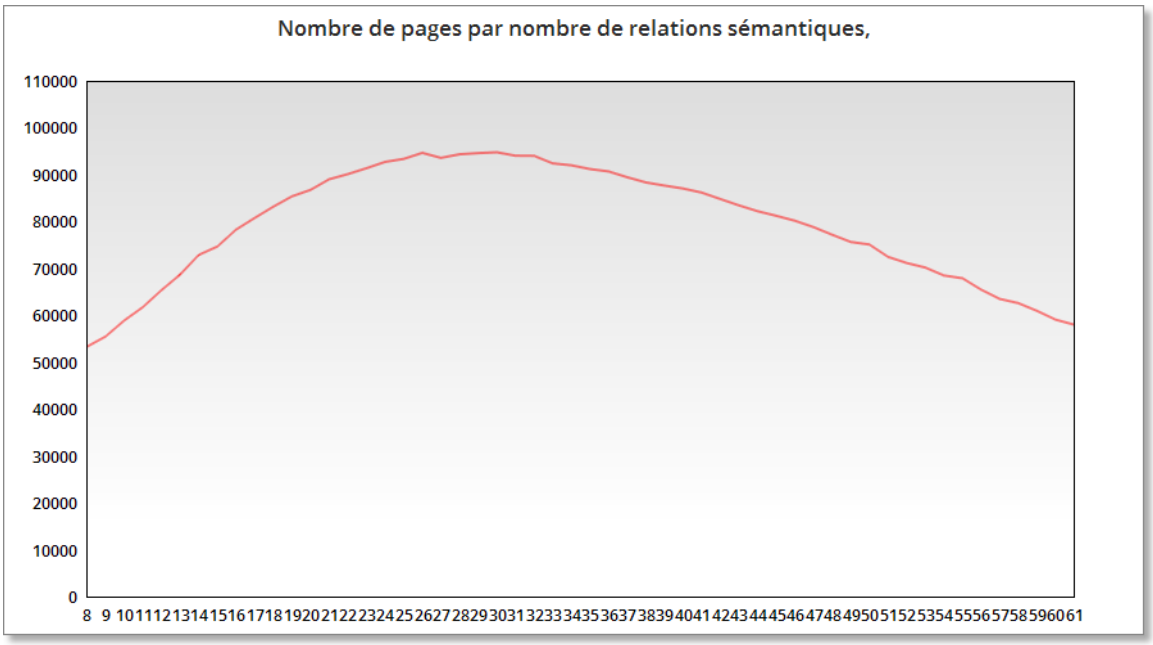

Fig.2. Nombre de relations sémantiques dans une page web.

Comme le montre le graphique de la figure 2, le plus souvent, une page web compte environ 30 relations sémantiques.

Analysons maintenant en figure 3 le nombre de relations sémantiques, en fonction de la position de la page dans un moteur de recherche.

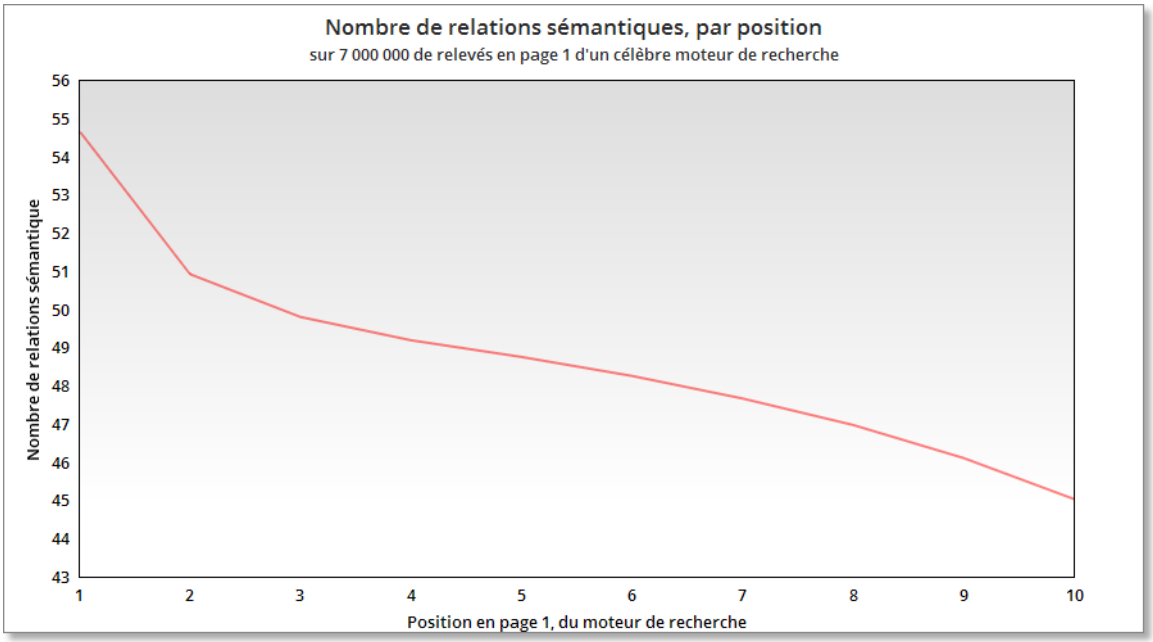

Fig.3. Nombre de relations sémantiques dans une page web en fonction de la position sur Google.

On voit clairement sur ce graphqiue qu'en règle générale, les pages en première position comptent plus de relations sémantiques que les pages en dixième position. Pour simplifier, on peut dire que les pages les mieux classées exploitent généralement plus le champ lexical de l'expression visé.

Notre outil extrait donc, dans un premier temps, les champs lexicaux d'une page et les trie du plus grand au plus petit (selon le nombre de relations sémantiques qui les composent).

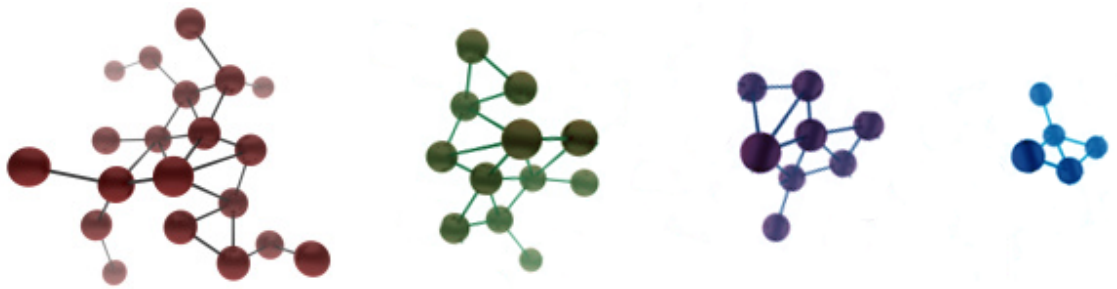

Fig.4. Champs lexicaux, composés de relations sémantiques, du plus grand au plus petit.

Prenons l'exemple de la page web http://www.metiers.internet.gouv.fr/metier/consultant-en-referencement-naturel. La figure 5 montre les champs lexicaux que l'outil a extrait.

Fig.5. Champs lexicaux détectés pour la page http://www.metiers.internet.gouv.fr/metier/consultant-en-referencement-naturel.

Des champs lexicaux extraits, le plus riche en relation de cette page est "référenceur". Le champ lexical "Techniques de référencement" n'étant que 11ème, il compte donc moins de relations.

Un moteur de recherche pourra en déduire que cette page traite principalement de "référenceur", et plus accessoirement de "Techniques de référencement".

Une des premières erreurs que l'on observe lors d'analyses est ce problème de ciblage : on suppose qu'un texte traite d'un sujet principalement, or il le traite accessoirement. Ce qui est très différent et influera énormément sur le positionnement.

Une page doit idéalement compter l'expression visée parmi ses premiers champs lexicaux. Exemple : Si une page est construite pour obtenir un bon classement sur "consultant SEO", alors ce champ lexical devra faire partie des premiers de la page.

Fig.6. Exemple évident de problème de ciblage sémantique d'une page web sur la requête "consultant SEO" (cas imaginaire et quelque peu caricatural :-)).

Dans les faits, bon nombre de pages comptent le champ lexical visé, noyé sous des champs lexicaux parasites. Les champs lexicaux parasites sont des champs qui n'apportent aucun soutien sémantique (par absence de relations) au champ lexical ciblé.

Pour améliorer le classement d'une page, on commence donc par rechercher d'éventuels problèmes de ciblage sémantique pour les corriger en fonction de la requête souhaitée.

A ce stade, nous avons donc :
• Extrait les champs lexicaux d'une page,
• Trié ces champs par nombre de relations sémantiques.
• Appris à détecter un problème de ciblage sémantique, c'est-à-dire une absence de cohérence entre l'objectif d'une page, et les champs lexicaux qui la composent.

Le classement d'un moteur de recherche est le fruit d'une sélection. Cette sélection tient naturellement compte du contenu des pages. Nous pouvons supposer qu'un moteur de recherche, dès lors qu'il fait rentrer des pages dans le haut de son classement, leur accorde une certaine crédibilité. En procédant à une extraction des champs lexicaux de ces pages.... nous pouvons établir un **référentiel**, en analysant ces pages, les champs lexicaux qui les composent, leurs poids et leurs agencements.

Ainsi, en comparant les champs lexicaux d'une page à ce référentiel, nous pouvons déterminer un certain nombre de champs lexicaux qui pourrait **soutenir** (sémantiquement) notre champ lexical cible (la requête sur laquelle ont veut se positionner). Par analyse, il est possible de déterminer si ces nouveaux champs lexicaux sont **parasites**, ou au contraire **soutiennent** le champ lexical ciblé.

Ainsi, après avoir déterminé quels sont les champs lexicaux les plus à même de soutenir la cible, nous pouvons établir des recommandations rédactionnelles. Il s'agit alors de sculpter les champs lexicaux, les agrandir, les réduire, en faire passer certain devant d'autres, etc. Le fait de s'alimenter directement auprès des pages bien classées maximise les probabilités d'une bonne acceptabilité de la part du moteur de recherche visé.

Pour résumer, la technique utilisée par notre outil 1.fr consiste à:
● Extraire, classer et mesurer les champs lexicaux de votre page.
● Extraire, classer, mesurer et filtrer les champs lexicaux des résultats de moteur de recherche sur ses 100 premiers résultats.
● Définir un référentiel idéal de soutien sémantique.
● Établir des recommandations éditoriales.
● Fournir les outils pour sculpter vos champs lexicaux (agrandir, réduire...).

Un exemple : requête "consultant SEO"

Imaginez que nous ayons appliqué cette technique à 3 pages relevées sur la requête "consultant SEO". Peut-être la vôtre…

La Figure 7 indique le type de recommandations qui pourrait résulter de cette analyse.

Fig.7. Exemple de champs lexicaux analysés pour une page web se positionnant en première page sur Google pour la requête "consultant seo"

La couleur verte indique que la situation eest bonne. L'orange détecte un "warning" (avertissement). Lorsque le champ est rouge, le champ lexical nécessite une action en priorité. Pour cet exemple de site en première page, on trouve une excellente couverture des champs lexicaux qui soutiennent la requête "consultant SEO". Le ciblage est satisfaisant. Cette page met donc toutes les chances de son côté.

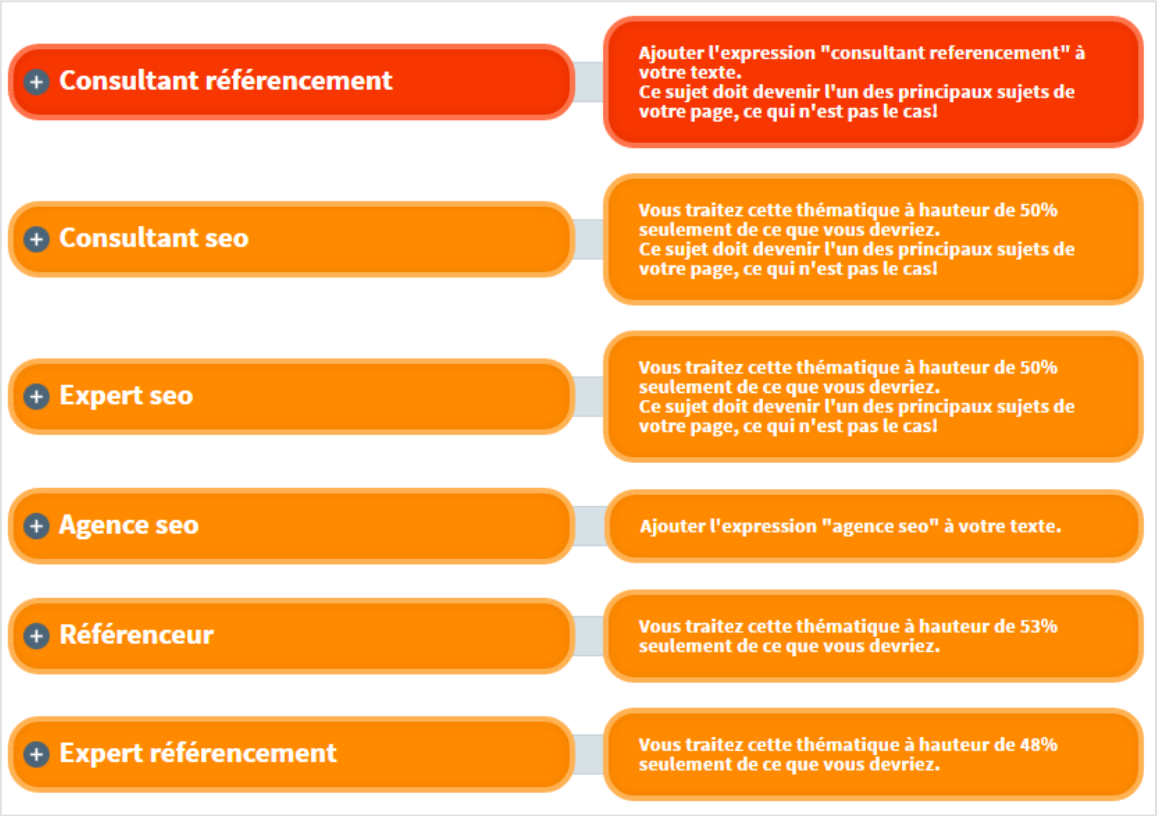

Fig.8. Exemple de champs lexicaux analysés pour une page web se positionnant en deuxième page sur Google pour la requête "consultant seo"

Sur l'exemple de la figure 8, on trouve un site qui se positionnene en 2 ème page de Google. On note un problème de ciblage et un manque de richesse sémantique. Les champs lexicaux semblent sous-exploités (voir commentaires dans les bulles rouges et oranges).

Fig.9. Exemple de champs lexicaux analysés pour une page web se positionnant en cinquième page sur Google pour la requête "consultant seo"

La figure 9 montre les champs lexicaux d'un site qui se positionne en cinquième page de Google, toujours pour la même requête : on note manque de richesse sémantique et des champs lexicaux là aussi sous-exploités.

La technique consiste donc maintenant à tenter de déterminer quels sont les champs lexicaux attendus par un moteur de recherche, et mesurer le fait que notre page remplit bien ce cahier des charges. Ou pas... Autrement dit, utilise-t-on les bons champs lexicaux, dans le bon ordre et dans de bonnes proportions ?

Il nous a fallu un an de développement pour permettre d'automatiser entièrement cette technique, et vous pouvez maintenant la tester en analysant vos pages sur le site http://1.fr. N'hésitez pas à nous dire ce que vous en pensez !

Pierre-Yves Landanger *est gérant de la société Webinfo-LTD. Pour en savoir plus : http://1.fr/*