L'indexation des syntagmes : la fin du raisonnement par mots clés ?

Retour au sommaire de la lettre

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

Jusqu'à maintenant, les moteurs de recherche comme Google, Yahoo! ou Bing fonctionnaient sur la base d'un système d'index inversé : les mots contenus dans les pages web sont alors détectés avec, pour chacun d'eux, la liste des pages dans lesquelles ils apparaissent, accompagnés d'un certain nombre d'informations complémentaires. Mais de nouvelles méthodes se mettent en place, autour de l'indexation des syntagmes ou groupes de mots (contrairement aux mots isolés analysés jusqu'à maintenant), ce qui pourrait profondément changer le paysage du référencement dans les années qui viennent. Cet article tente d'expliquer comment ces systèmes fonctionnent et leur incidence en termes de SEO à moyen terme...

Un brevet publié par Google récemment a mis en lumière un changement potentiel que beaucoup de spécialistes du SEO subodoraient : le passage d'un index à base de mots clés, à un index dont les entrées sont composés de groupes de mots (ou syntagmes). Nous allons expliquer dans la suite de cet article pourquoi un tel changement représente une évolution fondamentale dans les moteurs de recherches et pourquoi la structure de l'index joue un rôle important dans le fonctionnement d'un moteur.

Nous verrons également que l'indexation des syntagmes a des conséquences potentielles importantes sur les stratégies de SEO.

L'indexation des phrases : attention au faux ami !

Les linguistes anglo saxons emploient facilement le terme anglais "phrase" pour décrire un groupe de mots. Le terme "phrase" en français se traduit mieux par "sentence" en anglais. Pour éviter toute ambiguïté, on emploiera plutôt dans cet article, soit le terme "groupe de mots", soit le terme technique "syntagme". Un syntagme constitue l'étape intermédiaire entre le mot et la phrase : c'est un groupe de mots qui forme une unité par son sens et par sa fonction, à l'intérieur de la phrase.

Par exemple dans la phrase "Le chien du voisin a aboyé toute la nuit" on pourra isoler les syntagmes "Le chien du voisin" "a aboyé" et "toute la nuit".

Donc il faut faire attention lorsqu'on lit des articles en anglais sur ce sujet : quand les auteurs parlent d'indexer des phrases, en réalité leurs ambitions sont bien plus modestes et ils cherchent à indexer des groupes de mots, parfois très réduits en nombre.

Le processus d'indexation dans un moteur de recherche

Dans un moteur de recherche, les recherches ne se font pas dans la collection des pages téléchargées lors du processus de crawl. Les pages sont analysées pour en extraire des données qui sont ensuite arrangées dans une énorme base de données organisée sous la forme d'un **index inversé**. Cet index inversé ressemble à l'index des mots clés que l'on trouve à la fin d'un livre : les mots présents dans le livre sont indiqués, accompagnés du numéro de page où on peut trouver ce mot, et même parfois de l'emplacement dans la page. Cette structure en forme d'index inversé explique pourquoi on a pris l'habitude d'appeler la "base de données" d'un moteur un "index". Par contre, les index dans un moteur de recherche comme Google ont une structure beaucoup plus complexe : ils stockent non seulement l'emplacement des termes, mais aussi le poids de ces termes dans le document et de nombreux autres critères précalculés qui permettent de classer les pages de résultat.

Cette structure en index inversé présente un avantage majeur : elle permet d'avoir sous la main des listes de pages contenant des mots clés, ce qui accélère et facilite considérablement la recherche par mot clés et la génération des pages de résultats.

Le processus d'indexation pas à pas

Ce qui est décrit ci-après est un processus très simplifié.
Les pages sont parcourues par un *parser* (un analyseur syntaxique) qui analyse la syntaxe du code de la page pour en comprendre la structure. Classiquement, le parser peut soit éliminer des tags html qui vont gêner l'indexation, ou au contraire permettre de traiter différemment les mots placés au sein de certaines balises comme les <title> ou les <h1> pour attribuer des poids renforcés à ces termes.

Friends, Romans, countrymen.	So let it be with Caesar

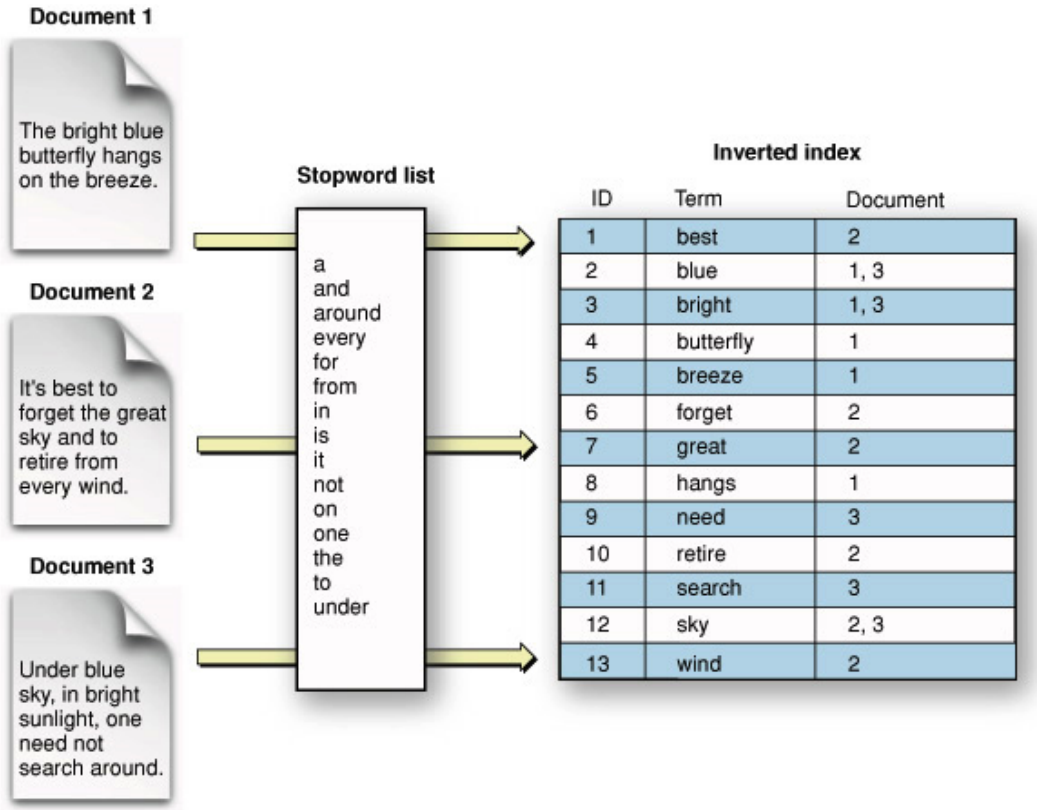
Le contenu textuel est ensuite découpé en éléments simples (baptisé "*tokens*", le processus s'appelle la "*tokenisation*"). Classiquement, les "tokens" correspondent à des mots clés.

Friends	Romans	countrymen	So

Ensuite les token sont *normalisés* : les caractères parasites sont enlevés, les majuscules transformées en bas de casse, etc... Il est habituel également à ce stade de ne pas indexer les mots vides (*stopwords* en anglais) : c'est-à-dire les mots de liaison ou utiles pour la syntaxe, mais qui ne sont pas porteurs de sens (La, de, se etc...)

friend	roman	countryman	so

Enfin, on construit un index, dont les entrées sont constituées de chaque mot, accompagnés de l'indication du document et de l'emplacement du document où l'on peut trouver ce terme :



Une fois l'index réalisé, il devient possible de chercher facilement dans la base de données les pages qui contiennent une occurrence de tel ou tel mot. En cherchant dans l'index à l'aide d'opérateurs booléens ("ET" ou "OU"), il devient possible également de chercher des groupes de mots.

Les Limites des bases à bases de mots clés isolés

Les index à base de mots clés sont caractéristiques de ce que les spécialistes de l'IR (*Information Retrieval*) appellent l'approche en "sac de mots". Cette approche consiste à "tokenizer" les textes, et ensuite à les traiter dans le moteur de recherche sans tenir compte de l'existence de relations entre les mots (relations sémantiques, ou syntaxiques).
Les algorithmes des moteurs de recherche font appel ensuite à des méthodes statistiques, qui partent du principe que les termes sont indépendants entre eux d'un point de vue statistique. C'est en pratique une simple approximation, qui donne des résultats acceptables. Le problème c'est qu'il existe en fait des relations entre les termes, qui proviennent soit de la "musique de la langue" (c'est à dire en général de la grammaire) soit du sens des mots qui se retrouvent logiquement associés dans un texte sur une thématique donnée.

Cela signifie que pour améliorer la qualité du moteur de recherche, il faut être capable d'identifier que certains groupes de mots sont effectivement liés entre eux, alors que d'autres ne le sont pas. Un tel moteur sera capable de reconnaître que la phrase *"certains quartiers avec des architectures modernes ont vu le jour à côté de la vieille ville"* ne parle pas de l'*"architecture moderne de la vieille ville"*. Les moteurs en général renverront sans problème le premier document sur la requête *"architecture moderne vieille ville"*.

Mais avant même de pouvoir résoudre des problèmes aussi subtils, un moteur avec un index à base de mots clés isolés risque de renvoyer des résultats non pertinents sur des cas beaucoup plus basiques.

Le problème des groupes de mots correspondants à des entités nommées

Parfois, le groupe de mots clés tapé dans la requête décrit une seule et même entité : Par exemple Michael Jackson, les Etats Unis d'Amérique, le Président de la République, Villeneuve la Garenne, la première dame. Pour éviter de créer du "bruit" dans le système, il est intéressant de savoir identifier la présence de chacun de ces groupes dans le texte, sous la forme d'une expression figée (les mêmes termes dans le même ordre). Cela évitera de sortir dans les résultats une page qui parle de Michael Jordan et de Paul Jackson en réponse à la requête "Michael Jackson".

Les expressions figées

Il existe aussi dans la langue de nombreuses expressions figées : les mots se suivent, toujours dans le même ordre. En fait, même en l'absence de graphies comme des tirets dans le cas de mots composés (ex : rez-de-chaussée), ces groupes de mots se comportent comme une unité linguistique unique : par exemple "carte blanche", "voeu pieux", "à bâtons rompus". Evidemment, si un moteur ne sait pas reconnaître ce type d'expressions, il renverra dans ses pages de résultats un grand nombre de pages hors sujet.

Les obstacles à l'indexation des syntagmes

Il semble donc intéressant de passer à un système capable d'indexer des groupes de mots (syntagmes) au lieu de simples mots clés isolés. Pendant très longtemps, les méthodes d'indexation de syntagmes ont buté sur un écueil considérable : lorsque l'on indexe des groupes de mots, au lieu de mots isolés, la taille de l'index explose littéralement. La mémoire nécessaire pour identifier les combinaisons de trois, quatre, cinq mots est également un obstacle.

Par exemple, si l'on part du principe que l'on veut indexer toutes les combinaisons de cinq mots, et que le corpus (l'ensemble des textes à indexer) contient 200 000 termes différents, dans ce cas on aura 3,2x10^26 syntagmes possibles (un 3 suivi de... 26 zéros). Ce chiffre dépasse les capacités de tout système existant, et même imaginable.

Dans la pratique, toutes les combinaisons de syntagmes ne sont pas utiles dans l'index. Voici un exemple donné par Google, à propos d'une base de textes issus de pages web qu'il met à disposition pour les recherches des linguistes (base Web 1T 5-gram) :

- La base de départ contient : 1 024 908 267 229 de tokens (c'est-à-dire de termes, éventuellement présents plusieurs et même de très nombreuses fois dans l'ensemble des textes (corpus).
- Le nombre d'unigrammes (termes uniques pris isolément) s'élève à 13 588 391.
- Le nombre de bigrammes (couples de termes présents) s'élève à 314 843 401.
- Le nombre de trigrammes (triplets de termes présents) s'élève à 977 069 902.
- Le nombre de 4-grammes à 1 313 818 354 !
- Etc.

Dans cet exemple, le nombre ne suit pas une logique combinatoire : le nombre de n-grammes identifié correspond à des groupes de mots constituant des séquences non pas aléatoires mais qui présentent une certaine fréquence d'apparition en commun (fréquence de cooccurrence ou plus précisément fréquence de "collocation"). Cela reste malgré tout intéressant de noter que l'on est obligé de multiplier le nombre de lignes dans l'index par 100 pour parvenir à stocker des expressions contenant jusqu'à 4 mots.

L'indexation des syntagmes facilitée par les progrès de l'infrastructure chez Google

Il est intéressant de noter à ce stade que les premiers brevets et articles sur l'indexation des syntagmes sont apparus chez Google après le dernier changement d'infrastructure majeur avant Caffeine, à savoir l'infrastructure Teragoogle (plus connue sous le nom qui lui a été donnée par les observateurs extérieurs : "BigDaddy". Teragoogle a été implémenté en 2006). Teragoogle a été conçu par Anna Patterson (qui depuis a quitté Google pour lancer le moteur de recherche Cuil). Teragoogle a permis à Google de multiplier par dix le nombre de pages présent dans l'index (passant d'un index primaire estimé à quelques milliards de pages à un index comptant des dizaines de milliards de pages). On peut supposer aussi que la nouvelle infrastructure Caffeine peut, potentiellement, offrir à Google l'opportunité d'aller plus loin encore dans la quantité d'informations traitées dans son index.

L'indexation des syntagmes reste malgré tout fortement limitée par le problème du volume d'informations à traiter. Outre les changements d'échelle dans l'infrastructure qu'une telle approche impose, la recherche dans un index plus gros est également un défi du point de vue du temps de réponse des moteurs.

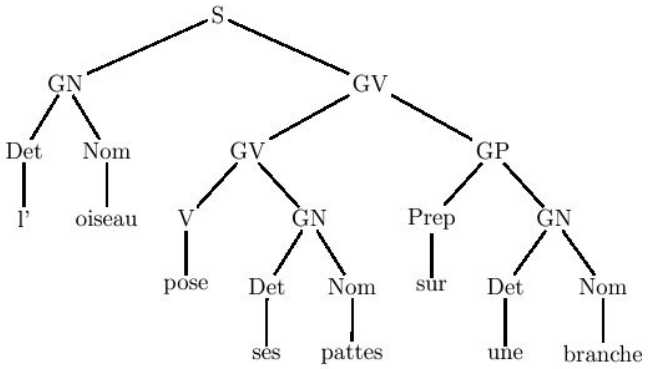
Pour éviter une inflation inutile du volume de données à traiter, il est indispensable de chercher à trier, parmi les syntagmes, ceux qui méritent une indexation, et ceux qui peuvent être ignorés du système (parce que leur identification dans les requêtes et dans les documents n'améliore pas la qualité des résultats retournés).

Les Méthodes d'indexation des syntagmes

Il existe deux approches classiques pour arriver à isoler les syntagmes utiles.

Les méthodes syntaxiques

La première approche consiste à analyser la syntaxe de la phrase, pour isoler des groupes de mots ayant des fonctions grammaticales différents (classiquement, on va chercher à isoler des syntagmes verbaux ou des syntagmes nominaux). Cette méthode a de nombreux avantages, notamment parce qu'elle permet de rattacher les termes dans les documents au bon groupe de mots lorsque le rattachement est ambigu.

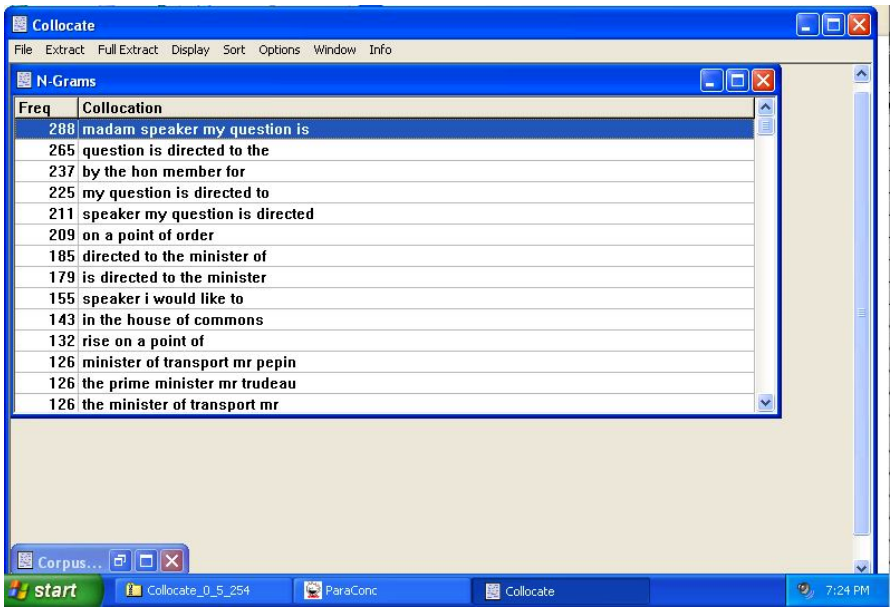


Un arbre syntaxique. L'une des nombreuses méthodes utilisées dans l'analyse syntaxique

Par contre, cette méthode peut s'avérer lourde et lente. Il n'est pas facile non plus de créer des analyses syntaxiques efficaces pour toutes les langues de la Terre, ce qui pose un problème lorsque l'on veut développer un moteur aussi universel que Google. Par ailleurs, le web est rempli de pages contenant des syntagmes isolés, et qui ne sont pas contenus dans des phrases de type sujet-verbe-complément. Dans ce type de cas, l'approche syntaxique n'apporte pas grand chose de plus.

Les méthodes statistiques

L'approche statistique, quant à elle, aà l'avantage de fonctionner dans plusieurs langues et de permettre de repérer des termes avec une grande fréquence de collocation même dans des pages web pauvre en contenu rédigé. Elle permet aussi de repérer des "pics" locaux . Par exemple "moyenne pondérée" est une expression figée que l'on rencontre dans des textes parlant de statistiques, ailleurs les deux termes peuvent être présents mais séparés.

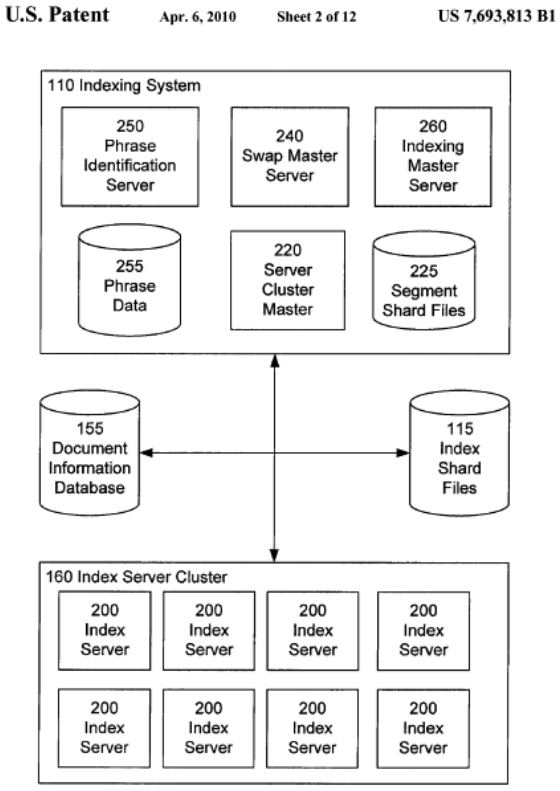


Un écran d'un outil d'analyse de fréquences de syntagmes dans un corpus : Collocate

L'Approche décrite dans le brevet de Google

Un brevet attribué à Google et datant de 2007 a été publié il y'a quelques semaines. La méthode d'indexation des syntagmes décrite dans ce brevet est clairement plutôt une méthode statistique.

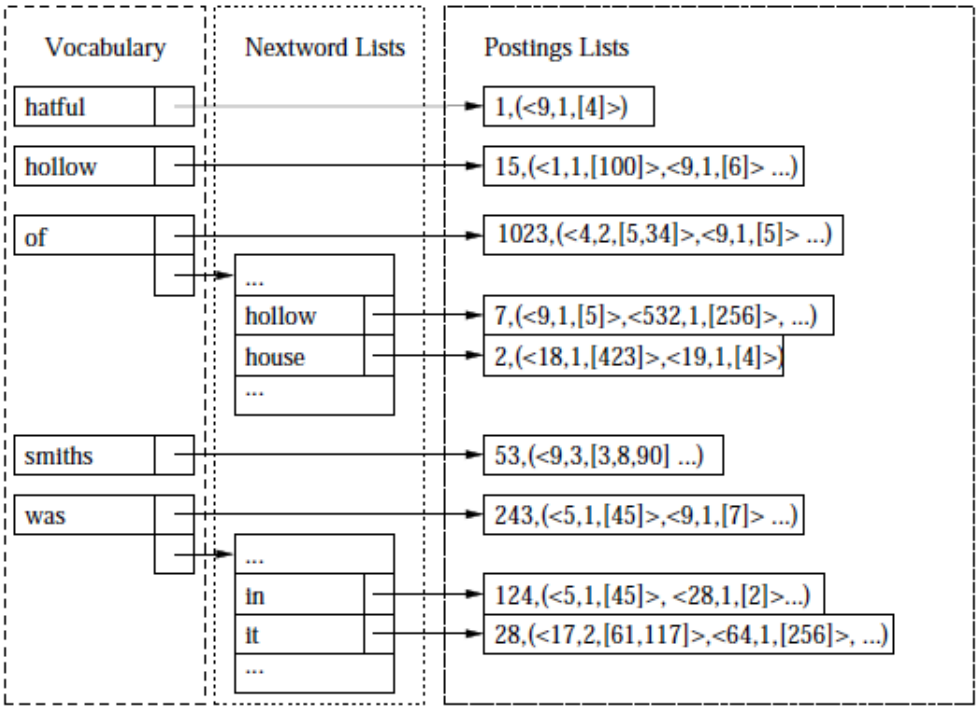
La première étape consiste évidemment à identifier la présence d'un syntagme dans la requête, et non d'une suite de mots clés indifférenciés. Il est clair que dans la requête "nouveau président de la Cour des Comptes", l'utilisateur cherche le nom du nouveau président de la Cour des Comptes, et non une page qui contient les termes "Comptes, cour, président et nouveau". La méthode consiste à découper la requête en syntagmes possibles (méthode des n-grammes ?), à mesurer la probabilité de trouver ces syntagmes dans les documents, à comparer les scores des différents syntagmes, pour finalement, moyennant quelques ajustements, décider si un ou plusieurs syntagmes utiles ont pu être identifié dans la requête.



Une fois un syntagme ainsi isolé, il faut pouvoir le retrouver dans l'index. Cela signifie que ce syntagme doit avoir une entrée dans l'index, et avoir été traité par le processus d'indexation.

L'identification des syntagmes à indexer se fait à partir de l'analyse des mesures de fréquence de cooccurrences en tenant compte de l'ordre des mots.

Les brevets cités dans la bibliographie exposent différentes méthodes possibles pour parvenir à bâtir un tel index, et l'architecture qui va avec. On ne sait pas avec certitude aujourd'hui si Google utilise un index de syntagmes, mais il semble néanmoins que cette approche soit devenue techniquement possible, même à l'échelle de Google.



Ci-dessus : une des solutions possibles. Un index normal combiné avec un index de syntagmes. L'index de syntagmes fonctionne avec une logique de prédétermination du mot suivant qui accélère l'extraction de l'information utile dans l'index.

Qu'est-ce que cela change d'un point de vue SEO ?

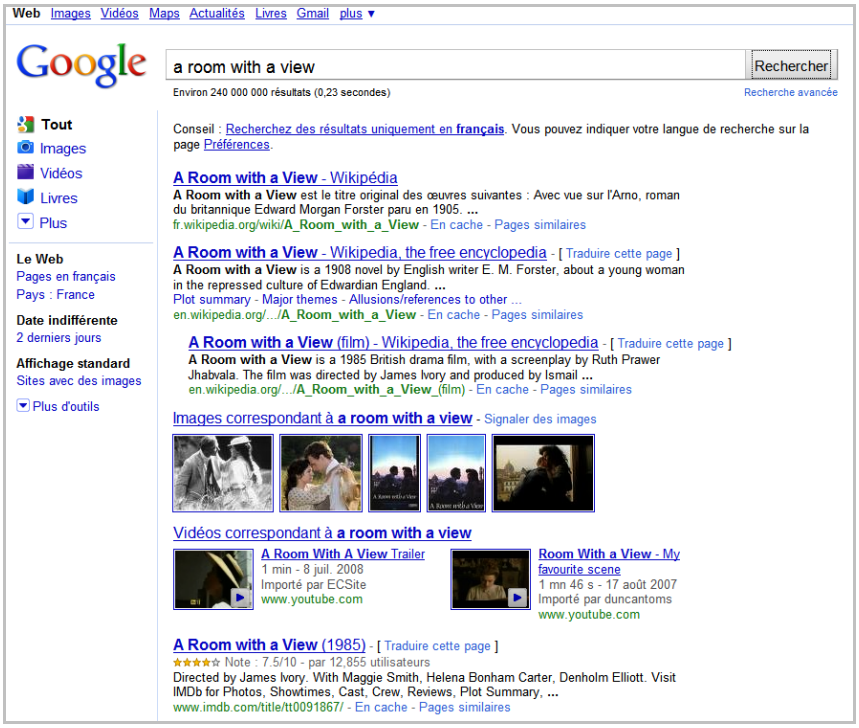
Un moteur de recherche qui ne travaille plus en "sac de mots" mais avec des syntagmes classe les résultats de manière très différente sur certaines requêtes.

Cela signifie que l'**ordre des termes** commence à avoir une importance bien plus grande, et qu'à l'inverse les méthodes de type "bourrage de mots clés" fonctionnent beaucoup moins bien. Dans certains cas où on peut s'attendre à ce que les utilisateurs tapent des **syntagmes fréquents** dans une requête, on peut aussi s'attendre à ce que le système classe en premier les pages qui contiennent ce syntagme ou une variante très proche. Cela obligerait à reconsidérer parfois de manière drastique les règles d'écriture des balises <title>, <Hn>, etc.

Les Évolutions à Attendre

Il est difficile de savoir si Google ou Bing utilisent déjà cette possibilité. Même un moteur qui fonctionne en "sac de mots" peut être sensible à la distance entre les mots et à l'ordre des mots.

Ce qui est clair, c'est que certaines fonctionnalités comme la reconnaissance des entités nommées aboutissent déjà à de la reconnaissance de syntagmes. Certains indices démontrent également que Google est déjà capable de travailler avec certains syntagmes. Par exemple, la requête "*a room with a view*" renvoyait, il y a trois ans, le même type de pages que "*room with view*". (le "a" est un mot vide, en théorie, non pris en compte dans la requête). Aujourd'hui, les réponses renvoyées correspondent clairement au roman et au film, et ne renvoient plus de pages de sites immobilier :



A room with a view : Google ne renvoie sur la première page que des résultats correspondant au syntagme complet "A room with a view", qui correspond à un film ou au roman éponyme. Il y a quelques années, des pages contenant "room" et "view" auraient aussi étaient renvoyées dans les premières positions.

Il est donc tout à fait possible que Bing, Google ou Yahoo n'utilisent pas ou peu ces techniques aujourd'hui. Mais la tendance lourde des moteurs de recherche est malgré tout de chercher à améliorer la précision de leurs résultats. Pour cela, il est indispensable de mieux comprendre l'intention de l'utilisateur exprimée par la requête. La reconnaissance de syntagmes dans la requête est une étape clé du processus, notamment s'il s'agit d'une entité nommée

(reconnaître que Villeneuve la Garenne est une ville, et qu'on ne cherche pas de pages parlant de lapins de garenne). La nécessité d'indexation des syntagmes apparait simultanément.

On peut donc s'attendre à voir de plus en plus d'applications de ces techniques dans les moteurs de recherche. C'est une étape importante, car beaucoup de traitements autour de la sémantique deviennent possible à partir de tels index. Mais c'est une étape difficile, qui nous amène à franchir encore un seuil dans la monstruosité des infrastructures nécessaires pour construire un moteur de recherche capable de rivaliser avec un Bing ou un Google...

BIBLIOGRAPHIE

Liens utiles
Un outil pour extraire des n-grammes dans un corpus de textes anglais :
http://www.kwicfinder.com/BNC/explore.html

Collocate, un outil d'analyse de fréquences d'apparition de syntagmes
http://www.athel.com/colloc.html

BREVETS

Detecting spam documents in a phrase based information retrieval system
Invented by Anna Lynn Patterson
US Patent Application 20060294155
Published December 28, 2006
Filed: June 28, 2006
http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.html&r=1&f=G&l=50&s1=%2220060294155%22.PGNR.&OS=DN/20060294155&RS=DN/20060294155

Integrating External Related Phrase Information into a Phrase-based Indexing Information Retrieval System
Invented by Anna L. Patterson
Assigned to Google
US Patent Application 20090070312
Published March 12, 2009
Filed September 7, 2007
http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.html&r=1&p=1&f=G&l=50&d=PG01&S1=20090070312.PGNR.&OS=dn/20090070312&RS=DN/20090070312

Autres brevets attribués à Google et/ou à Anna Patterson sur les mêmes sujets :

- *Multiple index based information retrieval system* (20060106792)
Assigned to Google
http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.html&r=1&f=G&l=50&s1=%2220060106792%22.PGNR.&OS=DN/20060106792&RS=DN/20060106792

- *Phrase-based searching in an information retrieval system* (20060031195)
Assigned to Google
http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.html&r=1&f=G&l=50&s1=%2220060031195%22.PGNR.&OS=DN/20060031195&RS=DN/20060031195

- *Phrase-based indexing in an information retrieval system* (20060020607)
http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.html&r=1&f=G&l=50&s1=%2220060020607%22.PGNR.&OS=DN/20060020607&RS=DN/20060020607

- *Phrase-based generation of document descriptions* (20060020571)
http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.html&r=1&f=G&l=50&s1=%2220060020571%22.PGNR.&OS=DN/20060020571&RS=DN/20060020571

- *Phrase identification in an information retrieval system* (20060018551)
http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.html&r=1&p=1&f=G&l=50&d=PG01&S1=20060018551.PGNR.&OS=dn/20060018551&RS=DN/20060018551

- *Index server architecture using tiered and sharded phrase posting lists*
Invented by Pei Cao, Nadav Eiron, Soham Mazumdar, Anna Patterson, Russell Power, and Yonatan Zunger
Assigned to Google
US Patent 7,693,813
Granted April 6, 2010
Filed March 30, 2007
http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,693,813.PN.&OS=pn/7,693,813&RS=PN/7,693,813

Philippe Yonnet, ***Global SEO Strategist, WEB DMUK (Londres) – Easyroommate / Vivastreet***

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://blog-abonnes.abondance.com/2010/06/lindexation-des-syntagmes-la-fin-du.html