Google utilise-t-il vraiment la méthode LDA *(Latent Dirichlet Allocation)* dans son algorithme ?

Retour au sommaire de la lettre

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

La méthode LDA (pour Latent Dirichlet Allocation) est une méthode qui permet de détecter et d'isoler des concepts et des relations sémantiques entre différents termes dans les documents. Elle a supplanté depuis plusieurs années LSI (Latent Semantic Indexing) souvent présentée - à tort - comme au coeur de l'algorithme de Google. Cet article a pour but de vous présenter LDA et pose la question de son éventuelle utilisation par le moteur de recherche leader, et bien sûr des implications que cela amène en SEO / référencement naturel...

Il y a bientôt cinq ans, j'avais eu l'occasion de dénoncer l'utilisation de la méthode LSI *(Latent Semantic Indexing)* comme argument de vente par des agences SEO, essentiellement indiennes et américaines. LSI était une méthode permettant d'isoler, au milieu du "bruit", les relations sémantiques entre des termes. Or l'utilisation de LSI par Google semblait peu crédible, compte tenu des inconvénients et des limites de la méthode.

Une autre méthode l'a supplantée assez vite (dès 2003) : la méthode LDA *(Latent Dirichlet Allocation)*. Or, LDA fournit des résultats beaucoup plus facile à réutiliser dans un algorithme de moteurs, et l'utiliser à grande échelle est envisageable. Ce qui est clair, c'est que LDA est réellement utilisée aujourd'hui pour des applications d'extraction d'information, en particulier sur la découverte des "sujets" abordés dans les documents.

SEOMoz a largement contribué à populariser LDA l'année dernière en annonçant que les classements de Google et LDA étaient remarquablement corrélés (http://www.seomoz.org/blog/lda-and-googles-rankings-well-correlated). Nous verrons plus loin ce que l'on doit penser de cette affirmation.

Mais commençons d'abord par rappeler ce que ce sont ces méthodes de calcul, et à quoi elles servent exactement.

Latent Semantic Indexing : une méthode déjà obsolète en 2005

La méthode LSI (aussi parfois appelée LSA), est toujours présentée par des agences, soit comme la clé de l'algorithme de Google (*"Google utilise LSI pour classer ses résultats, donc nous allons optimiser votre site pour l'algorithme LSI"*), soit comme mot magique pour valider que leurs méthodes de référencement sont "à la pointe" (*"nos algorithmes sophistiqués utilisent l'algorithme LSI pour calculer les optimisations appropriées"*). Soyons clairs : ces arguments sont fallacieux, et relèvent même dans certains cas de la tromperie manifeste (la méthode LSI n'est pas vraiment utilisée pour déterminer le contenu optimisé, on cherche juste à adapter le texte au contexte pour qu'il réponde à ce que l'on a compris de la méthode LSI).

Cette méthode présente trois inconvénients majeurs : tout d'abord, ses résultats sont difficiles à interpréter, et les calculs sont particulièrement coûteux s'ils doivent être réalisés à grande échelle. Mais le principal inconvénient de LSI est sans doute son incapacité à traiter correctement les cas de polysémie (si un mot à plusieurs sens, ces "sens" se retrouvent mélangés dans les corrélations). Ce qui ne plaidait pas pour une intégration de cette méthode dans l'algorithme de Google.

LSI était en fait déjà une méthode dépassée en 2005, au moment où certains gourous du SEO l'ont promue au rang de dernier gadget à la mode pour le référencement.

A chaque fois qu'une nouvelle technique est inventée dans le domaine de l'extraction d'information, celle-ci finit toujours par être présentée comme le nouvel algorithme de Google (ou au moins le rouage essentiel qui explique les nouveaux comportements). LDA ne faillit pas

à la règle... Mais commençons tout d'abord par étudier à quoi sert cet algorithme et comment il fonctionne.

LDA : une méthode statistique pour découvrir des classes de termes dans un document

L'objectif de la méthode LDA est de découvrir des "groupes" cachés (latents) à l'intérieur des documents. Ces groupes ou classes sont découverts par des méthodes statistiques en calculant des corrélations entre des évènements comme l'apparition de termes particuliers dans les documents.

Contrairement à la méthode LSI, le modèle utilisé est "probabiliste", c'est à dire que l'on essaie de déterminer une "probabilité" d'appartenance à un groupe ou à une classe. Chaque "classe" représente en réalité, lorsque la méthode est appliquée à des textes, les différents "sujets" abordés dans le document.

Voici un exemple simple d'analyse effectuée avec la méthode LDA par l'excellente équipe d'Antidot :

Topic 92	Topic 103	Topic 68	Topic 48	Topic 90
Renault	Numérique	Haïti	Barack Obama	Scrutin
Flins	Apple	Port au Prince	Démocrates	Réforme
Clio	Google	Séisme	Massachusetts	Collectivités
Carlos Ghosn	Mobile	Haïtiens	Sénat	Conseiller territorial
Usine	Microsoft	ONU	Scott Brown	Élus
Turquie	Opérateurs	Humanitaire	Républicains	Sénat
Bursa	Taxe	Blessés	Maison blanche	Départements
Production	Iphone	Secours	Ted Kennedy	Collectivités locales
Automobile	ARCEP	Morts	Congrès	

Extrait de : http://blog.antidot.net/2010/02/10/latent-dirichlet-allocation/

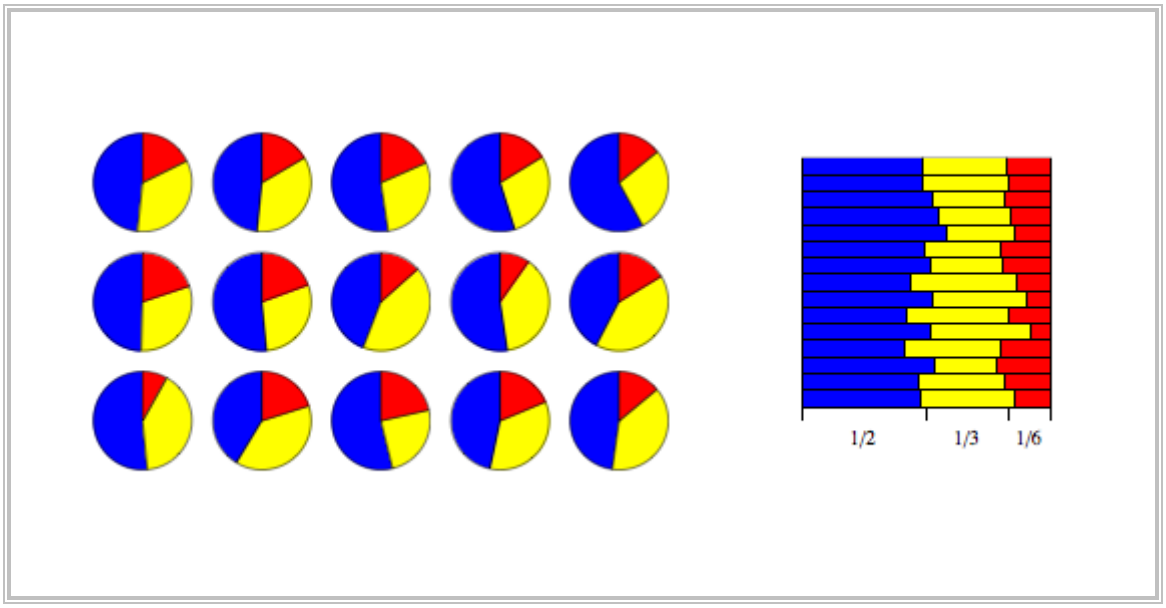
Quelque chose saute immédiatement aux yeux en lisant ce tableau : les "sujets" (topics) ne sont pas nommés ! C'est en fait l'un des avantages majeurs de la méthode : les "classes" regroupant les termes sont découvertes automatiquement. Un humain parviendrait à "nommer" le sujet ainsi identifié, mais la méthode LDA ne va pas jusque-là.

On voit aussi que les termes peuvent apparaître dans plusieurs classes. C'est logique, puisque le même mot peut avoir plusieurs sens, et donc être associé, suivant le contexte, à des sujets différents. Cela fonctionne y compris dans les cas où la différence de sens est subtile : le terme Sénat est identifié comme appartenant à deux sujets différents, selon qu'il fait allusion au sénat américain, ou à sa version française !

Allocation de dirichlet latente ? Qu'est-ce que ça veut dire ?

Nous avons déjà évoqué la signification du terme "latente" : la méthode permet de découvrir automatiquement des "sujets" (cachés, "latents") au milieu des textes d'un corpus. Mais que vient faire "dirichlet" dans cette histoire.

En fait, c'est une allusion à une méthode de calcul utilisée en probabilités, et dérivée d'une découverte du mathématicien allemand Johann Peter Gustav Lejeune Dirichlet. En fait, avec la méthode LDA, la distribution des classes (des sujets) est supposée respecter une loi *a priori* correspondant à une distribution de Dirichlet. Pour les spécialistes, précisons que cette approche s'inscrit dans les méthodes probabilistes bayésiennes.



Ci-dessus une illustration de la distribution de Dirichlet. Lorsque l'on désire découper des ficelles (toutes de longueur initiale 1.0) en K pièces de différentes longueurs, où chaque pièce a, en moyenne, une longueur désignée mais qui peut varier , on obtient les résultats ci-dessus qui obéissent à loi de Dirichlet.

Pourquoi LDA est plus efficace que LSI ?

La méthode LDA repose sur un modèle mathématique probabiliste beaucoup plus solide que la méthode LSI/LSA. Un certain nombre d'hypothèses de travail sous-jacentes dans la méthode LSI étaient empiriques, sans fondement mathématique réel.

En fait, une autre approche (probabiliste celle-là) a succédé directement à la méthode LSI : la méthode pLSI *(probabilistic Latent Semantic Indexing)*. Mais pLSI (qui ressemble beaucoup à LDA) présente toujours quelques inconvénients majeurs. Tout d'abord l'analyse isolée d'un document nouveau est difficile. Ensuite, elle n'est pas parfaitement "scalable" : elle s'appuie sur un corpus (un groupe de documents) d'entrainement, et plus ce corpus est grand, plus le calcul devient lent et complexe. Pourtant la taille du corpus détermine la finesse de la méthode.

LDA est une "généralisation" de pLSI, qui permet au contraire d'évaluer des documents nouveaux sans difficulté, et reste "scalable".

LDA part du principe qu'un document contient un mélange de différents sujets, ce qui produit des résultats beaucoup plus utiles dans la pratique et plus proches de la réalité.

Peut-on utiliser LDA dans un algorithme de classement de moteur de recherche ?

Evidemment, on peut se demander - légitimement, ici - si LDA peut servir dans l'algorithme de classement d'un moteur de recherche.

Dès lors que l'on considère une requête comme un document (c'est un texte très court), on peut essayer de déterminer si les classes auxquelles appartiennent les termes de la requête et les termes des pages à classer sont similaires.

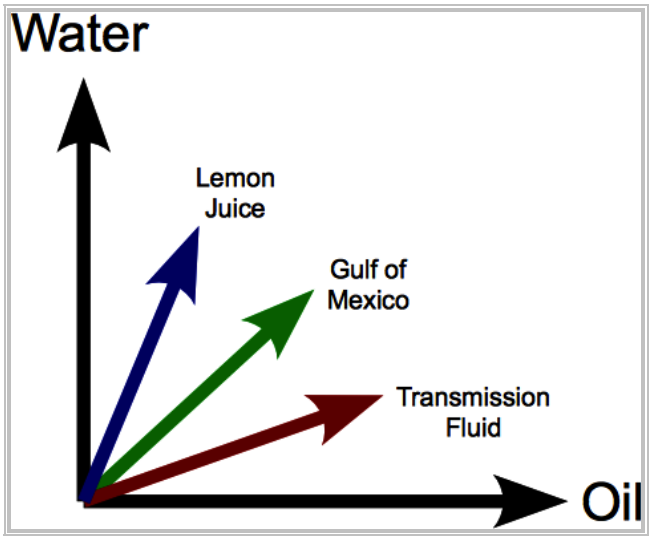
Cette approche a un avantage : comme LDA est une méthode excellente pour "séparer" les différents sens des termes en les reliant à des "sujets" différents, cette approche permet d'augmenter la pertinence des réponses en éliminant les documents qui, certes, contiennent les termes demandés, mais dans un autre contexte.

Mais en dehors de cette capacité de "désambiguïsation", LDA améliore-t'il systématiquement la pertinence des résultats ? Absolument pas. Cette approche est intéressante si on essaie de répondre à la question "trouve moi des documents traitant du même sujet que...", mais si la demande est "montre- moi les documents les plus pertinents sur la requête "kw1 kw2 ... kwn", des approches différentes voire plus classiques sont bien plus efficaces et génèrent, en particulier, beaucoup moins de bruit (le bruit est constitué de documents renvoyés parce qu'évalués comme pertinents par le système, mais que l'utilisateur final jugera non pertinents).

En outre, si l'algorithme LDA est plus "scalable", il reste relativement coûteux en ressources, et certaines applications ont encore un ratio coût/bénéfices défavorable par rapport à des approches plus classiques.

Ce qui signifie que si LDA est utilisé, c'est probablement en tant que signal secondaire, en combinaison avec d'autres critères, et sans doute sous une forme assez différente, puisque la réutilisation "brute" des données produites par la méthode reste peu efficace à des fins de classement.

L'une de ces méthodes alternatives possibles est le *cosinus LDA*, qui reprend les principes du Cosinus de Salton : au lieu de définir la position d'un document dans l'espace vectoriel des termes, on fait la même chose dans un espace vectoriel de sujets. Pour Salton, les coordonnées sur chaque axe sont déterminées par leur poids tf*idf, pour le cosinus LDA, par la probabilité d'appartenance à une classe/sujet donnée.



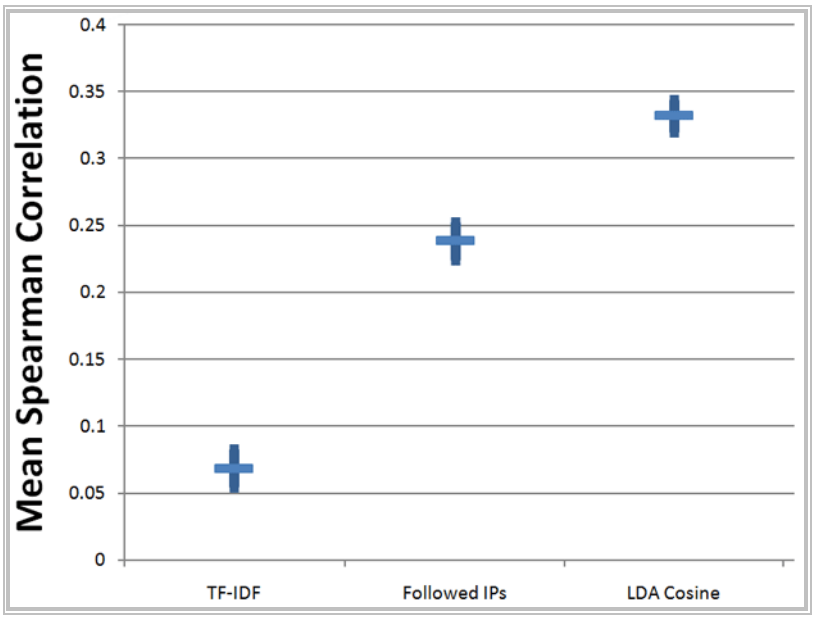
Une illustration ultra-simplifiée du cosinus LDA : dans un espace à deux dimensions, les coordonnées de trois documents sont calculées projetées sur deux axes correspondant à deux sujets : l'huile et l'eau (oil et water). Un document qui parle de "fluide de transmission" est plus proche de l'huile que de l'eau, quant au document qui parle de jus de citron, il est effectivement plus proche de l'eau que de l'huile. La mesure du cosinus de l'angle permet de mesurer une distance angulaire et indirectement la proximité thématique des documents.

Pourtant, on l'a vu, l'équipe de SEOMoz, réputée sérieuse dans le petit monde du SEO, a trouvé une forte corrélation entre LDA et les classements observés dans les pages de résultat de Google ? Mais qu'en est-il véritablement ?

LDA et l'algorithme de classement de Google : ce que révèle vraiment l'étude de SEOMoz

Le 6 septembre 2010, SEOMoz publie sur son blog un billet qui secoue le Landernau du SEO (enfin ceux qui suivent l'actualité des sites anglo saxons de SEO) : *"LDA et les classements de Google sont remarquablement bien corrélés"* (http://www.seomoz.org/blog/lda-and-googles-rankings-well-correlated).

Ben Hendrickson dans son étude a comparé les niveaux de corrélation pour trois approches :
- tf*idf ;
- le nombre de backlinks provenant d'ip différentes (en éliminant les nofollow) ;
- le cosinus LDA (décrit précédemment).



En réalité, les résultats de cette étude étaient faux : Ben Hendrickson a reconnu ensuite que le coefficient de corrélation trouvée n'était pas égal à 0.32, mais à 0.17 ! Ce qui veut dire que selon leurs propres tests, le poids des backlinks est toujours un meilleur facteur d'explication des classements que LDA.

De toute manière, cette étude ne démontrait en réalité pas grand chose.

Déjà, on compare des choses de dimension différentes (en langage commun, on compare des choux avec des carottes). Tf*idf est un poids, comme LDA. Il aurait fallu comparer le cosinus de Salton avec le cosinus LDA. En fait, SEOMoz ne nous dit pas comment ils calculent le tf*idf d'un document. C'est ennuyeux, car ce n'est pas très habituel de calculer un poids tf*idf pour un document.

Ensuite, en ne testant que deux critères "on page" et un critère "off page", on a peu de chances de découvrir quelle méthode est réellement utilisée. Il faudrait pouvoir tester d'autres critères et essayer de trouver des coefficients de corrélation significatifs.

Enfin, on trouve tous les biais habituels des études de ce type réalisées par des sociétés de référencement : elles ont l'air d'études faites "en condition de laboratoire". Elles ont l'air d'obéir à des protocoles scientifiques. Mais en réalité, leurs conclusions reposent sur du sable. "Corrélation n'est pas raison" : trouver une corrélation forte entre l'évènement A et l'évènement B ne signifie pas que A est la cause de B. D'ailleurs une forte corrélation n'est en soi pas forcément une donnée significative.

Déjà B peut être la cause de A (en l'occurrence l'algorithme de Google peut favoriser l'apparition de pages proches du sujet de la requête dans ses résultats, et qui donc présente un fort cosinus LDA, sans que LDA soit utilisé !)

Ensuite, A et B peuvent être causés par C. Par exemple, si on trouve une forte corrélation entre un bon classement sur un mot clé et la présence de ce mot clé dans le nom de domaine, cela signifie-t'il qu'il faut avoir le mot clé dans le nom de domaine ? Apparemment oui, mais le doute s'installera vite quand on aura observé aussi que :

- la corrélation est forte également entre présence des mots clés dans les anchor texts des backlinks, et bonnes positions sur ces mots clés ;

- que si le mot clé est présent dans le nom de domaine, les anchor texts contiennent aussi fréquemment le mot clé.

Donc qu'elle est la véritable cause de ces bonnes positions ? La présence des mots clés dans les anchor texts, ou dans les noms de domaine ? En fait ces études ne le disent pas.
Ces études statistiques sont parfois faites avec rigueur et scientificité (mais pas toujours, et les protocoles et méthodes utilisés sont rarement complètement rendus publics). Ce qui par contre pose toujours problème, ce sont les conclusions qui en sont tirées. Trouver des corrélations permet de trier entre les relations de causalité possibles et celles qui sont improbables. Mais jamais de conclure sur l'existence réelle d'une relation de causalité.

Prenons date : quelles sont les candidats pour le prochain algorithme à la mode chez les agences SEO ?

Voici un échantillon des méthodes avancées en traitement du langage naturel, et qui n'ont pas encore été découvertes ou exploitées par les gourous SEO. Si on vous les présente demain comme l'explication "évidente" du comportement du prochain algo de Google, au moins vous saurez que ce n'est pas une nouveauté :

- MASHA ("*Multinomial ASymmetric Hierarchical Analysis*") et HPLSA ("*Hierarchical Probabilistic Latent Semantic Analysis*") qui sont des avatars plus sophistiqués de la méthode pLSI/pLSA.
- la géométrie de diffusion (notamment parce Stéphane Lafon travaille chez Google maintenant).
- la NMF : *non négative matrix factorization.*
- les projections aléatoires (*random projections*).
- *Hidden Topic Markov Models*.
- ... ?

Quelles réelles conséquences a l'existence de LDA sur les stratégies de SEO ?

En fait peu importe que LDA soit réellement implémentée par Google dans son algorithme. Ces avancées technologiques indiquent clairement la direction vers laquelle évoluent le traitement du langage naturel et les moteurs de recherche. C'est à dire vers :

- un traitement plus efficace des informations sémantiques, notamment à des fins de désambiguïsation des requêtes et des termes contenus dans les documents.
- l'analyse des "sujets" abordés dans les documents, en plus de l'analyse des termes présents dans les documents (une approche que l'on appelle le "topic model").

Mais LDA n'a pas plus de conséquences sur la manière d'optimiser une page que d'autres techniques qui vont dans le même sens.

Par exemple, l'analyse syntaxique (grammaticale) des phrases d'un document, combiné avec un étiquetage sémantique permet aussi de résoudre les cas de polysémie. L'indexation des phrases peut également permettre de sélectionner correctement des documents parlant du "bon sujet" (par exemple, des documents parlant du président de la République française, sur la requête "Président de la République" qui est reconnue comme une expression monolithique, et non la combinaison de "président" et "république")

Cela signifie donc qu'il sera de plus en plus difficile de positionner une page sur la requête "baladeur mp3" juste en truffant la page d'occurrences de ces termes. Il faut avoir une page qui traite des baladeurs mp3 ou qui vend des baladeurs mp3. Cela signifie également qu'il sera plus difficile de sortir sur des requêtes "par accident", notamment grâce à des mots ayant plusieurs sens.

Cela veut-il dire qu'il ne faut pas s'intéresser à ces nouvelles méthodes ? En tant que source d'information sur le fonctionnement de Google sans doute. Savoir si elles sont implémentées

ou non ne donne aucune indication opérationnelle sur la manière d'optimiser son site. Elles sont par contre très intéressantes :

- pour améliorer un moteur de recherche interne ;
- ou pour créer des suggestions pertinentes sur ces pages : "ces autres pages parlent du même sujet".

N'oublions pas que toute avancée dans le traitement du langage, le data mining et l'analyse textuelle peut aussi bien servir à Google qu'aux webmasters...

Bibliographie

Billet sur le blog de Sébastien Billard : "Ne prenez pas LSI pour des lanternes"
http://s.billard.free.fr/referencement/?2006/10/09/296-ne-prenez-pas-lsi-pour-des-lanternes-par-philippe-yonnet

Article de Joseph O'Day
http://www.betternetworker.com/articles/view/marketing/seo/boiling-puppies-how-search-engines-really-work

Articles du blog antidot sur LDA
http://blog.antidot.net/tag/latent-dirichlet/

Critique de E. Garcia sur les méthodes statistiques utilisées par les agences SEO qui publient des études sur la corrélation entre critères et classements dans Google
http://irthoughts.wordpress.com/2010/11/08/on-statistical-significance-and-seo-statistical-%E2%80%9Cstudies%E2%80%9D/

L'article fondateur de la méthode : "Latent Dirichlet allocation"
Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John. ed. . Journal of Machine Learning Research 3 (4–5): pp. 993–1022.
http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf

***Philippe YONNET** , Directeur SEO international, Twenga.*

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://blog-abonnes.abondance.com/2011/09/google-utilise-t-il-vraiment-la-methode.html