L'apport de la sémantique et de la linguistique statistique pour le SEO

SEO Camp'us -4 et 5 février 2009

Philippe YONNET

Directeur du pôle métiers – Aposition

Président de l'association SEOCamp

Comment classer les pages : première idée

CALCUL DE TF * iDF

Exemple simplifié

Vers un meilleur critère de poids

tf*idf

tf = fréquence des termes dans le document

idf = inverse du nombre de documents dans lequel le terme est présent

Comment classer les pages : première idée

Nombre d'occurences du terme dans la page :

Poids = 1

Poids = 3

Comment classer les pages : première idée

Problème : les documents ne contiennent pas le même nombre de mots

Poids = 0,01

Poids = 0,003

Poids du terme = fréquence = « densité du mot clé »

Critère de poids retenu :
nombre d'occurrences
divisé par le nombre de mots du document

Comment classer les pages : première idée

Problème :
les mots n'ont pas la même fréquence d'apparition
dans la langue

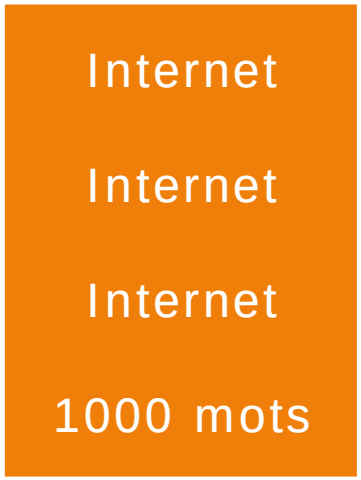
Combien de pages contiennent le mot clé **internet** d'après Google ?

2 110 000 000

Combien de pages contiennent
le mot clé **globicéphale** d'après Google ?

9 530

Exemple de calcul sans et avec tf*idf



Densité 3 pour mille

Densité 1 pour mille

Exemple de calcul sans et avec tf*idf

Internet

Internet

Internet

1000 mots

Index de Google
20 milliards de pages
(?)

10^9 pages

Globicéphale

1000 mots

DF[internet] =

2 x 10^9 / 20 x 10^9 = 0,1

DF[globicephale] =

10^4 / 20 x 10^9 = 5 x 10^-7

Exemple de calcul sans et avec tf*idf

Internet

Internet

Internet

1000 mots

Index de Google
20 milliards de pages
(?)

10^9 pages

Globicéphale

1000 mots

TF*iDF[internet] =

0,003 / 0,1 = 0,03

TF*iDF[globicephale] =

0,001/5 x 10^7= 2000 !

2000 >>>> 0,03

Pourquoi il faut abandonner la densité de mots clés

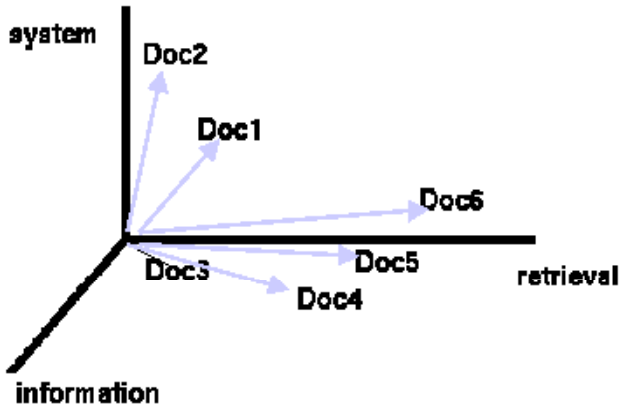
C’est un critère qui n’est plus utilisé par les outils de recherche

Pertinent que pour les requêtes à un seul terme

Induit le « keyword stuffing » facilement détectable

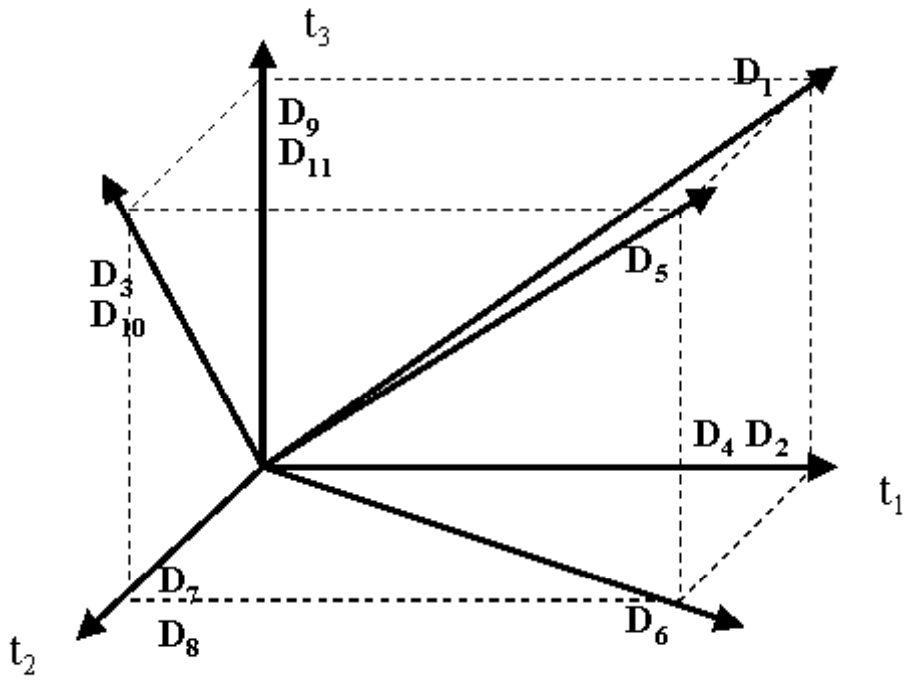
Le principe du Cosinus de Salton

Documents dans un espace à 3 dimensions :



Les documents proches dans l'espace
ont un contenu similaire

Le principe du Cosinus de Salton



En réalité, il y’a autant de dimensions que de “termes”
C’est un espace à n dimensions

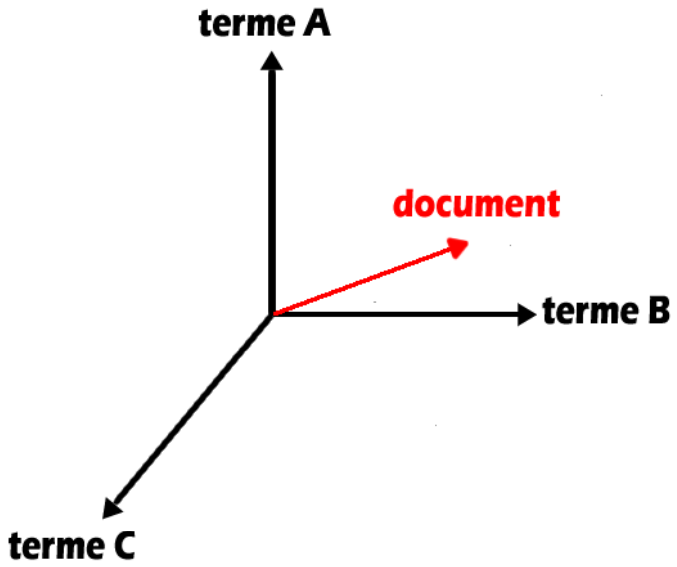
Le principe du Cosinus de Salton

CALCULER LE POIDS D'UN TERME
DANS UN DOCUMENT
tf*idf

Il existe de nombreuses formules différentes

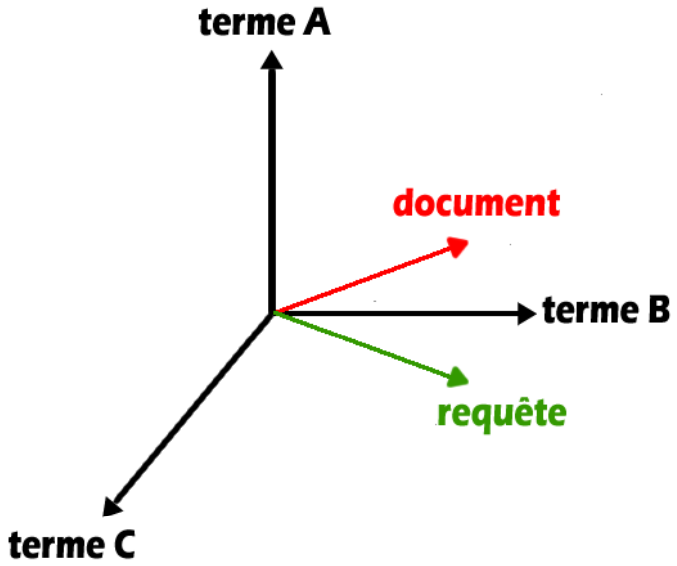
$$w_{ik} = \frac{tf_{ik} \log(N / n_k)}{\sqrt{\sum_{k=1}^{t} (tf_{ik})^2 [\log(N / n_k)]^2}}$$

Le principe du Cosinus de Salton



- Tout document peut être situé dans l'espace vectoriel de Salton, par un vecteur de coordonnées sur les axes correspondant à chaque terme de l'index

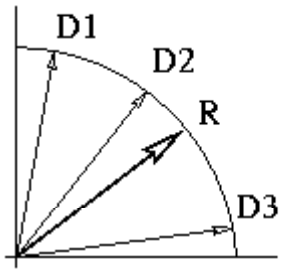
Le principe du Cosinus de Salton



Une requête est un document composé de quelques termes uniquement. Elle a donc aussi des coordonnées dans l'espace de Salton

Le principe du Cosinus de Salton

- Un calcul de distance (cosinus) entre la requête et les documents permet de classer les pages en fonction de leur proximité sémantique avec la requête



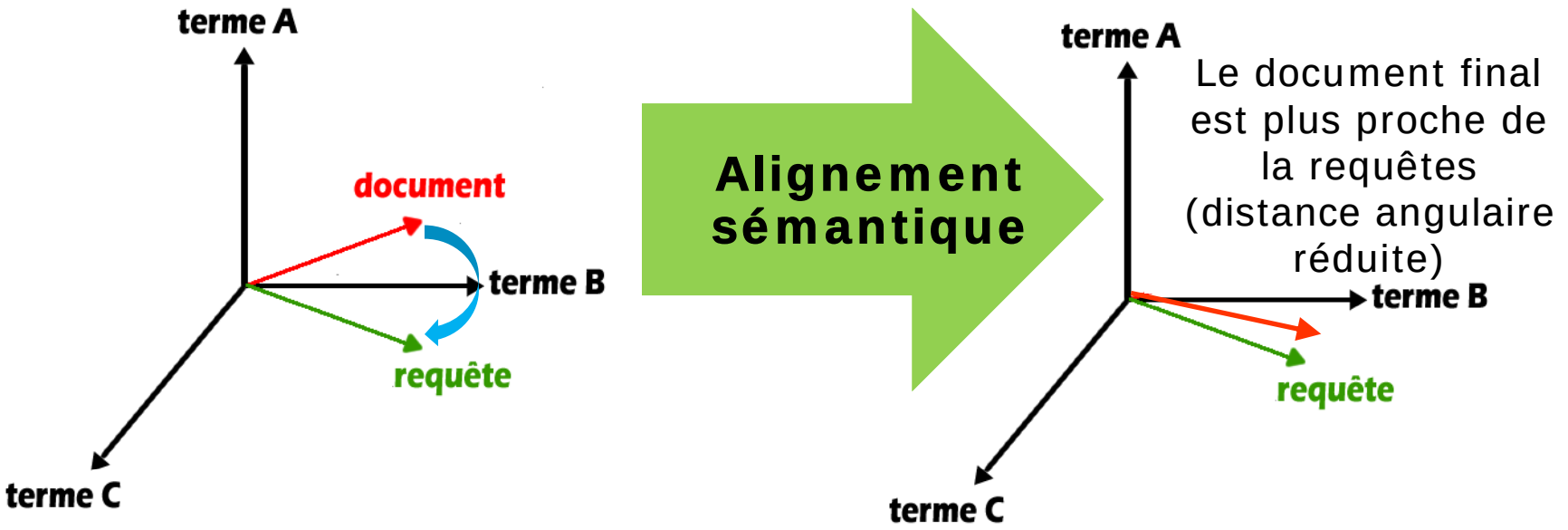
Trois documents et une requête dans le modèle d'espace vectoriel

D2 est le document le mieux classé

Les autres applications de la sémantique et de la linguistique statistique au référencement

Les bases théoriques de l'alignement sémantique

L'alignement sémantique consiste à changer le contenu textuel des pages pour les « orienter » comme la requête



L'enrichissement sémantique

Objectif : positionner des pages sur des requêtes secondaires non présentes dans la page à l'origine

Exemple : la page parle d'assurance voiture, mais pas de garantie sur un véhicule.

⇒Identifier les termes et expressions caractéristiques d'une thématique donnée
⇒Les outils habituels sont d'un faible secours sur des pages très techniques

La méthode : analyse d'un corpus de pages en rapport avec la thématique, avec analyse des fréquences et taux de cooccurrence

Nota bene : on peut utiliser les résultats d'un moteur pour avoir une première approche grossière

Les outils de normalisation

Problème souvent insoluble lorsque l'on traite avec des contenus générés par l'utilisateur.

Objectif : regrouper les pages parlant du même sujet pour un objectif de PR sculpting + d'ergonomie

EX : Beatles, The beatles, beetles, BEETLES, Beatels, Beatels (The)

⇒Comment tout regrouper dans une catégorie « The Beatles » ?

La solution : un outil de normalisation exploitant les infos de contexte

La génération de variantes graphiques, morphologiques

Comment, en fonction d'une liste de termes, générer toutes les variantes de graphie, toutes les variantes morphologiques ?

Masculin féminin : lion,lionne
Singulier, pluriel : lions, lion
Graphie : clé, clef, rez-de-chaussée, rez de chaussée
Conjugaison : est, être, suis, sommes, étions
Combinaison de termes : assurance voiture, assurance véhicule, assurance automobile

=> Utile pour le SEO, et pour le SEM

La reconnaissance des entités nommées

Identifier dans les pages ce qui correspond à des noms de personnes de villes, des marques, des noms de produits, etc…

Deux applications directes :
- Être capable d'agréger des contenus en temps réel sur des personnes, des marques, des produits, des lieux inconnus l'instant d'avant
- Générer des regroupements automatiques à l'aide d'un outil de balisage (taggage) automatique.

Le balisage automatique

Système qui prend un document en entrée, en lit le contenu, le « qualifie », et ajoute un attribut sous forme d'un tag dans la base du site.

Ces tags peuvent être utilisés pour regrouper des pages ou créer des liens connexes intelligents.

Exemple : identifier automatiquement la thématique d'une page pour la classer
Exemple : identifier une page qui parle d'une personne
Etc…

L'exploitation des thésaurus et des ontologies

Une ontologie est une base de données qui stocke sous forme de triplets deux termes reliés entre eux et la nature de la relation sémantique qui les relie.

En ajoutant des formes de relations sémantiques « spécial SEO », on crée des bases opérationnelles pour créer, même à grande échelle, des bases de donnée de liens pertinents à ajouter dans les pages

- Liens de navigation verticaux (niveau supérieur, inférieur)
- Liens vers des pages « connexes »
- Navigation thématique « bypassant » les silos verticaux

La génération de contenus automatiques

Créer des contenus à partir de règles de syntaxe simple et d'une base de variantes (expressions interchangeables, termes contextuels etc…).

Le challenge est de générer des textes « lisibles » par un internaute

La clusterisation – la classification automatique

Deux utilisations classiques :

-Qualifier des pages (voir des sites) pour le netlinking

⇒Objectif : savoir de quoi parle la page, sans la lire

⇒Regrouper des contenus non structurés : messages de forum, blogs, contenus aspirés, agrégés etc…

Et de nombreuses choses encore…

Si vous avez des questions

n’hésitez pas