Algorithmes de recherche Internet

Arnaud Forasacco

05/02/2021

Contents

1 Introduction

Les moteurs de recherche sur Internet doivent faire face à de très fortes contraintes. Les outils de recherche doivent répondre à des millions de requêtes par jour alors que la quantité de documents qu'ils doivent analyser est gigantesque. L'entreprise Google indexe près de 4,3 milliards de pages Web. La détection d'une mise à jour de ces documents est elle aussi source de grandes difficultés. Il n'y a pas de centralisation de ces données et un moteur de recherche doit par conséquent scruter en permanence le réseau en vue de détecter ces changements. Dans cette section, nous examinerons dans un premier temps les variables caractérisant Internet et les documents qui composent ce réseau. Ensuite nous étudierons plusieurs algorithmes de détection de pertinence employés dans les moteurs de recherche.

Ce réseau diffère fondamentalement des autres réseaux connus de par sa nature ouverte et son utilisation grandissante. Le développement d'Internet et plus particulièrement du World-Wide Web est complètement incontrôlé. Cette situation favorise une augmentation continue du nombre de pages et de liens disponibles et mène à la création d'un vaste réseau très complexe.

2 Le Concept TF-IDF

TF-IDF (term frequency-inverse document frequency) est une mesure statistique qui évalue la pertinence d'un mot par rapport à un document dans une collection de documents. Cela se fait en multipliant deux mesures : le nombre de fois qu'un mot apparaît dans un document et la fréquence inverse du mot dans un ensemble de documents. Il a de nombreuses utilisations, surtout dans l'analyse de texte automatisée, et est très utile pour noter des mots dans des algorithmes de machine learning pour le traitement du langage naturel (PNL).

TF-IDF a été inventé pour la recherche de documents et la recherche d'informations. Cela fonctionne en augmentant proportionnellement au nombre de fois qu'un mot apparaît dans un document, mais est compensé par le nombre de documents qui contiennent le mot. Ainsi, les mots qui sont communs dans tous les documents, tels que ceci, quoi et si, se classent bas même s'ils peuvent apparaître plusieurs fois, car ils ne signifient pas grand-chose pour ce document en particulier.

TF-IDF pour un mot dans un document est calculé en multipliant deux métriques différentes : Le « term frequency » d'un mot dans un document. Il existe plusieurs façons de calculer cette fréquence, la plus simple étant le décompte brut des occurrences d'un mot dans un document. Ensuite, il existe des moyens d'ajuster la fréquence, par la longueur d'un document, ou par la fréquence brute du mot le plus fréquent dans un document. La fréquence inverse du document du mot sur un ensemble de documents. Cela signifie à quel point un mot est commun ou rare dans l'ensemble du jeu de documents. Plus il est proche de 0, plus un mot est courant. Cette métrique peut être calculée en prenant le nombre total de documents, en le divisant par le nombre de documents contenant un mot et en calculant le logarithme. Ainsi, si le mot est très courant et apparaît dans de nombreux documents, ce nombre se rapprochera de 0. Sinon, il se rapprochera de 1. La multiplication de ces deux nombres donne le score TF-IDF d'un mot dans un document. Plus le score est élevé, plus ce mot est pertinent dans ce document particulier.

Le score TF-IDF pour le mot t dans le document d de l'ensemble de documents D est calculé comme suit :

$$tf\,idf\,(t, d, D) = tf\,(t, d)\,.\,idf\,(t, D)$$

Où :

$$tf\,(t, d) = \log\,(1 + freq\,(t, d))$$

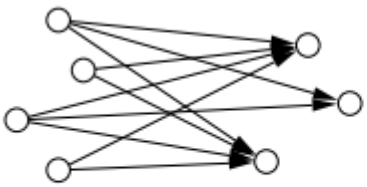
$$idf\,(t, D) = \log\,\left(\frac{N}{count\,(d \in D{:}t \in d)}\right)$$

3 L'algorithme HITS (Hyperlink-Induced Topic Search)

L'algorithme HITS est un algorithme de « link analysis ». Le but de l'algorithme HITS est de déterminer les hubs et les autorités qui renforcent leurs relations mutuellement sur un sujet donné.

L'algorithme est tel qu'il assigne deux scores portant sur : + Son autorité, qui estime la valeur du contenu de la page + Son hub, qui estime la valeur de ses liens à d'autres pages

Ainsi Jon Kleinberg (développeur de l'algorithme) dénombre les bons hubs comme des pages pointant vers beaucoup de bonnes autorités et les bonnes autorités comme des pages pointées par beaucoup de bons hubs. Cette dénotation de bons hubs et de bonnes autorités fait apparaître une troisième catégorie de pages ayant un grand nombre de liens entrants provenant de documents n'ayant aucune particularité. Ces pages, que nous nommons pages indépendantes, sont considérées comme universellement populaires et n'apportent que peu ou pas d'intérêt.



Hubs *Autorités*

Page indépendante

Une justification intuitive de l'autorité conférée à une page en fonction de la structure des liens l'entourant peut être donnée en considérant qu'un fort taux de jugement humain entoure l'ajout d'un lien hypertexte dans un document. En quelque sorte, l'auteur du document estime que la page vers laquelle il construit un lien évoque un sujet similaire à son souhait et paraît intéressante. Pour déterminer les hubs et les autorités d'un sujet donné, l'algorithme HITS se base sur un sous-graphe d'Internet S qui doit répondre aux conditions suivantes :

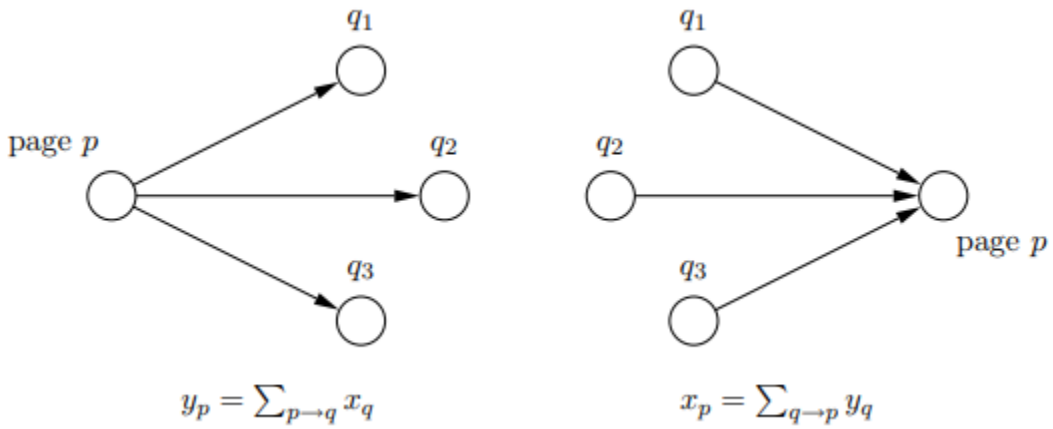
- S_σ est relativement petit,
- S_σ est riche en pages pertinentes,
- S_σ contient la totalité (ou la plupart) des plus importantes autorités.

En gardant S_σ petit, l'application d'algorithmes non triviaux peut s'effectuer sans s'occuper du temps de calcul nécessaire à la réalisation de la tâche. Les deux derniers points nous permettent de nous assurer d'avoir de bonnes chances de déterminer les bonnes autorités correspondant à la requête σ. Pour construire un tel graphe, l'algorithme utilise une requête de mots-clés (σ) afin de prendre en compte un petit nombre de pages (environ 200) depuis un moteur de recherche traditionnel à base d'index inversé. Cependant, cet ensemble de pages, noté R_σ, ne contient pas nécessairement l'ensemble des autorités du sujet σ.

Par exemple, il y a de fortes chances pour que les réponses à la requête « moteur de recherche » ne correspondent pas aux grands sites de moteurs de recherche : ces pages ne contiennent en effet que très rarement les mots-clés recherchés. On peut toutefois espérer que ce sous-graphe contienne des liens vers des autorités ou des hubs importants. R_σ est alors étendu en ajoutant les nœuds pointés par le sous-graphe et les nœuds pointant le sous-graphe afin d'obtenir un graphe augmenté S_σ. Les bons hubs et les bonnes autorités peuvent être extraits de ce graphe en donnant une définition numérique à la notion de hub et d'autorité.

L'algorithme HITS associe un vecteur de potentiels d'autorités non-négatifs $< x >$ et un vecteur de potentiels de hubs non-négatifs $< y >$ pour toutes les pages appartenant à S_σ. Ainsi, une page p possédant un fort potentiel d'autorité x_p, respectivement un fort potentiel de hub $ y_{p}$, sera vue comme étant une bonne autorité, respectivement un bon hub. Les poids sont ensuite mis à jour de la façon suivante : si une page est pointée par plusieurs bons hubs, nous incrémentons son potentiel d'autorité. Ainsi, pour une page p donnée, la valeur de son poids x_p est mise à jour par la somme des potentiels y_q de toutes les pages q pointant sur p.

Ces opérateurs sont décrits dans la figure ci-dessous :



La sortie de l'algorithme HITS est alors composée de deux listes ordonnées décroissantes de pages en fonction des potentiels respectifs de hub et d'autorité.

4 L'algorithme *PageRank*

L'idée principale de cette méthode est de simuler le comportement d'un internaute naviguant de manière aléatoire sur Internet.

La probabilité qu'il visite une page donnée est d'autant plus grande que cette page est pointée par beaucoup d'autres pages au travers de leurs liens hypertextes. Pour accéder à une page Web sur Internet, on peut d'une part l'atteindre directement en connaissant son adresse et d'autre part suivre un lien hypertexte d'un autre document.

Le calcul du PageRank est donc de la pertinence d'une page intègre ces deux éléments au travers d'une probabilité d. d représente en quelque sorte la probabilité que l'internaute aléatoire s'ennuie sur une page et décide de choisir une autre page au hasard.

$$PR(p_j)_t = (1-d) + d \sum_{\substack{i=1 \\ p_i \to p_j}}^{n} \frac{PR(p_i)_{t-1}}{C(p_i)}$$

Dans cette formule, $PR(p_j)_t$ représente la valeur du PageRank à l'itération t pour la page pj . $C(pi)$ est défini comme le nombre de liens sortants de la page pi, le paramètre d prend ses valeurs dans l'intervalle [0−1] et est généralement placé à $d = 0.85$ d'après des études statistiques menées par Lary Page.

Cette équation montre la récursion permettant de calculer le PageRank pour tous les nœuds du graphe représentant Internet.

Afin d'améliorer les performances de cet algorithme, une variation a été introduite conduisant à modifier, dans l'évaluation du *PageRank* d'une page, l'importance de l'apport de chaque lien au calcul final. Des heuristiques de modifications du poids de chaque lien ont notamment été étudiées. Ces heuristiques sont principalement basées sur l'analyse des textes entourant et présents sur les différents liens hypertextes contenus dans les pages Web et sur l'utilisation de mesures de similarités basées sur le modèle vectoriel et la mesure TF-IDF.

5 L'algorithme SALSA (Stochastic Approach for Link-Structure Analysis)

L'algorithme alternatif SALSA a été développé par Lempel et Moran. Son but est similaire à celui de l'algorithme HITS décrit dans la section : il cherche à déterminer les meilleurs pages correspondant à un sujet donné en les caractérisant en hubs et autorités.

La méthode choisie pour résoudre ce problème est toutefois différente de celle utilisée dans l'algorithme HITS. Elle consiste à parcourir au hasard les nœuds du graphe d'Internet en suivant les liens les reliant avec une distribution de probabilité uniforme. Les pages les plus visitées correspondent alors aux solutions recherchées.

Ainsi, en parcourant le graphe au hasard des liens, la probabilité de visiter une page autorité est grande. Cette idée de détermination stochastique de l'intérêt d'une page se rapproche de celle utilisée dans le calcul du PageRank mais diffère dans la séparation des résultats en hubs et autorités pour un domaine donné. La détermination des probabilités d'appartenance de chaque page du graphe à l'une des deux catégories possibles se fait à l'aide de chaînes de Markov.

Exemple élémentaire de chaîne de Markov, à deux états A et E. Les flèches indiquent les probabilités de transition d'un état à un autre.

La première étape consiste à construire un sous-graphe d'Internet correspondant à une requête donnée, sur lequel s'effectuera la recherche de pages pertinentes. Cette construction se base sur celle utilisée dans HITS et requiert les mêmes propriétés.

Afin de déterminer le potentiel de hub et d'autorité de chaque page, il faut déterminer la probabilité de visite de ces pages en suivant les liens du graphe de manière aléatoire.

Le parcours des arcs dans le sens initial permet de déterminer les potentiels d'autorités des pages : on mesure ici la probabilité qu'une page soit pointée par beaucoup d'autres pages. Le parcours du graphe en suivant les arcs dans leur sens inverse nous donne les potentiels de hub de chaque page : on mesure, dans ce cas, la probabilité qu'une page pointe vers beaucoup d'autres pages. Deux chaînes de Markov sont alors analysées : une chaîne de hubs et une chaîne d'autorités. Les états de transitions de ces chaînes sont générés en effectuant le parcours aléatoire du graphe.

Mais, contrairement à un parcours classique des liens, deux liens hypertextes sont traversés :

- En choisissant selon une probabilité uniforme d'aller sur une page pointée par la page actuelle, et
- En choisissant selon une probabilité uniforme d'aller sur une page pointant vers la page actuelle.

Les potentiels d'autorités sont alors définis comme étant la distribution stationnaire de la chaîne de Markov effectuant d'abord un pas (1) puis un pas (2), tandis que les potentiels de hubs sont définis comme la distribution stationnaire de la chaîne effectuant d'abord un pas (2) puis un pas (1).

Formellement, les deux matrices de transition des deux chaînes de Markov sont définies ainsi :

1. La matrice déterminant les potentiels de hub $\tilde{H}$,

$$\tilde{h}_{i,j} = \sum_{k:k \in S(i) \cap S(j)} \frac{1}{|S(i)|} \cdot \frac{1}{|E(k)|}$$

2. La matrice déterminant les potentiels d'autorité $\tilde{A}$,

$$\tilde{a}_{i,j} = \sum_{k:k \in E(i) \cap E(j)} \frac{1}{|E(i)|} \cdot \frac{1}{|S(k)|}$$

Dans ces formules, $E(i)$ décrit tous les nœuds du graphe pointant vers le nœud i, et donc les pages que nous pouvons atteindre depuis la page i en suivant un lien dans le sens inverse. $S(i)$ décrit tous les nœuds que nous pouvons atteindre depuis le nœud i en suivant un lien de i.

Une probabilité de transition $\tilde{a}i, j > 0$ indique qu'une certaine page k pointe à la fois vers les pages i et j et que, de plus, la page j peut être atteinte depuis la page i en deux pas : en parcourant le lien de la page k vers i dans le sens inverse et en suivant ensuite le lien de la page k vers la page j.

Le renforcement mutuel par la structure des liens du graphe de l'algorithme HITS fait qu'il pose problème dans certains cas, notamment ceux identifiés par l'effet TKC (Tightly-Knit Community). Cet effet apparaît lorsqu'une communauté de pages obtient un très bon score par les algorithmes d'établissement de pertinence par mesures topologiques, bien qu'elles ne fassent pas autorité sur le sujet donné. Ces communautés sont petites mais très fortement interconnectées entre elles. Il a toutefois été prouvé que bien que l'algorithme HITS soit affecté par ce problème, SALSA arrive à bien évaluer ces communautés.

6 L'algorithme PHITS

D'autres approches de détermination de hubs et autorités ont également été testées. L'algorithme PHITS est un algorithme statistique afin de déterminer ces deux catégories.

Le modèle qui a été construit tente d'expliquer deux types de variable, les citations c d'un document d en fonction d'un petit nombre de variables communes z qui sont appelées les aspects ou les facteurs. Ces variables communes peuvent être considérées comme des sujets ou des communautés de pages.

Le modèle peut alors être décrit statistiquement : un document $d \in D$ est généré avec une probabilité $P(d)$, le facteur, ou sujet $z \in Z$ associé à d est choisi en fonction d'une probabilité $P(z|d)$, et étant donné ce facteur, des citations $c \in C$ sont générées en fonction de la probabilité $P(c|z)$. La probabilité de chaque paire (document, citation) (d, c) est alors décrite par :

$$
\begin{aligned}
P(d,c) &= P(d)P(c|d) \\
P(c|d) &= \sum_z P(c|z)P(z|d)
\end{aligned}
$$

En considérant la matrice A représentant les paires (document, citation) décrites dans la section 1.2.3, où l'entrée $A[i,j]$ est non nulle si le document i possède un lien vers le document j, la probabilité de la matrice de citation A est la suivante :

$$
L(A) = \prod_{(d,c)\in A} P(d,c)
$$

Le problème consiste alors à trouver les valeurs de $P(d)$, $P(z|d)$ et de $P(c|z)$ qui maximisent la fonction de probabilité $L(A)$ des données observées. Afin de résoudre ce nouveau problème, les auteurs proposent d'utiliser l'algorithme EM (Espérance-Maximisation) de Dempster.

« Il s'agit d'un algorithme itératif qui permet de trouver les paramètres du maximum de vraisemblance d'un modèle probabiliste lorsque ce dernier dépend de variables latentes non observables. »

Ce modèle entièrement probabiliste à l'avantage d'apporter plus d'informations que le modèle utilisé par l'algorithme HITS. Une analogie peut être toutefois faite en considérant les autorités sur un sujet donné comme la probabilité conditionnelle $P(c|z)$ qui indique de quelle manière un document c est cité depuis une communauté z. Mais d'autres informations peuvent être extraites du modèle comme par exemple la probabilité $P(z|c)$ qui nous permet de connaître la communauté à laquelle appartient un document c donné, ou encore la découverte des documents caractéristiques d'une communauté donnée en déterminant le produit $P(z|c)$ * $P(c|z)$.

Comme nous avons pu le voir, les techniques de recherche d'informations sont très variées et prennent en compte des éléments très différents. À l'heure actuelle, l'analyse de la topologie d'Internet est un élément prépondérant dans la détection de la pertinence d'un document. Ce système permet en effet de récupérer un certain jugement humain inscrit dans les liens hypertextes : lorsqu'une personne décide d'inscrire un lien hypertexte dans une page Web, elle considère que le document pointé par ce lien apporte une information utile.