Moteurs de recherche : des algorithmes sans contrôle en quête de compréhension ?

Éric Bruillard

Moteurs de recherche et réseaux sociaux : des algorithmes sans contrôle ?

Eric Bruillard, EDA, université de Paris
eric.bruillard@u-paris.fr (Environ 15000 signes)

En ligne : http://www.apden.org/Moteurs-de-recherche-des-algorithmes-sans-controle-en-quete-de-comprehension.html

Des algorithmes dans notre vie quotidienne

Le mot « algorithme » fait figure d'épouvantail depuis quelques années, sorte de monstre malveillant tirant les ficelles derrière les applications offertes sur les téléphones portables. Pourtant, la notion d'algorithme est très ancienne (voir par exemple, « Histoire d'algorithmes ») et correspond simplement à une suite finie et non ambiguë d'instructions et d'opérations permettant de résoudre une classe de problèmes[1].

Elle est d'abord associée aux mathématiques et le mot lui-même viendrait d'un mathématicien persan du 9e siècle nommé Al-Khwârizmî. Les élèves de maternelle y sont familiarisés. Ainsi, ils construisent des objets selon une règle, par exemple en élaborant un collier avec des billes de différentes couleurs et des successions à respecter, ce qui les prépare notamment à la notion de numération. Les techniques opératoires qu'ils apprennent à l'école primaire, comme l'addition et la soustraction, sont des algorithmes qui transforment les nombres via leur écriture décimale. Notons un algorithme *littéraire*, celui que Georges Perec a écrit pour la revue *L'Enseignement programmé*, paru en décembre 1968 : « L'art et la manière d'aborder son chef de service pour lui demander une augmentation », adapté ensuite pour le théâtre sous le titre *L'augmentation*. Cet algorithme est brillamment développé, décrivant les différentes actions à mener pour obtenir l'augmentation convoitée [2].

Un peu plus tard, au moment de l'enseignement de la programmation, divers auteurs cherchaient à montrer que, comme Monsieur Jourdain faisait de la prose sans le savoir, nous suivons tous des algorithmes sans en avoir conscience : en gros des séquences d'actions avec quelques formes de contrôles avec « tant que » ou « jusqu'à ce que » pour aboutir à un résultat final. Les recettes de cuisine constituent depuis fort longtemps un exemple très commenté. Il y a d'un côté les ingrédients qu'il va falloir utiliser et de l'autre une description de la suite d'actions à réaliser, avec différents tests et contrôles : attendre l'ébullition de l'eau, que le beurre soit fondu mais pas noirci, que la pâte soit onctueuse… Les descriptions des recettes montrent d'ailleurs certaines limitations : comment juger que le mélange est bien réalisé, que la poêle est assez chaude, qu'il y a suffisamment de sel, etc. Certains ajoutent des « trucs », d'autres les gardent pour eux, Les recettes sont plus ou moins complexes, des tours de main peuvent s'avérer nécessaires, et le plus souvent le résultat n'est pas garanti, même si on suit scrupuleusement l'algorithme décrit dans la recette.

[1] https://fr.wikipedia.org/wiki/Algorithme
[2] http://lecture.cafeduweb.com/lire/13001-art-maniere-aborder-son-chef-service-pour-lui-demander-augmentation---georges-perec.html

L'intelligence artificielle est maintenant convoquée, avec des machines pouvant exécuter des recettes de grands chefs cuisiniers : le *Gastronomy Flagship Project* de Sony[3] est dédié à la gastronomie, une Intelligence artificielle pour la création des recettes, un robot pour les réaliser et un réseau social de cuisiniers pour une création communautaire[4] ; *Moley*, la première cuisine entièrement robotisée au monde[5] ou le *Bot Chef* de Samsung[6] : « Non seulement le robot cuisine des repas complets, mais il vous indique quand les ingrédients doivent être remplacés, suggère des plats en fonction des articles que vous avez en stock, apprend ce que vous aimez et nettoie même les surfaces après lui-même ».

En fait, les techniques d'intelligence artificielle peuvent donner aux appareils une certaine capacité d'agir par eux-mêmes et, au lieu de toucher un écran ou manipuler des boutons, nous parlerons à nos appareils constamment à l'écoute dans nos maisons et nos bureaux. La promesse ultime est dans leur capacité à prédire ce que nous voulons avant même qu'on le demande. L'inquiétude que l'on peut ressentir, ne vient pas d'activités menées par des humains suivant des algorithmes, mais nait lorsque ces derniers nourrissent des systèmes ou des machines automatiques qui prennent des décisions ou orientent nos choix et nos visions du monde. Nous allons tenter d'en expliquer la nature autour des questions de recherche d'information.

Des moteurs de recherche : des algorithmes au pouvoir grandissant

Au démarrage de l'Internet grand public, avec le développement du web, il a fallu inventer des techniques pour aider les humains à trouver les pages qui pouvaient les intéresser.

Pour cela, il fallait indexer un grand nombre de pages pour pouvoir les retrouver et déterminer un lien entre la demande d'un utilisateur et ces pages. Les techniques documentaires classiques, à partir des descripteurs des pages, s'avéraient peu performantes, notamment parce que l'indexation des pages, assurée par les auteurs eux-mêmes, ne suivait pas les règles classiquement recommandées en documentation. En outre, le fait que le web soit un hypertexte, c'est-à-dire que les pages sont liées entre elles, n'était pas pris en compte. Cela a été la grande force de Google, d'une part de constituer des parcs de machines pour indexer les pages web et d'autre part d'élaborer un algorithme reposant sur les caractéristiques des pages (notamment les caractères qu'elles contiennent) et les liens entre les pages[7]. En effet, les moteurs doivent palier à un défaut structurel du web : les liens hypertextes ne sont pas bidirectionnels et si vous lisez une page, vous ne pouvez pas savoir quels liens pointent sur cette page. Les moteurs de recherche ont récupéré ces informations et elles s'en servent pour déterminer l'*intérêt* d'une page. En gros, plus il y a de liens pointant sur une page, plus il y de chances pour qu'une page soit intéressante (surtout si ces liens proviennent de pages également jugées intéressantes !).

Notons que cette caractéristique a orienté des techniques courantes pour améliorer le référencement d'une page et a également été détournée via ce que l'on nomme le *Google bombing* ou le *bombardement Google*. Cela consiste à multiplier les liens vers une page ou un site avec la même expression dans l'origine du lien. Cette expression, souvent négative voire

[3] https://ai.sony/projects/
[4] https://www.futura-sciences.com/tech/actualites/intelligence-artificielle-ia-sony-veut-defier-top-chef-84723/
[5] https://moley.com/
[6] https://www.cnetfrance.fr/news/ifa-2019-on-a-cuisine-avec-le-robot-chef-de-samsung-39890207.htm
[7] Sur Page Rank, l'algorithme de tri utilisé par Google, voir Delahaye (2007) ou Guerraoui (2014).

dénonciatrice ou diffamatoire, devient un mot clé pour conduire au site (voir quelques exemples de détournements[8]).

En outre, contrairement aux algorithmes habituels en documentation qui fournissent de manière exhaustive les réponses à une demande, le très grand nombre de pages susceptibles de répondre à une requête, qui peut être un mot quelconque, conduit à trier les réponses à afficher. C'est ce tri, le choix effectué automatiquement par la machine, qui va s'avérer déterminant.

La question qui s'en sut est de trouver les meilleurs algorithmes, c'est-à-dire ceux qui *satisfont* le mieux les utilisateurs. Pour cela si on dispose d'autres informations et pas uniquement le ou les quelques mots tapés au clavier, on sera certainement mieux armé. La notion de « profil » va s'imposer : si on sait que la demande vient d'un élève d'une classe de cinquième, ce qui convient à l'un devrait convenir à l'autre ; de même au sein d'une entreprise, etc. Bien évidemment, plus on connait le demandeur, son pays, sa langue, ses opinions, ses goûts, ses préférences, etc., plus on peut avoir de chances de le satisfaire. Mais que veut dire « satisfaire » : pour un usager, qu'il soit content afin de le faire rester puis revenir ? Pour un annonceur, pousser les internautes vers son site ? Ne s'agit-il pas alors, dans les choix opérés, d'exploiter des biais cognitifs : donner à utilisateur ce qui confirme ce qu'il croit déjà, ce que d'autres qui lui ressemblent trouvent intéressants, etc. Et au lieu d'attendre qu'il formule une requête, pourquoi ne pas directement lui proposer ce qui est susceptible de l'intéresser. C'est l'un des modes privilégiés des réseaux sociaux, pousser l'information vers des utilisateurs, selon les personnes ou les sites auxquels il s'est abonné, passant de la recherche d'informations aux systèmes de recommandations.

On arrive aux bulles de filtres : ce qui est rendu visible à chacun ou à chacune dépend des données différentes données collectées sur lui ou elle. Il ou elle serait installé dans une « bulle » unique, construite à la fois par les algorithmes et par ses propres choix (« amis » sur les réseaux sociaux, sources d'informations, etc.), optimisée pour sa personnalité supposée[9] (voir aussi Doctorow, 2007, pour une vision dystopique).

Une autre évolution des moteurs de recherche est leur transformation en moteurs dits de réponse : les moteurs de recherche ne répondaient pas directement aux questions posées par les utilisateurs mais fournissaient des pages, à charge pour les demandeurs de les lire et d'en extraire les informations recherchées. Il s'agit maintenant de leur fournir le plus directement possible des éléments en lien avec leur question, voire directement une réponse[10]. Dans les travaux de recherche actuels, il s'agit d'aller encore plus loin : des données à l'information, aux connaissances (Abiteboul, 2019), ce que Google offre déjà dans des encarts à droite de la page de résultats.

Détourner pour mieux comprendre les fonctionnements des algorithmes

Pour résumer, côté documentation, on est passé :

[8] Voir https://fr.wikipedia.org/wiki/Bombardement_Google ou https://optimiz.me/google-bombing/ ou https://smartkeyword.io/seo-netlinking-google-bombing/
[9] Selon Eli Pariser, voir https://fr.wikipedia.org/wiki/Bulle_de_filtres
[10] https://fr.wikipedia.org/wiki/Moteur_de_réponse

- des bases de données de documents : un corpus bien délimité avec une liste de champs (auteur, titre, éditeur, mots clés, résumé…), des langages documentaires et une réponse exhaustive et sûre à une requête (aux erreurs de saisie près),
- aux pages web : un corpus mal délimité (couverture du web), du texte structuré indexé, des liens entre les pages, une réponse non exhaustive, un algorithme de tri des réponses.

Avec Internet, on a en gros trois manières de « tirer » de l'information : (1) utiliser un moteur de recherche ou un outil de recherche, il faut alors décrire ce que l'on cherche, ou fournir un fragment (image, son, texte) ; (2) naviguer en cliquant sur des liens proposés sur la page que l'on consulte, mais on dépend de ce qui est présent sur la page ; (3) consulter ce que le système ou notre réseau a décidé de nous envoyer.

Dans tous les cas, un processus a permis de filtrer l'information et de plus ou moins la personnaliser, en fonction de ce que l'on peut connaître de nos souhaits et de nos besoins. Mais pour satisfaire un humain, n'est-il pas finalement plus facile de le rendre prédictif, de le façonner petit à petit de façon à ce qu'il en arrive à souhaiter ce que l'on prévoit pour lui ? Heureusement, ce n'est pas aussi simple que cela.

Les professeurs documentalistes ont un rôle important à jouer, pour aider les utilisateurs à mieux appréhender les différents instruments de recherche à leur disposition et à exercer leur jugement sur leur fonctionnement et sur ce qu'ils proposent.

D'abord, ne pas oublier que la « pertinence » est avant tout un jugement sur l'adéquation entre des besoins de connaissance et des documents qui « contiennent » des informations (une capacité que les machines sont encore loin d'avoir) et qu'il y a différentes voies pour les trouver : taper une requête, choisir via une image, cliquer sur un lien visible, etc. Ainsi, choisir une page à partir d'une image est une technique qui peut être performante : avec Google images, on peut survoler rapidement de très nombreux exemples et les images elles-mêmes reflètent souvent un niveau de technicité. Par exemple, avec la requête « cycle de l'eau », les images proposées, leur complexité et le lexique visible, donnent des indices sur le niveau de classe auquel elles peuvent correspondre (plus facile à choisir qu'à partir de la lecture des brefs résumés fournis). En tous cas, il faut développer des techniques d'enquêtes et faire preuve de réflexivité : utiliser plusieurs méthodes, comparer, etc. On ne peut pas le faire tout le temps, mais faire quelques séances « marquantes », au cours desquelles on fera des choses inhabituelles et qui devraient surprendre, fournit des activités fructueuses.

Il est certes difficile d'enseigner le fonctionnement des machines informatiques qui nous environnent, mais on peut se donner l'objectif d'essayer de mieux le comprendre, et pour comprendre il faut détourner (Bruillard, 2020).

Ainsi, affiner le jugement de pertinence est un objectif important, mais il est plus facile de détecter des choses non pertinentes que des choses pertinentes. Il faut apprendre à rejeter, puis apprendre à classer, des documents meilleurs que d'autres relativement à une requête dans un contexte particulier.

Aller chercher des choses inattendues : ne pas prendre uniquement les recherches dans une vision de service mais en conduire avec des demandes inhabituelles : faire des jeux (voir par exemple Simonnot, 2008), essayer de faire taire le moteur (plusieurs mots qui ensemble de

donnent aucun résultat), rechercher un poème via une image, un avocat marron qui ne soit pas un fruit via Google images, etc. Se méfier des *nudges* (Boissière et Bruillard, 2021)[11].

Refuser la similarité : à l'instar des banques qui vous demandent un profil (prudent, risqué…) et gèrent selon le profil, pourquoi n'y aurait-il pas un profil de *rechercheur* (curieux, aventureux, scolaire…) ? Pourquoi ne pas demander des choses éloignées de ce que l'on connait ? Quelles poésies recommander à un adolescent amateur de science-fiction ?

Que la formation aux instruments de recherche ouvre de multiples perspectives, qu'elle soit le lieu d'ouvertures et de rencontres inattendues, au-delà de la satisfaction de critères simplistes de performance, n'est-ce pas ce que l'on pourrait souhaiter ?

Références

Abiteboul Serge (2019). Des données à l'information, aux connaissances. Vidéo https://www.youtube.com/watch?v=9Sk42Fy6lMo

Boissière Joël et Bruillard Éric (2021). L'école digitale. Une éducation à vivre et à construire. Collection Sociologie. Armand Colin, 368 p.

Bruillard Éric (2020). L'écriture inclusive ouvre des liens surprenants. Réflexions en didactique de l'informatique. STICEF, Volume 27, n°1, 2020. http://sticef.univ-lemans.fr/num/vol2020/27.1.4.bruillard/27.1.4.bruillard.html

Chabert Jean-Luc, Barbin Evelyne, Guillemot Michel, Djebbar Ahmed, … (2010). Histoire d'algorithmes. Du caillou à la puce. Belin.

Delahaye Jean-Paul (2007). Un moteur de recherche, pour le meilleur et pour le pire. Interstices, INRIA. https://interstices.info/un-moteur-de-recherche-pour-le-meilleur-et-pour-le-pire/

Doctorow Cory (2007). *enGooglés*, Traduction de Scroogled (Septembre 2007 ? Magazine Radar) par Valérie Peugeot, Hervé Le Crosnier et Nicolas Taffin pour C & F éditions. https://cfeditions.com/scroogled/scroogled.html

Guerraoui Rachid (2014). Un algorithme : PageRank de Google. Blog Binaire, Le Monde. https://www.lemonde.fr/blog/binaire/2014/12/01/un-algorithme-pagerank-de-google/

Simonnot Brigitte (2008). Quand les moteurs de recherche appellent au jeu : usages ou détournements ? Questions de communication, 14 | 2008, p. 95-114. http://journals.openedition.org/questionsdecommunication/752

[11] Développée par l'économiste comportemental Richard Thaler et le juriste Cass Sunstein, la théorie du nudge désigne une méthode d'influence cherchant à modifier des comportements humains, notamment des choix à faire, sans contrainte, ni obligation, ni sanction. https://fr.wikipedia.org/wiki/Nudge_(livre)