Une (très brève) histoire des moteurs de recherche

Dès lors qu'on avait inventé l'écriture, on chercha un support assez maniable et résistant pour entreposer et échanger les messages sans les altérer. Le papyrus fut préféré à la pierre ainsi qu'aux tablettes de cire ; il était plus léger, maniable et ne fondait pas au soleil. On mit au point des meubles destinés à stocker les rouleaux auxquels on avait accroché des étiquettes indiquant de quoi il était question, de manière à éviter d'avoir à dérouler les rouleaux un par un chaque fois qu'on chercherait une information. A ces étiquettes méta-informatives s'ajoutèrent des catalogues de fiches, dont le premier fut confectionné au troisième siècle avant Jésus-Christ par le poète grec Callimachus [Eliot et Rose, 2009].

On positionna les meubles chargés de rouleaux de papyrus dans des bâtiments construits pour les recevoir. On donna aux bibliothécaires les clefs de ces bâtiments en leur confiant la mission d'archiver les documents et de réguler les allers et venues des visiteurs. Etant donné l'importance revêtue par l'information en matière de religion, de science, d'économie et de droit, le rôle du bibliothécaire était éminemment politique [Kaser, 1962]. Son pouvoir s'exerçait sur les lecteurs, dont il pouvait influencer le savoir, et sur les auteurs, dont il pouvait atténuer ou amplifier l'autorité.

Au deuxième siècle avant Jésus Christ, lorsque les Pergaméniens (dont les terres se situaient sur la côte orientale de l'actuelle Turquie) se vantèrent de posséder une collection de rouleaux plus riche que celle d'Alexandrie, les Egyptiens cessèrent de les fournir en papyrus. Cet embargo décidé par Ptolémée conduisit les sujets du roi de Pergame, Eumène II, à fixer les informations sur la membrane intérieure de la peau de bête. Cette méthode connut un succès que l'on peut en partie expliquer par le fait que tous les pays pratiquaient l'élevage alors que le papyrus ne pouvait pas pousser partout. Le « parchemin » (en latin *pergamenum* : pergaménien) ne pouvait être roulé sans risquer d'altérer le contenu, aussi décida-t-on de le découper en rectangles qu'on plierait et relierait par le côté. Le codex, ancêtre du livre, était né [Langville et Meyer, 2006]. On mit au point de nouveaux meubles adaptés au stockage de ces objets parallélépipédiques et on décida d'inscrire les méta-informations sur leurs tranches plutôt que sur des étiquettes risquant de se détacher.

La recherche d'information est une démarche indifféremment sociale et technique, liée au type de contenu autant qu'au format du contenant, ainsi qu'à leurs producteurs, leurs propriétaires, leur localisation et leur environnement économique et politique. Il ne peut y avoir de processus de recherche si ceux qui ont produit le contenu et stocké le contenant n'ont pas agi de manière à ce que l'information puisse être retrouvée par celui qu'elle intéressera. Autrement dit, la recherche d'information et *le désir de communication* sont consubstantiels [Duguid, 2008]. C'est parce qu'on souhaite mobiliser l'information dans le temps et l'espace, et parce qu'on organise socialement et techniquement les modalités de sa communication, que la recherche est possible.

Mécanisation et automatisation

A partir de la fin du dix-neuvième siècle, alors que la recherche d'information s'effectuait encore grâce à des catalogues de fiches cartonnées produites à la main, certains documentalistes, en observant la mécanisation touchant de plus en plus de secteurs d'activité, rêvaient au jour où ils pourraient confier à la machine le soin d'effectuer à leur place un travail qu'ils jugeaient ingrat et répétitif. Ce fut le cas du belge Paul Otlet, avocat pacifiste fondateur en 1895 de l'*Institut international de bibliographie* à Bruxelles, où il avait conçu avec son collaborateur Henri La Fontaine le projet de constituer le livre universel du savoir [Mattelart, 2009, p. 24]. Dans son testament philosophique, *Traité de documentation* (1934), Paul Otlet

imagina une table de travail qui ne serait plus chargée d'aucun livre. « À leur place, écrivait-il, se dresse un écran et à portée un téléphone. Là-bas au loin, dans un édifice immense, sont tous les livres et tous les renseignements. De là, on fait apparaître sur l'écran la page à lire pour connaître la réponse aux questions posées par téléphone, avec ou sans fil. Un écran serait double, quadruple ou décuple s'il s'agissait de multiplier les textes et les documents à confronter simultanément ; il y aurait un haut parleur si la vue devait être aidée par une donnée ouïe, si la vision devait être complétée par une audition. Utopie aujourd'hui, parce qu'elle n'existe encore nulle part, mais elle pourrait bien devenir la réalité pourvu que se perfectionnent encore nos méthodes et notre instrumentation. Et ce perfectionnement pourrait aller jusqu'à rendre automatique l'appel des documents sur l'écran, automatique aussi la projection consécutive » [Levie, 2006].

Ce scénario se montra prémonitoire. Bientôt, les systèmes d'archivage et de projection automatiques virent en effet le jour. La première tentative est attribuée à Emanuel Goldberg. Dans les années 1920, il déposa une série de brevets décrivant une machine dont le but était de consulter mécaniquement un catalogue de documents enregistrés sur microfilm. Grâce à un ingénieux jeu de lumière, Goldberg formulait une requête en appuyant sur un bouton, après quoi la machine projetait le document correspondant [Buckland, 2006]. L'utilisation du microfilm se perfectionna dans les années 1930, notamment avec les travaux d'Helen et Watson Davis, de Rupert Draeger puis de Vannevar Bush. Elle atteignit son apogée avec le *Rapid Selector* de Ralph Shaw capable de consulter 78 000 entrées par minute [Sanderson et Croft, 2012].

Alors que ces innovations avaient lieu, l'ordinateur faisait son apparition à la suite des travaux d'Alan Turing, Claude Elwood Shannon, George Stibitz et Konrad Zuse. De nouveaux procédés de stockage et de traitement de l'information furent mis au point, basés sur l'utilisation d'impulsions électriques et du code binaire. Le « *bit* » devint à l'information ce que l'atome était à la matière : particule élémentaire [Negroponte, 1995]. Comme dans le cas du passage du rouleau de papyrus au codex de parchemin, dès lors qu'une nouvelle façon de traiter et de transmettre l'information avait été mise au point, il fallut concevoir de nouveaux procédés de stockage et de nouvelles procédures d'accès.

Sciences de la recherche d'information

La première fois qu'un ordinateur fut utilisé pour rechercher une information remonte sans doute à 1948, lorsque John Edwin Holmstrom présenta à la *Royal Society Scientific Information Conference*, en Grande-Bretagne, une machine surnommée « Univac » (*UNIVersal Automatic Computer*). En plus de stocker l'information sur bande magnétique, Univac était capable d'apparier automatiquement des codes thématiques et des références en consultant un catalogue au rythme de 120 mots par minute.

Ce fut également en 1948 que le terme « Science de la Recherche d'Information » (SRI) fut employé pour la première fois par l'informaticien Calvin Mooers. Il désignait à la fois le processus concret permettant de convertir l'expression d'un besoin d'information en une liste de documents susceptibles de combler ce besoin [Mooers, 1951, p. 25], et le champ des sciences appliquées désormais consacré à l'étude des méthodes qui permettraient de réaliser et d'améliorer ce processus [Baeza-Yates et Ribeiro-Neto, 2011].

En 1958, la première conférence dédiée à la SRI se tint à Washington. Il s'agissait pour les pionniers – Cyril Cleverdon, Brian Campbell Vickery, Peter Luhn – de définir le meilleur moyen d'indexer des références à un temps t de sorte qu'il soit possible de les mobiliser à $t+x$. Il fallait pour cela concevoir un instrument capable de consulter automatiquement les informations contenues dans une collection de documents numérisés et de générer des

descripteurs pour chacun d'eux : mots-clés, auteur, date. Le dispositif chercherait ensuite à apparier les termes d'une requête effectuée par un utilisateur aux descripteurs générés pour finalement suggérer une liste de documents potentiellement pertinents [Van Rijsbergen, 1979, p. 1].

Classification par mots-clés

Les chercheurs en SRI ne se contentèrent pas de numériser les fiches cartonnées et les procédures afférentes. Ils observèrent les alternatives à la classification décimale développée par Melvil Dewey en 1876 puis complétée par Henri Lafontaine et Paul Otlet au début du vingtième siècle. Cette classification se basait sur un rangement des documents par disciplines, thèmes et sous-thèmes. Désireux de sonder d'autres voies, Mortimer Taube et ses collègues proposèrent au début des années 1950 une méthode nommée « *Uniterm* », consistant à associer librement des mots-clés à chaque document et à effectuer des recherches sur la base du sujet traité plutôt que sur celle du champ disciplinaire [Taube *et al*., 1952]. Cyril Cleverdon, pionnier de la SRI, se livra à une comparaison rigoureuse des résultats donnés par la classification décimale et l'alternative *Uniterm*, et montra que la nouvelle méthode fonctionnait mieux que l'ancienne [Cleverdon, 1959].

Peter Luhn suggéra quant à lui d'attribuer à chaque document un score de pertinence correspondant à une requête donnée, de manière à hiérarchiser les listes de documents générées par le système par ordre de pertinence supposée [Luhn, 1957 ; Maron *et al*., 1959]. Le système de classement par mots-clés et par ordre de pertinence finit par s'imposer, concordant avec l'arrivée des ordinateurs dans les bibliothèques. Le changement de support et des conditions de stockage avait ainsi entraîné, comme dans le cas de l'invention du parchemin par les Pergaméniens, un changement des procédés d'indexation et de recherche.

Analyse statistique des textes

A la fin des années 1950, Peter Luhn montra que la fréquence d'utilisation d'un mot dans un document ainsi que sa position par rapport aux autres mots permettaient de mesurer le degré de pertinence du document pour une requête donnée [Luhn, 1958]. Ces deux paramètres (fréquence et position) permirent de développer des moteurs de recherche qui ne se contentaient plus d'apparier une requête à des méta-informations attachées au document par l'auteur ou le documentaliste, comme l'étaient les mots-clés du système *Uniterm*, mais directement aux informations contenues à l'intérieur du document.

Au début des années 1970, Karen Spärk Jones montra que la fréquence d'apparition d'un mot dans une collection de documents était inversement proportionnelle à l'intérêt du mot, cela car les mots les plus répétés étaient en général des articles, des pronoms ou des auxiliaires, tandis que les mots les moins répétés avaient une plus forte probabilité d'être significatifs [Jones, 1972]. On put croiser les paramètres définis par Luhn et Jones — position du mot, fréquence d'apparition dans le document et fréquence d'apparition dans la collection de documents — et augmenter considérablement l'efficacité des procédés de recherche automatisée.

Dans les années 1980, les chercheurs en SRI travaillèrent au couplage de l'analyse statistique des textes avec des approches probabilistes, notamment le modèle booléen et la théorie des ensembles flous [Salton et McGuill, 1983]. Ces méthodes permirent d'augmenter la probabilité, pour une requête donnée, de réussir à générer automatiquement une liste de résultats pertinente.

En France, des travaux en SRI furent menés à partir des années 70 par des chercheurs comme Christian Fluhr, du Commissariat à l'énergie atomique, qui travailla à l'élaboration d'un moteur de recherche utilisant une analyse syntaxique des textes et des requêtes, et par les équipes de l'Institut national de recherche en informatique et en automatique, qui publièrent deux ouvrages extrêmement riches concernant la recherche d'information sur Internet [Le Moal *et al*., 1996, 2002].

La recherche sur le web

Dès les débuts du web, à partir de 1993, des chercheurs essayèrent d'appliquer les travaux effectués en SRI pour développer des procédés d'indexation automatisés. Mais la différence majeure entre une collection documentaire classique et le web résidait dans le fait que sur ce dernier, la publication n'obéissait pas, ou trop peu, à des normes établies *a priori*. Les documents pouvaient être mis en ligne sans que rien n'obligeât leur auteur à respecter une quelconque convention, ce qui compliquait considérablement l'élaboration d'un traitement automatisé.

Un autre problème de taille est apparu dans le cas des premiers moteurs consacrés au web : les concepteurs de sites devaient prévenir à chaque fois qu'ils créaient une nouvelle page s'ils voulaient que celle-ci puisse figurer dans l'index. Autrement dit, l'action de l'éditeur en amont était une condition *sine qua none* pour que le moteur puisse fonctionner.

En 1993, les logiciels *JumpStation*, *World Wide Web Worm* et *Repository-Based Software Engineering* furent mis au point. Ils archivaient les titres, les accroches et les adresses URL des pages, mais les méta-informations n'étaient pas indexées de manière à pouvoir être retrouvées si le chercheur ne connaissait pas exactement les termes employés dans ces titres, accroches et adresses URL. Là encore, l'éditeur devait y songer au moment de publier les documents.

Le défi que constituait le web pour la SRI était également lié au nombre de documents, qui augmenta de manière exponentielle. Il n'y avait qu'une centaine de sites web pendant l'été 1993, puis quatre fois plus à la fin de l'année et encore quatre fois plus au milieu de l'année 1994 [Sanderson et Croft, 2012], pour atteindre 600 000 sites en 1996 [Battelle, 2005, p. 40]. Moins de dix ans plus tard, en 2003, le nombre de documents en ligne était de plusieurs dizaines de milliards [Picarougne, 2004, p. 21].

Au fur et à mesure qu'augmentait le nombre de documents, des progrès furent effectués en matière de paramétrage des systèmes d'indexation et des procédures automatisées. De nombreux logiciels virent le jour dans les années 1990 : *Excite* (1993), *Infoseek* (1994), *Lycos* (1994), *Webcrawler* (1994), *Altavista* (1995), *Excite* (1995), *Echo* (1996), *Ask Jeeves* (1997), *Google* (1998) et *AllTheWeb* (1999). Le premier à scanner des pages entières fut *Webcrawler*, qui permit de raffiner en nombre et en qualité les descripteurs. *Altavista* fut quant à lui le premier moteur à permettre au grand public d'effectuer des recherches en langage naturel (« Quelle est la couleur du cheval d'Henri IV ? »). Les méta-moteurs, combinant les résultats de plusieurs moteurs, firent leur apparition avec *Hotbot* (1996) et *Dogpile* (1996).

Succès des annuaires et des portails

Malgré le nombre de solutions proposées, les moteurs ne fonctionnaient pas encore suffisamment bien dans la deuxième moitié des années 1990 pour répondre à toutes les requêtes [Brown et Duguid, 2000, p. 41-44]. Le mieux, pour s'y retrouver, semblait encore de consulter les documents disponibles en naviguant de page en page, puis de les thésauriser en éditant manuellement un annuaire de liens hypertextes. C'est ce qui fit le succès de *Yahoo*.

David Filo et Jerry Yang, doctorants à Stanford University, rendirent disponibles en février 1994 des liens vers leurs sites favoris classés par catégories et sous-catégories, en accompagnant chacun d'eux par une description rédigée à la main. Ils nommèrent la plateforme : *Yahoo*. Très rapidement, l'annuaire reçut des centaines de milliers puis des millions de visiteurs par jour.

La stratégie dite « de portail » se développa après l'arrivée sur le web des acteurs historiques du monde des médias et des télécommunications [Meisel et Sullivan, 2000 ; Blevins, 2004]. Elle consistait à proposer sur un même site des liens vers les contenus appartenant au propriétaire du site en question, ainsi qu'un annuaire de liens pointant vers d'autres contenus et un moteur de recherche plus ou moins efficace [Van Couvering, 2008].

La mise en réseau des ordinateurs à très grande échelle et la possibilité pour leurs utilisateurs de publier et de lier des documents avaient ainsi abouti à un retour du classement par thèmes et sous-thèmes, effectué manuellement. Le succès des annuaires dura jusqu'à la fin des années 1990, après quoi le nombre de documents était devenu trop important pour que cette méthode ne soit pas aléatoire et insatisfaisante.

Analyse des liens

A partir de 1996, des chercheurs comme Robin Li et Jon Kleinberg planchèrent sur une nouvelle méthode d'appariement automatique entre une requête et des documents, qui consistait à exploiter les informations contenues dans la structure hypertextuelle. En répertoriant les liens entre documents, et en ajoutant cette information aux descripteurs fournis par l'analyse statistique des textes, la SRI réussit à obtenir des résultats plus pertinents. Ce fut le début de ce qu'Olivier Ertzscheid nommerait plus tard « la recherche d'information augmentée » [Ertzscheid *et al.*, 2009].

Pionnier en la matière, Jon Kleinberg détermina l'importance que l'on devait attribuer à la position des pages les unes par rapport aux autres en se référant au mode d'organisation du *Science Citation Index* (SCI), qui consiste à attribuer une valeur à une publication proportionnelle au nombre de publications qui la citent. Il développa pour le laboratoire Almaden d'IBM l'algorithme HITS (*Hyperlink-Induced Topic Search*) dont le fonctionnement se basait sur l'identification de certaines pages comme des « *authorities* », quand de nombreuses pages pointaient vers elles, et d'autres comme des « *hubs* », lorsqu'elles contenaient des liens vers les *authorities* [Kleinberg, 1999 ; Kleinberg *et al*., 1999]. Chaque document se voyait ainsi attribuer deux valeurs : l'une utilisée pour déterminer sa pertinence (*authority*), l'autre pour indiquer la pertinence des liens pointant de ce document vers d'autres documents (*hub*).

Le projet CLEVER (*Clientside Eigenvector Enhanced Retrieval*) permit d'améliorer le fonctionnement de l'algorithme HITS en raffinant la définition de ce qu'étaient des *hubs* et des *authorities*. D'autres ajustements furent proposés, à la suite notamment du projet SALSA (*Stochastic Approach for Link-Structure Analysis*), lancé à Haïfa en 1997, qui permit d'appliquer un méta-algorithme aux *hubs* et aux *authorities* afin de retenir ceux qui, parmi eux, étaient les *hubs* et les *authorities* les plus pertinents.

HITS fut utilisé par plusieurs moteurs de recherche destinés au grand public, en particulier *Teoma*, *Wisenut* et *Webfontain*, lancés au début des années 2000.

PageRank

En 1997, Larry Page conçut un logiciel nommé « *BackRub* », capable de suivre les liens allant d'une page vers une autre page, et de trouver, pour chaque page, les liens pointant vers elle. Larry Page utilisa les informations collectées à propos du maillage hypertexte pour concevoir, avec Sergey Brin, doctorant comme lui à l'Université Stanford, l'algorithme de SRI PageRank. Reprenant les idées formulées par Jon Kleinberg, ils considérèrent que chaque lien pointant vers une page, et identifié par *BackRub*, était un vote pour cette page. Plus une page recevait de votes, plus elle serait considérée comme étant pertinente, et plus son vote, lorsqu'elle pointerait elle-même vers d'autres pages, aurait de la valeur [Brin et Page, 1998 ; Page *et al*., 1999].

Brin et Page couplèrent leur analyse avec une méthode d'analyse statistique [Bianchini *et al*., 2005]. Le texte de chaque page était analysé en entier tandis qu'une attention particulière était consacrée au texte des liens eux-mêmes, appelé « ancres ». Le moteur fut nommé *Google*, en référence au terme mathématique « *googol* », qui désigne le nombre 10 à la puissance 100.

La particularité de *Google*, en plus de donner des résultats jugés plus satisfaisants que ceux des moteurs *Excite* et *AltaVista* [Battelle, 2005], était de considérer le web comme un système social à part entière. « Chaque document du corpus est considéré comme un membre d'un réseau ou d'une société stratifié(e), et ce avant même que la moindre requête n'ait été formulée. Le concept central en SRI de *pertinence* – lié à un "besoin d'information" spécifique – est complété par le concept sociométrique de *statut* et d'*autorité* » [Rieder, 2012].

L'algorithme PageRank est à la fois « politique » [Introna et Nissenbaum, 2000] et « moral » [Cardon, 2013, p. 65]. De nombreuses critiques lui sont adressées, notamment en raison du fait qu'il ait tendance à avantager les pages déjà visibles [Hindman *et al*., 2003]. Car en effet, plus un contenu est visible sur *Google*, plus il attire de visiteurs, et plus les probabilités sont fortes pour qu'il attire des liens, ce qui renforcera sa visibilité dans les listes de résultats. Au final, il apparaît que 90% de l'autorité PageRank est possédé par 10% des sites web [Pandurangan *et al.,* 2006, *cit. in* Cardon, 2013].

PageRank : à la croisée des disciplines

Le principe du PageRank s'enracine à la fois dans les traditions de la sociométrie et de la scientométrie. Dans un remarquable travail généalogique, Bernhard Rieder montre comment la formule renvoie aux travaux sociométriques de Moreno dans les années 1930, à leur outillage mathématique par Elaine Forsyth et Leo Katz au sein de la théorie des graphes, ainsi qu'aux méthodes bibliométriques conçues par Gabriel Pinski et Francis Narin à partir des travaux d'Eugène Garfield sur le *Journal Impact Factor* et le *Science Citation Index* [Rieder, 2012].

« Même si les échanges entre ces deux traditions [sociométrique et scientométrique] n'ont pas été très nombreux, a expliqué le sociologue Dominique Cardon à la suite de l'article de Bernhard Rieder, celles-ci convergent au moins sur la question qui sera décisive dans l'élaboration du PageRank : définir des métriques destinées à décrire les formes relationnelles du social. Que ce soit par le truchement de l'influence en sociométrie ou de la citation en scientométrie, un déplacement s'opère pour ne pas faire porter l'analyse sur des objets fixes et autosuffisants, qu'il s'agisse d'acteurs sociaux ou de documents, mais sur les relations qu'ils entretiennent les uns avec les autres » [Cardon, 2013, p. 67]

Dépositaires de traditions scientifiques préexistantes au web, les concepteurs de *Google* ont puisé leurs ressources dans différents champs disciplinaires de manière à combler les lacunes de la seule analyse statistique dès lors qu'il s'agissait d'automatiser le processus de recherche d'information sur le web. Au final, le fonctionnement du moteur est le fruit d'une fécondation

croisée entre sciences informatiques, sociologie des réseaux, sciences du langage, mathématiques probabilistes et théorie des graphes.

Fin de la transparence

Les ingénieurs employés par Google et par ses concurrents continuèrent à perfectionner leurs moteurs. Toutefois, plus la SRI progressait, et moins il était possible pour les observateurs avertis comme pour le grand public de connaître dans le détail les progrès effectués. Les années 2000 furent marquées par le glissement du paradigme de la transparence académique vers celui du secret industriel.

Ce glissement s'explique étant donné l'enjeu financier d'une part, qui n'était plus le même sur le web que dans les bibliothèques, et d'autre part parce que les concepteurs des moteurs craignaient que les éditeurs de contenus, dans le cas où ils en auraient trop su à propos des critères algorithmiques permettant d'évaluer la pertinence des documents, ne tentent de biaiser artificiellement les résultats en leur faveur.