Archivage du Web

Pierre Senellart

Archivage du Web

Pierre Senellart

Contenu de l’article

Le World Wide Web est la plus vaste source d’informations ayant jamais existé. Mais ces informations sont très *volatiles* : la moitié du contenu disparaît en moins de quelques années. Ce caractère éphémère de l’information est encore plus évident dans le Web social, dont le contenu est fourni par les utilisateurs du Web, et dans lequel l’accès à l’information, contrôlé par quelques grandes entreprises (cf. VII.6), n’est parfois plus possible après un certain délai.

L’*archivage du Web* est un processus de collecte, de sélection, d’enrichissement, de stockage, de préservation et de mise à disposition des informations du Web actuel, afin qu’elles restent accessibles aux utilisateurs dans l’avenir. L’objectif de cette démarche est de permettre, par exemple, à un historien dans trente ans de pouvoir étudier la manière dont un événement politique a été commenté par les parties prenantes, les médias et les simples utilisateurs du Web ; à un juge de pouvoir décider dans cinq ans si telle action était en violation des termes d’utilisation d’un service Web tels qu’ils étaient formulés à l’époque des faits ; ou encore, à un sociologue de réaliser dans vingt ans une étude diachronique d’une communauté à travers les traces que cette communauté a laissées sur le Web.

Des organisations à but non lucratif, telles qu’Internet Archive aux États-Unis (qui met à disposition une archive temporelle de plusieurs centaines de milliards d’URL), Internet Memory en Europe, des bibliothèques nationales comme la BNF et d’autres organismes, tels que l’INA en France, œuvrent à l’archivage de parties du Web au sein de l’IIPC (*International Internet Preservation Consortium*). Ils utilisent pour cela des logiciels de parcours du Web nommés « robots Web » ou « crawlers ». Malgré cela, une grande partie de l’information volatile du Web n’est pas archivée et devient inaccessible aux générations futures.

Un Web infini et en constante mutation

Il n’y a pas de limite au nombre de pages Web existantes, chacune associée à une URL (adresse Web) distincte. Par exemple, un service d'agenda en ligne comprenant un lien « mois suivant » permet de parcourir le calendrier de mois en mois et donc de cliquer un nombre arbitraire de fois sur le lien, ce qui engendre de nouvelles pages Web, toutes à des URL distinctes. Une telle succession sans fin de pages Web est appelée « piège à robot ».

Cette infinité virtuelle du Web rend impossible la tâche d’archivage de l’intégralité du Web. Il faut donc mettre en place des stratégies de collecte sélective, de filtrage, qui se concentrent sur certaines pages, et implémenter des robots qui évitent de tomber dans les pièges que nous venons de décrire, soit en limitant le nombre de pages par site Web, soit en mettant en place des parcours du Web les plus *larges* possibles. Une alternative, plus ambitieuse, est de se focaliser sur les pages Web *importantes*, ou sur celles traitant d’un sujet donné. Cette tâche est délicate, car on ne peut connaître

l’intérêt d’une page Web qu’une fois celle-ci récupérée. Il faut donc mettre en place des techniques de fouille de données (cf IV.7) et d’apprentissage statistique (cf. IV.9) pour estimer cet intérêt.

Par ailleurs, outre son infinité, le Web est en constant changement. Certaines informations disparaissent quelques heures, voire quelques minutes, après avoir été publiées. D'autres pages Web, à l’inverse, ne changent pas pendant des décennies. Il est crucial pour un robot Web de capturer cette dynamique, en revisitant régulièrement les pages changeant régulièrement, tout en visitant plus rarement celles qui sont plus stables. Mais cela demande de pouvoir prédire le taux de changement d’une page, pour en déduire un taux de revisite optimal. Trop de revisites inutiles entraînent un coût important en utilisation du réseau et en temps de calcul ; trop peu de revisites peuvent rendre une archive Web incomplète.

Les contraintes de collecte, d’accès et de stockage

Les robots Web traditionnels ont une vision un peu archaïque : ils considèrent que chaque page Web distincte d’un site est formée de contenu statique distinct de tous les autres contenus se trouvant sur le même site. Or les sites Web modernes présentent l’information d’une manière de plus en plus dynamique, avec, par exemple, des commentaires chargés au milieu d’une page Web au moment où l’utilisateur clique sur un bouton, au lieu de les présenter sur une page à part. Ces sites Web présentent également l’information de manière redondante : plusieurs URL distinctes peuvent inclure le même contenu. Le robot doit donc être capable de réaliser des interactions complexes avec une page Web (ce qui nécessite des technologies avancées, comme un « orchestrateur de navigateur ») et d’identifier les redondances d’un site de manière à le parcourir plus efficacement.

De plus, le Web actuel est devenu en grande partie un Web social et la majorité des contenus sont publiés par des internautes sur des sites de réseaux sociaux, comme Twitter ou Facebook. Il s’avère délicat, voire impossible, de collecter l’information de ce genre de site avec des outils traditionnels. À la place, il est souvent indispensable d'utiliser les interfaces de programmation (API) fournies par les sites de réseaux sociaux, qui permettent de récupérer une description structurée de contenu. Mais ces API viennent presque toujours avec de très fortes restrictions d’accès. Par exemple, il n’est pas possible de faire plus de 450 recherches sur Twitter toutes les 15 minutes. Il est donc indispensable d’optimiser l'utilisation des API de réseaux sociaux, afin d'archiver le plus grand nombre d’informations pertinentes dans un temps le plus court possible.

Enfin, une fois une archive Web constituée, se pose le problème du stockage pérenne. Un premier problème est un problème classique auquel sont confrontés les bibliothécaires et archivistes, quel que soit le contenu : comment faire en sorte que les supports physiques de l’information ne se détériorent pas et restent lisibles pour les générations futures ? Cela passe par de la réplication de l’information, des sauvegardes régulières sur supports optiques ou bande magnétique en plus de la copie principale de l’archive sur disque, ainsi que par l’utilisation de formats standard, comme WARC, pour représenter le contenu et méta-données d’une archive. Un problème plus spécifique au Web est celui de la taille gigantesque des archives, en particulier pour des archives génériques du Web. Ainsi, on estime à plusieurs pétaoctets la taille de l’archive constituée par Internet Archive. Le stockage et l’accès à l’archive requièrent donc une architecture distribuée, formée de grappes de plusieurs milliers de machines.

Bibliographie

- M. Faheem et P. Senellart, « Crawl intelligent et adaptatif d'applications Web pour l'archivage du Web ». *Ingénierie des Systèmes d'Information*, vol. 19, nº 4, p. 61-86, 2014.
- G. Illien. « Une histoire politique de l'archivage du web. » *Bulletin des bibliothèques de France* nº 2, 2011. http://bbf.enssib.fr/consulter/bbf-2011-02-0060-012
- J. Masanès. *Web Archiving.* Springer, 2006

Affiliation

Pierre Senellart. Informaticien. Professeur à l'École normale supérieure, Département Informatique, Paris. pierre@senellart.com

Illustration

Légende : Architecture d'un robot Web

Formats : Fourni aux formats PDF et SVG, le SVG peut-etre

Source et licence : Je suis l'auteur du schéma que j'ai conçu spécifiquement pour cet article. Les sous-composants utilisés sont tous libres de droits (issus de la bibliothèque Open Cliparts https://openclipart.org/).

Glossaire

- API : *Application Programming Interface*. Interface permettant l'accès à un service (par exemple un site de réseau social) par un programme.
- Archivage du Web : Processus de collecte, sélection, enrichissement, stockage, préservation et mise à disposition des informations du Web actuel, afin qu'elles restent accessibles aux utilisateurs dans l'avenir.
- Étude diachronique : Étude (par exemple d'un corpus ou d'un phénomène social) sur une longue période, en s'intéressant en particulier aux variations apparaissant à travers le temps.
- Orchestrateur de navigateur : Logiciel permettant de commander par un programme l'interface d'un navigateur tel que Mozilla Firefox ou Google Chrome, afin de reproduire une navigation sur un site Web de manière similaire à ce que ferait un internaute.
- Ordonnanceur (robot Web) : Partie d'un robot Web responsable de la sélection de l'ordre dans lequel les URL doivent être téléchargées.
- Robot Web (ou crawler) : Logiciel parcourant le Web de manière autonome afin de collecter méthodiquement des pages Web.
- URL : *Uniform Resource Locator*. Identifiant unique d'une ressource sur le Web, également appelée « adresse Web », par exemple http://www.example.com/chemin/vers/information.
- WARC : *Web ARChive format.* Format standard, normalisé par l'ISO, pour la représentation d'archives Web.

- Web social :Partie du Web formée des sites de réseaux sociaux et plus généralement de l’ensemble des sites permettant à des internautes arbitraires de fournir du contenu.