

« Vers un web de connaissances »

Selon **Serge Abiteboul***, le web de demain ne permettra plus seulement de trouver des informations. Il créera de la connaissance en nous assistant dans les tâches de recherche et de raisonnement.

A quoi ressemblera le web de demain ?

Peu de gens avaient prévu les évolutions actuelles. Cela nous incite à rester modestes dans l'anticipation. Aujourd'hui, le web est un réseau de documents écrits dans un langage, le html [qui permet de créer des liens entre les pages, ndr]. Des moteurs de recherche permettent de localiser ces pages afin que des êtres humains y trouvent l'information pertinente. Demain, la tendance serait d'aller vers un web non plus seulement d'informations, mais de connaissances.

On parle de web sémantique.

Qu'est-ce que cela veut dire ?

Une connaissance, c'est de l'information structurée, organisée. Cela peut aller d'un simple numéro de téléphone jusqu'au parcours complet d'un individu ou à l'organisation d'un voyage sur mesure. L'idée est que les machines nous remplaceront dans des tâches de recherche d'information, voire de raisonnement.

Par exemple, aujourd'hui, si j'ai besoin de retrouver votre numéro de téléphone, je vais fouiller dans un ancien e-mail pour savoir où vous travaillez, puis aller sur le site de votre employeur et y dénicher le fameux numéro. Demain, il me suffira de poser la question à un logiciel : quel est le numéro de téléphone professionnel de monsieur X ? Mais pour que les machines nous aident, il faudra que le web soit construit pour elles, c'est-à-dire autour d'un langage compréhensible par des algorithmes.

Devra-t-on réécrire tout le web ?

Il y a deux façons d'arriver à la publication de connaissances



FRANÇOIS GUENET

« Sans cesse émergent de nouvelles et passionnantes utilisations »

sur le web. Soit directement, dans des pages web. On utilise pour cela des balises, comme en html, pour que les logiciels s'y retrouvent. Par exemple, en html, le repère indique au navigateur d'écrire le texte en gras. On peut imaginer une balise <numéro de téléphone> qui indiquera que le contenu est un numéro de téléphone. Cela suppose de définir les mêmes balises pour tout le monde, ou d'être capable de traduire d'un langage de balises à un autre. L'autre voie consiste en l'extraction de connaissances directement des pages existantes. Cette « fouille de texte » marche de mieux en mieux, mais avec lenteur et des erreurs. Pour reprendre notre exemple, il s'agit de créer des robots informatiques capables de reconnaître dans du texte des associations entre des

personnes et leurs numéros de téléphone. Cette approche, plus complexe et moins fiable, présente par contre l'avantage de ne pas devoir attendre des individus et des entreprises qu'ils acceptent de jouer le jeu et de mettre en ligne non seulement de l'information, mais aussi ces connaissances.

D'autres tendances se profilent-elles ?

Des bannières comme « web 2.0 », « web sémantique », « web 3.0 » prolifèrent. On perd du temps à comprendre ce qui se cache derrière ; le plus souvent du marketing. Elles témoignent cependant de la richesse du web. De nouvelles utilisations, passionnantes, émergent sans cesse. La mode est aux réseaux sociaux, aux échanges d'informations entre internautes. Une autre tendance

est le « web des objets », où le moindre objet a une adresse internet. On imagine ainsi des réfrigérateurs qui vous informent directement de ce qu'il reste sur leurs étagères, voire passent des commandes.

En quoi la recherche en informatique peut-elle contribuer à ces développements ?

Le web actuel est déjà un beau défi. Ce n'est pas la même chose de faire fonctionner un algorithme sur quelques machines ou des milliards de pages sur des millions de serveurs. La recherche en informatique est là pour répondre à ces défis. Ce n'est pas un hasard si un moteur de recherche comme Google est sorti d'un laboratoire universitaire, celui de Stanford. Les innovations autour du web posent des questions. Par exemple, la sécurité des informations sur un vaste réseau n'est pas du tout garantie. En ce qui me concerne, un problème fondamental me préoccupe. Les bases théoriques sur lesquelles repose la gestion de l'information datent des années 1960. Elles ne sont pas adaptées à la structure d'information distribuée actuelle à la base du web. Nous nous en sortons, mais c'est plus de la cuisine que de la science. Trouver des fondements mathématiques adaptés aux données du web, c'est une question particulièrement stimulante.

PROPOS RECUEILLIS
PAR DAVID LAROUSSE

* Directeur de recherche à l'Inria Saclay, responsable de l'équipe Gemo sur la gestion de données et de connaissances distribuées.