Un aperçu du Web sémantique

Introduction

Lorsque Tim Berners-Lee a jeté les bases du Web actuel en 1989, son objectif était de simplifier l'accès aux informations disponibles sur l'Internet naissant au moyen d'un seul logiciel de visualisation appelé navigateur (*browser*).

En septembre 1998, ce chercheur anglais a proposé une nouvelle vision du Web, le Web sémantique, dont l'idée est d'organiser les informations disponibles de telle sorte qu'elles deviennent exploitables également par les machines.

Les bénéfices attendus sont d'une part une meilleure efficacité dans la recherche d'informations, grâce à l'élimination des mauvaises réponses des moteurs de recherche causées par les ambiguïtés du langage naturel (polysémie) et d'autre part une meilleure compréhension du contenu par les machines, grâce à la formalisation des informations.

L'ambition du présent texte n'est pas d'expliquer dans le détail le fonctionnement du Web sémantique — c'est encore un domaine de recherche où tous les problèmes n'ont pas été résolus. Il s'agit plutôt d'en donner un bref aperçu pour en comprendre les origines et les fondements théoriques. C'est pourquoi le jargon technique a été réduit ici au maximum. De même, nous n'aborderons pas les outils nécessaires à l'exploitation des ressources du Web sémantique (agents intelligents, langages d'interrogation, etc.).

Avant toute chose, il nous semble utile de revenir sur la signification même des mots « Web » et « sémantique » avant d'évoquer les technologies mises en œuvre.

Quelques définitions

Rappelons que le Web, qui s'est imposé dans le langage courant par rapport à l'expression initiale *World Wide Web*, est un système d'information constitué de documents multimédia reliés entre eux par des

hyperliens[1]. Il ne faut donc pas confondre le Web avec l'internet (*Interconnected Networks*), autrement dit l'infrastructure technique qui relie physiquement les différents réseaux d'ordinateurs dans lesquels sont stockés ces informations.

Pour un linguiste, la sémantique est l'étude du sens d'un texte à partir de la combinaison des mots. En intelligence artificielle, la sémantique « porte sur la capacité d'un réseau à représenter de la manière la plus humaine possible des relations entre des objets, des idées ou des situations »[2].

Le Web sémantique serait donc un système d'information constitué de documents multimédia (le web) dans lequel ces documents sont unis entre eux par un lien porteur de sens (la sémantique).

Quelle différence avec le Web que nous connaissons aujourd'hui puisque les documents sont déjà liés entre eux ? En fait, les hyperliens actuels indiquent juste l'emplacement d'une ressource et n'ont qu'une seule signification : *telle ressource X « est reliée » à telle ressource Y*.

Pour bien comprendre l'importance de cette différence *a priori* minime (le fait d'ajouter simplement une signification, un sens à un lien), il nous faut regarder du côté des sciences cognitives et plus précisément de l'intelligence artificielle dont l'objectif vise la reproduction par une machine des processus mentaux humains comme la compréhension, la perception, ou la décision.

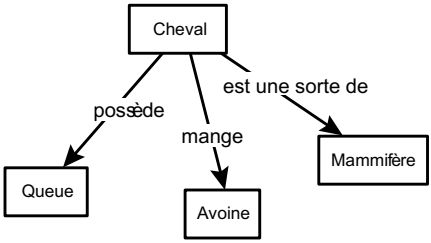
Modéliser la mémoire

L'une des façons de modéliser les connaissances nécessaires à une machine « intelligente » pour mener à bien une tâche repose sur un postulat selon lequel tout individu raisonne et apprend par association d'idées. Ainsi, notre mémoire pourrait être représentée par un ensemble de concepts (les idées) reliés entre eux (les associations) de façon hiérarchique (ex. un cheval est une sorte de mammifère) ou non (ex. un cheval possède une crinière et mange de l'avoine).

---

[1] Initialement appelés liens hypertextes, c'est-à-dire étymologiquement « liens au-delà du texte » d'après la racine grecque *huper*. Aujourd'hui, il est techniquement possible de créer des liens depuis n'importe quelle partie d'un document (image, animation Flash, etc.), d'où le nom plus générique « hyperlien ».

[2] cf. Le Grand dictionnaire terminologique (www.granddictionnaire.com)



Une telle représentation de la mémoire a été développée par Ross Quillian en 1968 dans un article intitulé *Semantic Memory*.

En ajoutant simplement une signification aux hyperliens connectant deux ressources, le Web sémantique devient l'équivalent d'une mémoire sémantique, autrement dit une représentation de la connaissance humaine à l'échelle mondiale.

Encadré : Le Web sémantique comme super cerveau ?

Il est tentant de penser que le Web sémantique viserait à transformer le Web actuel en un gigantesque cerveau (*Super Brain*). Il y a cependant encore loin de la coupe aux lèvres.

D'abord, le modèle de mémoire sémantique souffre de plusieurs défauts pour représenter précisément notre mémoire : difficulté à gérer les connaissances dynamiques, complexité à décrire et utiliser des connaissances procédurales, etc.

Ensuite, le Web sémantique s'attache à faciliter l'accès aux informations, mais ne définit pas comment celles-ci doivent être traitées.

Les composants du Web sémantique

Si l'on assimile le Web à une bibliothèque, une façon pragmatique de définir les moyens d'accéder facilement à l'information consiste à s'inspirer des méthodes de travail (classification, etc.) et des outils développés par les documentalistes (thésaurii…).

Du thesaurus à l'ontologie

Un thesaurus est un sous-ensemble du langage utilisé dans la vie quotidienne, qui précise les relations entre les mots et les phrases (ex. termes génériques, termes spécifiques, termes préférentiels ou non,

termes associés, etc.). En général, ce thesaurus est restreint à un domaine précis comme la santé, l'éducation ou des documents gouvernementaux. Le fait qu'il normalise la signification des termes garantit ainsi qu'un concept sera toujours représenté de la même façon, indépendamment de la langue. C'est d'ailleurs l'une des raisons pour lesquelles les logiciels de traduction assistée par ordinateur (TAO) reposent justement sur un thesaurus.

Puisqu'un thesaurus relie des concepts par des liens porteurs de sens, c'est donc aussi à sa façon une mémoire sémantique. Transposé dans le domaine des sciences de l'information, cela devient une ontologie[1], autrement dit une « structure hiérarchique des connaissances établie à l'aide d'un ensemble de concepts précis pour créer un vocabulaire convenu permettant l'échange d'information »[2].

Encadré : Exemples d'utilisation d'une ontologie[3]

- Portails Web — Règles de Catégorisation utilisées pour améliorer la recherche ;
- Collections Multimedia — Recherche sur le Contenu pour des éléments média (sauf texte) ;
- Gestion d'un site Web institutionnel — Organisation et Taxinomique Automatique des données et des documents ; Association des éléments de différents sites (acquisitions ou fusion) ;
- Gestion des Droits et Contrôle d'Accès

Des métadonnées pour « étiqueter les ressources »

Les ressources disponibles sur le Web peuvent être des textes complexes (mémoires de thèse, page d'accueil d'un portail thématique…), ou encore des données multimédia dont le sens échappe aux machines.

Tandis qu'un individu saura reconnaître par exemple le titre d'un morceau de musique en écoutant juste le début du fichier MP3 correspondant, il

[1] Ce sens donné au mot ontologie est relativement récent. Utilisée depuis des siècles par les philosophes, l'ontologie désigne la partie de la métaphysique qui étudie la nature et des relations entre les composantes existentielles.
[2] Source : thesaurus de la Bibliothèque Nationale du Canada.
[3] Source : http://www.w3.org/2003/08/owlfaq.html.fr

est nécessaire d'aider les machines à comprendre la nature de chaque ressource.

Une solution consiste à associer à ces dernières des informations complémentaires, par exemple le nom de l'auteur et le titre de l'œuvre musicale évoquée précédemment. De telles données qui décrivent d'autres données sont appelées des métadonnées.

La mise en œuvre

Les technologies utilisées

Toutes les technologies actuellement impliquées dans la mise en place du Web sémantique reposent sur XML (*eXtended Markup Language*), un langage de description de données formalisé en 1998. Sa capacité à créer d'autres langages, tous exploitables par un nombre réduit d'outils, simplifie considérablement les développements de nouvelles applications. De plus en plus d'applications utilisent XML pour la sauvegarde de leurs données, par exemple les suites bureautiques les plus récentes (*Microsoft Office 11*, *OpenOffice.org* et bientôt *KOffice.org*). À terme, il sera certainement possible d'accéder à toutes les données d'une organisation de façon unifiée.

Les applications

Les ressources actuellement disponibles sur le Web ont été générées soit par des hommes (sites Web personnel, carnets Web, etc.) soit par des machines (bases de données en ligne telles qu'horaires de train, pages jaunes, etc.).

Les premières, avec leur contenu non structuré, risquent de ralentir l'arrivée du Web sémantique. En effet, la mise en œuvre des métadonnées exige une grande rigueur, pas toujours exigible de la part des contributeurs. Par exemple, sur les cinq cents membres que compte le *World Wide Web Consortium*, seuls dix-huit disposaient d'un site Web conforme aux spécifications élaborées par ce même consortium, soit 3,6%[1]. De même, les propriétés des documents bureautique sont rarement remplies par les utilisateurs.

[1] Source : http://www.markokarppinen.com/20020222.html

En revanche, il sera plus facile de déployer le Web sémantique si les ressources sont gérées déjà par d'autres machines (données structurées). C'est pourquoi le commerce électronique est l'un des principaux bénéficiaires du Web sémantique, ainsi que les échanges d'informations entre systèmes informatiques.