L'intelligence du web

L'information utile à portée de lien

Jérôme Euzenat, INRIA & LIG

Le web en lui-même n'est pas un dispositif particulièrement complexe. Il s'agit en principe d'un ensemble de documents distribués sur un réseau dans lesquels il est possible de naviguer. Cependant, très tôt, les documents ont pu être engendrés dynamiquement lors de leur accès. Le même principe permet à des programmes d'accéder aux serveurs du web offrant ainsi des services web. De surcroit, l'interprétation de langages de scripts au sein des navigateurs permet de déporter l'interaction avec les données plus près de l'utilisateur. Le web peut alors être vu comme un ensemble de programmes communiquant entre eux. Même si une certaine tendance au sein des applications mobiles est de se passer du navigateur au profit d'applications légères, c'est bien le même principe qui est mis en œuvre. L'avènement des téléphones programmables avec accès données (oui, les SmartPhones) ne fait qu'étendre le périmètre du web.

Mais qu'en est-il de l'intelligence artificielle dans ce cadre ? Les utilisations toujours plus séduisantes que nous faisons du web s'appuient-elles sur des techniques ressortant de notre domaine ? Comme toujours la réponse n'est pas simple : les techniques d'intelligence artificielle ne sont diffusées auprès du grand public qu'embarquées dans des applications plus complètes (technologies du web, interface homme-machine, manipulation des données, etc.). Par ailleurs, elle utilisent de plus en plus de techniques provenant d'autres domaines (recherche opérationnelle, théorie des graphes, etc.).

Je considère ci-dessous deux angles particuliers permettant de répondre à ces questions : les applications fondées sur l'analyse de grandes quantités d'information et le développement d'un web sémantique. J'essaye ensuite de montrer comment l'IA peut contribuer aux tendances actuelles du web : web des données et réseaux sociaux.

La «sagesse des foules»

L'un des aspects particuliers du web est la quantité de données qui y réside et qui permet, grâce à un traitement massif, d'en extraire des informations fiables quant à ce qui peut être considéré comme populaire. Cette technique est en premier celle utilisée par le moteur de recherche de Google dans lequel les liens sont utilisés comme autant de votes en faveur des pages les plus populaires.

Cette pure analyse du réseau se complète d'une analyse du contenu manipulé : les mots apparaissant dans les liens, puis dans les pages sont autant d'indices du contexte de pertinence des liens. L'utilisation massive du moteur de recherche produit une rétroaction précieuse.

Diverses techniques issues de notre domaine, fouille de données, apprentissage automatique, souvent hybridées avec d'autres, statistiques, combinatoires, sont mises à profit pour faire parler les graphes. La puissance de la masse d'information est bien illustrée par le site 20 questions (http://www.20q.net/), utilisant des techniques de réseaux de neurones ou d'apprentissage d'arbres de décisions. Un courant, Web intelligence, jouant sur les deux sens du mot intelligence en anglais, s'est développé en liant les techniques d'intelligence artificielles et les technologies de l'information comme le web.

Disposer d'une telle quantité d'information et de la capacité d'analyse, peut être utilisé de manière très rentable, soit pour placer des annonces ciblées (Google), soit pour promouvoir d'autres produits que l'on vend (Amazon). Les techniques de fouille de données, en caractérisant le comportement des internautes dans des contextes précis, permettent de mieux cibler les annonces promotionnelles.

Ce principe s'applique aussi à plus petite échelle : les sites de questions-réponses (Stackoverflow) ou de recommandation de sites (Reddit) sont aussi fondés sur l'utilisation des avis des utilisateurs, et leur caractérisation, pour classer les meilleures réponses ou suggestions et ainsi amener un meilleurs confort d'utilisation. La qualité des réponses entraine une audience accrue qui entraine une meilleure qualité.

L'application de ce principe permet d'accéder à des recoins insoupçonnés de l'intelligence artificielle. Ainsi, la traduction automatique fondée sur l'apprentissage statistique implémentée dans Google translate. Elle utilise aussi une gigantesque base documentaire sur laquelle apprendre des traductions et le retour des internautes corrigeant les résultats.

Le web sémantique

Il y a plus de 10 ans, Tim Berners-Lee, l'homme à l'origine du web, lançait l'idée du web sémantique. Le mot sémantique n'était pas forcément le mieux choisi car il est utilisé de différentes manières, mais ce ne sera pas le premier ni le dernier. Il s'agissait de mieux qualifier l'information disponible sur le web de telle sorte que l'on puisse développer des algorithmes permettant d'en tirer parti. Les deux résultats recherchés étaient une recherche améliorée (il est plus facile d'obtenir l'information recherchée) et une plus grande interopérabilité (il est possible de l'utiliser en

conjonction avec d'autres informations). L'exemple paradigmatique du web sémantique était celui de la composition automatique de services pour planifier un voyage.

Le travail autour du web sémantique a permis de développer, souvent sous l'égide du W3C, des langages permettant de représenter formellement la connaissance dans le contexte distribué qu'est le web. Le cœur des technologies du web sémantiques est donc composé de :

RDF (Resource description framework) pour représenter l'information sous forme de graphes ;

RDFS/OWL (Web ontology language) pour modéliser le domaine de ces informations ;

SPARQL (Sparql query language) pour intéroger les sources d'information ;

Alignements pour exprimer les relations entre ontologies.

Il existe d'autres langages complémentaires (SKOS, RDFa) venant compléter ceux-ci. Des analyseurs, raisonneurs, évaluateurs pour ces langages sont disponibles et utilisés. Les technologies sémantiques développées pour le web sémantique ne sont pas cantonnées à l'implémentation d'un web sémantique. Ainsi, elles sont impliquées dans une myriade d'applications particulières :

Services web sémantiques pour décrire les services ;

Systèmes pair-à-pair sémantiques pour annoter les ressources partagées ;

Réseaux sociaux sémantiques pour représenter les relations sociales ;

Bureau sémantique pour décrire l'information personnelle (agenda, répertoire, etc.) ;

Intelligence ambiante pour exprimer l'information provenant de l'environnement ;

Web des données (liées) pour publier les données.

On pourrait estimer qu'autant de types d'applications tend à fragmenter le domaine et à entraver l'avènement d'un web sémantique. Mais l'utilisation de technologies communes fait que les sources d'informations développées dans un cadre sont exploitables dans un cadre différent. Ainsi, une ontologie géographique développée pour exporter l'information dans le web des données est utilisable dans une application de recherche d'information ou d'intelligence ambiante.

Les technologies du web sémantique offrent donc un espace d'adressage gigantesque (permettant de référer à des ressources bien définies) et des formalismes interopérables : une gigantesque base de connaissance. L'inférence, à cette échelle, est peu raisonnable, mais beaucoup de travaux ont pour but de lier harmonieusement raisonnement local et raisonnement global. De nouveaux problèmes, comme ceux liés à l'hétérogénéité de contenu entre sources de connaissances autonomes viennent se poser aux chercheurs.

Les mêmes compromis qu'en représentation de connaissance sont à l'œuvre dans ces langages : expressivité contre efficacité (ou décidabilité). Mais le gigantisme et la décentralisation du web repose le problème de la robustesse, offertes par les techniques statistiques évoquées dans la section précédente, contre la précision, apportée par les techniques du web sémantique. C'est certainement en articulant les deux approches que l'on obtiendra des résultats spectaculaires.

Le web des données

Le web sémantique n'a pas bénéficié de l'effet d'entrainement du web. Sans doute l'élaboration d'ontologies et l'expression de l'information dans les langages du web sémantique est-elle une étape plus difficile que l'encodage en HTML. À l'instar du web, le web sémantique ne sera intéressant que lorsque suffisamment de données y seront disponibles.

C'est pourquoi le principe du web des données ont été proposés. Il s'agit de publier des données sur le web de telle sorte que :

- Elles se réfèrent à des ressources identifiées ;
- Que les identificateurs permettent d'obtenir une description de la ressource ;
- Que les données publiées contiennent des liens vers d'autres jeux de données publiés.

Les technologies du web sémantique sont mises à profit dans ces jeux de données en identifiant les ressources par des URI, déréférencés en RDF via HTTP et en exprimant les liens à l'aide du prédicat OWL `sameAs`. Ces principes ont été adoptés par d'important pourvoyeurs de données. C'est le cas des dépositaires de données publiques, d'abord dans les pays anglo-saxons avec la publication des données data.gov et data.gov.uk mais maintenant dans le monde entier. C'est aussi le cas d'entreprises de presses telles Reuters, la BBC ou le New-York Times qui proposent leurs articles dans lesquels les entités nommées sont identifiées uniquement. De nombreuses sources de données d'intérêt général existent permettant de disposer d'identifiants de référence pour de nombreuses entités. On mentionnera dbpedia qui extrait ses descriptions de wikipedia ou geonames capable d'identifier des localisations géographiques.

À partir de ces données soigneusement publiées dans les formalismes précis du web sémantique, de nombreuses applications se développent qui tirent parti des liens entre les différents jeux de données. Ces applications vont de la visualisation de ces données sur des cartes, des échelles temporelles ou des graphiques à leur traitement plus auto-

matique par l'ajout d'ontologies ou de règles capable d'inférer à partir des données disponibles.

La disponibilité de telles quantités de données permet de surcroit d'appliquer les techniques du web d'une manière beaucoup plus assurée. Il est ainsi possible de chercher dans les données disponibles des corrélation entre les votes d'un député et les financements qu'il reçoit de sociétés (pas en France, c'est vrai).

Libérer l'intelligence des réseaux sociaux

Un des aspects très populaires du web à l'heure actuelle est celui des réseaux sociaux (Facebook, Twitter, LinkedIn, etc.). Leur principe est de matérialiser les liens interpersonnels. Le développement d'interfaces très faciles à utiliser a très certainement contribué au succès de ces applications.

Malheureusement, celles-ci restent fermées : les serveurs des sociétés développant ces applications captent les profils et relations de leurs utilisateurs. Ceci revient à recréer un web correspondant à chacun des réseaux sociaux au lieu du système d'information distribué qu'il était à l'origine. Le modèle de ces sociétés est de tirer parti des techniques développées dans les parties précédentes pour vendre des services à valeur ajoutée à des tiers.

Ces services de réseaux sociaux posent en particulier deux problèmes :

- Le manque d'interopérabilité : ces réseaux étant, selon la formule consacrée, des jardins clos, passer de l'un à l'autre est rendu coûteux. Il faut donc choisir le sien ou se résoudre à maintenir manuellement plusieurs profils. Cela restreint l'exploitation de cette «sagesse des foules» à quelques opérateurs ou à des applications cantonnées à un système particulier (ainsi l'expérience de classification fellows sur facebook : http://fellows-exp.com/).
- Le manque de respect de la vie privée : les utilisateurs sont amenés à confier aux sociétés en question, des informations qu'ils ne voudraient confier qu'à certaines personnes (leurs amis). Ceci est un problème qui se révèle lorsque les sociétés confient, volontairement ou involontairement, cette information à des tiers.

Ces problèmes ne sont pas uniquement théoriques et tendent à s'accentuer dans l'utilisation des téléphones : il devient de plus difficile de contrôler quelle information stockée sur ces appareils est accessible et par qui (applications, opérateurs, autres utilisateurs, etc.). En général, la lecture des conditions d'utilisations n'aide pas à en avoir plus. Pourtant, ces deux aspects devraient pouvoir être pris en compte par des technologies dédiées et, en particulier, celles développées dans le cadre du web sémantique. D'une part, on l'a vu, il existe des technologies versatiles pour exprimer le type d'information que l'on exprime sur les réseaux sociaux. RSS (Really Simple Syndication) a d'abord été développé en RDF, FOAF (Friend-of-a-friend) est un schéma de base pour décrire les réseaux sociaux (profils et relations). Elles peuvent permettre cette interopérabilité.

Bien entendu, cette interopérabilité n'est pas sans danger : c'est très clairement une menace pour le caractère privé de l'information.

Mais les technologies sémantiques sont aussi particulièrement adaptées pour exprimer les politiques de chaque individu quant à la divulgation de l'information dont il dispose et pour contrôler de manière simple la diffusion de son information, personnelle ou non, dans des sphères particulières. En effet, elles sont suffisamment puissantes pour décrire très précisément les conditions de diffusions («aux membres de ma famille restreinte et à mon médecin») et suffisamment flexibles pour définir des catégories adaptées («à mes connaissances se trouvant dans un rayon d'un kilomètre»). Elles permettent donc de définir des politiques de diffusion qui expriment un contrôle très fin de l'information : «mes disponibilités au travail peuvent être communiquées à mes collègues pour organiser une réunion de travail, pas à un fournisseur pour me démarcher».

En ce qui concerne l'implémentation de telles perspectives, il existe déjà des solutions, intégrées dans l'infrastructure du web, telles que FOAF+SSL permettant d'identifier les connexions, et donc les demandes d'informations, en fonction des relations indiquées à l'aide de fichiers FOAF. Les travaux développés en systèmes multi-agents peuvent aussi être utilisés pour négocier l'accès à l'information, soit sur la base de politiques, soit sur celle de principes plus généraux.

Cette courte présentation est délibérément dénuée de références. Le lecteur intéressé pourra utiliser le web pour trouver de nombreux points d'entrée.

Addendum : le meilleur chemin du sous-bois au palais

Crédit photo : © Hans Hillewaert, 2005 / CC-BY-SA-3.0

On m'a demandé d'illustrer mon propos par quelques applications. En particulier, l'une des suggestions était : « J'ai pris une photo d'un champignon et j'aimerais savoir ce que c'est ». Que peut le web sémantique ou la sagesse des foules dans ce cas ? Il me semble que cela dépend. Identifier un objet à partir d'une photographie est un problème délicat. D'autre part, il faut savoir ce que signifie identifier. La réponse peut être apportée avec différents niveaux de détails : «c'est un champignon», «c'est un cèpe», «c'est un *Boletus edulis*», «c'est le champignon photographié par Hans dans la forêt de Tillegem» me semblent des réponses appropriées. Enfin, la question n'est peut être pas la vraie question qui pourrait être : puis-je le ramasser (légalement, sanitairement ?), comment puis-je le cuisiner ?

À l'heure actuelle, il est possible de confier cette photographie à la sagesse des foules : la soumettre à un forum consacré à la mycologie ou à la cuisine, risque d'amener la réponse appropriée assez rapidement. Au pire, elle ouvrira une controverse (vraisemblablement plus dans le forum consacré à la cuisine) qui ne pourra être qu'enrichissante. Une fois identifiée la soumettre à un moteur de recherches ouvrira de nouveaux horizons ; il sera aussi plus aisé de chercher l'objet au sein du web sémantique. Dbpedia offrira un premier identifiant (http://dbpedia.org/resource/Boletus_edulis) à partir duquel naviguer. L'information risque d'y être plus pauvre et moins adaptée que celle fournie dans le web des humains. Cependant, l'utilisation du web des données à partir de tels identificateurs permettra plus surement de faire le lien entre la cuisine, la mycologie, la randonnée et la médecine que l'utilisation des forums. On peut noter que sous dbpedia il appartient à la catégorie "French_cuisine" alors que la page en français de wikipedia ne mentionne que brièvement les utilisations culinaires.

Mais le web sémantique, s'il permet de répondre à une telle demande de la part d'un spécialiste (à l'aide de SPARQL) nécessite des applications plus adaptées à l'accès du grand public. Une telle application pourrait permettre d'enrichir automatiquement ma base de photographies de champignons (pour l'instant à partir de leurs noms). Il est alors possible de les identifier, d'en ajouter les caractéristiques à partir de dbpedia, d'identifier les recettes dans lesquelles il intervient et de signaler l'endroit et la date à laquelle le spécimen a été photographié.

Malheureusement, le gestionnaire de photographies du téléphone de Hans, n'est pas capable d'échanger de l'information avec son navigateur autrement qu'en envoyant le cliché. Il ne peut pas non plus, produire des annotations du cliché qui permettront à d'autres utilisateurs du web (ou à elle-même lorsqu'elle cherchera des photos de champignons) de le trouver facilement.