L'Extraction d'Information Ouverte (OIE) : le Nouveau Graal des Moteurs de Recherche

Retour au sommaire de la lettre

Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

Depuis que les moteurs de recherche existent, leur procédure d'interrogation est immuable : taper des mots clés dans un formulaire de recherche. Depuis quelques années, des outils comme Siri ou Google Now changent petit à petit la donne en intégrant des interfaces vocales à ces investigations. Mais tout cela pourrait aller bien plus loin à l'avenir avec les techniques d'OIE ou Extraction d'Information Ouverte, qui mettent en relation les mots, les faits et les concepts et qui pourraient révolutionner notr efaçon de nous adresser aux moteurs. Les obstacles sont encore nombreux mais les objectifs passionnants...

En avril 2013, Google a fait l'acquisition d'une nouvelle startup à l'origine de l'application mobile **Wavii**. Cette acquisition est apparue comme une tentative de « contrer » l'acquisition par Yahoo ! de Summly. Mais en réalité, la technologie « embarquée » dans l'application Wavii est tout à fait différente : il s'agit tout bonnement de l'*Open Information Extraction* (Extraction d'Information Ouverte), une approche qui pourrait tout bonnement être à l'origine des moteurs de recherche du futur. On peut donc logiquement soupçonner Google d'avoir voulu mettre la main sur certains brevets, et certaines compétences, pour pouvoir utiliser l'OIE non seulement dans une application mobile (qui n'est déjà plus disponible), mais également pour améliorer son « knowledge graph » qui est aujourd'hui déployé dans une version très limitée.

Mais qu'est-ce exactement que l'OIE ? A quoi cela sert-il et pourquoi est-ce susceptible de révolutionner le fonctionnement des moteurs de recherche ? C'est ce que nous allons essayer de vous expliquer dans cet article.

Ne pas confondre « Information Extraction » et « Information Retrieval »

Mais commençons par un « caveat » (un avertissement) : en français, nous avons pris la mauvaise habitude de traduire « information retrieval » par « extraction d'information ». Le problème est qu'il existe un autre secteur de recherche en informatique, baptisé en anglais « information extraction ». Et les deux domaines de recherche sont très différents.

L' « **information retrieval** » fait allusion aux techniques de recherche d'information dans les documents utilisées par les moteurs de recherche actuels. Par la suite, nous utiliserons les termes « **informatique documentaire** », ou « **recherche documentaire informatisée** ».

L' « **information extraction** » fait allusion à des techniques plus ou moins automatiques de « **web mining** » (« fouille du web »), permettant de créer à partir des textes trouvés dans les pages web des bases de faits exploitables. Dans la suite de cet article, nous désignons ce domaine de recherche par les termes « **extraction d'information** ».

L'extraction d'information : un concept ancien mais souvent ignoré

L'extraction d'information n'est pas un domaine de recherche récent : les premières tentatives de création de bases de connaissances à l'aide de documents publiés sur le web remontent à une quinzaine d'années, et des approches similaires avaient été étudiées dès les années 80 pour des documents « hors ligne ». Les premières applications concrètes datent d'une dizaine d'années, et portaient sur deux domaines spécialisés

essentiellement : la création de bases de faits médicaux et pharmaceutiques, et les bases de données juridiques.

Les différents domaines de l'extraction d'information

Depuis ses premiers balbutiements, l'extraction d'information s'est différenciée en plusieurs domaines spécialisés, conduisant à la création d'approches spécifiques.

La reconnaissance des entités nommées (NER)

Le domaine qui a fait l'objet du plus grand nombre de recherches et d'application est sans conteste la reconnaissance des entités nommées (*Named Entity Recognition* en anglais, NER). L'objectif est d'identifier dans les textes les mentions de termes qui « nomment » des choses réelles, concrètes (pas des concepts), c'est-à-dire des noms de lieux, de personnes, de société, des marques, etc.

Comme les noms sont ambigus, les techniques de NER ont pour objectif non seulement de reconnaître une entité nommée, mais de l'identifier correctement (par exemple, dans la phrase « Michael Douglas est intervenu en moins d'une heure », identifier ce Michael Douglas comme un plombier du Bronx, et non comme l'acteur américain).

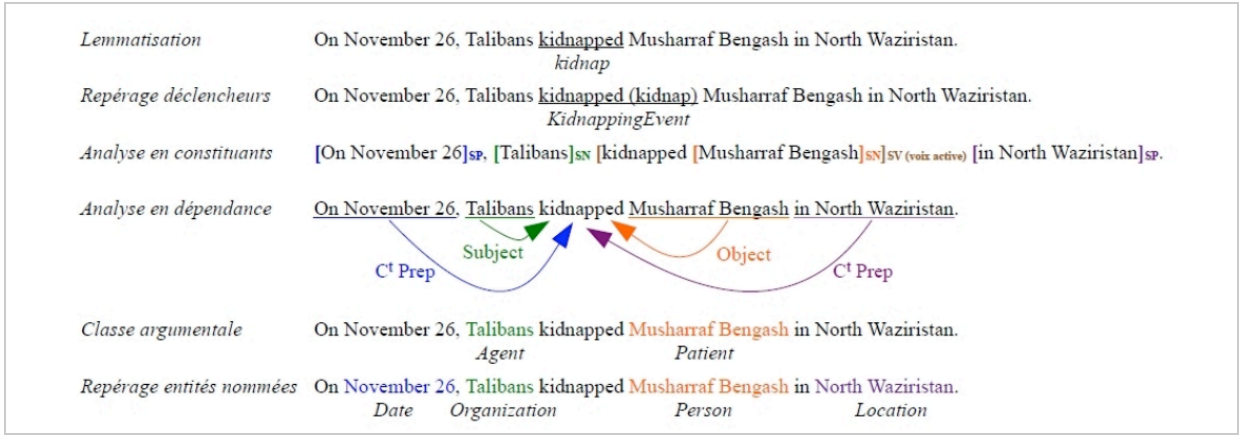
La résolution des co-références (COR)

La résolution des co-références cherche à établir les relations entre une entité nommée et toutes les désignations utilisées sur le web, comme par exemple : François Hollande, le président Hollande, le Président de la République, le Président de la République Française...

La COR est indispensable pour améliorer les applications de la NER.

L'extraction des évènements (EE)

Les techniques d'extraction d'évènements visent à reconnaître la mention d'évènements dans les textes, et à créer une base de données structurée à partir de ces évènements, comportant un certain nombre d'informations associées à l'évènement comme les différents noms de l'évènement, les dates de l'évènement, les protagonistes etc.



Les étapes successives de l'extraction d'un évènement dans un texte (implémentation typique dans l'outil GATE de l'université de Sheffield)

L'extraction des relations (RE)

L'extraction automatique des relations est un domaine de recherche beaucoup plus complexe. Il s'est développé sensiblement depuis sept à huit ans seulement. Il s'agit ici de créer automatiquement des « ontologies », c'est-à-dire des bases de données stockant des items et les relations reliant ces items sous forme de triplet { objet 1 ; relation entre objet 1 et objet 2 ; objet 2}. Par exemple {« roue » ; « est un composant de » ; « automobile »}.

Jusqu'à une époque récente, les ontologies étaient le plus souvent construites à la main par des documentalistes. Mais les progrès récents dans les techniques d'apprentissage automatique ont permis des approches plus automatisées, le plus souvent néanmoins dans des domaines très spécifiques (bases médicales, pharmaceutiques/biologiques / chimiques, bases juridiques). Le « knowledge graph » de Google est une application directe de « Relation Extraction » automatisée.

L'extraction de relation automatique : vers le moteur de recherche du futur

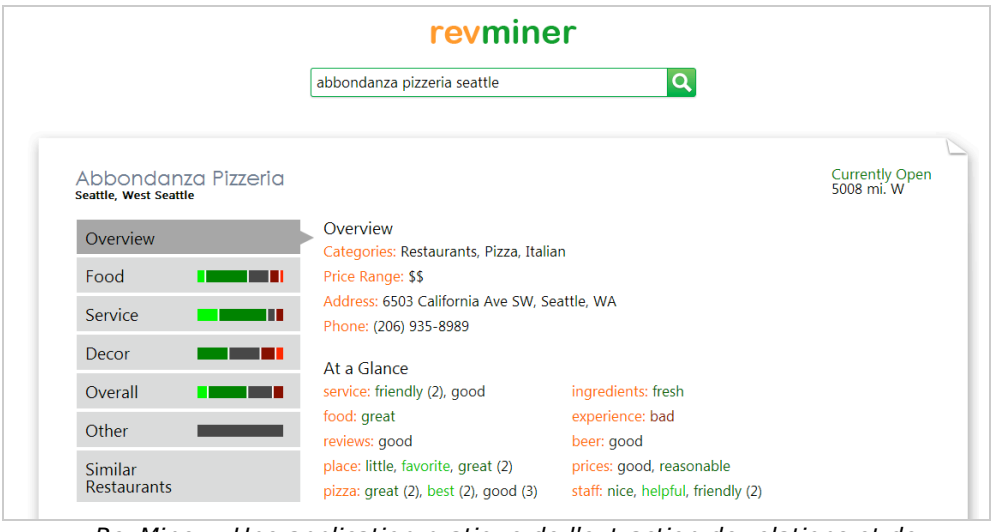
Le fait de disposer, grâce aux techniques d'extraction automatique de relations, de « bases de faits » très étendue, incluant également un grand nombre de relations (sémantiques bien sûr, mais pas uniquement) permet de rêver à des moteurs de recherche beaucoup plus pratiques que ceux d'aujourd'hui.

Imaginez un « Siri » ou un « Google Now » beaucoup plus sophistiqué, à qui vous pouvez poser n'importe quelle question en langage naturel, et espérer recevoir régulièrement une réponse précise, exacte, et pertinente. Une telle technologie trouverait aussi une application immédiate en étant embarquée dans les Google Glass. Cette évolution sonnerait la mort progressive des systèmes présentant « dix liens bleus », ou c'est à l'utilisateur de lire des pages web pour espérer y trouver la réponse à sa question.

Une application des bases de faits que l'on peut construire par extraction d'information, a connu récemment une forte couverture médiatique, lorsqu'en 2011 IBM a décidé d'inscrire son programme « Watson » comme candidat au jeu télévisé américain « Jeopardy ». Watson a réussi à battre deux champions humains, prouvant le niveau de maturité des systèmes de questions réponses modernes.

Hélas, jusqu'à une époque récente, les techniques d'extraction de relations, et d'extraction d'information d'une manière générale, se sont révélés limitées dans leur extension par l'obligation d'avoir recours à une expertise humaine pour obtenir une bonne qualité des bases construites à l'aide de ces techniques. Dans la pratique, ces techniques ont donc été limitées à des domaines d'expertise précis, et toute généralisation semblait impossible.

Dans un premier temps, l'apport des algorithmes d'apprentissage automatique n'a pas résolu le problème de manière significative. Certes, on a remplacé la reconnaissance de « *patterns* » (patrons) construite à la main par des spécialistes, par des algorithmes apprenant à isoler automatiquement des patrons, mais ces algorithmes restaient au mieux semi-supervisés, et ne fonctionnaient que dans un domaine précis, pour chercher un type d'information défini à l'avance.



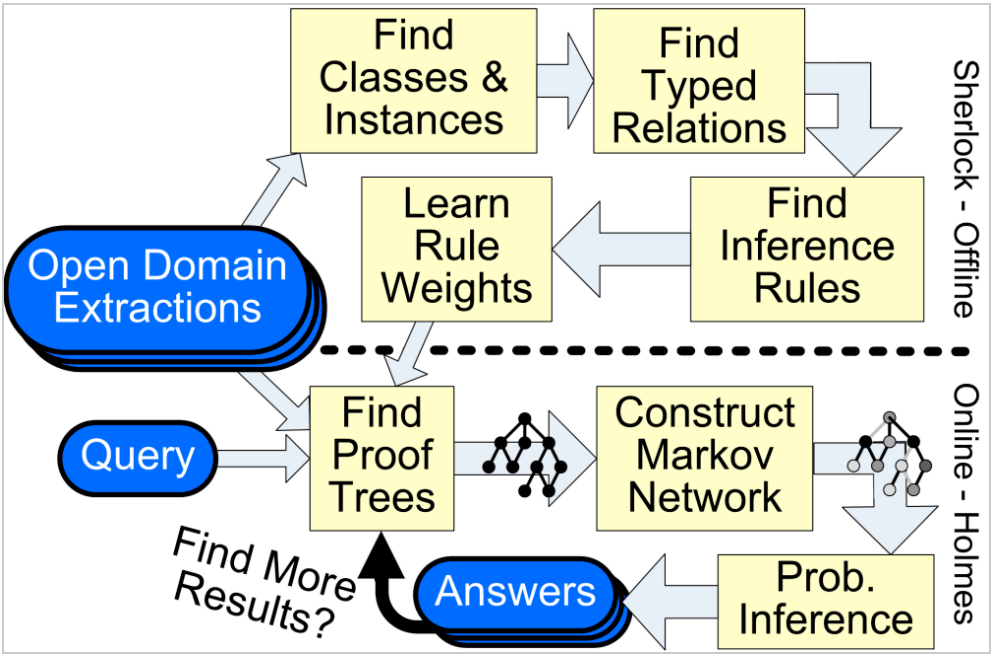
RevMiner : Une application pratique de l'extraction de relations et de l'extraction d'information de manière générale. L'analyse des avis sur les

restaurants permet d'extraire un grand nombre d'informations utiles, restituables dans un moteur de recherche. Cet outil a été mis au point par le laboratoire « Turing » de l'université de Washington, dirigée par Oren Enzioni.

L'extraction d'information ouverte : une technique non supervisée présent dans la technologie « Wavii »

Depuis quelques années, plusieurs équipes de recherche se sont lancées dans la mise au point d'approches capables de construire des bases de faits sans supervision (ou presque) et fonctionnant de manière universelle, c'est-à-dire sans définir à l'avance le type d'information ou de relations à collecter, et sans se focaliser sur un domaine particulier.

L'une des équipes les plus avancées dans ce domaine est celle d'Oren Enzioni, qui a « inventé » en 2007 une approche de ce type baptisée « extraction d'information ouverte » (*Open Information Extraction*). Le « open » fait référence ici au caractère non ciblé et sans *a priori* de sa méthode. En effet, l'OIE ne cherche pas à trouver un certain type d'informations, dans un domaine particulier, mais cherche à collecter tous types d'information, et sans se limiter à un domaine ou une thématique prédéfinie.



Un schéma simplifié d'une implémentation d'OIE aidée par des logiciels d'apprentissage automatique baptisés « Sherlock » et « Holmes » mis au point au sein de l'Université de Washington.

Depuis son article fondateur en 2007, Oren Enzioni a déjà mis au point une deuxième génération d'outils d'OIE, beaucoup plus performants, et plus « génériques ».

Il s'avère que c'est cette approche que l'on retrouve dans un brevet attribué à Wavii et à Oren Enzioni, ce dernier étant officiellement intervenu comme « conseiller technologique » pour la Startup. L'OIE est donc sorti des laboratoires pour trouver son application dans un outil pour smartphone proposé au grand public.

Qui est Oren Enzioni ?

Oren Enzioni est un professeur d'informatique à l'Université de Washington. C'est un spécialiste de l'extraction d'information et du web mining. Il est le fondateur et le directeur du Centre Turing de l'Université de Washington. C'est également un « serial entrepreneur », il a participé à de nombreux

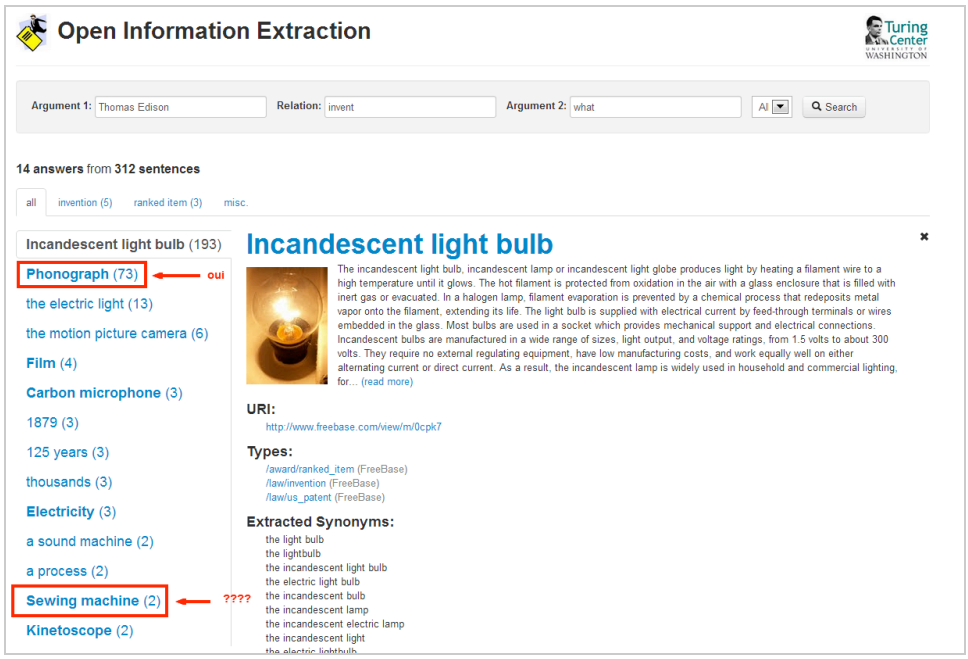
projets de startups depuis une dizaine d'années. Il est notamment l'un des cofondateurs du moteur de comparaison de produits Decide.com, mais a également participé aux projets Netbot, Metacrawler, ClearForest (éditeur de Calais, produit racheté par Reuters, plus connu dans sa version gratuite Open Calais). Il a fondé FareCast, un moteur d'analyse de prix pour les billets d'avions, racheté depuis par Microsoft (la technologie est intégrée dans Bing). Il figure parmi les associés de Madrona Venture Group, une société de « venture capital ». Il a conseillé plusieurs entreprises, dont Google, Microsoft, et plus récemment Wavii.

Les défis à relever par l'OIE dans le futur

Si Wavii démontre que l'*Open Information Extraction*, et les autres approches d'extraction d'information non supervisées, représentent une évolution riche de promesses pour les moteurs de recherche vers des systèmes de « questions réponses » universels et performants, nous n'en sommes qu'au début et il reste pas mal d'obstacles à résoudre.

Tout d'abord, la plupart des outils avancés dans ce domaine ont été conçus pour l'anglais (c'est vrai pour l'OIE). Il existe aussi des tentatives similaires pour le français et d'autres langues européennes. Mais personne n'a encore développé d'approches génériques pour de nombreuses langues, les outils sont en général conçus pour fonctionner dans une seule langue. Le premier défi à relever est donc de créer des outils universels !

L'exactitude de l'information extraite constitue un autre challenge : la source d'information utilisée (le Web) contient des informations erronées, qui sont susceptibles d'être extraites au même titre que les « bonnes informations ». Voici par exemple ci-dessous un exemple tiré de l'outil « ReVerb », un outil créé par l'équipe d'Oren Enzioni qui permet de faire des recherches dans la base de relations extraite par OIE. On voit que l'invention du Phonographe est attribuée à Thomas Edison, mais aussi celle de la machine à coudre, ce qui est une erreur commune chez les anglophones (en France on l'attribuerait à Thimonnier).



Reverb, un outil permettant de faire des recherches dans une base de faits et de relations extraites par OIE. Cet outil est accessible sur le site du laboratoire Turing de l'université de Washington.

Parfois, ce sont les systèmes d'extraction d'information qui se trompent, en raison de leurs limitations. Si un tel système tombe sur phrase contenant « Kentucky Fried Chicken », il y a de fortes chances qu'il en déduise que l'état du Kentucky a frit des poulets... La réduction de la proportion de ces faux positifs, de ces extractions

inappropriées, est un enjeu majeur. Hélas les progrès à effectuer dans ce domaine sont importants.

Les performances des outils d'extraction de données posent aussi un problème majeur pour une utilisation grand public et universelle : ces outils sont lents, très lents. Cela constitue un obstacle sérieux pour disposer de bases mise à jour en temps réel. En l'état actuel de ces technologies, une base constituée sur le world wide web entier a toutes les chances d'être obsolète avant d'avoir pu être déployée !

Quels changements en attendre pour les moteurs de recherche du futur ?

Néanmoins, on peut parier sur une intégration progressive et rapide de ces technologies dans les moteurs de recherche. Elles peuvent servir à « doper » de manière significative les applications de type Knowledge Graph, en élargissant de manière considérable les informations collectées, et les types de relations supportées entre les faits.

Elles peuvent aussi améliorer sensiblement les applications de recherche à base d'interface vocale comme Siri / Google Now, ou à interface gestuelle + réalité augmentée comme les Google Glasses, certaines applications de la Kinect etc.

Rappelons aussi ici que souvent l'extraction d'informations entre dans la réalisation d'applications de recherche répondant à certains types de requêtes transactionnelles : comparateurs de prix ou de services, comparateurs de vols, ou analyse prédictive de prix.

Cette évolution sonne t-elle le glas des fameux « liens bleus » et de la « boîte pour taper des mots clés » dans les moteurs de recherche, comme le clame le Professeur Enzioni dans un article publié dans Nature (« Search Needs a Shake Up ») ? C'est sans doute prématuré d'annoncer cela pour demain. Les approches de type « recherche documentaire informatisée » et « extraction d'information automatisée » sont de toute façon complémentaires, et ne sont pas 100% substituables. Cela annonce plutôt l'avènement de nouveaux systèmes de « questions réponses » enfin efficaces et utiles.

Siri et Google Now éduquent (involontairement) les utilisateurs à renoncer au paradigme de la recherche par mots clés pour revenir à une formulation des questions en langage naturel. Mais gageons qu'il faudra encore des années pour que le nouveau paradigme annoncé prenne une part significative dans les usages...

Bibliographie

Pages web

La bio complète d'Oren Enzioni :
http://homes.cs.washington.edu/~etzioni/bio.html

Lien vers l'outil ReVerb :
http://reverb.cs.washington.edu/

L'article d'Oren Enzioni dans Nature :
http://turing.cs.washington.edu/papers/Nature_search_shake-up.pdf

La fiche produit de Watson (IBM) :
http://www-03.ibm.com/innovation/us/watson/putting_watson_to_work.shtml

Et la video de son passage dans Jeopardy :
IBM's Watson Supercomputer Destroys Humans in Jeopardy
http://www.youtube.com/watch?v=WFR3lOm_xhE

Publications scientifiques

Open information extraction from the web.
Banko, M., Cafarella, M., Soderland, S., Broadhead, M., Etzioni, O.

In: Proceedings of the 20th International Joint Conference on Artifical intelligence, Hyderabad, pp. 2670–2676. Morgan Kaufmann, San Francisco (2007)
http://turing.cs.washington.edu/papers/ijcai07.pdf

The tradeoffs between open and traditional relation extraction.
Banko, M., Etzioni, O
In: Proceedings of ACL-08: HLT, Columbus, pp. 28–36. Association for Computational Linguistics, Columbus (2008)
http://turing.cs.washington.edu/papers/acl08.pdf

Regroupement sémantique de relations pour l'extraction d'information non supervisée
Wei Wang, Romaric Besançon, Olivier Ferret, Brigitte Grau
http://www.taln2013.org/actes/www/TALN-2013/actes/taln-2013-long-026.pdf

Open Information Extraction: the Second Generation
Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam
http://turing.cs.washington.edu/papers/etzioni-ijcai2011.pdf

Philippe YONNET , ***Directeur de l'agence Search-Foresight / Groupe MyMedia. Président de l'association SEO Camp (http://www.seo-camp.org/)***