Le « Knowledge Vault » de Google : évolution ou saut quantique ?

Retour au sommaire de la lettre

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

Le Knowledge Graph, basée sur l'outil Freebase, a été lancé il y a deux ans par Google et permet au moteur de recherche, par détection d'entitées nommées, de fournir directement des réponses à certaines requêtes informationnelles. Mais un nouveau concept, imaginé pour combler les lacunes du Graph, est en train de se mettre en place dans les laboratoires de recherche de la firme de Mountain View. Ce Knowledge Vault n'en est encore qu'à ses balbutiements mais il n'en est pas moins prometteur...

Le 16 mai 2012, Google annonçait sur son blog le lancement d'une nouvelle fonctionnalité qui marquait un tournant dans l'histoire du moteur de recherche : le Knowledge Graph (http://www.abondance.com/actualites/20120521-11478-knowledge-graph-google-officialise-son-moteur-semantique.html). Ce « graphe des connaissances » était présenté comme une immense base de données de faits qui permettrait à Google d'afficher directement toute une série d'informations utiles sur sa page de résultats.

Deux ans plus tard, il semble que Google ait atteint les limites permises par les techniques utilisées par le Knowledge Graph. La base de faits couvrant une partie limitée des informations cherchées par les internautes, Google cherche donc à l'étendre en utilisant des techniques plus sophistiquées. Cette nouvelle approche, Google l'a baptisée « Knowledge Vault » (http://www.abondance.com/actualites/20140902-14231-knowledge-vault-futur-recherche-google.html), et elle risque de changer radicalement le fonctionnement du moteur sur de nombreuses requêtes et dans de nouveaux cas d'utilisation.

Mais commençons par rappeler ce qu'est le Knowledge Graph, et pourquoi il a vite atteint ses limites.

Le « Knowledge Graph » et la réponse instantanée sur les requêtes informationnelles

Traditionnellement, on sépare les requêtes des internautes en trois grands types :
- Les requêtes navigationnelles (trouver l'URL d'une page bien déterminée. Exemple un utilisateur qui cherche la page d'accueil du Monde.fr, ou un post particulier sur un blog).
- Les requêtes transactionnelles (télécharger un logiciel, acheter un produit…).
- Les requêtes informationnelles (rechercher une information sur le Web : ces requêtes sont les plus fréquentes).

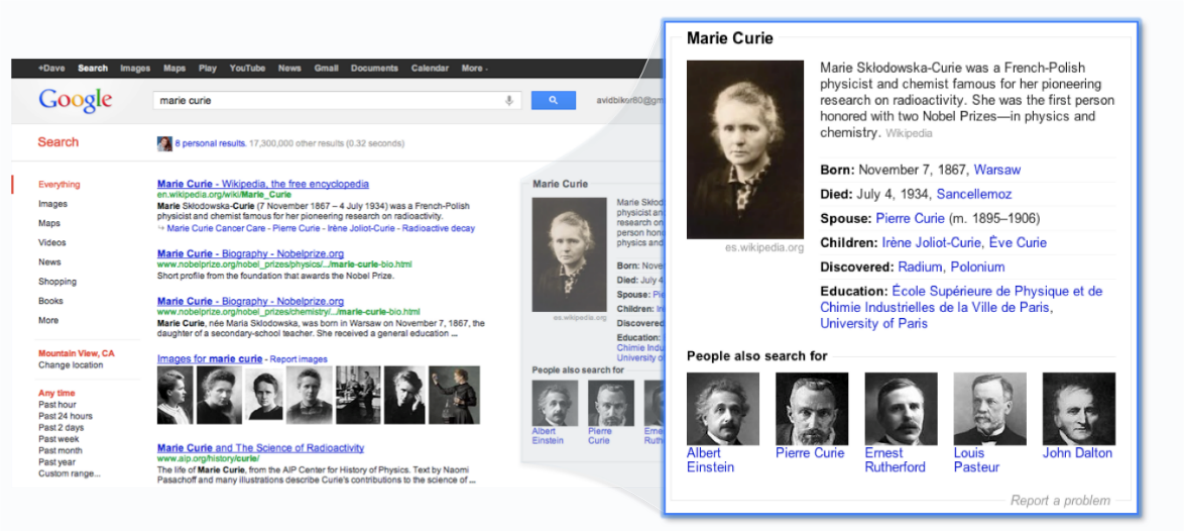
Pendant longtemps, les moteurs de recherche ont été conçus avant tout pour un usage navigationnel : les moteurs proposaient une liste de résultats sous formes de liens bleus cliquables. Ceci est effectivement parfait pour une requête navigationnelle, mais si l'utilisateur cherche une information, il lui faut cliquer sur les liens et chercher lui-même dans les pages ainsi ouvertes si l'information recherchée y figure bien (ce qui n'arrive pas toujours).

Entre 2004 et 2012, des tentatives plus ou moins réussies de création de « moteurs de réponse » ont vu le jour. Un « moteur de réponses », par opposition au modèle de moteur de recherche classique, donne directement la réponse à une question posée.

Mais pour que cela fonctionne, il faut que deux conditions soient remplies :
- Il faut comprendre la requête (la « question » de l'internaute) avant de pouvoir répondre.
- Il faut disposer de la réponse, ce qui suppose de disposer de bases de données particulièrement riches.

La première condition a nécessité des progrès dans la reconnaissance des entités nommées (les noms de lieux, de personnes, les raisons sociales de société). L'une des difficultés de l'exercice consiste à reconnaître l'entité figurant dans la requête quand sa désignation est ambigüe : quand une personne fait une requête contenant les termes « Taj Mahal », cherche-tt-elle des infos sur un monument ou sur le musicien ? S'agit-il de Michael Jackson le chanteur ou son homonyme plombier à New York ?

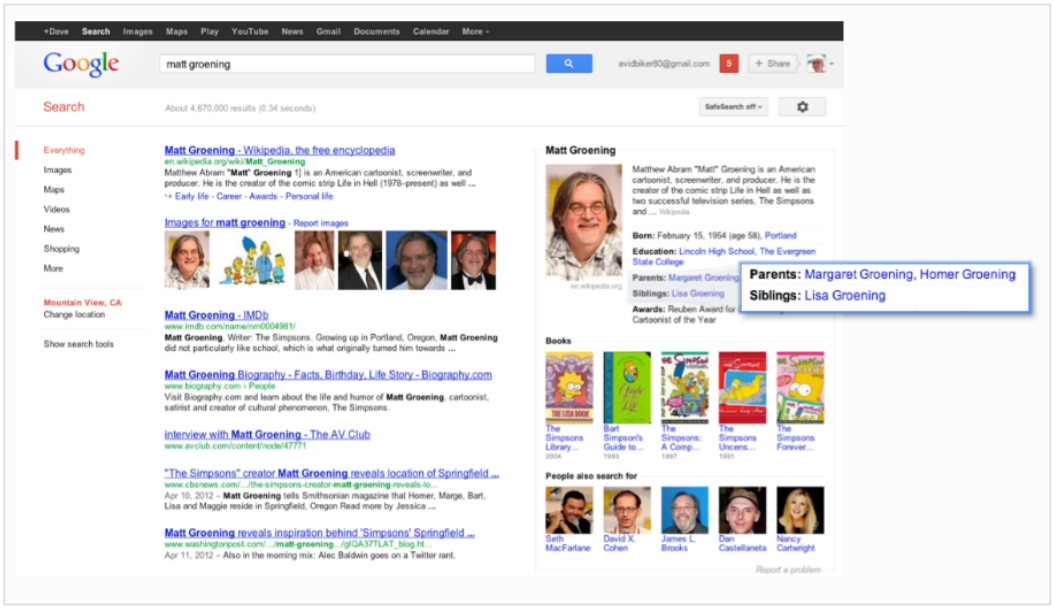
La seconde condition, Google a réussi à la remplir en rachetant Freebase, un projet de base de données « Open » rassemblant une base d'informations sur les entités nommées d'une taille impressionnante.



Exemple de données issues du projet Knowledge Graph : une requête sur "Marie Curie" fait apparaître une fiche complète sur la scientifique dans la partie droite de la page de Google, avec bio, dates de naissance et de décès, et de nombreux autres faits et informations.

Google a continué à développer Freebase à marche forcée, lui permettant de lancer le « Knowledge Graph » officiellement en mai 2012. Le Knowledge Graph est donc avant tout une extension de Freebase.

A fin 2013, la base du Knowledge Graph contenait :
- 500 millions d'entités (les nœuds du graphe) ;
- 3,5 milliards de faits (les arêtes du graphe) ;
- 1 500 types d'entités ;
- 35 000 types de relations.



Le Knowledge Graph stocke des entités, des faits à propos de ces entités, et des relations entre ces entités. Dans les informations sur Matt Groening, on trouve des faits (couple attribut <-> valeur : comme sa date de naissance) mais apparaissent également les relations entre Matt et ses parents (Matt Groening a pour parent Margaret Groening) .

Les limites de Freebase et donc du Knowledge Graph

Le cœur du Knowledge Graph est donc la base ouverte Freebase, qui est alimentée par un certain nombre d'acteurs du web. Malgré la diversité des sociétés et organismes qui l'alimentent (on y trouve aussi bien la CIA avec son célèbre « book of facts » que Netflix ou Wikipedia), des pans entiers de connaissance en sont absents.

Un aperçu des contributeurs les plus connus de la base Freebase.

Pour couronner le tout, les « fiches » y sont souvent incomplètes, ce qui limite beaucoup leur intérêt. En ce qui concerne les personnes célèbres par exemple, leur nationalité est non renseignée dans un cas sur quatre, leur lieu de naissance dans 29% des cas, et leur profession dans 32% des cas !

Relation	% unknown in Freebase
Profession	68%
Place of birth	71%
Nationality	75%
Education	91%
Spouse	92%
Parents	94%

Un aperçu du taux de complétion des faits dans les fiches Freebase pour les entités de type « célébrités ».

Le taux de couverture finalement assez faible de Freebase limite considérablement les applications du Knowledge Graph. Les chercheurs de Google ont donc commencé à chercher des alternatives au crowdsourcing pour compléter leurs bases.

L'extraction des faits et des relations : un problème complexe

Depuis une dizaine d'années, des techniques ont été inventées pour extraire des « faits » à partir de l'analyse textuelle des pages web. Il faut avoir recours à une analyse syntaxique fine pour « désambiguïser » le contexte et identifier clairement les entités, les attributs associés aux entités et leur valeur. On peut aussi utiliser la hiérarchie DOM HTML (notamment la présence de tableaux ou de listes) et la mise en forme pour améliorer la détection. L'approche fonctionne de manière relativement satisfaisante.

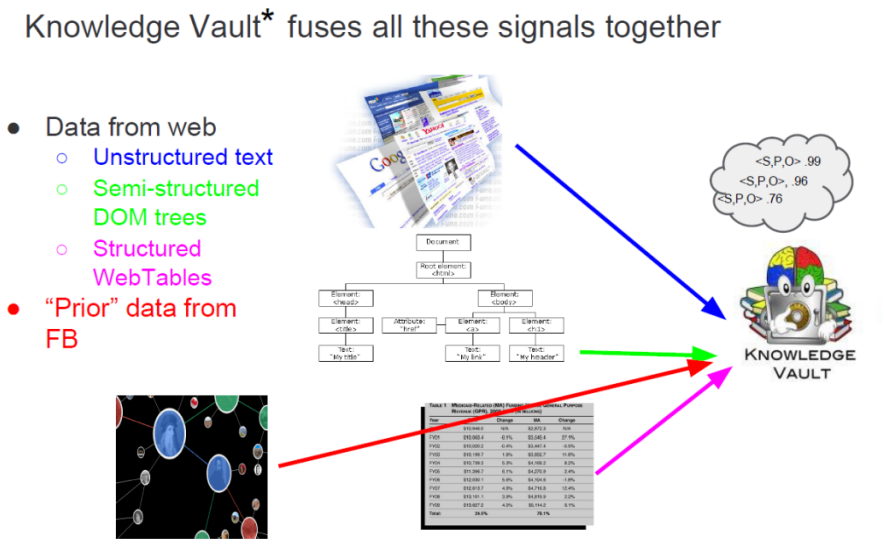
Les techniques d'extraction d'information ont de plus fait des progrès dans plusieurs domaines. Au début, il fallait impérativement savoir ce que l'on cherchait pour le trouver : on ne cherchait que certains types d'entités, et que les valeurs de certains attributs liés à ces types d'entités. En réalité, la typologie des entités nommées n'est pas limitée aux types habituellement cherchés, les attributs sont très divers, et les relations extractibles également beaucoup plus nombreuses dans la pratique que celles que l'on pourrait imaginer *a priori*.

Les progrès ont concerné deux aspects qui permettent d'envisager une plus grande efficacité des méthodes d'extraction :
- **La détection automatique de candidats pour de nouveaux types d'entités**, d'attributs et de relations (notamment avec les méthodes d'OIE : *Open Information Extraction*). Les résultats sont loin d'être parfaits, mais cela augmente la couverture (le taux de succès) de l'extraction.
- **La détection automatique de motifs**, qui permet d'apprendre à extraire des informations sur plus de pages web sans qu'un humain ait besoin de paramétrer le système d'extraction pour chaque site web utilisé comme source d'infos.

Mais le problème est que les informations ainsi extraites ne sont pas toujours fiables, car toutes les pages web ne sont pas des sources dignes de confiance et exemptes d'erreur. Donc s'il est possible d'exploiter les données extraites du Web sur le papier, et d'accéder ainsi à un

volume de « faits » beaucoup plus conséquent, dans la pratique c'est le meilleur moyen pour stocker dans les bases de connaissance des données qui vont se révéler ridiculement fausses.

La nouvelle approche de Google : le Knowledge Vault



Pour résoudre ce problème, des chercheurs de Google, avec à leur tête Kevin Murphy, ont imaginé une nouvelle approche pour compléter la base de faits existante en utilisant des données extraites du web. Cette approche fait encore largement appel à Freebase, mais pour calculer une évaluation de la confiance que l'on peut accorder aux données extraites du web. C'est cette approche qu'ils ont appelé le Knowledge Vault, par opposition au Knowledge Graph, dans un article scientifique publié fin 2013.

Cette publication a été remarquée par la revue New Scientist, qui a publié un papier fort remarqué sur le sujet. Mais l'article du New Scientist laisse penser que les travaux sont terminés, les solutions trouvées et que le « Knowledge Vault » va bientôt supplanter le « Knowledge Graph ». La réalité est légèrement différente...

Kevin Murphy a communiqué quelques indications sur la volumétrie des données récupérées grâce à cette méthode fin 2013. Selon lui, il s'agit de la plus grande base de faits extraits automatiquement jamais collectée. Et cette base est dix fois plus grande que les expériences déjà menées à bien.

Mais en réalité, le volume des données extraites du Web étaient encore en deçà des résultats obtenus pour le Knowledge Graph. On peut certes imaginer que le travail s'est poursuivi et qu'aujourd'hui, ces seuils ont été dépassés, mais l'équipe de Kevin Murphy ne prétend pas avoir trouvé LA solution pour développer de manière sérieuse la base de faits utilisée par Google.

C'est une étape, et non un aboutissement

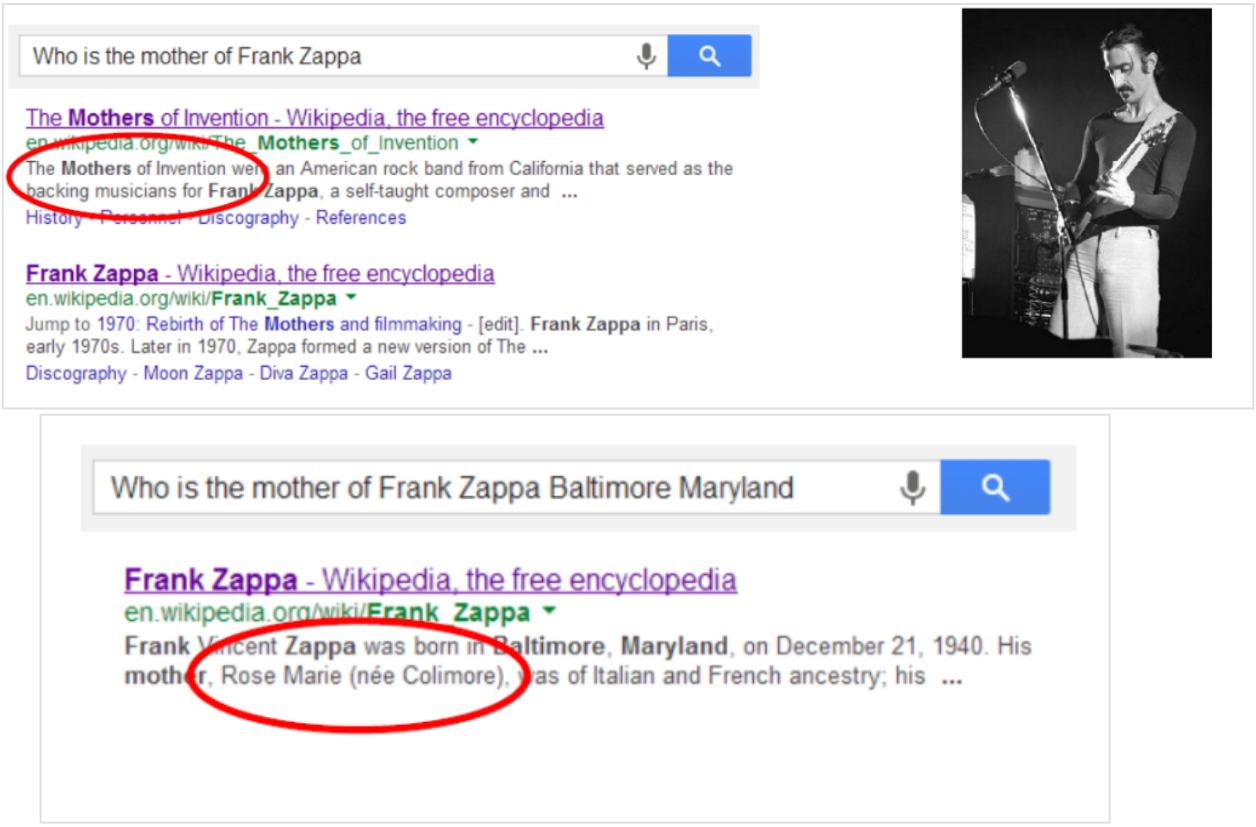
Au contraire, dans une conférence délivrée dans le cadre du séminaire CKIM fin 2013, Kevin Murphy a listé toutes les pistes qu'ils comptaient encore exploiter pour améliorer la couverture de ce système. Et sur ces points, ils en étaient encore à tester l'approche et la qualité comme la quantité de résultats obtenus.

Interroger le web

L'une des pistes pour extraire des faits non identifiés par des méthodes classiques, consiste à utiliser… le moteur de recherche pour les identifier. Une requête appropriée permet de faire

remonter une page qui contient l'information. Cette technique reproduit ce que les internautes faisaient avec un moteur navigationnel : chercher les URL qui sont susceptibles de contenir l'information.

La (très grande) difficulté ici est d'identifier la requête qui fera remonter l'information. Par exemple, la requête « Who is the mother of Franck Zappa » ne fait pas remonter de pages contenant la bonne information. Mais la requête « Who is the mother of Frank Zappa Baltimore Maryland » fait remonter l'information exacte, parce que Baltimore Maryland est son lieu de naissance.



Interroger les gens

L'autre piste, c'est d'étendre la logique de crowdsourcing en interrogeant les internautes pour découvrir les faits qui ne peuvent pas être extraits du Web (soit parce techniquement la donnée n'a pas pu être extraite, soit parce qu'elle n'était pas présente sur une page web). L'idée principale dans ce cas est de solliciter des feedbacks d'utilisateurs ou de les inciter à « compléter » Freebase. Une autre piste serait, *via* un lien sponsorisé, de diriger des internautes vers un « quiz ». Mais il faut aussi tenir compte de la « fiabilité » de ces fournisseurs d'information.

Cette méthode suppose donc que l'on trouve une méthode efficace pour évaluer les contributeurs, ce qui n'est pas tout à fait trivial.

Les principaux problèmes non encore totalement résolus

Pour couronner le tout, Kevin Murphy liste les défis qui restent à relever avant de disposer d'un système d'extraction universel et fiable :

- **Les entités manquantes** : la volumétrie des entités non détectées est encore considérable.

- **Les relations manquantes** : les méthodes actuelles sont surtout incapables d'extraire toutes les relations utiles.

- **Les informations implicites** : on trouve souvent des textes qui contiennent des faits qu'un humain saura extraire, mais pas une machine, parce que le « fait » n'est pas explicite, mais implicite : il découle du contexte. Par exemple, sur un site dédié à Victor Hugo, la page « bibliographie » contiendra potentiellement beaucoup de faits intéressants, mais si le terme « Victor Hugo » ne figure pas explicitement en regard de chaque livre (ce qui est probablement le cas) un extracteur automatique passera à côté de la relation « livres écrits par Victor Hugo ».

- **L'évaluation de la qualité des sources** : les méthodes d'extraction automatiques de faits retournent trop de données erronées. N'utiliser que des sources fiables serait évidemment un progrès mais les méthodes infaillibles pour identifier ces sources restent à inventer.

- **La prise en compte des contextes fictionnels** : l'un des principaux pièges que recèle l'extraction de données à partir du web est que dans certains cas, le fait n'est pas réel car la page parle d'une « fiction ». Or le web est rempli de sites qui reprennent ces informations fictionnelles. Par exemple, un système d'extraction automatique de faits pourra croire que la profession d'Abraham Lincoln, c'est d'être chasseur de vampires, parce que de nombreux sites parlent du film « Abraham Lincoln Vampire Hunter ».

Qu'est-ce que le Knowledge Vault peut changer ?

Néanmoins, à terme, les recherches en cours vont forcément aboutir à un changement d'échelle dans la volumétrie et la diversité des faits et relations collectées. Cela signifie d'une part qu'une proportion très importante de requêtes informationnelles pourraient recevoir une réponse directe (et diminuer encore le trafic envoyé depuis les moteurs de recherches aux sites web). Et d'autre part que l'on pourrait assister à une amélioration sensible du confort d'utilisation de la recherche vocale (de type Siri, Voice Search, Cortana, et qui suppose que l'on soit en mesure de renvoyer une info directe à l'internaute, et non une liste de liens bleus). Les bases de faits sont également très utiles pour améliorer sensiblement des applications de type « Google Now ».

De quelle échéance parle-t-on ici ? En fait, l'élargissement des bases de connaissance sera progressif, il ne faut pas s'attendre à des sauts quantiques, mais plus probablement à un processus graduel d'amélioration.

Du coup, qu'est-ce que cela implique en matière de SEO ? Comme ces applications de type « moteurs de réponse » court-circuiteront la visite de votre site s'il contient aujourd'hui des informations susceptibles d'être extraites, ce sont des typologies entières de sites qui reposent aujourd'hui sur le SEO qui risquent de recevoir beaucoup moins de trafic, au risque de remettre en cause des modèles économiques entiers (annuaires, bases encyclopédiques, bases de données en ligne, etc.). Faciliter l'extraction des informations sur votre site ne fera évidemment qu'accentuer le phénomène à votre détriment. Le gêner ne solutionnera pas forcément le problème si l'information que vous cherchez à garder pour vous et vos visiteurs est facilement extractible sur le site d'un de vos concurrents…

Mais il faudra au moins quelques années avant que les usages de recherches aient évolué de façon suffisamment notable pour que le fonctionnement en « moteur de réponses » ait sérieusement supplanté le « moteur navigationnel ». En attendant, l'impact négatif sur le trafic SEO de certains sites commence à être observable… Et on peut prédire que le phénomène ne fera qu'augmenter.

Bibliographie

L'article du New Scientist
Google's factchecking bots build vast knowledge bank
http://www.newscientist.com/article/mg22329832.700-googles-factchecking-bots-build-vast-knowledge-bank.html?full=true#.U_tDckjPbSC

La présentation de Kevin Murphy au CKIM 2013
From big data to big knowledge
http://cikm2013.org/slides/kevin.pdf

Publications scientifiques

Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion
Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, Wei Zhan, Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043
http://www.cs.cmu.edu/~nlao/publication/2014.kdd.pdf

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 Général de l'agence Search-Foresight, groupe My Media (http://www.search-foresight.com).***