Patrick Constant (Pertimm) : "La sémantique peut rapidement être intégrée dans les moteurs de recherche web"

Retour au sommaire de la lettre

Patrick Constant est PDG et co-fondateur de Pertimm, une société française spécialisée dans la recherche et le traitement d'information. Sa technologie, qui a pour ambition d'"humaniser l'ordinateur", équipe déjà des sites comme Pagesjaunes.fr ou Meetic entre autres. De plus, la société est partie prenante dans le projet Infom@gic qui a notamment pour ambition de proposer l'année prochaine un moteur de recherche sur le Web francophone basé sur des techniques faisant appel à la sémantique. Explications...

Pertimm (http://www.pertimm.com/fr/) est une société basée à Asnières, spécialisée dans la recherche et le traitement d'information. Nous avons rencontré Patrick Constant, son Président Directeur Général, qui a bien voulu nous expliquer l'activité de son entreprise ainsi que son intégration et son rôle dans le projet Infom@gic qui vise notamment à mettre en place des solutions de recherches avancées dans plusieurs domaines.

M. Constant, merci de nous recevoir et de bien vouloir répondre à nos questions. Pour commencer, pouvez-vous nous présenter Pertimm en quelques mots ?

Bien sûr. Pertimm est une société de 20 personnes, créée en 1997 par trois ingénieurs, à savoir Jean Poncet, Xavier Mignon et moi-même. Notre volonté, il y a plus de dix ans de cela, était de créer des outils de recherche tenant compte de la spécificité et de la diversité des besoins des entreprises. La société connaît une forte croissance depuis quelques années puisque notre CA était de 500 K€ en 2005, de 1 M€ en 2006, de 1,4 M€ en 2007 et tout semble indiquer qu'il sera encore meilleur en 2008. Nous sommes basés à Asnières et Valence en France ainsi qu'aux Etats-Unis (Las Vegas)... Nous avons aujourd'hui des clients très prestigieux comme le CNAM TS, le CNRS, les Pages Jaunes, EADS, l'IRSN, la Nasa ou Meetic, entre autres.

Pour bien comprendre votre activité, peut-être pouvez-vous nous présenter ce que vous faites pour certains clients ? Le site PagesJaunes.fr par exemple ?

Tout à fait. Nous sommes arrivés chez Pages Jaunes en 2006, alors qu'ils testaient la solution Fast pour leur site web. Ils ne semblaient pas satisfaits de Fast et désiraient tester d'autres possibilités. Nous avons proposé une dualité "produit packagé + développement spécifique". Nous avons alors réalisé un moteur de recherche pour le site Pages Jaunes en six mois, car l'outil devait absolument être prêt pour la date de lancement qui était fixée au 19 septembre 2007. Impossible de reculer l'échéance... Notre approche a été de fournir un moteur qui soit rapide et évolutif en fonction des futures décisions des Pages Jaunes. En ce sens, la solution que nous avons mise en ligne était 60 fois plus rapide que ce que proposait Fast... Notre technologie tourne donc sur le site des Pages Jaunes (http://www.pagesjaunes.fr/) depuis six mois et nous n'avons pas eu de retour négatif dessus depuis son lancement. C'est très stable...

Toute la chaîne de recherche est motorisée par Pertimm ?

Non, en fait, l'interprétation de la demande de l'internaute reste du ressort des Pages Jaunes. Lorsqu'un usager du Web va sur leur site, il tape une requête, celle-ci est interprétée, "digérée" par un algorithme qui la met en forme afin d'interroger les bases de données grâce à la technologie Pertimm. Mais nous travaillons constamment à cette interprétation de la requête avec les équipes de Pages Jaunes afin de l'améliorer, notamment dans des domaines que nous connaissons bien comme la linguistique, la phonétique ou l'analyse fine des champs textuels. Il y a certainement encore des choses à faire en termes de sémantique et d'ontologie. On s'y attache...

Quel volume cela représente-t-il ?

Le site web des Pages Jaunes traite environ 11 millions de requêtes par jour. Il est le 6ème moteur de recherche en France et se rapproche des 26 millions de requêtes traitées chaque jour en France par Google. Mais il faut voir qu'une requête Pages Jaunes génère environ 10 requêtes Pertimm car nous effectuons des investigations sur de nombreux domaines afin de fournir plusieurs informations à Pages Jaunes qui décidera ensuite lesquelles afficher. Ainsi, nous gérons 100 millions de requêtes quotidiennes sur nos systèmes. J'ai calculé que Google en gérait, pour sa part, environ 300 millions par jour dans le monde. Nous ne sommes pas si loin en termes de performances. D'autant plus que nous répondons en 200 millisecondes au plus pour les requêtes les plus complexes, ce qui est un résultat plutôt intéressant...

Un autre exemple ?

Oui, par exemple sur Meetic (http://www.meetic.fr/), le site de rencontres bien connu. Nous gérons leur base de données de 10 millions d'utilisateurs, chacun d'eux étant géré sous la forme d'un profil de 80 champs... La mise en place de notre produit a augmenté la vitesse de traitement d'un facteur 500 et a permis de gérer aux mieux quelques difficultés ciblées. Par exemple, le champ "*online - offline*" était très complexe car on assiste à 10 000 connexions/déconnexions par minute sur le site. Il fallait donc gérer ce champ au plus vite, pour rester cohérent, ce que nous avons fait.

Mais je pourrais également citer comme exemple de réalisation un système de veille sur la base de flux RSS pour EADS, une gestion des droits des utilisateurs pour Neuf Cegetel (renvoyer des pages de résultats proposant des documents correspondant au profil de la personne ayant lancé la requête), un système adapté à divers niveaux de sécurité pour l'IRSN, qui travaille dans le domaine du nucléaire. Nous travaillons également pour le Ministère de la Défense, etc. Les besoins sont la plupart du temps très différents d'un client à l'autre et nos produits doivent être assez souples pour s'adapter au mieux à toutes les problématiques en termes de formats de documents, de bases de données, de technologies propriétaires, de sécurité, etc.

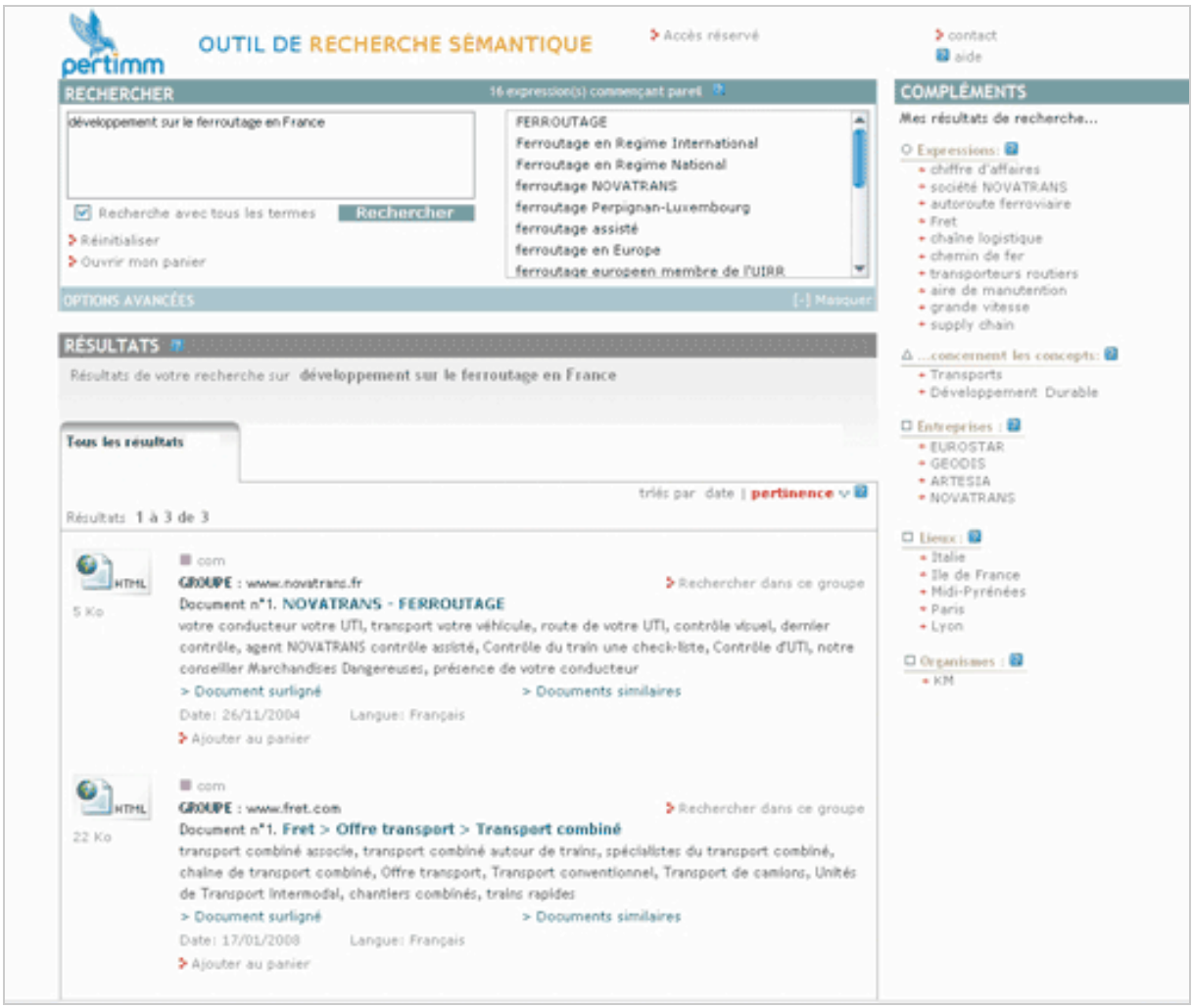
L'une des approches sur lesquelles nous travaillons également, avec le produit PDBA (*Pertimm DataBase Accelerator*) est d'harmoniser un système de bases de données existant, comme nous l'avons fait pour Meetic. Les bases de données de nos clients sont souvent énormes, parfois "patchées", quelquefois complexes. Le but est alors de donner la possibilité au client d'orchestrer de la meilleure façon possible la gestion de son système d'information.

Quelle est la base de votre technologie ?

Au coeur de nos produits se trouve une technologie d'indexation propriétaire. Les trois fondateurs de Pertimm sont des spécialistes de ces domaines. Jean Poncet, par exemple, avait créé une base de données objet dans les années 70... Pour la petite histoire, il a également largement participé aux concerts multimedia de Jean-Michel Jarre dans les années 80. Pour ma part, je suis plus spécialisé en linguistique et sémantique, etc. Nous avons chacun apporté notre brique au projet. Le but est de créer un moteur de recherche qui serait capable de s'approcher, dans son fonctionnement, d'un être humain. "Humaniser l'ordinateur", c'est notre grande ambition...

Je prends souvent la métaphore du libraire pour tenter d'expliquer notre approche : plus vous expliquez à un libraire humain ce que vous voulez, mieux il va vous comprendre et plus il trouvera le livre qui vous convient parfaitement. En clair, il analyse les informations au fur et à mesure que vous les lui fournissez pour obtenir une pertinence plus grande... En revanche, sur un moteur de recherche, plus vous tapez de mots clés, et plus vous rajoutez des opérateurs booléens AND entre eux et donc plus vous risquez de rater des documents intéressants qui répondraient à votre attente...

L'idée de Pertimm est de se baser sur des concepts que nous appelons "Pertimmiseurs" que nous utilisons dans le cadre d'un *mapping* sémantique. Nous pensons que l'approche statistique, basée sur des études "mathématiques" sur la présence de mots dans des pages, est très intéressante, mais reste limitée. Nous avons, pour notre part, la volonté de mettre en relation le sens que l'internaute veut donner à sa recherche ET le sens qui émane d'un document. Comprendre les requêtes et les documents, puis les faire se rencontrer... Encore une fois, "humaniser l'ordinateur"... Vaste pari...



Une requête en langage naturel avec présentation des résultats et suggestions sémantiques associées. Moteur de recherche Pertimm Enterprise Edition sur une recherche Internet.

Quels sont les principaux problèmes rencontrés ?

Le principal est certainement la puissance CPU nécessaire pour mener de tels calculs. Google, sur des requêtes à un seul mot, renvoie, grâce à des analyses statistiques, des résultats sans faire aucun calcul. C'est génial. Bravo ! Mais s'il devait extraire le sens d'une dizaine ou d'une quinzaine de mots pour renvoyer des résultats pertinents, la donne, en termes de puissance nécessaire, ne serait plus la même. Or, c'est ce qui serait pourtant intéressant : demander à l'internaute 10 ou 15 mots qui lui viennent à l'esprit et qui caractérisent sa demande, puis en extraire le sens et effectuer des recherches sur cette base... On pourrait ainsi combler le chaînon manquant entre le cerveau de l'utilisateur et le formulaire de recherche du moteur. "*Dites-nous ce qui vous passe par la tête, nous en faisons notre affaire pour trouver des résultats pertinents*". Voila le moteur de recherche de prochaine génération ! Mais à quel prix en puissance CPU ? C'est aujourd'hui possible sur un intranet, cela reste un vrai challenge sur le Web...

Mais je reste persuadé que, si Google a donné de "mauvaises habitudes" à l'internaute en lui demandant de saisir un faible nombre de mots clés, la situation peut encore évoluer. Un test intéressant se faisait dans les années 80/90, celui dit du "magicien d'Oz" (*NDLR : un exemple d'une telle expérimentation sur le site de l'IRIT :* http://www.irit.fr/rjcihm04/Papers/15_RJCIHM_Merlin_Papier.pdf *sinon tester la requête "ergonomie magicien d'oz" sur un moteur de recherche au choix :)*) : on mettait un internaute devant un ordinateur en lui demandant de taper des questions. Et, en fait, ce n'était pas l'ordinateur qui répondait, mais un autre être humain qui dialoguait, à distance, sans que le "cobaye" ne le sache... On s'apercevait très vite que le réflexe de l'usager était très vite de "*chatter*" avec l'ordinateur pour aller plus loin dans la requête... Pourquoi un moteur de recherche ne *chatterait*-il pas avec ses usagers pour mieux comprendre leurs attentes et donc mieux les servir ?

Le but est bien de mettre en relation, non pas un mot et un document, mais une personne qui cherche et une personne qui a conçu un document qui apporte une réponse... Le tout dans des temps de réponse acceptables. En quelques mots : pouvoir proposer une sémantique optimisée dans des délais optimisés. Tout un programme. Passionnant...

Google a donc donné certaines habitudes de recherche aux utilisateurs...

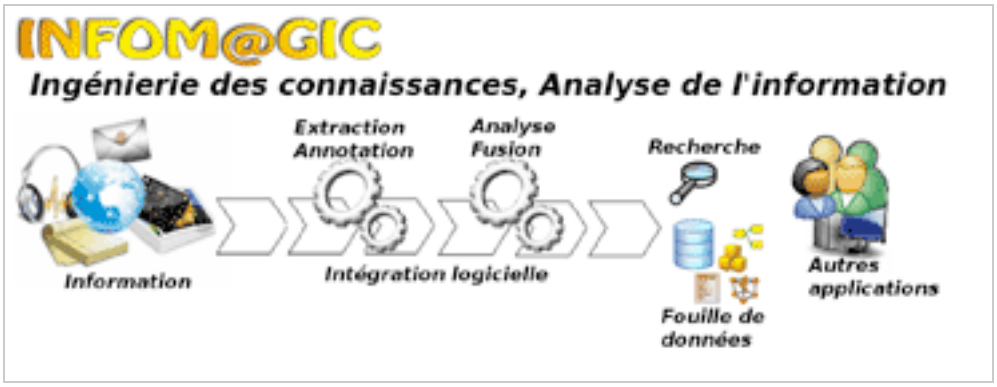
Oui, et plus globalement tous les moteurs de recherche depuis que le Web existe... On tape un à quatre mots clés, rarement plus, et c'est tout. Le langage naturel n'est plus utilisé. Il va être complexe et peut-être long de faire changer les habitudes, mais cela reste possible... On peut également travailler sur la base de la première requête effectuée pour aider l'usager à continuer. Par exemple, sur le moteur que nous avons mis en place pour l'IRSN, l'utilisateur tape une première requête et l'outil propose des concepts associés qui vont l'aider à mettre en place la suite de sa recherche...

Dans la façon de "comprendre" les requêtes, un module de traduction du langage SMS va-t-il devenir indispensable ?

Oui, cela fait partie de nos projets à venir et plusieurs chercheurs, en France, travaillent déjà sur le sujet. Sur Meetic, par exemple, dans les *chats*, le langage est souvent de type SMS. Il faut interpréter ce langage, ne serait-ce que pour des raisons de surveillance de ce qui se dit afin de traquer toute tentative de pédophilie, prostitution, etc.

Parlons du projet Infom@gic. Qu'en est-il ?

Infom@gic (http://www.industrie.gouv.fr/portail/ministre/comm.php?comm_id=6449) est un programme de 50 millions d'euros qui a été mis en place en quatre phases. Une phase 0 nous a permis d'instaurer un pré-cahier des charges. Une phase 1, en 2006, puis une phase 2, en 2007, nous ont permis d'avancer sur le projet et nous sommes donc, depuis le 15 décembre 2007, dans la dernière phase qui va durer 18 mois.



Pour reprendre le texte de présentation, "*le projet Infom@gic a pour ambition de créer une plate forme de recherche couvrant trois axes fondamentaux de l'ingénierie de la connaissance : les moteurs de recherche avancés, l'extraction de connaissances élaborées et la fusion d'informations multimédias*". Il s'agit d'un projet CapDigital (http://www.capdigital.com/), qui (toujours en reprenant les textes officiels) "*a pour ambition de faire de l'Île-de-France l'un des premiers pôles mondiaux des industries du contenu numérique, en renforçant la place des entreprises françaises, des chercheurs et des créateurs sur les marchés internationaux, de créer de nouveaux emplois dans ce secteur d'avenir, et de concourir au rayonnement de la culture française et européenne*". Il regroupe 21 partenaires dont Thales, EADS, Xerox, l'INA, des laboratoires de recherche comme le LIP6, LIPN et sept ou huit PME dont Pertimm.

Il existe de nombreux projets dans Infom@gic mais pour notre part, nous nous concentrons préférentiellement sur un projet de moteur de recherche qui a notamment pour but d'injecter de la sémantique dans la recherche web. Nous avons déjà fait une première maquette avec 12 millions de pages web extraites de site en ".fr" et les premiers tests que nous avons faits sont excellents. La pertinence renvoyée est bien meilleure que celle d'un Google, par exemple, à base de documents comparable. Cela a permis de valider le concept auprès de nos financeurs et de continuer nos travaux sur cette voie.

Cela pourrait signifier qu'à la fin de la quatrième phase d'Infom@gic, mi-2009, un moteur de recherche francophone intégrant une couche sémantique pourrait voir le jour ?

Oui, tout à fait. De notre côté, la technologie devrait être prête et nous mettrons en place un prototype en ligne, accessible à tous. Si les résultats sont bons, nous le laisserons en ligne. Mais si un autre acteur veut utiliser cette technologie dans un cadre plus vaste, pourquoi pas ?

Infom@gic propose des projets de recherche dans de nombreux domaines, certainement ?

Bien sûr, les transferts de technologie sont nombreux et les projets divers, dans l'extraction de textes en mode avancé, dans le domaine des statistiques, en *voice to text* ainsi que dans la recherche d'images similaires et la recherche de sens dans les images par extraction sémantique. Nous travaillons dans de nombreux domaines d'avenir...

Enfin, je voudrais citer deux projets auxquels nous travaillons également en ce moment. **Eneide** (http://www.capdigital.com/xwiki/bin/view/Projet/ProjectENEIDE) dans CapDigital qui a pour but l'amélioration des Environnements Numériques de Travail (ENT) dans le domaine de l'éducation, en partenariat notamment avec le groupe Hachette Livre. L'autre projet est **Trends** (*Trends Research Enabler for Design Specifications* - http://cordis.europa.eu/ist/kct/trends_synopsis.htm) qui a pour but, au travers de l'analyse du Web et des images disponibles en ligne, de définir des "tendances" en termes de couleurs, de formes, de textures, etc., dont pourraient se servir des architextes, des designers, etc. Des projets passionnants également...

Merci, M. Constant, pour avoir bien voulu répondre à nos questions.

Interview menée par Olivier Andrieu le 26 février 2008 dans les locaux de Pertimm à Asnières
Merci à Silvy Bikok, responsable marketing de Pertimm, pour son aide.

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://abonnes.abondance.com/blogpro/2008/03/patrick-constant-pertimm-la-semantique.html