Interview de Jérôme Pesenti (Vivisimo)

Retour au sommaire de la lettre

Vivisimo (http://www.vivisimo.com/) est devenu en quelques années l'un des acteurs majeurs du domaine du "search" aux Etats-Unis. La principale innovation proposée par cette société est une solution de "clustering" permettant de classifier "à la volée" des résultats de recherche en différentes thématiques. Nous avons interviewé Jérôme Pesenti, co-fondateur de la société, qui nous présente l'historique de l'entreprise, sa gamme de produits, sa technologie et ses avantages...

- Bonjour Jérôme Pesenti et merci de répondre à nos questions. Pouvez-vous vous présenter à nos lecteurs?

Bien sûr. Issu de l'École Normale Supérieure de Paris, je suis arrivé aux Etats-Unis il y a six ans, dans le cadre de la coopération. J'ai mis en place, à cette occasion, un projet de clustering de données pour le "*computer science department*" de l'université de Carnegie Mellon. Devinant tout le potentiel commercial de ce type de développement, nous avons eu l'idée, avec Raul Valdes-Perez, qui travaillait également dans cette université (il était en fait le "*faculty*" qui m'avait invité à cette université et c'est lui qui m'avait proposé le projet sur le clustering), de créer la société Vivisimo pour vendre des solutions de "clustering" basées sur nos travaux. Raul est devenu CEO de l'entreprise et moi "*chef scientist*". L'histoire était lancée :-) Nous avons concentré nos efforts au départ sur les solutions de clustering, puis sur des systèmes de "métasearch" (métamoteur) avant de développer un moteur complet, notamment pour réseaux intranet. Ces trois produits constituent notre gamme d'outils qui, nous l'espérons, sont à même de répondre à toutes les problématiques de recherche des entreprises. N'oublions pas également nos deux sites "vitrines", Vivisimo.com pour l'aspect "professionnel" et Clusty.com pour le grand public.

- Pouvez-vous nous présenter la technologie proposée par Vivisimo en quelques mots ? Qu'est-ce qui en fait son originalité ?

Nous proposons donc, comme je vous le disais, une suite de produits (clustering, métasearch, moteur) complémentaires. Le plus connu est bien sûr notre système de clustering. Ce n'est pas une idée nouvelle puisqu'elle date de 40 ou 45 ans... Mais les techniques de clustering de cette époque ne prenaient pas assez en compte l'aspect linguistique du problème, se contentant de traiter les documents comme des données quelconques. Le problème vient du fait que ces techniques génèrent des clusters qui ne sont pas homogènes d'un point de vue conceptuel et trop difficiles a "digérer" par l'utilisateur. Verity et IBM, qui avaient développé ce type de techniques à l'époque, s'en contentent cependant toujours aujourd'hui.

L'idée de Vivisimo a été de prendre en compte un "mix" entre techniques probabiliste, statistique, et linguistique pour créer des "clusters" (ou ensembles de données) plus homogènes conceptuellement. L'analyse est effectuée "à la volée", au moment de la requête, ce qui demande des puissances de calcul assez considérables. Nos logiciels sont écrits en langage C et très fortement optimisés pour obtenir des temps de réponse performants. Par exemple, on pourra sans problème analyser et "clusteriser" mille documents en moins d'une seconde. Nous avons entièrement développé notre technologie "intra muros", ce qui nous permet de la maîtriser sur le bout des doigts mais également de connaître ses éventuelles limites.

Chaque cluster généré par Vivisimo est décrit par une phrase simple ou un couple de mots (alors que les techniques statistiques utilise 10 mots ou plus). L'information retournée à l'utilisateur est optimalement concise et non redondante, communiquant en un minimum de mots le contenu de l'ensemble des documents. Cela permet à l'utilisateur d'obtenir une vue d'ensemble de centaines

de documents en quelques secondes. Cela permet aussi de ne pas avoir de problème de "precision/recall" comme dans les solutions de classification. Les labels des clusters sont extraits directement des documents, garantissant l'adéquation des catégories avec les documents qu'elles contiennent.

- Avez-vous pu mesurer l'apport qu'apporte le clustering à l'internaute, au travers d'études ?

Ce sont surtout des revues comme eWeek ou Infoworld, ainsi que des universités qui ont effectué ce travail. Voici quelques liens qui vous en diront plus sur ces tests certainement plus objectifs que tout le discours que nous pourrions avoir au sujet de nos produits :

http://www.infoworld.com/article/05/05/23/21TCvelocity_1.html
http://www.nwc.com/showArticle.jhtml?articleID=162100481
http://www.eweek.com/article2/0,1759,1683141,00.asp

Je vous propose - ainsi qu'à vos abonnés - également cette très intéressante étude menée par l'université du Maryland qui compare les interfaces de Vivisimo et de Grokker :

http://abonnes.abondance.com/archives/2005-06/vivisimo-study.pdf

En fait, les avantages du clustering sont de plusieurs ordres :

- Les clusters donnent une vue d'ensemble de la recherche effectuée et des thématiques disponibles sur une recherche.

- Elle permet d'aller aux centres d'intérêt majeurs plus rapidement en écartant des informations inutiles dans le cadre de vos attentes.

- Elle permet également de découvrir des thèmes auxquels on n'avait pas pensé et donc de "s'ouvrir l'esprit" à d'autres voies de recherche.

Globalement, sur Google par exemple, tout est mélangé, si vous tapez "jaguar", vous aurez à la fois des liens correspondant à la marque automobile, mais également au félin ou à la console Atari du même nom...

Dans ce cadre, nous sommes à l'opposé des systèmes de "personnalisation" tels que les développent des outils comme Google, Yahoo! ou AskJeeves. Nous pensons qu'il faut laisser le choix à l'internaute en lui présentant mieux l'information. Pourquoi, si un internaute tape "jaguar" sur un moteur, allons-nous privilégier les voitures sous prétexte que ses recherches antérieures avaient trait à Ferrari ? Peut-être est-ce qu'il s'intéresse également aux animaux. Notre but est de classifier l'information à la volée et de lui présenter les différents choix possibles, pas d'en éliminer ou d'en privilégier certains à sa place...

- Cette technologie est-elle implémentable sur n'importe quel résultat de moteur ?

Oui tout à fait. Si l'aspect "visible" de notre travail s'oriente plutôt autour de nos sites web Vivisimo.com et Clusty.com, une grande majorité de nos équipes travaillent plutôt sur l'implantation de nos solutions sur des intranets. Tout site web, qu'il soit "motorisé" ou non, peut avoir accès à notre technologie. Rappelons d'ailleurs que notre système de clustering ne s'adapate pas uniquement à un moteur de recherche. Il peut tout à fait prendre en compte un corpus d'un millier de documents, par exemple, et le classer en "clusters" sans qu'il n'y ait véritablement de saisie de requête comme sur un moteur...

Notre outil est bien évidemment multilingue (anglais, français, langues européennes, japonais, arabe, etc.). Point important : il sera surtout intéressant sur des "gros" sites gérant au moins de 50 à 100 000 documents, afin que les analyses soient les plus fines possible.

Beaucoup de nos clients fonctionnent également grâce à notre outil de "recherche fédérée" (meta-search) sur des sources d'information parfois très hétérogènes. L'idée de cet outil est de créer des "connecteurs" pour chaque source d'information à "attaquer". Chacun de ces connecteurs va s'adapter à la source en question : se connecter (au besoin de mots de passe si nécessaire), envoyer la requête et récupérer les résultats, puis générer un fichier au format XML normalisé en sortie pour traitement, synthèse globale et affichage. Chaque source peut être une base de données professionnelle, un moteur de recherche web, un moteur intranet, etc. Nous installons à chaque fois notre technologie chez le client. Nous ne fonctionnons pas en "ASP". Les problématiques sont souvent assez spécifiques car, sur un intranet, la notion de pertinence est souvent plus complexe que sur le Web. Par exemple, il n'existe pas réellement de notion de PageRank (analyse des liens des documents entre eux) sur un réseau privé. De plus, il n'est pas question de "laisser tomber" tel ou tel document qui poserait problème. L'exhaustivité est réellement la règle. C'est l'une des contraintes des intranets que l'on retrouve moins sur le Web...

- Comment expliquez-vous que, finalement, peu de moteurs majeurs aient intégré le clustering à leurs pages de résultats jusqu'à maintenant ?

Bonne question. :-) Lorsque nous avons développé nos produits, il y a cinq ans de cela, nous avons rencontré tous les acteurs du "search" aux Etats-Unis : AltaVista, Google, Yahoo!, etc. La réponse était souvent la même : "not invented here"... La plupart des moteurs ont plutôt vocation à implémenter des solutions qu'ils ont développé eux-mêmes. Et il semblerait que sur le plan du clustering, ils ne soient pas encore très avancés. Nous savons, par exemple, que Google travaille sur une solution de ce type, mais que leurs tests ne sontpas encore assez convaincants pour que leur solution soit mise en ligne. AOL utilise en ce moment notre clustering combiné avec les résultats de Google sur le site aol.com, servant plusieurs dizaines de millions de requêtes "clusterisées" par jour! Le clustering est en train de devenir "main stream"!

Il faut cependant voir que l'implémentation d'une solution de clustering se heurte également à des notions d'interface utilisateur. Où afficher les "clusters" si toute la partie droite de l'écran est déjà occupée par les liens sponsorisés ? Si on met les clusters sur la gauche, cela retrécit d'autant l'espace alloué aux résultats web. La question n'est pas simple à résoudre et préoccupe certainement la plupart des moteurs actuels...

- NorthernLight (première version), Vivisimo-Clusty, Exalead, selon vous toutes ces approches sont-elles similaires ?

Toutes ces approches sont différentes. En préambule, il est important de bien noter que notre système de clustering ne s'apparente pas à un "catégorisation" faisant intervenir un être humain pour définir au préalable des catégories et des règles d'insertion de ces documents dans les dites catégories. Tout est automatique et effectué "on the fly", à la volée... C'est certes plus subjectif au niveau de la qualité des thématiques créées, mais c'est beaucoup plus facilement gérable au quotidien tout en restant d'une excellente pertinence.

Il existe donc plusieurs "subtilités" plus ou moins importantes entre les différentes approches évoquées dans votre question. NorthernLight, par exemple, s'appuyait sur un fort travail humain développé par des documentalistes. Les concepts étaient attribués aux documents par des bibliothécaires, "à la main", au moment de l'indexation. Ils avaient, à leur grande époque, plus de 50 personnes allouées à cette tâche. Mais leur concept n'a pas survécu à l'explosion du Web et du nombre de documents disponibles en ligne... On n'était pas dans la classification pure où tout est manuel, mais l'automatisation finalement assez faible de leur outil a certainement amené leur échec.

Exalead, pour sa part, semble s'aider des données de leur index pour créer des pré-catégorisations automatisées au moment de l'indexation. De notre côté, tout est entièrement automatique et mis en place "à la volée". Nous ne faisons pas de "pré-traitement" sur les données à "clusteriser". Il existe également d'autres approches, comme celle d'AskJeeves, qui va plutôt se baser sur ses historiques de recherche (ses "logs") pour proposer sur ses pages de résultats telle ou telle voie pour affiner ses requêtes...

- Je dispose d'un moteur de recherche et désire intégrer la technologie Vivisimo. Quelle est la procédure ? Quels sont les coûts ?

C'est très facile en fait. Nous pouvons intervenir de deux façons distinctes en règle générale : soit nous nous situons entre l'internaute et le serveur du client : nous "captons" la requête, la renvoyons au serveur puis analysons et mettons en forme les résultats renvoyés. Cela se fait sur la base de la création d'un "connecteur" spécifiquement développé pour le site de notre client. L'intégration de notre solution se compte alors en heures. C'est la plupart du temps très rapide, notamment si le format géré par le site est, de façon nâtive, du XML. Nous pouvons également nous intercaler entre le serveur et le fournisseur de résultats. Si le serveur envoie une requête à un moteur distant, nous interfaçons notre solution à ce niveau pour prendre en compte et formaliser les données à transmettre.

Au niveau des coûts, cela oscille entre 10 000 $ pour une licence annuelle et 25 000 $ pour une licence illimitée dans le temps. Mais certains cas sont plus complexes et demandent le développement de nouveaux "connecteurs". On s'adapte à la demande... L'idée est de rajouter une brique à l'édifice sans casser tout le mur...

- Si la technologie de Vivisimo semble très pertinente en anglais, elle semble également moins performante en français. Je me trompe ?

En fait, les sites Vivisimo.com et Clusty.com sont aujourd'hui clairement optimisés pour la langue anglaise, ce qui est assez logique, la majeure partie de notre marché se trouvant aux Etats-Unis ou dans le monde anglophone. Mais nous savons sans problèmes traiter d'autres langues comme le français. Nous avons pour projet de développer le site Clusty, notamment, dans un grand nombre de langues pour démontrer l'aspect multilingue de nos solutions. Nous espérons les mettre en place le plus vite possible.

- Quel est le modèle économique de Vivisimo ?

90% de notre chiffre d'affaires vient de l'intégration de notre solution chez nos clients, notamment sur des intranets. Nous avons bien sûr des sites de démonstration qui sont plutôt des vitrines de notre savoir-faire. Mais notre ambition n'est pas de concurrencer Google ou Yahoo! sur le Web, en tout cas pas aujourd'hui.

- Vivisimo, c'est combien de personnes à l'heure actuelle ?

Vivisimo emploie 25 personnes dont la plupart sont basées géographiquement aux Etats-Unis. Mais nous avons d'ores et déjà des représentants en France, en Grande-Bretagne et en Allemagne. L'Europe est l'un de nos axes stratégiques de développement dans un proche avenir et ce dès 2005. L'idée majeure sera de s'étendre en fonction des marchés que nous acquérons. Nous avons déjà une dizaine de clients en Europe et notre ambition et de mettre en place des solutions commerciales fortes sur place ainsi qu'un support technique compétent. Et, en tant que français, j'espère bien que l'Hexagone sera une place majeure d'implantation pour Vivisimo :-)

Merci, Jérôme Pesenti, pour vos réponses (et merci également à Denis Harscoat pour son aide).