

Entretiens avec Jerome Pesenti, Mike Richwalsky et Pierre Caron

[Retour au sommaire de la lettre](#)

Nous avons interrogé trois experts participant au développement de métamoteurs pour en savoir plus sur les technologies utilisées et les grands enjeux sur ce marché. Jerome Pesenti (de Vivisimo), Mike Richwalsky (de Gnosh) et Pierre Caron (de Magellan MetaSearch) ont accepté de répondre à nos questions...

Jerome Pesenti – Co-fondateur – Vivisimo (<http://vivisimo.com/>)

Quand Vivisimo a-t-il été lancé ? Qui est à l'origine du projet ?

C'est en juin 2000 que Vivisimo Inc a été lancé. C'est le résultat de recherches effectuées auparavant au sein du département d'informatique de l'université Carnegie Mellon. Un site Web de démonstration avait également été créé antérieurement (cluster.cs.cmu.edu).

Initialement, le projet a débuté l'été 1998 sous l'impulsion de Raul Valdes-Perez, lequel a été presque aussitôt rejoint par Jerome Pesenti un chercheur français en post-doc. Ensuite, Christopher Palmer, un étudiant en doctorat a rejoint l'équipe et le trio a fondé Vivisimo. Les trois fondateurs sont toujours en charge du projet.

Le but initial était de transmettre aux utilisateurs une meilleure notion du contenu qui se trouve dans des centaines ou des milliers de résultats de recherches sans avoir à cliquer sur "page suivante" ou à faire descendre la page tout en essayant d'identifier mentalement les différents thèmes.



Quel type de contenu peut être recherché à l'aide de Vivisimo et de Clusty ?

Le même contenu que celui qui peut être recherché grâce à d'autres moteurs tels que MSN, Yahoo, Google etc... Clusty accorde cependant plus d'importance à certaines sources de contenu à haute valeur ajoutée comme Wikipedia que nous indexons suivant nos propres règles. Cela permet d'exploiter les images (de Wikipedia), les liens hypertextes insérés dans les articles etc... Le tout est transmis à l'utilisateur dans les résultats de recherche.

Pouvez vous décrire les technologies principales mises en oeuvre pour votre outil ?

Nous utilisons les technologies suivantes :

- i. le clustering (regroupement simultané en catégories d'items textuels).
- ii. la méta-recherche (envoi simultané d'une requête à de multiples moteurs de recherche puis agrégation et re-hiérarchisation des résultats obtenus auprès de tous ces moteurs).
- iii. un crawler/moteur de recherche, que nous utilisons pour indexer Wikipedia, l'Open Directory, et d'autres sites web majeurs d'informations.

Selon vous, comment les outils de méta-recherche peuvent-ils monétiser la recherche ?

Comme tous les autres moteurs de recherche : en plaçant des liens sponsorisés au sein des résultats de recherche.

En quoi Vivisimo Velocity est-il important pour Vivisimo ?

Vivisimo s'est fixé pour mission de transformer l'approche globale vis-à-vis de la recherche sur le Web et pas seulement sur Clusty.com ou sur son aîné Vivisimo.com. Velocity est une plateforme logicielle permettant de construire des portails de recherche de nouvelle génération. Nous l'utilisons en interne pour Clusty.com et Velocity est également utilisé par plus de 100 autres clients dans les secteurs du web, de l'édition, de l'administration et de l'entreprise privée. Velocity est donc l'ADN et le sang qui font vivre la société.

Comment voyez vous l'avenir des métamoteurs de recherche verticaux ?

De plus en plus, les méta-moteurs verticaux spécialisés sur des sujets spécifiques vont se développer. Pour des sujets spécifiques, l'utilisation d'un moteur de recherche ordinaire n'est pas

l'approche la plus efficace. Un prototype de méta-moteur vertical sera lancé par Vivisimo en janvier prochain.

Mike Richwalsky – Web Administrator - Allegheny College
(<http://webtools.allegheny.edu/gnosh/>)



Quand Gnosh a-t-il été lancé ? Qui est à l'origine du projet ?

Gnosh a été lancé au printemps 2005. Nous continuons d'y ajouter de nouvelles fonctionnalités et d'étendre la liste des sites que nous recherchons. C'est donc un projet en évolution constante.

L'idée de Gnosh est apparue lors d'une réunion du Social Software User Group (SSUG) au Middlebury College. Lors de cette réunion en 2005, nous nous sommes demandés pourquoi il n'existe pas d'outil simple d'utilisation permettant de rassembler et de présenter l'information provenant de tous les services disponibles sur le Web. Nous avons donc décidé d'en mettre un au point. Ken Bolton du Vassar College et moi-même nous sommes rapidement mis au travail sur une version de démonstration. Puis mes étudiants employés ont rejoint l'équipe et se sont mis au travail. Les étudiants ont joué un rôle prédominant dans le développement de Gnosh. Ils ont mis au point le système de gestion des utilisateurs et des modules de cache. Ils ont également développé un logiciel de réseau social.

Pouvez vous décrire les technologies principales qui ont été développées pour cet outil de métarecherche ?

La principale technologie que nous avons mise au point est un système de cache plutôt réussi qui nous permet de présenter rapidement et efficacement l'information aux utilisateurs. Nous avons aussi écrit du code qui permet de correspondre avec différents types d'API et d'obtenir les données dont nous avons besoin et de les formater correctement. L'autre partie de notre travail qui nous satisfait est la mise au point d'une couche technologique "sociale". Les gens ont la possibilité de créer un compte utilisateur et de sauvegarder leurs termes de recherche favoris. Ils peuvent aussi obtenir une liste des termes qu'ils ont utilisés lors de leurs recherches précédentes. Ces deux fonctionnalités sont à la disposition des utilisateurs sous la forme de fils RSS. Les utilisateurs peuvent également savoir quelles autres personnes ont fait des recherches en utilisant des termes similaires et les enregistrer comme "amis".

Quel type de contenu peut être recherché ?

Gnosh peut rechercher un large éventail de sources. En voici une liste : Google, Google News, Google Blog, MSN, Yahoo! (et des choses comme les images Yahoo et la recherche contextuelle), Amazon, BlogPulse, Del.Icio.Us, Digg.com, Feedster, Flickr, IceRocket, Plazoo, Technorati. Lorsque de nouveaux outils et API apparaissent, nous les ajoutons rapidement. Nous sommes très heureux d'avoir réussi à ajouter Technorati car cet outil nous a posé quelques problèmes.

Quelles sont les spécificités techniques de Gnosh ?

Tout est construit en PHP et utilise une base de données MySQL.

Quel est l'intérêt des réseaux sociaux pour la métarecherche ?

C'est une bonne question. J'utilise Gnosh à titre personnel et il est agréable de pouvoir voir en un coup d'œil ce que les gens pensent d'un sujet. Par exemple, vous pouvez faire une recherche à propos d'une "pièce Honda". Vous visualisez les résultats de votre recherche concernant cette "pièce Honda" mais vous pouvez également visualiser des résultats périphériques qui vous donnent d'avantage d'informations. En particulier, vous pouvez savoir ce que les gens ont écrit dans leurs blogs sur les pièces Honda, vous pouvez voir des images de ces pièces et vous pouvez mettre dans vos favoris les infos trouvées. Le réseau social complète la recherche. Nous pensons cependant que la recherche liée à ce réseau social ne remplacera jamais la recherche traditionnelle. La plupart du temps, interroger Google suffit à obtenir rapidement les résultats voulus. Gnosh présente un avantage surtout lorsque vous voulez des informations périphériques sur un sujet.

Pierre Caron – Association EGE – Magellan (<http://www.ege.fr/>)

Quand Magellan MetaSearch a-t-il été lancé et par qui ?

Magellan Metasearch est un projet open-source que j'ai lancé en août 2004. Il y a eu plusieurs contributeurs ponctuels dans l'histoire de l'outil, mais je suis le seul développeur qui assure le suivi du projet.

Objectif principal : en finir avec le business model basé sur le "paiement à la source". De nombreuses plateformes de veille proposées par des éditeurs français et internationaux fonctionnent sur ce modèle : le client paye très cher la plateforme "de base" et doit également payer chaque nouvelle extension et chaque nouvelle source de données interrogée. Ce modèle ne convient pas aux PME qui n'ont pas les budgets suffisants pour s'offrir de telles plateformes. Par ailleurs, ce modèle est celui de la dépendance technologique vis-à-vis de l'éditeur de la plateforme -- il devient très difficile de remettre en question la relation commerciale avec l'éditeur une fois la plateforme installée, si c'est lui seul qui possède les clefs pour modifier l'outil. Au contraire, avec Magellan Metasearch, tout développeur peut lui-même étendre l'outil pour les besoins spécifiques de l'entreprise.



Quelles sont ses principales caractéristiques ?

Principales caractéristiques de l'outil :

- * basé sur des technologies ouvertes (langage Perl, base de données MySQL).
- * principe d'ouverture : chaque développeur peut modifier l'outil et l'étendre à volonté.
- * langage d'interrogation évolué : opérateurs booléens, opérateurs spécifiques aux moteurs interrogés pour tirer parti de toutes les fonctionnalités disponibles (alors que d'autres moteurs comme Copernic Agent se contentent de prendre le plus petit dénominateur commun entre les moteurs).
- * vérification des résultats : test des liens, dédoublonnage, contrôle du contenu.
- * système d'alerte automatique : gestion de listes de diffusion par e-mail, consultation par RSS.
- * clustering automatique des résultats basé sur les meta-informations collectées (nom de domaine, pays, auteur, date...), permettant un filtrage rapide des résultats et une élimination des résultats non pertinents.
- * système multi-utilisateurs : gestion des autorisations, des configurations, des interfaces graphiques (une plateforme unique peut très bien être déployée pour plusieurs clients qui auront chacun leur propre interface graphique, leurs propres sources, etc.).

Quel est pour vous l'intérêt d'un métamoteur offline ?

L'interface utilisateur est généralement plus réactive.

Par ailleurs, avoir l'outil sur son poste de travail (ou sur un intranet d'entreprise) donne la possibilité de l'administrer au plus près des besoins du veilleur, et de se détacher du besoin de "plaire au plus grand nombre" typique des métamoteurs en ligne sur Internet.

Enfin, un outil en ligne déployé sur Internet pour le grand public doit être "léger" et performant pour une utilisation par de nombreux internautes en simultané : il s'adapte donc mal à des traitements lourds qui, dans un contexte d'utilisation de masse, auraient pour conséquence un déficit de disponibilité du service pour les autres utilisateurs.

Merci pour vos contributions !