CMU ou la recherche Internet en grappe

Traditionnellement, deux modes de recherche dominent sur les outils de recherche de l'Internet: la recherche catégorielle sur les annuaires, et la requête par mots clés sur l'ensemble des outils.

Cette seconde méthode s'accompagne d'outils algorithmiques qui permettent le traitement des pages Web soumises aux outils de recherche et indexées dans leurs bases, et un classement des résultats par ordre de pertinence. Google, par exemple, utiliserait une centaine de facteurs afin de classer les pages des sites qu'il propose en résultat, avec l'originalité d'avoir été l'un des premiers à intégrer le nombre de liens qui pointent vers un site parmi ses critères de classement. Le 'clustering engine' de Vivisimo Créé par trois étudiants diplômés de Carnegie Mellon University, Chris Palmer, Raul Valdes-Perez et Jerome Pesenti, Vivisimo se présente comme un métamoteur qui va rechercher ses résultats dans plusieurs bases d'outils de recherche. Son originalité provient d'un classement parallèle des résultats obtenus en catégories. Le '*clustering engine*', ou moteur de recherche en grappes, de Vivisimo dispose donc de la faculté d'explorer les descriptions proposées par les outils de recherche interrogés, et d'en extraire les 'co-textes' afin de définir des thématiques génériques. Une démarche que nous avons testé, et qui nous a surpris par la pertinence des catégories proposées, sur des requêtes en anglais, mais aussi en français. **Une nouvelle étape dans les méthodes de recherche** La puissance de Vivisimo dans la catégorisation de ses résultats n'a pas échappé aux spécialistes anglo-saxons des outils de recherche. Search Engine Watch l'a classé '*Best Meta Search Engine*' en 2002. Imaginons une recherchée sur une grande ville, au delà des classements traditionnels qui privilégient généralement les hôtels et autres sites à caractère commercial, Vivisimo proposera des rubriques 'Hôtels' bien entendu, mais aussi 'Sport', 'Journaux' ou 'Loisirs'. Certains spécialistes y voient une alternative à Google. Un raccourci peut-être un peu osé, mais la technologie déployée par Vivisimo mérite que l'on y prête une attention intéressée. **Le clustering, le futur des outils de recherche ?** Raul Valdes-Perez adopte l'image d'une bibliothèque pour décrire les résultats proposés par son moteur: sur un outil de recherche classique, les livres proposés en réponse à une requête sont fournis en vrac (avec parfois les best-sellers en tête, et souvent ceux qui ont payé pour figurer au plus haut). Avec Vivisimo, en plus de ce classement, le moteur propose les rubriques thématiques dans lesquelles sont classés les livres. La performance ne vient donc pas des sites proposés, mais de la méthode de classement développée. Dans le cas de Vivisimo, quatre critères sont utilisés : la concision des titres, leur précision, leur distinction, et la proximité avec les requêtes humaines, qui permet de s'écarter des choix mécaniques. Une recherche sur Vivisimo se déroule donc en deux étapes : l'agrégation des titres et descriptions des sites proposés en résultats, puis la création et le classement des sites dans des rubriques par des algorithmes mathématiques qui utilisent une base de connaissances sur les synonymes, les abréviations et les différentes formes de mots. **Une méthodologie en expansion** Créé en juin 2000, Vivisimo a capitalisé sur ses algorithmes, omettant volontairement de les publier dans des revues scientifiques. Ses premiers clients s'appellent la NASA, l'université de Stanford ou la revue de l'American National Association. Ainsi que de nombreuses sociétés biomédicales, pour lesquelles le regroupement des informations en grappes thématiques prend une dimension stratégique. Une technologie que nous devrions rapidement retrouver sur des outils de recherche majeurs de l'internet. A tester sur www.vivisimo.com