Yves Simon (Fast) : "Les portails d'entreprise doivent prendre conscience de l'importance de la recherche"

Retour au sommaire de la lettre

Yves Simon, responsable de l'activité "moteurs de recherche" pour la société Hemisphere Intelligence Informatique est également responsable du déploiement commercial de Fast en France. Une bonne occasion pour faire un point avec lui sur l'avenir du "search" tel qu'il l'envisage selon trois mots clés principaux : usages, pertinence, intelligence. Voici ses réponses à nos questions.

Responsable du déploiement commercial de Fast en France et expert en recherche d'information depuis de nombreuses années, Yves Simon nous confie son sentiment : l'avenir du Search passera par une prise en compte croissante de la dimension d'usage fondée sur une pertinence augmentée, qui transformera nos outils de recherche en véritables moteurs d'accès à l'information intelligents au service de l'utilisateur.... Laissons-lui la parole...

Pouvez-vous vous présenter en quelques mots ?

Bonjour, je suis aujourd'hui responsable de l'activité Moteurs de recherche pour la société Hemisphere Intelligence Informatique depuis plus d'un an et donc du partenariat avec l'éditeur Fast Search&Transfer. J'interviens ponctuellement sur des missions de spécifications fonctionnelles autour des technologies Moteurs de recherche.
Je suis issu d'un parcours dans le monde des services à l'international et d'un programme de MBA (IMHI 96') cofondé par l'ESSEC et Cornell University USA. J'ai cofondé en 2000 la société Amoweba, porteuse du projet de moteur de recherche distribué et parallèle Human-Links qui a ensuite absorbé Novadis-Services et Voyez-Vous (Mapstan) avant de m'impliquer dans Social Computing en 2004. Mon profil est accessible sur Viaduc.

Dans votre white paper publié sur le site de social-computing, vous évoquez la position centrale acquise par les moteurs de recherche. Pouvez-vous nous en dire plus ?

Oui, une position centrale acquise grâce à la très forte volumétrie de contenus rendant l'exploration systématique impossible et ouvrant donc le chemin à un usage dominant des outils de restitution faisant appel à des traitements automatiques et de plus en plus sophistiqués pour interpréter les requêtes des utilisateurs (que j'appelle alors consultants).

Il existe en réalité deux types de moteurs de recherche : les moteurs de recherche internet qui proposent des index issus d'un travail de crawl. Ces moteurs amènent les utilisateurs sur des sites où ils vont naviguer et être de nouveau amenés à rechercher de l'information en fonction de leurs besoins et de la typologie des services proposés. L'autre type de moteur de recherche est donc celui que l'on trouve sur un site web, un portail d'entreprise ou un intranet. Les premiers indexent le web, les seconds une base de contenus finis.

Sur le Web le calcul de pertinence est majoritairement généré à l'aide d'un ranking défini en fonction de l'autorité des résultats qui est concrétisée par des hyperliens pondérés en fonction du contenu. L'accent est mis sur la capacité de traiter une multitude de requêtes en temps quasi réel (inférieur à la seconde). L'approche de traitement du contenu est généralement moins développée car le volume des informations à traiter est très important et donc coûteux en ressources machine...

Que font de plus les moteurs d'entreprise ?

Les moteurs d'entreprise fédèrent généralement des sources de contenus hétérogènes résidant dans l'intranet, sur des bases de données externes et sur le web. Ils doivent pouvoir identifier les formats d'origine du document et normaliser ces informations pour les restituer à différents publics. Les traitements envisageables dans ce contexte permettent une normalisation des contenus plus fine, définissant ainsi un environnement d'indicateurs de pertinence plus précis que celui (souvent unique) dont disposent les moteurs qui indexent le web. Cette normalisation génère

un contexte de pertinence qui permet d'affiner considérablement la restitution en fonction de différents publics.

Certains moteurs web utilisent des traitements linguistiques tout comme c'est le cas pour certains moteurs d'intranets. Quelle est alors la différence de traitement ?

Il est vrai que sur internet comme sur un intranet on peut également procéder à une expansion de la requête utilisateur sur la base de traitements linguistiques ce qui permet alors d'augmenter le rappel et d'apporter une réponse à toute requête formulée par un utilisateur.

Le contexte de l'information est autre dans un Intranet et comme nous l'avons vu précédemment, la multitude des sources et des contextes d'utilisation métier font la différence, obligeant les éditeurs de moteurs d'intranet à porter leurs solutions sur de multiples environnements logiciels et matériels et à répondre en fonction de différents besoins utilisateurs. Il n'y a donc pas une boite de recherche pour tous mais des accès personnalisés à l'information en fonction de contraintes métier. Par ailleurs, sur un intranet, le contenu des documents peut être indexé linguistiquement, ce qui n'est pas le cas sur le web.

D'autre part, lorsque les moteurs du web se concentrent sur la capacité de montée en charge de leurs systèmes, les moteurs Intranet s'intéressent aux besoins métiers et souvent à des corpus informationnels structurés et non structurés très volumineux et hétérogènes ainsi qu'à l'augmentation de la qualité de la pertinence perçue par les utilisateurs.

Vous évoquez le concept de FUSE. Pouvez-vous nous en dire plus ?

Oui. C'est un concept créé par Yahoo. FUSE signifie : Find, Use, Share and Expand. On pourrait dire en français : Trouver, Utiliser, Partager, Augmenter. Cela correspond bien à ce que les utilisateurs attendent aujourd'hui des moteurs de recherche. L'idée est, au-delà de la recherche d'information, de fournir à l'utilisateur un ensemble d'indicateurs qui lui permette d'affiner ses résultats et d'obtenir en quelques clics la réponse qui correspond précisément à sa recherche. Cela consiste donc à proposer à l'internaute un ensemble de navigateurs connexes qui permettent de pénétrer dans le corpus informationnel sur la base de l'usage qu'il souhaite en faire ainsi que des fonctions de productivité et d'interactivité comme commenter, noter, transmettre, enregistrer, tagger etc...

Qu'appelez-vous navigateurs connexes, s'agit-il d'une aide contextuelle ?

Oui, ce sont des navigateurs générés automatiquement en fonction des entités identifiées dans le corpus de réponse à une requête et qui fournissent à l'utilisateur une aide contextuelle correspondant à son besoin d'affinage des résultats. Cette aide contextuelle permet notamment à l'utilisateur de naviguer dans la liste de résultats. Les indicateurs contextuels peuvent prendre la forme d'entités nommées, de dates ou de types de fichiers. Il peut s'agir également d'indications de recommandation du document par un autre utilisateur. Il s'agit d'une pratique qui rejoint la notion de folksonomie, utilisée par exemple par le moteur de news collaboratif Wikio. Mais sur un intranet on peut aller beaucoup plus loin et prendre en compte des notions supplémentaires telles que les commentaires des autres consultants et leurs profils, les notes ou encore des avis de pertinence du résultat par rapport à une thématique.

Que dire du « E » de FUSE qui correspond à Expand ?

Il s'agit d'exploiter la création d'informations par les consultants sur l'information brute. C'est une source dérivée de pertinence qui se nourrit d'interactions tout en les augmentant, et qui met le consultant en synergie avec l'outil. Cela peut passer par la personnalisation des résultats en fonction d'un profil comme le fait par exemple Yahoo. Ce profil peut être constitué volontairement par l'internaute d'une part, mais il peut être également enrichi de ses habitudes de navigation et du type d'information qu'il a coutume de rechercher. Le moteur renvoit alors à l'utilisateur les informations qui sont susceptibles de l'intéresser. Cette notion comprend également la capacité à utiliser des systèmes d'alerte par mail ou SMS, le filtrage d'information à partir de mots-clés ou encore l'abonnement à des flux RSS ou XML. L'idée est d'étendre au maximum les capacités d'interactions entre l'outil et les consultants et idéalement entre les consultants directement !

Est-ce l'internet de demain ou celui d'aujourd'hui ?

Tous les ingrédients technologiques sont déjà là et ils sont désormais validés par l'usage, et l'usage ne se décrète pas, ne se ralentit pas. C'est l'Internet d'aujourd'hui.

Que dire de la pertinence des moteurs de recherche d'entreprise ? Vous évoquez le fait que 12% des requêtes lancées sur des portails d'entreprise ne retournent aucune réponse. Comment expliquer un tel taux de silence ?

Par un manque d'équipement de technologies de qualité sur les sites web, e-business et Corporate. Historiquement, c'est sans doute le résultat d'une faiblesse de compréhension de la notion de recherche d'information par les décisionnaires ainsi que d'une évolution très rapide des usages. Un moteur de recherche se définit basiquement par deux fonctionnalités qui tiennent en deux mots : recherche et restitution. Cela peut très bien se faire à partir d'un script SQL très simple qui fonctionne sur l'indexation full text d'une base de contenus. C'est ce type d'outil qui était mis à disposition des utilisateurs sur les portails d'entreprise, il y a une dizaine d'années. Cependant les bases de données ont considérablement cru en volume comme en complexité et on se trouve maintenant face à des corpus pouvant dépasser un Tera octets. Cela a pour résultat d'aboutir à un dérapage des temps de réponse et à un taux de silence important aux requêtes émises. Les utilisateurs dans les organisations ne s'y trompent pas et le taux d'usage de l'Intranet sera souvent fonction du taux de réponse aux requêtes formulées, à la pertinence des réponses et à la rapidité des temps de réponse. Il va sans dire que de tels outils ne peuvent plus traiter avec efficacité de tels volumes de données et qu'il devient dès lors indispensable de faire évoluer l'outil ou de le remplacer purement et simplement.

On peut alors choisir d'utiliser une solution de type Google (Boite de recherche) ou bien de véritablement traiter ce problème en faisant l'acquisition d'un vrai moteur de recherche d'entreprise au service de la productivité des travailleurs du savoir dans un cadre stratégique voulu et non plus contraint.

Peut-on parler dans certains cas de « point de lassitude » de l'internaute ?

Certainement. C'est dommage, quand on pense que les utilisateurs qui ne dépassent pas la page d'accueil pourraient être convertis en utilisateurs satisfaits et fidélisés. Je crois qu'il est véritablement indispensable que les acteurs des portails d'entreprise prennent conscience de l'importance du rôle que peut jouer la brique de recherche d'information qu'ils mettent à disposition des visiteurs sur leur site. Il en va de même en interne, en ce qui concerne le partage et la circulation de l'information, même si les enjeux sont différents sur un intranet. Il faut également prendre conscience du fait, d'une part, que les personnes qui utilisent le web depuis longtemps sont de plus en plus exigeantes sur la qualité des résultats comme sur les fonctionnalités de partage mises à leur disposition. D'autre part les nouveaux utilisateurs qui ne sont pas toujours à l'aise avec l'outil informatique ont encore moins de propension à naviguer dans un univers où on ne leur propose pas une aide contextualisée. Si on ne veut pas convertir ces nouveaux utilisateurs en déçus de la société de l'information, il faut absolument leur fournir les outils qui correspondent à leurs besoins.

Comment faut-il procéder pour analyser à la fois l'origine du silence et l'origine du bruit ?

Pour commencer, il est indispensable d'analyser les logs du serveur afin de lister les mots-clés qui ont servi de base aux requêtes des utilisateurs. Lorsque l'entreprise ne présente pas à l'internaute, sur une requête donnée, les résultats qu'il serait souhaitable de présenter, il s'agit alors d'une recherche contre-productive et qui peut aboutir à ne pas vendre le produit qu'on aurait souhaité vendre en ligne. Lorsque le problème est identifié, il est possible d'effectuer volontairement un lien entre les requêtes formulées par les utilisateurs, sur la base et les documents qu'on voudrait présenter en réponse à ces requêtes. Ce lien, qui se traduit dans la plupart des cas par un gain de productivité, ne peut être effectué que si on dispose d'un moteur de recherche d'entreprise qu'il est possible à la fois de paramétrer et d'administrer. Il s'agit en fait de décliner l'offre de contenus dans une direction différente, en fonction des attentes des utilisateurs à un moment donné. Cette étude permet d'envisager l'évolution d'une gamme de produits en ligne, en fonction des attentes des utilisateurs.

Sur les portails on constate la plupart du temps que moins de cinq pour cent des contenus proposés génèrent 90% du chiffre d'affaires en ligne, alors que, inversement, 95% des contenus proposés ne font l'objet que de requêtes occasionnelles. L'entreprise aura naturellement tendance à

privilégier les résultats qui génèrent le meilleur retour sur investissement. C'est sur ce type de contenus, néanmoins fluctuants, que la réputation d'un site est bâtie. Mais il ne faut pas perdre de vue, pourtant, qu'il est également nécessaire d'apporter un résultat qualitativement identique, à toutes les requêtes formulées sur un portail.

Peut-on parler de standardisation de la qualité des moteurs de recherche, et si oui, sur la base de quels critères ?

Il y a trois problématiques à identifier et traiter, que ce soit sur le web ou pour un moteur d'entreprise : la volumétrie des gisements d'information, la dynamique des contenus, la densité de l'audience. L'identification de ces différents critères permet un dimensionnement de l'architecture sur laquelle on est alors amené à travailler.

Il est également nécessaire, comme évoqué précédemment, de paramétrer le moteur ou la base afin de « pousser » en avant, les réponses aux requêtes pour lesquelles il existe effectivement des résultats. Il faut également utiliser les processus de normalisation proposés par l'outil et paramétrer cette normalisation de façon suffisamment fine, pour proposer le bon résultat, à la bonne personne, au bon moment.

Vous évoquez la normalisation des bases documentaires. Peut-on parler également de normalisation des usages ?

On peut en effet parler de normalisation, pas nécessairement des usages en tant que tels, mais de l'étude de l'usage. Il s'agit en fait de mettre en valeur certains usages. En effet, lorsque la côte de popularité d'un usage est mise en avant cela permet de capitaliser sur cet usage dans le but d'augmenter la pertinence de la réponse à un besoin en information. Il faut également utiliser les indicateurs d'usage. On peut en effet comptabiliser sur un intranet le nombre de fois où un document a été par exemple imprimé, transmis, annoté ou encore mis à la corbeille ! Alors une certaine normalisation de l'usage... Peut-être, mais je crois qu'il s'agit d'abord et avant tout d'une prise en compte raisonnée de l'usage.

Faisons un peu de prospective. Quel est l'avenir des moteurs de recherche intranet ou intrasites ?

La tendance actuelle passe par une prise en compte croissante de la dimension sociale de l'outil de recherche d'information. L'usage d'un système crée de l'information à très forte valeur ajoutée et qu'il faut l'utiliser comme base pour repositionner la notion de pertinence. L'interactivité et la capacité à interfacer l'outil de recherche d'information avec d'autres applications d'entreprise, notamment des outils de gestion de contenus ou un système collaboratif, est également une tendance très forte ainsi que la spécialisation des outils en fonction des besoins.

Le moteur de recherche est une brique de base qui permet au système d'être plus réactif dans la restitution des informations aux utilisateurs, la validation de la conformité de certains contenus, ou encore la sécurisation des échanges. Le moteur de recherche est au cœur de la dynamique de travail en réseau. Je crois d'autre part, qu'à moyen terme on trouvera de plus en plus de moteurs de recherche embarqués dans des applications tierces.

L'interactivité est également itérative. L'utilisateur ayant trouvé un contenu qui correspond à son besoin en information sera parfois amené à produire d'autres contenus en rapport avec cette information qu'il sera amené ensuite à proposer à l'indexation du moteur de recherche. L'interactivité passe donc aussi par la création de contenus à forte valeur ajoutée. Ce qui est important et devient désormais une tendance forte, c'est de bien monitorer l'usage d'un système de recherche d'information.

Par ailleurs, le marché de la recherche d'information devient de plus en plus important à mesure que les besoins en information des internautes augmentent. Un champ très important de recherche d'information va progressivement s'ouvrir avec l'obligation légale d'archivage de données, notamment dans des domaines comme l'administration ou le secteur bancaire. L'avenir de la recherche d'information passera aussi par une prise en compte croissante des formats numériques multimedia comme l'audio ou la video. Ainsi, la recherche d'information a encore de très beaux jours devant elle.

Quel est votre rôle au sein de Hemisphère ? Etes-vous responsable de la diffusion de Fast ?

Fast est un acteur de la recherche d'information qui travaille sur tous les continents et a prouvé la qualité de sa technologie. Fast est présent en France depuis maintenant plus d'un an et a déjà remporté des références significatives. L'entreprise dispose également d'un solide réseau de partenaire sur le territoire. Dans ce cadre, Hemisphere Intelligence Informatique, SSII française, a passé avec l'éditeur Norvégien, un accord de partenariat pour la revente des technologies Fast Search&Transfer pour la France. Ce contrat est un prolongement de l'engagement d'Hemisphere Intelligence Informatique depuis 1997 sur la ligne de produits NextPage (racheté par Fast courant 2004). Je suis responsable du développement de ce courant d'affaires pour Hemisphere Intelligence Informatique.

Pouvez-vous citer quelques références ?

Fast est un acteur de la recherche d'information qui s'est bien positionné sur le marché Français. La société compte à ce jour de nombreuses références dont Alstom, Carrefour, PriceMinister ou les Pages Jaunes qui ont récemment sélectionné cette technologie. La société compte également plus de 3500 implémentations au niveau International.

Comment se décline cette technologie ?

Elle se décline en fonction de typologies de marchés. La plate-forme éditée par FAST s'appelle ESP v5 pour Enterprise Search Platform. C'est une plate-forme sur laquelle ont été bâties des Search Derivative Applications, c'est-à-dire des applications de recherche d'information métier, correspondant à certains types d'usages ou besoins en information, pour la recherche sur des sites web, des intranets, des applications e-commerce ou d'annuaires ainsi que des applications de surveillance de marchés par exemples.

Il s'agit plutôt dans ce cas d'un usage de veille ?

Oui, en effet. Fast dispose d'une brique logicielle dédiée à la veille stratégique qui s'appelle MarketTrack. Elle s'utilise notamment dans le cadre de cellules de veille marketing, juridiques, financières, R&D …

Quelle est l'originalité de l'offre Fast par rapport aux autres offres du marché ?

C'est difficile à résumer en quelques mots. ESP 5.0 est une plateforme proposant des traitements à la fois sémantiques et statistiques de grande qualité. Le moteur est capable de détecter soixante-dix-sept langues, d'indexer tous les formats de fichiers connus dont des formats audio et video, et propose des traitements tels que l'analyse du sens d'un texte à travers Contextual Insight. Le moteur propose aussi différentes stratégies de compréhension de la requête utilisateur permettant d'augmenter leur expérience d'usage. La plateforme présente d'excellentes performances de montée en charge sur les 3 axes de la volumétrie de documents, de la dynamique de l'index (mises à jour) et du volume de requêtes utilisateurs.
Ce n'est pas une boite noire. Il est par exemple possible de régler tous les indices de pertinence de façon à obtenir les meilleurs résultats possibles, de personnaliser des traitements de normalisation, d'interfacer des thésaurus externes et des dictionnaires d'entités simplement. Le moteur est à la fois configurable et personnalisable en fonction des besoins.

Comment accompagnez-vous vos clients en ce qui concerne le paramétrage du moteur ?

L'offre par défaut est fournie avec un ensemble de traitements standard. Il y a surtout et avant tout un travail de réflexion de fond et d'analyse stratégique, qui est mené avec le client pour déterminer ses besoins et les usages possibles du moteur en fonction de son activité. Nous paramétrons et révisons ensuite l'application en fonction de l'usage qui en est fait.

Fast dispose d'un module d'évaluation de la pertinence. Comment ce module fonctionne-t-il ?

Schématiquement, il s'agit d'une évaluation technique de la pertinence. Les critères de pertinence d'ESP 5 sont les critères classiques utilisés par les campagnes d'évaluation TREC, à savoir la précision et le rappel. La précision est le pourcentage de documents pertinents retournés par le

moteur, par rapport au nombre total de documents retournés. Le rappel est le pourcentage de documents retournés par rapport au nombre de documents pertinents dans une base. ESP 5 utilise par ailleurs des éléments intrinsèques au résultat pour calculer un ranking qui tient compte notamment de balises telles que l'auteur, la date de mise à jour ou encore la thématique du document, le prix d'une entrée catalogue. La plateforme ESP 5 effectue également des mesures d'autorité du document dans son contexte informationnel sur la base des travaux de Kleinberg et Barabasi. Les aspects plus techniques sont nombreux et pourraient sans doute faire l'objet d'un séminaire.

Avez-vous mis en place un module d'assurance qualité pour vos utilisateurs ?

La qualité est un concept dynamique et se mesure dans le temps. Ce n'est pas une mesure statistique. Il faut parler de mesure du degré de satisfaction de l'utilisateur. Cette mesure, pour être au plus près des besoins, s'effectue au cas par cas et selon le type d'utilisateur et le type d'information proposée. Il existe certes des critères de qualité normés dont il faut tenir compte mais il y a aussi la mesure de la qualité perçue par l'utilisateur et qu'il semble indispensable d'identifier comme de mettre en valeur, pour demeurer au plus près des attentes des consultants. C'est un module important à mettre en œuvre sur les projets, même si nous ne sommes que force de proposition dans ces domaines.

Croyez-vous qu'il soit indispensable de mettre en place des actions de certification de la qualité pour les moteurs de recherche, qu'il s'agisse de moteurs web ou de moteurs d'entreprise ?

Oui. Il est fondamental de parvenir à travailler avec les éditeurs sur ce sujet là afin que l'offre soit la plus transparente possible pour le client final. L'objectif est de mieux servir l'utilisateur dans le but de générer un chiffre d'affaires plus important et permettre une mise en valeur des contenus grâce à une meilleure qualité d'accès à l'information. Dans une telle perspective, qui s'intéresse à l'usage du moteur de recherche, il est alors possible de mesurer un retour sur investissement et donc la qualité fait la légitimité de la technologie.

Peut-on parler de ROI (retour sur investissement) de l'évaluation ?

Oui, absolument. Ce ROI peux simplement se constater à partir d'un calcul sur l'évolution du taux d'usage d'un moteur de recherche. Il convient d'évaluer le système afin que sa pertinence demeure au plus près des usages des utilisateurs de ce système. Lorsque les consultants trouvent plus facilement des contenus qu'ils n'auront pas matière à recréer, on peut constater un gain de productivité lié à l'amélioration de la qualité du système.

Quelques mots en guise de conclusion ?

Il existe aujourd'hui tout un ensemble d'actions connexes à la recherche d'information et qu'il est intéressant de proposer aux utilisateurs, dans la mesure où cela correspond à un besoin en information et que cela leur permet d'augmenter la qualité de leur expérience d'utilisation de l'outil en réseau. Ce sont également autant de critères de pertinence nouveaux et complémentaires qui émergent alors de l'usage et qui sont pris en compte par le moteur pour organiser la liste des résultats et faire ressortir l'information la plus directement utile sur la base d'une pertinence augmentée.

Je crois que, au regard des nouvelles fonctionnalités mises aujourd'hui à la disposition de l'internaute, il devient désormais presque plus pertinent de parler de technologies d'accès à l'information que de moteurs de recherche.

Interview menée par Marianne Dabbadie
Directrice Innovation i-KM
Laboratoire GERIICO – EA 1060