Les moteurs de recherche entre humanisation et automatisation

Les moteurs de recherche sont nommés discrètement assistants de recherche, voire logiciels d'infrastructure et leur vie est en grande partie sous-marine puisqu'ils sont encore à 50% vendus en OEM, intégrés à des portails, à d'autres moteurs de recherche, voire à des logiciels de CRM ou d'ERP. Cependant, nul ne conteste plus leur importance ni le fait qu'ils ont de nombreux utilisateurs, professionnels et grand public, même si les uns et les autres ont encore des reproches à leur faire.
Deux grandes écoles se détachent parmi les éditeurs de moteurs de recherche.
Il y a ceux qui comme Verity pensent que la technologie est arrivée à un palier, et que les progrès viendront d'une humanisation de la technique en faisant intervenir des experts humains, dans le cadre du KM, qui sous forme de réseaux sociaux, recommanderont tel ou tel document.
Pour les autres, tel Autonomy ou Influo, la solution ne peut être que tout automatique et des technologies nouvelles, comme les réseaux neuro-mimétiques, permettront de gagner en efficacité dans la sélection de l'information et la pertinence, transformant en dinosaures, les moteurs de recherche plein texte tel K2 de Verity! Pour Peter de Bie d'Inxight Software, d'ailleurs, le moteur de recherche n'est pas la solution, mais "le problème", puisqu'il remonte trop d'information non utilisable parce qu'indifférenciée. Il ne s'agit plus seulement de chercher/trouver, mais de catégoriser, de regrouper et de classifier. Tous de parler de thesaurus et de taxonomie des métadonnées, confirme Laurent Le Foll, Directeur Adjoint Sinequa qui a été longtemps Directeur Général Europe du Sud et créateur de Verity France. Que les moteurs soient plein texte ou non, leurs concepteurs veulent utiliser les champs – titre, titraille, auteurs, date et autres pour améliorer l'extraction des connaissances. C'est XML qui va permettre de structurer et de donner du sens à ces métadonnées. C'est ainsi qu'Arisem propose un moteur sémantique qui crée un document XML à partir de chaque document source. Celui-ci est ensuite indexé en mode "full text" par le moteur Search Server de Microsoft.
En principe, les moteurs de recherche ont une architecture similaire. Au centre un index que l'on peut décrire schématiquement comme un fichier en deux colonnes. La première est une liste de termes, la seconde contient, pour chaque terme, l'adresse des documents auxquels ce terme est lié. En aval, une base de données (base de références) contenant les documents indexés ou leur adresse. En amont, l'interface d'interrogation, une page web le plus souvent. Le moteur est deux temps : indexation et interrogation. Il y a ensuite différentes catégories de moteurs de recherche, plein texte, avec mode d'interrogation booléen (et, ou, sauf), langage naturel, statistique et probabiliste, linguistique et sémantique . Les solutions reposant sur l'analyse sémantique ne comprennent pas les mots, mais les concepts. Il ne sera pas facile cependant de classer les moteurs par catégorie, la plupart adoptant une combinatoire de ces technologies pour arriver au résultat optimal.
Une première différenciation peut en revanche être faite, entre moteurs purement Internet et moteurs qui sont des solutions d'entreprise. Les Altavista, Google, Lycos, Voilà, etc.. nouvelle forme des annuaires électroniques ne sont pas parvenus jusqu'ici à pénétrer le marché des solutions d'entreprise. Il est beaucoup plus facile de traiter des fichiers HTML ou faire des recherches simples sur de gros volumes d'information que de traiter des documents multi-formats pour répondre à des questions posées par des professionnels qui sont beaucoup plus élaborées. "La technologie à mettre en place nécessite d'être beaucoup plus performante, en terme de format de documents, volumétrie, les exigences ne sont pas les mêmes," confirme Laurent Le Foll.

Le marché des solutions d'entreprises est selon le Gartner Group et autres analystes dominé par Verity, à lui seul, 35% du marché,(45% en fait, depuis qu'il a racheté Inktomi), Autonomy venant loin derrière en seconde position avec une vingtaine de pour cents qui lui sont actuellement très disputés, en raison de sa politique de prix, qui en fait le moteur le plus cher du marché.
Ce marché considéré comme mûr par Didier Donnat de Verity est en phase de concentration. Le rachat d'Inktomi n°4 des moteurs de recherche par Verity en est une illustration. Les trois moteurs les plus linguistiques, Arisem, Sinequa, Lexiquest ont eux aussi changé de main cette année, Sinequa étant conforté par son rachat par un fournisseur de contenu, La Revue Fiduciaire, et Lexiquest par SPSS.
Pour Laurent Le Foll, la concentration se poursuivra, et il ne restera qu'un très petit nombre de moteurs de recherche de taille internationale. Mais il y aura place pour un petit nombre de petits éditeurs de produits innovants, dont on peut compter une vingtaine sur le marché français, les plus importants, étant Sinequa suivi de Digimind, Pertimm, Albert, Influo.
Par le fait qu'il est omniprésent, et qu'il ne donne pas entière satisfaction, le moteur de recherche a été considéré comme le parent pauvre, (expression d'Alain Beauvieux d'Albert), des logiciels Internet/Intranet. Il ne cessera de l'être qu'en augmentant son efficacité et apportant une véritable valeur ajoutée aux clients. Mais il serait dès à présent erroné de ne pas reconnaître que bon an mal an, les outils offerts aux veilleurs, voire aux sites de e-commerce (et dans ce cas, ils font de l'aide à la vente), ont progressé, à la fois du point de vue de l'automatique et de celui du support humain.
Mireille Boris

1
Albert et la GMIL, grammaire indépendante des langues

"L'outil de recherche doit être d'une efficacité absolue!…Et ne doit pas rester le parent pauvre de l'Internet / Intranet» prévient Alain Beauvieux – Directeur Général Go Albert France.
Sa jeune entreprise (deux ans, quarante clients dont la Coface, Autovalley, Saab, Hospit-Hall) appartient au groupe suisse industriel Tag. Situé à Montpellier, au cœur des vignes, le Centre de Recherche & Développement d'Albert compte aujourd'hui 30 chercheurs.
Issu d'une longue phase de tests , AMI Albert Meaning Interpreter constitue le socle commun à trois produits - AMI Enterprise Discovery (Intranet), AMI Website Access (sites et portails Internet) et AMI for Domino (interface de recherche plein texte pour utilisateurs Lotus Notes). " Avec AMI for Domino nous ne jouons pas les trublions mais apportons une offre complémentaire à celle d'IBM", résume Alain Beauvieux .
Adoptant une technologie mixte de linguistique et de statistique croisées, AMI exploite le plein texte et les métadonnées.
Il permet de s'exprimer en langage courant, fédère la recherche multi-sources, fait de la veille, de l'assistance de navigation, de l'aide au commerce électronique. AMI permet d'augmenter le taux de connexion/vente car il donne la possibilité à l'internaute de trouver un article par sa référence, son nom ou sa marque sans en connaître l'orthographe exacte.
Outre le français et l'anglais, la version 3.5 d' AMI inclut un support étendu pour l'allemand et le hollandais et s'intéresse aux langues asiatiques. Une GMIL, grammaire minimale indépendante des langues, a été créée. Cette option permet d'indexer un document non seulement sur les mots qu'il contient, mais également sur des groupes de mots qui en donnent le sens. GMIL extrait d'un document les groupes de mots significatifs (groupe nominal, compléments d'objet, compléments circonstanciels...), et améliore la pertinence des recherches. Les termes composés de plusieurs mots (ex : "pomme de terre") sont désormais considérés comme des concepts à part entière et indexés en tant que tels. Ceci permet, notamment, une meilleure gestion des synonymes, des sigles et des acronymes. Les documents retournés peuvent désormais être regroupés par thèmes et "clusterisés". Ces thèmes sont calculés dynamiquement en fonction de la requête et des documents retournés. Ils permettent à l'utilisateur de mieux comprendre les résultats obtenus.
XML aide à structurer les métadonnées et les meilleures performances sont obtenues sur XML natif. "Des deux écoles, automatiser, humaniser, je suis de la première," reconnaît Alain Beauvieux. Le prix d'entrée d'AMI Enterprise Discovery est de 50000 euros. La proportion de solutions d'entreprise, et celle d'accès à des sites web est l'ordre 2/3 et 1/3. Partenaire SQLI, AMI est Open Source. Toutes ses solutions tournent sous Linux, Solaris, et Windows.
www.albert.com/

2
Autonomy et les logiciels d'infrastructure intelligents

Plus qu'un moteur de recherche, Autonomy est capable de catégoriser l'information, de générer des taxonomies, de créer des hyperliens, de profiler les employés, de personnaliser l'information et d'envoyer des alertes. Après IDOL (Intelligent Data Operating Layer), son nouveau module CEN(Collaboration and Expertise Networks) met l'accent sur la collaboration, "parce que le marché comprend l'importance de la collaboration", précise Richard Gaunt, co-fondateur et CTO d' Autonomy.
Sur les 210 membres de sa compagnie , 70 travaillent exclusivement en R&D. Leur objectif est de "donner du sens à un monde non structuré".

Veille : - Le choix du concept Autonomy mérite une explication qui permettra de mieux situer votre entreprise.

RG : - Nous nous sommes lancés en 1986 avec cette idée qu'il existe des morceaux de logiciel autonomes pouvant prendre des décisions pour vous. Nous pensions à l'époque que des agents autonomes, détectant nos besoins pourraient faire nos courses et nous les livrer avant même que nous n'y ayons pensé. Nous développons des briques logicielles capables de comprendre un document, un paragraphe et de signaler que ces derniers sont similaires à ce qu'on recherche, voire à ce que l'on devrait rechercher. La technologie sous-jacente repose sur des modèles mathématiques, inférences de Thomas Bayes et théorie de l'information de Claude Shannon. Suppléant aux faiblesse des méthodes informatiques traditionnelles, ces algorithmes de reconnaissance des formes tendent à faire réaliser plus aisément par l'ordinateur ce que font naturellement les hommes dans leur interactions avec l'information non structurée, qu'il s'agisse de texte ou de langage parlé. Notre logiciel totalement indépendant des langues sait retranscrire la voix en texte. Nous traitons la voix aussi facilement que le texte.

Veille : - Où réside la valeur principale d'Autonomy comparé à un moteur de recherche classique ?
RG : - Les moteurs de recherche du marché nécessitent des efforts manuels pour l'extraction des éléments les plus simples . La conséquence en est le coût. Le personnel perd son temps à chercher. Parce qu'il comprend la signification, l'idée d'un paragraphe, Autonomy résout le problème. Il attrape l'information et la livre à l'utilisateur en contexte. Avec un moteur de recherche, on sait ce qui est arrivé, avec Autonomy on peut faire des découvertes. CEN constitue une nouvelle façon d'organiser la compréhension des êtres, construite sur ce qu'ils font et non ce qu'ils nous disent ou pensent qu'ils savent. Nous apportons de la valeur en mettant l'information en contact avec la personne qui doit la recevoir.

Veille - Le KM n'effraie plus, il est devenu vital. Les éditeurs de logiciels de gestion de contenu, de changement l'utilisent comme une fenêtre pour présenter leurs produits. Le WCM se nomme désormais ECM. La démarcation entre couches logicielles devient complexe et mouvante. Est-ce gênant pour vous?
RG - Nous correspondons à toutes ces définitions. Mais celles-ci sont restrictives et de ce fait tendent à dévaluer les produits. Nous préférons aller vers le marché en décrivant ce que nous sommes.
Or, nous sommes plus qu'une application de portail. Nous répondons à un besoin fondamental : livrer l'information à la personne dans le contexte du problème qui est le sien.

Aucun de nos concurrents n'approche le problème de notre façon. Microsoft a des produits de gestion de contenu, mais rien de ce que nous faisons, comme prendre une information sous-jacente très complexe et la rendre simple. Nous maintenons une compréhension du document source et supportons un très grand nombre de référentiels en même temps. Nous supportons tous les portails standards. Nous avons les APIS pour nous connecter à Lotus, Exchange, Documentum, Vignette, Siebel. Nous complétons les principaux produits CRM, ERP, KM, B2B, B2E.

Veille : - Les noms de vos clients sont prestigieux, Intel, Sprint, EDS, Crédit Lyonnais, US Department of Commerce, HP, Nestlé, AT&T, US Army, JD Edwards, Sun, Nasa, Texas Instruments, TF1, Pechiney…et les secteurs couverts sont très différents.
RG : - IDOL et CEN sont complètement génériques. Nos clients viennent de secteurs très différents et la façon dont ils utilisent le produit est également très diversifiée. Certains s'en servent en gestion de contenu, collaboration, contrôle de la concurrence, règlement de litiges comme dans les procès faits à l'Industrie du Tabac. Toutes les agences américaines de lutte contre le terrorisme sont équipées de notre produit.
Mireille Boris

(encadré)

"Autonomy est à ce jour une compagnie très bénéficiaire dans un espace unique. Nous avons à cause d'elle été amenés à définir un nouveau niveau d'architecture logicielle, l'"intelligent infrastructure layer", la couche intelligente d'infrastructure. C'est ici qu'il va y avoir de la compétition dans l'avenir. A ce jour, Autonomy jouit d'une avance certaine. Sa façon de gérer le contenu de différents référentiels et de les mettre en contexte, est très différente de celle de sociétés de KM telle Invention Machine qui sont des outils sémantiques basés sur des règles et dont le mode de raisonnement est prédéterminé."
Daniel W. Rasmus du Giga Information Group

(encadré)

Une cartographie de l'expertise chez BAE

BAE SYSTEMS, auparavant British Aerospace, emploie130000 personnes dont 35000 ingénieurs sur110 sites pour un CA de 13 milliards de £ en produits Défense.
"La quantité d'information reçue y devenait un fardeau.80% des employés en réseau passaient 30 minutes par jour à rechercher de l'information et 60% une heure ou plus à dupliquer le travail des autres," explique Ian Black, directeur de la communication d'Autonomy.
Une solution de gestion de l'information pour diminuer le gaspillage d'efforts et donner plus de valeur aux ressources d'information tenues sur Intranet a été recherchée. Le choix de BAE s'est porté sur IDOL et plus récemment CEN.
Autonomy agrège le contenu de nombreuses sources en différents formats, structurés et non structurés, y compris l' Intranet et ses10000 nouveaux éléments par jour. Ce contenu est automatiquement catégorisé et des hyperliens sont insérés au contenu, à la volée ce qui facilite l'accès des employés à "leur" information.
Autonomy cartographie les forces et les faiblesses de l'expertise et réduit le temps de recherche de l'information de plus de 90%. Mieux, il est devenu le moteur central de la Virtual University (e-learning) qui permet d'apprendre à partir de l'expérience d'un autre.
7 mois ont suffi pour générer un ROI.

Légendes des illustrations

Richard Gaunt, co-fondateur et CTO

Ian Black, Directeur de la Communication Autonomy

CEN Cluster Mapping
Spectographie des relations entre les intérêts du personnel et les sources d'information pour identifier en temps réel les zones clé d'expertise

Network Profiler
Réseau d'e-mails à l'intérieur d'un réseau collaboratif avec création de profils

CEN Visualization
Visualisation CEN illustrant la relation d'un expert avec d'autres à l'intérieur de la communauté des connaissances

www.autonomy.com/

3
Digimind et sa plate-forme de veille

Digimind est une société française fondée en juillet 98. Son métier, la veille stratégique – 60% édition de logiciel, 40% conseil.
Le premier produit à avoir été développé est v-strat, outil intranet de partage et diffusion d'infos veille. Strategic Finder, ensuite, est un méta-moteur qui interroge des moteurs de recherche professionnels du web invisible par secteur d'activité professionnelle, sur les bases de données spécifiques à un secteur. C'est un petit outil individuel. Pour l'utiliser, il suffit de savoir utiliser un navigateur Internet.
Monitor, lui, prévient dès qu'un site est modifié.
Evolution est une plate-forme de veille globale, de surveillance du web, de bases de données, de forums de discussion. Elle possède les outils de traitement et de partage, de catégorisation, de classement, de commentaire des informations. Multiplate-forme, elle tourne aussi bien sous Solaris, Aix, Windows et Linux. Elle s'adresse aux documentalistes et responsables de veille, aux utilisateurs (marketing et commercial, comités directeurs), aux services informatiques. Elle réalise le sourcing, la collecte, le filtrage, la diffusion de l'information.
La technologie de catégorisation est celle de la société Amoweba, une technologie statistique, auto-apprenante, indépendante des langues. Des réseaux neuronaux sont à l'origine d'une technologie de clustering incrémental. Chaque nouveau document permet de mieux définir les nouvelles alertes.
En amont, les outils de filtrage, ne gardant que l'information pertinente, ont été développés par Digimind.
Evolution a un an et demi mais représente 12 années hommes de recherche et développement. Elle s'adresse surtout aux décisionnaires, et va au devant de l'utilisateur final.
Parmi ses 200 clients, petites, grandes entreprises et institutions, on peut citer Aventis Pasteur, Groupama, Total .
Elle répond à des problèmes de R&D pour identifier de nouveaux produits et de nouveaux fournisseurs, ou à des problèmes commerciaux, pour obtenir des avantages compétitifs sur le marché, et surveiller la concurrence.
Le mode d'interrogation est convivial.
Un push e-mail quotidien apporte le résumé de toutes les nouveautés, surlignées.
La personnalisation peut être faite par le destinataire ou par un service central, l'entreprise choisit. On crée facilement un espace de veille avec les collègues, et d'une seule personne à un groupe, le système monte en puissance
Evolution, plate-forme serveur est interrogée à partir d'un navigateur web.
Digimind travaille à l' amélioration constante des outils amont de collecte et de filtrage et à l'ergonomie.
"Nous voulons apporter de la valeur au client. Si le client reconnaît la valeur, il est en mesure d'acheter, " souligne Patrice François, directeur associé. "Augmentation du CA, des parts de marché, diminution des délais de conception, on s'efforce d'apporter un ROI à nos clients. L'administrateur traduit les questions en requêtes. L'utilisateur final est destinataire d'infos ciblées par rapport à ses attentes."
Strategic Finder est vendu 457 euros. Evolution se situe entre 5000 et 150000 euros. Il comprend 8 modules dont 7 d'acquisition et un de gestion de communauté. La solution se décline par fonction, marketing, R&D, documentation, veille, et par secteur, telecom, pharmacie, énergie, automobile, etc..

Evolution au Ministère de l'Intérieur

Le Centre d'Etudes et de Prévision du Ministère de l'Intérieur est une petite structure – 4 chargés de mission sous la direction d'un préfet, qui dépend directement du Ministre. Béatrice Fournier y est chargée de mission pour l'intelligence économique et la veille. Elle suit les activités de Digimind depuis sa création et a opté pour un service de veille automatisé. "Plus besoin d'aller voir les sites qui nous intéressent!"

Béatrice Fournier a débuté avec "Monitor", et au renouvellement de son abonnement, a choisi "Evolution", avec lequel il est possible de valider les informations que l'on diffuse automatiquement aux utilisateurs. "Je suis l'administrateur et j'envoie les informations aux chargés de mission et aux directeurs. J'ai un portail personnalisé, mes correspondants également." Les agents qui lui sont le plus utiles ont été mis en alerte les premiers. Il y a 150 sites en alerte. Il en faudrait beaucoup plus. "Notre champ de compétences est extrêmement large, - sécurité, problèmes des préfectures, citoyenneté. Ce sont uniquement des sites Internet, en partie institutionnels (1er Ministre), associatifs (syndicats de police), presse d'information générale, abonnements à des revues de presse. 400 alertes, n'est-ce pas trop pour une seule personne? "Ce sont des sites qui ne changent pas forcément tous les jours. Exemple : les sites de centres de recherche en sciences sociales; mais quand ils bougent, c'est très important pour nous."

Béatrice Fournier est d'ores et déjà satisfaite du produit. Elle a commencé une évaluation avec les utilisateurs qui reçoivent les alertes par e-mail, et auxquels cela permet de réagir plus rapidement. Le sourcing est correctement fait. "Des améliorations sont possibles. Je n'ai pas accès au web invisible (toutes les pages qui ne sont pas indexées avec des mots-clés), aux news, aux forums. Et sur les fonctions de base, j'aimerais avoir plus de statistiques – pouvoir croiser des statistiques sur des sites et des agents, ce qui est tout à fait possible."

Pour la montée en puissance, il faut encore attendre 4 à 5 mois, ensuite le système sera bien rôdé, la vision du produit sera plus complète. "Je n'avais pas le budget mais je pense que l'accompagnement proposé par Digimind est tout à fait intéressant. Il est précieux de pouvoir bénéficier de cette partie conseil également. Pour parler en francs, mon budget ne dépasse pas 50000F, avec le conseil, il devrait être de 70000 F."

www.digimind.fr/

Influo et les signaux faibles

Issu de 15 ans de R&D au CNRS, redéveloppé par Sensoria, Influo occupe une dizaine de personnes en R&D. Il est un produit de la technologie des réseaux neuro-mimétiques la plus récente. Celle-ci mime le fonctionnement du cerveau humain dans ses manifestations les plus subtiles telle la capacité d'oubli ou l'intervention de connaissances de différents niveaux en même temps. Ce type de simultanéité permet d'identifier un mot alors qu'il y a beaucoup de bruit. Le moteur, totalement automatisé, "comprend" les documents et "comprend" les requêtes. Il est possible pour interroger d'utiliser les opérateurs booléens et avec "sauf", enlever un pan entier du contexte. (Exemple : la sexualité sauf la pornographie!)
Clients cibles : les éditeurs de contenu et les entreprises qui veulent faire de la veille et ont un besoin d'automatisation.
Le moteur assiste dynamiquement l'utilisateur avec une évolution du contexte dans le temps. Exemple : Lewinsky a un lien avec Clinton en 99, plus en 2003! A chaque nouvelle information, la connaissance change. Cf (illustration): la cartographie de l'évolution des bulles sémantiques construites autour du terme Sécurité.
1/ Influo apprend les mots, 2/ il tisse les liens pour tous les mots. 3/ Il remet en cause sa connaissance en fonction des nouveaux arrivages. Dynamique, la connaissance est de plus en plus stable…Lien pertinent, le mot est sur du blanc, signal faible, gris, non pertinent, noir.

Après un premier produit, vendu au groupe Tests en partenariat avec Eurocortex pour 01net, Influo travaille à une nouvelle version de plus en plus automatique qui est le résultat de cette expérience.
"Notre philosophie, c'est l'automatisation à outrance et la tranquillité d'esprit, "plug and find", explique le CEO Ralph Villoing. "Notre marge de progression est dans l'automatisation. Nous sommes de ceux qui vont le plus loin dans l'automatisation. Nous n'avons pas cette crainte de laisser la main à un système. En veille, nous avons une valeur ajoutée importante pour détecter les signaux faibles, même si le mot n'apparaît qu'une seule fois.
L'accès à l'information va/doit déterminer le ROI. Jusqu'ici les outils n'étaient pas assez efficaces."
La V2 met le moteur en situation. Son échelle de notation pour la pertinence, va de 1 à 95%, et il "ose" mettre des notes basses. Demain, l'utilisateur pourra constater que c'est un outil qui lui rend service. Le prix d'Influo se situe entre 15000 et 70000 euros en fonction du nombre de documents à indexer.

"Influo a de réels atouts pour la veille"

Michaël Thevenet, journaliste, responsable de la plate-forme 01net a veillé aux performances d'Influo sur son site :

Veille : - En fonction de votre expérience, en quoi Influo est-il différent d'autres moteurs de recherche? Où est son "plus"?
MT : - Lorsque j'ai effectué un tour du marché pour 01net, fin 1999, j'étais à la recherche d'une solution "technologique" susceptible d'attirer les internautes par la pertinence des résultats proposés. Les différentes offres d'alors étaient très similaires : des moteurs de recherche classiques (Verity, Autonomy & co) auxquels des surcouches sémantiques étaient

ajoutées pour améliorer la pertinence des résultats. Point commun de ces offres : des coûts très élevés en acquisition mais aussi en maintenance (thesaurus à valider par des humains, en particulier).
Influo apportait un outil entièrement automatique, intégrant des fonctions de suggestions orthographiques et sémantiques, pour le prix d'un moteur de recherche classique. Autre avantage, le logiciel fonctionnait sur une petite configuration (un serveur PC sous Windows) là où les concurrents demandaient des configurations plus musclées.
De plus, influo permettait un traitement en temps réel des nouveaux articles alors que le moteur Verity que nous utilisions alors imposait de créer un index temporaire pour la journée, index qui était fondu dans l'index général durant la nuit.

Veille : - Pouvez-vous donner un exemple d'extraction de connaissance ?
MT : - Si je tape 'thucruk', aucune occurrence du mot n'existe dans les articles publiés sur le site, mais le moteur me propose (en suggestion orthographique) 'tchuruk', qui est bien le nom que je voulais taper. Personne n'a expliqué au moteur que 'thucruk' pouvait signifier 'tchuruk'. C'est la lecture des articles de 01net (où le patron d'Alcatel a souvent été cité) qui lui a permis d'extraire cette connaissance.
Si je tape 'portable', le moteur me renvoie plusieurs milliers de réponses : la requête est ambiguë. S'agit-il d'un portable PC ou d'un téléphone portable ? Dans la liste des suggestions sémantiques, je vois 'téléphone'. C'est bien un téléphone portable qui m'intéresse, je sélectionne le mot et trouve alors un bon millier d'articles qui ne parlent plus que de téléphones, et dans les suggestions sémantiques, je vois encore comment affiner ma recherche : au mot 'téléphone' sont liées les notions 'WAP', 'SFR', 'messagerie', etc.. Là encore, personne n'a indiqué au moteur que 'portable' était lié à 'téléphone', ni que 'téléphone' était lié à 'WAP', 'SFR' ou 'messagerie'. C'est bien le moteur qui a extrait cette connaissance du fonds documentaire.

Veille : - A-t-il été facile à installer?
MT : - Cela a été une des briques les plus simples à installer dans la plate-forme technique. En fait, les développements ont été très réduits : deux semaines, dont une a été perdue du fait d'une mauvaise configuration du serveur (l'erreur était du côté de notre intégrateur qui avait installé une version de Windows qui n'était pas supportée par Influo). Le plus long a été de mettre au point le front office, c'est-à-dire de trouver le mode de présentation des résultats accompagnés des suggestions sémantiques.

Veille : - Comment ont réagi les journalistes, et les lecteurs de 01 Net?
MT : - Les journalistes ont rencontré quelques problèmes de recherche d'articles : parce que la prise en compte du nom des auteurs n'a été que tardivement activée dans la stratégie d'indexation ; là encore c'est un choix de l'équipe de 01net et non le fait du moteur.
Les internautes ont plébiscité le moteur : la part de pages consultées sur le site à partir de requêtes faites au moteur a triplé dans les six mois qui ont suivi la mise en ligne du moteur. Cela est très probablement dû, d'une part, aux suggestions orthographiques, d'autre part, à la qualité des suggestions sémantiques qui accompagnent l'affichage des listes de résultats .

Veille : - Que pensez-vous d'Influo pour une application de veille?
MT : - Influo a de réels atouts pour la veille : rapidité d'indexation des nouveaux documents et pertinence de la connaissance construite, le tout de manière entièrement automatique. La principale difficulté ne vient pas du moteur lui-même mais de la conception de l'interface de consultation des résultats. Dans le cadre d'une application de veille, il me semble tout à fait

possible de construire une interface simple qui permette d'effectuer une traque efficace sur mots-clés (certainement plus efficace et plus pertinente que ce qu'offre, par exemple, Google) dans le cadre d'un Intranet : par exemple en extrayant en temps réel les connaissances des documents arrivant sur un serveur de l'Intranet alimenté par un robot. De l'expérience que j'en ai , la pertinence attribuée par le moteur d'Influo aux documents est digne de confiance, ce qui constitue un atout irremplaçable pour des veilleurs victimes du fameux syndrome "information overload".
www.influo.com/

5
Inxight Software et la visualisation des résultats

Cette entreprise américaine de Californie (100 personnes dont 10 en Europe, à Anvers et Grenoble), spin off de Xerox, a été fondée en 96-97 par des chercheurs de Xerox autour de deux technologies, linguistique (extraction de concepts) et visualisation (qui a donné lieu à un brevet).
"Nous sommes spécialistes en catégorisation, extraction, taxonomie, visualisation des résultats, explique Peter de Bie, ingénieur commercial, et sommes passés de 80% à 60% de ventes en OEM, adressant directement les entreprises fournisseurs de contenu, (Reuters, AFP), et pour extraction des entités nommées, le monde de la pharmacie, du gouvernement, des brevets (European Patent Office).
Inxight Software propose trois logiciels. Inxight Smart Discovery, est un produit de structuration, de navigation dans de grands volumes de documents, tels des référentiels comme Documentum, Lotus Notes, Plumtree qui peuvent utiliser conjointement un autre index tel celui de Verity.
Smart Discovery est plus qu'un moteur de recherche, puisqu'il permet de naviguer dans les données, d'extraire des concepts de la collection de documents, de filtrer les entités (noms de personnes, informations géographiques, métadonnées), de réaliser de petits résumés dans le concept de l'utilisateur. C'est le profil de l'utilisateur qui va déterminer le contenu du résumé.
A Smart Discovery succède un autre produit Inxight Categorizer, vers lequel évoluent les clients lorsqu'ils sont plus expérimentés. Le troisième produit breveté, Inxight Vizserver construit des arbres de visualisation pour données non structurées- Table Lens permettant de son côté de visualiser les données structurées.
Cette offre se caractérise par la pertinence de sa catégorisation et le caractère unique de sa visualisation.
Pour l'extraction des concepts qui fait appel à des systèmes d'apprentissage, Inxight Software combine techniques statistiques et linguistique. Inxight Software passe de l'extraction des entités nommées, à l'extraction des relations entre entités nommées (fact finding), suite au rachat de la technologie Whizbang! Labs.
"Un moteur n'est pas la solution mais le problème, " souligne Peter de Bie," avec de trop longues listes de résultats. La recherche fédérée ajoute des sources et multiplie les listes de résultats. Il faut résoudre le problème des grandes collections de documents par la taxonomie et la structuration de l'information. Nous optons pour le tout automatique et ne proposons pas d'outils collaboratifs, préférant un partenariat avec IBM sur la collaboration."
Inxight Software se vend sur serveurs de 1 à 4 processeurs de 100000 à 300000 euros.
Plus Intranet qu'Internet, Inxight Software ne structure pas tout l'Internet, mais les sources intéressantes pour améliorer l'information en interne.

Un produit hors cadre

En France, Inxight Software est distribué de manière exclusive par la société de conseil et d'ingénierie Netfective. "Inxight est un produit hors cadre", note Anwar Guerch, responsable du développement. "Il répond à un besoin très spécifique des entreprises. C'est un outil de mise en perspective de l'information. Il donne une vue hyperbolique. Il permet non seulement d'accéder à un contenu via un certain nombre de requêtes, mais de visualiser la structure de l'information." Netfective vend Inxight en direct à des clients qui ont une certaine taille, ou du moins dont l'information a une taille critique, tels Peugeot, Renault, l'AFP, les Intermarché.
C'est un produit transverse et multi-usage, - veille, gestion documentaire, on l'intègre à des

systèmes de workflow. Le partenariat est né d'une proximité humaine et d'une adhérence réelle aux fonctionnalités de l'outil. "Par sa combinaison visualisation-sémantique, cet outil est unique. Il n'a pas d'équivalent". Renault songe à l'utiliser en prospective, avec une notion d'historique, dans sa gestion de la sous-traitance, pour raccourcir le temps de recherche dans le dédale des informations échangées avec les sous-traitants.

Mireille Boris

www.inxight.com/

6
Sinequa et l'espace vectoriel

Le traitement automatique des langues, est le cœur de métier de Sinequa depuis vingt ans. Sous le nom de Cora jusqu'en 2000, la société composée d'une quinzaine d'ingénieurs produisait des correcteurs grammaticaux, des générateurs de textes, des résumeurs automatiques et un outil phare Darwin, logiciel de gestion documentaire intégrant de l'analyse linguistique au niveau syntaxique.
Suite à l'explosion d'Internet, Cora a sorti en 1998, la version 1 de son moteur de recherche Intuition, avec plusieurs innovations, la principale étant l'analyse sémantique. L'équipe a trouvé le moyen de représenter le sens global d'un document sous une forme mathématique, en s'appuyant sur les différents sens des mots du document. Un brevet a été déposé aux Etats Unis. Intuition repose sur un dictionnaire sémantique à partir duquel le langage est modélisé dans un espace vectoriel indépendant de la langue. Ce dernier comprend environ 800 dimensions (ou concepts) et 400000 lemmes. Intuition identifie plus d'un million de termes différents et identifie le sujet du verbe et le complément. Parmi ses grandes fonctionnalités : la recherche intelligente et multilingue, la comparaison et la classification de documents similaires, l'identification des attentes d'un utilisateur selon les questions qu'il formule et les documents qu'il consulte.
Avec ses profils personnalisés, Intuition sait construire un vecteur sémantique des centres d'intérêt de l'internaute pour de la publicité ciblée ou bien du push sélectif. La recherche "intuitive" contribue à la navigabilité des sites.
Intuition de Sinequa est encore vendu à 50% en OEM. " OEM et vente directe sont complémentaires" affirme Luc Manigot, directeur scientifique, "cependant le contact direct avec le client permet de comprendre où sont les vrais besoins".
Racheté courant 2002 par la Revue Fiduciaire, Sinequa conserve une totale indépendance technique." Le moteur est déployé sur le site de la revue fiduciaire. Tout se passe vraiment très bien. Cette société nous ouvre un certain nombre de portes. Mais nous avons un moteur extrêmement polyvalent. Nous ne faisons pas d'Intuition un modèle dédié au domaine juridique! Ce moteur est différent des autres, parce qu'il regroupe toutes les fonctionnalités des autres et un peu plus…"
Depuis qu'il est Revue Fiduciaire, Sinequa a gagné comme nouveaux clients, le Monde, l'Humanité, Ouest France et la COB.
"Nous souhaitons accélérer notre développement," déclare Laurent Le Foll, Directeur Général Adjoint. " Je ne donnerai pas d'objectifs chiffrés, mais nous sommes en mesure de répondre à des secteurs bien identifiés comme la Presse. Pour les sites Internet de e-Commerce, nous avons les algorithmes pour "désambiguiser" les questions afin que l'utilisateur trouve les produits qu'il cherche. Leroy-Merlin, un de nos plus anciens clients, avec La Redoute, signale que ses prospects sont de plus en plus à l'aise avec les questions…Nous correspondons enfin au marché des Intranets d'entreprise.
Dans le débat, humanisation, automatisation, nous sommes en tant que Sinequa dans une position intermédiaire. Nous essayons d'automatiser le maximum de ce qui est possible, tout en laissant la main à l'utilisateur. Il n'y a pas une chapelle d'un côté, une chapelle de l'autre. Les dictionnaires linguistiques, les analyseurs, les bases de connaissances ont été développés par des humains. Tout ce travail humain a été capitalisé. Le service d'indexation de DIVA va valider ce qui a été fait par le moteur pendant la nuit. Nous laissons toujours une part à l'adaptation. Nous sommes sceptiques sur le 100% automatique."

DIVA Press, à la lisière de la veille

S'appuyant sur Sinequa, DIVA Press a réalisé une très belle application de contextualisation de l'information, à la lisière de la veille. Ses outils d'extraction de connaissance facilitent en particulier la lecture rapide de gros volumes de documents.
Historiquement proche du monde de la presse,créée en 1999 au sein de l'AGEFI, la plate-forme DIVA a en base électronique la presse quotidienne nationale, les hebdomadaires et mensuels économiques et financiers, des fils de news, etc.. Elle propose à ses utilisateurs une gamme extrêmement variée de produits, qui vont de la recherche documentaire au panorama de presse et aux outils de veille sur profil. Pierre Briand, 32 ans, directeur de la stratégie et du développement du groupe Finintel se voit confier la direction générale de DIVA-Press, et entend bien faire progresser le numéro 1 de la distribution électronique de la presse économique et financière de langue française (le Financial Times en plus a été pressenti…).
Recherche documentaire à partir d'un fonds d'archives, alimentation de panoramas de presse sur intranet, outils de veille active sur profil via des alertes électroniques ont été packagés dans trois séries d'offres.
Dans "Picking", l'utilisateur trouve DIVA Solo pour les consultations ponctuelles et les recherches documentaires à partir d'un fond d'archives. DIVA Expert en plus envoie des alertes quotidiennes.
Dans "Clipping", la DIVA revue de presse est proposée ainsi que DIVA Line, chaînes d'information pré-packagées. Il s'agit de la diffusion collective par mode Intranet des panoramas de presse et des informations stratégiques pour les entreprises vers le responsable d'un centre de documentation ou d'une cellule de veille. Panoramas de presse, dossiers sont préparés par des consultants éditoriaux qui affûtent les outils de veille automatique.
Dans "Ticking", c'est la solution technologique, DIVA marque blanche, elle-même, qui est proposée.
On notera les alertes sur profil, le profilage pour veille avancée, la constitution de panoramas de presse par l'utilisateur ou par les consultants, la contextualisation et l'enrichissement de l'information présentés dans un espace ergonomique à trois dimensions, les pushs contextuels, la génération de pertinence par filtrage et contextualisation, l'aide à la lecture rapide par surbrillance de mots clefs. Au cours d'un tri dynamique et sémantique, les réponses sont fournies par ordre de pertinence.
M.B.
www.sinequa.com/

7
Pertimm et la sémantique contextuelle.

Le nom de la marque est Pertimm, pour "pertinent et immédiat". Son logo représente un petit martin-pêcheur (Dipper) qui rapporte des pierres précieuses de ses plongées. La société a été créée en 1997 en France par une équipe de trois ingénieurs, Patrick Constant, Xavier Mignon et Jean Poncet qui ont uni leurs compétences dans la linguistique, l'intelligence artificielle, les bases de données et les systèmes d'exploitation temps réel. Elle est implantée depuis deux ans aux Etats Unis également, à Orlando, Floride et emploie une douzaine de personnes.

Son offre technologique s'appuie sur une "invention" (un brevet avec 81 points a été déposé aux Etats Unis) dont l'objet est de permettre la navigation dans les contenus, par des requêtes conceptuelles, les plus proches possibles d'une conversation humaine. L'utilisateur peut naviguer dans l'ensemble des informations disponibles, en conversant avec le système et en gardant le contrôle sur les concepts qui guident sa conversation.
Ce système translingue (Pertimm est au format Unicode) permet à partir d'une requête exprimée dans une langue d'obtenir des réponses dans plusieurs autres langues.
"Nous nous sommes toujours défendus de dire que Pertimm était un moteur de recherche, en raison de la mauvaise réputation des moteurs de recherche! Nous proposons un outil qui permet de voir le contenu et de prendre des décisions en fonction du contenu qu'on a vu," déclare Xavier Mignon, Directeur Marketing.
Pertimm est un outil d'accès, de visualisation, de navigation dans le contenu. Les américains l'ont qualifié d'outil de sémantique contextuelle. Il fonctionne en effet sur la notion de contexte. Les fichiers sont découpés de manière sémantique, en éléments qui portent un sens. C'est une stochastique. Le contexte fait la richesse des contenus. Mais Pertimm ne cherche pas à gérer le sens. C'est l'utilisateur qui va donner le sens.
Pertimm est basé sur l'analyse morpho-syntaxique du contenu. De la statistique est ensuite appliquée aux résultats d'analyse linguistique, des réseaux lexicaux également ainsi que certaines techniques neuronales qui permettent d'aller chercher plus vite de l'information dans un contenu. A point nommé, la fonction mathématique qui paraît la meilleure est utilisée.

Pertimm propose trois produits ou plutôt trois formes de commercialisation du produit.
Historiquement, les APIS (Application Program Interfaces) de Pertimm s'adressent à des partenaires, revendeurs ou intégrateurs tels Ennov ou Alogic qui prennent la technologie et l'intègrent dans leurs produits.
Ensuite Pertimm Web s'adresse aux grands comptes. Il fédère des Intranets. Les 800 serveurs http du CNRS sont consultés à travers un serveur Pertimm via un browser. Pertimm permet aussi aux 20.000 personnes de l'ANPE de trouver l'information dans la documentation interne de ses 22 régions qui seront bientôt 26. A la BNF, il fédère des serveurs Lotus/Notes et des serveurs http au travers d'une seule requête. Pertimm est "Day One". Aucun besoin de formation. Une journée suffit pour que tout le personnel l'utilise. On ne touche pas aux serveurs existants. Pertimm qui est incrémental et décrémental en quasi temps réel est installé sur une autre machine. Il est en principe illimité. Il n'impose pas de limites inférieures à celles de la machine sur laquelle il est installé.
Troisième mode de commercialisation, le service WebProDipper. Les sociétés qui font de la veille s'adressent à des spécialistes de la veille. Pertimm intervient comme fédérateur de ces abonnements. Il va permettre de les filtrer et de n'avoir qu'une seule requête. Il uniformise les

abonnements. Il sert de tampon par rapport à ces contenus. Bureau Van Dijk vient de créer avec Pertimm un portail qui s'appelle Piste, dans le domaine de l'agro-alimentaire. Différents partenaires EDF du secteur de l'énergie, sont fédérés par un accès Pertimm. A l'Office Européen des Brevets, Pertimm apporte la partie gestion de web dynamique; ses outils permettant de dire si une information est nouvelle ou non. WebProDipper est de l'ASP (Application Service Provider). Le site technique de Pertimm est hébergé en salle blanche à Courbevoie chez LD-Com. Pertimm y dispose d'une ligne de 100 Mb sécurisée par plusieurs firewalls. Comportant près de 20 serveurs bi-processeurs sous Linux, il héberge le service de WebProDipper.
L'effort de développement de Pertimm le plus important actuellement porte sur l'analyse d'image, pour l'optimisation de recherches combinées, texte et images, en particulier dans le domaine de l'imagerie médicale et de l'aide au diagnostic.
Le produit Pertimm est diffusé en location annuelle, dont le prix est fonction du nombre de fichiers et d'utilisateurs. Pour 3000 utilisateurs et 100000 fichiers, il faut compter 30000 euros/an.
Mireille Boris

8

Verity et la recommandation d'experts

Les entreprises disposent de contenus sur des supports divers et variés, dans des systèmes de gestion documentaires, des systèmes de fichiers, sur d'autres petits Intranets, sur des ERP, CRM, dans des bases de données…Cette information est associée à un certain niveau de sécurité, par utilisateur, par métier, avec des autorisations gérées à travers un annuaire d'entreprise.

–"Premier point : Verity K2, numéro 1 des moteurs de recherche, (420 personnes dans le monde, 70 en Europe, 14 en France), se positionne comme le point d'accès de cette information disponible", explique Sylvie Pichot, consultant avant-vente.

Le moteur de recherche va d'abord proposer une indexation, et un référencement de toutes les indexations. Verity fournit les passerelles spécifiques dédiées à chaque type de source. Il supporte plus de 250 formats de fichiers. "La première étape de la vie du moteur consiste à être l' accès unique et fédérateur d'informations qui peuvent être écrites dans 26 langues".

Il propose un ensemble de dictionnaires pré-renseignés, des listes de mots vides de sens modifiables par le client, des dictionnaires de règles qui vont permettre de prendre en compte les différentes formes de conjugaison possibles, les pluriels irréguliers et enfin des paramètres de type structurel pour définir ce que c'est qu'une phrase dans nos langues européennes. "Le moteur est dit plein texte, mais ne se limite pas à des recherches de type plein texte. On peut aussi faire de la recherche sur champ, sur zone, une fois qu'a été établie l'infrastructure de référencement".

Les multiples fonctionnalités de Verity K2 ont été regroupées en 3 tiers fonctionnels, le premier tiers décrit précédemment ayant pour but la découverte de l'information. Le deuxième tiers, ensuite, va permettre d'organiser cette information, la catégoriser, la classifier, en extraire des concepts, créer des réseaux sémantiques. L'utilisateur va pouvoir naviguer dans l'information sans forcément exprimer une requête. "Le troisième tiers où sont les fonctionnalités les plus avancées, va apporter une connotation humaine au moteur de recherche. On va parler de réseaux sociaux, de détection de communautés virtuelles d'utilisateurs, avec un même sujet, un même métier. On va recommander à ces utilisateurs des documents complémentaires, et faire de la recommandation d'experts, pour arriver à un haut niveau de personnalisation de l'interface de recherche."

A la recherche paramétrique qui aide à réduire le périmètre des listes de résultats, Verity joint la recherche fédératrice qui a vocation d'ouverture vers l'extérieur. Verity K2 fusionne les retours d'information sur une seule et même page. Des partenariats avec des fournisseurs de contenu tel Lexis-Nexis permettent de proposer aux clients d'élargir leur information avec les news diffusées par ces sociétés.

Le point fort d'une solution comme Verity K2 c'est cet ensemble de méthodes à disposition de l'utilisateur, conciliables les unes avec les autres. Il est important de pouvoir proposer à l'utilisateur des plans de catégorisation qui lui soient adaptés. Le point de vue d'un veilleur est différent de celui d'un marketeur ou de celui d'un membre de la direction. Verity K2 construit plusieurs plans de catégorisation.

"Le dernier tiers prend en compte un retour des utilisateurs, de façon implicite ou participative, et l'utilisateur va pouvoir voter. La liste de résultats traditionnels est complétée par de la recommandation complémentaire. On sait recommander les documents, on sait aussi recommander des experts. On humanise le retour d'un moteur de recherche."

"Faire intervenir les gens pour humaniser le produit."

Didier Donnat , Directeur Général Verity

Veille : - Pourquoi proposer des réseaux sociaux, et faire intervenir le KM?

DD : - Pour leur efficacité, bien sûr! Nous faisons intervenir les gens pour humaniser le produit. Cela a toujours été dans notre stratégie. La boîte noire, le tout automatique n'a jamais été notre culture. On favorise au maximum l'ouverture technique, fonctionnelle et l'apport humain, des utilisateurs, des experts, pour définir des centres d'intérêt, de catégorisation personnalisée au sens strict du terme, en fonction de l'individu.

Veille : - Que devient Inktomi chez vous?

DD : - Il y avait deux activités chez Inktomi, un service pour le web racheté par Yahoo et une activité produits Inktomi Search Enterprise, que nous avons rachetée. C'est un produit concurrent sur certains secteurs. Inktomi, rebaptisé Verity Ultraseek, nous permet en fait de revenir sur un marché qu'on avait un peu laissé de côté, des projets d'entrée de gamme, départementaux, Internet. On est peu sur Internet. Quand un client déploie pour 50000 utilisateurs en interne, on prévoit la part d'Internet, mais ce n'est pas notre cheval de bataille. Nous sommes focalisés sur les portails d'entreprise et les Intranets. Inktomi nous permet de ré-aborder les sites Internet, les PME, les applications plus verticales. Les solutions sont tout à fait complémentaires en terme de fonctionnalités et de cibles applicatives. Nous avons en outre récupéré 2500 clients dans le monde, lors du rachat. Nous avons développé un connecteur pour relier l'infrastructure K2 au moteur Ultraseek. Les documents indexés et organisés par Ultraseek sont ainsi récupérés, fédérés par K2. Ultraseek devient une source d'information comme une autre, au travers de la recherche fédératrice. Si les clients Ultraseek ont des besoins K2, de mise en place de réseaux sociaux, on va fédérer cela.

Pour les nouveaux clients, en fonction des besoins, c'est Ultraseek ou K2. Les algorithmes que nous développons, permettent d'élargir la couverture fonctionnelle. Ultraseek apporte la maîtrise de XML, ce qui permet d'utiliser des métadonnées. A horizon trois ans nous allons fusionner les avantages des deux offres…98% des demandes de techniques de recherche sont couvertes par Verity K2 et Ultraseek.

M.B.

www.verity.com/