Chronique

de la recherche sur le web

par Marc Duval 14 août 2002, vol.2, n°7

Tout le web francophone ou le web selon Voila

Dans le monde de l'Internet et des sites de recherche, Voila est unique. Il est à la fois portail, répertoire et «moteur de recherche». Ces trois chapeaux ont fait en sorte que Voila est devenu une source de malentendu. La notoriété de l'automate de recherche a pris le pas sur les deux autres fonctions.

L'automate de recherche est né en 1996 chez la société Echo. A cette époque héroïque, il était l'éponyme de la société. Christophe Dupont, Christophe Ruelle et Michel Bisac sont les piliers de la société. Christophe Dupont est le concepteur, Christophe Ruelle l'informaticien et Michel Bisac le président.

En 1997, l'automate Echo est intégré au site PagesWeb de France Télécom. Puis, il prend son essor international avec la participation financière de France Télécom dans la société.

En juillet 1998, France Télécom inaugure son portail Voila. L'orthographe du nom du portail est encore hésitant. Tantôt, il s'écrit avec un accent grave, tantôt sans accent. En fait, il s'écrira sans accent pour refléter son adresse et ses ambitions internationales. Dès lors le portail Voila et l'automate ne font qu'un dans l'esprit des internautes. Pourtant son vrai nom est «Tout le web francophone», un nom pas si commode, que l'on surnommera KE, sigle de «knowledge engineering» signifiant «ingénierie des connaissances», «génie cognitif» ou la «cognitique».

Dès le début, Echo innove avec la présentation du résultat de recherche. Il est le premier automate à délivrer un descriptif des mots en contexte, la page indexée et une fonction de marqueur. Google reprendra le format du descriptif, appelé kwic pour keyword in context, et la page indexée. Il a tout ce qu'il faut pour rivaliser avec ses concurrents américains qui ont un pied à terre en France: AltaVista France, Lycos France, HotBot France, MSN France et même Yahoo! Plus, il a une plate-forme pour les professionnels de la recherche: Voila pour les geeks. Sa trousse de recherche comprend les opérateurs booléens, de proximité et d'adjacence, un formulaire avec la recherche par tous les champs disponibles: texte, hyperlien, texte de l'hyperlien, domaine générique et géographique, nom de domaine, URL, serveur, balises «keywords», «title» et «description», texte de substitution d'une image, fichiers multimédia. Mieux, il introduit la recherche thématique: une véritable liste de mots-clés. Tout est transparent, sur une même page. Son aide est aussi appuyée par la notoriété d'un spécialiste en la matière: M. Olivier Andrieu du site abondance.

En 1999, les ambitions de Voila sont internationales. Il est prévu d'essaimer en Europe et en Amérique et d'atteindre les milliards de documents. Il réalise sa première ambition en ouvrant le site américain «voila.com» qui recherchera entre autres des fichiers Windows, Mac et Unix. Voila aura des interfaces en Allemagne, en Belgique, au Danemark, en Espagne, Italie, au Pays-bas, au Portugal et au Royaume-Uni. Malheureusement, l'aventure de «voila.com» s'arrêtera en 2001. L'adresse sera redirigée vers «Voila pour les Geeks». L'ambition internationale sera remise entre les mains de Wanadoo implantée en Belgique, en Espagne, en France, en Hollande, au Liban, à Madagascar, au Maroc, au Sénégal et au Royaume-Uni.

En 2002, sa banque de données compte 60 millions de pages couvrant les principaux pays de la francophonie. Sa seconde ambition d'atteindre le

milliard de documents ne sera pas atteinte. Pour s'approcher du nombre magique, il utilisera le répertoire de l'Open Directory Project passé à la moulinette de son indexeur. Cette seconde banque de données se nommera le Guide mondial.

En août 1999, le portail Voila est classé 4e en France par NetValue et Wanadoo est 1er. Un an plus tard, Voila est 2e derrière Wanadoo. En 2001, Voila est 4e et Wanadoo toujours premier. En juin 2002, le classement est identique.

La société 1ère position nous donne d'autres statistiques. En septembre 2001, Voila était en 3e position avec 18% de la fréquentation, derrière Yahoo et Google. En juin 2002, toujours 3e, il voit sa fréquentation baissée à 13%.

Bon an mal an, les portails Voila et Wanadoo qui utilisent KE se maintiennent dans le peloton de tête.

En juin 2002, Voila change sa philosophie. Il se «recentre» sur la France et ses capacités de recherche rétrécissent comme une peau de chagrin. Voila semble avoir jeté la serviette. Il concède la victoire à Google qui accapare plus de 50% du trafic.

Son logiciel de recherche est simplifié et des opérateurs sont retirés: les booléens OU/OR, ET/AND, SANS/NOT, et les opérateurs d'adjacence NEAR et PROCHE. Ses fonctions d'affichage deviennent des options placées dans la page de personnalisation. Le «classement par popularité» et la recherche de fichiers FTP disparaissent. Finalement, le résultat de recherche devient médiocre à moins de connaître à fond son fonctionnement.

Cette transformation s'appuie sur les habitudes de recherche de l'internaute moyen. Ce dernier recherche avec un minium de contrainte: il utilise trois mots et moins sans stratégie de recherche spécifique ni d'affichage sophistiqué. Voila répond aussi à la concurrence en insérant la correction d'orthographe et des références provenant des actualités et de l'encyclopédie Hachette. Jusqu'ici, il peut répondre adéquatement à la majorité des internautes, mais son talon d'Achille est la reconnaissance des caractères.

Les faiblesses et les forces de Voila

Voila n'est pas qu'un automate de recherche. Il est aussi un répertoire. Or la recherche en page d'accueil est effectuée autant dans le répertoire que dans l'automate. C'est pourquoi les faiblesses et les forces toucheront aussi le répertoire.

En changeant d'orientation, Voila/KE s'est affaibli. Plusieurs de ses caractéristiques ont disparu. Six faiblesses ont été relevées: a) la non reconnaissance des accents; b) l'absence des opérateurs d'union et d'adjacence; c) la rigidité et le dysfonctionnement du formulaire; d) le cloisonnement des banques de données; e) la recherche sur la page source et f) la publicité.

Du côté des forces, quatre sont à retenir: a) le nombre de documents; b) la délimitation des sections; c) la recherche thématique et d) la recherche avancée.

Les faiblesses de Voila

Six faiblesses ressortent de l'utilisation de Voila:

1. Non reconnaissance des accents
2. Absence des opérateurs d'union et d'adjacence
3. Rigidité et dysfonctionnement du formulaire.
4. Cloisonnement des banques de données.
5. La recherche sur la page source
6. La publicité

1. Les accents et les majuscules

Voila/KE n'a jamais reconnu les accents, ni les majuscules. Dans la plupart des cas, il n'y a pas d'incidences. Il peut même être un avantage comme la recherche des graphies «événement» et «événement» - c'est le parti de Voila. Toutefois, les accents sont un apport à l'optimisation de la pertinence. Dans Voila/KE, cote et côte, faîte et faite, marché et marche, intègre et intégré, etc. amènent leur lot de bruit. Les majuscules ont la même fonction que les accents. Paris, Pâris et paris donneront les mêmes résultats alors qu'un automate acceptant les accents et les majuscules saura distinguer, la ville de Paris, le héros Troyen Pâris et le nom commun «paris» dans une requête.

La non reconnaissance des accents et des majuscules est d'autant plus apparente que KE a maintenant un vérificateur d'orthographe qui n'est pas encore au point. Par exemple «évènemant» n'est pas reconnu comme une faute, mais «evenemant», et «événemant» le sont.

Non seulement, Voila/KE ne tient en compte ni les accents, ni les majuscules, il ne reconnaît pas non plus les traits d'union. La requête «pied-à-terre» aura le même résultat que «pied à terre» ou «pied terre». Pour que le mot «pied-à-terre» - et ses semblables - soit reconnu comme tel, il doit être mis entre guillemets anglais comme ceci: “pied-à-terre” ou “pied-a-terre”.

2. Les opérateurs de recherche

L'ancienne version de KE avait les opérateurs booléens d'union et d'adjacence. La nouvelle version de Voila/KE fait disparaître les opérateurs booléens d'inclusion, les opérateurs d'exclusion et ceux d'adjacence. KE recherche tous les mots d'une requête sans exception.

Voici un tableau d'après une requête modulée par les variantes des opérateurs booléens.

classement moteurs recherche =
. . . . . 8654 réponses
+classement +moteurs +recherche =
. . . . . 8654 réponses
classement ET moteurs ET recherche =
. . . . . 8654 réponses (le mot ET est commun)
classement AND moteurs AND recherche =
. . . . . 1537 réponses

***

+classement +moteurs -recherche =
. . . . . 1595 réponses
classement ET moteurs SANS recherche =
. . . . . 4422 réponses
classement ET moteurs NON recherche =
. . . . . 4018 réponses
classement AND moteurs AND NOT recherche =
. . . . . 202 réponses
classement AND moteurs NOT recherche =
. . . . . 202 réponses

***

moteurs NEAR recherche =
. . . . . 405 réponses
moteurs PROCHE recherche =
. . . . . 7055 réponses
(automates OU moteurs) ET recherche =
. . . . . 655 réponses
(automates OR moteurs) AND recherche =
. . . . . 158 réponses

3. Le formulaire

Le formulaire de recherche de Voila a toujours été fixe. Ce type de formulaire ne permet pas d'entreprendre une recherche ciblée comme c'est le cas avec celui de HotBot. Le formulaire fixe est utilisé par Google, AOL France et Lycos France. Mais chez Voila, ce formulaire n'est pas au point: la recherche d'exclusion ne répond pas à tout coup.

Le résultat de la recherche a aussi un comportement erratique dès qu'une requête fait appel à un champ de recherche. De plus, le logiciel efface le champ dès la deuxième page; le résultat de recherche est remis à zéro en utilisant une recherche simple.

4. Les banques de données

Voila a 3 banques de données principales sur le formulaire de recherche: «Sélection de sites», «Tout le web francophone» et le «Guide mondial». Ces trois banques ont chacune leurs particularités.

Avec Sélection de sites, un mot au singulier trouvera aussi ce mot au pluriel; pas avec KE.

Lorsque vous utilisez une stratégie de recherche en page d'accueil avec des préfixes, elle sera appliquée au Guide du Web et à KE, mais pas au Guide mondial.

D'autre part, le libellé de chacune des banques n'est pas suffisamment clair à prime abord. Pour être en mesure de faire la distinction entre les trois banques de données, l'internaute doit faire l'effort de consulter l'aide. Encore là, la clarté n'est pas au rendez-vous. La «Sélection de sites» devient le «Guide du Web» qui est le nom attitré de cette banque. Il serait plus approprié d'unifier les noms. «Sélection de sites» devrait porter celui de «Guide du Web» ou «Guide du web francophone» et le «Guide mondial» aurait avantage à être renommé «Annuaire ODP» ou à être identifié en bas des pages comme le fait la plupart des autres

sites de recherche comme Google. Même l'utilisation des icônes n'apporte davantage d'informations.

5. Recherche sur la page source

Comme plusieurs automates, Voila/KE fait sa recherche sur la page source. Ce mode de recherche repère les termes dans le texte visible et invisible de la page. Certains rédacteurs connaissent cette particularité et en profitent pour tromper KE. D'autres méconnaissent et induisent un bruit en utilisant les commentaires dans un hyperlien. En affichant la page source, le mot invisible apparaîtra dans la zone «title» de ce code: <a HREF="URL" title="Titre [texte invisible]="_top">Texte visible. Ce mode de recherche était adapté il y a quelques années parce que les codes sources étaient moins développés ou moins utilisés, mais aujourd'hui ces codes sont courants. Le passage de la souris sur un hyperlien ou sur une image ou l'ouverture d'une liste déroulante les révèle en donnant un renseignement supplémentaire sans alourdir la page. Mais pour un internaute moins aguerri, le fait de ne pas voir ses termes de recherche lui fait croire au mauvais fonctionnement de l'automate.

La recherche sur la page source entière devrait être proscrite dans une recherche simple et exclusive en recherche avancée. L'internaute moyen devrait avoir accès aux seules pages dans lesquelles les termes de recherche sont visibles.

6. La publicité

La publicité sur Voila est parfois insistante. Elle l'est d'autant lorsqu'elle est sous forme d'une fenêtre pop-up qui surgit lors d'une recherche. Voila n'utilise pas les liens commandités. A choisir, ces derniers seraient préférables.

Les forces de Voila

Voila a quatre forces qui surpassent ses concurrents:

1. Banque de données en français.
2. Délimitation des sections.
3. Recherche thématique
4. La recherche avancée

1. Banque de données en français

Voila/KE a la plus grande banque de données de pages web avec environ 60 millions de pages. Comparativement, Google a une banque d'environ 6 millions de pages web en français estimés, AltaVista, 28 millions, AntiSearch, 25 millions et MSN (Inktomi), 14 millions. L'estimation a été réalisée en utilisant la préposition «à». Cette lettre est un mot vide dans la plupart des banques de données en français, mais pas en anglais.

2. Délimitation des sections

Bien que les sections des banques de données n'aient pas des noms explicites, Voila délimite clairement sa «Sélection de sites», son automate «Tout le web francophone» et le «Guide mondial». Il est le seul site à le faire aussi distinctement.

3. La recherche thématique

La recherche thématique est l'arme secrète de KE. Ce mode de recherche est unique et mériterait une plus grande attention. C'est elle qui apporte le taux de pertinence le plus élevé du web. Couplée à la recherche par champ et au filtre parental, elle devient imbattable.

4. La recherche avancée

La recherche avancée avec stratégie demeure un choix de recherche intéressant bien qu'elle ait perdu un peu de lustre avec le retrait des opérateurs booléens d'union et d'adjacence. Ce qui ne l'empêche pas d'être efficace avec l'exclusion de termes. Surtout, elle demeure intacte page après page contrairement à la recherche par formulaire.

***

8 conseils pour un meilleur résultat de recherche

Si vous voulez faire une recherche fructueuse, utilisez la page des préférences. Voici huit conseils qui vous aideront n'ont plus à chercher mais à trouver.

1. Utilisez la recherche en page d'accueil pour la première approche. Le résultat de la recherche vous donnera une vue d'ensemble des différents services. Vous pourrez raffiner votre recherche en utilisant le formulaire détaillé.
2. Activez le filtre parental; il filtrera la majorité des sites pornographiques.
3. Activez le filtre orthographique; bien qu'il ne rivalise pas avec celui de Google, il saura vous rappeler certaines fautes.
4. Choisissez 20 références (ou résultats) par page ou plus; vous ne perdrez pas votre stratégie de recherche du formulaire.
5. Choisissez d'afficher par titres si vous optez pour plus de 30 références par page. Notez que les options d'extrait et de marqueur ne fonctionnent pas avec cet affichage.
6. Choisissez l'affichage du marqueur. Cette exclusivité chez KE épargne du temps à repérer les termes de recherche. Un bémol: le marqueur fonctionne seulement avec une recherche simple, sans expression. Chez d'autres automates de recherche, cette option est dans la barre d'outils.
7. Habituez-vous à la recherche par champ (liste déroule à droite de «placés») du formulaire pour réduire le bruit au lieu de la recherche simple.
8. Souvenez-vous que la recherche de fichiers est indicative et non nominative. Cette recherche est efficace si une page porte sur un seul sujet. Pour avoir cette condition, la recherche par le champ titre est recommandée.

Conclusion

Voila a subi plusieurs métamorphoses. Dans l'ensemble, il a la stature d'un grand site de recherche. Il a des atouts indéniables. Avec la dernière version, il s'est affaibli en se «recentrant» sur le marché français. La convivialité a fait place à des contrariétés qui pourraient voir l'écart de l'utilisation augmenté avec Google. Pour avoir un résultat de recherche fructueux, l'internaute doit être aguerri et utilisé les fureteurs Internet Explorer et Netscape 6.2.

L'amélioration d'un automate de recherche demande un investissement que France Telecom ne semble plus vouloir supporter. Les concurrents directs de Voila que sont AntiSearch et DeepIndex peuvent voir d'un bon oeil son recul. Avec un développement dynamique, tous peuvent aspirer à la place de «Tout le web francophone» de Voila.

En tant que portail, Voila/Wanadoo arrive bon premier. Mais en tant que site de recherche, il est à un carrefour: développer «Tout le web francophone» ou se retirer. Les statistiques d'utilisation de l'automne devraient être déterminantes pour son avenir. Enfin voilà, faisons le voeu que la nouvelle version de recherche soit une retraite et non un retrait.

Marc Duval
Service de recherche documentaire DSI
1280-4, bd de Montarville.
Longueuil. Québec. Canada
J4B 8B4
marcduval@dsi-info.ca
http://www.dsi-info.ca