Chronique

de la recherche sur le web

par Marc Duval 4 novembre 2002, vol. 2, n° 9

Copernic Agent Professional, l'agent semi-professionnel

En 1996, la société québécoise Agent Technologies Corporation entrait sur la scène de l'Internet. En 1997, elle lance son métamoteur Copernic. Sa mission: explorer le web à travers une dizaine des principaux sites de recherche. En moins d'un mois, la société lance une nouvelle version qui intègre son métachercheur à Internet Explorer de Microsoft. Les principales caractéristiques du logiciel Copernic sont: agrégation du résultat de recherche des principaux sites de recherche, élimination des doublons et les liens morts, sauvegarde des requêtes et des résultats de recherche, classement des références par pertinence et par divers critères de tri. Dès son arrivée, il reçoit éloges et prix.

L'année suivante, Agent Technologies étend la recherche sur 33 sites regroupés par domaines: actualités, affaires, courriels, divertissement, gouvernement, informatique, sport et voyage. Quelques mois plus tard, des améliorations sont apportées à l'interface telles qu'un assistant de recherche et le partage amélioré des résultats de recherche en différents formats de fichiers. L'algorithme est aussi bonifié. Le nombre de sites de recherche est augmenté à 125.

En 1999, Agent technologies Corporation devient Copernic Technologies inc. La même année, une nouvelle version est lancée. L'interface est à nouveau remodelé. Les termes de recherche reçoivent l'option de surlignage. L'affichage «zoom» est ajouté. Elle intègre plus de 130 sites de recherche dont ceux pour les MP3.

En octobre, Copernic Technologies inc. annonce la version 2000 déclinée en Copernic 2000, Copernic 2000 plus et Copernic 2000 pro. Une autre interface est dessinée. L'intégration avec Internet Explorer est enrichie d'une recherche à partir d'un mot cliqué sur une page web. La version pro est gratifiée de plus de 250 sites de recherche regroupés sous 32 domaines. Aussi, la recherche s'étend au web couvrant d'autres langues et d'autres pays.

Au printemps de 2000, Copernic 2000 se voit ajouter la recherche en «langage naturel» pour la version en anglais et une «garde-robes» d'apparence. Les versions en espagnol, allemand, portugais et italien s'alignent à côté de celles en français, en anglais et en japonais. D'autres produits sont lancés dont Copernic Summarizer.

En 2001, c'est l'explosion dans l'accessibilité des sites de recherche avec plus de 1000 sources. Aperçu de pages web, traduction, recherche automatisée appelée abusivement «veille», insertion de Copernic Summarizer sont autant de nouveautés.

En 2002, la société Copernic Technologies inc. est rebaptisée Copernic. Le 8 octobre 2002, Copernic 2001 devient Copernic Agent Professional, la version 6. Cette mouture reçoit plusieurs nouveautés quant à l'affichage des références: intégration de Summarizer avec l'identification des concepts, groupement et affichage des pages selon plusieurs critères dont par domaines, par langues et par dates. Le gestionnaire de recherches automatisées est modifié: il regroupe toutes les

recherches en une seule heure choisie par le chercheur alors que le précédent gestionnaire avait un horaire variable. Il s'enrichit de la surveillance de page appelée «veille de page» et de l'envoi de la recherche par fichier joint.

Le module de recherche est revisité: l'analyse comprend la vérification et la suppression des liens identiques et invalides, la détection de pages identiques et de la langue de la page, repérage de la taille et de la date de modification des pages et extraction des concepts.

Copernic a reçu des améliorations indéniables. Mais qu'en est-il réellement? Durant l'évaluation d'un mois, 127 recherches automatisées ont été menées, composées de 99 surveillances de pages et 28 recherches. Le fureteur utilisé a été en grande partie Netscape 7 qui est, à mon avis, supérieur à Internet Explorer pour la recherche. Internet Explorer a été utilisé pour vérifier l'intégration de Copernic à ce fureteur. Neuf faiblesses et sept forces ont été relevées lors de ce mois.

Les neuf faiblesses sont:

1. La limitation des langages de recherche.
2. L'absence de recherche par champs.
3. La gestion déficiente des pages identiques (doublons).
4. La déficience dans l'affinage du résultat.
5. Une interface encombrée.
6. La restriction des pages à surveiller.
7. La déficience dans l'envoi des rapports de recherche.
8. L'intégration dans Internet Explorer
9. L'inadéquation entre le site de recherche et le métachercheur

Les sept forces sont:

1. La surveillance automatique de changements dans des pages Web.
2. Le résumé des pages Web trouvées.
3. La création des catégories personnelles.
4. La recherche dans le résultat de recherche.
5. La création de sous-dossiers.
6. L'envoi des rapports de recherche par fichier joint.
7. Annotation des recherches

Les faiblesses de Copernic Agent Professional

1. La limitation du langage de recherche

La principale faiblesse de Copernic est sa limitation du langage de recherche. Copernic n'exploite pas ce qui fait la richesse des langages de recherche. Son langage se résume aux opérateurs booléens de base: ET, OU, expression. On remarquera qu'il n'a pas l'exclusion ni d'opérateur d'adjacence. Dans ces conditions, nous ne pouvons pas faire des recherches spécifiques et se rendre dans les recoins d'une banque de données. Les stratégies de recherche sont limitées aux algorithmes des automates de recherche. Aucune recherche en profondeur ne peut être faite. L'utilisation du langage est réservée au résultat de la recherche dans lequel l'exclusion est présente.

2. L'absence de recherche par champs

La recherche par champs est une fonction importante pour des recherches au-delà du bon vouloir de l'automate. En se privant de cette fonction, Copernic ne permet pas de rechercher uniquement par titres, par sites, par domaines, par mots dans l'URL ni par hyperliens ou par hypertexte, ni par fichiers multimédias intégrés à une page web et encore moins par formats pour les «pdf».

L'absence de recherche par champs se fait aussi sentir dans la «recherche avancée».

3. Déficience de la gestion des pages identiques

Copernic met dans ses atouts la suppression des pages identiques. Il est vrai qu'il fait la suppression mais seulement avec sa recherche dans les automates de recherche traditionnels. Il a failli dans sa prétention dans la recherche des manchettes. Copernic n'est pas en mesure de repérer et d'éliminer les pages qui ont des différences mineures. C'est le cas des fils de presse repris par plusieurs journaux. Un autre cas est soulevé par Google News et AlltheWeb qui ajoutent le temps d'affichage dans ses références. Les articles reçoivent un temps d'affichage modifié tout au long de la journée avec la mention «4 hours ago», «6 hours ago», etc. Dans le résultat de recherche, Copernic trouve le même article et l'affiche en double ou en triple alors qu'il est unique dans Google News. Copernic semble ajouter des références qu'il juge nouvelles à sa liste au lieu de refaire une nouvelle liste d'où la duplication.

4. Déficience dans l'affinage du résultat

Copernic a «amélioré» l'affinage du résultat. L'analyse comprend 3 niveaux: minimale (vérifie et élimine les liens identiques), intermédiaire (vérifie et élimine les liens identiques ou invalides, repère les tailles et les dates de modifications des pages web) et optimale (vérifie et élimine les liens identiques ou invalides, repère les tailles et les dates de modifications des pages web, détecte les langues des pages et les pages identiques). L'analyse optimale comprend aussi trois fonctions: «conserver seulement les résultats contenant les mots clés, extraire les concepts clés des pages web et sauvegarder les pages sur disque pour consultation hors ligne». Parmi ces fonctions, seule celle de conserver les résultats contenant les «mots clés» est inadéquate. Le résultat de cette opération bloque la consultation des références éliminées. Ces références-ci aident le chercheur à affiner sa stratégie. Pour l'affinage du résultat, il ne doit pas utiliser cette fonction et faire son affinage a posteriori. Heureusement, Copernic a ajouté la recherche dans le résultat. Malheureusement, il n'exploite pas la recherche dans les concepts qui ne contiennent pas toujours les mots de la requête. Il ne fait pas non plus de recherche par champs.

5. Interface encombrée

L'interface comprend trois barres d'outils, une barre de commandes et trois sections. Lorsque toutes les barres sont affichées, l'interface devient encombrée. Les fenêtres peuvent être dimensionnées mais elles n'ont pas de fermeture automatique. Il faut ouvrir la fonction d'affichage pour activer une fenêtre.

Pour enlever cette impression d'encombrement, les barres d'outils sont amovibles et flottantes avec la souris. Il n'en demeure pas moins que l'écran est surchargé et peu maniable.

6. Restriction des pages à surveiller

La surveillance automatisée des pages est appelée «veille de page». Cette fonction est une innovation pour Copernic. Cependant, elle est restreinte aux pages statiques. Pendant un mois, plusieurs pages ont été mises sous surveillance. Parmi celles-ci, trois d'entre elles étaient des requêtes dans Yahoo!, HotBot et Google. La surveillance des requêtes est un excellent moyen pour voir la dynamique des références et leurs environnements. Les pages de Yahoo! et de HotBot ont été très bien rendues contrairement à Google. Copernic n'a pas réussi ce test ni celui de la surveillance de la page d'accueil de Branchez-vous!

Une autre déficience de la surveillance est l'impossibilité de limiter dans une page ce qui doit être surveillé. Plusieurs pages ont seulement la date du jour qui change. Il en résulte qu'une page modifiée de façon mineure viendra se joindre aux autres dans la corbeille de courrier. Pour pallier ce désagrément, il aurait pu y avoir un tableau des modifications comme le fait d'autres logiciels de surveillance de pages comme Morning Paper.

Une autre absence est la page de référence. La fonction de surveillance de page signale ce qui est ajouté, mais pas ce qui est retiré. La page de référence permet la comparaison, à défaut que cette comparaison soit faite par Copernic.

7. Déficience dans l'envoi des rapports

Le test d'envoi des rapports a été mené avec Netscape 7. Copernic a ajouté l'envoi des rapports en fichier joint à son envoi dans le corps du message. Cette initiative est particulièrement appréciée. Cependant, elle aurait besoin d'être peaufinée avec deux améliorations. La première est le nom des fichiers, la seconde est l'option d'envoi.

Tous les rapports joints portent le même nom. Il faut les renommer pour savoir au moins de quelle recherche il s'agit si les fichiers sont sauvegardés. Ceci est un moindre mal pour quelques fichiers mais rendu à plus de 10 par jour, si ce n'est à plus de cent, il y a perte de temps à les renommer.

L'envoi des fichiers a une alternative: ou tous les fichiers sont envoyés dans le corps du message ou ils sont joints. Les pages surveillées ont besoin d'être sauvegardées, le fichier joint aide à bien faire ce travail. Les résultats de recherche n'ont pas ce besoin. Ils sont consultés dans le corps du message et les documents hyperliés peuvent être sauvegardés. Certaines de ces recherches doivent être expédiées à un tiers. Copernic manque de souplesse à ce chapitre.

Enfin, la préparation des requêtes avait inclus les concepts dans l'affichage. Ces derniers étaient absents dans l'affichage du résultat posté alors que le tri avait été fonctionnel.

8. Intégration à Internet Explorer

Copernic est déjà intégré à Internet Explorer. Avec la version 6, de nouvelles caractéristiques apparaissent: surveillance de page, historique, raccourci pour remonter de niveau dans un site en plus du surlignage, de la fenêtre de recherche et des flèches de navigation dans la page et dans le résultat de recherche. La barre d'outils de Copernic est utile, mais elle n'est pas gigogne, c'est son principal défaut. S'il y a un manque d'espace, des fonctions disparaissent. Il faut la mettre pleine largeur pour les avoir toutes. C'est pourquoi, la barre d'outils de Copernic cohabite mal avec d'autres barres comme celles de Yahoo! ou de Google. Il aurait été plus sage que cette barre soit intégrée au volet de recherche.

9. Inadéquation des résultats de recherche

Une comparaison a été faite entre les réponses du résultat de recherche de quelques sites web et ceux de Copernic. Cette comparaison a donné des différences. Nous savons que les sites nationaux sont adaptés au pays. Copernic ne semble pas y être sensible et on y trouve des incohérences.

La recherche sur Altavista France ne donne pas le même résultat que sur AltaVista France/Copernic parce que Copernic ne cherche pas dans la banque de données françaises mais dans le web.

La recherche avec HotBot-France/Copernic est surprenante. En fait, elle se fait sur HotBot. Ce qui donne des résultats différents quant on sait qu'HotBot utilise Inktomi et HotBot France, Fast Search.

Avec AOL France, il se connecte sur AOL.fr/Exalead, mais pas sur AOL.fr/Google. Avec Voila, Copernic l'a divisé en Voila (Tout le web francophone) et Voila Sélection (le répertoire). Il n'a pas fait cette distinction avec Lycos France qui a un répertoire spécifique. Pour le Canada, il utilise Looksmart et non Looksmart Canada. Le cas de Nomade est particulier. Dans deux cas, les requêtes - «voiture» et «immobilier» - avaient des réponses différentes entre la liste du répertoire Nomade et celle de Nomade/Copernic, dans un second - Harry Potter, le répertoire recensait 40 sites et Copernic en avait récupéré 5. Est-ce symptomatique ou anecdotique?

Que Copernic ait développé un algorithme tenant compte de la langue et des domaines géographiques pour les automates internationaux, soit. Mais qu'il ne tienne pas toujours compte des banques données adaptées à chaque pays, c'est tout autre.

Les forces de Copernic Agent professional

Copernic Agent Professional a 6 forces: surveillance automatique de changements dans des pages web, création de résumés des pages Web trouvées, possibilité de créer des catégories personnelles, recherche dans le résultat, création de sous-dossiers, envoi des rapports de recherche par fichier joint.

1. Surveillance automatique de changements dans des pages Web

La surveillance automatique de changements dans des pages web est une réussite «imparfaite», une réussite tout de même. Elle est efficace pour les pages statiques qui sont majoritaires. Les changements sont surlignés et la majorité des sauvegardes des pages sont conformes aux originales.

2. Création de résumés des pages Web trouvées

Copernic a intégré dans son CAP son Summarizer. Cette intégration se révèle un succès avec l'ajout des concepts dans les références. Elle permet de mieux les apprécier.

3. Possibilité de créer des catégories personnelles

La création de catégories personnelles de sites de recherche n'est pas à dédaigner. Il est possible de regrouper dans une catégorie plusieurs domaines et même plusieurs sites de recherche. La souplesse de cette fonction est remarquable. L'ajout de sites de recherche extérieurs à Copernic lui donnerait une grande flexibilité.

4. Recherche dans le résultat

En comparaison avec la version précédente, la recherche dans le résultat est une bonne amélioration bien que ce ne soit pas suffisant. Le formulaire de recherche avancé demeure fruste. En fait d'avancé, il est plutôt de base en regard des formulaires d'un AltaVista et d'un HotBot. Toutefois, il constitue un tremplin pour une version 7. L'effort est louable et digne de mention.

5. Création de sous-dossiers

La création de sous-dossiers est ce petit quelque chose qui plaît. Rien de révolutionnaire, elle est tout à fait pratique.

6. Envoi des rapports de recherche par fichier joint

Copernic 2001 avait déjà l'envoi des rapports par courriel dans le corps du message. La version 6 ajoute le fichier joint. La recherche envoyée de cette façon garde les caractéristiques de l'original notamment pour le tri. En ajoutant le tri par titre, il est facile de repérer les pages identiques passées au travers des mailles du filtre.

7. L'annotation des références

Il arrive à l'occasion de mettre une note sur une référence. Copernic a très bien intégré cette fonction. Celle-ci se trouve dans la fenêtre des paramètres de la référence. La note créée se place sur le côté gauche de la référence sous la forme d'un «post-it» jaune.

Petit hic, l'annotation de la recherche est loin d'être évidente comme le «post-it» de la référence. Elle est cachée dans les «propriétés» de la requête et il n'y a aucune icône pour nous la rappeler.

Conclusion

Au cours des années, Copernic a choisi le chemin du développement horizontal. Il a augmenté le nombre de ses sites de recherche et il les a regroupés. Il a intégré son logiciel Summarizer lui donnant une plus-value. Il a ajouté la surveillance de pages web et l'envoi de fichiers joints. Il a amélioré sa recherche interne avec l'ajout d'un formulaire. Toutes ces améliorations nous ont fait oublier qu'il a perdu son traducteur.

En fait, bien que Copernic soit un logiciel reconnu et primé, il a échoué dans sa mission en prenant le chemin de la croissance horizontale au lieu de développer le noyau de son logiciel - la recherche à partir des grands sites de recherche comme AltaVista, Google et Yahoo. Il a laissé de côté la recherche par champs et le développement d'un langage de recherche structuré. Son développement horizontal est méritoire mais il ne peut cacher ses défauts.

Copernic Agent Professional est-il professionnel? Non, il n'est pas professionnel pour un spécialiste de la recherche ou un chercheur exigeant. Il ne permet pas d'exploiter toutes les possibilités d'un automate de recherche comme rechercher les documents «pdf».

Copernic Agent Professional est-il professionnel? Oui, il l'est pour tous les autres professionnels qui ne font pas métier de chercheur. Sa version de base satisfera l'internaute moyen, ses deux versions payantes l'honnête professionnel. Il continuera de recevoir des lauriers des magazines populaires. Mais tel qu'il est, il ne remplacera ni un AlltheWeb, ni un Google avec leur formulaire détaillé. Il n'aura de «professional» que le nom en attendant qu'il arrive à maturité... pour un professionnel de la recherche.

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