La recherche d'information sur le web

Jean-Pierre Lardy

La recherche d'information sur le web

Par Jean-Pierre LARDY
UNIVERSITE CLAUDE BERNARD LYON I - URFIST
courriel : lardy@univ-lyon1.fr
serveur : http://urfist.univ-lyon1.fr et http://www.adbs.fr/adbs/sitespro/lardy/risi.htm

Sous une apparente simplicité, la recherche d'information sur le web recèle de nombreuses difficultés qui conduisent le plus souvent à une sous-utilisation de ces richesses.
L'histoire d'Internet et du web permet de mieux comprendre les problèmes rencontrés. Le réseau a été développé en milieu universitaire comme outil de communication et de partage non marchand de l'information circulant entre les chercheurs. Au milieu des années 90 le réseau s'est ouvert à toute la société : d'outil de spécialiste il s'est retrouvé en quelques années entre les mains du grand public. Le web, grâce à sa simplicité d'édition et de consultation, a rendu l'Internet convivial et accessible à tous. Cependant malgré cet engouement qui a conduit à un développement énorme de l'offre, rien n'a été fait pour garantir qu'un document publié sera retrouvé, visible et lisible. Ainsi cet immense réservoir d'information n'a rien à voir avec une bibliothèque malgré les analogies que certains ont avancées. L'information y est hétérogène aussi bien dans son contenu que dans sa forme, peu contrôlée, volumineuse, dispersée, sans classement ni archivage. Que diraient les clients de grandes surfaces si les milliers de produits du magazin étaient offerts dans une telle anarchie ?

Depuis une dizaine d'années des outils de recherche d'information, nombreux et variés, destinés au grand public, ont été développés pour explorer le web. Mais comment les utiliser au mieux, leur accorder notre confiance et ne pas gaspiller notre temps, noyé sous les réponses ?

1- Comment utiliser au mieux les outils de recherche ?

Des outils de recherche de nature différente

Deux grandes approches sont utilisées pour la production d'outils de recherche :

- L'approche manuelle qui mobilise des documentalistes (surfers) pour signaler dans des **annuaires** ou **guides et listes thématiques** des sites. L'intérêt principal est la sélection des sites et leur classement par thèmes au détriment de l'exhaustivité,
- L'approche automatique des **moteurs de recherche** ou **index de pages** qui permet de signaler beaucoup plus d'information et garantit une bonne mise à jour au détriment d'un quelconque classement ou indexation humaine.

En fait ces deux approches se complètent et sont de plus en plus associées :

- tel **annuaire** fait appel à un *moteur* quand il n'a pas de réponse par exemple **Nomade** ou **Yahoo** et *Google*,
- tel *moteur* complète ses réponses par celles d'un **annuaire** par exemple *Google* et **l'Open Directory**.

Une syntaxe (presque) commune

Comme souvent le formalisme des requêtes s'appuie sur des variantes syntaxiques d'un outil à l'autre. Cependant nous pouvons utiliser sans risque d'erreur important la syntaxe suivante dans tous les formulaires de recherche simple :

Opération		
Rechercher une expression		
Imposer un terme		
Exclure un terme		
Tronquer un terme		

Attention, les signes + et – doivent être collés à gauche du terme concerné. Cette écriture permet d'effectuer des recherches simples, suffisantes dans la plupart des cas.

Un peu de méthode

Avant toute chose, il faut réfléchir au vocabulaire : faire la liste des synonymes, repérer les ambiguïtés. Les variations lexicales sont nombreuses et les outils de recherche ne sont d'aucune aide. A part quelques annuaires spécialisés qui s'appuient sur des langages documentaires, nous devons utiliser le langage naturel et ses imperfections (NB : les fautes d'orthographe sont très fréquentes dans les pages html).

Ainsi il est préférable de commencer une recherche dans un annuaire : l'information sélectionnée y est catégorisée ce qui permet de surmonter les problèmes dûs aux ambiguités sémantiques de tout langage humain (synonymie, polysémie). Pour cela nous croiserons les avantages du classement et la recherche par index. Par exemple pour une question sur les conséquences de la maladie de la vache chez l'homme, nous sélectionnerons le thème principal **SANTE** et lancerons la recherche dans l'index en cochant **dans cette catégorie**. Le nombre restreint de sites proposés est aussi un avantage.

Les moteurs de recherche seront utilisés pour des questions précises. Le problème principal est alors le nombre toujours important de pages trouvées.

Apprenez à déchiffrer les résultats

L'organisation en catégories des annuaires permet de repérer assez vite les réponses pertinentes. Il n'en est pas de même avec les moteurs de recherche. De nombreuses méthodes automatiques de tri des résultats ont été développées ces dernières années : tri par pertinence pour AltaVista, Fast ou Voila, tri par popularité pour Google ou classement dynamique en catégories pour NorthernLight. En pratique les utilisateurs parcourent seulement les 10 à 20 premiers résultats. Il est donc intéressant :

- d'une part de reformuler les questions pour un même outil,
- d'autre part d'utiliser plusieurs moteurs de recherche.

Les méthodes de tri ne tiennent compte en aucun cas du sens des textes et leur qualité est très variable d'une question à l'autre.

2- Peut-on leur accorder notre confiance ?

Une réponse facile serait : « mais que faire d'autre sinon parcourir les liens hypertextes ». Il est vrai que ces outils présentent de nombreuses lacunes et fournissent énormément de bruit en réponse. Cependant l'expérience montre qu'il est possible en réflechissant bien et en évitant de se laisser disperser de trouver ce que l'on cherchait. Les outils « grand public » sont donc utiles mais l'avenir est aux outils de recherche spécialisés. Nous citerons comme exemple l'annuaire des sites médicaux francophones CISMEF ou le moteur des sites éducatifs français SPINOU. A partir de corpus plus homogènes et moins volumineux, ils donnent des réponses beaucoup plus pertinentes.

3- Ne pas gaspiller son temps ?

Le reproche principal fait à la recherche d'information sur le web est le temps passé. Il est le fait principalement des professionnels de la documentation habitués par la force des choses (et des coûts) à limiter au maximum la durée des sessions de recherche. En fait la perte de temps provient plus des invites « à la promenade » des pages web que des outils de recherche eux-mêmes. C'est ce qui fait tout le charme de ce média.

4- Conclusion

Tel qu'il est, le web ne sera jamais un espace de recherche performant. Cependant malgré ses imperfections dûes à un développement mal maîtrisé, il reste une formidable aventure ouverte au plus grand nombre.

Bibliographie

- **CERISE** - Conseils aux étudiants pour une recherche d'information spécialisée efficace
  http://web.ccr.jussieu.fr/urfist/cerise/
- **RIsI** - Recherche d'information sur l'Internet
  http://www.adbs.fr/adbs/sitespro/lardy/risi.htm
- **InfoSphère** - Apprendre à faire une recherche d'information efficace
  http://www.bibliotheques.uqam.ca/InfoSphere/
- **Détective de l'Internet** – Un cours pour évaluer la qualité des ressources de l'Internet
  http://www.desire.org/detective/detective-fr.html

Quelques outils de recherche

Annuaires		*Moteurs de recherche*	
CISMEF	www.cismef.org	AltaVista FR	fr.altavista.com
LookSmart Fr	www.looksmart.fr	FAST	www.alltheweb.com
Nomade	www.nomade.fr	Google	www.google.fr
Open Directory	dmoz.ch	Lokace	www.lokace.com
Yahoo Fr	fr.yahoo.com	NorthernLight	www.northernlight.com
Yoodle	www.yoodle.ch/fr/default.asp	Spinou	www.cndp.fr/spinoo/
		Voila	www.voila.fr