Les moteurs de recherche

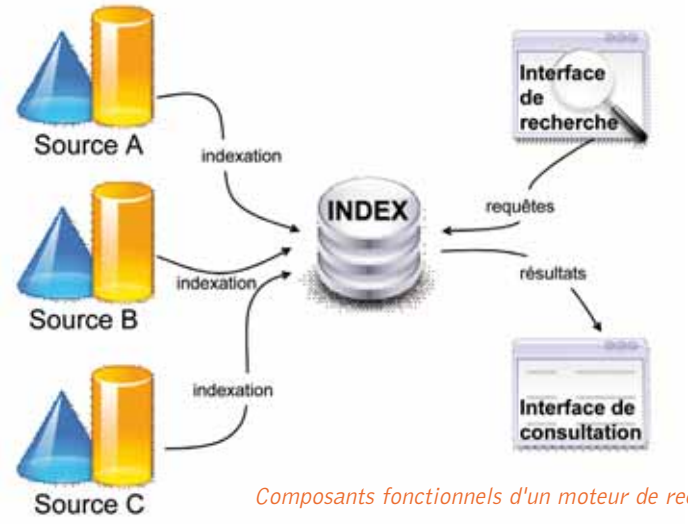
Traditionnellement, la recherche dans les catalogues de bibliothèques est conçue sur le principe de l'interrogation d'une base de données : on indique, pour les différents champs (titre, auteur, sujet, etc.), les chaînes de caractères que l'on souhaite retrouver ; on obtient en réponse tous les résultats contenant ces chaînes de caractères, et uniquement ceux-ci. Les usages du web et le succès inconditionnel des moteurs de recherche nous conduisent aujourd'hui à repenser notre approche de la recherche d'information au sein d'un catalogue. La recherche par mots-clés règne en maître, et c'est l'outil de recherche qui remplit la délicate mission de faire le tri dans les innombrables résultats trouvés. Faut-il pour autant remettre en cause les principes de structuration des données qui sont ceux des catalogues de bibliothèques ? Et comment s'approprier ce nouvel outil qu'est le moteur de recherche ?

Qu'est-ce qu'un moteur de recherche ?

Aujourd'hui, quand on utilise le terme de moteur de recherche, on peut entendre deux choses différentes : l'outil et le service. Ainsi, quand on dit que Google est un moteur de recherche, c'est au service qu'on se réfère : Google fournit un service qui offre la possibilité d'effectuer des recherches plein texte, par mots-clés, sur un vaste ensemble de pages web. Bien sûr, derrière le service, il y a l'outil : une infrastructure logicielle qui permet de mener à bien cette tâche. Ce sont les moteurs de recherche en tant qu'outils qui nous intéressent ici : imaginons qu'au lieu d'interroger des pages web, nous interrogions des notices de catalogue...

Un moteur de recherche comporte plusieurs composants :

- un ou plusieurs connecteurs, qui collecte(nt) les données auprès d'une ou plusieurs sources ;
- l'indexeur, qui analyse le texte et constitue l'index ;
- la ou les interface(s) de recherche, qui permet(tent) de formuler des requêtes ;
- la ou les interface(s) de consultation, qui permet(tent) d'afficher les résultats.



Composants fonctionnels d'un moteur de recherche

Au contraire d'une base de données qui exploite directement la structuration des données par champs, le moteur de recherche passe par une étape intermédiaire, la constitution d'un index, dans lequel elle stocke pour chaque terme indexé un certain nombre d'informations.

L'index a pour rôle de stocker par anticipation les mots qui feront l'objet des recherches des utilisateurs, suivant une structure déterminée : par exemple, un index inversé permet de passer d'un document qui contient des mots à une liste de mots avec l'indication des documents qui les contiennent.

L'indexation est également le moment où le moteur de recherche va procéder à des analyses pour préparer le calcul de la pertinence, afin de classer les résultats d'une recherche. Pour cela, il peut utiliser différentes méthodes de calcul, ou algorithmes. Ainsi, lors de l'indexation, le moteur de recherche pondère les différents champs : il attribue une valeur plus ou moins élevée à un mot en fonction du champ dans lequel il le trouve.

Autres exemples : pour certains moteurs, un mot rare dans un texte aura plus de valeur qu'un mot qui revient très souvent, un mot trouvé dans un champ très court aura plus de valeur que le même mot dans un champ très long.

Une fois l'index constitué, on peut exploiter le résultat par l'intermédiaire de l'interface de recherche, en formulant notamment des recherches par mots-clés qui permettront de retrouver les entrées d'index associées au terme demandé.

Exploiter la structuration des données

Les moteurs de recherche ont été à l'origine imaginés pour permettre d'interroger des documents en plein-texte, non structurés : en particulier des textes qui constituent les pages web. Les moteurs sont conçus pour extraire l'information de ces textes et, grâce à des moyens statistiques, identifier ceux qui ont le plus de chances de répondre efficacement à une question posée. Cela signifie-t-il pour autant que la structuration de l'information n'a plus d'importance ? Pour un catalogue de bibliothèque, c'est loin d'être le cas.

Dans un récent rapport intitulé *Online catalogs : What librarians and users want*[1], OCLC rappelle que même si la recherche par mots-clés, ou recherche simple, est l'instrument préféré des usagers, ils jugent également utile de pouvoir effectuer des recherches avancées, ou recherches par champs, et de pouvoir affiner leurs résultats en utilisant les facettes[2]. Or, recherche avancée et navigation par facettes reposent sur la structuration fine des données telle que nous la connaissons dans les formats MARC. La Bibliothèque nationale d'Australie a également montré dans un article daté de 2007[3] que l'on pouvait exploiter la structuration des données en MARC pour paramétrer le calcul de pertinence.

À la BNF, deux expériences ont permis de mettre ces principes en application : le développement de la nouvelle interface de Gallica[5] entre 2007 et 2009, et la mise en œuvre de la recherche dite par mots dans le catalogue général[6]. Ces deux outils utilisent le logiciel libre Lucene pour indexer des données en XML, qui sont fournies à partir d'une conversion des notices en Intermarc du catalogue. L'objectif est de proposer aux usagers une recherche plus conforme aux habitudes

La navigation par facettes dans Gallica

du web, tout en exploitant la structuration des notices bibliographiques pour classer les résultats par pertinence, et offrir des fonctionnalités comme la recherche avancée ou, dans Gallica, la navigation par facettes.

La souplesse offerte par l'outil est extrêmement précieuse pour s'orienter dans la masse des ressources disponibles, ou pour rechercher des informations particulières. Ainsi, par exemple, on peut aujourd'hui afficher dans Gallica tous les documents publiés entre deux dates, sans avoir besoin de préciser les mots de l'auteur ou du titre, puis utiliser les facettes pour filtrer dans ce corpus des types de documents, des thèmes (ceux-ci étant définis par un plan de classement basé sur la classification Dewey) ou des formats (possibilité de limiter la recherche aux ouvrages pour lesquels le mode texte est disponible). Dans le catalogue, certaines informations qui étaient auparavant cachées car non indexées par la base de données sont désormais exploitables en recherche, comme les équivalents anglais des autorités RAMEAU ou les zones de notes. L'outil est plus à même de traiter la masse, et permet d'éviter les effets de seuil (requêtes ramenant « trop » de réponses).

Perspectives

La combinaison entre une interface de recherche simple et un système de navigation par facettes est en train de devenir **l'interface naturelle des OPAC dits « nouvelle génération »[6].** On voit donc bien se définir une convergence entre les interfaces de catalogues de bibliothèques et les usages des moteurs de recherche sur le web.

Les enjeux qui restent associés à ce type d'outils résident maintenant dans le travail sur l'amélioration de la pertinence des classements de résultats, et sur l'ajout de fonctionnalités dites sémantiques qui permettront d'automatiser la structuration pour certains contenus. Dans le cadre du projet européen TEL Plus, une expérimentation va porter sur l'utilisation de technologies sémantiques, comme la reconnaissance d'entités nommées, pour relier des textes non structurés avec des référentiels, comparables à nos notices d'autorité. La disponibilité croissante des référentiels de bibliothèques sous une forme compatible avec les technologies de la famille RDF comme **SKOS[7]** devrait favoriser la montée en puissance de ce type d'outils.

Dans ce contexte, il est important d'observer que les classements de pertinence et les traitements des données (type navigation par facettes) tendent à révéler les insuffisances de la qualité des données, ou leur hétérogénéité naturelle, et nécessitent de fréquents ajustements. Le moteur de recherche n'est pas un outil clef en main ; il faut imaginer que derrière le succès du calcul de pertinence de Google, il y a des armées d'ingénieurs qui scrutent au quotidien le traitement des données et ajustent les algorithmes à l'évolution du web. De la même façon, les catalogues nouvelle génération ne donneront des résultats satisfaisants pour les usagers que dans la mesure où les bibliothécaires continueront à se mobiliser pour veiller à la cohérence des données, à leur structuration fine, et à l'intelligence des outils qui les exploitent.

E. Bermès

emmanuelle.bermes@bnf.fr

BNF www.bnf.fr
Direction des services et des réseaux
Département de l'information bibliographique et numérique
Emmanuelle Bermès ☎ 01 53 79 42 40
Quai François-Mauriac
75706 PARIS CEDEX 13

1 Publié en avril 2009 : http://www.oclc.org/reports/onlinecatalogs/default.htm

2 Sur la navigation par facettes, voir Marc Maisonneuve et Cécile Touitou, « Une nouvelle famille d'Opac. Navigation à facettes et nuages de mots » BBF 2007, t. 52, n° 6 : http://bbf.enssib.fr/consulter/bbf-2007-06-0012-002

3 « Relevance ranking of results from MARC-based catalogues : from guidelines to implementation exploiting structured metadata » par Alison Dellit et Tony Boston, Bibliothèque nationale d'Australie, février 2007 : www.nla.gov.au/nla/staffpaper/2006/documents/Boston_Dellit-relevance-ranking.pdf

4 http://gallica.bnf.fr

5 Cette recherche est disponible depuis avril 2009 : http://catalogue.bnf.fr

6 Pour n'en citer que quelques uns : celui de la bibliothèque nationale d'Australie ; Libris, le catalogue collectif des bibliothèques suédoises ; la Queens Library de New York ; WorldCat ; etc.

7 Récemment, la Library of Congress a publié son site Authorities & Vocabularies : http://id.loc.gov/authorities/. Une version expérimentale de Rameau est également disponible en SKOS : http://stitch.cs.vu.nl/rameau