Ouvrir les boîtes noires : un outil pédagogique pour une approche critique de la recherche d'information en ligne

Cyrille Suire
cyrille.suire@univ-lr.fr
Laboratoire Informatique, Image et Interaction (L3i), Université de La Rochelle, France

Axel Jean-Caurant
axel.jean-caurant@univ-lr.fr
Laboratoire Informatique, Image et Interaction (L3i), Université de La Rochelle, France

Charles Illouz
charles.illouz@univ-lr.fr
Centre de Recherche en Histoire Internationale et Atlantique (CRHIA), Université de La Rochelle
France

Introduction

Les rapports entre les technologies du numérique et les Sciences Humaines et Sociales (SHS) sont aujourd'hui profondément renouvelés par le développement des Humanités numériques. Les mutations en cours permettent en particulier de repenser les pratiques d'enseignement et d'éducation au numérique. Voir en particulier le très récent ouvrage rédigé de manière collaborative lors du Edcamp qui s'est tenu à Paris les 1er et 2 septembre 2016 (Bourgatte et al., 2016). Parmi ce vaste chantier, la problématique de la recherche et de l'accès à l'information semble primordiale. Elle représente en effet une part importante du travail quotidien des chercheurs et des étudiants en SHS et se trouve profondément bouleversée par le tournant numérique. Celui-ci se matérialise par l'explosion du nombre de ressources disponibles en ligne et par une plus grande hétérogénéité des documents et des moyens d'accès. Si la disponibilité immédiate de documents jadis inaccessibles, faute de moyens ou de temps, est un atout précieux pour la recherche actuelle en SHS, il n'en reste pas moins que les activités de recherche et d'accès à l'information requièrent des compétences pointues et une expérience significative pour être véritablement maîtrisées. Le numérique pose à cet égard des problèmes spécifiques qui doivent être pris en compte dans la formation des étudiants.

Des travaux récents ont montré que la numérisation des documents et leur mise à disposition en ligne étaient loin de permettre un accès universel et transparent au matériau de la recherche (Milligan, 2013). Des problématiques techniques, méthodologiques et cognitives limitent notre compréhension de l'accès à l'information (Cardon, 2013). Sur le plan technique, la distinction généralement opérée entre « information seeking », l'ensemble des processus et pratiques des utilisateurs pour répondre à un besoin d'information et « information retrieval », les méthodes et techniques informatiques qui permettent au système de répondre à ces besoins est fortement significative (Buchanan et al., 2005). Il existe un fossé entre les besoins des utilisateurs et les technologies utilisées pour y répondre. Il est ainsi souvent difficile de comprendre les relations de causes à effets entre les critères de recherche (inputs) saisis par l'utilisateur, et les résultats (outputs) fournis par le système. Ce fossé sémantique se décompose en de multiples problématiques. Sans faire ici une liste exhaustive, les exemples ne manquent pas. Les technologies de reconnaissance optique de caractères (OCR), par exemple, génèrent des erreurs qu'il est complexe d'identifier et de mesurer. L'indexation des documents, quant à elle, repose sur des catégories développées par d'autres, dont les utilisateurs n'ont bien souvent pas connaissance. Les paramètres des algorithmes de pertinence, de personnalisation et de classement des résultats sont inaccessibles alors même qu'ils régissent la manière dont sont générés et classés les résultats des recherches en ligne.

Ces transformations subies par les données (des inputs aux outputs), inconnues des utilisateurs, sont des boîtes noires. Les étudiants doivent avoir conscience de leurs effets et doivent pouvoir les intégrer à leur appareillage critique. L'outil que nous développons poursuit deux objectifs:

Dresser des ponts entre les opérations menées par les systèmes d'accès à l'information et l'idée que s'en font les utilisateurs. Notre outil a ainsi pour vocation de mettre en lumière et permettre d'expliquer l'ensemble des transformations que subissent les données et les documents dans le cadre de

l'interaction entre un utilisateur et un système de recherche d'information.

Mettre les étudiants en situation expérimentale de recherche d'information et permettre d'en visualiser les résultats individuels et collectifs. Ces résultats doivent ensuite pouvoir être réinvestis dans une discussion interrogeant la démarche et la méthode de recherche d'information dans un contexte numérique.

Environnement technologique et fonctionnalités

L'outil que nous développons est un moteur de recherche et une interface de bibliothèque numérique reposant sur le framework libre Hydra (Apache Solr, Fedora Commons et Blacklight), capable d'indexer des sources primaires comme secondaires. Il est conçu pour être utilisé lors de séances de cours dirigées par un enseignant où les étudiants doivent répondre à un besoin d'information défini par le formateur.

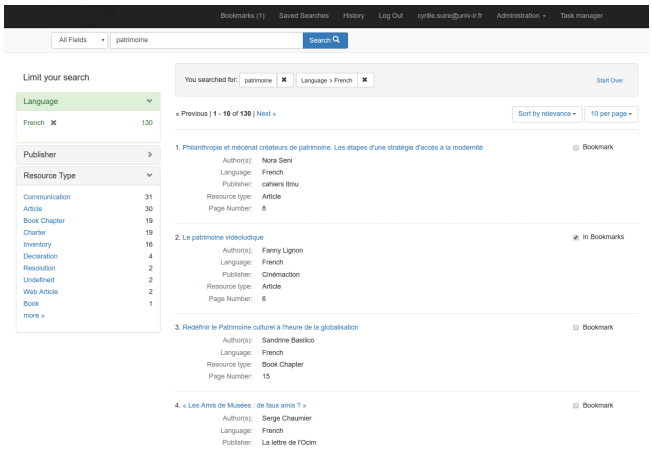

Figure 1 : Interface principale de l'outil de recherche

Sur la base du framework Hydra, nous ajoutons des composants qui documentent la requête effectuée par l'utilisateur et fournissent des informations supplémentaires dont :le résumé de la requête transmise au moteur de recherche (mots, expressions et filtres); les paramètres du moteur (métadonnées interrogées ou ignorées, variations linguistiques utilisées ou ignorées, etc.); les procédures de classement des résultats utilisées (critères de pertinence, pondérations utilisées, etc.).

Cette documentation supplémentaire permet d'afficher à l'utilisateur des compléments d'information pour chaque requête qu'il effectue dans le moteur de recherche. Il lui est alors possible de comprendre comment a été décomposée sa requête, quels paramètres extérieurs à son contrôle ont été appliqués et comment ont été calculés les résultats.

L'enseignant qui dirige une séance peut par ailleurs agir sur certains paramètres du moteur de recherche, pour modifier dynamiquement la manière dont celui-ci réagit aux inputs des utilisateurs. Il est par exemple possible de modifier le comportement de l'algorithme de pertinence, de changer la présentation des résultats ou encore d'activer et désactiver l'usage de certaines métadonnées. Ces options ont un impact important sur le comportement du moteur. Les étudiants peuvent immédiatement le mesurer et réfléchir à l'influence de ces paramètres cachés sur leur pratique de recherche.

Durant la séance, l'outil garde également une trace de l'activité globale de recherche de chaque utilisateur. Cette trace est élaborée grâce aux logs de l'application, que nous enrichissons de données liées au comportement de recherche des utilisateurs (Suire et al., 2016). Ces informations servent d'abord à générer des représentations personnelles et collectives des recherches effectuées durant la séance. Il est ainsi possible de débattre avec les étudiants des résultats obtenus, en se fondant sur les différentes approches et stratégies de recherche qu'ils ont développées. Par ailleurs, les données d'usage collectées par l'application sont utiles à l'évaluation de l'outil. Nous pouvons par exemple comparer le comportement des étudiants avant et après la formation, sur des tâches de recherche d'information de même nature et ainsi évaluer la pertinence pédagogique de nos développements. Nous complétons par ailleurs cette évaluation par des questionnaires et des entretiens qualitatifs, permettant de mesurer l'intérêt des étudiants pour la problématique, l'approche et les outils mis en place.

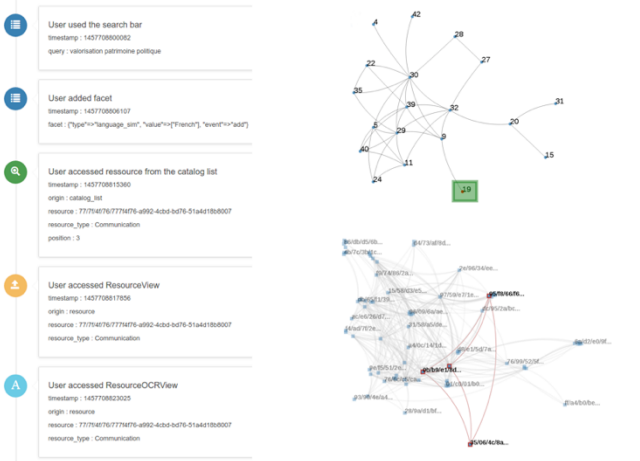

Figure 2 : Exemples de représentation visuelle: à gauche, un extrait du processus de recherche d'un utilisateur (ici identifié par le numéro 19) et à droite une représentation en réseau des documents consultés par cet utilisateur au regard de l'activité de l'ensemble du groupe.

Contexte expérimental et applications

Bien que l'outil soit encore en développement, nous menons des expériences afin d'évaluer son intérêt pédagogique et son fonctionnement. Nous expérimentons l'outil avec des groupes de 50 étudiants, en 2e année de cycle universitaire (Ces étudiants peuvent être considérés comme débutants, aussi bien en terme de connaissance du domaine qu'en terme de compétences en recherche d'information (Jenkins et al., 2003)), qui débutent leur spécialisation en Histoire. Lors d'une séance de 2 heures, ils bénéficient d'abord d'une courte présentation du fonctionnement de notre outil. Les étudiants ont ensuite 60 minutes pour élaborer une problématique de recherche à l'aide d'un corpus d'environs 300 documents hétérogènes (textes scientifiques ou institutionnels, documents iconographiques, etc.) relatifs à la thématique de la préservation du patrimoine. La séance se termine sur un échange de 40 minutes autour des indicateurs fournis par notre outil.

Lors de nos premières expériences, que nous présenterons plus en détail lors de notre communication, les thématiques de ces échanges ont été nombreuses. A titre d'exemple, les étudiants ont été surpris de leur tendance à se focaliser sur les premiers résultats calculés par le moteur de recherche, alors même que de nombreux documents pertinents se trouvaient dans les pages suivantes. Les capacités de représentations graphiques de notre outil ont également permis de mener une discussion constructive sur les différentes stratégies qu'il convient d'adopter face à ce type de besoin d'information. Les développements en cours permettront prochainement de simuler d'autres situations de recherche d'information courantes en SHS (Ellis, 1989; Savolainen, 2016). Ces premières expériences ont toutefois déjà témoigné de l'intérêt de mettre les étudiants en situation expérimentale, pour les engager dans une pensée critique de la recherche d'information dans un contexte numérique.

Bibliographie

Bourgatte, M., Ferloni, M. and Tessier, L. (2016). Quelles humanités numériques pour l'éducation ? - Éditions MkF http://www.editionsmkf.com/produit/edcamp-icp.

Buchanan, G., Cunningham, S. J., Blandford, A., Rimmer, J. and Warwick, C. (2005). "Information seeking by humanities scholars." International Conference on Theory and Practice of Digital Libraries. Springer Berlin Heidelberg, pp. 218-29.

Cardon, D. (2013). Présentation, Réseaux, 177, pp. 9-21.

Ellis, D. (1989). A behavioural approach to information retrieval system design. Journal of Documentation, 45(3): 171-212.

Jenkins, C., Corritore, C. L. and Wiedenbeck, S. (2003). Patterns of information seeking on the Web: A qualitative study of domain expertise and Web expertise. IT & Society, 1(3): 64-89.

Milligan, I. (2013). Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997-2010. The Canadian Historical Review, 94(4): 540-69.

Savolainen, R. (2016). Contributions to conceptual growth: The elaboration of Ellis's model for information-seeking behavior. Journal of the Association for Information Science and Technology, 68(3) : 594-608.

Suire, C., Jean-Caurant, A., Courboulay, V., Burie, J.-C. and Estraillier, P. (2016). "User Activity Characterization in a Cultural Heritage Digital Library System." Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries. (JCDL '16). New York, NY, USA: ACM, pp. 257-58.