Documentation et bibliothèques

DOCUMENTATION BIBLIOTHÈQUES

SYSTÈME DE RECOMMANDATION BASÉ SUR LES NOTICES BIBLIOGRAPHIQUES MARC 21 : ÉTUDE DE CAS À BIBLIOTHÈQUE ET ARCHIVES NATIONALES DU QUÉBEC (BANQ)

Outside of the Theme—Recommendations using the MARC 21 Bibliographic Records: A Case Study at Bibliothèque et Archives nationales du Québec (BanQ)

Pascal Laforce and Sylvie Ratté, Ph.D.

Volume 64, Number 2, April–June 2018

URI: https://id.erudit.org/iderudit/1059160ar
DOI: https://doi.org/10.7202/1059160ar

See table of contents

Publisher(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (print)
2291-8949 (digital)

Explore this journal

Cite this article

Laforce, P. & Ratté, S. (2018). SYSTÈME DE RECOMMANDATION BASÉ SUR LES NOTICES BIBLIOGRAPHIQUES MARC 21 : ÉTUDE DE CAS À BIBLIOTHÈQUE ET ARCHIVES NATIONALES DU QUÉBEC (BANQ). *Documentation et bibliothèques*, *64*(2), 40–50. https://doi.org/10.7202/1059160ar

Article abstract

This article suggests a system of recommendations based on the filtering of documents linked to MARC 21 bibliographic records and the borrowing history of the user as recorded in the integrated library system. The project was undertaken in 2017 at Bibliothèque et Archive nationales du Québec. The high level of precision supports the potential and the reliability of the system. The results demonstrate that the use of a system of recommendation adapted to public libraries improves the quality of service to users, with minimal impact on the operational systems already in place.

SYSTÈME DE RECOMMANDATION BASÉ SUR LES NOTICES BIBLIOGRAPHIQUES MARC 21 : ÉTUDE DE CAS À BIBLIOTHÈQUE ET ARCHIVES NATIONALES DU QUÉBEC (BANQ)

Pascal LAFORCE
Faculté des sciences
Université du Québec à Montréal
laforce.pascal.2@courrier.uqam.ca

Sylvie RATTÉ, Ph.D.
Département de génie logiciel et des TI
École de technologie supérieure
sylvie.ratte@etsmtl.ca

RÉSUMÉ | ABSTRACT

Le présent article propose un système de recommandations fondé sur le filtrage des documents associés à des notices bibliographiques MARC 21 et l'historique des emprunts de l'abonné provenant du système intégré de gestion de bibliothèque. L'expérience a été menée en 2017 à Bibliothèque et Archives nationales du Québec. Le taux de précision élevé obtenu lors de la période d'évaluation prouve le potentiel et la faisabilité du système. Les résultats obtenus montrent que l'utilisation d'un système de recommandation adapté aux bibliothèques publiques permet d'améliorer la qualité des services offerts aux usagers avec un minimum d'impact sur les systèmes opérationnels déjà en place.

Outside of the Theme—Recommendations using the MARC 21 Bibliographic Records: A Case Study at Bibliothèque et Archives nationales du Québec (BanQ)

This article suggests a system of recommendations based on the filtering of documents linked to MARC 21 bibliographic records and the borrowing history of the user as recorded in the integrated library system. The project was undertaken in 2017 at Bibliothèque et Archive nationales du Québec. The high level of precision supports the potential and the reliability of the system. The results demonstrate that the use of a system of recommendation adapted to public libraries improves the quality of service to users, with minimal impact on the operational systems already in place.

Introduction

La recherche de documents dans une bibliothèque est parfois une tâche difficile pour ses usagers en raison du nombre très important de documents disponibles en tous genres et formats. Les usagers ont des intérêts précis et un temps limité faisant en sorte qu'une proportion relativement faible de la collection les intéressera réellement. Les bibliothèques organisent souvent leurs espaces publics selon un système de classification similaire à Dewey, selon un niveau de subdivision plus ou moins profond, suivi d'un tri par ordre alphabétique des documents selon le nom de l'auteur principal. Ce type de classification tend à être trop complexe pour les usagers lorsque la subdivision est grande, car le nombre de classes à mémoriser devient trop important. Un système à classification unique n'est pas forcément idéal puisque l'usager s'intéresse peut-être plus à un même auteur, un même personnage, une même région géographique, une date de parution récente, un même sujet secondaire, etc. faisant en sorte que le système de classification physique ne répond pas à tous les besoins. La collection de la bibliothèque est en constant changement puisque de nouveaux exemplaires sont acquis alors que d'autres sont élagués ou constamment en circulation, ce qui n'aide pas les usagers à se familiariser avec la collection.

Les bibliothèques proposent généralement un catalogue d'accès public (CAP) en ligne qui permet à l'usager d'effectuer des recherches par mots-clés et facettes, et de naviguer d'une fiche descriptive à l'autre à l'aide de caractéristiques communes. Les CAP présentent toutefois l'inconvénient de nécessiter beaucoup de temps de la part de l'usager alors que celui-ci est de plus en plus pressé. Nous pensons que les CAP auraient avantage à proposer un mode de recherche alternatif en fonction de recommandations basées sur l'historique de lecture de l'usager qui nécessiterait un minimum d'effort, un peu comme le font déjà Netflix et Spotify pour les films et la musique.

De nombreux domaines d'affaires ont intégré des systèmes de recommandation à leur site Web transactionnel afin de faire découvrir à leurs usagers et clients des produits et services similaires à ceux qu'ils ont déjà consultés ou consommés, avec comme objectif de stimuler les ventes. Certains de ces systèmes ont été développés et perfectionnés depuis plus de 20 ans et leur utilité ne fait plus de doute dans le domaine du commerce électronique. Leur usage tend à se démocratiser dans d'autres secteurs d'activités depuis que ces techniques sont enseignées dans les écoles de commerce et de technologie. Les études du groupe de travail DELOS/NFS (Calan & Smeaton 2003) et de Nicholson (2006) ont montré que les bibliothèques tireraient avantage de ces techniques. Pourtant, seules quelques bibliothèques publiques ont vraiment réussi à intégrer ce type d'outils à ce jour.

Le présent article tentera de démontrer les avantages des systèmes de recommandation pour les bibliothèques et leurs usagers, ainsi que la faisabilité technique d'intégrer ces systèmes aux logiciels et aux technologies couramment utilisés dans le milieu documentaire à partir de l'expérience menée en 2017 à Bibliothèque et Archives nationales du Québec (BAnQ). Plus spécifiquement, l'article expliquera comment :

- constituer un dictionnaire de termes riches à partir des notices bibliographiques du Système intégré de gestion de bibliothèque (SIGB) ;
- bâtir le profil des abonnés de la bibliothèque à partir de leur historique de prêts ;
- calculer la similarité des documents avec le profil ;
- intégrer le système au SIGB ou à d'autres systèmes ;
- évaluer la qualité du système à partir de la rétroaction explicite des usagers.

L'article est structuré en quatre sections et débute par une revue de la littérature sur les diverses études et les enjeux d'application de ce type d'outil dans les bibliothèques publiques. Le matériel et la méthodologie employés à BAnQ sont ensuite décrits. Une discussion sur les résultats obtenus et les améliorations possibles au système conclut l'article.

Revue de la littérature

Peu d'articles traitant des systèmes de recommandation appliqués aux bibliothèques publiques ou même du forage de données en général dans ce domaine d'application ont été publiés jusqu'à ce jour. Cela s'explique principalement par le fait que le sujet nécessite la compréhension de plusieurs domaines de recherche très différents dont ceux de la science de l'information, de la recherche informationnelle, de l'apprentissage automatique et de l'interface humain-machine.

Vers la fin des années 1990, le groupe de travail international et multidisciplinaire DELOS/NSF a été formé afin de répondre aux défis techniques croissants que représentait l'arrivée des bibliothèques numériques. La question des systèmes de recommandation et de personnalisation a été abordée par un comité de 2001 à 2003 afin d'identifier les sujets de recherches prioritaires (Calan & Smeaton 2003). Il en est ressorti que les bibliothèques numériques deviennent de plus en plus hétérogènes et qu'il devient alors difficile de trouver un modèle unique qui s'applique bien à tous les genres et formats. On y mentionne que la recherche sur les systèmes de recommandation était jusque-là essentiellement basée sur la recherche de meilleurs critères de similarité. Le groupe souligne qu'il est aussi important de tenir compte des aspects de collaboration Web 2.0, de la confidentialité des données, de la personnalisation du système en fonction des intérêts changeants des usagers, des problèmes avec le démarrage à froid des systèmes de recommandation et des coûts prohibitifs pour la mise en place de ces systèmes pour la majorité des bibliothèques. Plusieurs des enjeux technologiques ont été résolus depuis, mais la question de l'intégration de l'aspect collaboratif et social demeure d'actualité, car on s'interroge encore sur la confidentialité des données recueillies sur les usagers et sur la qualité de l'information que ceux-ci partagent volontairement avec la bibliothèque.

Le chercheur en science d'information Scott Nicholson a écrit de nombreux articles dans les années 2000 sur le forage de données et son application possible pour les bibliothèques. L'essentiel de ses articles fait état de la pertinence et de l'intérêt qu'auraient les bibliothèques à se doter d'entrepôts de données et de puissants outils de fouille de données afin d'améliorer la qualité des services à leurs usagers et la prise de décisions à partir d'une meilleure compréhension de leur communauté d'utilisateurs. Un article paru en 2006 traite du sujet de la sensibilité de la fouille des données

recueillies sur les abonnés par les bibliothèques et propose une approche pour que leur utilisation soit acceptable (Nicholson 2006). On y mentionne la nécessité de la création du terme « *bibliomining* », qui combine les techniques du forage de données, de la bibliométrique, des statistiques et de l'entreposage de données afin de faciliter l'indexation de ce sujet de recherche dans la littérature. L'auteur mentionne que les techniques de filtrage collaboratif et de filtrage par contenu peuvent être utilisées et combinées pour créer des liens supplémentaires entre les documents de la collection. Il suggère aussi d'utiliser les métadonnées des documents se trouvant dans les notices bibliographiques MARC 21 pour l'élaboration de recommandations par filtrage de contenu. Différentes données recueillies sur l'usager, dont ses termes de recherches, ses documents consultés, ses événements de circulation et ses demandes de références, sont de bons candidats pour fournir une rétroaction implicite sur l'appréciation ou non d'un document. Finalement, l'auteur appuie la création du projet ouvert Digital Reference Electronic Warehouse (DREW) avec comme objectif d'offrir un modèle d'entreposage de données commun à toutes les bibliothèques afin de faciliter la recherche et les comparaisons entre elles.

En 2008, les chercheurs allemands Michael Mönnich et Marcus Spiering ont publié un article sur le développement entre 2002 et 2007 du système BibTip par la German Research Foundation. Le système a été commercialisé et déployé dans plusieurs bibliothèques allemandes (Mönnich & Spiering 2008). Il a été basé sur la consultation des notices bibliographiques dans le catalogue d'accès public de la bibliothèque plutôt que sur les événements de circulation. Un algorithme de règles associatives « a priori » a été utilisé afin d'analyser des modèles dans la consultation des notices au cours des sessions de travail. Le résultat est une liste de recommandations par document plutôt que par abonné. Les avantages de l'approche sont que la confidentialité des données ne pose pas de problème et que l'utilisation des consultations permet d'analyser beaucoup plus de données et de recommander des documents qui ne sont pas disponibles à la circulation, comme les ouvrages de référence. Les auteurs mentionnent que le système fonctionne mieux quand les transactions sont nombreuses dans la base de données. Il semble y avoir un effet de loi de Pareto où une liste de 200 000 recommandations est suffisante pour couvrir 80 % des consultations sur un catalogue d'un million de documents. Les résultats d'un sondage effectué entre 2005 et 2006 auprès des utilisateurs du système ont montré un taux d'appréciation de 84 %.

Les chercheurs néerlandais Robin van Meteran et Maarten van Someren ont quant à eux développé en 2000 le système Personalized Recommender System (PRES) basé sur les techniques du filtrage par contenu à partir d'une base de données d'articles traitant de conseils de rénovation (van Meteran & van Someren 2000). Les auteurs ont utilisé les méthodes *tf-idf* et le modèle vectoriel afin d'élaborer un dictionnaire de termes à partir du contenu des articles, de construire un profil et de comparer la similarité des documents avec une distance cosinus. Un système d'évaluation implicite a été utilisé afin d'évaluer automatiquement les documents appréciés par l'usager authentifié. L'avantage du système est que les résultats sont personnalisés selon l'historique et les préférences individuelles des usagers. Les auteurs mentionnent les difficultés d'obtenir une précision constante à partir des textes eux-mêmes en raison du vocabulaire très vaste utilisé, ce qui complique la construction du dictionnaire. Ils recommandent d'ailleurs de combiner les techniques de filtrage collaboratif à celles du filtrage de contenu afin d'obtenir le meilleur des deux approches.

Les auteurs Ji, Liu & Qi (2015) ont de leur côté comparé les systèmes de recommandations de livres basés sur les approches par contenu, par collaboration et par règles d'association. Ils ont conclu que l'approche par filtrage collaboratif perd de son efficacité à mesure que la bibliothèque possède plus de documents et d'usagers avec des profils différents. L'approche par contenu est quant à elle perçue par les auteurs comme étant plus difficile à mettre en place, car elle nécessite des métadonnées et un dictionnaire de termes de qualité. Les auteurs ont ultimement retenu l'approche par règles d'associations et la méthode « a priori » pour sa simplicité et la facilité d'analyser les résultats. Ils proposent d'utiliser les techniques de *clustering* sur les utilisateurs afin d'éviter les problèmes de démarrage à froid au moment de l'ajout d'un nouvel utilisateur au système.

Les étudiants Mathieu Dion, Jonathan Muschale et Pascal Laforce ont tenté en 2017 une expérience de conception d'un système de recommandation avec une approche collaborative en utilisant l'algorithme des plus proches voisins (k-NN) afin de tenter de recommander des documents sans utiliser les métadonnées des documents (Dion, Mushalle & Laforce 2017). L'approche était simple, mais les résultats obtenus n'ont toutefois pas permis d'obtenir un taux de précision supérieur à 15 % en raison de la collection de documents trop vaste pour le nombre d'emprunts effectués par les abonnés. Les auteurs en ont conclu que l'utilisation des caractéristiques des documents permettant d'obtenir une fonction de poids augmenterait significativement le taux de précision.

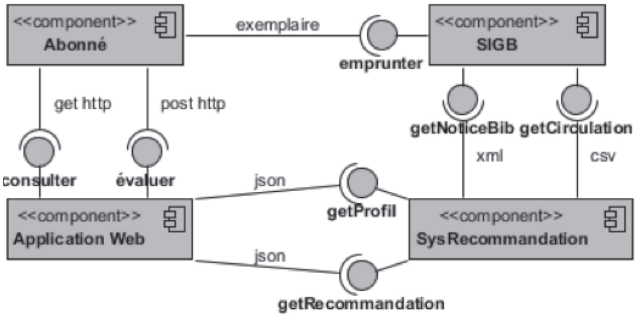
Le système de recommandation développé dans le cadre de cette expérience tentera d'aborder les problèmes et les enjeux soulevés par les études et expériences précédentes. Le système utilisera principalement les techniques du filtrage par contenu afin d'extraire les caractéristiques des documents et utilisera une méthode d'évaluation implicite basée sur les événements de circulation du SIGB.

Matériel utilisé

La Figure 1 montre l'architecture générale du système de recommandation développé à BAnQ.

FIGURE 1

Architecture du système de recommandation



Données provenant du SIGB

Le système de recommandation repose sur les données provenant du SIGB dont les notices bibliographiques et leurs étiquettes, la liste des abonnés de la bibliothèque et la liste des événements de circulation.

Les notices bibliographiques sont essentiellement des fiches décrivant les documents que possède la bibliothèque dans sa collection. Les SIGB utilisés en Amérique du Nord utilisent majoritairement le format de description MARC 21 développé par la Library of Congress depuis les années 1950 (Library of Congress 2017). Seulement 650 000 des 5,5 millions de notices bibliographiques que contient le SIGB de BAnQ sont disponibles à la circulation et sont utilisés par le système de recommandation. Le format XML est généralement utilisé pour l'échange des notices MARC 21 afin de faciliter l'interopérabilité des systèmes.

Le format MARC 21 définit plusieurs centaines de propriétés différentes par un numéro unique communément nommé étiquette. Il y a des étiquettes pour définir les auteurs, le titre, les numéros d'identification, les sujets, les termes d'indexation, les résumés, etc. Les étiquettes peuvent elles-mêmes être éclatées en sous-zone afin de donner plus ou moins de précision sur un élément en particulier. Les sous-zones contiennent des chaînes de caractères généralement courtes représentant la valeur de la propriété. Le système de recommandation utilise ces chaînes de caractères pour constituer un dictionnaire de 250 000 termes uniques. Les termes conservés doivent être partagés par plusieurs documents pour être utiles au système de recommandation et jusqu'à cinq millions d'association terme-document sont possibles.

Les renseignements personnels des abonnés sont généralement conservés dans un système de gestion de la clientèle, mais les SIGB nécessitent seulement un identifiant unique pour fonctionner. Il y a un peu plus d'un million d'usagers inscrits à BAnQ utilisant divers services, mais seulement 200 000 abonnés actifs profiteraient des recommandations.

Les événements de circulation représentent les transactions de prêt, de renouvellement, de retour qu'un abonné effectue sur un document. BAnQ cumule près de 100 millions de transactions depuis son ouverture, mais le système a seulement besoin de connaître le dernier emprunt d'un abonné actif pour constituer son historique, ce qui représente environ 10 millions d'enregistrements.

Outils et technologies

Le SIGB fournit un moteur de recherche intégré afin de forer les notices bibliographiques et de découvrir les étiquettes ayant un potentiel pour le système de recommandation.

Les notices bibliographiques sont extraites du SIGB par l'intermédiaire des fonctionnalités d'exportation automatique qu'offre le SIGB. Chaque notice est exportée dans un répertoire de travail en XML et le langage XPATH permet d'extraire des portions spécifiques afin d'alimenter le dictionnaire. Des extractions en CSV (*comma-separated values*) de la liste des abonnés et de leurs prêts sont aussi utilisées.

Le modèle de données du système de recommandation est réalisé en SQL dans une base de données relationnelle Oracle Enterprise. Les fonctions d'agrégations standard et les fonctions analytiques spécialisées d'Oracle sont utilisées dans les requêtes SQL du système afin de calculer le poids des termes et la similarité des documents. Un paquetage en PL/SQL permet de gérer le processus de recommandation et de formater le résultat des recommandations en objets JSON (*JavaScript Object Notation*). Le principal avantage que revêt l'utilisation des technologies SQL est qu'elles sont plus facilement accessibles par les équipes techniques des bibliothèques que les outils de statistiques couramment utilisés par les statisticiens, comme R et Matlab.

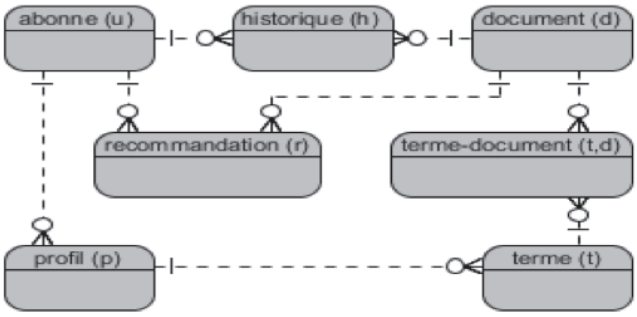
Enfin, un prototype Web développé avec Oracle Application Express permet d'afficher les résultats à partir de services Web RESTful développés pour le système de recommandation. Le choix de cette technologie n'est que pour le développement rapide d'un prototype et d'autres technologies pourraient facilement être utilisées en remplacement.

Méthodologie

Le processus de conception du système de recommandations est constitué d'activités touchant le prétraitement, la modélisation et le post-traitement. La Figure 2 présente le modèle entité-relation qu'utilise le système de recommandation.

FIGURE 2

Modèle entité-relation



Activités de prétraitement

La première étape consiste à analyser et à comprendre le format de fichier MARC 21 que supporte la grande majorité des SIGB utilisés au Québec afin de confirmer le potentiel de la technique du filtrage par contenu. Cette recherche est appuyée par le travail d'un comité d'experts du catalogage. Le comité sélectionne les notices bibliographiques qui sont d'intérêt pour le système de recommandation. Le résultat est un vecteur $D = (d_1, d_2, ..., d_n)$ de n documents représentant les notices bibliographiques sélectionnées.

La deuxième étape consiste à sélectionner et à extraire les étiquettes de la norme qui sont d'intérêt pour le système de recommandation. Les auteurs, les éditeurs, la catégorie Dewey, les sujets et les termes d'indexation seront utiles alors que les notes, les résumés et les numéros ISBN ne le seront pas. Le résultat est un vecteur $T = (t_1, t_2, ..., t_n)$ de n termes créés à partir des étiquettes et des sous-zones sélectionnées des documents actifs. Le vecteur X des termes-documents est également créé.

La troisième étape consiste à extraire les abonnés actifs et leur historique de prêt. Seul un identifiant suffit au système de recommandation en raison de l'approche de filtrage par contenu qui a été choisi. L'historique de prêt est constitué à partir des événements de circulation du SIGB des m derniers mois. En effet, un abonné n'emprunte pas directement un document, mais plutôt un des exemplaires de ce même document. Les données peuvent ainsi être réduites à une simple paire « NoAbonné — NoDocument ». La constante m permet d'éliminer les prêts qui ne sont peut-être plus représentatifs des goûts de l'abonné. Le résultat est le vecteur $U = (u_1, u_2, ..., u_n)$ de n abonnés et le vecteur $H = (h_1, h_2, ..., h_n)$ des n documents empruntés par l'abonné.

Activités de modélisation

La première étape consiste à bâtir le profil personnel de chaque abonné à partir des termes du dictionnaire qui sont associés aux documents empruntés. La liste de ces termes est obtenue par la jointure du vecteur des termes-documents X et du vecteur des documents empruntés de l'abonné H. Seul l'historique des 18 derniers mois est utilisé afin de dynamiser davantage l'algorithme. Le résultat de la jointure est une collection de termes qui se répètent plus ou moins selon les documents se retrouvant dans l'historique de l'abonné et le travail de catalogage qu'ont effectué les bibliothécaires dans le SIGB. La collection est ensuite réduite pour ne conserver que la fréquence brute $f_{t,\,h}$ de chaque terme. Le résultat est un vecteur $P = (p_1, p_2, ..., p_n)$ des n termes se trouvant dans le profil de l'abonné.

La deuxième étape consiste à calculer le poids w des termes du profil en utilisant une variation de la technique de recherche d'information *tf-idf*. Elle permet d'attribuer à un terme plus de poids lorsque celui-ci revient souvent dans un même document (ou dans un même profil) et moins de poids s'il revient trop souvent dans la collection. La fonction utilisée est la suivante :

$$w = tf(t,d) * idf(t,D) * tag(t) * delai(h)$$

La fonction *tf* correspond à la fréquence brute du terme parmi les documents empruntés par l'abonné. Elle est couramment normalisée par un logarithme afin d'éviter qu'un seul terme n'occupe trop de place.

$$tf(t,d) = 1 + \log(f_{t,d})$$

La fonction *idf* représente la fréquence inverse du document parmi la collection qui correspond au nombre total de documents qui utilisent le terme présent ou non dans l'historique de l'abonné.

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

La fonction *tag* retourne un nombre réel entre 0 et 1 attribué préalablement à chaque étiquette MARC 21 par le comité d'experts. La fonction *delai* représente le nombre de mois depuis la dernière utilisation du terme dans l'historique de l'abonné. Le résultat est le vecteur P pondéré de l'abonné. Pour des besoins de performance dans la prochaine étape, il peut s'avérer nécessaire de réduire le vecteur à la liste des top-N termes.

La troisième étape consiste à comparer les documents de la collection avec le profil de l'abonné. Pour ce faire, une fonction de similarité basée sur la distance euclidienne est utilisée.

$$sim(d,P) = \sqrt{\sum_{t \in P : d \in p}^{|P|} w_{t,d}^2}$$

Le travail consiste, pour chaque document, à obtenir l'intersection des termes se trouvant dans le document avec les termes se trouvant dans le profil de l'abonné. Il faut ensuite additionner les poids calculés précédemment dans le profil. Le résultat est une liste de documents triés du plus proche au plus loin avec le profil.

La quatrième étape consiste à multiplier la distance de similarité par l'attribution d'un poids qui vise à augmenter la pertinence du document quand l'un des trois cas suivants se présente :

- La langue du document correspond à la langue utilisée le plus souvent par l'abonné ;
- Le document est réputé populaire auprès des abonnés ;
- Le document fait partie des nouveaux arrivages du catalogue.

Il s'avère également nécessaire de retirer de la liste les documents déjà empruntés par l'abonné et les documents avec des titres trop similaires. Le résultat est une liste ajustée des documents similaires au profil de l'abonné.

Activités de post-traitement

La première étape consiste à développer quelques services Web qui permettent de consulter le résultat du système. Le premier service Web retourne la liste des documents disponibles dans la collection. Le deuxième retourne la liste des termes du dictionnaire. Le troisième retourne l'historique d'un abonné. Le quatrième retourne la liste des termes présents dans le profil de l'abonné. Le cinquième retourne la liste des documents recommandés par le système pour un abonné.

La deuxième étape consiste à intégrer les services Web du système de recommandation au SIGB, à l'espace « Mon dossier » du portail institutionnel, à une application externe ou aux courriels de diffusion envoyés aux abonnées. Pour les besoins du projet à l'étude, un prototype d'application Web a été développé afin de montrer la faisabilité du projet.

La troisième étape consiste à permettre aux abonnés de donner une rétroaction directe sur les recommandations proposées par l'entremise de l'application Web. Cette rétroaction se résume par une évaluation booléenne « le document m'intéresse » ou « le document ne m'intéresse pas ». Les données recueillies permettent d'évaluer la précision du système de recommandation grâce à la participation volontaire des usagers.

Résultats et discussion

Sélection des participants

Le système de recommandation a été déployé dans un environnement de test et mis à la disposition d'une dizaine de participants pendant une semaine. Les participants étaient des informaticiens, des bibliothécaires et des directeurs travaillant pour BAnQ qui trouvaient intéressant d'intégrer éventuellement un système de recommandation aux logiciels de la bibliothèque. Les participants étaient invités à décrire sommairement l'usage qu'ils font de leur carte d'abonné afin d'aider à évaluer subjectivement les résultats du système. Les participants étaient plutôt représentatifs des usagers de la bibliothèque et avaient chacun des profils d'emprunt très différents les uns des autres. En effet, certains ont emprunté plus de 1 000 documents sur une période de 10 ans alors que d'autres ne sont abonnés que depuis quelques mois. Certains ont un profil d'emprunt très spécifique (ex. : romans fantastiques ou guides de voyage) alors que d'autres empruntent divers types de documents (ex. : roman, film, musique, jeux vidéo, etc.). Certains n'empruntent que pour eux-mêmes alors que d'autres empruntent pour toute leur famille avec leur carte.

Le Tableau 1 présente le nombre de documents empruntés par les participants et trois mots décrivant sommairement leurs centres d'intérêt de lecture afin de les comparer au résultat du système de recommandation.

TABLEAU 1

Nombre d'emprunts et description des participants

Abonné	18 mois	Total	Description
A	39	543	Individuel, Philosophie, Cyclisme
B	33	438	Famille, Romans, Revues
C	25	178	Individuel, Romans, Guide voyage
D	8	125	Individuel, Jeux vidéo, Films
E	317	1 549	Famille, Cuisine, Romans jeunesse
F	54	54	Individuel, Science, Philosophie
G	174	1 158	Individuel, Jeux vidéo, Films
H	39	236	Famille, Jeux vidéo, BD
I	24	156	Individuel, Films, Fantastique
J	16	25	Individuel, Musique, Séries télévisées
K	49	179	Individuel, Films, Séries télévisées
L	82	741	Famille, Romans, Fantastique

Les Tableaux 2, 3 et 4 présentent des extraits du résultat du système de recommandation pour le participant *L*. Le participant *L* emprunte régulièrement pour lui-même des romans fantastiques et des films de science-fiction en plus d'emprunter des ouvrages très variés pour sa femme et ses quatre enfants âgés de 10 à 18 ans.

TABLEAU 2

Extrait de l'historique du participant L

Titre	Date emprunt
1984 / Georges [c.-à-d. George] Orwell ; traduit de l'anglais par Amélie Audiberti.	2017-07-14
Je t'aime encore comme ça / Francesco Gungui ; traduit de l'italien par Faustina Fiore.	2017-07-10
Brume : nouvelles / Stephen King ; traduites de l'anglais (États-Unis) par Michèle Pressé et Serge Quadruppani.	2017-07-06
Harvest of time / Alastair Reynolds.	2017-06-08
La Shoah / Thomas Cussans ; traduction : Jean-Pierre Dauliac	2017-05-22
Les yeux du dragon / Stephen King ; traduit de l'américain par Évelyne Châtelain ; illustrations intérieures de Nicolas Duffaut.	2017-05-22
Radioprotection en radiologie médicale : 101 questions pour comprendre et agir / Hervé Leclet, Martine Madoux.	2017-02-17
Land art de printemps / par Marc Pouyet.	2016-07-26

TABLEAU 3

Extrait du profil du participant L

Groupe	Terme	TF	DF	Total
SUJ	Rayonnement	5	24	8,3736
AUT	Gungui, Francesco	2	3	6,9338
DEW	616.989705	2	2	6,3774
SUJ	Torrance, Dan (Personnage fictif)	1	2	6,2949
SUJ	Guerre spatiale	5	454	6,2799
SUJ	Radioactivité	2	39	6,1076
AUT	Corey, James S. A.	2	8	5,9822
SUJ	Astéroïdes	3	39	5,9574
AUT	Orwell, George	2	47	5,3959
SUJ	Vérone (Italie)	1	17	5,3863
SUJ	Land art	3	56	4,7604
SUJ	Science-fiction	5	3 903	4,7301
SUJ	Mesures	5	494	4,7159
SUJ	Web	1	76	4,6592
SUJ	Doctor Who (Personnage fictif)	1	106	4,6283

Évaluation qualitative des résultats

Les participants étaient ensuite invités à faire parvenir leur appréciation générale du système par courriel. Une forte majorité des répondants ont donné une évaluation d'ensemble positive du système, bien que perfectible. Les commentaires reçus mentionnent que certaines thématiques présentes dans leur profil y occupent trop de place.

TABLEAU 4

Extrait des recommandations du participant L

Titre recommandé
Star Trek. Vers les ténèbres [enregistrement vidéo] = Star Trek. Into darkness / réalisation, J. J. Abrams.
The expanse : roman / James S. A. Corey ; traduit de l'anglais (États-Unis) par Thierry Arson.
Conduire un véhicule de promenade / [publication réalisée par la Société de l'assurance automobile du Québec ; recherche et rédaction, Michèle Jean et Micheline Briand avec la collaboration de Diane Hamel… et al.]
Harlock, space pirate [enregistrement vidéo] = Albator, corsaire de l'espace / réalisation, Shinji Aramaki ; scénario, Harutoshi Fukui, Kiyoto Takeuchi.
Prometheus [enregistrement vidéo] / réalisation, Ridley Scott.
Aurélie Laflamme : les pieds sur terre = Aurélie Laflamme : somewhat grounded / réalisation, Nicolas Monette.
L'Académie Jedi / Jeffrey Brown ; texte français d'Isabelle Allard.
Chroniques lunaires / Marissa Meyer ; traduit de l'anglais (États-Unis) par Guillaume Fournier.
Goldorak [enregistrement vidéo] : l'intégrale / histoire originale et conception des personnages, Go Nagai ; réalisateur, Tomoharu-Katsumata ; producteur, Toei Animation Co., Ltd. -
La comédie atomique : l'histoire occultée des dangers des radiations / Yves Lenoir.
Land art d'automne / par Marc Pouyet.
E. T. l'extraterrestre [enregistrement vidéo] = E. T. the extraterrestrial / réalisation, Steven Spielberg.

Après analyse, certaines catégories documentaires comme les jeux vidéo semblent en effet occuper trop de place dans les recommandations en raison du facteur de popularité qui a été inclus dans l'algorithme et qui est anormalement élevé en raison de la grande popularité des jeux vidéo. Le système ne peut reconnaître la console de jeu de l'abonné vu que cette information n'a pas été retenue pour l'élaboration du système de recommandation. D'autres participants font état du fait que le système leur recommande des romans jeunesses en raison des emprunts que ceux-ci ont faits par le passé pour leurs enfants alors qu'ils auraient préféré des recommandations ne s'appliquant qu'à eux. D'autres soulignent que leurs centres d'intérêt ont changé avec le temps et qu'ils souhaiteraient personnaliser les différents facteurs de normalisation des résultats. Finalement, l'ensemble des participants souhaitent que le système permette le retrait manuel d'un ou de plusieurs documents issus de trois sources : 1) historique de l'abonné ; 2) documents recommandés déjà consultés par l'usager sur une autre plateforme (ex. : l'usager a déjà vu le film au cinéma) ; 3) documents recommandés non pertinents pour l'abonné.

Évaluation quantitative des résultats

À la suite de cette première expérience, des ajouts ont été faits au système de recommandation afin d'implémenter un mécanisme d'évaluation « le document m'intéresse » et « le document ne m'intéresse pas ». L'utilisateur peut ainsi retirer les documents de son historique ou de ses recommandations. Cette évaluation explicite permet d'éviter que des documents empruntés pour un besoin très spécifique par l'abonné ne soient utilisés pour calculer son profil par la suite. Le second avantage de cet ajout est qu'il permet de cumuler des données statistiques sur le taux de précision du modèle. Les documents évalués positivement sont ainsi considérés comme des vrais positifs et les documents évalués négativement comme de faux positifs. La liste des réservations de l'abonné peut quant à elle servir pour le calcul du rappel où les documents réservés, mais non recommandés par le système, sont considérés comme de faux négatifs. L'algorithme de recommandation a été ajusté afin de simplement ignorer les documents identifiés comme étant inintéressants par l'ajout de filtres au moment de la création du vecteur *P* et des recommandations. L'expérience de BAnQ a donné des résultats similaires à ceux obtenus par R. van Meteran et M. van Someren avec le système PRES. L'ajout de l'évaluation directe a permis au système d'apprendre les préférences de l'abonné plus rapidement et de retourner de meilleures recommandations tout en nécessitant moins d'effort de calcul. À la suite des modifications, une deuxième ronde d'évaluation a été effectuée avec cinq des participants originaux. Les participants ont cette fois été invités à passer en revue les documents de leur historique et de leur liste de recommandations afin d'identifier ceux qui ne sont pas ou ne sont plus pertinents à l'abonné.

Le Tableau 5 présente le résultat de l'évaluation de l'historique des participants.

TABLEAU 5

Évaluation de l'historique

Abonné	Historique	Éval.	Taux	Positives	Négatives	Retraits
A	543	73	13,40 %	49	24	4,40 %
B	438	84	19,20 %	68	16	3,70 %
D	125	124	99,20 %	98	26	20,80 %
G	1 158	26	2,20 %	25	1	0,10 %
K	179	167	93,30 %	154	13	7,30 %
Total	**2 443**	**474**	**19,40 %**	**394**	**80**	**3,30 %**

Les résultats montrent que les participants ont réussi à évaluer 19,4 % de leur historique pour ultimement retirer un total de 3,3 % des documents du système de recommandations. Les résultats tendent à montrer que plus les participants passent en revue leur historique, plus les documents sont retirés du système de recommandation. Cela permet de conclure que les intérêts personnels des abonnés changent avec le temps et que l'hypothèse initiale de restreindre le système de recommandation à l'analyse des 18 derniers mois est en soi une bonne idée. Les commentaires reçus des participants indiquent que l'ajout d'une préférence personnelle au dossier de l'abonné leur permettant de contrôler le nombre de mois à retenir serait souhaitable.

Le Tableau 6 présente le résultat de l'évaluation quantitative des recommandations par les participants.

TABLEAU 6

Évaluation des recommandations

Abonné	Recom.	Éval.	Pos.	Nég.	Réserv.	Précision	Rappel
A	25	25	16	9	8	64,0 %	32,0 %
B	25	16	12	4	1	75,0 %	6,3 %
D	25	25	16	9	0	64,0 %	0,0 %
G	25	25	24	1	4	96,0 %	16,0 %
K	25	25	15	10	14	60,0 %	56,0 %
Total	**125**	**116**	**83**	**33**	**27**	**71,6 %**	**23,3 %**

Les résultats indiquent que 83 des 116 recommandations évaluées étaient pertinentes aux participants permettant d'obtenir un taux de précision moyen tout à fait acceptable de 71,6 %. L'analyse des 27 documents réservés par les participants a permis de constater qu'aucun n'était toutefois présent dans la liste des recommandations permettant d'obtenir un taux de rappel moyen de 23,3 %. L'analyse des données individuelles sur les réservations permet de constater qu'il faudrait augmenter le nombre de recommandations calculées par le système avant de diminuer le taux de rappel. Cela se ferait toutefois au détriment du taux de précision qui dégringolerait en flèche. Le nombre de recommandations est un paramètre sensible qui devrait potentiellement varier d'un abonné à l'autre, car les abonnés n'ont pas toujours suffisamment de documents dans leur historique pour que le système de recommandation puisse bâtir le profil avec suffisamment de confiance. Il serait sans doute préférable d'utiliser une approche de rang par degré de similarité minimal plutôt que d'utiliser l'approche des 25 meilleurs résultats comme le suggèrent Yan et Garcia-Molina (1994) dans leur étude.

Évaluation de l'atteinte des objectifs

La constitution d'un dictionnaire de termes riches à partir des notices bibliographiques du SIGB a été grandement facilitée par le fait que les bibliothécaires annotent rigoureusement les notices bibliographiques dans le catalogue et qu'ils utilisent le répertoire des vedettes-matières qui est maintenu par un fournisseur externe. Les termes associés aux notices bibliographiques sont tenus à jour par les bibliothécaires, ce qui évite aussi d'avoir des différences importantes dans la qualité des résultats de recherche. Le système d'indexation en place évite l'utilisation des méthodes LSA et de lemmatisation des termes comme dans van Mereten et van Someren (2000). La complexité de l'activité est principalement due à la sélection des bonnes étiquettes à utiliser dans le format MARC 21 et à la pondération que l'on souhaite accorder. La pondération utilisée pourrait varier énormément d'une bibliothèque à l'autre en fonction des particularités de la collection et des valorisations de catégories documentaires souhaitées par le ystème de recommandation.

La constitution du profil des abonnés avec la méthode *tf-idf* a bien fonctionné. Le succès est dû à la bonne qualité du dictionnaire de termes utilisé et à la bonne réutilisation des mêmes termes dans les notices du SIGB. Les étiquettes d'une notice étant souvent subdivisées en plusieurs souszones (ex. : « France >> Histoire >> XIVe siècle >> Romans, nouvelles, etc. ») font en sorte qu'un abonné peut spécifier son profil s'il possède dans son historique plusieurs documents traitant d'une même subdivision. Il faut éviter que les subdivisions ne prennent trop de poids, car certaines catégories documentaires n'ont pas toujours la même richesse et d'autres ont certaines combinaisons de termes qui reviennent constamment ensemble. Des améliorations dans l'implémentation de la méthode *tf-idf* sont encore à évaluer afin d'ajouter un facteur subdivision du terme par étiquette. Aussi, il reste à mettre en place des mécanismes de *clustering* d'abonnés pour éviter le problème de démarrage à froid comme le suggéraient Ji, Liu, Song et Qi (2015). R. van Meteran et van Someren (2000) concluaient également que la combinaison des techniques de l'approche collaborative avec l'approche par contenu comme le font Amazon.com et Netflix permettrait d'améliorer les résultats par l'injection de termes absents au profil de l'abonné en fonction des intérêts partagés par la communauté.

Le calcul de la distance de similarité a été plus complexe que prévu initialement, car la distance cosinus, la plus utilisée en sémantique vectorielle, est plus difficile à implémenter en langage SQL au moyen de fonctions d'agrégation et de jointures que le sont d'autres mesures de distance. Une distance euclidienne a été utilisée en remplacement et a tout de même donné de bons résultats. Les premiers tests ont permis de découvrir que les différentes éditions d'un même ouvrage (ex. : édition 2010, 2011 et ses traductions) ont chacune leur propre notice bibliographique et un numéro de document différent. Le système relevait la très grande similarité entre ces notices et avait alors tendance à recommander les éditions alternatives plutôt que d'autres documents. Le problème a été en grande partie réglé en appliquant un filtre sur le titre abrégé des documents plutôt que sur leur numéro au moment du retrait de l'historique. Il reste toutefois des améliorations à apporter à ce filtre, car il arrive que des documents trop similaires aient un titre suffisamment différent pour qu'ils soient recommandés quand même par le système.

L'intégration du système de recommandation au SIGB n'est pas encore complétée, mais l'intégration des services Web RESTful a été démontrée comme étant très simple à réaliser dans une application Web servant de prototype. Dans la plupart des cas, il suffit de fournir le numéro de l'abonné en paramètre pour obtenir les données générées par le système de recommandation. Il reste toutefois des travaux à faire pour protéger la confidentialité des données. Des discussions avec le fournisseur du SIGB à BAnQ sont prévues dans les prochains mois.

L'évaluation de la qualité des recommandations à partir de la rétroaction explicite des usagers a été plutôt simple à mettre en place. Les participants de l'étude étaient enthousiastes à l'utiliser puisque cela leur a permis d'obtenir de meilleures recommandations. Pour l'heure, seules les

évaluations négatives ont été utilisées pour élaguer les résultats, mais des améliorations au système permettraient d'y intégrer les évaluations positives (ex. : en accordant un poids plus important aux termes provenant d'un document intéressant) et en incluant les réservations des abonnés dans le processus d'élaboration du profil des abonnés. L'ajout de personnalisation des facteurs de normalisation des données (temporelle, fréquence, rareté, étiquette, popularité, etc.) permettrait d'améliorer le taux de précision obtenu et d'optimiser le système en fonction des préférences de ses utilisateurs. L'évaluation du rappel dans un système de recommandation appliqué aux bibliothèques est difficile à utiliser puisqu'il arrive régulièrement que plusieurs documents traitant d'un même sujet soient considérés par l'abonné comme étant des documents alternatifs plutôt que des documents complémentaires (ex. : un roman publié existant sous différents formats).

Conclusion

L'expérience à BAnQ montre qu'il est possible d'utiliser les données provenant d'un système de gestion intégré de bibliothèque et d'y intégrer un système de recommandation de documents personnalisés à chaque abonné en utilisant une approche de filtrage par contenu. Les notices bibliographiques contiennent des données idéales pour la constitution d'un dictionnaire de termes sans les problèmes de réduction, typiques dans le filtrage par contenu. La construction d'un profil par abonné à partir de son historique fonctionne bien, mais il s'avère nécessaire d'apporter une attention particulière aux centres d'intérêt changeant des usagers et au problème de démarrage à froid. Le calcul de la similarité des documents nécessite des adaptations particulières en raison de la présence de proches doublons dans la collection des bibliothèques. L'intégration du système de recommandation au SIGB ou au portail de l'organisme ne présente pas de problèmes techniques particuliers du moment où l'entente de service avec le fournisseur permet ce type d'intégration. L'expérience a montré qu'il est possible d'obtenir un bon taux de précision lorsqu'on intègre les évaluations directes des abonnés puisqu'il permet à celui-ci de s'approprier le système de recommandation et de le personnaliser à son goût. L'expérience à BAnQ est somme toute concluante et l'organisme y voit un potentiel pour améliorer ses services aux usagers. Les outils et technologies utilisés pour le projet sont répandus dans le milieu des bibliothèques publiques québécoises et le projet devrait être reproductible dans plusieurs autres milieux documentaires.

Quelques pistes d'améliorations possibles au système de recommandation seront à l'étude dans les prochains mois. La sélection et la pondération des étiquettes provenant des notices bibliographiques seront réévaluées afin de régler certains problèmes liés à quelques catégories documentaires. L'ajout d'un mécanisme de *clustering* d'abonné sera évalué ainsi que l'ajout d'un paramètre pour personnaliser le nombre de mois à retenir pour l'évaluation. D'autres travaux seront nécessaires afin d'éliminer les cas de proches doublons restants. Des modifications devront être demandées au fournisseur du SIGB afin d'y intégrer une section de recommandation dans le catalogue d'accès public. Des travaux sont prévus pour inclure une pondération supplémentaire dans les évaluations positives au moment de la constitution du profil de l'abonné. Finalement, BAnQ envisage de présenter le système de recommandation au milieu documentaire dans le cadre d'un colloque et de leur offrir une coquille réutilisable du projet qui leur permettrait d'implémenter le système dans leur propre environnement.

Le système de recommandation conçu pour BAnQ démontre le potentiel de réutilisation des données opérationnelles cumulées par les systèmes intégrés de gestion des bibliothèques ainsi que la facilité de les mettre en place avec les outils et les techniques de forage de données modernes qui ont fait leurs preuves dans d'autres secteurs d'application.

Remerciements

Nous remercions M. Christian Desrosiers pour ses conseils sur la recommandation de produit.

Nous remercions M. Jonathan Muschale pour ses recherches sur les algorithmes et les outils d'apprentissage automatique et sa participation au premier projet de recommandation.

Nous remercions M. Philippe Chabot, M. Christian Eilers, M. Éric Éthier, M. Jean-François Gauvin, M. Jean-Bruno Giard, Mme Colette Langlois, M. Antonio-Otavio Mello, Mme Pascale Montmartin, M. Paulo Leduc, Mme Marielle St-Germain, M. Stéphane Tellier et M. Réjean Thibeault pour leur participation aux essais du système de recommandation.

SOURCES CONSULTÉES

Callan, J. & A. Smeaton. 2003, mai. Personalisation and Recommender Systems in Digital Libraries Joint NSF-EU DELOS Working Group Report. *ERCIM News*.

Nicholson, S. 2006. The basis for bibliomining : Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. *Information Processing & Management* 42 (3) : 785-804.

Mönnich, M. & M. Spiering. 2008. Adding Value to the Library Catalog by Implementing a Recommendation System. *D-Lib Magazine* 14 (5/6).

Meteran, R. van & M. van Someren. 2000. *Using Content-Based Filtering for Recommendation*. Amsterdam : NetlinQ Group.

Ji, W., S. Liu, Y. Song & J. Qi. 2015. Research of Intelligent recommendation system based on the user and association rules mining for books. *5th International Conference on Computer Sciences and Automation Engineering* : ICCSAE, 294-299.

Dion, M., J. Mushalle & P. Laforce. 2017. *Recommandation de produit chez BAnQ*. École de technologie supérieure.

Library of Congress. 2017. MARC 21 format for bibliographic data. 1999 Edition Update No 1 (October 2000) through Update No 24 (May 2017). Consulté le : 24 juillet 2017. <www.loc.gov/marc/bibliographic/>.

Yan, Tak W. & H. Garcia-Molina. 1994. Index Structures for Information Filtering Under the Vector Space Model. In *Proceedings of the Tenth International Conference on Data Engineering*. Washington : IEEE Computer Society, 337-347.