Données et industries des contenus, modélisation et prototypage d'un système de recommandation pour le catalogue d'un service d'information documentaire

Amine Sennouni

HESAM UNIVERSITÉ

le cnam

ÉCOLE DOCTORALE ABBÉ GRÉGOIRE

Laboratoire Dispositifs d'Information et de la Communication à l'Ère Numérique, Paris, Île-de-France (DICEN-IDF)

THÈSE

présentée par : **Amine SENNOUNI**

soutenue le : **1er Juillet 2021**

pour obtenir le grade de : **Docteur d'HESAM Université**

préparée au : **Conservatoire national des arts et métiers**

Discipline : **Sciences humaines et humanités nouvelles**

Spécialité : **Sciences de l'information et de la communication**

DONNEES ET INDUSTRIES DES CONTENUS, MODELISATION ET PROTOTYPAGE D'UN SYSTEME DE RECOMMANDATION POUR LE CATALOGUE D'UN SERVICE D'INFORMATION DOCUMENTAIRE

THÈSE dirigée par :

Mme CHARTRON Ghislaine

et co-encadrée par :

M. BACHR Ahmed Abdelilah

Jury

Mme Ghislaine CHARTRON, Professeur des universités, CNAM Paris	Directrice de thèse
M. Ahmed Abdelilah BACHR, Professeur des universités, ESI Rabat	Co-encadrant
Mme Brigitte SIMONNOT, Professeure des universités, Université de Lorraine	Rapportrice
M. Imad SALEH, Professeur des universités, Université Paris 8	Rapporteur
M. Mostafa BELLAFKIH, Professeur des universités, INPT Rabat	Président/Examinateur

THÈSE

Affidavit

Je soussigné / soussignée, Amine SENNOUNI, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Mme. Ghislaine CHARTRON (directrice) et de M. Ahmed Abdelilah BACHR (co-directeur), dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect de la charte nationale de déontologie des métiers de la recherche.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Paris, le 19/07/2021

Signature

Affidavit

I, undersigned, Amine SENNOUNI, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific direction of Mrs. Ghislaine CHARTRON (thesis director) and of de Mr. Ahmed Abdelilah BACHR (co-thesis director), in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with the French charter for Research Integrity.

This work has not been submitted previously either in France or abroad in the same or in a similar version to any other examination body.

Place Paris, date 19/07/2021

Signature

Dédicaces

Je dédie cette thèse, comme preuve de respect, de gratitude et de reconnaissance à :

- Mes chers parents, Saida et Ahmed Rachid, pour tous leurs sacrifices, leur amour et leur tendresse éternels ;
- Mon frère, Yassine, pour son soutien et encouragement continus ;
- Mon oncle, Mehdi, de m'avoir été d'un soutien inconditionnel, sur tous les plans, durant ma thèse en France ;
- Toute ma chère famille pour son affection et son soutien ;
- Mes meilleurs amis pour leur aide, leur temps, leurs encouragements, leur assistance et leur soutien ;
- Mes encadrants, qui m'ont aidé énormément dans l'aboutissement de cette thèse par leurs orientations et conseils.

A tous ceux qui ont contribué de près ou de loin à la réalisation de cette thèse.

Merci infiniment.

Remerciements

Il m’est particulièrement agréable avant de présenter mon travail, d’exprimer toute ma gratitude envers les personnes qui, de près ou de loin, m’ont apporté leur sollicitude.

Mes remerciements vont, tout d’abord, à Mme Ghislaine CHARTRON, professeur au Conservatoire National des Art et des Métiers (CNAM), directrice de l’Institut des Sciences et Techniques de la Documentation (INTD) et directrice de ma thèse de doctorat, pour sa disponibilité, ses remarques pertinentes, sa rigueur, ses conseils avisés et ses directives bien placées.

Je remercie également M. Ahmed Abdelilah BACHR, professeur à l’Ecole des Sciences de l’Information (ESI) et co-encadrant de ma thèse de doctorat, pour son suivi, ses orientations et son accompagnement.

Je remercie aussi Pr. Brigitte SIMONNOT, Pr. Imad SALEH et Pr. Mostafa BELLAFKIH pour leur disponibilité et de m’avoir fait l‘honneur d‘accepter de faire partie du jury de cette thèse.

Enfin, j’adresse mes remerciements à M. Salaheddine BAHJI, directeur de l’ESI, pour son soutien et accompagnement et M. El Hassan LEMALLEM, ex-directeur de l’ESI et conseiller du Monsieur le Haut-Commissaire au plan de m‘avoir permis de vivre cette aventure de recherche. Mes remerciements vont également à l’ensemble du corps professoral et administratif du CNAM Paris et de l’ESI Rabat pour leur aimable collaboration.

Résumé

Les systèmes de recommandations font partie des modèles d'apprentissage automatique qui transforment la recherche de l'information. Ce concept reste nouveau dans les services documentaires dédiés à la recherche, et rares sont les travaux qui traitent de ces systèmes dans les Services d'Information Documentaire (SID) physiques et destinés à une catégorie précise d'utilisateurs, à savoir les chercheurs.

C'est dans ce cadre que s'inscrit notre thèse qui a pour objectif principal de concevoir et mettre en œuvre un modèle de système de recommandation basé sur les données implicites d'un échantillon des utilisateurs de l'Institut Marocain de la Recherche Scientifique et Technique (IMIST), en s'appuyant sur l'approche de filtrage collaboratif. Trois sous-objectifs ont été fixés à notre travail :

- Diagnostiquer l'existant et identifier le besoin de l'IMIST en termes d'usage d'un système de recommandation ;
- Modéliser un système de recommandation pour l'IMIST ;
- Mettre en œuvre et évaluer le prototype de moteur de recommandation adapté à l'IMIST.

Pour contextualiser notre travail, nous avons commencé, tout d'abord, par rendre compte de l'importance de la data dans l'optimisation des processus et des services des industries culturelles d'une manière générale puis pour les services documentaires en particulier. En lien avec notre terrain, nous avons ensuite apprécié le besoin de l'IMIST pour un système de recommandation, avant de procéder à la modélisation de celui-ci d'un point de vue fonctionnel et technique.

Notre contribution est ensuite la mise en place d'un prototype de système de recommandation adapté, reposant sur les données implicites des utilisateurs, afin de leur fournir des recommandations pertinentes d'ouvrages à partir des données d'usage implicites et d'autres données déclaratives en rapport avec la spécialisation dans un champs disciplinaire donnée.

Le prototype du système de recommandation proposé, distingue l'utilisateur anonyme, qui bénéficie des recommandations selon les notices consultées, et l'utilisateur abonné, pour lequel le système offre des recommandations personnalisées selon son profil. Le système intègre une conversion les données implicites des usagers en données explicites afin de les exploiter. L'implémentation du système a été effectuée dans un environnement Spark dédié à l'apprentissage automatique à l'aide du langage Scala et moyennant le modèle des moindres

carrés alternés (ALS) de factorisation matricielle, appartenant aux méthodes de filtrage collaboratif à base de modèle.

Mots-clés : système de recommandation, apprentissage automatique, données implicites, filtrage collaboratif, service d'information documentaire, catalogue d'ouvrages, OPAC, industries des contenus, valorisation des données, Maroc.

Résumé en anglais

Recommender systems are among the promising fields of machine learning, which have revolutionized the information retrieval. This concept remains new in documentary structures, and rare are the works that deal with this point.

The context of our thesis is the Moroccan Institute for Scientific and Technical Information (IMIST), having as an objective to design and implement a recommender system based on log data, according to the collaborative filtering approach.

Three sub-objectives were used to guide our work:

- Diagnose the existing situation, and identify the need for the IMIST in terms of the use of a recommender system;
- Modeling a recommender system for the IMIST;
- Implement and evaluate the prototype of recommender system.

To achieve these objectives, we first began to identify the IMIST's need for a recommender system, then we design and carry out a prototype based on user's implicit data to provide recommendations.

This prototype distinguishes between the anonymous user, who benefits from the recommendations according to the notices consulted, and the subscriber user, for whom the system offers personalized recommendations according to his profile.

It should be noted that it was necessary to convert the implicit data of the users into explicit data in the form of scores, in order to exploit them.

The implementation of the system was done in a Spark environment, using the Scala language and the ALS Train Implicit model of the matrix factorization.

Key words: recommender system, machine learning, implicit data, collaborative filtering, documentary services, books catalog, OPAC, content industry, data valuation, Morocco.

Table des matières

Liste des sigles et des abréviations

Le sigle ou l'abréviation	Le développement
AJAX	Asynchronous JavaScript and XML
ALS	Alternating Least Squares
API	Application Programming Interface
CDD	Classification Décimale de Dewey
CNIL	Commission nationale de l'informatique et des Libertés
CSV	Comma-Separated Values
DC	Dublin Core
ETL	Extract Transform Load
FOAF	Friend Of A Friend
GAFAM	Google, Apple, Facebook, Amazon et Microsoft
IMIST	Institut Marocain de l'Information Scientifique et Technique
INRA	Institut National de Recherche Agricole
ISBN	International Standard Book Number
K- NN	K- Nearest Neighbors
LOD	Linked Open Data
MAE	Mean Absolute Error
MARC	Machine Readable Cataloging
MSE	Mean Squared Error
NMAE	Normalized Mean Absolute Error
OCLC	Online Computer Library Center
OPAC	Online Public Access Catalog
RDD	Resilient Distributed Dataset
RDF	Ressource Description Framework
RGPD	Règlement Général de Protection des Données

RMSE	Root Mean Squared Error
RTB	Real Time Bidding
SID	Service d'Information Documentaire
SIGB	Système Intégré de Gestion de Bibliothèque
SKOS	Simple Knowledge Organization System
SR	Système de Recommandation
TDM	Text and Data Mining
UMAE	User Mean Absolute Error
VIAF	Virtual International Authority File

Liste des tableaux

Liste des figures

Liste des annexes

Introduction générale

Notre thèse s'inscrit dans le cadre de la personnalisation de l'expérience utilisateur dans les services des structures documentaires. En effet, la recherche de documents dans un Service d'Information Documentaire (SID) représente parfois une tâche délicate pour un utilisateur, au regard du nombre important de documents disponibles confronté au besoin précis, souvent inscrit dans un temps disponible limité, et pour lequel une partie infime de la collection sera pertinente.

Les SID mettent en place des systèmes classificatoires, composés de plusieurs niveaux de subdivision rendant leur maitrise complexe pour les utilisateurs. En outre, ces collections sont évolutives, dans la mesure où de nouvelles ressources font leur entrée et d'autres sont désherbées ou en circulation, ce qui ne plaide guère pour la familiarisation des utilisateurs avec le fonds du SID. Le catalogue d'accès public (OPAC) en ligne permet à l'utilisateur d'effectuer des recherches et de repérer à partir de mots clés ou de sujets les fiches descriptives des documents. En revanche, ces catalogues publics en ligne demandent souvent beaucoup de temps à l'utilisateur.

Nous pensons que les catalogues en ligne des SID gagneraient à être plus précis, en proposant des recommandations basées sur l'historique de l'interaction de l'utilisateur avec le catalogue, un peu comme le font déjà Amazon, Netfix et Spotify pour les films et la musique. De nombreux secteurs d'activité ont greffé des systèmes de recommandation sur leurs sites web transactionnels, en vue de garantir à leurs clients des expériences personnalisées où les produits et services similaires sont mis en avant, dans l'intention de stimuler les ventes. Ces systèmes ont été développés depuis plus de 20 ans, notamment dans le domaine du commerce électronique.

Notre terrain de thèse est l'Institut Marocain de l'Information Scientifique et Technique (IMIST), disposant d'une unité documentaire physique, qui a pour vocation de mettre à la disposition des milieux scientifiques l'information scientifique et technique dont ils ont besoin pour développer leurs activités. Pour y parvenir, la personnalisation des catalogues en ligne par le biais d'un système de recommandation devrait plaider pour plus de précision, de pertinence et d'optimisation du temps de recherche quant à la sélection et l'accès à l'IST par les chercheurs.

Dans les années 1990, le groupe de travail international et multidisciplinaire DELOS/NSF[1] a travaillé sur les enjeux techniques liés à l'arrivée des bibliothèques numériques, notamment les systèmes de personnalisation ont été évoqués par un comité, où l'accent a été mis sur la complexité des bibliothèques numériques et l'actualité brûlante de l'intégration des données recueillies sur les utilisateurs.

Pour ma part, j'ai commencé à m'intéresser à la question de la personnalisation en recherche de l'information dans le cadre de mon mémoire de master. À cette occasion, j'ai rencontré une dizaine de décideurs dans le secteur des télécommunications au Maroc, pour les interroger sur leurs comportements informationnels en matière de surveillance de l'environnement. Le constat était celui que la majorité de ces décideurs avait un besoin de personnalisation du dispositif de surveillance de l'environnement de leur entreprise, en fonction de leurs intérêts changeants et de leur profil.

Par ailleurs, cette problématique fait appel à plusieurs disciplines de recherche : la recherche de l'information, l'apprentissage automatique et l'interface homme-machine.

Dans quelle mesure les systèmes de recommandation pourraient-ils personnaliser et améliorer la recherche et l'accès aux ressources documentaires et informationnelles pour les utilisateurs des structures documentaires ?

I. Problématique de recherche : objectifs et questions de recherche

Les récents progrès des technologies de l'information et des réseaux de communication ont donné à l'information de nouveaux contours, tout en reconnaissant sa valeur scientifique, technique et d'usage. De surcroît, les progrès techniques de numérisation et de partage ont facilité sa production, son exploitation et sa circulation.

Au Maroc, les chercheurs sont confrontés à un environnement qui n'aide pas au développement de la recherche scientifique, en commençant par les faibles budgets qui y sont alloués, en

[1]LAFORCE, P. ; RATTE, S. 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), 2018.

passant par les restrictions financières et techniques d'accès et d'exploitation de l'information scientifique et technique.

La recherche d'information documentaire est une activité qui a pour objectif de permettre à l'utilisateur d'obtenir des documents pertinents, de manière à ce qu'ils répondent à ses besoins d'information. La phase d'indexation du corpus des documents produit des représentations par un système informatique. L'utilisateur formule par ailleurs sa requête, et cette dernière est interprétée afin d'obtenir une représentation de la requête qui sera confrontée à la représentation des documents[2].

Dans le cadre du modèle basique de la recherche d'information : un utilisateur exprime son besoin sous forme de requête et le résultat retourné par le système demeure identique, quel que soit l'utilisateur. En effet, le contexte de l'utilisateur n'est pas pris en compte, ainsi que la particularité de la requête.

Au demeurant, les outils de mesure d'appariement tentent de mettre en relation le besoin informationnel de l'utilisateur avec le corpus documentaire disponible, et d'aider le système de recherche d'information à cibler les documents potentiellement pertinents pour cet utilisateur. Le mécanisme de mesure d'appariement le plus répandu est celui de la similarité, qui évalue le degré de ressemblance entre les représentations des documents et la représentation de la requête. La recherche sur le catalogue en ligne d'un système intégré de gestion des bibliothèques ne permet pas une précision en termes de résultats dans les deux modes de recherche (simple ou avancée), l'utilisateur se trouve face à un nombre important de notices listées dans plusieurs pages de résultats à l'image d'un moteur de recherche.

Pour pallier ces problèmes, il existe les systèmes hypermédias adaptatifs visant à personnaliser les contenus en fonction du contexte de l'utilisateur, comme les systèmes de recommandations en plein essor dans de nombreux domaines (les sites d'e-commerce, de musique, de vidéo, de presse, des jeux etc.).

Les services d'information documentaire destinés aux chercheurs revêtent une importance particulière, ils contribuent à la qualité de la production scientifique et technique, notamment

[2] BACCINI, Alain ; DEJEAN, Sébastien, DESIRE KOMPAORE, Nongdo [et al.]. Analyse des critères d'évaluation des systèmes de recherche d'information. *Technique et Science Informatiques*, Hermès Science Publications, 2010.

pour l'Institut Marocain de l'Information Scientifique et Technique (IMIST), centre documentaire destiné aux chercheurs. Ces derniers se trouvent confrontés aux problèmes évoqués ci-dessus sur la pertinence de l'information proposée, situation dégradant leur démarche de recherche et impactant leur satisfaction des services.

Notre thèse vise à contribuer à répondre à ce besoin en proposant la conception et la mise en œuvre d'un système de recommandation adapté à un service d'information documentaire scientifique.

Concrètement, notre modèle de recommandation sur les catalogues en ligne sera développé à partir des logs d'usage des utilisateurs chercheurs, représentant des données implicites plutôt qu'à partir de leurs appréciations explicites majoritairement sollicitées dans les systèmes de recommandation déployés dans le contexte des industries culturelles et créatives. Ce choix des données collectives d'usage réfère à l'importance des communautés de recherche et a fortiori à l'importance des communautés de pratiques scientifiques.

Par ailleurs, dans cette introduction, il nous semble important de préciser la dimension « big data » sous-jacente à notre travail et à la proposition du système de recommandation qui en tiendra compte, dimension qui renouvelle les services liés au repérage des contenus dans de nombreux secteurs.

Notre thèse s'ancre donc dans le mouvement big data pour les considérations suivantes :

- Le big data et l'intelligence artificielle sont deux technologies liées, notre proposition d'un modèle et prototype de recommandation repose sur des algorithmes qui sont, aujourd'hui, capables de traiter des données en temps réel. Les systèmes de recommandation sont des dispositifs d'apprentissage automatique dont l'efficience dépend de la quantité de données traitées. Autrement dit, le big data alimente les algorithmes de recommandations par des données susceptibles de permettre à la machine d'apprendre des règles, pour agir de manière automatique. Le big data permet d'accélérer la courbe d'apprentissage d'un système de recommandation : plus il reçoit de données, plus il apprend et devient précis.

- La contribution de notre thèse, quoiqu'elle soit inscrite dans le contexte d'un jeu de données *log* ne couvrant qu'un échantillon de 500 utilisateurs chercheurs (ce qui ne vérifie pas le premier V (volume) du big data) vise à permette les évolutions que connaîtra le prototype par la suite,

avec l'augmentation du nombre des utilisateurs et de leurs données. C'est pourquoi, notre contribution intègre, d'un point de vue technique, la mise en place d'un environnement big data, composé d'un dispositif de stockage distribué et d'un outil de traitement de données. Ces deux composants sont à la base du stockage des jeux de données *log* d'entrée et de l'entraînement de notre modèle de recommandation.

- La dimension du temps réel est un autre défi « big data » auquel tente de répondre notre thèse, puisque dans un catalogue en ligne d'un SID, les interactions des utilisateurs avec ce catalogue peuvent être continues et en évolution régulière, tracées par des quantités de données importantes, qui seront exploitées pour créer des expériences sur-mesure et personnalisées grâce au système de recommandation proposé. Dans cette perspective, il fallait prévoir que les technologies mobilisées soient en mesure de traiter correctement ce volume promis à une croissance exponentielle.

Nous allons désormais préciser les objectifs et questions de recherche, présenter les méthodologies de travail utilisées et résumer la structuration de notre thèse.

L'objectif principal de la thèse est de développer un service de personnalisation du catalogue en ligne, représentant un vecteur majeur de l'information scientifique et technique des chercheurs. Cet objectif doit conduire à la modélisation et au prototypage d'un système de recommandation destiné aux chercheurs de l'Institut Marocain de l'Information Scientifique et Technique (IMIST) du Maroc.

Mais cet objectif final ne peut être poursuivi sans une contextualisation plus large de l'évolution des services d'accès à des contenus et une contextualisation plus large de l'évolution des données disponibles et de leur valorisation en services spécifiques dans les SID. Ainsi, à notre objectif final, et pour valider sa pertinence, nous avons également voulu inscrire dans cette thèse deux autres objectifs qui vont permettre de mettre en contexte de façon diachronique et comparative l'évolution des services dédiés à l'accès à des contenus pour l'utilisateur. Tous ces services sont en prise avec l'émergence de types de données particuliers, disponibles à un moment donné.

Nos trois objectifs seront donc les suivants :

Objectif 1 : apprécier la valorisation de la data de façon transversale aux industries culturelles et étudier son rôle dans la proposition de nouveaux services innovants aux utilisateurs ;

Objectif 2 : mettre en évidence, pour les SID, la valorisation de la data jusqu'à présent et marquer la nouvelle dimension qu'ouvre le big data ;

Objectif 3-principal : apprécier les services existants de l'IMIST et la valorisation de la data faite jusqu'à présent par les professionnels de l'information et de documentation, et proposer un système de recommandation associé au catalogue en ligne pour personnaliser les interactions de chaque utilisateur avec le fonds documentaire.

Les questions de recherche sont des énoncés interrogatifs, qui permettent de formuler et d'expliciter les objectifs fixés. Elles visent à guider et surtout opérationnaliser la recherche, en déterminant les champs d'intervention de la thèse.

Pour atteindre nos objectifs, nous proposons de répondre aux questions de recherche suivantes :

Objectif 1 : apprécier la valorisation de la data de façon transversale aux industries culturelles et étudier son rôle dans la proposition de nouveaux services innovants aux utilisateurs :

- ✓ Quels sont les secteurs composant les industries culturelles ?
- ✓ Quelle est la nature de la data exploitée dans différentes plateformes numériques des industries culturelles ?
- ✓ De façon comparative, quels usages est fait de la data dans les différents secteurs des industries culturelles ?
- ✓ Quels sont les enjeux de l'utilisation de la data dans les industries culturelles ?

Objectif 2 : mettre en évidence, pour les SID, la valorisation de la data jusqu'à présent et marquer la nouvelle dimension qu'ouvre le big data :

- ✓ Quelle est la nature de la data exploitée dans les SID ?
- ✓ A quel niveau la data sert à éclairer les comportements des utilisateurs et leurs attentes dans les SID ?
- ✓ Quel est l'apport des données structurées et non structurées aux services fournis aux utilisateurs des SID ?
- ✓ Quelle valeur apporte le web de données aux SID ?
- ✓ Quels sont les défis posés par le big data dans les SID ?

- ✓ Dans quelle mesure le big data contribue à l'innovation et fait évoluer le rôle des professionnels de l'information dans les SID ?

Objectif 3- principal : apprécier les services de l'IMIST existants et la valorisation de la data faite jusqu'à présent par les professionnels de l'information et de documentation, et proposer un système de recommandation associé au catalogue en ligne pour personnaliser les interactions de chaque utilisateur avec le fonds documentaire.

- ✓ Quelles sont les composantes de l'offre d'information scientifique et technique de l'IMIST ?
- ✓ Quelles sont les plateformes numériques permettant aux chercheurs de l'IMIST d'accéder à cette offre ?
- ✓ Quelles sont les données cumulées sur les utilisateurs chercheurs ?
- ✓ Sont-ils satisfaits des dispositifs déployés pour l'accès et l'exploitation de l'offre ?
- ✓ Quel modèle de système de recommandation à concevoir sur la base des préférences des utilisateurs en général et des chercheurs en particulier de façon à leur permettre une meilleure expérience sur les catalogues en ligne ?
- ✓ Quelles données disponibles ? Quelle approche choisir et quels mécanismes d'implémentation technologiques à mettre en place ?
- ✓ Quelle évaluation d'un tel système de recommandation pouvons-nous réaliser ?

II. Les méthodes, cadre conceptuel et contraintes de recherche

Afin d'atteindre les objectifs de recherche fixés, nous avons sollicité pour chacun d'eux des méthodes de recherche correspondantes :

Objectif 1 : Nous mobilisons pour cet objectif la méthode documentaire et la dimension comparative, qui conduit à réaliser une synthèse d'information sur le sujet de la data dans les industries culturelles et à dresser une analyse comparative de l'usage, des enjeux et des services crées ou optimisés à partir de la data dans différents secteurs.

Objectif 2 : Nous mobilisons également pour cet objectif la méthode documentaire pour réaliser une synthèse des principales dimensions de la data dans les SID, afin de cerner les principaux enjeux et usages déployés à partir de la data structurée et non structurée dans les SID. Une vision diachronique est visée pour cet objectif, le « big data » en constitue une nouvelle étape.

Objectif 3 : Dans un premier temps, nous mobilisons la méthode d'enquête, qui vise à obtenir des informations précises sur un sujet donné au moyen l'enquête du terrain par le biais d'un mini-questionnaire administré auprès de professionnels de l'information de l'IMIST, et dont l'objectif est d'étudier l'usage de la data dans l'entité documentaire de recherche en question. Puis dans un second temps, après avoir précisé le cadre conceptuel des systèmes de recommandations, nous mobilisons une méthode constructiviste d'ingénierie qui procède à l'élaboration d'un modèle pour la conception, puis la réalisation d'un prototype informatique qui sera finalement soumis à une première évaluation. Des méthodes statistiques quantitatives d'apprentissage basées sur les données de logs d'utilisateurs sont alors convoquées.

Nous terminons cette introduction, en précisant certaines contraintes que nous avons rencontrées pour conduire ce travail :

- La rareté des données et des travaux académiques sur ce sujet de recherche dans le monde francophone d'une manière générale et dans le contexte du Maroc en particulier ;
- La complexité des algorithmes qui sont à la base des systèmes de recommandation ;
- Une certaine difficulté à définir et à disposer des données d'entrée servant à alimenter un service de recommandation dans le contexte étudié.

III. La structure de la thèse

Notre thèse est structurée de la manière suivante :

- Une première partie sur la data dans les industries du contenu : analyse comparative de points saillants, elle comporte deux chapitres :

Chapitre 1 : il délimite le champ des industries culturelles et appréhende leur transformation portée par le numérique et le big data.

Chapitre 2 : il rend compte de l'usage des données dans les secteurs constituant les industries culturelles dans une optique comparative (édition, médias, musique et e-learning).

- Une deuxième partie : la data dans les services d'information documentaire : nouveaux services, elle comprend deux chapitres :

Chapitre 1 : il met en relief les apports des données structurées et non structurées à la classification, aux possibilités de recherche et autres services pour les utilisateurs.

Chapitre 2 : il traite du rôle de la data dans l'optimisation des services d'information documentaire à l'ère du big data, aussi bien au niveau des professionnels de l'information et de la documentation, qu'au niveau des utilisateurs.

- Une troisième partie : les SR et leurs développements, elle regroupe trois chapitres :

Chapitre 1 : il appréhende la modalisation des systèmes de recommandation au niveau des processus et représentation des utilisateurs.

Chapitre 2 : il précise les différentes approches des processus de recommandation (à base de contenu, à base de filtrage collaboratif, hybride), en expliquant le fonctionnement, les avantages et les limites de chaque approche.

Chapitre 3 : il traite des algorithmes de recommandation à base de filtrage collaboratif, le plus répandu dans les systèmes de recommandation et qui sera mobilisé dans notre modèle fonctionnel et technique. Le chapitre présente aussi les fonctions d'évaluation des systèmes de recommandation, en citant certains exemples.

- Une quatrième partie : Etude de cas, conception et modélisation d'un SR pour le catalogue de l'IMIST, elle comporte cinq chapitres :

Chapitre 1 : il présente notre terrain expérimental, à savoir l'Institut Marocain de L'information Scientifique et Technique (IMIST), en présentant son offre et ses ressources documentaires et informationnelles.

Chapitre 2 : il identifie les besoins des professionnels de l'information et de la documentation de l'IMIST à travers un questionnaire dont l'objectif est de connaître l'usage actuel des données et de valider le besoin de personnalisation d'accès aux ressources documentaires. Il présente également les grandes lignes de notre prototype de recommandation, en comparaison avec d'autres SR des SID.

Chapitre 3 : ce chapitre est consacré à la conception du modèle de recommandation proposé pour optimiser l'accès et l'exploitation des ressources documentaires pour une population constituée de la moitié des chercheurs adhérents à l'IMIST au titre de l'année 2015. Ce modèle prend en entrée des données *log* implicites recueillies à partir de l'interaction des utilisateurs avec le système et repose sur la classification supervisée basée sur l'algorithme des K plus proches voisins (*K Nereast Neighbors-KNN*) et la factorisation matricielle par l'approche de recommandation de filtrage collaboratif.

D'un point de vue opérationnel, le modèle utilise l'extraction des données implicites collectées sous forme de fichiers logs, intègre une phase de nettoyage puis de transformation en données explicites exprimées par des plages de notes. Deux scénarii sont envisagés : celui d'un utilisateur authentifié et celui d'un utilisateur anonyme.

Chapitre 4 : il présente l'architecture technique de la solution et la mise en œuvre du prototype par le biais de l'algorithme de recommandation (ALS) adapté à notre modèle relevant de la

bibliothèque (Mllib) contenant des algorithmes d'apprentissage automatique sur l'environnement Spark.

Chapitre 5 : il soumet le prototype à plusieurs évaluations : le calcul de fonctions d'erreur et l'évaluation explicite par des utilisateurs sous forme de questionnaire.

Nous terminerons en soulignant certaines limites du prototype réalisé et en dressant quelques perspectives à envisager.

Partie I : La data dans les industries culturelles : analyse comparative de points saillants

Chapitre 1 : Les transformations liées aux données dans les industries culturelles

Introduction

Les industries culturelles et de contenu sont en plein "choc de données". La personnalisation de l'expérience devient une tendance majeure, le rapprochement entre un contenu et l'expérience du spectateur, lecteur, joueur est plus grand.

La valorisation de la data conduit aujourd'hui à la reconsidération de la valeur client dans les industries culturelles et de contenu, notamment pour relever le défi de la prescription pour les publics et les audiences, la dissémination des contenus et le sponsoring/financement de la création.

1. Industries culturelles et nouveaux enjeux

L'expression "industrie culturelle" a vu le jour à travers les travaux d'Adorno et de Horkheimer (1947)[3] face aux menaces appréhendées de l'application des techniques de reproduction industrielle à la création et à la diffusion massive des œuvres culturelles.

En 1947, Adorno et Horkheimer ont examiné la standardisation du contenu et la prédominance de la recherche de l'effet qui résultent de l'application des techniques de reproduction industrielle à la création culturelle.

Guibert définit les industries culturelles comme « des biens culturels (musiques, images, films, livres) sont un type de biens informationnels dont les caractéristiques diffèrent selon le point de vue de l'acteur concerné : créateur, utilisateur ou éditeur. Dans la perceptive des créateurs, les biens culturels sont reproductibles. Malgré leur nature intangible, ces biens sont copiés et stockés sur des supports physiques (comme des CD, pour la musique, ou des livres) une fois l'original créé »[4].

[3] ADORNO, Theodor et HORKHEIMER, Max. *Raison et mystification des masses*, Allia, 2012, 104 p., traduction d'Eliane Kaufholz, ISBN : 978-2-84485-436-0.

[4] GUIBERT, Gérôme ; REBILLARD, Franck et al. *Médias, culture et numérique. Approches socioéconomiques*, Paris, Armand Colin, coll. « Cursus », 2016, 237 p., ISBN : 978-2-200-61454-6.

Au début, les processus de standardisation et de reproduction industrielle concernaient juste le cinéma, alors que les autres secteurs culturels étaient marqués par la prédominance d'une production de type artisanal avec une individualisation de l'œuvre. A la fin des années 70, la situation a profondément évolué. De nouveaux médias se sont développés, au premier rang desquels la télévision, et la marchandisation de la culture est devenue courante. Aussi, le concept est utilisé au pluriel plutôt que le singulier, désignant, par-là, une pluralité des secteurs économiques davantage qu'un processus unique.

Le champ des industries culturelles n'englobe pas toute la culture. Le terme industrie exclut les beaux-arts et l'architecture ainsi que le spectacle vivant, pour ne retenir que les biens faisant l'objet d'un processus industriel de production permettant la reproduction à partir d'un prototype. De fait, le cœur des industries culturelles est composé des entreprises d'édition (édition de livres, de presse, de phonographie ou de multimédias) et de production de films (industrie cinématographique) ou de contenus audiovisuels (ces derniers étant destinés majoritairement aux diffuseurs à la télévision).

Les données personnelles revêtent ainsi une importance particulière, puisqu'elles permettent entre autres d'optimiser le temps de visite sur un site web ou sur un site réel, de comprendre les comportements et prédire les attentes des utilisateurs. Une opportunité pour les secteurs de la presse, du cinéma ou du jeu vidéo d'enrichir leur offre tout en optimisant les coûts. En outre, le big data permet une vitesse importante dans l'analyse et la génération des données, profitant ainsi de l'évolution des nouveaux outils d'analyse de ces données.

Par ailleurs, Bustamente a précisé que les transformations des industries culturelles résultent principalement de l'adoption généralisée des technologies. Or, il est important de voir les transformations actuelles par rapport aux médias traditionnels et des structures organisationnelles. Il argue « La politique de communication a un défi majeur, pour tenir en compte les besoins du public face aux enjeux dictés par le mercantilisme » [5].

Les appareils connectés aux réseaux de téléphonie mobile et à internet sont des sources de données majeures. Ainsi, les smartphones, les tablettes et les ordinateurs émettent des données

[5]BUSTAMENTE, E. Cultural Industries in the Digital Age: Some Provisional Conclusions. *Media, Culture and Society*.2004, Vol. 26, n°6, p. 803-820.

relatives à leurs utilisateurs lors de la navigation internet, l'utilisation des moteurs de recherche, les messages laissés sur les réseaux sociaux, le téléchargement et l'utilisation d'applications, la publication en ligne de photos et vidéos, les achats sur des sites de vente en ligne, etc.

Outre les objets connectés, les données proviennent d'autres sources telles que : les données démographiques, les données scientifiques et les données liées à la fréquentation des lieux publics, etc. Toutes ces données donnent des informations sur les usagers, leurs appareils, leurs déplacements, leurs centres d'intérêt, leurs habitudes de consommation, leurs loisirs et leurs projets, etc…L'augmentation permanente du nombre des utilisateurs d'internet et de téléphones mobiles fait que le volume des données numériques croît de manière fulgurante[6].

Les industries culturelles font intervenir plusieurs acteurs (producteurs, fabricants, créateurs, annonceurs, diffuseurs et consommateurs) qui utilisent des outils divers, selon la nature de chaque acteur, par exemple les producteurs mobilisent des contrats, alors que les consommateurs utilisent des catalogues de contenus, la personnalisation et les recommandations, comme l'illustre la figure suivante[7] :

[6] Ibid., p.31

[7] CNIL (Commission Nationale de l'Informatique et des Libertés). Industries créatives, contenus numériques et données : les contenus culturels vus au travers du prisme des données Le graal de la recommandation et de la personnalisation. *CAHIERS IPINNOVATION & PROSPECTIVE* [en ligne].2015, n°3. Disponible sur https://www.cnil.fr/fr/innovation-prospective/publications [Consulté le 25 juin 2016].

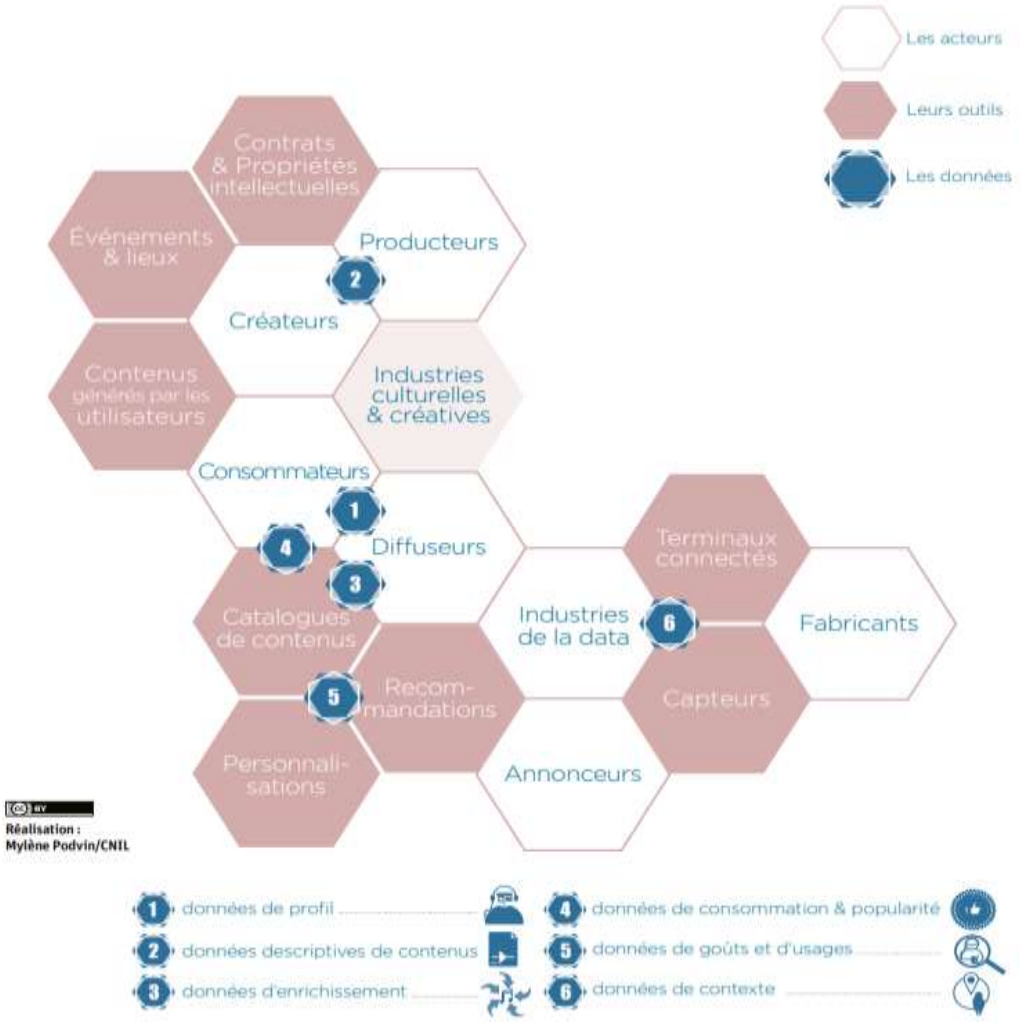

Figure 1 : L'écosystème des industries culturelles : acteurs, outils et l'usage des données (CNIL, 2015)

Les données utilisées dans les industries culturelles peuvent provenir de sources diverses, puisque certaines sont produites par les créateurs et producteurs, alors que d'autres sont enrichies par les diffuseurs, produites algorithmiquement ou générées par les utilisateurs à travers leurs traces d'interaction avec les systèmes (tags, playlists, notes, commentaires), ces données de création ou de consommation de contenus culturels et créatifs sont ventilées par catégorie[8] :

- Données personnelles : il s'agit des données d'identité, de contact et du profil sociodémographique ;

[8] Ibid., p.32.

- Données descriptives de contenu : elles regroupent les données de catalogage (artiste, auteur, éditeur), d'identification (genre, thème), les données techniques (format, compression) et les données juridiques (droits d'auteur ou d'accès) ;
- Données de consommation : elles renseignant sur les goûts collectifs des utilisateurs, comme la tendance ou la popularité d'une chanson, le nombre de consultations, le total des achats, les commentaires et retours sur les réseaux sociaux, etc. ;
- Données d'enrichissement : elles comprennent des données d'approfondissement, telles que les biographies, mes critiques, les notes et évaluations, etc. ;
- Données concernant l'usage de chaque utilisateur : elles peuvent être très synthétiques (type, quantité, historique de consultations...), mais aussi très détaillées (les passages surlignés d'un livre, la vitesse de lecture…) ;
- Des données de contexte : il s'agit de l'horodatage relatif à l'enregistrement de l'instant durant lequel une action a été effectuée, la localisation ou toutes les données issues de capteurs (les émotions, l'état physiologique, l'humeur, etc.).

2. Numérique, données et économie de la culture

La croissance des industries culturelles et créatives est de plus en plus liée à la transformation numérique de ce secteur. Les revenus du secteur ont augmenté de plus de 22 milliards d'euros entre 2003 et 2013, comme le montre la figure 2 [9] :

[9] EY FRANCE. *Données personnelles et Big data*. Disponible sur : http://www.ey.com/FR/fr/Industries/Media---Entertainment/Comportements culturels-et-donnees-personnelles-au-coeur-du-Big-data>. [Consulté le 22 Avril 2015].

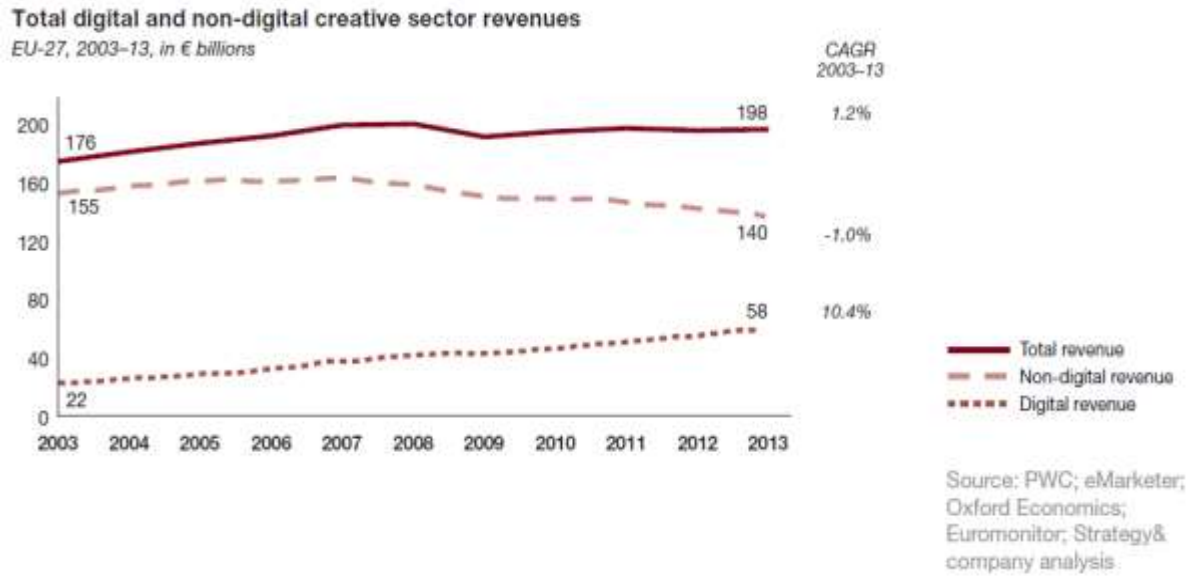

Figure 2 : Les revenus des secteurs numériques et non numériques en Europe 2003-2013 (EY France, 2015)

Alors que les revenus liés au non-numérique ont connu une baisse de plus de 14 milliards d'euros sur cette période (à 142,70 milliards d'euros), la composante numérique a enregistré une augmentation de 36,7 milliards d'euros (avoisinant 59,3 milliards d'euros)**.** Les consommateurs ont majoritairement tendance désormais à trouver des contenus médiatiques et de divertissement sur internet et à accéder à des produits en format numérique comme un téléchargement de film, un enregistrement audio ou un jeu[10].

Le consommateur se voit proposer plusieurs services, de plus en plus diversifiés et sur mesure à travers l'augmentation du nombre de chaînes de télévision, des vidéos en ligne auto-diffusées, des catalogues de musique en ligne, des plateformes de streaming ou de téléchargement, des recommandations variées.

Le temps consacré aux médias en ligne sur Internet a pris une place importante dans la journée de l'internaute, tout en se diversifiant car les media consultés hier ne sont pas forcément ceux d'aujourd'hui**.** La figure 3[11] illustre cette tendance (enquête européenne de 2013 déjà ancienne).

[10] Ibid., p.34.

[11] EY FRANCE. *Données personnelles et Big data*. Disponible sur http://www.ey.com/FR/fr/Industries/Media---Entertainment/Comportements culturels-et-donnees-personnelles-au-coeur-du-Big-data>. [Consulté le 22 Avril 2015].

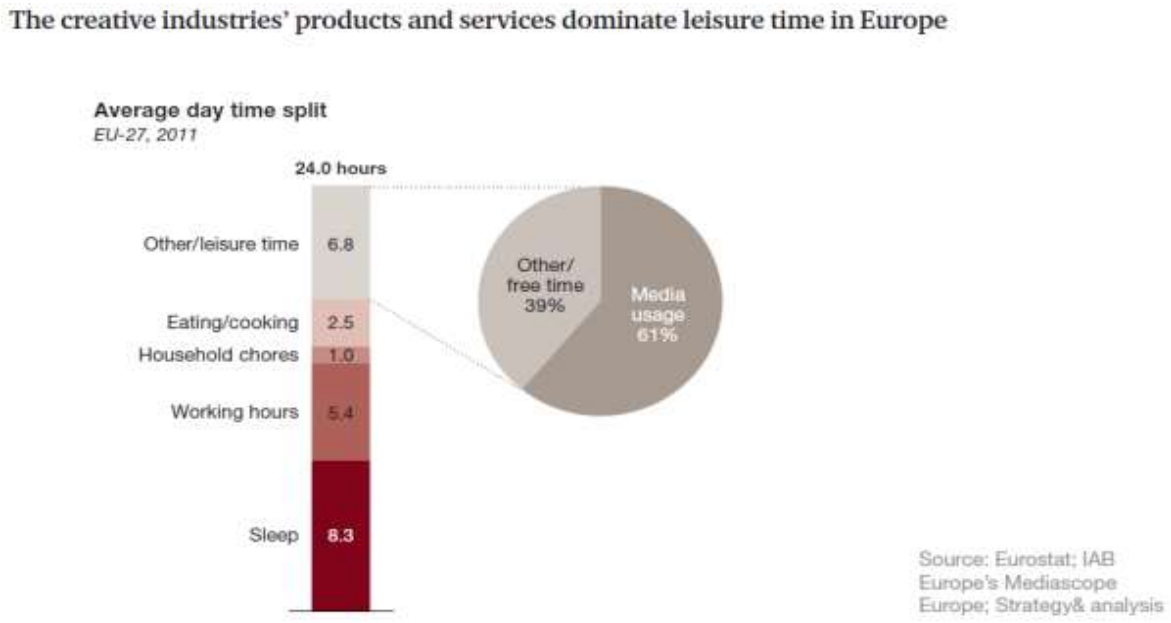

Figure 3: La place des produits culturels dans le temps des activités quotidiennes des citoyens en Europe (EY France, 2015)

Il ressort de la figure ci-dessus que l'usage des industries culturelles en général et plus particulièrement les médias s'est fait une place importance dans le quotidien des européens. Ainsi, les médias représentent 61% du temps quotidien consacré aux loisirs.

Les technologies convoquées par les industries culturelles à l'ère du numérique sont à la fois transversales à l'Internet (les architectures client-serveur, cloud, big-data, etc.) et spécifiques aux contenus (les standards attachés aux contenus et les technologies d'accès, comme les moteurs de recherche et les réseaux sociaux).

Lire un journal en ligne, regarder un film en streaming, visiter une exposition virtuelle, autant de pratiques culturelles impulsées par la transformation numérique de la culture, tenant compte du public et du modèle économique, comme le précise Chartron « La transition du statut de prototype innovant vers des services durables nécessite, d'une part, la rencontre d'un public à une échelle significative et, d'autre part, la consolidation d'un modèle économique pérenne »[12].

[12] CHARTRON, Ghislaine. *Cours l'économie de l'information, les différentes filières des industries culturelles, transformation numérique*, 2019.

Pour Chantepie [13] la production des contenus culturels s'est transformée à travers la dématérialisation des supports physiques (CD, DVD, papier, etc.), qui permet de reproduire et diffuser sur des réseaux marqués par des progrès des débits, ce qui traduit la mutation de l'infrastructure technique de distribution, induisant des modifications sur l'ensemble des chaînes de valeurs et une recomposition des modèles économiques, conduisant à la déflation des prix unitaires des biens culturels numériques.

Dans le même ordre d'idée, l'économie du numérique a transformé les industries culturelles de par ses piliers de gratuité, de l'accès, des services et de collaboration. En effet, la chaîne de valeur de ces industries a intégré le marketing digital, qui a renouvelé la relation client tant sur le plan des interfaces homme-machine que l'architecture de l'information, afin de garantir la performance de l'accès à l'information, tout en mettant en avant l'expérience de l'utilisateur, matérialisée par la combinaison de facteurs relationnels et de facteurs émotionnels attachés aux interfaces numériques et les dimensions de l'engagement des utilisateurs/clients, la découvrabilité, la segmentation, l'éditorialisation et l'image-ination des contenus[14].

3. Le big data et l'économie de la culture, nouvelle donne

Lorsque que nous parlons du big data culturel, nous considérons très souvent des données très variées, qu'elles soient structurées ou non structurées comme par exemple la matrice de chiffres représentant les liens entre profils d'un réseau social.

De nombreuses entreprises ont su exploiter les changements d'échelles et de paradigmes adossés à la prolifération de la data, pour créer de nouveaux services. Il s'agit d'un marché à forte croissance qui s'oriente vers des valorisations très diversifiées des données. Le marché mondial du big data était estimé à 6.3 milliards de dollars en 2012. C'est encore un marché

[13] CHANTEPIE, Philippe ; LE DIBERDER, Alain. « IV. L'exploitation numérique : tout change », Philippe Chantepie éd., *Révolution numérique et industries culturelles.* La Découverte. 2010, pp. 55-72.

[14] CHARTRON, Ghislaine. *Edition et publication des contenus : regard transversal sur la transformation des modèles*. Lisette Calderan, Pascale Laurent, Hélène Lowinger, Jacques Millet. Publier, éditer, éditorialiser, nouveaux enjeux de la production numérique, De Boeck Supérieur, pp.9-36, 2016, Information & Stratégie. Disponible sur https://halshs.archives-ouvertes.fr/halshs-01522295/document [Consulté le 06 Août 2020].

jeune, mais en très forte croissance avec des revenus qui progressent de +40% en moyenne par an.

La figure ci-dessous[15] met l'accent sur la valeur élevée de la donnée personnelle culturelle numérique issue de plusieurs acteurs (utilisateurs, agrégateurs de contenu, infrastructures et établissements culturels) cadrés par différentes réglementations (notamment le RGPD en Europe), des processus de sécurité et de certification dessinant les contours d'un environnement de confiance et ouvrant les perspectives de grands investissements stratégiques.

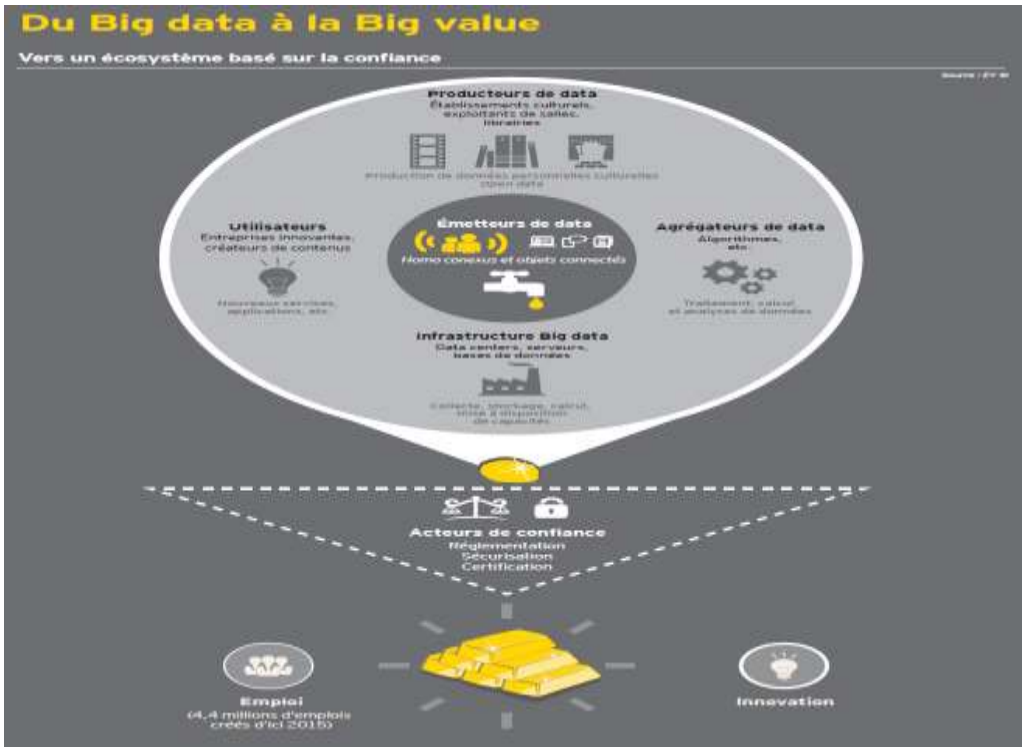

Figure 4 : Le big data culturel (EY France, 2015)

Mais nous pouvons poser la question suivante : à partir de quel volume parlons-nous de big data culturel ? La réponse n'est pas aisée car la frontière est floue. Néanmoins, nous évoquons le big data dans des contextes où le volume de données à gérer est hors du commun, à travers la prolifération des données d'usage, qui renseignent, selon Katz-Gerro, sur les goûts

[15] Ernst and Young France. *Données personnelles et Big data*. Disponible sur http://www.ey.com/FR/fr/Industries/Media---Entertainment/Comportementsculturels-et-donnees-personnelles-au-coeur-du-Big-data> [consulté le 22 Avril 2015].

les valeurs, les attitudes et les comportements dans les industries culturelles, en s'appuyant sur des méthodologies (variété, contexte et intégration) et des théories (production, consommation, politiques publiques et identités)[16].

Le secteur des industries culturelles est par ailleurs marqué par des phénomènes de concentration accélérés autour des GAFAM[17] et autres acteurs internationaux tels que Netflix opérant sur des marchés internationaux et non soumis aux mêmes réglementations fiscales que les acteurs nationaux.

Sur un autre registre, le croisement du big data et l'économie de la culture affecte le modèle économique. Ainsi, si le prix d'un livre ou d'un album musical change deux ou trois prix différents lors de son cycle de vie de vente, le big data a intégré la flexibilité liée à une définition dynamique du prix d'un bien culturel, et qui varie en temps réel, en fonction de la demande des utilisateurs.

De fait, le big data aide à une gestion des prix en temps réel, selon les données disponibles sur chaque utilisateur, de ses goûts, du moment de découverte et de consommation et de l'écran utilisé, etc.

La Maison Blanche a publié un rapport sur la corrélation entre le big data et la définition dynamique des prix des biens culturels, où elle cite incidemment une étude de Benjamin Shiller qui estimait que si Netflix utilisait des données comportementales pour personnaliser les prix de son abonnement, les profits supplémentaires pourraient atteindre jusqu'à +12%15[18].

[16]KATZ-GERRO, Tally. Cultural Consumption Research: Review of Methodology, Theory, and Consequence. *Internationational Review of Sociology : revue internationale de sociologie*. 2004, Vol.14, n° 1, p. 11-29.

[17] Il s'agit de cinq géants du web : Google, Amazon, Facebook, Apple et Microsoft.

[18] CNIL (Commission Nationale de l'Informatique et des Libertés). Industries créatives, contenus numériques et données : les contenus culturels vus au travers du prisme des donnéesLe graal de la recommandation et de la personnalisation. *CAHIERS IPINNOVATION & PROSPECTIVE* [en ligne], n°3 ,2015. Disponible sur https://www.cnil.fr/fr/innovation-prospective/publications [Consulté le 25 Juin 2016].

Conclusion

L’entrée de la culture dans l’ère du big data est certainement assimilable à certaines formes d’accélération de ces processus d’industrialisation où il va s’agir tout particulièrement de segmenter finement les offres et les produits au client.

La réglementation sur la protection des données personnelles est aussi un sujet vif dans ce secteur, avec des contextes souvent asymétriques entre pays, alors que l’exploitation des traces numériques est au cœur du développement des nouveaux services. Par ailleurs cette concentration des acteurs est accélérée par la nécessité de développer des infrastructures techniques pour gérer ces masses de données à l’international : datacenters, environnements de développement propres aux données non structurées, etc.

Chapitre 2 : La typologie de la data dans les différents secteurs culturels

Introduction

L'usage des données dans les industries culturelles (musique en ligne, vidéos à la demande, livre numérique et jeux vidéo) prend plusieurs formes, afin de configurer la création en fonction des besoins et goûts supposés des consommateurs.

Les industries culturelles et de contenu produisent et gèrent des données, qui interviennent à toutes les étapes de la production et de la diffusion des contenus, grâce à des outils et des techniques nourris par de larges corpus de données.

1. Les données structurées et non structurées

Plusieurs contenus ont été publiés sur le web. Outre les informations scientifiques, se sont développés des contenus journalistiques, fictionnels, musicaux ou encore audiovisuels, selon des modalités extrêmement diverses[19].

Les données structurées concernent des données structurées sémantiquement comme une notice bibliographique qui recense les informations, telles que les auteurs, le titre, la date de publication, le contenu d'une œuvre décrite en mots clés. Une œuvre elle-même peut désormais être considérée comme une donnée structurée quand elle est encodée dans un langage de type XML structurant le contenu en différents zones.

Au contraire les données non-structurées sont assimilables à des données de flux, qui se renouvellent rapidement, elles sont générées souvent par des interactions ou des capteurs, elles ne subissent pas de traitements sémantiques pensées en amont par l'homme ; par leur volume, elles sont donc assimilables à du big data. Les processus dynamiques de partages entre

[19] CHARTRON, Ghislaine et REBILLARD, Franck. *Modèles de publication sur le web*. Rapport d'activités ASCNRS 103, Jul 2004.

internautes génèrent aussi un flot de données non structurées qui portent de nouvelles valeurs exploitables dans des services inédits à grande échelle.

Caractéristiques des données structurées :

- Elles permettent d'ajouter des informations sémantiques qui aident notamment les robots à mieux comprendre les contenus culturels ;
- Nécessitent des traitements préalables à leur usage : tri, indexation et résumé etc. ;
- Elles proviennent désormais tant des bases de données internes qu'externes ;
- Les valeurs de ces données sont déterminées et connues à l'avance ;
- Les données sont stockées de manière organisée, facilement compréhensibles et repérables.

Nous ne pouvons pas parler de big data au sens originel du terme pour ce type de données.

Caractéristiques des données non structurées :

- Elles proviennent majoritaires d'interactions, de transactions continues, de capteurs, d'échanges entre internautes, à des échelles de masses critiques importantes ;
- Une valorisation majeure concerne l'alimentation par ces données d'algorithmes de prédiction et de recommandation personnelles provenant des traces numériques ;
- La personnalisation est un enjeu principal : proposition de contenus adaptés aux goûts et aux habitudes de chacun (playlists de Deezer ou Spotify, recommandations personnalisées de vidéos pour Netflix) ;
- Dans les jeux vidéo, les données incitent à acheter d'autres jeux, mais aussi à adapter le jeu en fonction des difficultés rencontrées par les joueurs ou des niveaux qu'ils apprécient (comme pour Steam) ;
- Pour les livres, les données d'usage aident notamment à mesurer le temps passé à lire le livre en relevant les passages les plus lus (c'est l'exemple d'Amazon qui arrive de la sorte à anticiper le potentiel de vente de certains ouvrages).

Ces données non structurées sont à considérer comme du big data [20] : citons par exemple l'émergence d'environnements immersifs qui déploient l'adaptation et la personnalisation des contenus culturels en temps réel par le biais des capteurs corporels (rendant compte de l'humeur, du degré du stress, des mouvements cardiaques de l'utilisateur etc.) pour ainsi proposer les contenus adaptés ou en concevoir de nouveaux plus innovants. Ces données de flux sont organisées en bases de données NoSQL (comme MongoDB, Cassandra ou Redis) qui implémentent des systèmes de stockage considérés comme plus performants que le traditionnel SQL pour l'analyse de données de masse (orienté clé/valeur, document, colonne ou graphe).

2. Data et les médias

La révolution digitale n'est pas uniquement le résultat du développement des infrastructures technologiques, mais elle concerne également la gouvernance et la valorisation des données au cœur de la convergence de plusieurs secteurs d'activité économique.

Les acteurs économiques s'intéressent aux données produites par leurs activités, pour tenter de mieux comprendre la valeur qu'ils apportent à leur écosystème. Chaque acteur doit maîtriser les bonnes pratiques du management des données structurées et non structurées associées à leurs activités opérationnelles ; les acteurs des industries culturelles, *a fortiori,* sont confrontés à de réels enjeux liés à l'exploitation et la valorisation de leurs données.

Toutefois, le manque de compétences techniques et une gouvernance de l'information en silos entravent souvent les initiatives initiées[21].

Dans la 15ème édition de l'étude annuelle « Global Entertainment & Media Outlook », sur les perspectives de l'industrie des médias et des loisirs, PwC prévoyait une croissance de 5 % en

[20] ROLLAND, Sylvain. Comment les algorithmes révolutionnent l'industrie culturelle. *La Tribune* [en ligne].2015. Disponible sur https://www.latribune.fr/technos-medias/comment-les-algorithmes-revolutionnent-l-industrie-culturelle-523168.html. [Consulté le 22 Août 2019].
[21] ALA-FOSSI, Marko., et al. The impact of the Internet on business models in the media industries. *The Internet and the Mass Media*. Los Angeles/London: SAGE, 2008, p.149–169.

moyenne par an entre 2013 et 2018. Les dépenses totales du marché sur le numérique (excluant l'accès à l'internet) progresseraient de 12,2% en moyenne par an entre 2013 et 2018 et représenter 65% de la croissance du marché. La publicité tient le haut du pavé et concerne directement l'exploitation des données. En 2018, 33% des revenus publicitaires mondiaux seraient digitaux, à comparer aux 17% des revenus consommateurs.

Les médias gèrent aujourd'hui une masse importante de données disponibles liées à leur audience, levier important de la compétitivité des entreprises. Le secteur des médias constitue également un terrain propice au développement de projets liés aux données, à cause de la multiplication des vecteurs de diffusion-réception connectés à internet (TV connectée, Set top box, mobile, web, tablette, objet connecté, etc…), le volume de données lié à cette consommation diversifiée ne cesse de croître.

Sur un autre registre, l'avènement du big data a révolutionné encore cette corrélation de la donnée et l'optimisation des services fournis au public, en jouant deux cartes principales :

Le contenu personnalisé :

L'accumulation des données importantes sur les consommateurs (intérêts, âge, genre…) déployées pour prédire et anticiper les comportements, pour ne faire acheminer à l'utilisateur que le produit adéquat, au bon moment au bon endroit.

Prenons un exemple, Netflix a produit la série "House of Cards" en fonction des goûts et centres d'intérêts de ses utilisateurs. Alors que le Huffington Post améliore l'expérience utilisateur de son site web moyennant le big data acquis et traité en temps réel.

L'Equipe fournit des offres sur mesure aux clients en utilisant des données sur les usages digitaux des consommateurs de la marque (abonnés et visiteurs gratuits). L'idée est de mieux vendre les produits existants (comme le papier, premier quotidien français) et d'innover.

La publicité ultra-ciblée :

Dans une logique de RTB (Real Time Bidding). Le big data permet de repérer les types de profils qui regardent quels programmes et quelles rubriques et ainsi ajuster les publicités. Ainsi si deux profils différents suivent un même programme, ils ne visionneront pas les mêmes spots publicitaires.

Ainsi, le projet lancé par le groupe Figaro vise à optimiser la vente de ses abonnements, avec une forte dimension publicitaire. A partir des recueils de données, il s'agit d'accompagner les annonceurs sur la connaissance client grâce au trafic généré.

Le digital représente une part grandissante dans les revenus des groupes de presse (50 % du CA pour le Figaro, environ 40 % pour l'Equipe)[22]. La publicité en ligne permet aux sites de gagner de l'argent et de diffuser des contenus gratuits et ceci grâce à la valorisation des données.

Les groupes de presse procèdent à l'analyse de leur lectorat sur la base des données relatives à la catégorisation socioprofessionnelle, ils achètent des données affinitaires et comportementales anonymes de plusieurs fournisseurs de données comme Google et Facebook[23].

3. Renouvellement de pratiques professionnelles : le data journalisme

Le data journalisme s'est affirmé ces dernières années comme une nouvelle pratique professionnelle mobilisant des compétences informatiques : « *S'il y a une révolution de toute façon elle ne vient pas de nous, journalistes, mais des pouvoirs publics qui pour leurs propres raisons mettent à disposition de plus en plus de données publiques. Grâce aux outils fournis par l'informatique, on peut beaucoup plus facilement transformer ces données et produire des nouveaux contenus* » (entretien avec un data journaliste pour un média *pureplayer,* 21 février 2017)[24].

En utilisant la data dans la presse, une nouvelle organisation des salles de rédaction a progressivement émergé. L'intégration de ces nouvelles sources d'information nécessite la maîtrise de processus d'identification, de collecte et de filtrage des données, de mise en forme et de restitution sous forme de data-visualisations et de scénarisations. Cette rencontre de la

[22] ZDNET-Social Media Club. *La data : un levier d'innovation pour les groupes de presse* [en ligne]. Disponible sur : https://www.zdnet.fr/blogs/social-media-club/la-data-un-levier-d-innovation-pour-les-groupes-de-presse-39835362.htm [Consulté le 12 Juin 2016].

[23]DYMYTROVA, Valentyna. *Data journalisme, entre pratique créative innovante et nouvelle médiation experte ? Une analyse conjointe des discours et des productions journalistiques.* XXI Congrès de la SFSIC Création, créativité et médiations, 2018.

[24] PAVLOVIC RIVAS, Marina. Médias et données : une influence sur la diffusion et la qualité de l'information ? .HEC Montréal | *« Gestion »*. 2017/1 Vol. 4, P. 80-81.

data avec la presse a donné naissance à de nouvelles pratiques axées sur trois niveaux de compétences : journaliste-designer-développeur (figure 5) [25].

L'exploitation des données dans le travail quotidien d'un data journaliste prend la forme d'une chaîne de traitement, qui consiste à récupérer et à intégrer les données, procéder à leur préparation et à leur nettoyage, puis les analyser pour les interpréter d'un ou de plusieurs points de vue possibles (calculs statistiques, croisement avec d'autres jeux de données, recherche de corrélations et visualisations intermédiaires). Dans la pratique, la préparation des données requiert notamment un recoupement des sources et des entretiens avec les spécialistes pour pouvoir statuer sur la pertinence des données d'entrée. Cette phase est chronophage. La créativité des data journalistes renvoie ensuite à leur capacité à créer des liens entre les données, à les croiser et identifier des relations originales qui n'étaient pas explicites auparavant.

La restitution revêt plusieurs formes allant de l'éditorial à la visualisation, en passant par les applications web. Néanmoins, la touche artistique et esthétique est sollicitée pour rendre ces données intelligibles et exploitables. De ce fait, les visualisations permettent des économies de temps par rapport à la lecture d'un article. De même, la personnalisation de la forme et du contenu au profil et au contexte de l'utilisateur peut être une orientation possible[26].

[25] CULTURE RP. *Les enjeux du data-journalisme pour la Presse : Part II. Data Driven Journalism Roundtable* [en ligne]. Octobre 2013. Disponible sur http://culture-rp.com/2013/10/17/les-enjeux-du-data-journalisme-pour-la-presse-part-ii/ [consulté le 20/06/2015].
[26] Ibid.

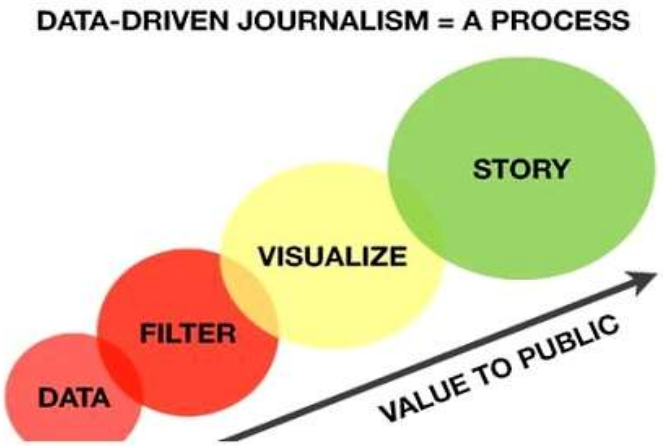

Figure 5: Le modèle de Data journalisme à trois compétences (CULTURE RP, 2013)

Les deux quotidiens de référence, précurseurs en matière de data-journalisme, sont *The New York Times* (USA) et *The Guardian* (GB), qui ont constitué des équipes réunissant cette triple compétence[27].

4. Data pour les médias : opportunité et défi

La data dans le secteur de la presse a permis à des groupes de coupler l'interactivité au journalisme d'analyse et d'investigation. A titre d'illustration, *The Guardian* a constaté que ses enquêtes du data-journalisme étaient les plus lues et sur lesquelles les lecteurs passaient le plus de temps en ayant la possibilité de télécharger des données brutes et d'interagir avec le media, ce qui contribue parallèlement à attirer les annonceurs et à générer de nouveaux revenus publicitaires[28].

Dans cette tendance, les groupes médias deviendraient des spécialistes de l'exploitation de données à la fois à travers le data-journalisme, mais également par la mise à disposition de

[27] CULTURE RP. *Les enjeux du data-journalisme pour la Presse : Part II. Data Driven Journalism Roundtable* [en ligne]. Octobre 2013. Disponible sur http://culture-rp.com/2013/10/17/les-enjeux-du-data-journalisme-pour-la-presse-part-ii/ [consulté le 20/06/2015].

[28] « Le data journalisme : entre retour du journalisme d'investigation et fétichisation de la donnée. Entretien avec Sylvain Lapoix », *Mouvements*.2014, vol. 79, no. 3, pp. 74-80.

données brutes (data store) ou la vente de data-visualisations et d'applications pour des entreprises ou des collectivités (figure 6) [29].

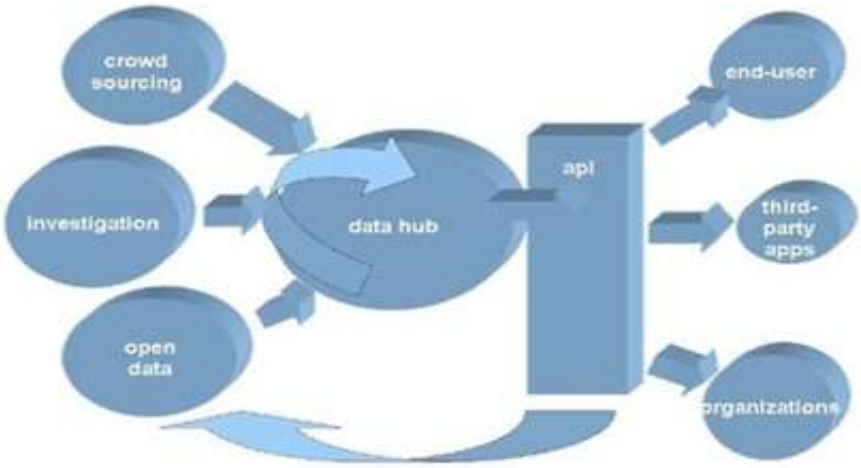

Figure 6: Le nouveau business modèle de la presse à l'ère de la data (CULTURE RP, 2013)

Dans ce business modèle, les médias collectent les données des différentes sources, de l'open data, ou les recueillent à partir des investigations particulières, pour les rendre accessibles aux utilisateurs finaux, aux développeurs ou organisations intéressés par la data.

Coté défi, la nature des actions mises en œuvre ou non par les acteurs des médias et de la presse conditionne la diffusion et la qualité de l'information accessible aux citoyens. A titre d'illustration, la circulation de l'information durant la campagne électorale présidentielle de 2016 aux États-Unis a été entachée par la désinformation sur les réseaux sociaux, un phénomène auquel nous n'échappons pas.

En pratique, la nouvelle forme d'investigation et d'exploitation de données dans les médias, permet de simplifier l'information et produire du sens, en vue de de faciliter l'analyse des sujets complexes avec un maximum de recul, de recoupement et de vérification de l'information, pour contourner le risque de désinformation au moyen du croisement entre le travail de recueil et de traitement des données et le travail journalistique.

[29] CULTURE RP. *Les enjeux du data-journalisme pour la Presse : Part II. Data Driven Journalism Roundtable* [en ligne], Octobre 2013. Disponible sur http://culture-rp.com/2013/10/17/les-enjeux-du-data-journalisme-pour-la-presse-part-ii/ [consulté le 20/06/2015].

L'usage de la data dans les médias s'intègre dans une logique de précision et d'exactitude**,** puisqu'il s'agit d'une analyse de données conditionnée par le récit à établir, ce qui prémunit les lecteurs du biais de l'absence de fiabilité de l'information.

En résumé, l'accessibilité des données doit être corrélée à l'œil critique éditorial du « quatrième pouvoir », pour tirer bénéfice des potentialités des données, mais dans un cadre plus éthique. De menace à ressource, concilier la disponibilité des données avec l'interaction des publics, requiert l'analyse de données dans leur processus de production et de diffusion du contenu journalistique en relation avec la mission du média.

5. Data dans le secteur de l'édition

Le secteur de l'édition est caractérisé par l'importance des métadonnées, à la fois pour le commerce du livre et pour l'accessibilité et la circulation dans les bibliothèques. Le contexte numérique a renouvelé les standards de ces métadonnées. En France, par exemple, le Syndicat national de l'Edition anime un groupe de travail expert sur ces questions en lien avec les acteurs internationaux[30].

La CLIL (La Commission de Liaison Interprofessionnelle du Livre) a procédé à la normalisation des données nécessaires à la description et la commercialisation des livres, pour en extraire des métadonnées riches, circulant dans des formats compréhensibles, utilisant les normes et les schémas ci-après :

- ISBN (International Standard Book Number) Numéro international normalisé permettant l'identification d'un livre dans une édition donnée sur tout support ;
- DublinCore : un schéma générique de métadonnées, destiné à la description des ressources numériques à travers quinze champs de métadonnées relatifs au contenu, à la propriété intellectuelle et à « l'instanciation » : titre, créateur, éditeur, sujet,

[30] ODEH, Souad et CHARTRON, Ghislaine. « Acteurs et économie des métadonnées du livre en France : analyse et avenir » [en ligne]. *Documentation et bibliothèques* 62.2016, n° 1, pp. 21–32. Disponible sur : https://www.erudit.org/fr/revues/documentation/2016-v62-n1-documentation02445/1035926ar/ [Consulté le 21 Août 2020].

description, source, langue, relation, couverture, date, type, format, identificateur, collaborateur et droits ;

- Norme ONIX 3.0 : il s'agit d'une norme permettant l'échange et l'interopérabilité des fichiers de métadonnées relatifs aux livres physiques et numériques ;
- Classification Thèmes CLIL (France) et THEMA (internationale) : c'est une classification thématique du livre physique, visant à structurer une grande librairie générale[31].

La convergence avec les métadonnées élaborées avec les SID est accélérée par le numérique mais certains référentiels restent encore différents, ancrés dans des pratiques professionnelles respectives[32].

Nous allons insister désormais sur les données non structurées et massives qui renouvellent actuellement le secteur de l'édition même si tous les éditeurs ne s'y investissent pas encore.

Les données comportementales des lecteurs sont devenues de nouveaux enjeux :

Plus précisément, les livres ont le défi de capter les attentes des consommateurs, avec une expérience personnalisée, ciblée et interactive**.** Pour y parvenir, il est possible de récupérer des données sur les divers aspects de l'engagement du lecteur avec les livres (données pertinentes provenant des réseaux sociaux, des capteurs d'internet des objets et des recherches Google etc.), et de mobiliser des outils analytiques comme Google Analytics et Shopify[33].

Au niveau de la capture des engagements des utilisateurs, la nature des données produites autour de l'objet « livre » dans le monde de l'édition concerne les avis, les notes et commentaires recueillis par les plateformes de vente ou les réseaux sociaux, le temps passé sur un livre. Un livre peu acheté mais lu rapidement en entier laisserait supposer que les lecteurs ont apprécié celui-ci malgré, peut-être, un manque de mise en avant par l'éditeur.

[31]Syndicat National de l'Edition. *Métadonnées et livres numériques* [En ligne]. Disponible sur :https://www.sne.fr/numerique-2/metadonnees-et-livres-numeriques/ [consulté le 23 Août 2019].
[32] ODEH, Souad et CHARTRON, Ghislaine.Op. cit., p.47.
[33] L, Bastien. Édition : le Big data au secours du marché du livre. *Le big data* [en ligne], *2016.* Disponible sur *:* https://www.lebigdata.fr/edition-big-data [consulté le 23 Août 2019].

Les données liées au comportement de lecture peuvent être approfondies : durée de lecture, pourcentage d'un livre lu et passages les plus marquants ouvrent la voie à la personnalisation et à la recommandation personnalisée. Aujourd'hui, des plateformes de streaming telles que Youboox recueillent des millions d'informations sur les habitudes de lecture et peuvent les partager avec l'industrie du livre pour en faire un atout.

En matière d'optimisation du marketing éditorial, les données sur le temps utilisé pour la lecture et les chiffres de vente d'un livre déterminent les futurs best-sellers et les éditeurs peuvent grâce à ces données anticiper le succès d'une série. Si le tome 2 n'est terminé que par 50 % des lecteurs contre 90 % pour le premier, cela pourrait remettre en cause une suite possible. De la même façon, un livre peu sélectionné, mais souvent toujours terminé par le lecteur peut questionner l'adéquation de sa couverture qui ne semblerait pas inciter les lecteurs à le découvrir. Le big data peut ainsi optimiser le marketing des éditeurs à travers l'analyse du comportement des lecteurs, aboutissant à la création de designs de couverture attractifs pour augmenter les ventes[34]**.**

Nous allons évoquer désormais deux exemples d'usages des données dans le secteur du livre dans une vision d'aide à la décision mobilisant des algorithmes d'intelligence artificielle.

- **L'exemple du service *Short edition* :**

Short Édition, est un éditeur communautaire de la littérature courte, qui s'apparente à un programme de recherche fondé sur la technologie du Machine Learning, qui entreprend le rôle d'un système intelligent aidant le comité éditorial dans l'évaluation de la qualité des documents.

Ainsi, l'élément clé d'apprentissage automatique correspond à un modèle où les individus réalisent une tâche et l'algorithme apprend à reproduire la tâche. L'algorithme peut par la suite exécuter la tâche sans intervention humaine avec un taux d'erreur mesurable. Ainsi, dans notre cas précis, l'algorithme prédictif de Short Édition évalue le travail d'un membre du comité éditorial (il se nourrit de ces interventions et évaluations) et donne une probabilité sur la qualité littéraire de nouveaux textes soumis.

[34] Ibid., p.50.

L'algorithme est utilisé comme une aide à la décision, mais ne vise pas à remplacer l'intervention humaine, matérialisée par l'avis du comité éditorial ; il est destiné à établir un pré-classement et une pré-évaluation des travaux soumis et filtrer les œuvres pour sélectionner les meilleures[35]. Short édition ne vise pas à transformer le comité éditorial en une machine ; il s'agit plutôt d'un dispositif de filtrage-assistant, dans une détection, moins de la qualité littéraire que de l'absence de qualités. De ce fait, l'ordinateur pointe dans les textes soumis un certain nombre de caractéristiques, qui seront alors confirmées par un cerveau humain.

Short Edition s'adapte à des environnements de *digital nomad*, qui permettent aux auteurs et lecteurs d'exercer leurs activités à distance, grâce à internet, en s'affranchissant des contraintes spatio-temporelles, puisqu'il s'agit d'écouter ou lire un panel d'histoires courtes, d'un trait et en moins de 20 minutes, tout en s'appuyant sur la force de l'intelligence artificielle.

La capture d'écran ci-dessous présente la page d'accueil de Short édition[36] :

Figure 7 : L'écran d'accueil de Short édition (Short Edition, 2015)

Le fonctionnement de cet algorithme de prédiction de la qualité littéraire requiert une base de données des œuvres proposées par Short et évaluées par cinq à dix lecteurs. À partir de ces premières données, la machine commence à apprendre et établir des liens entre la qualité et les exigences.

[35] SHORT EDITION. *Editeur communautaire* [en ligne]. Disponible sur : http://short-edition.com/fr/p/distributeurs-d-histoires-courtes [consulté le 20/06/2015].
[36] Ibid.

Une première base de 25.000 œuvres sert donc à alimenter les algorithmes, afin de lui fournir les bases d'une première série d'évaluation. De cette manière, Short édition dispose de nouvelles références, et de nouveaux critères lui permettant de juger de la qualité littéraire du texte considérant les éléments suivants :

- Les indicateurs sémantiques, le nombre de répétitions d'un mot, ou de plusieurs à l'intérieur du texte à travers des figures de style, comme l'anaphore par exemple ;
- L'ampleur du champ lexical, le recours à des termes argotiques, vulgaires ou des mots rares peut donner une idée sur la qualité globale de certains textes ;
- La longueur des phrases, des paragraphes (vers et strophes y compris), allant jusqu'à des mesures syntaxiques, comme le nombre de propositions relatives ;
- Le style, dans son ensemble, passe par une classification grammaticale des termes employés (quantité d'adverbes, d'adjectifs, de pronoms, de verbes, de noms, etc.) ;
- Les fautes d'orthographe et de ponctuation apportent un regard qualitatif ;
- Enfin, la machine recherchera la lisibilité du texte, à travers un examen de la cohésion du texte mesurée minutieusement à partir de l'emploi des connecteurs logiques et l'enchainement des idées.

Short Edition est, répétons-le, utilisé comme une aide à l'évaluation et au traitement des textes. Néanmoins, cette innovation aura aussi un grand intérêt pour les acteurs de l'édition (pré-évaluation des manuscrits) et pour les acteurs de la presse (médias ou contenus participatifs, nouveau modèle économique avec le numérique) face au flux de textes circulant.

Le choix des critères peut par contre être très contestable car subjectif.
Ce type d'algorithme peut aussi être convoqué pour mesurer le degré d'émotion qui se dégage d'une œuvre, en mesurant les indicateurs liés à la joie, à la peur, au plaisir, etc.

Tout compte fait, la note donnée par la machine importe moins que sa capacité à justifier cette note attribuée à la fin de l'analyse des textes soumis[37].

[37] SHORT EDITION. *Editeur communautaire* [en ligne]. Disponible sur : http://short-edition.com/fr/p/distributeurs-d-histoires-courtes [consulté le 20/06/2015].

L’exemple du service *Kobo* :

L’usage de la data dans le domaine de l'édition est au cœur du service Kobo qui propose des tablettes de lecture notamment pour les livres numériques. Cet usage voudrait répondre à l’anticipation des attentes des lecteurs, à la détermination du degré de succès d'un titre notamment.

Avec l'avènement de la lecture numérique, Kobo permet désormais de repérer les livres qui n'ont pas été ouverts et ceux ayant été lus jusqu'à la dernière page et à quelle vitesse, en stockant des données sur les lecteurs, source d’informations pour les éditeurs dans un second temps. Ainsi, la collecte et/ou du traitement de données sont conditionnés par les autorisations des lecteurs de ce service, en conformité avec le RGPD **(**le Règlement Général de Protection des Données) appliqué en matière du traitement des données de manière égalitaire sur tout le territoire de l’Union Européenne depuis 2018.
La capture d’écran qui suit, nous présente l’interface d’accueil de Kobo[38] :

Figure 8: L’écran d’accueil de Kobo (KOBO, 2015)

Kobo peut prendre des décisions sur la base des données cumulées, c’est le cas de la promotion d’un auteur moyennement connu ou inciter à transformer un livre en série, à la vue de l'engagement des lecteurs, etc.

Dans l’édition scientifique, les professionnels du secteur accordent une grande importance à l’analyse des insights relatifs à l’engagement du lecteur avec les articles, en déployant des outils

[38]KOBO. *Ebook* [en ligne]. Disponible sur https://www.kobo.com/ [consulté le 20/06/2015].

analytiques et des systèmes de données avancés sur les différentes plateformes numériques des éditeurs.

En effet, ces systèmes peuvent aider les éditeurs à paver la voie pour un traçage efficace des pratiques et des habitudes d'usage des articles et contenus scientifiques par les lecteurs en ligne à partir des données *log* relatives au parcours et actions de navigation sur les sites et les applications web, ainsi que les cookies ou les témoins de connexion envoyés par les serveurs de ces plateformes et stockés sur le terminal du lecteur, permettant d'enregistrer certaines données, afin de faciliter la navigation et de permettre certaines fonctionnalités

Concrètement, il s'agit, entre autres, de savoir les articles et contenus visités et ceux qui ne le sont pas, pour cibler en temps réel les achats, fixer les objectifs commerciaux et marketing, proposer des contenus pertinents et comprendre les affinités des lecteurs.

Chercher, stocker et miner des données sur les pratiques numériques de lecture est un processus, qui mobilise plusieurs sources hormis la plateforme de l'éditeur. Ainsi, les réseaux sociaux, les capteurs de l'internet des objets, les moteurs de recherche, les e-mails sont aussi sollicités

A titre d'exemple, l'éditeur Barnes & Noble s'est doté d'un entrepôt de données, pour obtenir une compréhension des habitudes de lecture au travers des analyses de terabytes de données internes, et l'éditeur Elsevier a utilisé les données de lecture numériques pour créer « l'article du futur », soit un article enrichi par des fonctionnalités, des images et des fichiers divers. L'éditeur a été capable d'utiliser ce modèle pour fournir à son audience une expérience dynamique et ciblée unique[39].

Le bilan de l'atelier " Culture, médias et numérique ", a mis en évidence les tendances lourdes relatives au secteur du document numérique (le livre augmenté, la valeur économique et sociétale, le renouvellement des modèles économiques, l'évolution de la régulation, la transformation des pratiques culturelles et de la fonction d'intermédiation) et les tensions majeures (dématérialisation et la tension entre acteurs économique et sur les usages), qui peuvent réorganiser le secteur avec la personnalisation, les nouveaux modèles de définition des

[39] VERLAET, Lise et DILLAERTS, Hans. « L'enjeu du web de données pour l'édition scientifique », *I2D – Information, données & documents*.2016, vol. 53, no. 2, pp. 49-49.

tarifs, les nouveaux acteurs technologique et la sociabilité des utilisateurs sur les réseaux sociaux[40].

6. Data dans le secteur de la musique

Les données structurées et non structurées sont au cœur du secteur de la musique, à la fois avec des enjeux économiques, techniques, politiques et intellectuels.

Aussi, les métadonnées représentent des informations d'enrichissement et d'annotation des contenus, réalisés en amont ou en aval de leur publication sur le web, en s'appuyant sur plusieurs approches : les ontologies, la documentation collaborative, la redocumentarisation de fonds audiovisuels, la veille informationnelle et l'utilisation de l'indexation libre à travers les folksonomies[41].

Dans le contexte d'un enregistrement musical, nous distinguons les différents types de métadonnées suivants :

- Les métadonnées de propriété : les informations sur les ayants-droits et propriétaires du contenu permettent la juste rétribution de la filière ;
- Les métadonnées de gestion (ou commerciales) : les informations sur les conditions de commercialisation telles que le prix, les différents supports disponibles, les autorisations de diffusion, etc. ;
- Les métadonnées descriptives : elles permettent de connaître les informations factuelles liées à un contenu (année de production, nom de l'œuvre, etc.) ;
- Les métadonnées d'enrichissement : il s'agit des informations annexes au contenu qui enrichissent et améliorent l'expérience de l'utilisateur (paroles, photos, vidéo, etc.) ;
- Les métadonnées acoustiques : les données résultant de l'analyse sonore du contenu (tempo, métrique, etc.)[42].

[40] CHARTRON, Ghislaine. Tendances lourdes et tensions pour les filières du document numérique. *Le "Document" à l'ère de la différenciation numérique*, Dec 2011, Maroc.

[41] BROUDOUX, Evelyne et SCOPSI, Claire. Métadonnées sur le web : les enjeux autour des techniques d'enrichissement des contenus. *Etudes de communication*. 2011, n°.36.
[42]CNIL (Commission Nationale de l'Informatique et des Libertés). Op.cit., p.39.

L'archivage et la centralisation de ces données dont certaines sont libres de droit, ont jusqu'ici toujours failli. La numérisation croissante de la musique, la diversification des supports d'écoute et surtout la complexité d'une industrie musicale qui intègre une multitude d'intermédiaires à tous les niveaux (sociétés d'ayants-droits, labels, producteurs, distributeurs physiques et digitaux, etc.) rend la tâche d'autant plus difficile. En France par exemple, les deux principales bases de données en la matière sont la BNF (dépôt légal) et les archives de Radio France.

La musique a été la première industrie touchée par la révolution internet. L'enjeu majeur est désormais celui de la construction des systèmes de recommandation pertinents, grâce à l'interprétation des données. D'autant qu'à l'inverse du secteur du livre, il n'existe pratiquement plus de disquaires avec qui échanger pour obtenir des conseils. De fait, les plateformes de musique s'alimentent souvent des données sociales issues des réseaux sociaux, pour générer des inputs aux algorithmes de recommandation[43], elles s'appuient également sur toutes sortes de métadonnées comme celles détaillées précédemment.

Les avancées de l'IA sont aussi notables en termes d'apprentissage automatique dans ce secteur. Ainsi, dans le contexte américain, les données des sites de vente de musique ont été analysées, en étudiant 83 critères par titre, pour identifier des corrélations et anticiper si un titre deviendra un tube ou non.

Deezer et Spotify tentent aussi de comprendre la musique comme les auditeurs la comprennent, en utilisant un apprentissage informatique à deux dimensions : analyse sonore et analyse de textes. Deezer et Spotify ont commencé par scanner en profondeur leur gigantesque base musicale de plus de 30 millions de titres et en dégager « Le timbre, le tempo, la présence de tel ou tel instrument, la nature acoustique, électrique ou électronique de la musique, la voix… ».

Selon Beyers[44] les deux opérateurs font appel aussi à l'indexation des textes traitant de la musique avec une fréquence quotidienne (des blogs, des chroniques de disques, des articles,

[43] ALEXANDER, Peter J., Peer-to-Peer File Sharing: The Case of the Music Recording Industry. *Review of Industrial Organization*. 2002, Vol. 20, n°. 2, p. 151-161.
[44] BEYERS, William, el al. *The Economic Impact of Seattle's Music Industry* [en ligne]. Office of Economic Development, 2004. Disponible sur :https://www.seattle.gov/Documents/Departments/FilmAndMusic/Seattle_Music_EIS_2008.pdf [Consulté le 23 Juin 2016].

des tweets, des paroles…), en vue d'extraire les mots-clés récurrents relatifs à un artiste ou un disque et alimenter l'algorithme.

Avec toutes ces données, Deezer et Spotify ont les moyens de personnaliser l'ensemble de l'expérience pour chaque utilisateur et de connaître ces internautes en profondeur avec « *un taste profile* » chez Spotify, qui établit pour chacun de ses abonnés - payants ou pas - une carte d'identité traçant ses habitudes. De la même manière, un algorithme trouvera le bon équilibre entre les tubes et des titres plus « découverte », afin de s'attaquer aux *casuals*, qui sont des auditeurs ayant tendance à mettre la radio dans la voiture ou à la maison, mais ils n'ont pas l'envie et la motivation d'aller découvrir des nouveautés par eux-mêmes.

Spotify tend à construire des radios personnalisées et des playlists thématisées (repas, fête, voyage…). Deezer va un peu plus loin en proposant « Flow », un unique bouton qui déclenche une playlist infinie, présentée sous forme d'une bande-son personnalisée de recommandations, construite à partir de l'historique des écoutes, les indications sur les musiques préférées et les préférences de genres et d'artistes.

Les *data scientists* exercent côte à côte avec les professionnels métier, dans le but d'optimiser les pratiques de marketing et de compréhension du consommateur.

La transformation numérique de l'industrie musicale, selon Moreau[45], a affecté les étapes de création, de financement et de production des œuvres, ainsi que le métier des différents acteurs de ce secteur dans ses diverses facettes (production, distribution/diffusion, promotion, etc.). Inscrite dans une optique disruptive, cette transformation numérique a donné naissance aussi à de nouveaux modèles d'affaires numériques (notamment le streaming).

Par ailleurs, les coûts de production et de distribution des œuvres musicales ont diminué de manière significative grâce aux TIC au même titre que les métadonnées nécessaires pour décrire et identifier une œuvre, puisqu'elles peuvent être utilisées ou produites par les utilisateurs, et la numérisation a changé le mode d'échange des œuvres musicales, qui sont passées du disque et du CD aux fichiers numériques susceptibles à être copiés et diffusés. De fait, le progrès technique permet, d'un côté, d'associer la valeur des œuvres musicales aux métadonnées liées

[45] MOREAU, François. The Disruptive Nature of Digitization: The Case of the Recorded Music Industry. *International Journal of Arts Management* ?. 2012.

à leur production et utilisation et, de l'autre, la gestion de ces métadonnées sur internet à travers des sites des différents acteurs de cette industrie au lieu du système des médias de masse[46].

7. Data dans le secteur du E-Learning

La data dans l'industrie du e-learning renvoie aux données créées par les apprenants pendant qu'ils suivent un cours en ligne ou un module de formation. Par exemple, si un usager est en interaction avec un module de formation, ses progrès, les résultats de l'évaluation, le partage social ainsi que d'autres données produites au cours de son apprentissage constituent un ensemble relevant du big data.

Ce champ désigné comme celui du *learning analytics* collecte, mesure et analyse les traces laissées par les apprenants dans leur environnement digital, en vue d'améliorer le processus d'apprentissage. Le *learning analytics* s'alimente de l'ingénierie des connaissances appliquée aux traces numériques d'activités, dans le but d'exploiter l'analyse des données d'apprenants dans la compréhension et l'optimisation de l'apprentissage[47].

Plus précisément, la data fournie aux professionnels du e-learning leur permet :

- De définir les besoins d'apprentissage des apprenants. Par exemple, les données permettent de vérifier si un scénario basé sur la réalité est plus efficace qu'une activité de résolution de problèmes ;
- De repérer les parties qui peuvent avoir besoin d'être affinées dans les cours ou les modules. Par exemple, si plusieurs apprenants passent une longue période, pour

[46] BOURREAU, Marc et GENSOLLEN, Michel. « L'impact d'Internet et des Technologies de l'Information et de la Communication sur l'industrie de la musique enregistrée ». *Revue d'économie industrielle* [en ligne].2006, n°116, p. 2-2. Disponible sur : https://www.cairn.info/revue-d-economie-industrielle-2006-4-page-2.htm [consulté le 23/08/2020].

[47] FRAOUA, K E., LEBLANC J-M., CHARRAIRE S., CHAMPALLE O., *Information and Communication Science Challenges for Modeling Multifaceted Online Courses,* Human Computer Interaction Orlando, 2019.

terminer un module particulier, cela signifie probablement que le module doit être amélioré, afin de le rendre plus facile et accessible pour les apprenants ;

- Elle fournit une analyse des modules e-learning qui sont le plus visités et dans le cas de l'apprentissage social des liens partagés avec d'autres apprenants. Par exemple, déterminer quel lien a été le plus partagé via Facebook, et ainsi personnaliser le cursus de formation à l'apprenant en fonction de caractéristiques (âge, sexe, niveau d'éducation, etc.). L'objectif est la prédiction des formations pointues cadrant avec les attentes et les objectifs d'apprentissage des apprenants ;
- Les données peuvent être reçues en continu. Par conséquent, les professionnels du e-learning peuvent apporter leurs modifications et changements, afin d'affiner leur stratégie de e-Learning en temps réel.

La data ne se résume pas seulement dans le volume de données, mais dans sa valeur aussi, du moment où elles peuvent être analysées, pour offrir aux organisations ou des professionnels du e-Learning la possibilité de déterminer comment l'apprenant reçoit des informations, à quel rythme et d'identifier les problèmes qui peuvent exister dans la stratégie e-learning[48].

Par ailleurs, les commentaires, les enquêtes et les discussions en ligne peuvent fournir un retour des apprenants sur la pertinence des cours et des modules. L'animateur de la plateforme peut faire les ajustements nécessaires pour améliorer la performance des apprenants ou encore tenir compte de leur souci de personnalisation du contenu, en fonction de leurs niveaux, leurs prérequis et leurs attentes[49].

En matière du suivi de l'apprenant et avec les données, les professionnels du e-learning acquièrent la capacité de suivre un apprenant tout au long du processus d'apprentissage, du début jusqu'à la fin. En d'autres termes, il devient possible de voir comment les apprenants ont abordé un test ou comment rapidement ils ont terminé un module difficile, pour suivre ainsi les comportements de l'apprenant de manière individuelle ou dans un groupe.

[48]FARIS, Robert et HEACOCK, rebekah. Measuring Internet Activity: a (selective) Review of Methods and Metrics. *Internet Monitor Special Report Series*. 2013, n°.2, 39 p.

[49] Ibid.

8. Synthèse sur l'usage de la Data dans les industries culturelles

La place de la data structurée et non structurée dans les industries culturelles dans les cinq secteurs des industries culturelles abordés précédemment est résumé dans le tableau suivant, insistant sur les enjeux et les services innovants associés :

Le secteur	Les données	Les enjeux	Les exemples de services internes et externes développés à partir des données
Médias : presse, audiovisuel	- Les données brutes et hétérogènes disponibles sur l'audience et les bases de données volumineuses, pour en extraire de l'information et la présenter de façon engageante au public.	- L'extraction des données sur les attentes et les préférences de l'audience de chaque programme visionné, chaque chanson écoutée, chaque avis laissé sur les réseaux sociaux ; - Le développement du data-journalisme : récupérer et intégrer les données, les nettoyer, puis les analyser d'un ou de plusieurs points de vue possible (calculs statistiques, croisement avec d'autres jeux de données, recherche de corrélations et visualisations intermédiaires) ; - L'amélioration de la créativité des data journaliste à travers la création de liens entre les données, leur croisement, pour identifier des relations originales qui n'étaient pas explicites auparavant ;	- Les pure players digitaux comme Google, Facebook, Netflix qui déploient de grands programmes orientés vers la connaissance très fine du client ; - Proposition du contenu le plus adapté à la cible recherchée et diffusion au bon moment sur les bons canaux ; - Prédiction fine (les films, les émissions, etc.) que l'audience plébiscitera ; - Publicité ciblée, différenciée en fonction des caractéristiques de leur profil, - Bases de données enrichies, d'infographies interactives, de timelines et de cartes rich-media, d'applications interactives ; - Mise à disposition de données brutes (data store).
Edition	De façon plus ancienne, les données commerciales relatives aux achats et les métadonnées utiles pour le repérage et le commerce des livres. Et plus récemment, les données comportementales de lecture collectées sur les plateformes et les appareils connectés des lecteurs.	- La corrélation les données de vente à des données comportementales, et à des données sur la lecture, avis, notes et commentaires recueillies sur les plateformes ou les réseaux sociaux ; - La normalisation des métadonnées nécessaires à la description et à la commercialisation des livres, utilisant les normes et des standards.	- La définition d'une politique éditoriale à partir de l'analyse des données d'usage et d'achat ; -Une optimisation marketing des éditeurs en étudiant les données d'achat des livres en fonction de plusieurs variables (profil, sexe, âge, localisation géographique, etc.) ; - Une compréhension fine des pratiques : la durée de lecture, la fréquence, le pourcentage d'un livre lu et les passages les plus marquants dans le but de prédire les préférences des lecteurs et leur recommander des livres correspondant à leurs goûts et tendances ; - L'aide à l'évaluation des manuscrits (Short edition) et au repérage des succès potentiels (algorithmes).

Musique	Les données sociales issues des réseaux sociaux et les métadonnées décrivant les morceaux de musique.	- La construction des systèmes de recommandation pertinents, grâce à l'interprétation des données ; - L'indexation des textes traitant de la musique avec une fréquence quotidienne (des blogs, des chroniques de disques, des articles, des tweets, des paroles…).	- Segmentation des contenus : Playlists thématisées (repas, fête, voyage…) ; Par exemple, Deezer et Spotify utilisent un apprentissage informatique à deux dimensions : analyse sonore et analyse de textes tendant vers la constitution d'une base musicale de plus de 30 millions de titres, sont identifiées des données sur le timbre, le tempo, la présence d'instruments, la nature acoustique, électrique ou électron-ique de la musique, la voix, etc. ; - Personnalisation en fonction des usagers : Customisation de l'expérience pour chaque utilisateur, définir pour chacun un *taste profile* ; -Algorithmes de recommandation affinés ; - Contextualisation automatisée du contenu, en fonction de l'individu, du collectif.
E-Learning	Les données créées par les apprenants pendant un cours en ligne ou un module de formation (les traces numériques).	- Le suivi des comportements d'apprentissage de l'apprenant de manière individuelle ou dans un groupe ; - La prise en compte du retour des apprenants sur la pertinence des cours et des modules, pour améliorer l'offre des cours et leur déroulement.	- Déploiement du *learning analytics* qui collecte, mesure et analyse les traces numériques d'usage laissées par les apprenants dans leur environnement digital, en vue d'améliorer le processus d'apprentissage ; - Permettre aux professionnels e-learning d'améliorer leurs contenus ; - Analyser des modules e-Learning les plus visités et les plus partagés avec d'autres apprenants ; - Prédire les modules et les cours futurs, susceptibles d'intéresser les apprenants.

Tableau 1: Synthèse comparative de l'enjeu de la donnée dans les secteurs d'industries culturelles[50]

[50] CNIL (Commission Nationale de l'Informatique et des Libertés). Industries créatives, contenus numériques et données : les contenus culturels vus au travers du prisme des données. Le graal de la recommandation et de la personnalisation. *CAHIERS IP INNOVATION & PROSPECTIVE* [en ligne]. 2015, n°3. Disponible sur : https://www.cnil.fr/fr/innovation-prospective/publications [Consulté le 25 Juin 2016].

Dans le tableau de synthèse ci-dessus, nous remarquons que les industries culturelles ont en commun le défi de la personnalisation et de la recommandation, qui placent la connaissance de l'utilisateur et la proposition de services personnalisés au centre de leurs principaux enjeux.

Cependant l'usage de ces données dans le cadre des activités des professionnels des industries culturelles et du contenu demeure encore limité. Une étude qualitative inscrite dans un appel à projet "Développement de l'emploi en Île-de-France" réalisée par Cap Digital[51] menée auprès de 78 professionnels (48% de dirigeants et DRH) des filières Edition, Presse, Audiovisuel/Cinéma, Publicité/Communication de l'usage de la data, a révélé que les professionnels ont principalement recours aux données pour la connaissance client (71%)**,** donc dans un objectif marketing et encore peu de services innovants et personnalisés.

Quant aux différences, il faut noter que les secteurs comme l'édition et la presse ont surtout considéré les données pour faciliter le travail du back-office entre les différents acteurs. La diffusion des publicités ciblées est aussi une priorité des médias de masse.

L'objectif de fournir des services de contenus affinés selon les caractéristiques des usagers (et de développer des fonctionnalités d'anticipation) marque plus le secteur de la musique et celui du *learning analytics* dans l'analyse que nous avons pu faire dans cette première approche.

[51] Chloé, L. *L'impact de la data sur les métiers des industries culturelles et créatives. Retour sur l'étude de Cap Digital*. In Blog MBA MCI [en ligne]. (Mis en ligne 2017). Disponible sur :https://mbamci.com/data-metiers-culturels-et-creatifs/ [Consulté le 24 Août 2019].

9. Les freins de l'exploitation de la data dans les industries culturelles

Le traitement et l'exploitation de la data dans les industries culturelles évolue au fur et à mesure du développement des technologies d'information et de la communication. En revanche, l'analyse de la data se confronte aussi à des freins généraux, que nous présentons dans le tableau qui suit[52] :

Le frein d'exploitation de la data dans les industries culturelles	L'explication
Le manque en RH	Les ressources humaines spécialisées dans le traitement des données dans les industries culturelles et créatives restent généralement très limitées (pas plus de 10 personnes) avec un manque de formation et de qualification, contrairement à d'autres secteurs d'activité comme la banque, l'environnement…
Les outils analytiques	Les acteurs dans ces industries disposent, en règle générale, de peu d'outils et de dispositifs dédiés au traitement et à la fiabilisation des données non structurées, pour en faciliter l'exploitation[53].
Le temps réel	Les entreprises culturelles tendent à utiliser les données clients rarement à des fins prédictives. Cependant, un tel usage devrait accroître la rapidité de traitement de ces données et optimiser la prise de décision.
La vision unifiée	La majorité des sociétés utilisent la data dans des silos internes, hérités de différentes générations des système d'informations, empêchant l'exploitation optimale des données clients. Chaque composante utilise les données issues de ses bases de données, afin de répondre à ses propres enjeux métiers. Mais la data ne peut pas alimenter la prise de décision stratégique, en raison de l'absence d'une vision unifiée et unique.

[52]Ernst & Young. *Les secteurs culturels et créatifs européens, générateurs de croissance* [En ligne]. Disponible sur : http://www.creatingeurope.eu/en/wp-content/uploads/2014/11/study-full-fr.pdf [Consulté le 15 Avril 2021].

[53] Ibid.

Le ROI	Le retour sur investissement des solutions dédiées au traitement de la Data dans la performance d'une industrie culturelle n'est pas encore perçu comme un enjeu prioritaire.
Le soutien des DG	La conjoncture économique défavorable combinée à l'absence de mesure du ROI alimentent la réticence du top management dans les industries culturelles et ne favorisent pas une exploitation optimale des données au sein de l'entreprise, notamment celles non matures.
La protection des données privées	La sécurité des données et la protection de la vie privée sont des sujets centraux à l'heure où les règlementations juridiques se renforcent et où les sanctions se multiplient.

Tableau 2: Freins pour l'exploitation optimale de la data dans la sphère des industries culturelles

Enfin, le besoin s'affirme dans ce secteur, comme dans d'autres secteurs, de profils poly-compétents au moment où le travail devient collaboratif et fait appel à des compétences hybrides (l'association de deux compétences d'univers professionnels différents), à l'image du journalisme qui conjugue l'enquête journalistique et la data-visualisation.

Conclusion

En conclusion, la data utilisée dans les industries culturelles aide à la création de nouveaux services et l'amélioration de la qualité de services existants, tout en tenant compte de l'analyse et la visualisation de données en temps réel. L'objectif est aussi celui d'améliorer la prise de décision par exemple via des tableaux de bord.

Il est vrai que pour saisir ces opportunités, les organisations du monde de la culture et de la production de contenu doivent s'équiper de nouveaux outils et s'adjoindre de nouvelles compétences.

Les décideurs doivent se saisir de ces enjeux notamment dans un contexte où le développement des dispositifs mobiles permet de plus en plus de générer des flux de données en temps réel sur les usages et leurs usagers.

Partie II : Data et les SID

Chapitre 1 : Croissance des données dans les SID et diversification des services associés

Introduction

Dans ce chapitre, nous allons désormais nous intéresser au statut des données dans les Services d'Information Documentaire (SID), dispositifs inscrits dans le secteur plus global de la culture et de l'enseignement.

Nous identifierons les différents types de données convoquées dans les SID, pointerons l'évolution des enjeux et des services progressivement développés, transformant ainsi le modèle de la bibliothèque. Nous insisterons dans le second chapitre sur la dimension big data, qui aujourd'hui renouvelle autrement les services notamment par une connaissance profilée des usagers.

1. Typologie des données dans les SID

Avant d'appréhender de manière fine les pratiques d'usage des données dans les services d'information documentaires, nous passons en revue des types des données, objet des éventuels usages, ainsi que leurs caractérisations synthétisées dans le tableau ci-dessous :

Type de données	Caractérisation des données
Données d'activité	- Les données statistiques relatives au prêt, au taux de rotation des collections, aux réservations, retours des documents, aux consultations sur place et à distance ; - Les données liées à l'action culturelle (expositions, contes pour enfants, rencontres avec des écrivains, conférences, colloques, etc.) et l'action sociale (ateliers d'alphabétisation, ateliers pour la recherche d'emplois, initiations aux recherches et aux nouvelles technologies, les partenariats associatifs, etc.) ; - Les données résultant des maillons de la chaine documentaire allant des acquisitions (veille documentaire pour le repérage, sélection des documents, achat, dons, legs, dépôt légal et l'échange entre bibliothèques), le bulletinage des périodiques, les données de conservation des collections courantes et des collections patrimoniales, les données d'élagage de collections.
Données de description	- Les données décrivant les usagers du point de vue sociodémographique, d'interaction et des préférences (le sexe, l'âge, l'adresse personnelle, le niveau d'étude, les sujets et les intérêts de recherche et de lectures, etc.) ; - Les données et métadonnées servant à caractériser les collections d'une bibliothèque et aidant à y accéder (les index, bulletins de sommaires et de résumés pour les numéros des périodiques, les bibliographies, les dossiers thématiques, etc.).
Données de référence	- Les données relatives aux opérations relatives au traitement intellectuel des documents : le catalogage, la description et l'indexation matière (description par des mots du contenu afin de permettre les recherches en se basant sur les thésaurus comme celui de l'UNESCO), les données des indices de classification (Classification Décimale de Dewey ou Universelle et les classifications spécialisées par domaine).

Tableau 3: La typologie des données exploitées par les services d'information documentaire

Le tableau met en évidence trois principales catégories de données dans les SID, à savoir les données sur les usagers, les données sur les collections et les données d'activités des bibliothèques.

D'un point de vue historique et identitaire, les bibliothèques ont principalement été attachées à l'élaboration des données de référence et de description des documents. La sociologie des publics a ensuite donné une importance croissante aux données caractéristiques des lecteurs,

aujourd'hui les données d'activités des lecteurs, de plus en plus traçables dans les collections numériques, renouvellent les enjeux sur la qualité des services rendus et attendus.

Les services d'information documentaires rassemblent et traitent généralement les données de manière traditionnelle (pourcentages, minimum, maximum, moyennes, etc.) pour éclairer leur décision.

Selon Jamene Brooks-Kieffer[54], il existe une nouvelle dimension des méthodes d'analyse. L'auteur commence par opposer la macro-évaluation basée sur un ensemble de variables, de façon à ce qu'elles décrivent un organisme, à la micro-évaluation qui se concentre sur la manière dont ces variables s'affectent les unes par rapport aux autres.

« Les manières traditionnelles avec lesquelles les bibliothécaires rassemblent et traitent les données vont rarement jusqu'à l'emploi des techniques d'analyse, de traitement, ou d'exploration qui seraient considérées comme une nécessité dans toute autre profession aussi riches en données que la nôtre. De telles techniques ne sont pas faciles à employer, mais elles produisent des résultats remarquablement informatifs, si ce n'est parfois inconfortable. Mais à la place, nous préférons n'avoir affaire qu'à la signification superficielle et rassurante de nos données, nous reposant sur une prédominance des variables quantitatives et les conclusions simples arithmétiques que nous pouvons en retirer. Ces conclusions sont rassurantes, car elles ne mènent que rarement à des résultats inattendus »[55].

A titre d'illustration de ce passage de la macro à la micro évaluation, Brooks-Kieffer prend l'exemple d'une situation particulière : celle où un directeur d'une bibliothèque universitaire cherche, à savoir dans quelle mesure les services de prêt répondent aux besoins en documentation des usagers distants de la bibliothèque.

Pour répondre à cette question, une macro-évaluation avance un indicateur de performance, qui est le nombre total de documents empruntés, distribués en fonction des codes postaux des usagers. Néanmoins, pour déterminer, par exemple, pourquoi les usagers distants empruntent

[54] AURORE, Cartier. *« Promesses et défis de l'open data en bibliothèque », dans : Lionel Dujol éd., Communs du savoir et bibliothèques* [en ligne]. Paris, Éditions du Cercle de la Librairie, « Bibliothèques », 2017, p. 97-108. DOI : 10.3917/elec.dujo.2017.01.0097.Disponible sur : https://www.cairn.info/communs-du-savoiret-bibliotheques--9782765415305-page-97.htm.[Consulté le 09 Juillet 2021]
[55] Ibid.

moins de documents que les usagers locaux, il faudra faire appel à une micro-évaluation avec une combinaison des données quantitatives et qualitatives provenant du Système Intégré de Gestion de Bibliothèque (SIGB) et des usagers, pour examiner les interactions des usagers distants avec le service d'information documentaire: les politiques mises en place au sein de l'établissement, les contraintes horaires et les contenus des cours, etc.

Un autre exemple illustre, peut-être davantage, les idées développées par Jamene Brooks-Kieffer, à l'image de l'étude réalisée en 2013 à la bibliothèque de la faculté du New Jersey[56], qui visait à évaluer la pertinence des acquisitions les plus récentes au regard de l'activité de l'établissement. Pour ce faire, il était question d'interroger la corrélation de trois variables différentes : les acquisitions récentes, les circulations récentes et les demandes de prêt entre bibliothèques également récentes, de façon à éclairer un certain nombre de points liés au développement de la collection (les besoins des usagers, les données d'usages et les demandes de documents par rapport au fond disponible)[57].

L'objectif de cette étude était de renouveler le dialogue entre les acquéreurs et les usagers et de renouveler la politique de développement des collections des services d'information documentaire.

2. L'exemple d'Online Computer Library Center OCLC) : évolution des services de la Data

L'OCLC gère l'évolution des collections physiques des bibliothèques dans le contexte de la numérisation de masse à travers la réutilisation des données bibliographiques, comme en témoigne l'exemple de WorldCat, qui est la plus grande base de données bibliographique du monde. Avec les informations de plus de 3 milliards de documents de bibliothèques. Depuis 2012, l'OCLC a favorisé un processus d'ouverture de ses données bibliographiques qui servent à la description catalographique.

[56] E. LINK, Forrest el al. *Mining and Analyzing Circulation and ILL Data for Informed Collection Devlopment* [en ligne]. Preprint de College & Research Libraries, 2015. Disponible sur : http://crl.acrl.org/content/early/2014/10/20/crl14632.full.pdf [Consulté le 10/10/ 2015].
[57] Ibid.

Par ailleurs, l'organisme de recherche s'est doté d'une section entièrement consacrée à l'extraction et à l'analyse de données (*Data Mining Research Area*), les objectifs assignés à cette section sont les suivants :

- Générer des présentations intéressantes et innovantes des données ;
- Fournir des informations pour répondre à un certain nombre de besoins en matière de prise de décision dans les bibliothèques, telles que le développement des collections, la numérisation et la conservation.

Les objectifs évoqués ci-dessus ont pris la forme de plusieurs projets, commençant par les collections des fonds orientées vers la numérisation et la conservation, en passant par la déduction par inférence des publics cibles ou des niveaux d'audience des ouvrages à partir des informations provenant des fonds, jusqu'à l'évaluation comparative de collections.

Aux éléments déjà évoqués, nous ajoutons l'exploitation des identités des auteurs et des éditeurs, pour fournir des services adaptés sur l'interface de recherche de WorldCat. Ils ont été présentés, en 2013 à Strasbourg, lors du meeting du conseil régional de l'EMEA, par Roy Tennant, gestionnaire principal des projets de l'OCLC[58]. Roy Tennant insiste particulièrement sur l'élaboration des « identités WorldCat » qu'il décrit comme algorithmiquement construite à partir de la base de données de WorldCat. Le principe des identités est de rassembler sur une même page l'ensemble des données concernant un auteur ou créateur, en les extrayant de la base de données à l'aide d'algorithmes et de programmes.

En 2017, OCLC a conduit un projet sur les données liées, permettant aux SID de créer des relations entre les ressources. Ont participé les responsables des métadonnées de plusieurs bibliothèques, des chercheurs et ingénieurs logiciel, afin de tester en conditions réelles un prototype consistant à attribuer aux différentes ressources un Uniform Ressource Identifier (URI), pour accéder directement à une description de données dans plusieurs sérialisations RDF (HTML, RDF/XML, Turtle, JSON-LD, N-Triples)[59]. Il s'agit d'un projet pilote fruit d'un partenariat entre des bibliothèques et les chercheurs, responsables produit et développeurs d'OCLC, auquel ont participé seize bibliothèques, en vue de tester un service de données liées

[58] TENNANT, Roy. *Managing the Digital Library*. New York: Reed Press, 2004. Print.
[59] OCLC. Projet pilote de données liées d'OCLC [en ligne], 2020. Disponible sur : https://www.oclc.org/fr/member-stories/ld-prototype.html [consulté le 25 Août 2020].

et optimiser le processus de description des ressources des bibliothèques par le biais des métadonnées. Ainsi, un prototype d'un service de comparaison a été mis en place, pour permettre de rechercher, distinguer ou modifier les données liées.

OCLC a développé une multitude d'applications web évolutives et disponibles, en s'appuyant sur les données de WorldCat grâce aux technologies big data inscrites dans le processus d'exploitation des bibliothèques dans un environnement cloud, permettant un traitement rapide et distribué de données sur différents serveurs.

Confrontés à un environnement en pleine mutation et à des utilisateurs plus exigeants, les services d'information documentaire ont amélioré leurs capacités en matière de partage des données et optimisé les chaînes de travail grâce à la plate-forme WorldShare. L'objectif est d'accéder à des données et fonctionnalités, visant à améliorer la visibilité et l'utilisation des collections à travers un environnement composé de plusieurs services Web[60] :

- **Recherche et expérimentation :** il s'agit d'aider les SID à représenter les relations entre les éléments bibliographiques et refondre les vedettes-matières, ainsi que transformer des fichiers d'autorités de bibliothèques en données liées ;
- **Les données et fonctionnalités de WorldCat :** elles permettent aux SID la gestion des métadonnées et l'accès aux données relatives aux ressources électroniques et aux bibliothèques ;
- **Les services web de prêt entre bibliothèques WorldShare :** ils aident les SID à partager des articles et les politiques de prêt ;
- **Les données et fonctionnalités des Services de gestion WorldShare** : elles sont destinées à optimiser les opérations des SID et les professionnels de l'information (les données liées aux acquisitions, aux prêts, aux activités de circulation et aux informations sur les comptes des abonnés) ;
- **WorldShare Conception de rapports** : il permet d'utiliser les données des SID, afin de générer des rapports statistiques personnalisés et des visualisations pertinentes. Ce module aide à prendre des décisions basées sur les données. Par exemple, il est possible pour un SID d'évaluer la pertinence des abonnements à certaines bases de données, déterminer les

[60] Ibid., p.72.

documents les plus consultés ou encore visualiser les données de prêt au sein d'un consortium.

Nous allons insister dans les parties suivantes sur les enjeux récents de certaines données dans les SID pour **l'interopérabilité des bibliothèques avec l'univers du Web, pour la qualité de la recherche d'information, pour l'enrichissement des catalogues en ligne, et enfin pour l'aide à la décision facilitée par la visualisation des données d'activité[61].**

Les données des SID sont actuellement enregistrées sous forme de notices en format lisible par machine (MARC). Il s'agit d'un standard métier établi dans le monde entier depuis plus de 50 ans. Toutefois, plusieurs autres acteurs de l'information utilisent des formats XML ou d'autres formats plus récents.

Les données ouvertes reliées ou *Linked Open Data* (LOD) adossées au Web se fondent, quant à elles, sur la publication des données sous une forme structurée et liée, en se basant sur de nouveaux formalismes tels que le *Ressource Description Framework* (RDF), qui permet de décrire les ressources web et leurs métadonnées à travers un standard Web partagé. Ainsi, des données bibliographiques de structures très diverses (MARC, DublinCore, EAD, etc.) peuvent être converties en un format unique basé sur le RDF, qui permet une recherche fédérée de façon plus performante que lors de l'interrogation de plusieurs bases de données séparées. Par exemple, la Bibliothèque Nationale de France (BNF) a introduit un service data.bnf.fr (http://data.bnf.fr/) sous ce format, qui permet d'agréger des résultats d'une requête de plusieurs sources simultanément (le catalogue des imprimés, la bibliothèque digitale et d'autres données non issues des catalogues traditionnels, etc.)[62].

L'utilisation d'identifiants uniques et pérennes pour les données contribue aussi à créer des relations stables entre les ressources. Aussi, plusieurs dispositifs ont contribué à l'amélioration de la visibilité des données des SID. A titre d'illustration, presque 80 % des visiteurs du service

[61]SIMONNOT, Brigitte. *L'accès à l'information : Moteurs, dispositifs et médiations*. Hermès Lavoisier, 2012, 249 p. Systèmes d'information et organisations documentaires. ISBN 978-2-7462-3829-9.

[62] BERMES, Emmanuelle. *Le Web sémantique en bibliothèque.* Avec la collaboration d'Isaac Antoine, Poupeau Gautier. Éditions du Cercle de la Librairie, « Bibliothèques », 2013, 176 p.

data.bnf.fr proviennent des moteurs de recherche, alors que d'autres arrivent sur le catalogue à travers des liens tiers[63].

3. La mise en place d'un projet de Linked Open Data (LOD)

La mise en place d'un projet Linked Open Data (LOD) en un service d'information documentaire, commence par la conversion des données préexistantes sous un format RDF (Ressource Description Framework). Un LOD exploite les apports des données structurées aux possibilités de recherche et aux services destinées aux utilisateurs des unités documentaires.

Il existe des bonnes pratiques inspirées de projets LOD, il s'agit notamment des projets[64] suivants :

- Le réseau de bibliothèques suédois Libris 2009 : il a mis en place une application web transformant les notices, pour publier les données de son catalogue sous le format RDF ;
- La Bibliothèque nationale espagnole 2013 : elle a développé un outil nommé MARiMbA dans le cadre de son projet Linked Data, permettant la conversion des données MARC21 en RDF ;
- Le projet européen LOD2 : le déploiement des technologies sémantiques, pour rendre les données publiques plus accessibles et les intégrer sur le web.

[63] HÜGI, Jasmin et PRONGUÉ, Nicolas. *Les bibliothèques face aux Linked Open Data* [en ligne]. Haute école de gestion de Genève, 2014.Disponible sur : http://doc.rero.ch/record/209598/files/M7-2014_memoire_HUGI-PRONGUE.pdf [Consulté le 24 Août 2020].

[64] PROJET LOD-B 1. *Processus de transformation des données en RDF : analyse des best practices* [en ligne], 2015. Disponible sur : http://campus.hesge.ch/id_bilingue/projekte/lodz/doc/processus_transformation.pdf [Consulté le 24 Août 2020].

Le processus d'un projet LOD se déroule, selon Hugi et Prongué[65], en huit grandes étapes :

Etape 1 : la revue de la littérature

Il s'agit de dresser un état de l'existant sur le web sémantique des services d'information documentaire, en rendant compte des évolutions techniques en matière des standards et modèles relatifs aux données bibliographiques.

Etape 2 : l'analyse des données

L'analyse des données dans un projet de LOD a pour objectif de déterminer les données qui seront publiées, en fonction des critères prédéfinis : la pertinence, la qualité, la quantité, etc.

Etape 3 : la modélisation

Cette étape correspond à la conception du modèle de données retenu et sur lequel reposent les différentes fonctionnalités dans une application sémantique.

Les données traditionnelles peuvent être structurées suivant un modèle très simple en notices bibliographiques et en notices d'exemplaires, ou un modèle innovant basé sur trois niveaux : œuvre, instance et item. L'IFLA (*International Fedration of Library Association*) a conçu, à son tour, un autre modèle baptisé FRBR (*Functional Requirements for Bibliographic Records*) organisé en quatre niveaux distincts : l'œuvre, l'expression, la manifestation et l'item.

Selon le modèle choisi, chaque ressource devra ensuite recevoir un identifiant de type URI et les données peuvent être reliées entre elles sous forme d'entités.

Etape 4 : le mapping

Il faut ensuite choisir un « vocabulaire » du web sémantique, dans le but de décrire au mieux les données à publier. Il est recommandé de recourir aux « vocabulaires » déjà existants sur le web, tels que le DublinCore (DC), pour décrire des ressources bibliographiques, Friend Of A Friend (FOAF) pour décrire les personnes et leurs relations et Simple Knowledge Organization

[65] HÜGI, Jasmin et PRONGUÉ, Nicolas. *Le virage Linked Open Data en bibliothèque : étude des pratiques, mise en œuvre, compétences des professionnels* [en ligne]. Haute école de gestion de Genève, 2014.Disponible sur : http://www.ressi.ch/num15/article_100 [Consulté le 24 Août 2020].

System (SKOS) destiné aux systèmes de connaissance, tels que les thésaurus, pour veiller à une meilleure interopérabilité des données. Le choix des vocabulaires doit tenir compte de leur pertinence et leur niveau d'utilisation.

Etape 5 : les liens externes

Ces liens pointent vers des référentiels du web sémantique disponibles en LOD et spécialisées dans un domaine. Ainsi, pour les descriptions de personnes et d'organisations, les liens peuvent renvoyer vers le référentiel VIAF (Virtual International Authority File) ou directement vers des fichiers d'autorités du SID[66].

La génération de liens externes peut se faire directement à travers des codes ou des descripteurs contrôlés préexistants, par exemple les codes du format MARC (MAchine Readable Cataloging).

Etape 6 : la transformation

Cette étape du processus consiste à remplacer le langage naturel/humain par un langage informatique, pour transformer les données.

Etape 7 : le contrôle qualité

Une fois les premières données RDF créées, il faut effectuer un contrôle rigoureux par le biais d'échantillons représentatifs des données, pour faire face aux erreurs.

Etape 8 : la publication

Il s'agit de la publication des données sur le web sous forme de fichiers téléchargeables depuis un serveur. Pour être des données LOD, elles doivent respecter les standards du web sémantique et se conformer aussi à une licence ouverte (Open Data).

En résumé, un projet LOD est une nouvelle opportunité d'optimiser l'usage des métadonnées, qui remet en question les fondements mêmes de la recherche d'information, tels que nous les connaissons actuellement. L'utilisation de méthodes de créativité ou la recherche exploratoire

[66]DUNSIRE, Gordon el al. Linked Data vocabulary management: infrastructure support, data Integration, and interoperability. *Information standards quarterly*. 2012. Vol. 24, n° 2/3, p. 4-13.

peuvent être de bons moyens, pour découvrir des opportunités de plus-value encore ignorées que les LOD peuvent offrir aux services d'information documentaires.

4. De meilleures possibilités de recherche

Les données structurées peuvent être liées à d'autres jeux de données de provenance externe, ce qui permet des requêtes enrichies et plus pointues. D'ailleurs, c'est le cas des données géographiques comme la latitude et la longitude, qui permettent de mettre en place de fonctionnalités de recherche au moyen d'une carte géographique. Un autre exemple concerne l'exploitation des liens entre personnes[67], comme en témoigne la recherche qui suit[68] :

Figure 9 : La recherche de relation entre deux personnes sur le portail finlandais Kulttuurisampo (Hugi, 2014)

[67]WENZ, Romain. Linked Open Library Data in practice: lessons learned and opportunities for data. *SWIB, 2012* [en ligne]. Cologne. 27 novembre 2012. Disponible sur <http://swib.org/swib12/slides/Wenz_SWIB12_108.ppt> [Consulté le 20 Octobre 2015].
[68]HUGI, Jasmin. Le virage linked open data en bibliothèque : étude des pratiques, mise en œuvre, compétences des professionnels. *Revue Electronique Suisse des Sciences de l'Information (RESSI)* .2014, n° .14.

La figure montre qu'une recherche entre les deux peintres Rubens et Miró, permet d'afficher les différents liens qu'entretient le premier peintre avec toutes les personnes enregistrées sur la base de données. Le second peintre est lié au premier par une relation de collaboration.

Les données structurées au format RDF permettent aussi une meilleure sérendipité, en permettant à l'utilisateur un accès à des ressources complémentaires comme des notices d'autorité d'autres bibliothèques ou des articles de Wikipédia. Le Centre Pompidou à Paris a relié toutes les données grâce au format RDF. De cette manière, une seule requête sur une personne retourne une liste des livres, des œuvres d'art, des dossiers pédagogiques et des événements en lien avec cette personne, comme l'illustre l'exemple ci-après [69] :

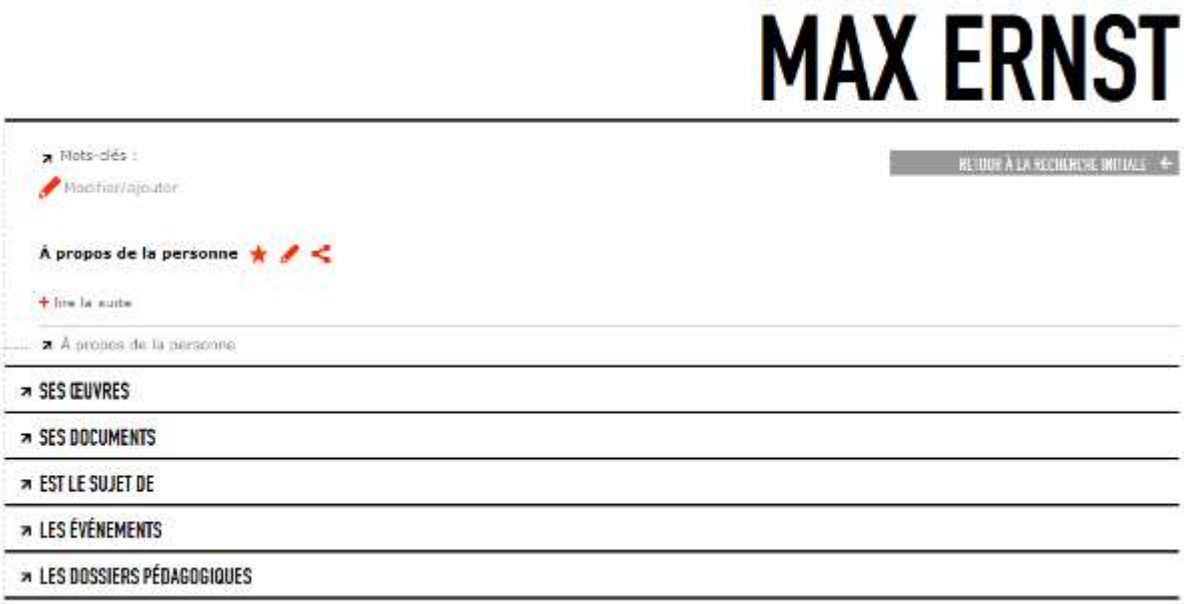

Figure 10: Une seule interface pour les ressources variées du Centre Pompidou, sur Max Ernst (Hugi, 2014)

Ces applications sont sous format RDF et aident l'utilisateur à trouver des informations qu'il ne cherchait pas dans ses requêtes initiales.

[69] Ibid., p.78.

5. Les catalogues documentaires augmentés

En 2007, une nouvelle génération des OPAC augmentés a vu le jour profitant notamment des mutations du web 2.0, qui véhicule de nouveaux outils et techniques au service des SID (répertoire de ressources, listes de notices, signalement de nouvelles acquisitions, foire aux questions, etc.).

Le catalogue augmenté est un catalogue enrichi, qui permet aux internautes de contribuer aux contenus et d'échanger entre eux avec la possibilité de personnalisation des services et des informations, le crowdsourcing et la sérendipité. Il présente une panoplie de fonctionnalités, telles que la page de couverture, le sommaire ou le résumé, la navigation à facettes, les notes, les critiques et le nuage de mots clés.

Aussi, Les OPAC augmentés utilisent des versions augmentées (couvertures d'ouvrages, extraits, commentaires d'usagers, etc.) comparés aux catalogues classiques, ainsi que l'émergence des outils d'appropriation ou d'externalisation d'écriture et de communication à travers l'exploitation de solutions grand public comme Netvibes ou Delicious, qui permettent la matérialisation de la bibliothèque dans les environnements numériques de ses utilisateurs de type Facebook.

Les Services d'Information Documentaire (SID) ont fait face à la surabondance de l'offre culturelle dans les différents canaux numériques, ainsi que la large accessibilité aux ressources informationnelles via internet. C'est ainsi que les SID ont emprunté les dispositifs de communication et de production de contenu offerts par le Web social à travers des applications et services 2.0 (blogs, wikis, flux RSS, indexation sociale, réseaux sociaux, etc.).

Avec les catalogues augmentés, les utilisateurs sont passés du stade d'un simple consommateur de l'information à un consommateur actif, placé au cœur d'une démarche interactive, personnalisée et ultra-ciblée. Les notions d'intelligence collective, de partage des savoirs et d'interopérabilité des applications ont dessiné de nouveaux contours à l'information en ligne ou en intranet, dans la mesure où les modes de collecte, de traitement et de diffusion des ressources informationnelles et documentaires associés au web classique sont ainsi fondamentalement remis en cause, comme le montre le tableau ci-dessous :

Fonctions du catalogue augmenté	L'explication
Web 2.0	Les sites communautaires, les blogs, les outils de partage de signets et les agrégateurs de fils RSS donnent la priorité à l'utilisateur plutôt qu'à la ressource diffusée.
Modèle de recherche d'information	Le modèle linéaire est en passe de disparaitre pour les modèles navigationnels basés sur les liens et l'individualisation des nœuds présents sur les différents réseaux sociaux. La navigation ne concerne pas uniquement les documents, mais s'étend pour concerner aussi les relations entre personnes, groupes ou communautés. Le système de navigation à base de pagination souvent fastidieux a été remplacé par le *scroll infini* plus pratique et adapté à tous les dispositifs mobiles (*responsive design*). Il repose sur la présentation de l'ensemble du contenu sur une page unique, qui convient juste de la parcourir du haut jusqu'au bas.
Analyse des ressources	L'utilisateur peut décrire des ressources à travers le contenu d'un billet sur un blog ou caractériser une photo ou une vidéo sur des plateformes dédiées voire même indexer librement des contenus par l'entremise des tags (folksonomie). Dans cette optique, chaque individu mobilise ses propres schémas de représentation mentale au lieu de se référer à un cadre théorique et à des outils préétablis.
Evaluation	L'utilisateur peut apprécier et émettre son avis qualitatif ou quantitatif sur les différents contenus, ce qui lui confère le statut d'un évaluateur voire un prescripteur. D'un point de vue collectif, une ressource popularisée sur les réseaux sociaux contribue à sa visibilité et élargit sa diffusion à grande échelle et sur le plan individuel, les feedbacks, les appréciations et les activités d'un internaute peuvent être suivis par l'ensemble de la communauté.
Actualités, personnalisation et recommandation	L'intégration d'un fil général d'actualité culturelle ou de la bibliothèque à la page d'accueil du catalogue et un agenda culturel, ainsi que la personnalisation de l'interface et des services selon un paramétrage de l'utilisateur, qui peut se voir recommander les documents correspondant à ses préférences et historique de consultations.
Les vignettes, les tables des matières, les liens externes et le texte intégral	L'association des notices bibliographiques aux vignettes de couverture des documents et les tables des matières sur les catalogues en ligne, permet une meilleure identification visuelle du document et une idée générale de son contenu.

	De même, les catalogues augmentés permettent de pointer vers la version électronique du texte intégral du document ou permettre son téléchargement sous format PDF, texte ou image depuis la notice courante du document ou l'article.
Echange et importation de notices	L'utilisation des normes et des formats unifiés de catalogage et des protocoles d'échange de notices comme celui de Z39.50 a facilité la réutilisation des notices clé en main en provenance d'autres catalogues. Possibilité d'importer les fiches des métadonnées depuis les bases de données des éditeurs sans avoir à les saisir manuellement, ainsi que les notices des archives ouvertes par l'intermédiaire des protocoles d'interrogation et de moissonnage des serveurs d'archives ouvertes, à l'image de OAI-PMH (Open Archives Initiative-Protocol Metadata Harverster).

Tableau 4 : Les fonctions remplies par un catalogue documentaire augmenté

Pour illustrer les fonctionnalités portées par les catalogues augmentés et présentées dans le tableau ci-dessus, prenons quelques exemples :

Dans le cadre de l'appel à projet 2010 « Services numériques culturels innovants » lancé par le ministère de la Culture et de la Communication, la bibliothèque de Toulouse a développé une application offrant l'accès à son catalogue adapté aux smartphones, ce qui correspond à une fonctionnalité de personnalisation et d'adaptation de l'affichage des catalogues à toutes les tailles des écrans des dispositifs mobiles, permettant ainsi aux utilisateurs de s'affranchir de la contrainte spatio-temporelle et consulter aisément les références relatives aux documents qu'ils cherchent, ou encore la Bibliothèque du Congrès qui assure une large diffusion de ses archives vidéo sur Youtube et celles sonores sur le portail iTunes.

De surcroit, les réseaux sociaux ont créé une relation privilégiée de proximité entre le professionnel de l'information et les utilisateurs, c'est le cas des bibliothèques universitaires d'Angers, disposant de deux pages Facebook fédérant plus de 1 000 amis, qui animent des débats sur les services qui leur sont offerts[70].

De plus, les utilisateurs peuvent désormais déposer des annotations, des tags ou des commentaires par rapport aux documents disponibles sur l'OPAC, comme c'est le cas de

[70] DUJOL, Lionel. CRÉER DES SERVICES INNOVANTS. Stratégies et répertoires d'action pour les bibliothèques, Partie II. *Impact du numérique sur l'offre de la bibliothèque : nouvelles approches professionnelles*, 2011.pp 70-80.

WorldCat, qui associe les notices des documents à des rubriques dédiées aux lecteurs, pour émettre une critique ou un ajouter un marqueur ou encore HubMed, qui représente un service web de la base de données PubMed, permettant des nouvelles fonctionnalités comme le « tagging » ou la catégorisation par facettes.

S'agissant de l'intégration du texte intégral dans la notice du catalogue augmenté, la bibliothèque de l'Université de Michigan aux Etats-Unis donne l'accès à cette version, comme l'illustre cette capture de la page d'accueil de son catalogue en ligne[71] :

Figure 11: Le lien vers le texte intégral d'un document dans la notice du catalogue en ligne de la bibliothèque de l'Université de Michigan (KAENEL De, 2007)

Il s'agit de visualiser le document numérisé en format image, texte ou PDF dans sa propre plateforme digitale à partir d'un lien figurant dans la notice complète du document.

Un autre exemple plus englobant, le catalogue en ligne « Saphir » spécialisé dans la documentation de la santé publique en Suisse utilise sur son catalogue en ligne les services web, pour afficher l'image de couverture sur la page de la notice complète des livres, comme l'indique la capture suivante[72]:

[71]KAENEL DE, Isabelle et IRIARTE, Pablo. Les catalogues des bibliothèques : du web invisible au web social (I), *Revue électronique Suisse des sciences de l'information* [en ligne], 2007. Disponible sur : http://www.ressi.ch/num05/article_028 [consulté le 27 Août 2019].

[72] Ibid.

Figure 12 : L'intégration du résumé, de la table des matières et de la vignette dans le catalogue en ligne Saphir spécialisé dans la documentation dans le domaine de la santé Publique en Suisse (KAENEL DE, 2007)

L'affichage automatique des vignettes des couvertures des documents sur un catalogue en ligne pour une meilleure identification visuelle pour les utilisateurs, consiste à établir le lien entre la technique AJAX (Asynchronous JavaScript and XML) qui permet à travers l'ISBN ou le code de langue du document en question, d'exécuter un programme écrit en langage de programmation JavaScript, incorporé dans une page web, est exécuté par le navigateur qui change une partie de la page, en y affichant ces vignettes.

6. La visualisation des données d'activités du SID : l'exemple du projet « Prévu »

Pour illustrer, un service assimilable à une aide au pilotage par l'analyse et la visualisation des données du SID, nous prendrons l'exemple du projet « Prévu », initié en 2015, par un doctorant

de l'université Paris 8, ayant pour objectif de visualiser les prêts, d'où le titre du projet « Prévu »[73].

L'accès aux données bibliographiques et de circulation des livres a permis de fédérer les données de prêts de la BU, après anonymisation du lectorat, conformément aux recommandations de la CNIL[74] :

Figure 13: La visualisation des données sur la plateforme « Prévu » (http://www.prevu.fr)

A partir de la figure ci-dessus, nous remarquons que les scénarios de visualisation fournissent des informations factuelles des données de prêt recueillies et du nombre de transactions, reflétant les événements marquants pour le SID et l'université. L'historique des prêts et des records rend parfaitement compte des rythmes universitaires et des périodes d'examens et affiche une courbe décroissante des inscrits, qui s'accompagne par l'augmentation des emprunts dans le même temps**.**

[73] DARQUIE, Gaetan. L'écriture de récits interactifs : outils dédiés et enjeux de conception. Thèse en sciences de l'information et de la communication, Université Paris 8, 2015.

[74] PREVU. *Visualisation des statistiques sur la bibliothèque universitaire de l'université Paris 8*[en ligne]. Disponible sur http://www.prevu.fr [consulté le 10 juin 2016].

Sur un autre registre, d'autres exploitations relatives au lectorat et ses pratiques sont devenues possibles :

- Qui sont les lecteurs ? Etudiants des 1ers cycles, doctorants, enseignants ?
- Quelle composante de l'université est la plus gourmande en documentation, ou à l'inverse, laquelle se fait plus discrète ?
- Le best-off de personnalités décisives dans l'histoire de l'institution (lecteurs actifs) ;
- Les stratégies documentaires de recherche des publics (les mots-clefs de recherche, repérer des typologies et des communautés de lecteurs) ;
- Les ouvrages les plus populaires de l'école doctorale et l'identification des secteurs où les livres sont les plus empruntés ;
- La plateforme Prévu met librement à disposition pour tout auteur ou tout ouvrage les données de prêt qui lui sont associées, ouvrant ainsi une voie toute tracée à l'ingénierie de la recommandation et des algorithmes.

Le projet « Prévu » traduit la libération des données, pour ne pas juste optimiser la circulation des emprunts, mais également les documenter et les valoriser. Les données d'usage du SID, en l'occurrence les données de prêts, permettent de ré-agencer et de redistribuer l'accès aux documents en fonction d'un ordre reflétant les intérêts intellectuels d'une communauté.

Plusieurs mécanismes sont déployés dans le but de se rapprocher de l'utilisateur : l'outil de recherche bibliographique « Prévu » propose un système d'indexation plus pratique donnant une vue d'ensemble sur les données générales du SID universitaire de Paris 8 ; l'ensemble des emprunteurs, l'ensemble des prêts effectués, etc. ce qui aide les utilisateurs à comprendre les tendances qui marquent la vie du SID à l'aide des représentations graphiques significatives[75].

Conclusion sur la croissance des données du SID

Il existe deux types de données : les données sur les données qui décrivent la nature des objets auxquels on se réfère, on les désigne comme des métadonnées. Cependant, en dehors de ces métadonnées, il existe les données relatives aux activités des SID : acquisitions, désherbages, fréquentations, circulations des documents, inventaires des collections, données de logs captées

[75] Ibid., p.85.

au moment où les utilisateurs se connectent au site internet de la bibliothèque ou bien à ses bases de données. Certaines données de ce dernier type sont assimilables à du big data, elles sont non structurées, volumineuses, arrivent en flux continu, nous reviendrons sur ces données dans le chapitre suivant.

Dans ce premier chapitre, nous avons aussi insisté sur le mouvement d'open data, et plus précisément sur le linked open data, marqué par l'échange, le libre partage et les liens des données. Les services d'information documentaire ont initié plusieurs initiatives sur l'ouverture de leurs données telles que les données bibliographiques, qui sont présentes sur le web de données et reliées entre elles, de façon à contribuer à une large réutilisation.

Chapitre 2 : Le Big data dans les SID

Introduction

Le phénomène du big data est l'un des grands défis informatiques de la décennie. Les données massives, ou big data, ont pris une importance particulière dans le contexte des services d'information documentaire d'une certaine ampleur pour les raisons suivantes :

- **Le coût du stockage :** Les solutions de *cloud computing* permettent une gestion des données élastiques suivant les besoins réels des entreprises ;
- **Les plateformes de stockage distribuées et les réseaux à très haut débit :** les données sont désormais stockées à des endroits distincts, et parfois non identifiés ;
- **Les nouvelles technologies de gestion et d'analyse de données :** Parmi ces solutions technologiques liées au big data, il faut citer Hadoop permettant le développement et la gestion d'applications distribuées de grandes quantités de données.

1. Le big data : dimensions et infrastructure

Le big data désigne donc un ensemble très volumineux de données qu'aucun outil classique de gestion de base de données ou de gestion de l'information ne peut vraiment stocker ou traiter.

Le recours quotidien aux services de télécommunication, aux services en ligne, aux équipements électroniques, les achats, la consommation de ressources produisent désormais de grandes quantités de données, auxquelles s'ajoutent celles que les autorités rendent accessibles dans le cadre de projets d'ouverture de données publiques (*Open Data*), certes de moindre envergure que les données d'usages cependant.

D'un point de vue économique, les données massives présentent un potentiel considérable. Chaque jour, nous générons 2,5 trillions d'octets de données, ces données ont de multiples origines : messages sur les sites de médias sociaux, images numériques et vidéos publiées en ligne, enregistrements transactionnels d'achats en ligne et de signaux GPS des téléphones mobiles…[76].

[76] LEMBERGER, Pirmin et BATTY, Marc. *Big data et Machine Learning : Les concepts et les outils de la data science.* 3e édition. Paris: Dunod, 2019. 272 p.

Le *big data* considère des données non structurées et hétérogènes pouvant être corrélées, ce qui en fait une source de nouvelles connaissances et de création de valeur pour les entreprises publiques et privées. Son utilisation concerne, par exemple, les études de marché automatisées et capables de réagir instantanément aux moindres fluctuations, la découverte d'abus concernant des transactions financières, les analyses web afin d'étendre et d'optimiser des campagnes de marketing, les diagnostics médicaux approfondis, etc.[77].

De surcroît, le big data permet aux entreprises de traiter de gros volumes de données en temps réel, et de surveiller le trafic réseau et d'analyser la qualité de service. Le big data s'avère un moyen d'anticipation des changements avec la possibilité de détecter les menaces et les opportunités, ainsi que les attentes des clients. En 2001, l'analyste du cabinet Meta Group (devenu Gartner) Doug Laney caractérisait le big data selon le principe des « trois V »[78] :

- **Volume :**

  Le big data est associé au volume important de données, oscillant entre des téraoctets et des pétatoctets ;

- **Vitesse :**

  Il s'agit de la fréquence de génération et de partage de données, qui concernent non pas leur création, mais leur traitement et diffusion aux utilisateurs en temps réel ;

- **Variété :**

  La diversité des données prend plusieurs formes : linguistique (plusieurs langues sur le web), structurelle (texte, image et vidéo…) et normative (XML, RDF…). Elle correspond aussi à la l'expansion des données semi ou non structurées.

Certains experts ajoutent un quatrième et cinquième « V », le premier étant la Véracité, qui correspond à la qualité et la crédibilité de la source du contenu afin de pouvoir exploiter ces données, et le second relatif à la Visibilité renvoie à la mise place de grandes plateformes big data, assurant une large utilisation des données.

---

[77] DELORT, Pierre. *« Big data : concepts et définition »,* éd., *Le Big data.* Presses Universitaires de France, 2015, pp. 29-47.

[78] CORINUS, Mickael ; DERY, Thomas. *Accès rapport d'étude sur le big data [en ligne]. Rapport d'étude sur le big data*. SRS Day, 2012. Disponible sur :<https://srs.epita.fr/wp-content/uploads/2014/02/big-data-rapport.pdf> [Consulté le 15 Septembre 2019].

Il s'agit ensuite de donner du sens aux données massives par le biais de l'analyse, en s'appuyant sur une infrastructure technologique adaptée, basée majoritairement sur Hadoop, plateforme open source réputée par sa puissance d'indexation, de transformation, de recherche ou d'élaboration de modèles sur de très gros volumes de données.

En ce qui concerne l'architecture, Hadoop fonctionne sur plusieurs serveurs, avec une grosse capacité de stockage. En pratique, Hadoop gère un système de réplication de façon à assurer une très haute disponibilité des données réparties sur les différents serveurs, palliant la défaillance possible de certains d'entre eux[79].

En effet, Hadoop comporte deux composants principaux[80] :

- **Hadoop Distributed File System** (HDFS) : le HDFS est un système de fichiers distribué, qui permet de stocker de très gros volumes de données sur un grand nombre de machines équipées de disques durs ;
- **Hadoop MapReduce** : un modèle de programmation pour le traitement de données à grande échelle.

De surcroît, la plateforme Hadoop est composée de plusieurs modules qui forment une architecture puissante résiliente aux pannes :

- **Apache Pig** : Pig est une plateforme de haut niveau pour la création de programmes MapReduce utilisés avec Hadoop ;
- **Apache HBase** : HBase est une base de données distribuée, non-SQL ;
- **Apache Spark** : Spark est un modèle de programmation qui n'est pas liée au modèle MapReduce de deux étages (Map et Reduce). Il promet des performances jusqu'à 100 fois plus rapides que Hadoop MapReduce, pour certaines applications ;
- **Apache Hive** : Apache Hive est une infrastructure d'entrepôt de données permettant l'analyse des données.

[79] Ibid., p.89.
[80] CORINUS, Mickael; DERY, Thomas. *Accès rapport d'étude sur le big data [en ligne]. Rapport d'étude sur le big data*. SRS Day, 2012. Disponible sur :<https://srs.epita.fr/wp-content/uploads/2014/02/big-data-rapport.pdf> [Consulté le 15 Septembre 2019].

2. Pouvons-nous parler de big data dans les SID ?

Les collections des Services d'Information Documentaire (SID) peuvent être réparties en quatre catégories : les abonnements de périodiques électroniques, les collections numériques patrimoniales, les archives rassemblées au titre du dépôt légal du Web, et les bouquets d'e-books, de titres de presse et de VOD.

Au fil du temps, les professionnels de l'information ont été amenés à s'adapter aux nouveaux supports de l'information du manuscrit à l'imprimé et aux documents numériques. Aujourd'hui encore, il s'agit de relever un nouveau défi, celui lié à la gestion des données massives ou le big data.

Le volume des collections numériques varie selon les SID. De fait, cette masse ne peut être assimilée aux gigantesques entrepôts de données factuelles de certains secteurs de recherche par exemple, ou même à ceux que gèrent les grands acteurs d'Internet. A titre d'exemple, la bibliothèque numérique américaine Hathi Trust, disposait en 2012 d'environ 450 Teraoctets pour dix millions de livres numérisés, à comparer avec les taux d'accroissement quotidiens d'acteurs comme Twitter (7 To par jour).

Il serait possible de qualifier de « big data » dans les SID, les données suivantes, même si la qualification reste discutable :

- Les données de *log* des sites web et des bases de données relatifs à l'utilisation des ressources documentaires et des services SID, en vue de proposer de nouveaux services aux usagers ou aux bibliothécaires ;
- Les données relatives aux statistiques d'usages permettant d'évaluer les différents produits et services en ligne, d'orienter les décisions d'acquisition, de planifier l'infrastructure technique et promouvoir les services de la bibliothèque. Or, ces données sont livrées avec une périodicité, un format et un traitement variables, hétérogènes, parfois à l'état brut ;
- Les données relatives aux rapports de fonctionnement, notamment les fréquentations physiques.

En effet, ce que nous pouvons considérer comme du big data dans les SID est plutôt lié aux données produites ou reçues sur les collections et les utilisateurs plutôt que la taille des

collections. Ces données (logs, statistiques, métadonnées, etc.) regroupent des données hétérogènes.

Pour illustrer ce croisement du big data avec les SID, prenons quelques exemples :

Le réseau des bibliothèques de Singapour composé de plusieurs bibliothèques publiques a développé l'utilisation du big data, en vue d'analyser les statistiques des prêts en relation avec les données bibliographiques et ainsi proposer aux utilisateurs des recommandations d'usage[81].

Le site (www.data.gouv.fr/paris) permet, selon Boustany[82], d'accéder aux données produites par les SID de la ville de Paris comme les collections des bibliothèques et fonds spécialisés et patrimoniaux, les titres les plus prêtés, etc., afin de les analyser et aider les professionnels de l'information à cerner les besoins des utilisateurs et à prendre les décisions pour améliorer les services fournis.

L'exploitation des données massives a aussi permis à de nouveaux services de voir le jour, telle que l'application mobile « Affluences », qui croise les données relatives aux horaires d'ouverture des SID à celles provenant des capteurs dans les files d'attente pour estimer en temps réel le temps d'attente prévisionnel[83].

Un autre exemple concerne le projet Counter (www.projectcounter.org) fournit des données d'usages des ressources numériques sous forme de rapports formatés et partagés permettant de mesurer l'utilisation des produits et services en ligne d'une manière crédible, cohérente et compatible, en s'appuyant sur les données fournies par le fournisseur.

[81] Figoblog. *Big Data et bibliothèques* [en ligne]. Disponible sur : https://figoblog.org/2015/01/13/big-data-et-bibliotheques/ [Consulté le 19 Juin 2020].
[82] BOUSTANY, Joumana. « Données massives, science et bibliothèques », Michel Netzer éd., *Les sciences en bibliothèque* [en ligne]. Éditions du Cercle de la Librairie, 2017, pp. 191-200. Disponible sur : https://www.cairn.info/les-sciences-en-bibliotheque--9782765415251-page-191.htm [consulté le 12 Août 2019].
[83] Ibid.

Le code Counter spécifie le format, le contenu, la périodicité et le mode de transmission des statistiques pour des revues et des bases de données il permet d'avoir des données précises sur les revues, comme le montre la figure 14[84] :

Exemple de *Journal report 1*

	ISSN print	ISSN online	Janvier 2001	Février 2001	Mars 2001	Cumul année
Total revues			6 637	8 732	7 550	45 897
Revue AA	1212-3131	3225-3123	456	521	665	4 532
Revue BB	9821-3361	2312-8751	203	251	275	3 465
Revue CC	2464-2121	0154-1521	0	0	0	0
Revue DD	5355-5444	0165-5542	203	251	275	2 978

Figure 14 : un exemple de reporting généré par le projet Counter
(www.projectcounter.org)

Ce code permet de générer plusieurs rapports relatifs aux interrogations sans accès au texte intégral, le format du document (HTML, PDF) ou le nombre de séances et les chiffres concernant les bases de données.

Le projet COUNTER peut être considéré du big data dans la mesure où il couvre l'utilisation des ressources électroniques des bibliothèques (revues, bases de données, livres, ouvrages de référence et bases de données multimédia) dans un contexte où le volume des données hétérogènes est important et la fréquence de mise jour des données est élevée.

Afin d'exploiter les statistiques de consultation dans l'identification des besoins réels des utilisateurs, OpenEdition plateforme des revues en libre accès, a mis en place un projet d'analyse des fichiers log des accès distants aux différentes revues, en vue d'analyser les flux de consultation grâce à des outils d'analyse big data, garantissant la fiabilité des données souvent compromise avec le moissonnage des robots[85].

A titre d'exemple, OpenEdition s'appuie sur des outils d'analyse d'usage comme Google Analytics, pour avoir une vue sur le trafic des visiteurs de la plateforme, leurs parcours en ligne,

---

[84] *PROJET COUNTER* [en ligne]. Disponible sur : https://www.projectcounter.org/. [Consulté le 28 Août 2019].

[85] FRANCO, Daniele. « Exploiter les données d'usages en bibliothèque : pourquoi faire ? : journée d'étude Enssib – Lyon, 14 janvier 2016 » [en ligne], *Bulletin des bibliothèques de France (BBF)*, 2016, n° 7. Disponible sur : https://bbf.enssib.fr/tour-d-horizon/exploiter-les-donnees-d-usages-en-bibliotheque-pourquoi-faire_65839 [Consulté le 31 Août 2020].

leurs sessions de connexion, le rythme de leurs consultations, les articles et revues les plus lus, etc.

Enfin, le projet EzPaarse né de la collaboration entre le consortium Couperin, le CNRS-Inist et l'université de Lorraine, a permis aux SID universitaires d'accéder à une meilleure information sur les ressources numériques et leurs utilisateurs. Il s'agit d'utiliser un proxy, une machine qui sert d'intermédiaire entre les machines d'un réseau et l'Internet, pour analyser les connexions aux ressources numériques, des graphiques sont produits chaque mois permettant de rendre compte des indicateurs d'accès. En pratique, plusieurs outils ont été sollicités comme Omniscope destiné au traitement des données statistiques, Ezagimus utilisé pour construire un entrepôt de données d'utilisation dédié au stockage sur le long terme et Kibana pour la visualisation des tableaux.

L'ouverture des SID sur la dimension du big data permet de disposer des instruments et des méthodes d'évaluation et de pilotage pertinents. De même, la visualisation facilite l'exploitation de l'analyse de données et rend compte des résultats et de l'activité du SID. Néanmoins, le big data dans les SID demeure discutable, du moment où ils n'empruntent que certaines techniques du big data et souvent ne traduisent pas la règle de 3 V dans le détail.

3. Le développement du data mining dans les SID

Le data mining se définit comme étant une exploration de données ou encore extraction de connaissances à partir de données, qui a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques[86].

L'exploration de données est assimilée à un destiné à extraire des informations intéressantes à partir des grandes quantités de données, qui sont stockées dans les bases de données ou d'autres formes de dépôts.

En principe, l'exploration de données peut être descriptive ou prédictive :

[86] Définition du datamining, [En ligne]. Disponible sur : https://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es [Consulté le 22 Septembre 2020].

- L'exploitation descriptive des données : elle renvoie au processus d'exploitation où l'information ou la connaissance utiles sont extraites à partir des bases de données ;

- L'exploitation prédictive des données : elle implique l'utilisation de l'apprentissage automatique, pour prédire des valeurs futures et certaines variables à partir des données stockées dans des bases ou des entrepôts de données.

En effet, les deux fonctions évoquées ci-dessus sont implémentées après un processus d'extraction de données désignée par le terme ETL [87] (fonction d'intégration de données consistant à extraire des données à partir d'un certain nombre de sources et de les transformer, et finalement de les stocker dans un entrepôt de données pour les préparer au data mining.)

Le data mining peut permettre aux SID l'amélioration des services à travers :

- La relation entre les données sur le prêt et les futures acquisitions des ressources documentaires ;
- La relation entre les données caractérisant les utilisateurs (démographiques, géographiques, socioprofessionnelles, etc.) et les ressources documentaires ;
- Les catalogues et les services en ligne d'une bibliothèque fournissent aux professionnels d'information un large ensemble de données sur comment et quand les services des SID sont exploités par les usagers, en vue de les optimiser et les améliorer.
- Les données cumulées sur la circulation des collections d'un SID, peuvent être déployées par le biais du datamining, pour en ressortir les tendances futures relatives aux besoins des utilisateurs[88].

Le data mining est une technique algorithmique qui s'est particulièrement développée avec le big data mais et elle s'applique aussi à des données structurées textuelles. Le rôle des SID ne se limite pas uniquement à trouver l'information, mais il peut concerner aussi l'analyse d'évolution des thèmes qui sont traités dans un corpus.

Les analyses data mining appliquées à des données textuelles structurées dans les SID peut fournir des visualisations représentant les fréquences d'apparition des termes au cours de

[87] Extract, Transform, Load.

[88]CHIGNARD, Simon. *Comprendre l'ouverture des données publiques*. FYP Editions. Collection entreprendre, 2012.

l'existence d'un document, et comparer la fréquence d'emploi des mots dans un périodique donné, pour extrapoler des tendances[89].

L'application des méthodes de fouille de texte et de données à la gestion des collections tend vers la complémentarité des techniques documentaires classiques consistant à indexer un document par le biais d'un plan de classement ou une liste d'autorité par de nouvelles techniques de fouilles de texte et de données permettant à une collection d'être organisée, en s'appuyant sur les algorithmes de « modélisation de sujet » ou de « topic modeling », dont l'objectif est d' interpréter les thématiques présentes dans un corpus et produire des listes de vocabulaires, où se trouvent rassemblés des mots dont la fréquence d'utilisation est corrélée. Ainsi présentées, ces listes mettent en avant les principaux thèmes traités dans le corpus, et surtout permettent de localiser une thématique précise abordée.

Nous allons illustrer nos propos par quelques exemples désormais.

Le projet américain intitulé « Mapping text » avait pour dessein de mettre en place une plateforme dédiée à la recherche et à la visualisation sur une collection de périodiques de l'Université du Nord Texas et l'Université de Stanford[90]. Le corpus numérisé représente 232 567 pages de gazettes dont la publication s'étale de 1829 à 2008 intégrait dix modèles de sujets composés de cent mots chacun, présentés aux usagers, en fonction de leur degré d'importance. De fait, les techniques de TDM ont permis d'extraire ces sujets à partir des mots-clés et leur importance dans un corpus important.

Dans le même ordre d'idée, pour Rostaing, le classement automatique supervisé des textes est sollicité dans le cas des collections volumineuses, pour identifier les documents ayant des caractéristiques communes après avoir entraîné l'algorithme à partir d'un corpus de départ. Par exemple, l'Institut National de Recherche Agronomique (INRA) a employé cette technique de classification automatique de textes basée sur les techniques de TDM supervisées ou un corpus de départ est sollicité pour entrainer l'algorithme, qui cherchera à regrouper dans un ensemble

---

[89] MASANES, Julien. *La magie des séries et des grands nombres : ce que l'on peut faire avec le Big Data et les archives.* Actes du séminaire IST Inria, octobre 2014.

[90] GILLIUM, Johann. *Big data et bibliothèques : traitement et analyse informatiques des collections numériques* [en ligne]. Mémoire d'étude. Enssib, 2016. Disponible sur : http://microblogging.infodocs.eu/wp-content/uploads/2016/04/66017-big-data-et-bibliotheques-traitement-et-analyse-informatiques-des-collections-numeriques.pdf [consulté le 22 Septembre 2020].

plus vaste les documents présentant des caractéristiques similaires. Ainsi, Il devient possible de retrouver des documents répondant à des caractéristiques précises.[91]

Une autre exploitation du data mining concerne la reconnaissance d'entités nommées, visant l'enrichissement des métadonnées des documents par l'alimentation semi-automatisée de thésaurus par l'identification de termes candidats qui, après validation d'un spécialiste, peuvent alimenter un thésaurus ou une ontologie[92]. Les textes des documents suivent un processus de traitement et d'analyse, allant de la séparation du texte en phrases et en unités simples jusqu'à l'étiquetage morphosyntaxique. A l'issue de ce travail, des extractions peuvent être effectuées, de façon à aider un analyste humain à déterminer les entités les plus pertinentes pour alimenter le thésaurus.

Un autre exemple concerne la plateforme ISTEX[93], permet à la communauté de l'enseignement supérieur et de la recherche française, d'accéder en ligne aux collections rétrospectives de la littérature scientifique multidisciplinaire et de les exploiter grâce à des techniques de fouille de textes[94] .

Ainsi, le type d'entités concernées par l'indexation automatique varie des noms de personnes aux lieux et dates. Pour le projet ISTEX, la reconnaissance des entités nommées à travers les mots situés dans leur contexte, permet un balisage automatique favorisant une recherche plus performante sur les corpus[95].

De plus, les techniques d'exploration des données ou de data mining prédictif sont utilisées pour créer des modèles d'apprentissage automatique (Machine Learning, ML), alimentés par les données disponibles et permettant de prédire les valeurs reliées de multiples variables. C'est le cas des algorithmes de moteur de recherche et les systèmes de recommandation qui visent des fonctions de personnalisation à partir de l'historique de l'utilisateur.

---

[91] Ibid., p.96.
[92] NOUVEL, Damien ; SOULET, Arnaud, Jean-Yves Antoine [el al.]. Reconnaissance d'entités nommées : enrichissement d'un système à base de connaissances à partir de techniques de fouille de textes. *Traitement Automatique des Langues Naturelles*. Jul 2010.
[93] https://www.istex.fr/
[94] BERARD, Raymond. Istex, vers des services innovants d'accès à la connaissance [en ligne]. Montpellier : ABES, 2012. Disponible sur : http://www.abes.fr/Ressources-electroniques2/Acquisitions/Licences-nationalesISTEX [consulté le 23 Septembre 2020].
[95] Ibid.

Les systèmes de recommandation pour les SID reposent sur les données structurées, provenant du catalogue en ligne du Système Intégré de la Gestion des Bibliothèques (SIGB) dont les notices bibliographiques et leurs métadonnées, la liste des abonnés du SID et l'historique des transactions de circulation et les données non structurées composées des systèmes d'évaluation explicite des usagers et les données *log.*

4. L'impact du big data sur l'évolution des compétences des professionnels des SID

Les professionnels de l'information dans les SID ont besoin de savoir comment tirer parti des différents types de données produites aujourd'hui dans le cadre des SID, et dans quelle mesure ces données peuvent contribuer à l'amélioration des services.

Ces professionnels ont besoin de savoir en quoi ces données volumineuses diffèrent des autres données traitées jusqu'à présent et ils ont besoin de comprendre les méthodes d'analyse et les logiciels/matériels qui sont mobilisés.

Au-delà de la gestion simple des documents, le big data interpelle, pour Collet-Thireau[96], les professionnels de l'information à développer une culture des données, qui repose sur de nouvelles compétences de plusieurs ordres :

- ✓ Les compétences informatiques : comprendre le web, la structure des documents numériques et les systèmes d'information et diffuser des données sur un portail ou une plateforme, tout en leur assurant un accès performant et les présentant sous forme de visualisations pertinentes ;

- ✓ Les compétences relatives au traitement des données : comprendre les sources de données (les bases de données, les progiciels intégrés de gestion, les fichiers logs, les cookies, les entrepôts de données, le cloud, etc.) et utiliser des méthodes de data mining

[96] COLLET-THIREAU, Katell et THOMAS, Jean-Paul, « Big Data et Open Data : quel impact pour les professionnels de l'information ? », *I2D – Information, données &* documents [en ligne]. 2015/4 (Volume 53), p. 9-10. Disponible sur : https://www.cairn-int.info/revue-i2d-information-donnees-et-documents-2015-4-page-9.htm [consulté le 24 Septembre 2020].

pour extraire de la valeur des données, les mettre en forme et les décrire à travers les métadonnées ;

- ✓ Les compétences juridiques : connaître les lois qui régulent le domaine d'exploitation des données (code de la propriété intellectuelle, les données personnelles, le Règlement Général sur la Protection des Données, droit d'auteur, etc.) ;

- ✓ Les compétences managériales : savoir gérer et animer un projet dans une approche systémique marquée par l'esprit d'équipe, tout en développant une culture de la donnée auprès des différentes parties prenantes.

En matière de formation et de développement des compétences, plusieurs stratégies sont aussi à favoriser :

- Conclure des partenariats avec des entreprises spécialisées ;
- Se familiariser avec les nouvelles méthodes de modélisation des données ;
- Maitriser le stockage distribué et le traitement parallèle de données ;
- Acquérir des services big data en mode cloud auprès de fournisseurs. A titre d'exemple, il est possible d'intégrer de fonctionnalités big data dans des applications métiers.

Conclusion

L'exploitation de la data au sein des industries culturelles d'une manière générale, et au sein des services d'information documentaires plus particulièrement, affecte l'identification et l'appréhension des opportunités de création de valeur.

L'apport réel de la data pour l'accès à la connaissance dans les SID correspond à l'évolution d'un accès arborescent et hiérarchique à un accès fragmentaire et multidimensionnel. La data peut aider à créer de nouveaux services et améliorer la qualité de services existants, en tenant compte de l'analyse et de la visualisation de données en temps réel. L'objectif est ici d'améliorer la prise de décision et d'approfondir la connaissance d'un sujet (par exemple via des tableaux de bord ou des résultats de datamining).

La disponibilité de données de plus en plus volumineuses ainsi que les outils d'analyse qui y sont associés ont permis une grande évolution des services d'information documentaire : moteurs de recherches, systèmes de recommandation, fouilles de textes, visualisations permettant une exploration optimale des connaissances avec la possibilité de dévoiler des liens entre des domaines de connaissance auparavant éloignés.

Dans cette logique, un SID peut s'adapter à différents contextes et créer des valeurs différenciées. La production de nouvelles données engendrera encore de nouveaux services, c'est par exemple le cas du développement des applications pour les dispositifs mobiles générant des flux de données en temps réel sur les usages et leurs utilisateurs. Ces données pourront être déployées dans le contexte des SID, pour créer des services innovants en vue d'optimiser et personnaliser l'accès des utilisateurs aux ressources disponibles.

Partie III : Les systèmes de recommandation et leur développement

Chapitre 1 : La modélisation des systèmes de recommandation, introduction

Dans le contexte des évolutions technologiques, les données à exploiter ou à analyser sont devenues très volumineuses. De nombreuses applications de big data ont été développées au cours des dernières années dont les systèmes de recommandation, qui permettent de faciliter cette recherche ainsi que l'extraction des informations pertinentes. Il s'agit de guider l'utilisateur dans son exploration des données afin qu'il trouve des informations pertinentes.

Les systèmes de recommandation ont été étudiés dans des domaines divers et variés comme le e-commerce, les industries culturelles et bien d'autres. Nous abordons ici le processus de recommandation, en mettant la focale sur le modèle utilisateur.

Les plateformes de services en ligne représentent une source de données importantes. De ce fait, la définition des profils pour les utilisateurs s'est affinée dans les systèmes de personnalisation, les systèmes adaptatifs, les systèmes de recommandation, les systèmes d'analyses comportementales, etc.

Un système de recommandation est un système de filtrage et de personnalisation de l'information, visant à présenter et à prédire des choix d'items susceptibles d'intéresser l'utilisateur, ceci pour des biens divers (livres, musique, films, etc.)[97].

La conception d'un système de recommandation suit un certain nombre d'étapes que nous allons détailler.

Un système de recommandation cherche à prédire les attentes et les intérêts des utilisateurs, ce qui nécessite de recueillir des données les caratérisant, en vue de construire un profil utilisateur.

La collecte se fait de deux manières distinctes :

- La collecte des données explicites (le filtrage actif) :

[97] ADOMAVICIUS, Gediminaset Youngok, KWON. New recommendation techniques for multicriteria rating systems.*IEEE Intelligent Systems*. 2007.

Comme son nom l'indique, cette approche consiste à ce que l'utilisateur donne lui-même et explicitement des données sur ses intérêts et ses préférences. Les utilisateurs notent, aiment et taguent des contenus proposés.

Cette approche présente l'avantage de construire l'historique du profil de l'utilisateur. *A contrario,* elle représente l'inconvénient du biais de la déclaration de l'utilisateur.

- La collecte des données implicites (le filtrage passif) :

Il s'agit d'une approche qui collecte de l'information sur l'utilisateur sans son intervention directe, à travers l'étude de ses comportements dans le passé. Cette collecte s'opère en fonction de plusieurs entrées :

- La fréquence de consultation ;
- Le comportement en ligne de l'utilisateur ;
- L'action sur le réseau social ;
- La liste des items/objets utilisés et consultés.

Le filtrage passif a la caractéristique de ne rien demander aux utilisateurs, puisque l'information est automatiquement récupérée. Par contre, elle peut poser le problème du biais d'attribution, puisque ce n'est pas forcément parce qu'un objet est acheté, qu'il est, par conséquent, apprécié par l'utilisateur.

Un modèle d'utilisateur correspond à une matrice, il joint les intérêts de l'utilisateur aux produits et objets disponibles (films, livres…). Toutefois, les besoins des utilisateurs évoluent, ce qui nécessite une mise à jour continuelle du système de recommandation.

Une fois le modèle utilisateur conçu, les algorithmes de recommandation interviennent pour chercher les similarités et les corrélations entre les différents objets/contenus et les profils d'utilisateurs, pour générer une liste des propositions[98]

Les systèmes de recommandation accumulent une très grande quantité de données sur un utilisateur (son comportement d'achat et de navigation et ses caractéristiques) et en les croisant

[98]RAMAN, Karthik. *Online learning to diversify from implicit feedback*. Qiang Yang, Deepak Agarwal, and Jian Pei, editors, Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12), 2012.p.705-712.

avec d’autres données (comportements de consommateurs similaires, attributs des biens), ils prédisent les biens qui pourraient être pertinents.

Chapitre 2 : Les approches de recommandation

Introduction

Dès l'apparition du domaine des systèmes de recommandation, plusieurs approches et méthodes ont émergé, certaines ont été étudiées, expérimentées et comparées. Parfois, plusieurs terminologies sont utilisées pour désigner une même méthode ou approche. C'est pour cela que certains chercheurs se sont intéressés à la classification de ces méthodes et proposent une terminologie unifiée.

La littérature sur les systèmes de recommandation montre une forme de consensus en ce qui concerne la ventilation des systèmes de recommandation en trois principales approches : les méthodes basées sur le filtrage collaboratif, les méthodes basées sur le contenu et les méthodes hybrides.

1. L'approche objet ou basée sur le contenu (*Content-Based Filtering*)

Il s'agit de proposer des contenus à partir des caractéristiques de ces contenus, en les corrélant avec les intérêts et les préférences des utilisateurs.

En outre, cette approche utilise les attributs (les caractéristiques d'un contenu), en vue de recommander des contenus similaires. A titre d'illustration, les thèmes généraux et les mots clés associés à un contenu / objet donné, sont comparés aux préférences de l'utilisateur[99].

[99] LU, Jie et al. Recommender system application developments: A survey, *Decision support system*, June 2015, Vol.74, P. 12-32.

Les avantages et les inconvénients de cette approche sont récapitulés dans le tableau ci-après[100] :

Avantages	Inconvénients
- Pouvoir fonctionner, même avec un seul utilisateur dans le système ; - Plus le système collecte les données sur les différents objets, plus il devient performant en matière de recommandation.	- Certains attributs peuvent être extraits automatiquement, ce qui fait qu'ils risquent de ne pas être révélateurs du contenu des objets en l'absence d'une validation humaine ; - La dose de subjectivité lors de la description d'un objet donné ; - L'analyse des items par l'extraction de propriétés n'est pas toujours facile (musique, multimédia, image ou même d'autres items textuels) ; - L'hyper-specialisation de la recommandation, en se basant sur le profil, tend à enfermer l'utilisateur dans ses goûts initiaux.

Tableau 5 : Les avantages et les inconvénients de la recommandation basée sur le contenu

Il est à signaler que cette approche de recommandation est moins utilisée comparativement à celle des proches voisins (filtrage collaboratif) et que les inconvénients de cette approche représentent les atouts de la seconde.

2. L'approche de la recommandation sociale ou de filtrage collaboratif (*Collaborative Filtering)*

Cette approche définit l'utilisateur dans le contexte des autres utilisateurs du système. Partant de la logique de recherche des similarités, le système prédit ce que l'utilisateur appréciera en

[100] KEMBELLEC, Gérald ; CHARTRON, Ghislaine et SALEH, Imad. *Les moteurs et systèmes de recommandation*. Paris : ISTE éditions, 2014, 228 p.

fonction des utilisateurs qui lui ressemblent, tout en liant son comportement passé à celui du futur.

En effet, le filtrage collaboratif s'opère de deux manières différentes[101] :

- **Une recommandation orientée utilisateurs similaires (voisinage proche) :**

Il s'agit d'utilisateurs regroupés en fonction de leurs similarités, en termes des préférences et des intérêts relatifs aux objets et items à recommander, à partir du calcul des notes attribuées par chacun des utilisateurs à des objets/ contenus. Le calcul s'appuie sur une matrice à deux axes, le premier correspond aux utilisateurs et le second aux notes attribuées aux objets par ces utilisateurs.

Dans le cas où des utilisateurs évaluent de la même manière ou presque un contenu donné, à travers la moyenne des ratings attribués, ils recevront des recommandations sur ce qu'ils aiment et apprécient les uns et les autres.

Pour ce faire, des algorithmes sont déployés comme celui de Pearson. Ce coefficient se calcule sur la base des valeurs figurant sur la matrice avec des valeurs extrêmes, le groupe représenté par « +1 » signifie une forte corrélation positive, et celui représenté par « -1 » correspond à une faible corrélation négative. Les meilleurs voisins sont identifiés à partir d'un seuil de corrélation[102].

Pour finir, il faut mentionner que cette approche rend difficile l'action du système de recommandation en temps réel, quand il s'agit de gros sites où il existe des millions d'utilisateurs, c'est la raison pour laquelle certains d'eux ont implémenté des techniques de traitement des données hors ligne[103].

---

[101] PECUNE, F., MURALI, S., TSAI [et al.]. *A Model of Social Explanations for a Conversational Movie Recommendation System*. In Proceedings of the 7th International Conference on Human-Agent Interaction, July 2019. Pp. 135-143. ACM.

[102] LU, Jie et al. Recommender system application developments: A survey, *Decision support system*, June 2015, Vol.74, p. 12-32.

[103] KEMBELLEC, Gérald; CHARTRON, Ghislaine et SALEH, Imad. *Les moteurs et systèmes de recommandation*. Paris : ISTE éditions, 2014, 228 p.

- **Une recommandation orientée contenus similaires :**

Par contraste avec l'approche précédente « users nearest neigborhood », celle « d'Item-to-Item » se fonde sur le contenu et non pas sur l'utilisateur, partant toujours du coefficient de Pearson. Cette méthode cherche la corrélation éventuelle entre des contenus similaires pour les proposer à l'utilisateur. Il s'agit de construire une matrice item-item déterminant des relations entre des objets similaires à ceux déjà explorés par les utilisateurs. Ainsi, cette matrice est utilisée pour proposer de nouveaux objets. Cette approche présente des atouts et son lot d'imperfections[104], comme le précise le tableau suivant :

Avantages	Inconvénients
- Le privilège de ne pas avoir besoin de connaissances (description) sur les contenus ; - La qualité des recommandations croit avec le temps en rapport avec l'évolution du nombre des utilisateurs et du nombre d'items	- *Scalability :* le problème des sites des millions d'utilisateurs, ce qui rend la recommandation en temps réel très difficile. - *Cold start :* il s'agit du démarrage à froid, quand le système n'a pas, au début, assez de données sur l'utilisateur et les objets. - *Sparsity* : le nombre de contenus est important, au point que parfois les utilisateurs se contentent de noter qu'une partie infime des contenus. - *Le shilling :* il s'agit de l'action malveillante d'influencer les recommandations, en créant des faux profils pour favoriser ou défavoriser certains objets par rapport à d'autres.

Tableau 6 : Les avantages et les inconvénients de la recommandation basée sur le filtrage collaboratif

Il est à signaler que cette approche de recommandation est plus commune que celle basée sur le contenu et que les imperfections de cette approche représentent les avantages de la première.

3. L'approche de la recommandation hybride *(Hybrid Filtering)*

La recommandation hybride est une combinaison des deux approches (objet et sociale), dans la perspective de combler les lacunes de chacune d'elles avec une collecte des caractéristiques des

[104]Ibid., p.107.

contenus et la prise en compte des « ratings » des utilisateurs dans le contexte d'une recommandation sociale. Ainsi, il est possible d'éliminer les manquements des deux approches, comme en témoigne le cas de la plateforme d'e-commerce Amazon.

4. Autres approches de recommandation

Il existe d'autres approches de recommandation plus pointues, telles que :

- **L'approche basée sur les connaissances (***Knowledge Based***)**

La base de connaissances acquise sur un domaine précis représente la matière première de la recommandation dans ce type de systèmes, en la confrontant aux connaissances et aux attentes des utilisateurs en termes des caractéristiques pour un item donné. C'est l'exemple des produits recommandés avec des spécifications précises liées au besoin des utilisateurs.

- **L'approche démographique (***Demographic Based***)**

Cette famille de systèmes adapte la recommandation à la zone géographique de l'utilisateur (le pays, la ville ou le quartier). La personnalisation concerne la langue d'affichage ou la recommandation des items en fonction de leur particularité et leur disponibilité dans une zone géographique donnée[105].

- **L'approche communautaire** *(Community Based)*

Il s'agit de recommander aux utilisateurs les items aimés par leurs amis. En effet, c'est une sorte de spécialisation du filtrage collaboratif, où la similarité est prise en compte uniquement dans le cas des utilisateurs amis[106].

[105] GAILLARD, Julien. *Systèmes de recommandation : adaptation Dynamique et Argumentation* [en ligne]. Thèse de doctorat. Université d'Avignon, 2014. Disponible sur : https://hal.archives-ouvertes.fr/tel-01168476/ [consulté le 12 Mars 2017].
[106] Ibid.

Conclusion

La recommandation est un processus dont les phases sont intimement liées à la collecte des données jusqu'à la génération des suggestions destinées aux utilisateurs.

Les principales approches de recommandation sont les suivantes :

- La recommandation basée sur le contenu ou objet : qui se fonde sur l'analyse des préférences du consommateur à partir propriétés décrivant les objets et les contenus ;
- La recommandation sociale ou collaborative, qui repose sur le comportement passé des utilisateurs similaires, qu'il s'agisse de la proximité d'intérêt que de la similarité des objets et des contenus ;
- La recommandation hybride : qui combine les trois approches précédentes, en capitalisant sur les atouts de chacune d'elles et comblant leurs lacunes et limitations.

Chapitre 3 : Les algorithmes de recommandation

Introduction

Un algorithme est une suite d'instructions structurée en différentes étapes qui vise à résoudre un problème.

En matière de recommandation, les algorithmes utilisent les caractéristiques d'un client ou d'un utilisateur, pour générer une liste de produits ou d'objets à recommander. Les plateformes peuvent se baser sur l'historique d'usage ou d'achat, ainsi que sur l'évaluation explicite, sur les données démographiques et la liste des préférences pour constituer les entrées à ces algorithmes.

L'utilisation des algorithmes de recommandation ont pour objectif de valoriser la quantité des données disponibles dans plusieurs objectifs :

- Amélioration de l'expérience utilisateur ;
- Amélioration continue des indicateurs clés (panier moyen, réduction des délais de recherche de contenus/produits, etc.) ;
- Analyse automatique des données et de leur filtrage pour des recommandations pertinentes.

Le présent chapitre présente les algorithmes les plus répandus en matière de recommandation.

1. L'apprentissage automatique

Les systèmes de recommandation font partie des branches de l'intelligence artificielle, dont l'objectif principal est d'aider les utilisateurs à identifier les items liés à leurs intérêts et préférences. En effet, l'intelligence artificielle consiste à produire des machines capables de simuler l'intelligence.

L'apprentissage automatique permet d'améliorer les performances des machines à résoudre des problèmes à forte complexité logique et algorithmique, en leur permettant d'apprendre à partir des données pour résoudre des tâches spécifiques (traduire un discours, distinguer dans une

messagerie électronique les courriels spams et non spams, etc.) sans être explicitement programmés pour ces tâches[107].

De même, l'apprentissage automatique est utilisé dans différents domaines, tels que le scoring, la recommandation, la détection de fraudes et d'anomalies, etc. et comprend plusieurs types[108] :

- **Apprentissage supervisé :** il consiste à apprendre une fonction de prédiction à partir d'exemples annotés. Nous considérons des couples entrée-sortie, les classes sont prédéterminées et les données sont étiquetées. Les données d'apprentissage (observations) sont accompagnées par des étiquettes indiquant leurs classes ;

- **Apprentissage non supervisé :** dans ce modèle d'apprentissage, l'algorithme doit découvrir par lui-même la structure des données, dans la mesure où les modèles ne sont pas disponibles à priori. L'algorithme d'apprentissage se charge, dans cette approche, de trouver les similarités en se basant sur la notion de proximité entre les données d'entrée afin de les regrouper automatiquement en sous-ensembles homogènes (clusters) ayant des caractéristiques jugées par l'algorithme d'apprentissage similaires et communes ;

- **Apprentissage semi-supervisé :** comme son l'indique, l'apprentissage semi-supervisé constitue une forme d'apprentissage hybride entre le supervisé et le non supervisé. En effet, il peut s'effectuer d'une manière probabiliste si les exemples sont étiquetés, ou non probabiliste quand il s'agit d'identifier la distribution sous-jacente des exemples dans leur espace de description quand les étiquettes manquent ;

- **Apprentissage par renforcement :** c'est une forme d'apprentissage supervisé où l'algorithme ne dispose pas d'une étiquette pour chacune des prédictions, mais plutôt d'une mesure de qualité de ses prédictions. Ainsi, l'algorithme apprend un

[107] AGGARWAL, Charu C. *Recommender Systems*, Springer International Publishing, 2016. 498 p.

[108] Ibid.

comportement, avec une observation de l'algorithme sur l'environnement et produit une valeur de retour qui guide l'algorithme d'apprentissage.

2. Les algorithmes de recommandation à base de contenu

La recommandation à base de contenu ou orienté objet est faite par des algorithmes qui vérifient si un contenu non encore exploré par l'utilisateur, pourrait éventuellement l'intéresser à partir de ce qu'il a déjà évalué positivement dans le passé. Un système de recommandation axé sur le contenu ou l'objet peut intégrer trois algorithmes distincts :

- Le système se contente de s'ancrer dans les généralités par vérification, si l'objet existe parmi les genres préférés de l'utilisateur par le biais d'une similarité binaire de 0 et 1 ;
- Le second algorithme de *Dice* se focalise plutôt sur les mots clés caractérisant le contenu en question avec la mesure du degré de similarité et les croisements possibles entre les mots clés préférés par l'utilisateur et ceux relevés et extraits des différents contenus. Toutefois, cette technique n'est pas tellement efficace, puisque les mots clés ne sont pas toujours révélateurs du contenu, le système de recommandation peut aussi proposer des documents intégrant les mots car ils y apparaissent avec une fréquence élevée sans être forcément représentatifs du contenu ;
- La méthode de Rochchio (le retour de pertinence) privilégie l'idée de rester à l'écoute du feedback de l'utilisateur, en lui permettant de juger les recommandations proposées par le système, une fois ce dernier reçoit les évaluations des utilisateurs, il les ventile en deux catégories : celles relevant d'un jugement positif, et celles relevant d'un jugement négatif ; par la suite le centre de chaque groupe des évaluations des utilisateurs est calculé, de telle sorte que la requête de l'utilisateur soit plus affinée et pertinente, du moment où il y a des contenus à retenir et d'autres à écarter.

3. L’algorithme de filtrage collaboratif à base de mémoire

La prédiction de la notation sur un objet cible repose, dans le cas du filtrage collaboratif à base de mémoire, sur les données stockées déjà dans la mémoire du système. En effet, ce type de filtrage est ventilé en deux approches principales :

- **L’approche centrée sur l’utilisateur :**

Dans ce cas, un utilisateur « X » est identifié par les appréciations qu’il a émises à propos de certains items. Ensuite, nous procédons à déceler les utilisateurs similaires à l’utilisateur cible « X » et à prévoir les notes d’évaluations non encore réalisées de l’utilisateur « X», en calculant les moyennes pondérées des notations de ce groupe d’utilisateurs[109].

- **L’approche axée sur l’objet (Item) :**

Il s’agit de prédire les notations sur un objet donné par un utilisateur « X ». Pratiquement, il importe de déterminer un ensemble S d’objets les plus semblables au nouvel objet Y considéré. Les évaluations des objets dans l’ensemble « S » sont utilisées pour prédire si l’utilisateur « X » appréciera l’objet « Y ».

4. L’algorithme de filtrage collaboratif à base de modèle

Le filtrage collaboratif à base de modèle utilise les données d’entrée, en vue de générer des modèles de prédiction sur les items susceptibles d’intéresser les utilisateurs. Une fois le modèle créé, le système n’a plus besoin de solliciter des données brutes, dans la mesure où un profil de l’utilisateur a été établi.

En effet, le filtrage basé modèle crée un modèle descriptif liant les utilisateurs, les objets et les évaluations. Ainsi, le filtrage collaboratif effectue, dans ce cas, la prédiction à partir du profil utilisateur ou ses précédentes évaluations. Par contre, le filtrage à base de mémoire solliciterait la totalité ou une partie des profils utilisateurs, en vue de générer une nouvelle prédiction

[109]AGGARWAL, Charu C. Op.cit., p.112.

relative à une note potentielle que l'utilisateur pourra attribuer à un objet donné en fonction des notes données par d'autres utilisateurs.

Ce filtrage à base de modèle s'opère par plusieurs phases :

- **Le calcul de la similarité**

Qu'il s'agisse de la similarité entre les utilisateurs ou entre les objets, plusieurs méthodes sont sollicitées à cet effet :

- **La similarité entre items (objets)**

L'idée de base du calcul de la similarité entre deux objets a et b consiste à isoler, d'abord, les utilisateurs qui ont évalué ces deux éléments, puis à appliquer une technique de calcul de similarité pour déterminer la similitude entre ces objets que nous notons "$\mathrm{sim}(a, b)$".

Ainsi, le tableau suivant explicite ce qui précède[110] :

La méthode	La définition
La probabilité conditionnelle **(la probabilité d'un événement sachant qu'un autre événement a eu lieu)**	Il s'agit de récupérer l'historique des transactions du client pour calculer la probabilité qu'un objet est susceptible d'intéresser les tilisateurs : $\mathrm{sim}(i, j) = P_{rB}[j \in B \mid \in B]$ Sachant que : "$\mathrm{sim}(i, j)$" : la similitude entre ces objets i et j B : l'historique d'achat de l'utilisateur
La similarité Cosinus	Elle renvoie au calcul de la similarité entre les objets ou les items, en calculant le cosinus des angles entre les vecteurs des notations des utilisateurs et ceux relatifs aux objets : $\mathrm{sim}(i, j) = \frac{r_i . r_j}{\| r_i \|_2 * \| r_j \|_2}$

Tableau 7 : Les méthodes de calcul de similarité entre les objets (items/articles)

[110]SHARMA, Richa, and RAHUL, Singh. Evolution of recommender systems from ancient times to modern era: A survey. *Indian Journal of Science and Technology*. Septembre 2016.

- **La similitude entre les utilisateurs :**

Par analogie avec les méthodes du filtrage collaboratif basées sur les objets (items, contenus) et au lieu de calculer la similitude entre deux éléments, nous nous concentrons sur la similitude entre deux clients ou utilisateurs, comme l'indique le tableau ci-dessous[111] :

La méthode	La définition
La corrélation de Pearson (Pearson correlation)	Cette méthode est utilisée pour calculer la similarité entre deux utilisateurs en se basant sur leurs notations. Sachant que : a et b des utilisateurs ; $r_{a,p}$: l'évaluation de l'utilisateur (a) au produit (p) ; $\overline{r_a}$et r_b : l'évaluation moyenne des utilisateurs ; P : Ensemble de produits notés par (a) et (b) ; Les valeurs se situent entre -1 (forte corrélation négative) et +1 (forte corrélation positive). La corrélation est représentée par la fonction ci-après : $$sim(a,b) = \frac{\sum_{p\in P}(r_{a,p} - \overline{r_a}) * (r_{b,p} - \overline{r_b})}{\sqrt{\sum_{p\in P}(r_{a,p} - \overline{r_a})^2} . \sqrt{\sum_{p\in P}(r_{b,p} - \overline{r_b})^2}}$$
La similarité Cosinus	Les utilisateurs sont considérés comme des vecteurs en algèbre linéaire et la similitude entre les vecteurs est mesurée au moyen de la distance cosinus de la manière suivante : $$sim(a,b) = \frac{r_a . r_b}{\| r_a \|_2 * \| r_b \|_2}$$ Sachant que : a et b des utilisateurs ; $r_a$et r_b : l'évaluation moyenne des utilisateurs ; $r = (a \quad b \quad c)$; $\|r\|_2 = \sqrt{a^2 + b^2 + c^2}$.

[111]Ibid., p.115.

La corrélation de Spearman	Il s'agit d'une méthode qui mesure la corrélation sur la base des rangs au lieu des notations, contrairement au coefficient de Pearson : $$\mathrm{sim}(a,b) = \frac{\sum(\mathrm{rank}_{ai} - \overline{\mathrm{rank}_a}) \,.\, (\mathrm{rank}_{bi} - \overline{\mathrm{rank}_b})}{\sqrt{\sum(\mathrm{rank}_{ai} - \overline{\mathrm{rank}_a})^2} \,.\, \sqrt{\sum(\mathrm{rank}_{bi} - \overline{\mathrm{rank}_b})^2}}$$ Sachant que : a et b des utilisateurs ; rank ai ou bi : le rang attribué par un utilisateur (a) ou (b) à un item (i) ; $\overline{\mathrm{rank}_a}$ $\overline{\mathrm{rank}_b}$: la moyenne des rangs attribués par les utilisateurs (a) et (b).

Tableau 8 : Les méthodes de calcul de similarité entre utilisateurs

- **La détermination des utilisateurs proches/voisins**

La détermination des voisins d'un utilisateur conditionne la pertinence de la prédiction. Ainsi, il existe des algorithmes qui identifient le top N voisins dont la similarité est supérieure à 0.

- **Le classement et la présentation des items à recommander**

Le système de recommandation procède au classement des items en fonction de leurs côtes prédites et/ou de leur popularité. En effet, une fois la sélection des top-N items effectuée, la partie la mieux classée sera recommandée aux utilisateurs[112].

Dans le même ordre d'idée, nous pouvons aussi citer le filtrage collaboratif à base de modèle qui correspond à une approche, qui sollicite les techniques de machine learning ou d'analyse de données (clustering, classification, factorisation matricielle, etc.), pour établir des modèles de prédiction des préférences des utilisateurs vis-à-vis de certains items et contenus, en calculant la valeur la plus probable d'un score ou une note sur la base de ce qui existe comme données sur l'utilisateur, c'est ainsi que le modèle utilisateur est créé et servira à la recommandation.

[112]YANG, Zhe, et al. *A Survey of Collaborative Filtering-Based Recommender Systems for Mobile Internet Applications*.IEEE Access 4. 2016, p.3273-3287.

En ce qui concerne l'étude de cas de notre thèse, qui sera traitée dans la partie suivante, **nous visons à déployer ce type de filtrage collaboratif à base de modèle** au travers les algorithmes de classification supervisée et de factorisation matricielle, compte tenu des points clés suivant :

- Les données *log* apportent plus de détails sur l'utilisateur que le contenu, puisqu'elles relatent principalement les interactions des utilisateurs avec le système et leur parcours de navigation ;
- Le filtrage collaboratif permet **de découvrir les données inaccessibles à un profil**, en vue de faire des prédictions précises, en utilisant l'historique de tous les utilisateurs d'un système ;
- Le filtrage collaboratif s'applique à des **langues et des secteurs différents** grâce à la flexibilité de ses algorithmes, traduite par la modification des paramètres de similarités et la considération des différents avis ;
- Les algorithmes de filtrage collaboratif peuvent **changer de taille et de besoin.** Ils s'adaptent à un nombre croissant d'utilisateurs et de contenus grâce aux capacités de passage à l'échelle, qui ont pour corollaire des résultats pertinents en temps réel.
- A cet égard, **la classification supervisée**, que nous allons solliciter dans notre thèse, consiste à définir les règles permettant de classer des objets dans des classes à partir de variables qualitatives ou quantitatives caractérisant ces objets. Quant à la factorisation matricielle, les notations des usagers pour les items sont rassemblées dans une matrice de notation composée des vecteurs d'utilisateurs et des items. Or, il est à signaler que chaque utilisateur ne peut pas noter tous les items. Par conséquent, cette matrice reste creuse, d'où l'importance de la prédiction pour la remplir entièrement[113].

Quant à **la factorisation matricielle**, elle consiste à décomposer en produits de plusieurs sous matrices, afin de réduire la dimension de la matrice, et ainsi diminuer la densité des données, ce qui devrait plaider en faveur d'une performance dans le cas des systèmes de recommandation. Mathématiquement, nous considérons une matrice unitaire **« T »** de notations « **R »**, contenant les évaluations d'un ensemble d'utilisateurs « **U** » pour un ensemble d'Items

[113] SHANG, Xuedong. *Recommandation personnalisée* [en ligne], 2015. Disponible sur : https://xuedong.github.io/static/documents/recommender.pdf [consulté le 22 Février 2016].

« **I ».** L'objectif de la factorisation matricielle est de trouver deux matrices : **« P »** correspondant aux utilisateurs, et « **Q »** renvoyant aux objets**,** tel que :

$$\hat{R} = P^T \times Q \approx R$$

Donc, la prédiction « R ui » du vote qu'un utilisateur « $\mathbf{P_u}$ » émetttera sur un Item « $\mathbf{q_i}$ » peut être calculée par la fonction suivante :

$$\hat{r}_{ui} = p_u^T q_i$$

En fait, plusieurs modèles d'apprentissage sont employés dans la factorisation matricielle, parmi lesquels nous citons, le modèle des moindres carrés alternés (*Alternating Least Squares*) plus connu sous l'abréviation d'ALS. Ce modèle agit sur les deux facteurs « $\mathbf{P_u}$ » et « $\mathbf{q_i}$ » pour construire un modèle d'apprentissage[114].

Il importe de signaler que ces deux algorithmes, la classification et la factorisation matricielle, reposent sur un apprentissage automatique supervisé, qui consiste à apprendre à un système à automatiser une fonction de prédiction à partir d'exemples annotés et connus. En fait, pour générer le modèle utilisateur dans le filtrage collaboratif à base de modèle, les données sont étiquetées et collectées, pour construire une fonction de prédiction. Il s'agit d'un modèle dont l'objectif est de généraliser, c'est-à-dire d'apprendre une fonction qui fasse des prédictions pertinentes sur des données relatives au profil de l'utilisateur ou ses notations précédentes et non pas présentes dans la base d'apprentissage.

5. Le type de données utilisées dans les systèmes de recommandation

Selon SHARMA et al.[115] l'étape de structuration de données regroupe trois catégories de données dans la modélisation d'un système de recommandation : données d'identité, données d'activité, et les données de sécurité, cette typologie de données est présentée dans le tableau ci-dessous :

[114] Ibid., p.118.
[115]SHARMA, Richa, and RAHUL, Singh. Evolution of recommender systems from ancient times to modern era: A survey. *Indian Journal of Science and Technology*. Septembre 2016.

Le type de données utilisé dans un système de recommandation	La caractérisation des données
Les données d'identité	Elles renvoient aux caractéristiques personnelles de l'utilisateur, qui sont généralement d'ordre statique et fournies de manière explicite par les utilisateurs (via des formulaires d'inscription par exemple). Elles concernent les données personnelles (nom, sexe, taille, couleur des yeux, etc.), les données démographiques (pays, ville, adresse, etc.), le cursus académique, l'historique des emplois et les distractions et les centres d'intérêt.
Les données d'activité	Elles correspondent aux traces d'activités issues de l'interaction entre l'utilisateur et le système d'information (parcours de navigation sur un site web, fichiers log d'un serveur web, fichiers log d'une base de données, activités sur un réseau social numérique, etc.). Les données d'activité sont généralement accompagnées par trois types de feedbacks. D'abord, ceux explicites qui fournissent le moyen d'évaluer, a priori et sans ambigüité, la pertinence d'une activité de l'utilisateur pour le calcul de ses centres d'intérêts (l'achat des produits ou la consultation des items, les boutons j'aime sur Facebook, etc.) Puis les feedbacks implicites se rapportant aux actions des utilisateurs, qui génèrent des contenus nécessitant des analyses plus profondes, en vue de calculer des centres d'intérêts (commentaires, tags ou annotations) et les feedbacks externes extraits par des capteurs externes à partir des activités liées aux organes physiques, comme les détecteurs de mensonge (au cours d'un entretien) ou du mouvement des yeux.
Les données de sécurité	Elles sont liées aux réponses données par l'utilisateur pour limiter ou autoriser le traitement de ses données, afin de construire son profil conformément aux législations en vigueur en matière du traitement des données personnelles.

Tableau 9: Les types de données utilisées dans les systèmes de recommandation

Source : SHARMA, Richa, and RAHUL, Singh. Evolution of recommender systems from ancient times to modern era: A survey. *Indian Journal of Science and Technology*. Septembre 2016 (avec adaptation).

Il est à signaler que, les données qui alimentent les systèmes de recommandation peuvent être liées au contexte, au contenu et à la sémantique. Une fois combinées elles dressent le profil des utilisateurs, en vue de leur faire parvenir des recommandations sur des items bien définis.

6. L'exemple de Spotify et de Netflix

Pour illustrer la dimension opérationnelle de ces algorithmes de recommandation, nous allons prendre l'exemple de deux grandes plateformes de streaming musique et vidéo, à savoir Spotify et Netflix. Le tableau 10 résume des caractéristiques majeures des services développés :

La plateforme intégrant un système de recommandation	**Les services fournis, les données utilisées et Les algorithmes déployés**
Spotify	- Il s'agit de proposer un service de streaming qui comporte un onglet « nouveautés personnalisées », proposant les derniers albums, en fonction des écoutes. Les radios qui permettent d'enchaîner les morceaux en partant d'un genre musical ou d'un artiste particulier donnent parfois d'étranges résultats, tout comme les recommandations d'albums. - Pour parvenir à ce résultat, Spotify utilise à la fois le filtrage collaboratif, qui se nourrit des données de goûts implicites sans solliciter aucune notation. Il s'agit des données de navigation dans les pages des artistes et des albums, le temps d'écoute moyen et les chansons enregistrées sur une playlist. Le content-based filtering, repose sur les données décrivant les contenus, qu'il s'agisse du contenu textuel (similarité entre les mots qui reviennent le plus, les noms des artistes ou des albums, les adjectifs utilisés, etc.) et du contenu sonore (le rythme, le tempo, le niveau de basses, etc.). Plusieurs centaines de millions de chansons sont ainsi recommandées sur la base d'autres millions de préférences de goût analysées[116].

[116] PUJOL, Lisa. *Parlons recommandation : Partie 2 / Le modèle Spotify*. *MEDIUM* [en ligne], 2018. Disponible sur : https://medium.com/delightblog/parlons-recommandation-partie-2-le-mod%C3%A8le-spotify-e81e6697c75e [Consulté le 29 Août 2019].

Netflix	Netflix est une entreprise américaine qui offre des films et séries télévisées en flux continu (streaming) sur internet et téléphone mobile. En effet, Netflix dispose d'un moteur de recommandation 'cinematch', qui propose aux usagers des films et séries susceptibles de les intéresser. Ainsi, la recommandation chez Netflix consiste à ce que chaque usager ayant visionné un film, donne son avis et attribue une note comprise entre 1 et 5 sur ce dernier. Les données colligées ont trait à la fois au comportement des utilisateurs (données de navigation), aux notations données aux films et séries et au contenu laissé par les utilisateurs (les commentaires et les notes). Les algorithmes les plus déployés sont les suivants : - Algorithme de classement personnalisé de vidéos : il suggère des vidéos pour chaque membre de manière personnalisée, en les affichant sur la page d'accueil sur environ 40 lignes ; - Algorithme de classement du top n vidéo : il génère des recommandations en haut de la page. Cet algorithme vise à trouver les meilleures recommandations personnalisées dans l'ensemble du catalogue pour chaque membre et se concentre seulement sur la tête du classement ; - Algorithme générant la barre « maintenant, à la mode » : il recommande les vidéos de tendance, allant de quelques minutes à quelques jours seulement, avec une dose de personnalisation selon le membre[117].

Tableau 10: Exemples de services des systèmes de recommandation de Spotify et Netflix

Caractérisons les algorithmes déployés et leur fonctionnement :

- **La recommandation *made in* Spotify**

Le système de recommandation de Spotify utilise à la fois le **collaborative filtering** et le **content-based filtering :**

Pour la première approche, Spotify exploite des données implicites au lieu d'un système de notation des contenus comme on peut le trouver sur Netflix ou sur YouTube. En effet, les données des utilisateurs ont trait à la navigation et à l'écoute.

A titre d'illustration, les morceaux ajoutés à la playlist d'un utilisateur ou encore le temps d'écoute moyen sur un genre musical précis, représentent des sources de données implicites, et

[117] GOMEZ-URIBE, Carlos A. et HUNT, Neil. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*. 2016, vol. 6, no 4, p. 13.

qui permettent d'effectuer des comparaisons entre les préférences des utilisateurs. Par exemple, si deux utilisateur X et Y ont écouté les morceaux A, B et C, il en découle que les goûts des utilisateurs X et Y sont partagés. De cette manière, l'algorithme de Spotify suppose que chaque chanson écoutée par l'un est susceptible d'intéresser l'autre.

Quoique cet exemple soit simplifié, il n'occulte guère le degré de complexité des algorithmes de système de recommandation greffés sur la plateforme musicale Spotify, qui s'alimentent en de centaines de millions de chansons recommandées sur la base d'autres millions de préférences des utilisateurs.

Quant à la seconde approche axée sur le contenu, Spotify se nourrit de données sur les contenus à la fois textuels et sonores qui sont l'objet de l'analyse.

Pour les données textuelles, Spotify s'appuie sur les techniques du Traitement Automatique des Langues ("TAL" ou en anglais **"**Natural Language Processing"**)** qui permettent d'analyser les données textuelles sur le web moyennant une veille, visant la collecte de ce qui se dit sur les artistes et les albums musicaux et les termes employés avec une pondération.

Ainsi dans l'exemple ci-dessous[118], la probabilité que l'artiste soit qualifié de « perky » est la plus haute :

[118] WHITMAN, B., "How music recommendation works – and doesn't work". *Variogram* [en ligne]. December 11, 2012. Disponible sur: https://notes.variogr.am/2012/12/11/how-music-recommendation-works-and-doesnt-work/ [consulté le 15 Avril 2017].

adj Term	Score
perky	0.8157
nonviolent	0.7178
swedish	0.2991
international	0.2010
inner	0.1776
consistent	0.1508
bitter	0.0871
classified	0.0735
junior	0.0664
produced	0.0616

Tableau 11 : Les poids des termes sur Spotify par le biais du Traitement Automatique du Langage (WHITMAN, 2012)

Il ressort du tableau que Spotify vise à rapprocher les contenus similaires en fonction de ces « top terms » (qui évoluent constamment) et baser ses recommandations sur ces rapprochements, à l'image du collaborative filtering.

Pour les données sonores, Spotify recueille les données qui ont pour objectif de caractériser des contenus à l'image du rythme, du tempo, etc., tout en évitant l'orientation des utilisateurs vers des contenus déjà très populaires. Ainsi, un jeune artiste débutant peut voir recommander ses chansons peu populaires et cumulant peu d'écoutes grâce aux caractéristiques sonores.

Le mécanisme de recommandation de Spotify vise à favoriser la découverte de nouveaux artistes, pour consacrer la diversité culturelle des utilisateurs et éviter l'enfermement dans une hyper-spécialisation de leurs goûts.

- **La recommandation *made in* Netflix**

Actuellement leader sur le marché de la Vidéo à la Demande (VOD), Netflix a développé des algorithmes, pour proposer des recommandations personnalisées et aider les abonnés à découvrir des séries TV et des films susceptibles de les intéresser.

Selon l'approche de l'analyse comportementale, il s'agit de calculer la probabilité qu'un utilisateur regarde un film ou une série sur le catalogue à l'issue d'un nombre grandissant de données provenant de la pratique du *binge-watching,* qui consiste à regarder plusieurs épisodes d'une série ou plusieurs films pendant une période relativement longue, pour extraire des données portant sur :

- Les interactions avec la plateforme (historique de visionnage et évaluations d'autres titres, par exemple) ;
- Les choix des autres utilisateurs, pour vérifier la similarité ;
- Les données liées à chaque film ou série (le genre, les acteurs, le réalisateur, etc.).

Netflix fait appel à d'autres données contextuelles, afin de mieux affiner les recommandations :

- Le temps passé sur un titre ;
- La période d'utilisation ;
- Les appareils utilisés pour regarder.

Les différentes données récoltées alimentent un algorithme de recommandation au moment où une recherche est lancée par un utilisateur, en vue de la comparer à d'autres recherches identiques ou similaires effectuées avant par d'autres utilisateurs. La page d'accueil d'un abonné se présente de la manière suivante[119] :

Figure 15 : La présentation des recommandations sur la plateforme Netflix (Système de recommandation de Netflix, 2020)

[119] Système de recommandation de Netflix [en ligne]. Disponible sur : https://www.futura-sciences.com/tech/questions-reponses/informatique-netflix-fonctionne-algorithme-recommandations-8640/ [consulté le 22/03/2020].

Sur la figure 15, nous constatons que les recommandations sont associées à chaque titre sur la base de la popularité, de ce que les utilisateurs ont aussi regardé en plus du film courant et les films du même genre.

Le système de recommandation de Netflix se fonde sur un processus composé de deux phases majeures[120] :

- La création d'un compte abonné sur Netflix est associée à une possibilité d'exploration des préférences de l'utilisateur, qui se voit proposer de certains titres à regarder, en vue de recueillir les données présentées ci-dessus. Toutefois, si l'utilisateur n'opte pas pour cette possibilité, le système de recommandation procédera à des suggestions basées sur la popularité. Mais, avec le temps, les derniers films regardés orienteront les futures recommandations.

- Une fois le système de recommandation initialisé, Netflix mise sur l'organisation et la structure des recommandations, pour rendre l'expérience plus personnalisée, en présentant les titres par catégorie ou section sur la page d'accueil en fonction des préférences de chaque utilisateur. Par exemple, les sections les plus recommandées sont affichées en haut de la page et les titres sont listés de gauche à droite[121].

Ainsi, la personnalisation concerne trois niveaux distincts :

- Choix des sections ;
- Choix des titres appartenant à la section ;
- Ordre de ces titres.

Pour étayer nos propos par des chiffres, 80 % des contenus lus sur Netflix découlent du système de recommandations, ainsi une grande partie des titres qui seront vus sont le fruit de l'algorithme de recommandation[122].

[120] GOMEZ-URIBE, Carlos A. et HUNT, Neil. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*. 2016, vol. 6, no 4, p. 13.
[121] Ibid.
[122] WHITHMAN, B., Op.cit., p.123.

Par ailleurs, Netflix traite aussi les données liées au contenu, en analysant les notes et les commentaires laissés par les utilisateurs, mais également par des marqueurs. Ainsi, chaque titre est associé à plusieurs tags et à une vignette de présentation, permettant d'identifier sa nature et ses caractéristiques, afin que l'utilisateur puisse le découvrir et prendre la décision de le regarder.

En résumé, il importe de souligner que ce nouveau mouvement de recommandation a offert aux clients un large éventail de choix et d'options. En revanche, ce nouveau mode de consommation a augmenté le volume d'informations que l'utilisateur peut investir avant de prendre sa décision d'achat ou d'usage[123].

7. Les modèles de développement des profils des utilisateurs dans un système de recommandation

Cette partie voudrait aborder l'élaboration du profil de l'utilisateur dans les systèmes de recommandation.

La modélisation de Souissi pour la création des profils utilisateurs (figure 16) [124] s'opère suivant quatre phases : la collecte de données, la préparation (ou structuration) des données, l'analyse de données et la représentation des données.

[123] CIOCCA, S., "How Does Spotify Know You So Well?".*Medium*. October 10, 2017.

[124] SOUISSI, Amen. *Modélisation centrée sur les processus métier pour la génération complète de portails collaboratifs* [en ligne]. Thèse de doctorat. Université des Sciences et Technologie de Lille - Lille I, 2013. Disponible sur : https://tel.archives-ouvertes.fr/tel-00935324/PDF/ThA_se_Amen-SOUISSI.pdf [consulté le 12 Septembre 2018].

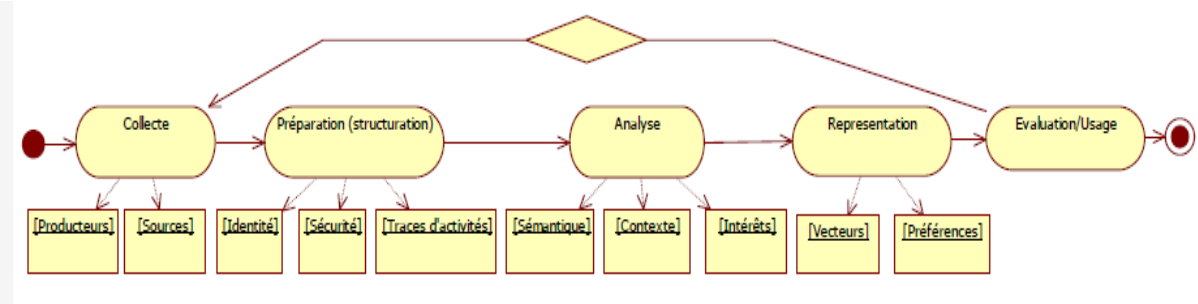

Figure 16 : La représentation sous forme de diagramme d'activités du processus de développement (quatre premières activités) et d'usage (activité évaluation/usage) des profils utilisateurs (SOUSSI, 2013)

Les profils construits sont évalués ou utilisés dans différents systèmes (personnalisation, recommandation, etc.), qui peuvent nécessiter la reprise du processus (étape évaluation/usage).

La collecte de données permet de construire les profils. Ces données peuvent être fournies par les utilisateurs de manière explicite (via des formulaires par exemple) ou de manière implicite (collecte automatique des traces d'activités) à partir de plusieurs sources : les systèmes d'exploitation (les fichiers log de bases de données, les logs de serveurs Web, etc.) et les applications (l'email, le social bookmarking, etc.). Toutefois, deux problèmes surgissent : la multiplicité des sources de données, notamment avec les technologies du Web 2.0 et la multiplication des identités numériques, ce qui génère beaucoup de données partagées au sein de diverses applications.

Au-delà des applications ou des systèmes représentant les sources de données, il est également intéressant de prendre en compte les utilisateurs qui produisent ces données pour construire le profil de l'utilisateur, soit à travers les traces d'activités de l'utilisateur, soit à partir des interactions entre l'utilisateur et le système, surtout pour le cas des utilisateurs inactifs et qui ne laissent que peu de traces sur le système.

La collecte de données peut être représentée par la figure suivante[125] :

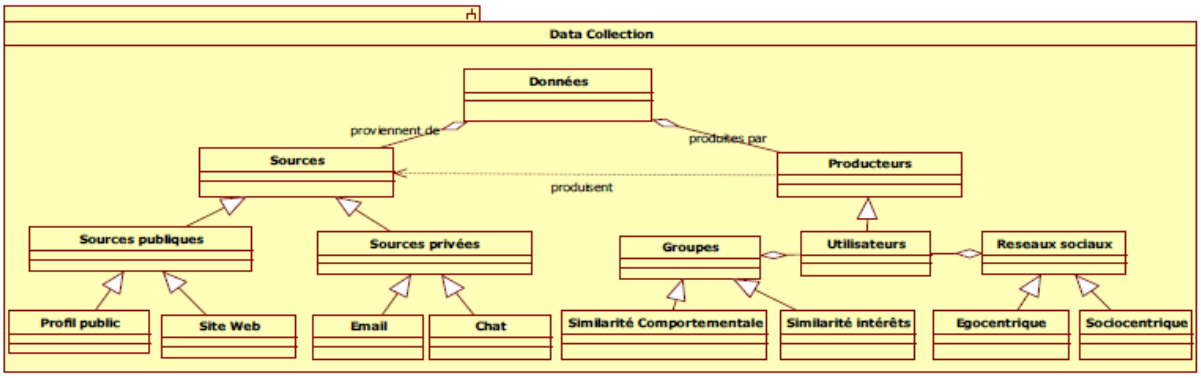

Figure 17 : Le modèle de données de la phase de collecte de données (SOUSSI, 2013)

Le grand enjeu demeure de recouper facilement ces données dans les systèmes d'information par fusion des identités partielles d'utilisateurs (OpenID, Shibboleth, CardSpace, Liberty Alliance, etc.) et la fiabilité des sources de données, qui renvoie à la dimension qualité attendue du modèle de profil.

La préparation des données consiste à constituer un jeu de données de qualité homogène, dont le format et la structure sont cohérents et bien définis. De fait, il s'agit de traiter les valeurs manquantes, aberrantes, nulles et homogénéiser les échelles des données d'identité, d'activité et de sécurité collectées sur l'utilisateur.

Selon Souissi, une fois le modèle de données établi, les règles d'association qui en découlent sont définies (les utilisateurs qui ont acheté le produit de gamme X, ont également acheté les produits de gamme Y). Donc intuitivement, pour un utilisateur ayant acheté un produit de gamme X, les produits de gamme Y peuvent être utilisés pour enrichir son profil).
La seconde approche de solution qui se développe vise à améliorer la précédente, en n'inférant le profil d'un utilisateur qu'à partir des individus qui lui sont similaires, mais aussi en qui il a pleinement confiance ou qui influencent réellement son comportement, cela implique l'usage de nouvelles données relationnelles entre les utilisateurs, les réseaux de confiance par exemple ou de manière plus générale les réseaux sociaux.

Les données d'identité et les données de sécurité sont explicitement renseignées par les utilisateurs. En revanche, les données sur les activités des utilisateurs sont soumises à des

[125] Ibid., p.127.

traitements appropriés, afin d'en extraire les centres d'intérêts pertinents de l'utilisateur, comme l'illustre la figure ci-après[126] :

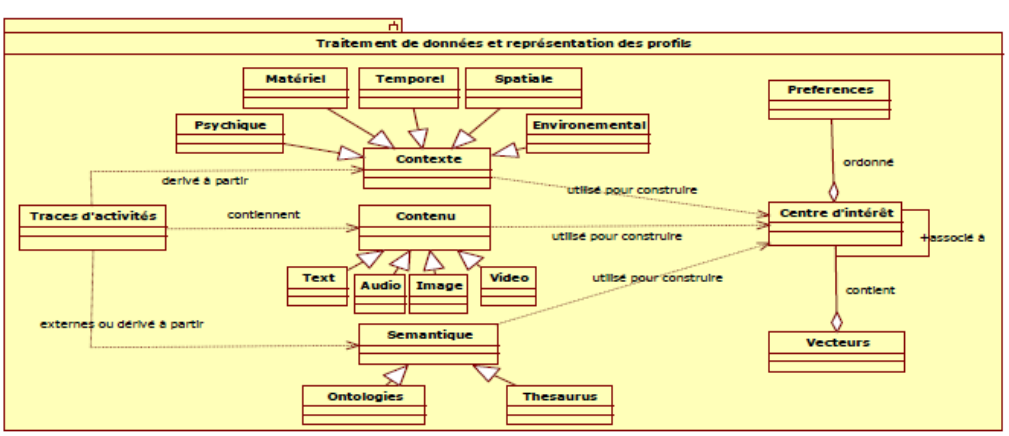

Figure 18 : Le modèle de traitement et de représentation des profils (SOUSSI, 2013)

En matière des traces d'activités, ces données peuvent être décomposées en trois grandes dimensions : contenu, contexte et sémantique

- Le contenu renvoie aux données extraites à partir des traces d'activités des utilisateurs et qui désignent le contenu généré par l'utilisateur (Exemple : un commentaire ou la description de la ressource avec laquelle l'utilisateur a interagi). Au-delà du texte, le contenu peut également désigner des données multimédias (audio, image, vidéo). La seule différence réside dans la nature des algorithmes de fouille de données nécessaires pour les analyses ;

- Le contexte renvoie aux facteurs ayant un impact sur l'activité de l'utilisateur. Dans la modélisation des profils utilisateurs, la notion du contexte est multiple, il peut être temporel (lié au temps), spatial (lié à la localisation géographique), matériel (lié au matériel utilisé, exemple : ordinateur, téléphone mobile, etc.) ou le contexte psychologique ou émotionnel (lié à l'humeur) ;

- La sémantique correspond aux analyses textuelles, qui donnent plus de sens à la terminologie manipulée selon des domaines précis. Il s'agit d'associer un terme à son

[126]SOUISSI, Amen. *Modélisation centrée sur les processus métier pour la génération complète de portails collaboratifs* [en ligne]. Thèse de doctorat. Université des Sciences et Technologie de Lille - Lille I, 2013. Disponible sur : https://tel.archives-ouvertes.fr/tel-00935324/PDF/ThA_se_Amen-SOUISSI.pdf [consulté le 12 Septembre 2018].

usage dans le contexte de l'utilisateur, en s'appuyant principalement sur des ontologies. Il peut s'agir de réutiliser des ontologies de référence existantes (Wordnet ou ODP par exemple) ou construire une nouvelle ontologie à partir des textes analysés.

8. L'évaluation des systèmes de recommandation

Comme les systèmes de recommandation relèvent de la recherche d'information, il est judicieux d'essayer utiliser des mesures d'évaluation existantes pour la recherche d'information au domaine de la recommandation.

L'évaluation de systèmes de recommandation peut se faire, selon GUY[127], selon trois méthodes distinctes :

- L'évaluation hors ligne : elle repose sur l'historique des comportements des utilisateurs jusqu'à un instant donné, en vue d'identifier leurs comportements futurs à partir des comportements passés. Pour y parvenir, les données relatives aux actions des utilisateurs (généralement des évaluations) sont réparties en données d'apprentissage et de test, de façon à ce qu'un sous ensemble des actions utilisateurs soit caché et que le système de recommandation cherche à le prédire sur la base du sous-ensemble des données d'apprentissage pour juger la pertinence du système de recommandation en question ;

- La deuxième méthode (études sur un échantillon) : elle consiste à recruter un groupe d'utilisateurs volontaires d'un système et d'observer le processus d'usage du système de recommandation lors d'une expérimentation menée. De la même manière, il est possible de demander les retours des utilisateurs sur leur interaction avec le SR. Le questionnaire permet aussi de collecter des données qualitatives, pour expliquer les résultats quantitatifs. Cependant, cette approche risque parfois de ne recruter qu'un petit nombre de participants, ce qui rend les conclusions peu généralisables ;

[127] GUY, Shani and GUNAWARDANA, Asela. Evaluating recommendation systems, Recommender Systems Handbook. *Springer US*. January 2011, p257-297

- Le dernier type est l'évaluation en ligne : elle concerne un échantillon des utilisateurs réels du système en temps réel, l'objectif étant d'observer leur comportement et le comparer au reste de la population. Ce test peut être une simple comparaison des chiffres d'audiences avant et après l'application du SR.

L'évaluation des systèmes de recommandation est un peu délicate, compte tenu de la nature des données sur lesquelles ils sont appliqués. Toutefois, nous pouvons présenter deux catégories de mesures mathématiques, s'inscrivant dans le cadre de l'évaluation hors ligne selon YANG [128]:

- **Les mesures de la précision prédictive :**

- L'erreur Moyenne Absolue (MAE : Mean Absolute Error) : elle mesure le taux de déviation des recommandations prédites, en comparaison avec les choix faits par l'utilisateur. Cette mesure est pertinente pour évaluer globalement la capacité du système de recommandation à prédire les notes des utilisateurs.

Cette mesure permet de calculer l'écart entre les scores prédits et les scores réels :

$$|\bar{E}| = \frac{\sum_{(c,s)\in W} |u^p(c,s) - u(c,s)|}{|W|}$$

- U (c, s) est le score réel
- Up (c, s) est le score prédit par le SR
- W = {(c, s)} est la paire *usager-objet* avec lequel le système de recommandation fait la prédiction.

- L'Erreur Moyenne Absolue Normalisée (NMAE : Normalized Mean Absolute Error) : c'est la version normalisée de MAE, présentée directement au-dessus. Elle permet de comparer deux échelles de notation qui ne sont pas égales :

[128]YANG, Zhe, WU, Bing, ZHENG, Kan, et al. A Survey of Collaborative Filtering-Based Recommender Systems for Mobile Internet Applications. IEEE Access. 2016, vol. 4, p. 3273-3287.

$$NMAE = \frac{MAE}{v_{max} - v_{min}}$$

Sachant que :

- MAE= L'erreur Moyenne Absolue
- V =Il est à préciser que les Vmax et Vmin sont respectivement la valeur maximale et minimale de notation.

- L'Erreur Moyenne Absolue par Utilisateur (UMAE : User Mean Absolute Error) : Elle consiste à déterminer le niveau de satisfaction par utilisateur. De fait, nous procédons, d'abord, au calcul d'une moyenne locale des valeurs prédites par utilisateur, et ensuite au calcul de la moyenne générale[129].

Cette fonction vérifie si le système satisfait le maximum de ses utilisateurs, puisqu'une bonne MAE peut être obtenue à partir de peu d'utilisateurs :

$$UMAE = \frac{\sum_{j=1}^{N} \frac{\sum_{i=1}^{N} j |p_{ij} - a_{ij}|}{N_j}}{N}$$

Il est à souligner que N représente le nombre total d'utilisateurs, Nj est le nombre des notes prédites pour l'utilisateur j.

- **Mesures de précision de classement :**

Ces mesures sont les mêmes utilisées dans le domaine de la recherche d'information, il s'agit notamment de :

- La précision : elle renvoie à la qualité des recommandations, il s'agit du nombre de suggestions pertinentes/ nombre total des suggestions, ce ratio est calculé comme suit :

[129] Ibid., p.132.

$$\text{Précision} = \frac{\text{nombre de } \mathbf{bonnes} \text{ recommandations}}{\mathbf{nombre\ total} \text{ de recommandations}}$$

- Le rappel : il correspond au nombre des suggestions pertinentes communiquées à l'utilisateur / le total de « bon produits » dans le système, ce ratio est calculé comme suit :

$$\text{Rappel} = \frac{\text{nombre de } \mathbf{bonnes} \text{ recommandations}}{\mathbf{nombre\ total\ de\ bons\ produits}}$$

- F-Mesure : c'est une synthèse des deux paramètres, pour remédier aux conflits entre les deux ratios dans certains cas, elle se présente de la manière suivante :

$$\text{F} - \text{mesure} = \frac{2 * \text{Précision} * \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Ainsi, si le système recommande un seul élément à un utilisateur et qu'il est pertinent, alors, la précision est de 1, alors que le rappel est peu élevé. Ce score est défini comme la moyenne harmonique entre le rappel et la précision[130].

9. Les systèmes de recommandation en e-commerce

A l'identique des vitrines « produits », les sites e-commerce poursuivent principalement l'objectif de générer des dividendes maximaux, en essayant d'orienter les clients vers des produits pertinents. Le rôle qui incombe aux vendeurs dans une boutique physique, peut être entrepris par un système de recommandation.

[130]YANG, Zhe, WU, Bing, ZHENG, Kan, et al. A Survey of Collaborative Filtering-Based Recommender Systems for Mobile Internet Applications. *IEEE Access*. 2016, vol. 4, p. 3273-3287.

La recommandation e-commerce[131] a pour objectif de :

- Permettre une grande visibilité sur les produits disponibles ;
- Attirer les internautes sur les plateformes marchandes ;
- Faciliter l'exploration des produits ;
- Augmenter les ventes et le chiffre d'affaires.

Les systèmes de recommandation s'alimentent généralement de plusieurs sources de données[132]:

- D'abord, l'information textuelle générale ou détaillée, fournie par l'utilisateur, de manière explicite (remplir un formulaire, etc.) ou implicite (collecte des informations de plusieurs sources) ;
- L'évaluation numérique selon une échelle donnée (1-5 par exemple) ;
- L'évaluation ordinale avec l'attribution d'un jugement ou un avis (neutre, avec, contre, etc.) ;
- L'évaluation binaire demandant à l'utilisateur de décider entre deux options ;
- L'évaluation unaire pour le cas des objets utilisés / achetés par l'utilisateur et ce qui reflète la qualité de cet objet ;
- L'émission des commentaires ou des tags sur liste de choix prédéfinis ou libre par l'utilisateur.

Les recommandations proposées à l'utilisateur dépendent de la connaissance des utilisateurs du domaine, de leur localisation géographique et du temps de la recommandation.

S'il est difficilement imaginable d'avoir un bon magasin sans bons vendeurs, il est tout aussi difficile d'imaginer un bon site e-commerce sans un moteur de recommandation efficace.

[131] Silicon Salad Marketing. *La recommandation, ou comment vendre plus avec des suggestions personnalisées* [En ligne]. Disponible sur : https://www.siliconsalad.com/blog/recommandation-e-commerce-les-nuls/ [Consulté le 22 Août 2018].
[132] OARD, Douglaset W., KIM, Jinmook, et al. Implicit feedback for recommender systems. In Proceedings of the AAAI workshop on recommender systems. 1998. p. 81-83.

En effet, la recommandation, en e-commerce, est basée sur trois étapes essentielles, selon le blog de l'agence web Silicon Salad[133] :

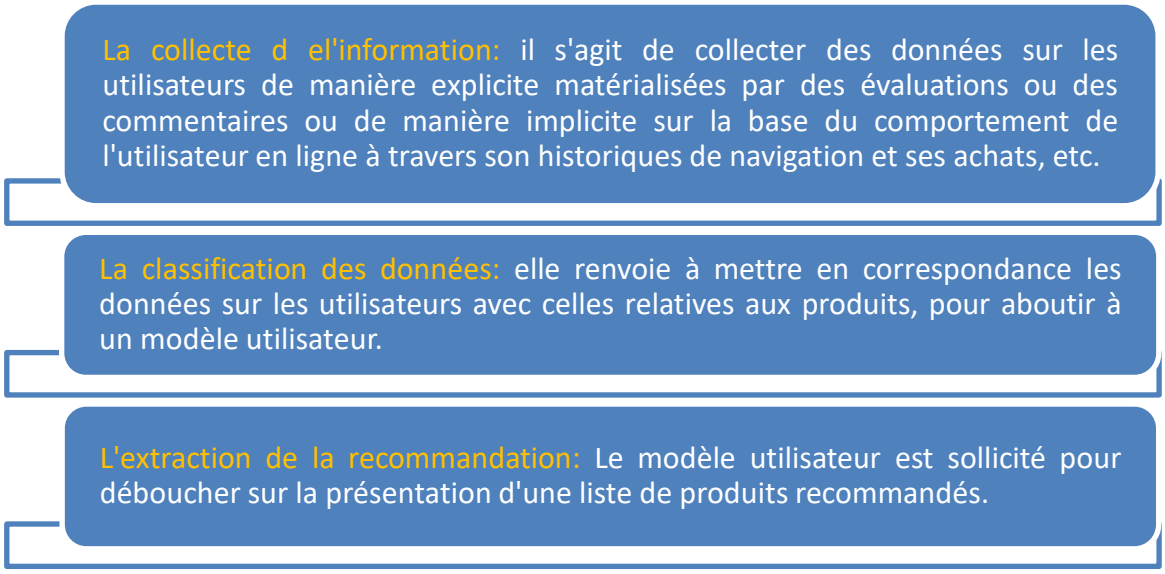

Figure 19 : Le processus de recommandation en e-commerce
(https://www.siliconsalad.com/blog/recommandation-e-commerce-les-nuls/**)**

Le choix ou la sélection d'objets ou d'items correspondant aux besoins des utilisateurs et des clients représente la première vocation des sites marchands et la solution optimale à une surcharge d'informations n'est autre que les systèmes de recommandation.

Le e-commerce est devenu une priorité inévitable pour la majorité des entreprises. De ce fait, l'augmentation du chiffre d'affaires du secteur a été multipliée par deux en 2012 par rapport à 2011. L'objectif est de susciter l'achat d'impulsion ou de deviner plus précisément ce que chaque internaute souhaite acheter et ainsi de conclure la vente et d'amplifier le panier moyen.

Depuis ses débuts, l'e-commerce s'est servi d'une personnalisation basée presque uniquement sur une connaissance des clients avec un niveau de filtrage collaboratif simple « les clients qui ont acheté X ont aussi acheté Y », « meilleures ventes », etc.

De nos jours, le niveau d'exigence du client est de plus en plus élevé, la personnalisation peut consister en la création de profils grâce auxquels les acheteurs indiquent le type de caractéristiques ou d'informations qu'ils privilégient. D'autres méthodes sont également utilisées, certaines simples et d'autres plus complexes :

[133]Silicon Salad Marketing. Op.cit., p.135.

• Le versioning[134]: présenter plusieurs versions d'une même page d'un site e-commerce en fonction des catégories de segmentation des clients ;
• La recommandation de produits : présentation ciblée des produits, en se basant sur des données anonymes (implicites, explicites ou les deux) collectées au sujet d'un acheteur ;
• Les solutions de filtrage interactives : présenter des informations spécifiques basées sur une consigne du client.

Ces méthodes permettent de tenir compte du comportement individuel des clients et des relations entre les produits ou les groupes de produits, afin de prévoir quels items ou objets pourraient les intéresser. Concrètement, la recommandation peut intégrer le versioning, en présentant une version distincte de la page du site e-commerce selon les clients, afin de présenter les produits, en fonction des préférences des clients, identifiées par le biais de leurs données collectées et leurs traces numériques.

- L'exemple d'Amazon.com :

Amazon est une entreprise du commerce en ligne, considérée parmi les pionniers de ce domaine. De fait, Amazon a bouleversé la sphère du e-commerce avec l'intégration des systèmes de recommandation sur sa plateforme, en jouant la carte de la personnalisation de l'expérience d'achat de chaque utilisateur, qui peut découvrir des recommandations de produits sur mesure.

Sur le plan technique et selon JAYASHRI[135], le moteur de recommandation d'Amazon utilise principalement l'approche du filtrage collaboratif basé sur les items ou les objets (d'item-to-items) à base de la mémoire (données déjà stockées dans la mémoire du système), d'où les fameuses expressions : "les personnes qui ont vu/acheté ces articles ont aussi vu/acheté ces articles". Les mécanismes déployés, pour y parvenir sont les suivants :

- La détermination de la correspondance entre un objet donné et d'autres plusieurs objets ou items ;

[134] La gestion des différentes versions d'un document.
[135] JAYASHRI, Sonawane ; SHIVANI, Gore ; LAXMI, Duggineni, et al. A Survey on Recommendation System Used in E Commerce. *International Journal of Advanced Research in Computer and Communication Engineering*. 2016, vol. 5.

- La prise en compte des notations explicitement fournies par les clients (variant de 1 étoile jusqu'à 5 étoiles) et des données de navigation et d'achat (temps passé sur un item, le panier des achats, etc.) ;
- La génération d'un tableau des éléments que les clients ont tendance à acheter ;
- La construction d'une matrice de produits à travers itération de toutes les paires d'éléments et du calcul de la similarité entre ces paires ;
- La proposition des suggestions les plus appropriées[136].

Dans le détail, les données d'achat, de notation et de navigation des clients représentent la principale source de données recueillies par le biais du mécanisme d'authentification de compte pris en charge par Amazon. En conséquence, une fois connectés à leur compte, les clients ont droit à des recommandations sur la page web principale de la plateforme, Un espace dédié aux recommandations « vos recommandations », permet aux clients de filtrer les propositions par ligne de produits, évaluer les produits recommandés et même noter les achats précédents.

Toutefois, il est utile de préciser que le système de recommandation d'Amazon utilise aussi l'approche hybride alliant le filtrage collaboratif et la recommandation basée sur le contenu. Ainsi, les recommandations sont, d'une part, fruit des propriétés de l'objet lui-même (filtrage basé sur le contenu) et, d'autre part, les comportements d'autres usagers (filtrage social).

Il est à signaler qu'en 2016, Amazon a basculé son système de recommandation en open source, il s'agit du système de recommandation DSSTNE (Deep Scalable Sparse Tensor Network Engine, qui est désormais accessible sur Github sous licence Apache2.

Pour le système de recommandation du site de vente des livres, Amazon propose une panoplie de services aux utilisateurs passionnés par les nouveautés du monde de l'édition et des livres de tout type :

- Le système de recommandation suggère les auteurs dont les livres sont les plus vendus ;
- Le système de recommandation propose les livres achetés par l'utilisateur autres que la transaction courante ;
- Le système de recommandation recommande des livres les plus achetés par les usagers.

[136] Ibid., p.137.

Il y a également d'autres services de notification des utilisateurs pour les nouveautés via email, en s'appuyant sur la base de leurs requêtes. Par ailleurs, l'interaction des utilisateurs est omniprésente à travers l'émission de leurs feedbacks (like, dislike, etc.) et les commentaires à propos des différents contenus[137].

10. Les systèmes de recommandation dans les industries du contenu et les SID

Pour faire face au défi de la personnalisation de l'expérience des internautes sur la toile, les systèmes hypermédia adaptatifs tendent à adapter les contenus au contexte de l'utilisateur, et en particulier la famille des systèmes de recommandations déployés dans plusieurs contextes (e-commerce, la musique, les vidéos, la presse, les jeux, etc.). Quant aux services d'information documentaire, nous soulignons qu'il existe un certain nombre de systèmes liés notamment aux bibliothèques numériques, permettant un accès à plusieurs niveaux pour un public élargi d'utilisateurs et de nouvelles opportunités pour la bibliothèque[138].

Au demeurant, cette nouvelle génération de bibliothèque a tiré parti des mutations des nouvelles technologies, afin de mieux servir les utilisateurs, et ceci dans une optique de personnalisation et de filtrage opérés sur le volume considérable des données disponibles.

Les systèmes de recommandation ont été sollicités, pour aider les utilisateurs dans l'exploitation des collections de documents, comme nous allons l'illustrer par les exemples présentés ci-dessous [139] :

---

[137] *« Recommendation Engines: How Amazon and Netflix Are Winning the Personalization Battle MarTech Advisor »* [En ligne]. Disponible sur : https://www.martechadvisor.com/articles/customer-experience/recommendation-engines-how-amazon-and-netflix-are-winning-the-personalization-battle/. [Consulté le 01 Septembre 2019].

[138] PORCEL, Carlos, MORENO, Juan Manuel, et HERRERA-VIEDMA, Enrique. A multi-disciplinar recommender system to advice research resources in University Digital Libraries. *Expert Systems with Applications* [En ligne]. 2009, vol. 36, n. 10. Disponible sur : https://www.sciencedirect.com/science/article/abs/pii/S0957417409003698?via%3Dihub [consulté le 12 Janvier 2018].

[139] Ibid.

- **L'ACM digital Library :**

L'ACM Digital Library met à la disposition des utilisateurs des recommandations via son portail : http://portal.acm.org , en s'appuyant sur deux techniques :

- Une technique basée sur le contenu des documents de recherche, qui repose sur la recommandation des articles similaires et connexes à chaque requête effectuée ;

- Une technique basée sur les données de navigation des utilisateurs, qui permet de recommander ce que les lecteurs d'un article courant ont également lu[140].

La recommandation sur ACM Digital Library intervient quand un utilisateur accède aux données bibliographiques relatives à un article sur une liste de résultats retournée, pour qu'une notification de recommandation apparaisse, comme nous pouvons le constater sur l'exemple suivant[141] :

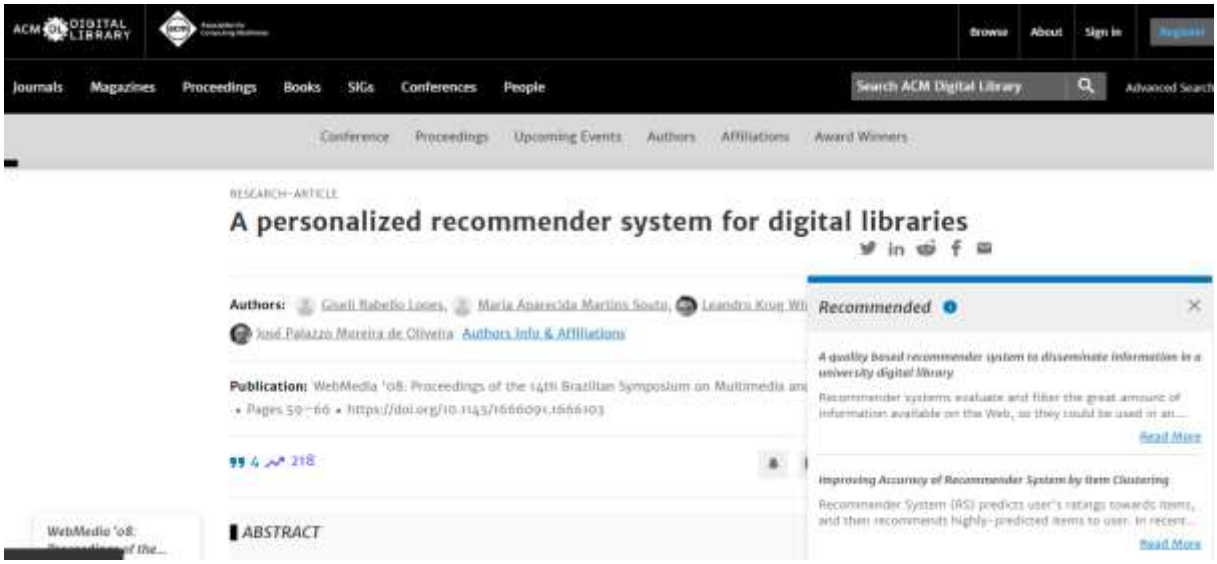

Figure 20 : La recommandation sur ACM Digital Library (https://dl.acm.org/journals)

[140] FRANKE, Markus; GEYER-SCHULZ, Andreas; NEUMANN, Andreas W. Recommender services in scientific digital libraries. *Multimedia services in intelligent environments*. Springer Berlin Heidelberg.2008, p. 377-417.
[141] ACM Digital Library. *Journals* [En ligne]. Disponible sur : https://dl.acm.org/journals [consulté le 06 Avril 2020].

- **Tech Lens + (solution de recommandation pour les SID numériques) :**

Le système de recommandation TechLens a été initié par GroupLens, qui est un laboratoire de recherche spécialisé en interaction homme-machine et relevant du département d'informatique et d'ingénierie de l'université américaine de Minnesota. Le système peut être greffé sur les services d'information documentaire numériques depuis son portail : https://grouplens.org/about/projects/. Tech Lens est système hybride qui combine deux approches :

- Une approche basée sur le filtrage collaboratif : la prédiction de l'avis d'un utilisateur sur un article donné repose sur les évaluations déjà soumises par d'autres utilisateurs similaires à propos de cet article, soit la mesure de la similarité entre les utilisateurs. En sus, les méthodes du filtrage collaboratif fonctionnent aussi sur la base du graphe de citations et les liens entre les citations, en calculant la similarité entre deux articles à partir des citations communes ;
- Une approche basée contenu : la génération des recommandations pour les articles consiste en une analyse textuelle des articles, permettant de calculer les scores de similarité les plus élevés.

- **BookPsychic**

BookPsychic est un système de recommandation simple pour les usagers de la bibliothèque, à l'image de Netflix ou Amazon, il permet d'évaluer les livres et les DVD et enregistre les préférences des utilisateurs, pour suggérer des listes de recommandation depuis l'adresse web : http://www.bookpsychic.com/. Voici l'aperçu de sa page d'accueil :

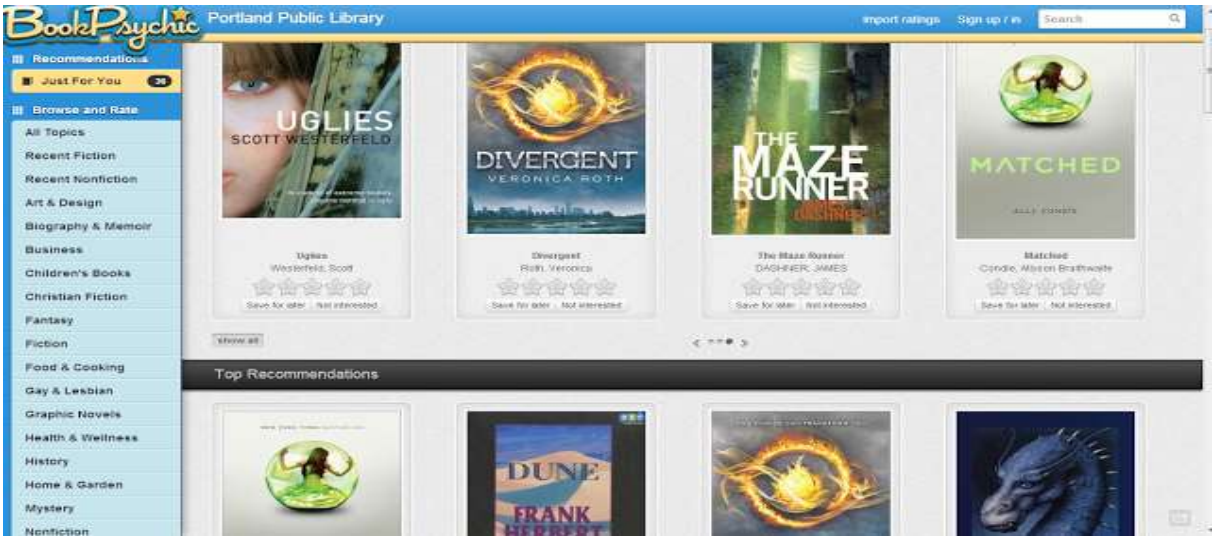

Figure 21: L'interface de BookPsychic système de recommandation pour une bibliothèque numérique (http://www.bookpsychic.com/)

En effet, il s'agit de choisir une bibliothèque qui est inscrite dans bookpsychic et de noter les livres par genre, en vue d'obtenir des recommandations.

- **Huddersfield Book Recommender system :**

Il s'agit d'un système qui permet aux utilisateurs de contribuer explicitement à la génération des recommandations, en ajoutant manuellement des titres similaires à des livres à partir du portail : https://library.hud.ac.uk/. La recommandation prend la forme suivante :

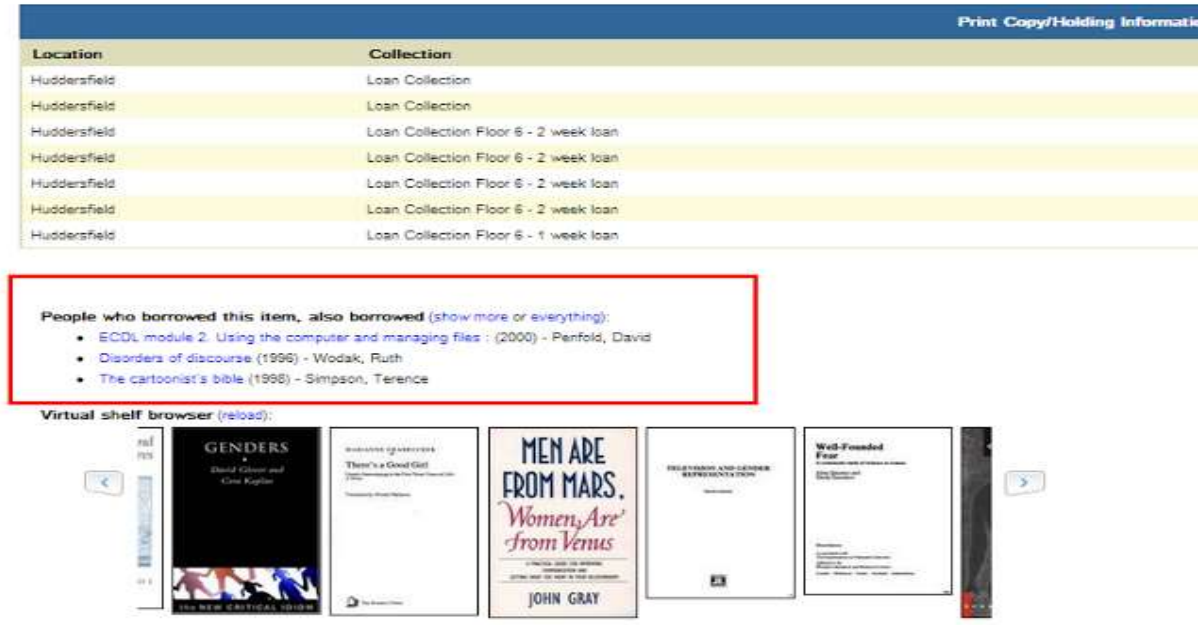

Figure 22: L'interface de Huddersfield Book Recommender System (système de recommandation pour une bibliothèque numérique) (https://library.hud.ac.uk/)

De même, le système permet à d'autres utilisateurs de suivre ce qu'ils peuvent évaluer dans les ressources disponibles.

- **Ex Libris-bX Recommender**

bX est un service web commercialisé par ExLibris (2009) basé sur les recherches effectuées au laboratoire national de Los Alamos au Mexique par Johan Bollen et Herbert Van de Sompel, et qui recommande des articles de recherche utilisant les données obtenues à partir des revues à travers les OpenURL « Uniform Ressource Locator », qui est un protocole standardisé, permettant à un utilisateur d'atteindre facilement les ressources qu'il a l'autorisation de consulter.

En sus, bX Recommender d'Ex Libris collecte les données d'utilisation des ressources documentaires et informationnelles provenant des scientifiques et des établissements universitaires du monde entier, afin d'aider les utilisateurs à bénéficier des recommandations d'articles et livres électroniques pertinents. Concrètement, si deux articles sont utilisés par un utilisateur lors d'une même connexion, le système analyse les liens existants entre eux, pour fournir aux utilisateurs d'autres articles sur le même sujet, servant à préciser ou à élargir les

sujets de recherche initiaux et leur permettre de trouver d'autres ressources grâce à la découverte inattendue.

Parallèlement, les recommandations de bX se nourrissent des données des résolveurs de liens entre les articles, déployés par les systèmes de découverte, les plateformes des éditeurs et toute autre source qui lie les utilisateurs à un texte complet par l'intermédiaire d'un résolveur de liens, alors que les articles peuvent provenir de différentes revues, éditeurs ou plateformes.

Il s'agit d'un service qui exploite la puissance de la communauté scientifique en réseau, pour générer des recommandations basées sur l'utilisation des articles. Voici un aperçu d'une interface d'une bibliothèque numérique intégrant le bX Recommender d'Ex Libris :

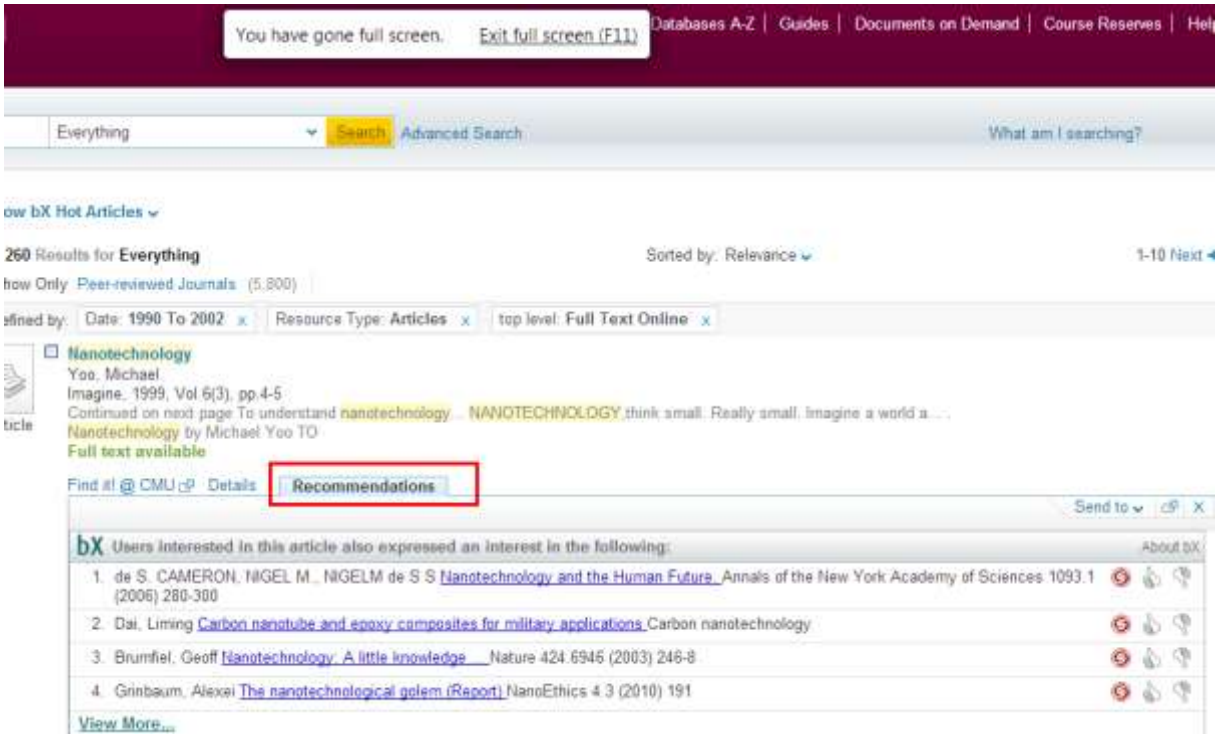

Figure 23: L'interface de recherche de la bibliothèque universitaire Michigan intégrant bX Recommender (https://www.exlibrisgroup.com/fr/produits/bx-recommender/**)**

Comme le montre l'aperçu sur l'interface de bX Recommender, les résultats d'une recherche sont ordonnés selon la pertinence, dépendant aussi bien de la spécialité de l'étudiant que du cycle dans lequel il suit ses études pour le cas de l'université de Michigan dans l'exemple ci-dessus. L'intégration de bX Recommander s'effectue par des API dans les interfaces des

catalogues en ligne, les outils de découverte Primo et Summon, les interfaces de résolveur de liens SFX et Alma et la solution de listes de lectures Leganto[142].

Le service de recommandation bX présente les articles retournés pour donner suite à une requête avec la recommandation d'autres articles similaires, comme en rend compte la figure ci-après[143] :

Figure 24 : Le service de recommandation bX implémenté sur une recherche des articles scientifiques (https://www.exlibrisgroup.com/fr/produits/bx-recommender/)

A partir d'une fiche de métadonnées d'un article, bX trouve d'autres articles similaires en intégrant les critères suivants :

- Des mêmes mots-clés ou des mots-clés différents ;
- Dans différentes bases de données ;
- Dans différentes revues.

Ce qui permet d'approfondir la sérendipité dans les résultats de recherche.

Les données *log* d'usage analysées à partir des résolveurs de liens, mettant en œuvre la norme Open URL et faisant le lien entre une citation bibliographique et le texte intégral représentent les sources de données du service bX en lui permettant :

[142] EX LIBRIS GROUP. *BX Recommender* [en ligne]. Disponible sur : https://www.exlibrisgroup.com/fr/produits/bx-recommender/. [Consulté le 01 Septembre 2019]
[143] Ibid.

- D’identifier les chemins de recherche d'informations des utilisateurs d’une manière standardisée ;
- De croiser et comparer l’usage de plusieurs réservoirs et bases de données d’articles scientifiques : le fichier central Primo, les bases de données en texte intégral, Google Scholar, PubMed, etc.,
- La valeur ajoutée des données sur bX réside dans le fait qu’elles soient agrégées à partir de plusieurs sources d'information et de communautés scientifiques, formant un échantillon représentatif de leurs activités[144].

Il est à noter qu’il existe aussi des services qui permettent de greffer un système de recommandation sur les bibliothèques numériques, comme c’est le cas d’Amazon Personalize :

- **Le web services d’Amazon “AMAZON Personalize”**

Amazon Personalize est un service de la famille des services web d’Amazon, qui est dédié à l’apprentissage automatique permettant aux développeurs de créer des recommandations à partir du service AWS : https://aws.amazon.com/fr/personalize/. Ainsi, des interfaces de programmation applicative (API) peuvent être greffées dans des systèmes intégrés de gestion de bibliothèques comme Koha[145].

Le web service traque les activités des utilisateurs d’une plateforme ou d’une application donnée (clics, pages vues, inscriptions, achats, etc.) et il peut être aussi alimenté par des articles à recommander ou par les données démographiques supplémentaires provenant des utilisateurs, telles que leur âge ou leur emplacement géographique.

Au niveau du processus de recommandation, le service web d’Amazon examine, d’abord, les données, élimine celles qui ne véhiculent pas de sens, puis sélectionne les algorithmes correspondants, pour créer un modèle de personnalisation ajusté aux données.

Le web service d’Amazon « Amazon Personalize » propose trois services :

[144] VELLINO, Andre. Recommending research articles using citation data, *Library Hi Tech*., 2015, vol. 33, No. 4, pp. 597-609.

[145] Capgemini. Service Amazon Personalize [En ligne]. Disponible sur : https://www.capgemini.com/service/amazon-personalize/ [Consulté le 05 Mai 2021].

- Les recommandations personnalisées

Les recommandations de produits visent à se conformer aux profils et aux habitudes d'un utilisateur et réussir une conversion d'achat. Ainsi, Amazon Personalize peut permettre aux plateformes de transformer le contenu en fonction du comportement, de l'historique et des préférences de chaque utilisateur, de façon à améliorer l'engagement et la satisfaction du client.

- La recherche personnalisée

Pour faire face aux nombres recherches infructueuses et non pertinentes effectuées par les utilisateurs, le service web d'Amazon propose des résultats de la recherche cadrant avec leurs intérêts pour le terme recherché.

- Les notifications personnalisées

Les messages marketing basés sur le comportement des utilisateurs garantissent un taux de conversion élevé, c'est dans cette optique qu'Amazon Personalize propose à chaque utilisateur de recevoir la communication marketing la plus pertinente au bon moment.

Après avoir passé en revue quelques solutions de systèmes de recommandation déployées dans des bibliothèques numériques repérées, nous résumons leurs caractéristiques et approches algorithmiques de recommandation dans le tableau 12 ci-dessous :

Critères / Solutions de SR	Bibliothèque physique	Bibliothèque numérique	Données explicites	Données implicites	Usager authentifié	Usager anonyme	Filtrage collaboratif à base de modèle (items et users)	Filtrage collaboratif à base de mémoire (items et users)
ACM Digital Library	✗	✓	✓	✗	✗	✓	✗	✓
Book Physic et Huddersfield Book Recommender system	✗	✓	✓	✗	✗	✓	✗	✓
Ex Libris Bx recommender	✗	✓	✓	✗	✗	✗	✗	✓
Web service « Amazon Personalize »	✓	✗	✗	✓	✓	✗	✗	✓

Tableau 12 : Comparaison des systèmes de recommandations de certaines bibliothèques numériques

Le tableau montre que, la majorité des systèmes de recommandation déployés sur plusieurs bibliothèques numériques ont tendance à être associés à des services d'information

documentaire numériques et non pas à un fonds avec une existence physique**. Par ailleurs, le mécanisme de collecte de données est majoritairement explicite, c'est-à-dire qu'il demande à l'utilisateur authentifié d'évaluer les différents documents pour lui proposer des recommandations, et non pas implicite qui recueille les données à partir des traces numériques d'interaction de l'utilisateur avec le système.**

De même, l'approche de recommandation de filtrage collaboratif à base de mémoire est la plus utilisée par ces systèmes, puisqu'il s'agit d'exploiter une base de données contenant les notations des utilisateurs sur les contenus et stockées dans la mémoire de serveur, pour être sollicitée en ligne au moment de la génération des recommandations**,** alors que l'approche de recommandation de filtrage collaboratif à base de modèle consiste à prétraiter et à entrainer un modèle, renfermant les profils des utilisateurs et à le solliciter en mode hors ligne, pour procéder à la prédiction, sans surcharge sur la plateforme ou le catalogue en ligne du service d'information documentaire.

Conclusion

Aujourd'hui, le volume de l'information augmente de façon considérable, de sorte qu'il n'est plus difficile pour les utilisateurs d'y trouver des contenus pertinents qui répondent à leurs attentes. Les systèmes de recommandation représentent un vecteur de personnalisation de l'information, ils permettent au travers des sélections de contenus assorties aux besoins des utilisateurs, de réduire la surcharge cognitive et les guider quant à la prise de décision relative au choix à opérer dans le panel des objets/items disponibles.

Un système de recommandation repose sur plusieurs approches, celle basée sur le contenu, celle se focalisant sur les items et une approche hybride. La recommandation est un processus dont les phases sont liées de la collecte des données jusqu'à la génération des suggestions destinées aux utilisateurs. Les systèmes de recommandation font l'objet d'une l'évaluation qui vise, généralement, à mesurer l'écart entre les suggestions et les besoins des utilisateurs par le biais de plusieurs métriques (l'erreur moyenne absolue, le recall, la précision et la F-mesure).

Le système de recommandation s'est développé avec le e-commerce, pour s'étendre à tous les domaines y compris les services et les centres documentaires, notamment pour leurs ressources numériques.

Partie IV : Etude de cas : conception et modélisation d'un SR pour le catalogue de l'IMIST

Chapitre 1 : Présentation de l'IMIST, de son offre et de ses services documentaires

Introduction

Un opérateur national de recherche a mission d'œuvrer pour la promotion, le développement et la valorisation de la recherche scientifique. C'est le cas du Centre National de la Recherche Scientifique et Technique (CNRST), qui met en œuvre des programmes de recherche et de développement technologique, développe l'infrastructure de la recherche et celle de la diffusion de l'information scientifique et technique et conclue des conventions et des partenariats de recherche avec des établissements et organismes de recherche publics ou privés.

Pour faciliter l'accès et la diffusion de l'information scientifique et technique, le CNRST a créé une structure dédiée, il s'agit de l'Institut Marocain de la Recherche Scientifique et Technique (IMIST), qui permet aux chercheurs et aux professionnels de disposer d'un hub d'accès au savoir scientifique et technique de manière adaptée à leurs besoins afin de suivre les mutations scientifiques et technologiques dans les différents domaines et disciplines.

1. Présentation du contexte et du corpus d'étude

L'Institut Marocain de l'Information Scientifique et Technique (IMIST) s'inscrit dans le plan quinquennal de développement 2000-2004 avec un budget d'investissement de 150 millions de dirhams. Il relève du CNRST en sa qualité d'opérateur de la recherche.

La vocation de l'IMIST est de mettre à la disposition des milieux scientifiques et industriels, l'information et la documentation scientifique et technique dont ils ont besoin, pour être à la pointe de leurs activités.

L'IMIST est doté d'une infrastructure logistique importante, pour remplir au mieux ses missions[146] :

[146] INSTITUT MAROCAIN DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE. *IMIST* [en ligne]. Disponible sur http://www.imist.ma [consulté le 10 juin 2020].

Figure 25: L'infrastructure et moyens à l'IMIST (http://www.imist.ma)

Cette structure documentaire est destinée à la communauté des chercheurs débutants et expérimentés :

Les étudiants :

- Élève ingénieur en 4ème année ;
- Étudiant en médecine à partir de la 4ème année ;
- Étudiant en 3ème année licence.

Les chercheurs et personnel scientifique :

- Étudiant en master recherche ;
- Doctorant ;
- Les enseignants chercheurs ;
- Les cadres des administrations publiques ;
- Les cadres des entreprises.

Quant au fonds documentaire, l'IMIST offre, d'une part, les ressources documentaires sous format papier en l'occurrence, des ouvrages, des périodiques, des thèses, des actes de congrès et rapports scientifiques et, d'autre part, des ressources électroniques CD/DVD et d'autres en ligne[147].

[147] INSTITUT MAROCAIN DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE. *Bases de données de l'IMIST* [en ligne]. Disponible sur http://www.imist.ma [consulté le 10 juin 2016].

Aussi, les produits et les services documentaires fournis par l'IMIST varient des catalogues aux bases de données et revues électroniques, comme le montre la figure 26 [148] :

Figure 26: les produits et les services fournis par l'IMIST (http://www.imist.ma)

L'IMIST met aussi à la disposition des universités et des chercheurs plusieurs bases de données spécialisées auxquelles il est abonné à partir d'une plate-forme centralisée d'accès à distance aux ressources électroniques internationales, telles que Scopus, MathScinet, Science Direct, Web of Science, Cairn.info, les Ebooks de Springer, Jstor et Aluka.

Il est à souligner que, le corpus que nous allons traiter dans notre thèse est relatif à l'exploitation des monographies enregistrées dans le catalogue en ligne, qui représente l'un des services majeurs fournis à la communauté d'universitaires et chercheurs adhérents à l'IMIST. Nous avons choisi de valoriser les données maitrisées en local, à savoir les fichiers log générés au moment de la consultation des notices sur le catalogue en ligne des monographies. Les données d'usage des collections de revues mises à la disposition des lecteurs sont des données pilotées par les éditeurs et les fournisseurs des bouquets de revues électroniques, difficilement disponibles de façon personnalisée pour l'objet de notre recherche. Nous avons préféré tester notre moteur de recommandation sur les ouvrages pour ce travail.

[148]Ibid., p.152.

2. Le système documentaire actuel de l'Institut Marocain de l'Information Scientifique et Technique

L'IMIST oriente son offre informationnelle et documentaire vers des utilisateurs composés des chercheurs, des enseignants et des étudiants chercheurs, et de façon moindre vers les professionnels du monde de l'entreprise.

En revanche, L'IMIST ne dispose pas d'un système de recommandation, l'orientation des utilisateurs se limite à un service de référence classique d'une bibliothèque, où l'utilisateur établit un échange (question-réponse) avec les professionnels de l'information, en vue de lui proposer des références ou documents susceptibles de répondre à son besoin informationnel.

Il est possible aussi d'acheminer à l'utilisateur les nouveautés documentaires cadrant avec un profil créé par l'utilisateur lui-même, et ceci en mode "push" à son adresse électronique, selon une fréquence paramétrée dans le cadre de la Diffusion Sélective d'Informations (DSI). De fait, l'utilisateur se trouve devant le scenario suivant (cf. figure 27) :

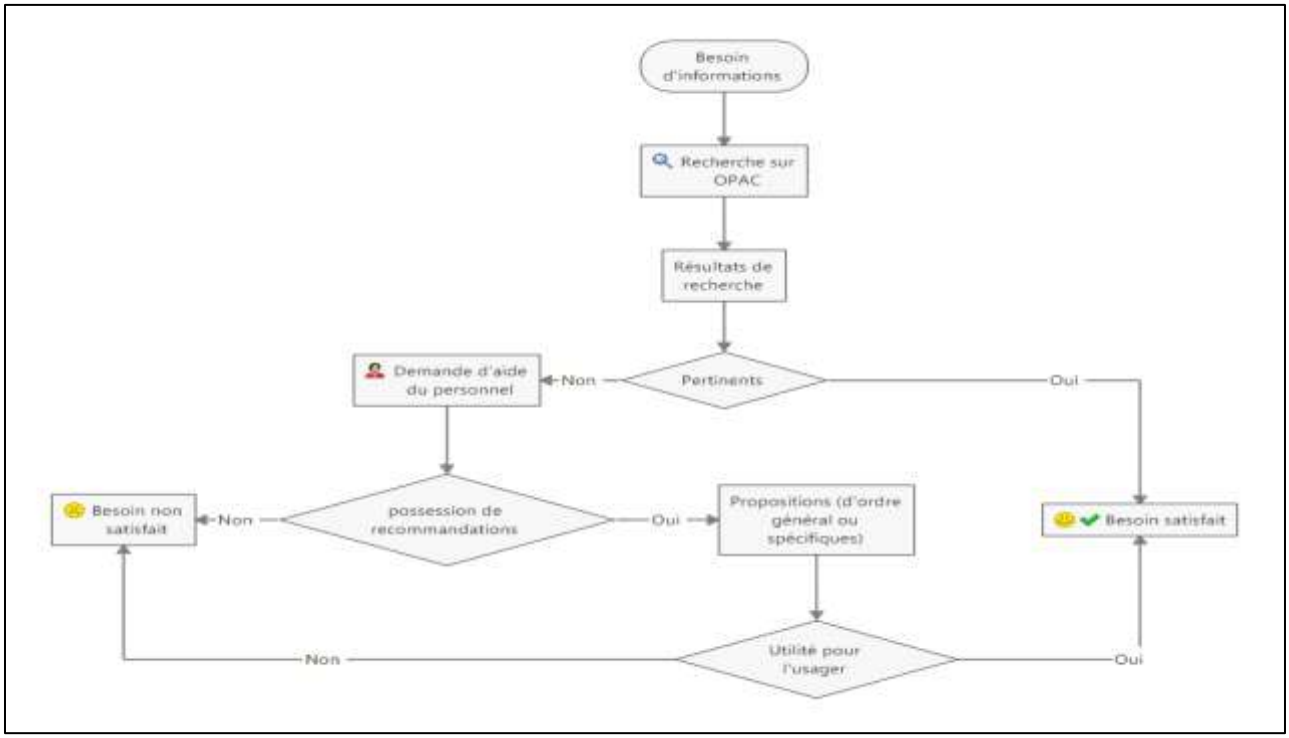

Figure 27 : Le processus de la recherche d'informations au sein de l'IMIST

La figure 27 relate le cheminement actuel d'un utilisateur exprimant un besoin d'information, son premier réflexe est d'effectuer une requête sur le catalogue en ligne (OPAC), les résultats de cette recherche dépendent de la requête et des mots clés employés. **Si l'utilisateur n'est pas satisfait par les résultats acheminés, il peut demander dans un deuxième temps l'aide du personnel du SID qui peut être fixé sur le sujet de la requête ou non et son aide peut s'avérer bénéfique ou inutile pour lui, pouvant ainsi générer une insatisfaction.**

Il importe de souligner que les utilisateurs ne sont pas toujours familiarisés avec la recherche documentaire et le temps consacré pour l'obtention d'un résultat satisfaisant importe pour une catégorie de chercheurs investis par ailleurs dans de lourdes démarches de recherche scientifique.

3. Le système intégré de gestion de bibliothèque au sein de l'IMIST

L'IMIST est doté du Système Intégré de Gestion de Bibliothèque (SIGB) open source baptisé « PMB » version 7.3.1, il s'agit d'un SIGB multilingue qui permet de gérer les activités nécessaires au fonctionnement d'une bibliothèque de manière informatisée, telles que la gestion des collections et des usagers, la circulation des documents, les acquisitions et l'édition de rapports statistiques, etc.

En effet, PMB prend en charge toutes les fonctionnalités standards d'un SIGB :

- La gestion des acquisitions : budgets, fournisseurs, commandes…
- Le catalogage et indexation : notices bibliographiques, notices autorités…
- Les périodiques : abonnements, bulletinage, dépouillement…
- La circulation : inscriptions, prêts…
- L'édition et statistiques : lettres de relance, statistiques de prêts…
- L'interface publique : Online Public Access Catalog (OPAC)…

Le système PMB rassemble trois composantes majeures :

- **Les interfaces :**
  - L'administrateur du système : Il permet de personnaliser et paramétrer le système ;
  - Le bibliothécaire : fonctions de catalogage, acquisition et toutes les fonctions de gestion interne ;
  - L'OPAC : Interface dédiée à l'utilisateur accessible via un navigateur, lui permettant d'effectuer ses requêtes, éventuellement réservations.
- **Le serveur :** qui fonctionne en tant que gestionnaire de transactions ;
- **La base de données :** qui regroupe les données sous une forme relationnelle (tables et champs) et se connecte au serveur pour héberger les informations.

Etant donné le scénario de recherche d'informations présenté ci-dessus et à travers nos réunions de préparation avec les responsables des services de l'IMIST qui seront en quêtés par la suite , nous soulignons plusieurs constats :

- L'orientation des professionnels de l'information de l'IMIST peut ne pas toujours correspondre au besoin des usagers ;
- La dépendance au personnel (sa disponibilité, ses informations et son savoir-faire) ;
- Une perte de temps lors de la recherche manuelle sur les rayonnages ;
- Un problème d'expression du besoin de manière intelligible par l'usager et sa compréhension par le professionnel ;
- Les recommandations sont d'ordre général, elles se limitent aux thématiques et leur repérage en termes de classes sur les rayonnages, elles se limitent aussi souvent à la référence la plus utilisée ou connue dans un domaine donné ;
- L'absence d'un suivi du feedback de l'usager à l'issue de la recommandation pour évaluer sa pertinence.

Au niveau du catalogue en ligne du SIGB, il est à noter :

- L'absence d'une fonction de recommandation ou de filtrage avancé d'information ;
- Les historiques de la session en cours du catalogue en ligne peuvent être investis dans un système de recommandation.

De surcroît, la recherche sur le catalogue en ligne du SIGB ne plaide pas en faveur d'une précision en termes des résultats de notices de documents retournées aux chercheurs, car dans les deux modes de recherche (simple ou avancée), l'utilisateur se trouve face à un flot important de notices, réparties en lots sur six à dix pages de résultats par analogie à un moteur de recherche. De fait, il s'agit d'un système de navigation organisé par pages, différent d'un scroll infini, reposant sur la présentation de l'ensemble du contenu sur une page unique, comme l'indique la figure 28.

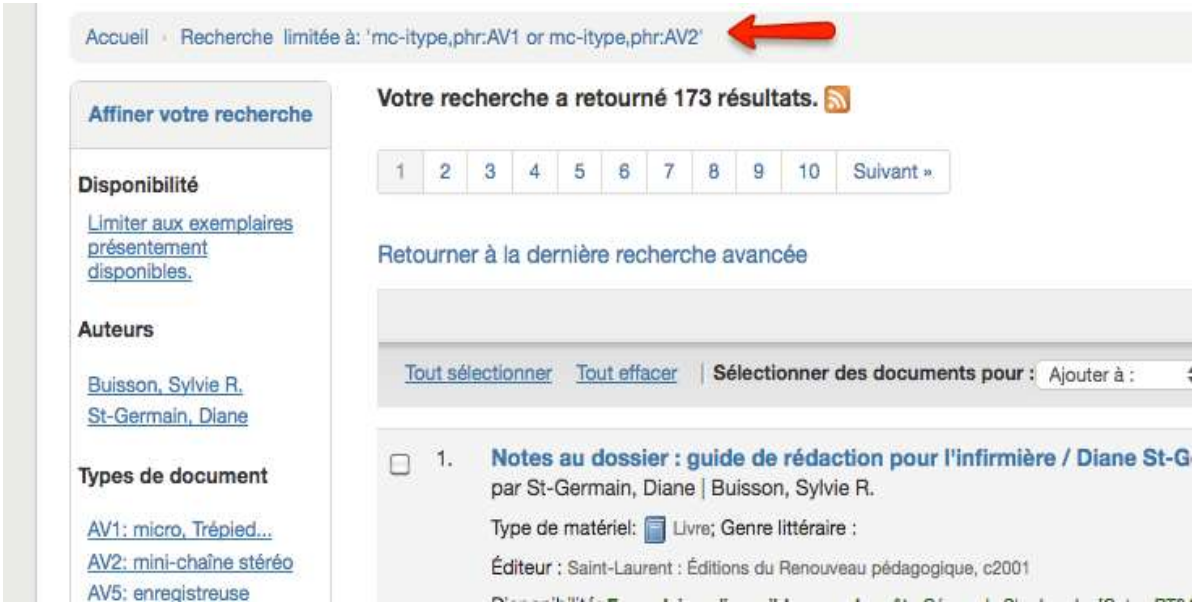

Figure 28: La page de recherche sur le catalogue en ligne du SIGB PMB, les résultats sont organisés par page

Le scénario présenté dans la figure 28, expose nos chercheurs de l'IMIST à une surcharge cognitive, qui ne facilite pas la décision concernant les documents pertinents à consulter, à visualiser la notice pour l'emprunter ou le réserver. Par conséquent, un temps important doit être investi par l'utilisateur et, par ailleurs, sans avoir la garantie de repérer le bon document, donc avec le risque d'un certain découragement, voire d'un abandon de la requête.

Au niveau de la politique d'acquisition :

La connaissance du suivi d'un certain nombre de recommandations pourrait signifier à l'IMIST les collections les plus sollicitées.

Conclusion

La recherche d'information scientifique et technique à l'IMIST s'organise par l'intermédiaire d'une offre informationnelle et documentaire destinée à des utilisateurs composés des chercheurs, des enseignants, des étudiants chercheurs et des professionnels du monde de l'entreprise. Cependant, l'orientation des utilisateurs est toujours pensée selon la logique d'un service de référence classique, visant à traduire les requêtes dans un langage documentaire, pour

repérer des documents pertinents. La seule personnalisation existante est celle de la diffusion sélective de l'information mais qui reste basée sur la définition d'un profil figé, qui n'évolue pas.

Dans notre thèse, nous cherchons à concevoir un modèle et un prototype de système de recommandation adaptif, favorisant l'accès et la découverte personnalisée dans un cadre universitaire où l'information demeure un atout majeur pour le développement de la production scientifique.

Chapitre 2 : La donnée et l'enjeu de sa valorisation

Introduction

Une analyse de l'existant est nécessaire pour nous permettre d'apprécier le positionnement de la data dans la chaine de fonctionnement actuel de l'IMIST et de recueillir les besoins portés par les professionnels de l'information et de la documentation afin d'améliorer l'offre des services. Nous avons ainsi recueilli la documentation disponible et conduit une petite enquête au sujet de l'exploitation des différentes données dans leurs services.

1. Exploitation actuelle de la donnée : une enquête générale sur l'usage des données à l'IMIST

- Les produits et les services de l'IMIST

La population cible de ce petit questionnaire est constituée des responsables de quatre services de l'IMIST, à savoir le service bibliothèque, le service chaine documentaire, le service veille et le service innovation[149]. La taille réduite de la population enquêtée s'explique par le fait que nous avons voulu cibler les responsables des 5 services de l'l'IMIST, quatre ont finalement collaboré pour répondre à notre questionnaire (voir Annexe 2 : questionnaire sur l'usage des données à l'IMIST).

De prime abord, les produits et les services ont fait l'objet d'une question ouverte auprès des enquêtés, ils ont mentionné différents services documentaires destinés aux utilisateurs et que nous résumons ci-dessous :

- Le prêt de documents ;

- La consultation de documents ;

- La recherche de références bibliographiques des Collections papier et numériques ;

- Les activités et les formations ;

[149] SENNOUNI, Amine. La Data au service de l'innovation dans les Services d'Information Documentaires (SID) universitaires nationaux. *Revue française des sciences de l'information et de la communication* [En ligne], 10 | 2016, mis en ligne le 01 janvier 2017. Disponible sur : http://rfsic.revues.org/2759 [consulté le 29 janvier 2017].

- L’accès aux collections papier et numérique ;
- L’accès au catalogue en ligne ;
- L’accès libre pour toutes les collections ;
- Le service WIFI disponible sur toutes les salles de lecture ;
- Le service questions/réponses ;
- La réservation des documents en ligne.

Les réponses mettent l’accent sur des produits et des services classiques offerts aux utilisateurs dans un SID, dans le cadre de la mission principale d’une bibliothèque, à savoir faciliter l’accès au savoir et à la connaissance à la population cible de l’IMIST.

Nous notons, dans ces réponses, l’absence de produits et de services d’information secondaire, rendant plus facilement repérable l’information primaire sur les rayonnages : il s’agit notamment des index, des bulletins de sommaires, des bulletins de résumés pour les revues, des guides d’orientation ou d’usage et des listes des nouvelles collections.

Nous relevons également l’absence d’une mention de personnalisation dans les services et produits fournis aux usagers, à l’image de la Diffusion Sélective de l’Information (DSI).

- L’accès au catalogue en ligne

Le catalogue documentaire central en ligne permet d’accéder aux collections d’ouvrages et de fouiller dans les ressources disponibles via l’interface utilisateur, composante d’une plateforme de gestion intégrée d’une bibliothèque, à savoir un catalogue en ligne public en l’occurrence PMB.

Le catalogue en ligne d’accès aux données bibliographiques de l’IMIST est hébergé sur le web, selon nos services enquêtés, et il est accessible depuis le navigateur de l’usager indépendamment du temps et de l’espace et sans les contraintes liées à un réseau local. Ainsi les services d’information documentaire ont plus de chances de recueillir des données des utilisateurs concernant leurs comportements informationnels.

- La provenance des données du SID

Les services cibles de notre enquête s’intéressent à produire et à extraire les données internes en priorité. Les professionnels de l’information et de la documentation de l’IMIST se limitent

aux données produites en interne, plus que celles externes représentées par des rapports et études gouvernementaux, livres blancs, données du prêt interbibliothèques, etc.

- Typologie des données internes et externes de l'IMIST

A la suite de la documentation recueillie et du questionnaire conduit, nous pouvons énumérer les types de données disponibles et citées dans l'enquête :

-Données sur les ressources documentaires (taux de rotation, taux de renouvellement des collections, âge des collections, découpage par classe, etc.) ;

-Données sur les ressources matérielles (superficie, rayonnages, les places assisses, les salles de lectures, les ordinateurs, les serveurs, les salles de traitement, le dispositif antivol, le matériel des côtes, etc.) ;

-Données sur les produits et services documentaires (le catalogue en ligne, les bulletins de sommaires et de résumés pour les revues, les dossiers documentaires thématiques, le matériel de reproduction, la segmentation en groupes de Diffusion Sélective d'Information, etc.) ;

-Données sur les conventions et partenariats avec des organismes du même secteur (le prêt interbibliothèques, protocoles, etc.) ;

-Données structurées et non structurées sur la liste des documents disponibles dans le SID ;

- Données sur les usagers (affiliations, sexe, historiques des prêts, etc.) ;

- Données bibliographiques des notices bibliographiques (auteur, titre, descripteurs, classification, etc.).

Les services enquêtés produisent régulièrement des données en interne, comme ils les reçoivent de provenance externe dans le cadre de leurs activités quotidiennes couvrant toute la chaine documentaire (l'acquisition, le traitement et la diffusion).

Pour finir sur ce point, nous notons la quasi-absence des données d'usages ou de données sur les interactions des utilisateurs avec le système, alors qu'elles devraient occuper une place

centrale dans un service d'information documentaire visant prioritairement la satisfaction de ses utilisateurs.

2. Vers une personnalisation des catalogues en utilisant les données

Trois services enquêtés sur quatre considèrent comme important, voire nécessaire de disposer d'un système de recommandation, en vue de faciliter l'accès aux ressources documentaires de manière plus personnalisée à travers l'exploitation des données portant sur les collections, mais également sur les statistiques et les données d'usage disponibles, au cœur de l'adaptation de l'offre documentaire aux chercheurs.

Selon cette logique d'exploitation de la data dans des unités documentaires de recherche, certaines bibliothèques nationales et universitaires en Europe ont déjà développé cet usage de la data pour la personnalisation des services documentaires et culturels. A titre indicatif, la Bibliothèque Nationale de France a lancé un projet de service de fourniture de corpus numériques à destination de la recherche, dit « Corpus », initié en 2016.

Ce projet a pour dessein de fournir des corpus personnalisés aux chercheurs, au niveau individuel ou d'un groupe de recherche avec le déploiement des outils du Texte et Data Mining (TDM), dans l'intention d'analyser les collections en profondeur. Il s'agit dans ce cas d'accompagner l'utilisateur dans la prédéfinition d'un corpus personnalisé et non pas de raffiner sa requête d'interrogation.

Des bibliothèques universitaires ont mobilisé des outils de TDM et des technologies numériques, comme c'est le cas de l'université de Leyde, d'Oxford ou encore Pittsburgh, Stanford et Columbia aux États-Unis et en France à l'Université de Bordeaux Montaigne, où la personnalisation et la structuration de corpus numérisés ont pris la forme de mise à disposition pour les chercheurs en géographie de corpus de cartes numérisées en lien avec leurs recherches[150].

[150] DELESPIERRE, Louis. *Les services personnalisés aux publics en bibliothèque universitaire, une exigence d'innovation et de transformation : l'exemple des services aux chercheurs*. Mémoire d'étude. ENSSIB, Mars 2019.

En guise de conclusion, les résultats de cette petite enquête au sein de l'IMIST a mis en relief l'intérêt du recours à la data, pour améliorer l'offre orientée vers l'utilisateur final, voire l'affiner et la personnaliser.

3. Les besoins formulés

Le questionnaire conduit ainsi que les échanges menés avec certains professionnels de l'IMIST ont permis de cerner les attentes et les besoins prioritaires quant aux services documentaires :

- Permettre aux utilisateurs de s'authentifier ;
- Faciliter pour les utilisateurs l'accès aux ressources documentaires ;
- Proposer à un utilisateur actif des recommandations de documents qu'il juge pertinents par rapport à ses besoins ;
- Faire un suivi de l'évolution du besoin des utilisateurs ;
- Avoir une vue globale sur les collections consultées, afin d'orienter la politique d'acquisition.

4. L'objectif de la solution envisagée

Nous ne pourrons pas répondre à tous ces besoins formulés. Dans cette thèse, nous avons considéré en priorité les besoins concernant la qualité des réponses à apporter aux requêtes des usagers.

Les systèmes de recommandation pouvaient ainsi accompagner cet objectif en aidant l'utilisateur à découvrir des documents auxquels il n'aurait pas songé.

Notre projet a donc été de développer un module de recommandations à base de jeux de données. Ce module une fois testé, pourrait intégrer par la suite le SID.

Le modèle proposé consiste à suivre les traces des utilisateurs, pour proposer des contenus qui peuvent être pertinents, il aura pour rôle :

- D'analyser l'activité et les habitudes de navigation des usagers ;
- De proposer des recommandations pertinentes aux usagers.

Etant basé sur le filtrage collaboratif comme approche de recommandation, le modèle va permettre :

- De faire profiter aux usagers les préférences des autres utilisateurs ;
- De favoriser le transfert des connaissances ;
- De faciliter la sélection et le filtrage de l'information.

5. Vers la modélisation d'un système de recommandation appliqué aux services d'information documentaire : constats et orientations

Dans notre thèse, nous aspirons concevoir un modèle de système de recommandation favorisant l'optimisation de la sélection et l'accès des usagers chercheurs au Maroc aux ressources documentaires, représentant un instrument pour la production scientifique. De ce fait, l'objectif est double :

- Améliorer la satisfaction de l'utilisateur, en lui proposant des contenus en adéquation avec ses intérêts ;
- Satisfaire un plus grand nombre d'utilisateurs, en élargissant le champ des intérêts couverts par les recommandations, et ce afin d'augmenter les chances d'obtenir au moins une recommandation pertinente pour chaque utilisateur.

Notre proposition de modèle repose sur trois constats, résultant de notre enquête à l'IMIST, et s'appuyant sur les concepts, approches et mécanismes d'évaluation investigués dans le chapitre relatif aux systèmes de recommandation :

Constat 1 : les différentes requêtes exécutées sur le catalogue en ligne des services d'information documentaire requièrent souvent du temps, pour vérifier et feuilleter de manière linéaire les pages des résultats. Ceci ne conduit pas forcément à des résultats pertinents, les

modèles classiques de recherche d'information ne tentent pas de trouver une similarité parfaite entre la représentation des requêtes et celle des documents, tout en tenant compte du profil de l'utilisateur ;

Constat 2 : un modèle de recommandation appliqué aux SID peut reposer sur la catégorisation des utilisateurs par groupes homogènes partageant des centres d'intérêt de recherche similaires, observables à travers leurs logs d'usage. En pratique, cela se traduit par une transformation de ces données *log* implicites en données explicites matérialisées par des notations. Puis, il pourra être calculé une recommandation basée sur la factorisation matricielle, et tenant compte de deux cas de figure : l'utilisateur anonyme et l'utilisateur authentifié ;

Constat 3 : le système de recommandation sert à améliorer la prise de décision en termes de sélection des documents pertinents à l'issue de leurs requêtes. Il aidera à repérer des documents en un temps réduit, pour en faire la matière première de leurs démarches de recherche scientifique.

Nous comparons dans le tableau 14 le modèle de recommandation que nous proposons à celui d'autres services que nous avons déjà évoqués précédemment en fonction de plusieurs critères qui permettent d'identifier la nature du SID, le type de données utilisées dans un système de recommandation, l'approche algorithmique suivie et le type de l'utilisateur auquel la recommandation est proposée, en vue de comparer ces exemples avec notre modèle de recommandation proposé :

Critères / Exemples de solutions de SR en Bibliothèques numériques	Bibliothèque physique	Bibliothèque numérique	Données explicites	Données implicites	Recommandation sur l'OPAC	Usager authentifié	Usager anonyme	Filtrage collaboratif à base de modèle	Filtrage collaboratif à base de mémoire
ACM Digital Library	✗	✓	✓	✗	✗	✗	✓	✗	✓
Book Physic et Huddersfield Book Recommender system	✗	✓	✓	✗	✗	✗	✓	✗	✓
Web service « Amazon Personalize »	✓	✗	✓	✗	✓	✓	✗	✗	✓
Notre modèle	✓	✗	✓	✓	✓	✓	✓	✓	✗

Tableau 13 : Comparaison des systèmes de recommandations de certaines bibliothèques numériques avec le prototype envisagé

Il faut noter d'emblée que la matrice comparative ci-dessus a pour objectif de dresser un positionnement de notre modèle de recommandation par rapport à d'autres systèmes de recommandation greffés sur des bibliothèques numériques et bases de données et n'aborde pas la performance à travers des valeurs de jugement supérieures ou inférieures.

Le modèle de recommandation que nous envisageons se distingue sur plusieurs critères des autres exemples :

- Il s'appuie sur **une préparation affinée des données d'entrée** de notre système de recommandation, où les enregistrements des données *log* ont été nettoyés, convertis en scores pour chaque type de données, puis en une moyenne globale révélant l'intérêt de l'utilisateur pour un type de documents, pour finalement aboutir à un fichier final, contenant un jeu de données de qualité homogène ;
- Il implémente **une approche de filtrage collaboratif à base de modèle**, qui consiste **à prétraiter et entrainer un modèle, considérant les profils des utilisateurs, et le solliciter en hors ligne, pour procéder à la prédiction, sans surcharge de la plateforme ou du catalogue en ligne.** Cette approche contraste avec l'approche de filtrage collaboratif à base de mémoire, qui requiert d'être associée à une base de données contenant les notations des utilisateurs sur les contenus et stockée dans la mémoire du serveur, pour être sollicitée en ligne au moment de la génération des recommandations ;

- Il est développé **sur la base des logs d'usage des utilisateurs, représentant des données implicites** plutôt qu'à partir de leurs appréciations explicites, méthode utilisée dans la majorité des systèmes de recommandation déployés pour les industries culturelles et de contenu et les services d'information documentaire ;
- Il propose des recommandations **à deux types de profil, à savoir un usager authentifié disposant déjà d'un compte et un usager anonyme qui vient d'interroger le système pour la première fois** ;
- Il est destiné à être intégré au **catalogue en ligne** du service d'information documentaire, dans la mesure où la recherche et la navigation sont optimisés par le biais du système de recommandation sans créer une plateforme additionnelle ;
- Ce modèle de recommandation est conçu pour **une collection physique de ressources documentaires composée d'ouvrages**, qui sont disponibles et consultables sous leur format papier depuis les rayonnages d'une unité documentaire, et non pas comme les systèmes de recommandation appliqués aux ressources numériques accessibles en texte intégral.

Conclusion

L'IMIST ne dispose pas d'un système de recommandation. L'orientation des utilisateurs demeure classique, le professionnel de l'information établit un échange avec l'utilisateur sur son équation de recherche, qui risque de ne pas être satisfaite, si le professionnel n'arrive pas à l'interpréter correctement ou de manière adaptée au besoin de l'utilisateur.

Nous avons diagnostiqué l'exploitation des données à l'IMIST, en nous basant sur la documentation, sur nos observations et sur l'enquête conduite auprès de quatre professionnels de l'information et nous avons constaté une sous-exploitation des données disponibles dans une optique de personnalisation pour les utilisateurs.

Le chapitre qui suit sera consacré à la conception du système de recommandation que nous proposons, pour répondre à ces besoins.

Chapitre 3 : Conception et modélisation

Introduction

Après l'étude de l'existant, nous allons détailler la conception du système de recommandation dans une structure documentaire de recherche. Nous allons concevoir un modèle de recommandation basé sur les données implicites des utilisateurs.

Dans un premier temps, nous préciserons le corpus de l'IMIST disponible, avant d'exploiter les données implicites des utilisateurs et les transformer en données explicites pour construire le prototype envisagé.

1. Le corpus de l'IMIST : les utilisateurs et leurs données

Après avoir présenté les missions et les services de l'IMIST dans le chapitre précédent, nous allons, à ce stade, examiner le corpus documentaire faisant l'objet de notre étude, à savoir les données de consultation du catalogue en ligne de l'Institut par 500 utilisateurs adhérents aux services de de la structure documentaire durant l'année 2015.

Le choix de cet échantillon s'est fait sur la base des utilisateurs les plus actifs en matière de consultation du catalogue en ligne au titre de l'année 2015 et pour lesquels des données sont disponibles. Cette population est aussi représentative de l'ensemble des adhérents de la structure, puisqu'ils représentent 45% des chercheurs à l'IMIST (professeurs, chercheurs, doctorants, etc.) au titre de l'année 2015.

Les données collectées sur les utilisateurs ont été obtenues à partir des statistiques de l'OPAC accessibles en backoffice de l'interface administration, qui permet d'interroger toutes les données issues de la navigation des utilisateurs sur le catalogue en ligne.

Les statistiques de l'OPAC s'obtiennent à partir de requêtes MySQL associées à la vue. Elles permettent de rendre compte de l'historique de l'interaction de l'usager avec le catalogue en ligne. Le service du système d'information de l'IMIST a confectionné, dans le cadre des rapports de suivi des activités et transactions des utilisateurs avec le catalogue en ligne, une

recension sous forme d'un fichier Excel construit sur la base de statistiques exécutées par l'administrateur[151] :

#mots_saisis();	Retourne l'expression saisie par l'usager	Texte
#url_ori();	Retourne l'adresse internet (url) de provenance de l'usager	Texte ou Texte large
#url_asked();	Retourne l'adresse internet (url) des pages demandées en consultation	Texte ou Texte large
#num_session();	Retourne le numéro de session	Texte
#login();	Retourne le login du lecteur	Texte
#adresse_ip();	Retourne l'adresse IP de l'usager	Texte
#timestamp();	Retourne la date et l'heure au format AAAA-MM-JJ HH:MM:SS	Date/Heure
#date();	Retourne la date	Date
#empr_categ();	Retourne la catégorie de l'emprunteur	Texte
#empr_statut();	Retourne le statut de l'emprunteur	Texte
#nb_total();	Nombre de résultats total	Entier

Tableau 14 : Les statistiques à exécuter pour relever les actions des utilisateurs sur l'OPAC du SIGB PMB (https://doc.sigb.net/pmb34/co/edit_stat_opac.html)

Les statistiques exécutées ci-dessus ont aidé à construire la base de connaissances, qui comporte des données relatives aux paramètres suivants : derniers documents consultés, les classes thématiques CDD correspondantes pour un utilisateur, le temps cumulé passé sur l'OPAC, le taux de rebond de la requête et finalement le nombre de fois que l'utilisateur a affiché la notice d'un document à partir des résultats obtenus.

[151] Documentation PMB [En ligne]. Disponible sur : https://doc.sigb.net/pmb34/co/edit_stat_opac.html [Consulté le 19 Septembre 2020].

Voici un exemple relatif à l'exécution d'une vue sur les statistiques de visite[152] :

Figure 29: Exemple de l'exécution de la statistique relative aux nombres de visites sur l'OPAC du SIGB PMB (https://doc.sigb.net/pmb34/co/edit_stat_opac.html)

Le choix de ces données d'usage extraites des fichiers log était contraint par le fait que l'IMIST ne dispose pas d'autres sources de données sur l'exploitation des ressources documentaires et informationnelles que ces fichiers issus du back office du catalogue en ligne. Il est à signaler que ces données ne pouvaient pas être exploitées directement sur notre solution, car elles sont implicites et requièrent une conversion en notations explicites.

[152] Ibid., p.170.

La figure 30 présente les valeurs relatives aux paramètres retenus pour une partie de l'échantillon :

ID utilisateur	Dernier Document consulté (références)	Classes documents empruntés	Les rebonds de requête	Affichage de docs dans les résultats	Le temps passé sur l'OPAC
1	Fouché, Pascal , Péchoin, Daniel , Schuwer, Philippe .Dictionnaire encyclopédique du livre. [Paris] : Éd. du Cercle de la librairie, impr. 2011	Classe 600	2 fois	8 fois	10 Min
2	Gaussier, Éric , Yvon, François .Modèles statistiques pour l'accès à l'information textuelle. Paris : Hermès science publications-Lavoisier, impr. 2011	Classe 200	0 fois	3 fois	6 Min
3	ISAAC, Henri , Sid Ahmed, BENRAOUANE .Guide pratique du e-Learning. : DUNOD, 2011	Classe 000	2 fois	9 fois	34 Min
4	Fijalkow, Jacques , Fijalkow, Éliane .La lecture. Paris : le Cavalier bleu éd., impr. 2010	Classe 300	3 fois	12 fois	25 Min
5	Poullet, Yves , Henrotte, Jean-François .Droit des technologies de l'information et de la communication 2011. Bruxelles : Larcier, DL 2011	Classe 500	1 fois	5 fois	14 Min

Figure 30: Un extrait du fichier dérivé des données log implicites relatives aux préférences des 500 utilisateurs de l'OPAC de l'IMIST au titre de l'année 2015

Les identifiants des classes CDD qui figurent sur les données *log* d'interaction des utilisateurs avec le catalogue en ligne renvoient aux disciplines et spécialités suivantes :

L'identifiant de la classe CDD	La classe CDD
000	Informatique, information, ouvrages généraux
100	Philosophie, Parapsychologie et Occultisme, Psychologie
200	Religions
300	Sciences sociales
400	Langues
500	Sciences de la nature et Mathématiques
600	Technologie (Sciences appliquées)
700	Arts, Loisirs et Sports
800	Littérature (Belles-Lettres) et techniques d'écriture
900	Géographie, Histoire et disciplines auxiliaires

Tableau 15 : Les dix classes de la Classification Décimale de Dewey

La figure 30 indique que les utilisateurs identifiés par un numéro ont tous **une classe thématique principale de la Classification Décimale Dewey (CDD), où s'inscrivent les ressources documentaires susceptibles de les intéresser, tout en notant le taux de rebond des requêtes, le nombre de fois que la notice détaillée d'un document a été affichée et le**

temps passé par l'utilisateur sur l'OPAC au titre des sessions cumulées des utilisateurs du 1er Janvier au 31 Décembre 2015.

Toutefois, nous notons que les utilisateurs chercheurs présentés selon leur appartenance à une classe disciplinaire de la CDD ont un statut de chercheur spécialiste d'une discipline donnée, **leurs recherches devraient s'inscrire dans une classe CDD bien précise, mais cela n'empêche pas que les résultats retournés par le catalogue en ligne peuvent concerner, en fonction de la formulation de la requête, des documents pluridisciplinaires.**

A ce niveau, nous remarquons que les données obtenues sont hétérogènes et que les préférences et les détails liés à l'interaction d'un utilisateur ou d'un groupe d'utilisateurs sur le catalogue en ligne sont différents.

2. La classification supervisée du corpus

La classification supervisée renvoie à une méthode statistique d'analyse des données. Elle vise à produire automatiquement des règles à partir d'une base de données d'apprentissage contenant des cas déjà traités et validés[153].

Cette méthode est utilisée, en vue de classifier l'échantillon des utilisateurs du catalogue en ligne de l'IMIST, pour que la recommandation tienne compte en priorité du profil de l'utilisateur.

En effet, notre classification supervisée fait appel à l'algorithme d'apprentissage supervisé K-Nearest Neighbors (K-NN ou KNN) ou les K plus proches voisins, dont le choix est lié au contexte des données déjà étiquetées représentées par des utilisateurs associés à des classes disciplinaires de la CDD. Ainsi, la classification *k*-NN correspond à une classe d'appartenance, de manière à ce qu'un objet d'entrée (un utilisateur) soit classifié selon le résultat majoritaire des classes d'appartenance de ses k plus proches voisins (les classes disciplinaires CDD).

En pratique, pour un élément X à classer, il est nécessaire, dans notre cas, de calculer la distance entre chacun des éléments x' et X, en retenant les K plus proches éléments/voisins, laquelle

[153] CHRISTOS, Bouras et Vassilis, TSOGKAS. A clustering technique for news articles using WordNet. *Knowledge based-systems*, 2012, vol. 36, p.115-128.

distance est représentée par la proximité en matière de la classe de la Classification Décimale de Dewey (CDD) traduisant l'intérêt scientifique exprimé par chaque utilisateur.

De même, l'algorithme du KNN ne prévoit pas une différence entre les observations d'apprentissage et les nouvelles, du moment où l'influence est identique pour chacun des voisins, indépendamment de leur degré de similarité.

Dans le contexte de l'utilisation du catalogue en ligne de l'IMIST, la classification concerne les données relatives aux domaines/disciplines de spécialité de notre communauté de 500 utilisateurs chercheurs regroupés par classe disciplinaire de Dewey selon des proportions différentes, comme en rend compte la figure 31 :

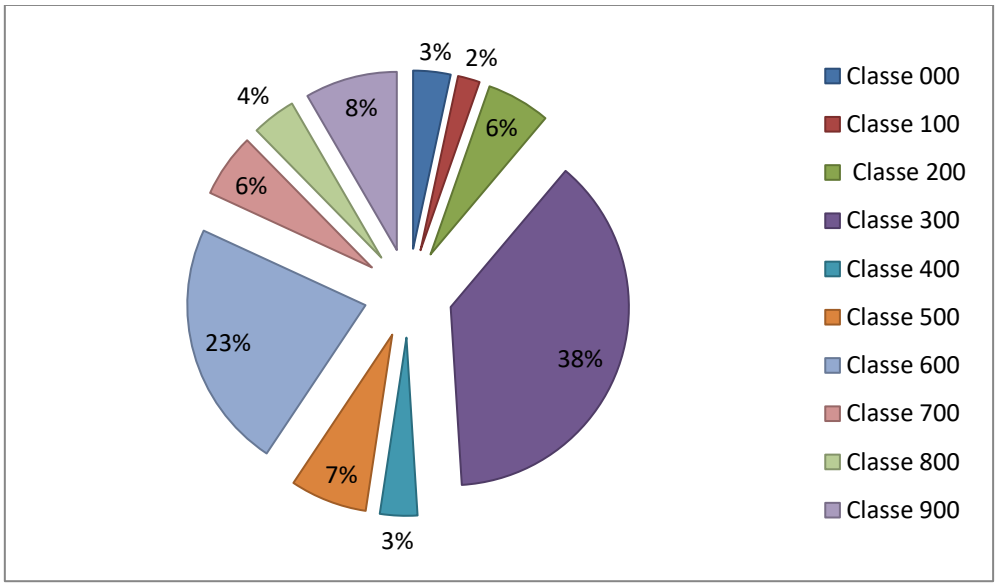

Figure 31 : Le nombre d'utilisateurs intéressés selon les classes CDD sur l'OPAC IMIST

Les domaines ou disciplines intéressant le plus l'échantillon des utilisateurs du catalogue en ligne de l'IMIST au titre de l'année 2015, sont les classes disciplinaires 300 traitant des sciences sociales et 600 couvrant les technologies et les sciences appliquées avec un pourcentage de 38% et 23% respectivement, soit 61% de notre échantillon des utilisateurs chercheurs de l'IMIST. Tandis que, les autres classes de cette classification encyclopédique, au nombre de 8, se partagent le reste des utilisateurs. La figure 30 mentionne donc 10 catégories principales d'utilisateurs.

Dans notre thèse, nous avons opté pour un type de données nominal représenté par des noms ou identifiants des classes selon la CDD, pour obtenir dix groupes de spécialité distinctes, comme nous pouvons le constater dans le tableau 16 :

Classe thématique CDD	Nombre d'utilisateurs intéressés
Classe 000	17 utilisateurs
Classe 100	10 utilisateurs
Classe 200	29 utilisateurs
Classe 300	190 utilisateurs
Classe 400	17 utilisateurs
Classe 500	35 utilisateurs
Classe 600	113 utilisateurs
Classe 700	29 utilisateurs
Classe 800	20 utilisateurs
Classe 900	40 utilisateurs
Total	**500 utilisateurs**

Tableau 16: La répartition de l'échantillon des utilisateurs de l'OPAC de l'IMIST par classe de la CDD

Par classification supervisée appliquée à notre population cible, nous obtenons les groupes des utilisateurs partageant le même centre d'intérêt scientifique, qui se connectent principalement sur le catalogue en ligne pour chercher dans une classe disciplinaire plus que dans d'autres.

Chaque utilisateur appartient à une classe, mais sans pour autant rejeter les croisements possibles entre sa spécialité et d'autres disciplines, d'où l'importance de la conception d'un modèle de recommandation, tenant compte de ce type de documents à prédire et à lui proposer.

Cette méthode diffère de celles qui donnent plutôt l'importance aux appréciations formulées par les utilisateurs par le biais de notations des documents sur une échelle donnée.

3. La « non pertinence » des résultats sur le catalogue en ligne

Le fichier log ou de journalisation a permis de tenir compte de deux paramètres de données relatifs **à l'affichage des différents résultats obtenus à l'issue d'une requête et le temps passé sur le catalogue pour défiler les résultats et en consulter certains.**

Dans l'échantillon de 500 utilisateurs examiné durant l'année 2015, nous avons décelé les extrémités pour les deux paramètres, comme le montre le tableau 17 :

ID utilisateur	Affichage de docs dans les résultats	Le temps passé sur l'OPAC
446	99 fois	102 min
3	10 fois	09Min

Tableau 17: Les extrémités d'affichage des résultats et le temps passé sur l'OPAC

Le tableau montre que le basculement de la notice d'un document vers une autre varie **de 10 à 99 fois**, tandis que le temps passé sur le catalogue, en vue de visualiser ces résultats oscille entre **9 et 102 minute**s.

Les deux extrémités maximales de ces deux paramètres sont élevées, traduisant une consultation de plusieurs notices de documents, ce qui la rend relativement longue.

En conclusion, la surcharge informationnelle est perceptible dans le temps mesuré pour accéder aux documents souhaités, d'où l'intérêt d'un modèle de recommandation, qui devrait limiter le risque que l'utilisateur soit désorienté quant au choix des ressources retournées dans les résultats.

4. Vers un modèle de système de recommandation appliqué aux services d'information documentaires (SID)

Nous avons vu précédemment que les modèles classiques de recherche de documents dans les services documentaires peuvent engendrer une surcharge informationnelle souvent non pertinente qui se traduit par des temps élevés de consultation et un certain nombre de rebonds lors de la requête.

Comme nous l'avons déjà indiqué dans la revue de littérature sur les systèmes de recommandation. Ces derniers s'alimentent de deux types de données :

- Données explicites : qui représentent les notations que les utilisateurs ont attribuées aux items, sur une échelle allant de 1 à 10 très souvent, celles-ci sont directement exploitables par le système de recommandation ;

- Données implicites : qui caractérisent le comportement de navigation de l'utilisateur et son interaction avec le système (requêtes effectuées, durée de la session, etc.).

L'implémentation de notre modèle repose sur l'approche de filtrage collaboratif passif à base de modèle, qui préconise l'exploitation des données implicites cumulées par les utilisateurs lors de leurs interactions avec le système. Ainsi, les utilisateurs proches voisins sont identifiés pour chaque utilisateur et organisés sous forme de groupes homogènes obtenus à l'issue de la classification supervisée opérée.

A cet effet, les recommandations collaboratives sont basées sur la qualité des documents les plus fréquentés dans une classe thématique Dewey par les utilisateurs appartenant à un même groupe, c'est-à-dire à un même domaine ou champs de recherche.

Le modèle que nous proposons, repose sur le prototypage d'un système de recommandation, qui part des centres d'intérêt communs en termes de type de documents recherché (classes de Dewey), pour générer des groupes d'utilisateurs (catégories par affinités communes et obtenues à l'issue de l'opération de la classification supervisée).

De fait, les utilisateurs d'une même catégorie devraient se voir proposer des recommandations de documents qu'ils n'ont pas consultés auparavant sur la base de l'approche de filtrage collaboratif à base de modèle, mais toujours dans le cadre de leur domaine de recherche de spécialisation. De cette manière, nous essayons de limiter à la fois le taux de rebond de requête et le temps d'une requête sur le catalogue en ligne.

La figure 32 propose une vue globale du modèle proposé :

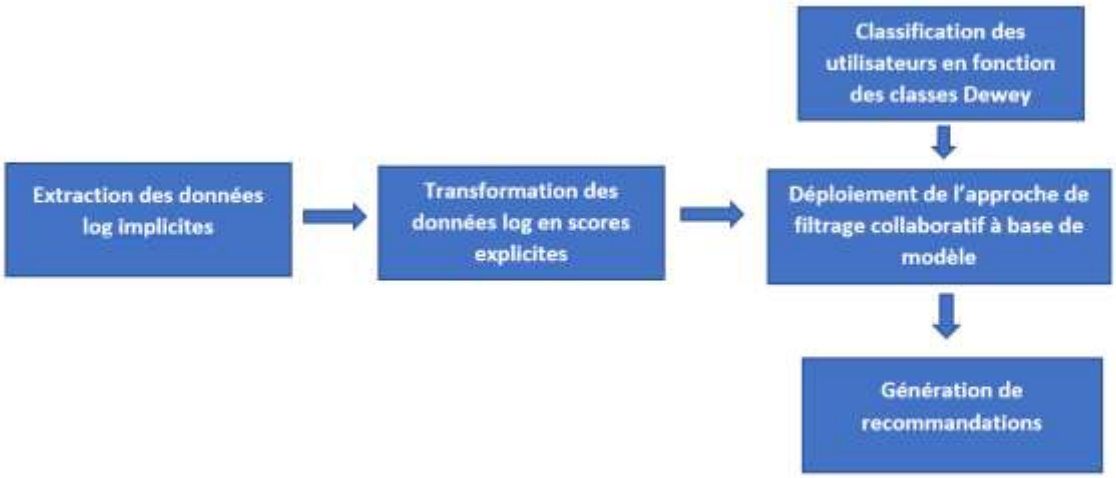

Figure 32: Une vue globale du modèle de recommandation proposé

Nous allons faire appel à **la classification supervisée** comme étant l'une des techniques d'apprentissage automatique, qui repose sur la structuration d'un ensemble d'objets/ d'items en différents groupes ayant des ressemblances et des affinités communes. Ainsi, les utilisateurs qui sont considérés comme similaires (similarité intra-classe) font partie de la même catégorie, alors que ceux qui sont dissimilaires (dissimilarité inter-classe) sont liés à des groupes différents.

Pour obtenir les recommandations, le système doit d'abord établir des groupes d'utilisateurs sur la base des classes CDD des documents consultés par les utilisateurs identifiés, pour mettre en œuvre **une corrélation entre les utilisateurs chercheurs qui s'intéressent aux ressources documentaires et un champ disciplinaire principal**, représenté par une grande classe au niveau du schéma classificatoire de Dewey et ceci en considérant la notice du dernier document visualisé, qui permet d'identifier la classe disciplinaire à laquelle appartient l'utilisateur. Les responsables de l'IMIST nous ont indiqué que le dernier document visualisé a satisfait l'utilisateur et qu'il traduit donc son affinité à la classe CDD de ce document.

La recommandation se fera, dans notre modèle, en proposant d'autres documents de la même classe consulté par d'autres utilisateurs.

Deux cas de figure seront à distinguer : **l'utilisateur authentifié et l'utilisateur anonyme.** Dans le premier cas un modèle est créé à l'utilisateur sur la base son profil, alors que dans le second cas, l'utilisateur aura droit à des recommandations, en fonction des requêtes qu'il est en train d'effectuer.

La corrélation conditionne la prédiction, qui consiste à son tour à prévoir les éventuels documents appartenant à la même classe thématique de l'usager, mais qu'il n'a pas encore consultés. En d'autres termes, l'obtention de la liste de recommandations est un résultat direct de cette confrontation entre la matrice items et la matrice des utilisateurs, faisant partie d'un même groupe.

La recommandation d'une liste de documents aux utilisateurs fait appel, selon notre modèle, aux données contenues dans les fichiers logs de consultation du catalogue en ligne, en commençant par le dernier document consulté, en considérant le temps cumulé passé sur l'OPAC, le taux de rebond de requête et le nombre de fois que l'usager a affiché la notice d'un document donné. Ces données implicites correspondent aux paramètres suivants :

- **Le dernier document consulté :** la notice du dernier document consulté et la classe disciplinaire de l'utilisateur est identifiée ;
- **Le nombre de rebonds de la requête :** le nombre de requêtes effectuées et dont les résultats n'ont pas été visualisés ou consultés et ayant conduit à des reformulations de la requête initiale ;
- **L'affichage du document dans les résultats :** les notices des documents visualisées dans les résultats retournés à l'utilisateur suite à ses requêtes ;
- **Le temps passé sur l'OPAC :** c'est la durée de la session de l'utilisateur.

Par la suite, **nous allons procéder à une conversion des données implicites en notes exprimant l'intérêt des utilisateurs** et ceci en attribuant une pondération à chaque paramètre des données recueillies, excepté les données relatives aux classes CDD dans lesquelles s'inscrivent les documents. En fait, cette pondération devrait dépendre du rapport étroit qu'entretient le paramètre en question avec l'intérêt documentaire du chercheur et sa préférence. Par exemple : le paramètre de l'affichage de la notice d'un document donné révèle davantage les préférences de l'utilisateur que le paramètre relatif au temps passé sur l'OPAC.

Par ailleurs, pour définir les plages de notes, **nous devrons diviser les données en intervalles de même amplitude**. Ainsi, **les différentes valeurs des quatre paramètres seront organisées**

suivant une échelle, qui correspondront à de nouvelles valeurs rendant les données explicites et exploitables et desquelles une moyenne globale par utilisateur sera calculée à partir des valeurs numériques définies pour les différents paramètres. La note générale représentera la donnée d'entrée de notre modèle de recommandation.

Il est à souligner que les trois paramètres qui seront utilisés pour calculer la note globale de l'utilisateur sont : le nombre d'affichages de notices des documents, le temps passé sur l'OPAC et le taux de rebond de la requête, alors que le dernier document consulté nous sera utile pour identifier la classe disciplinaire de l'utilisateur et constituer la base des documents qui seront objet de la recommandation.

La génération de la recommandation devrait se faire sur la base **des algorithmes de l'apprentissage automatique** et ceci dans deux cas de figure, d'abord, celui d'un utilisateur authentifié sur le catalogue en ligne, et pour lequel il suffit **d'exploiter son modèle** et ensuite celui anonyme pour lequel il s'agit de recommander les documents similaires à ceux qu'il est en train de consulter.

Concrètement, nous allons, dans notre modèle, faire appel à la factorisation matricielle, qui consiste à décomposer une matrice en plusieurs autres matrices. La matrice originale est le produit de ces matrices. En effet, dans sa forme triviale, la décomposition d'une matrice permet d'en obtenir deux, l'une reliée aux documents et l'autre associée aux utilisateurs représentés par leurs vecteurs correspondants.

Les notations des utilisateurs aux documents forment une matrice, composée par les vecteurs d'utilisateurs et des documents et appelée matrice de notation, vue que chaque utilisateur ne peut pas noter tous les documents, cette matrice reste alors creuse. La factorisation matricielle a pour objectif de réduire la dimension de la matrice et diminuer de la densité des données, pour offrir une meilleure performance au système de recommandation.

Pour procéder à l'implémentation de notre système de recommandation sur la base du modèle conçu, il a été nécessaire de comparer les grandes plateformes d'apprentissage automatique (Machine Learning). Il convient de souligner que la recommandation dans les deux volets relatifs à la corrélation et à la prédiction peut s'opérer par l'entremise d'un tableur et d'un langage de programmation.

En revanche, ils demeurent insuffisants, dans notre cas, qui requiert des traitements distribués dans la mesure où le nombre des utilisateurs augmente de manière continue, ainsi que les

documents et les notices, même si le prototype de cette thèse ne couvre qu'un petit échantillon de 500 utilisateurs. La recommandation devra se faire en temps réel et avec un jeu de données et des utilisateurs en évolution quotidienne (ressources documentaires et utilisateurs).

Le tableau 18 compare les caractéristiques majeures de deux librairies libres, Mahout et Spark, prenant en charge des algorithmes d'apprentissage automatique et de recommandation.

Critère de comparaison	**Spark**	**Mahout**
Fondateur	Spark est un projet de la fondation Apache (Open source) dédié au traitement Big data.	Apache Mahout est un projet open source de la fondation Apache dédié au traitement Big data.
Langages de fonctionnement	Java, Scala, Python.	Java.
Utilisation	La classification, la régression, le clustering, le filtrage collaboratif et la réduction de dimensions.	Le clustering et le filtrage collaboratif.
Algorithmes déployés dans les libraries	• Classification :logistic regression, naive Bayes, etc. • Regression: generalized linear regression, survival regression etc.; • Decision trees, random forests, and gradient-boosted trees; • **Recommendation: Alternating Least Squares (ALS);** • Clustering: K-means, Gaussian mixtures (GMMs); • Topic modeling : latent Dirichlet allocation (LDA).	• Filtrage collaboratif: User-Based Collaborative Filtering, Item-Based Collaborative Filtering, Matrix Factorization with ALS, Matrix Factorization with ALS on Implicit Feedback Weighted Matrix Factorization, SVD++; • Classification: Logistic Regression - trained via SGD Naive Bayesv /Complementary Naive Bayes Hidden Markov Models.

Tableau 18: Comparatif des libraires libres d'apprentissage Spark et Mahout

En nous appuyant sur cette comparaison préalable, nous avons choisi de travailler avec Spark, notamment parce qu'il dispose d'un environnement informatique plus stable que Mahout dans les algorithmes de filtrage collaboratif orientés item et utilisateur[154].

En effet, la génération de recommandations sur **Spark est liée à l'algorithme ALS (Alternating Least Squares), fourni par sa librairie Mllib,** offrant la plupart des algorithmes et utilitaires d'apprentissage automatique. La recommandation **se fait en deux phases importantes : la première phase consiste à créer le modèle et le sauvegarder en cas de besoin. La deuxième phase consiste à utiliser le modèle construit, pour générer les recommandations à l'utilisateur actif**. En pratique, le modèle ALS repose sur une approche de factorisation matricielle, qui permet principalement de traiter de grandes quantités de données d'une manière rapide et précise.

La figure 33 illustre ce processus de génération des recommandations, en utilisant l'algorithme ALS :

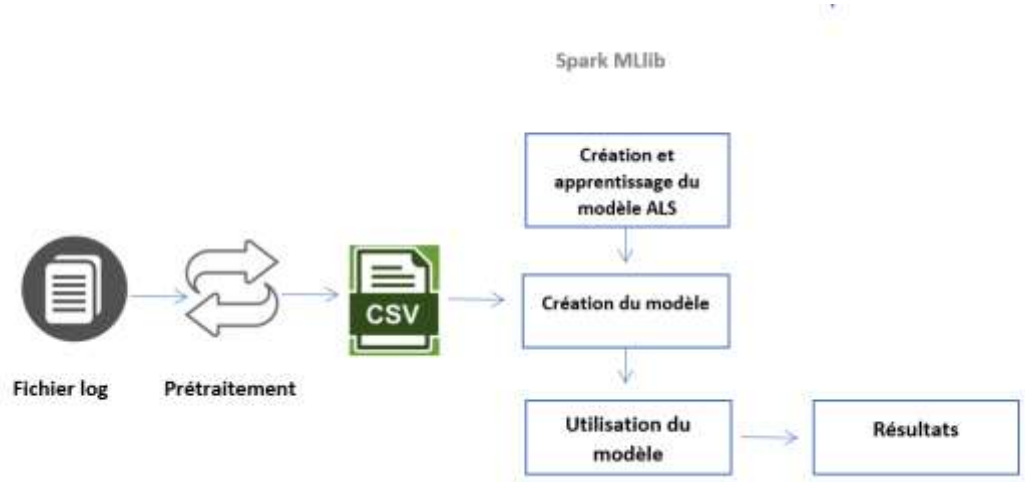

Figure 33: Le processus de génération de recommandations basé sur ALS

Comme le montre la figure 33, la première partie du processus est composée de quatre étapes essentielles : le prétraitement des données, le chargement des données, la préparation des données et enfin la création et l'apprentissage du modèle. Alors que la seconde partie consiste

[154] BESSE, Philippe et VIALANEIX, Nathalie. *Statistique et Big Data Analytics ; Volumétrie, L'Attaque des Clones* [en ligne], 2014. Disponible sur : https://hal.archives-ouvertes.fr/hal-00995801v3 [consulté le 06 Octobre 2016].

à employer ce modèle sur les données de l'utilisateur actif, puis générer des recommandations et procéder à leur évaluation.

L'algorithme ALS ou des moindres carrés alternés consiste à estimer la notation de chaque utilisateur pour tous les documents en se basant sur les notations existantes dans la matrice creuse. ALS tente d'estimer la matrice de R comme le produit de deux matrices utilisateurs et documents.

Le choix de l'algorithme ALS dans la factorisation matricielle pour implémenter une approche de filtrage collaboratif à base de modèle, s'explique par les considérations suivantes :

- Sa capacité à manipuler des grandes quantités de données ;
- Sa capacité à créer un modèle de profil pour les utilisateurs dans le cadre du filtrage collaboratif à base de modèle ;
- Sa capacité à réduire la taille de stockage de la matrice utilisateurs et documents ;
- Obtenir une matrice creuse de grande taille pour optimiser le calcul et le temps et améliorer la qualité des recommandations ;
- Procéder à la recommandation sans des surcharges sur le serveur ;
- Par ailleurs, il s'agit du seul algorithme permettant de faire de la recommandation dans la librairie Spark d'apprentissage automatique.

Le processus de l'algorithme sera mis en œuvre de bout en bout à travers notre prototype.

5. L'architecture fonctionnelle du modèle proposé

La figure 34 schématise le système de recommandation envisagé qui se « nourrit » initialement de la base de connaissance obtenue à partir des données *log* des utilisateurs.

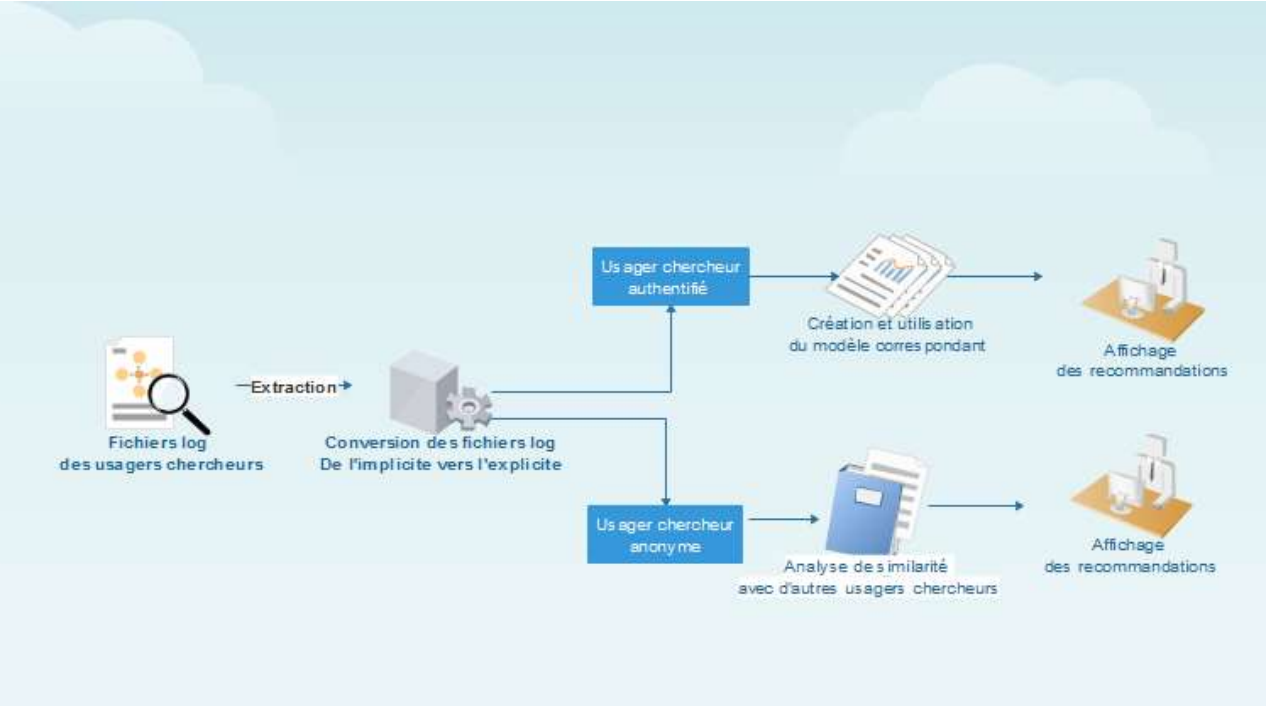

Figure 34: Le processus de recommandation à travers le modèle proposé

Les algorithmes sollicités dans ce processus s'inscrivent dans le cadre de l'approche de filtrage collaboratif à base de modèle, qui tend à construire un modèle de prédiction basé sur l'algorithme de factorisation matricielle et celui de la classification supervisée comme déjà évoqué. Ainsi, les données *log* implicites extraites (soient les données relatives au dernier document consulté, au nombre de rebonds de la requête, à l'affichage du document dans les résultats et au temps passé sur l'OPAC) seront transformées en données explicites traduites par des notations dans une matrice composée des vecteurs d'utilisateurs.

En pratique, plusieurs modèles d'apprentissage sont employés dans le filtrage collaboratif à base de modèle, il s'agit de la factorisation matricielle et du modèle des moindres carrés alternés (Alternating Least Squares) plus connu sous l'abréviation d'ALS.

Ce modèle agit sur les deux facteurs « $\mathbf{P_u}$ » et « $\mathbf{q_i}$ », respectivement les utilisateurs et les documents, pour construire un modèle d'apprentissage. Quant à la clasification, elle sert à classer les utilisateurs selon les spécialités ou disciplines de recherche conformément à la Classification Décimale de Dewey (CDD). Le processus de recommandation se déroule en plusieurs étapes que nous détaillons maintenant.

6. L'extraction de fichiers logs (Input)

Dans notre cas d'usage, les simulations de recommandations sont tributaires de l'exploitation d'un jeu de données que le service en charge du catalogue en ligne de l'IMIST a généré depuis les fichiers logs extraits du SIGB PMB.

Ces fichiers comprennent des paramètres de navigation des utilisateurs **(l'identifiant de l'utilisateur, le dernier document consulté, le nombre de rebonds de la requête, le nombre d'affichage du document dans les résultats et le temps passé sur l'OPAC)**.

7. La transformation des données implicites en données explicites (traitement)

Les données implicites recueillies ne sont pas directement exploitables et comparables, puisqu'elles ne sont pas homogènes. A titre d'exemple, le temps passé sur le catalogue en ligne ne peut pas être rapproché au nombre de rebonds de la requête et au nombre d'affichage de documents dans les résultats.

De surcroît, la similarité entre les profils de plusieurs chercheurs demeure une tâche délicate. A titre d'illustration, un utilisateur qui s'est connecté durant une période de 20 minutes et durant lesquelles il a effectué 13 requêtes, son profil ne peut pas être directement comparé à un autre chercheur ayant navigué sur l'OPAC pendant 11 minutes avec 10 requêtes effectuées.

Pour pallier l'hétérogénéité constatée sur notre jeu de données d'entrée, la méthode répandue dans le traitement des données implicites est de les transformer en notations semblables à celles collectées explicitement[155].

Par ailleurs, la particularité de notre modèle réside dans la combinaison de trois paramètres de données implicites pour estimer une note générale, qui exprime l'intérêt de l'utilisateur vis-à-vis des documents. Nous n'allons pas utiliser le paramètre relatif au dernier document consulté dans la formule de la note générale, vu qu'il n'est pas une

[155]OARD, Douglaset W., KIM, Jinmook, et al. Implicit feedback for recommender systems. In Proceedings of the AAAI workshop on recommender systems. 1998. p. 81-83.

valeur quantitative. Néanmoins, il nous est utile pour identifier la classe disciplinaire de l'utilisateur et constituer la base de données des documents à recommander.

Le schéma 35 rend compte des étapes du processus de transformation de données implicites en notations :

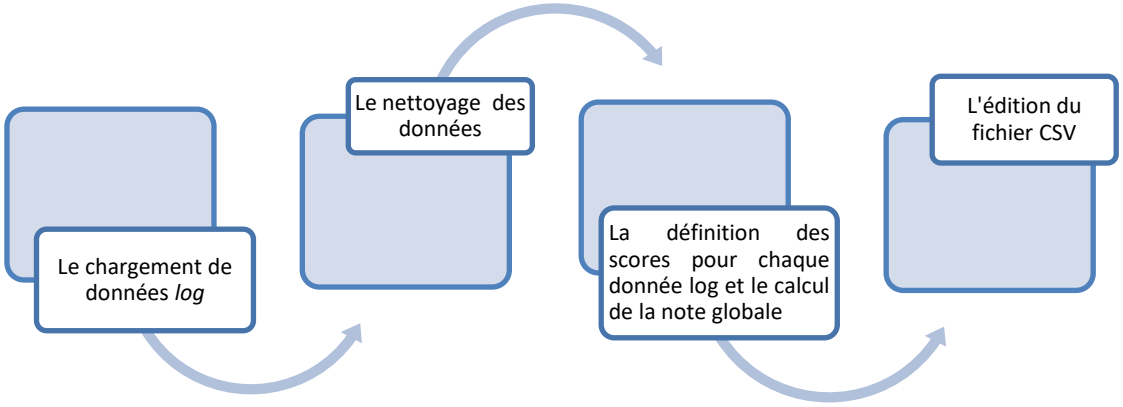

Figure 35 : Le processus de transformation de données implicites en scores

Pour réussir cette transformation, notre passage des données implicites aux données explicites, commence par nettoyer les données (données aberrantes, nulles et manquantes), définir, ensuite, les plages de notes à appliquer. Puis, il faudra calculer une note globale de l'intérêt d'un utilisateur sur la base d'une formule, tenant compte de tous les paramètres recueillis à partir des données *log*.

Nous obtiendrons un fichier final dans un format que nous pourrons exploiter par la suite dans notre modèle.

8. Le chargement de données

Les données sont souvent sous format texte (fichier *log*) exportées après des manipulations du service des systèmes d'information sous la forme d'une base de connaissance :

ID utilisateur	Dernier Document consulté (références)	Classes documents empruntés	Les rebonds de requête	Affichage de docs dans les résultats	Le temps passé sur l'OPAC
1	Fouché, Pascal , Péchoin, Daniel , Schuwer, Philippe .Dictionnaire encyclopédique du livre. [Paris] : Éd. du Cercle de la librairie, impr. 2011	Classe 600	2 fois	8 fois	10 Min
2	Gaussier, Éric , Yvon, François .Modèles statistiques pour l'accès à l'information textuelle. Paris : Hermès science publications-Lavoisier, impr. 2011	Classe 200	0 fois	3 fois	6 Min
3	ISAAC, Henri , Sid Ahmed, BENRAOUANE .Guide pratique du e-Learning. : DUNOD, 2011	Classe 000	2 fois	9 fois	34 Min
4	Fijalkow, Jacques , Fijalkow, Éliane .La lecture. Paris : le Cavalier bleu éd., impr. 2010	Classe 300	3 fois	12 fois	25 Min
5	Poullet, Yves , Henrotte, Jean-François .Droit des technologies de l'information et de la communication 2011. Bruxelles : Larcier, DL 2011	Classe 500	1 fois	5 fois	14 Min

Tableau 19: La forme du fichier à exploiter

Il est à rappeler que ces données *log* sont relatives à l'interaction de 500 utilisateurs chercheurs adhérents de l'IMIST lors des sessions journalisées durant l'année 2015.

9. Le nettoyage de données (traitement)

Il s'agit de détecter et nettoyer les incohérences présentes sur le fichier, c'est une opération qui se déroule en plusieurs étapes :

- **Le traitement des valeurs manquantes à travers l'analyse du fichier**. Effectivement, nous avons identifié trois enregistrements contenant des valeurs manquantes. Il s'agit d'une faible portion qui représente moins de 5% de la taille du fichier et, par conséquent nous avons éliminé ces enregistrements, nous n'avons conservé que les enregistrements entiers ;

- Sur un autre registre, **nous avons remplacé quelques valeurs nulles pour le paramètre du temps passé sur l'OPAC par la moyenne générale de ces sessions, soit 30 minutes, puisque cette valeur ne peut être nulle alors que les données**

relatives au nombre de rebonds de la requête, l'affichage des notices des documents et le dernier document consulté sont disponibles dans la base de connaissance.

10. La définition des plages de notes (traitement)

Pour définir les plages de notes, nous devons diviser les données en intervalles de même amplitude**. Le nombre d'intervalles est conditionné par l'échelle de notation à adopter. Pour y aboutir, deux échelles de notations sont définies : l'une sur 5 points et l'autre sur 10 points. Cette dernière a été retenue, pour garantir plus de précision avec des intervalles de 10 au lieu de 5. Dans un second temps, chaque intervalle sera traduit par une note, ce qui rend plus fiable l'estimation de la notation.**

Dans le même ordre d'idée, nous précédons à la substitution de chaque valeur contenue dans la plage de notes par la notation correspondante, comme le montre le tableau suivant :

	Borne Inférieure	Borne Supérieure	Borne Inférieure	Borne Supérieure	Borne Inférieure	Borne Supérieure
⇨ Note 1/10	[0 ;	0,8[	[0 ;	2,3[	[5 ;	7,4[
	[0,8 ;	1,6[	[2,3 ;	4,6[	[7,4 ;	9,8[
	[1,6 ;	2,4[	[4,6 ;	6,9[	[9,8 ;	12,2[
	[2,4 ;	3,2[	[6,9 ;	9,2[	[12,2 ;	14,6[
	[3,2 ;	4[	[9,2 ;	11,5[	[14,6 ;	17[
⇨ Note 6/10	[4 ;	4,8[	[11,5 ;	13,8[	[17 ;	19,4[
	[4,8 ;	5,6[	[13,8 ;	16,1[	[19,4 ;	21,8[
	[5,6 ;	6,4[	[16,1 ;	18,4[	[21,8 ;	24,2[
⇨ Note 9/10	[6,4 ;	7,2[	[18,4 ;	20,7[	[24,2 ;	26,6[
	[7,2 ;	8[	[20,7 ;	23[	[26,6 ;	29[

Tableau 20 : La définition des notations correspondant aux paramètres des données log recueillis

Il s'agit donc d'utiliser les différentes valeurs numériques extraites des données *log* avec des intervalles de même amplitude, pour en dériver, par la suite, une note selon la pondération (soit l'importance du type des données par rapport à l'intérêt d'un utilisateur).

Le tableau 20 donne un aperçu général sur le taux de rebond des requêtes, le nombre de fois que la notice d'un document a été affichée et le temps passé sur le catalogue en ligne, lesdites valeurs ont été comparées pour chaque type des données citées, en vue d'en déterminer les

écarts et en fixer des seuils Max et Min. De cette manière, dix notes sous forme d'intervalles ont été définies.

A titre d'exemple, les valeurs relatives correspondant au paramètre du taux de rebond des requêtes, les valeurs maximales et minimales correspondent respectivement à 0 et 8 fois, et les 10 notes étant des intervalles dont l'amplitude est de 0,8. Il en est de même pour les deux autres paramètres relatifs aux données d'affichage de la notice des documents et du temps passé sur l'OPAC.

Cette opération permet d'obtenir des notations sur une échelle unifiée, comme l'illustre le tableau 21, qui reprend l'exemple des données *log* des deux premiers utilisateurs de notre fichier d'entrée de 500 utilisateurs :

Identifiant de l'utilisateur	Notice du dernier document consulté	Note (rebonds de la requête)	Note (Affichage du document dans les résultats)	Note (temps passé sur l'OPAC)
1	Doc1	4	4	3
2	Doc 2	1	2	1

Tableau 21 : Extrait du fichier obtenu après la conversion des données en notations explicites

Nous avons attribué une note de 1 à 10 de manière que chaque note corresponde dans le tableau 21 à un intervalle de valeurs de même amplitude, et ce pour chaque paramètre (le taux de rebond des requêtes, le temps passé sur l'OPAC et le nombre de fois qu'une notice d'un document a été affichée).

Nous avons donc procédé à l'attribution d'une note de 1 à 10 pour les intervalles de chaque type des données *log*s implicites. A titre d'exemple, la note 1/10 correspond à l'intervalle [0 ; 0,8 fois [, pour les données relatives au taux de rebond des requêtes.

La base de connaissances dérivée du fichier log d'entrée dont un aperçu figure sur le tableau 22 indique que l'utilisateur dont l'ID est 1 a effectué 3 rebonds de requête, ce qui renvoie à une note de 4 /10 et il a affiché 8 notices de documents, ce qui correspond à la note de 4/10 et il a passé sur l'OPAC 10 minutes pour une note de 3/10 conformément à l'échelle des intervalles de notations de chaque type des donnés implicites figurant dans le tableau 21. Finalement, nous obtenons un fichier structuré, qui comprend des notations sur une échelle de 10 points au niveau des différents paramètres extraits de la base de connaissance dérivée du fichier log.

Pour la mise en cohérence des données, le logiciel XLSTAT spécialisé dans les statistiques, qui représente une extension d'Excel, a permis d'opérer les différentes phases de transformation. En effet, les données ont été enregistrées dans un fichier CSV (données tabulaires séparées par des virgules ou des points virgule), contenant 3 colonnes : « l'identifiant de l'utilisateur », « le titre du document » et « la note générale estimée » dont nous allons expliciter le calcul ci-après.

Ce fichier sera scindé en deux parties, la première comporte les identifiants des utilisateurs, des documents et les notes attribuées et la seconde affecte les ISBN des différents documents aux identifiants des notices correspondantes. Ce fichier sera exploité dans l'algorithme ALS de Spark pour générer les recommandations et se présente de la manière ci-après :

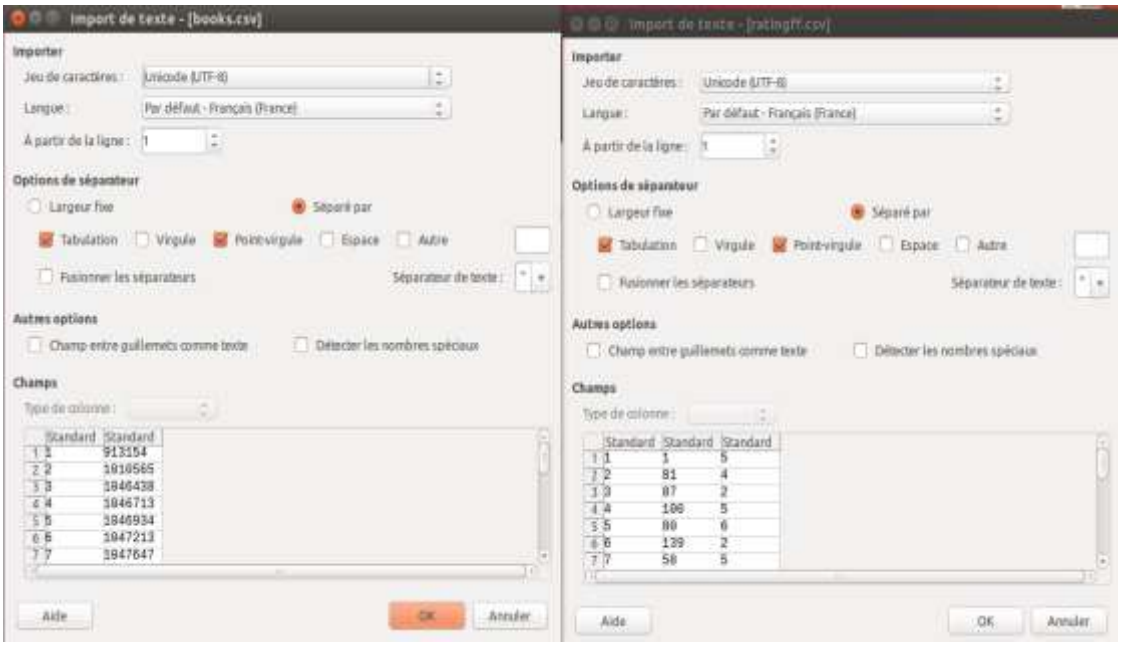

Figure 36: La structure des fichiers de test CSV (documents et évaluations)

Les deux fichiers sous format CSV seront sollicités par Spark tout au long des simulations de notre prototype de recommandation. En pratique, Spark devrait faire appel à ces fichiers à travers l'exécution du script Scala, en vue d'effectuer les traitements permettant de créer les modèles des utilisateurs servant à la génération des recommandations.

11. Le stockage sous le format final

L'étape suivante fut de calculer une note générale, qui prend en compte les 3 notations relatives aux paramètres considérés. **Après des discussions avec l'administrateur de base de données, nous avons convergé sur le fait que le paramètre d'affichage du document dans les résultats aura une pondération élevée, puisqu'il exprime au mieux l'intérêt de l'utilisateur chercheur vis-à-vis du document et la discipline ou la thématique dans laquelle il s'inscrit.**

Il convient aussi de mentionner que le fait de se contenter du dernier document consulté dans le fichier log ne peut pas poser un problème à notre modèle, dans la mesure où nos utilisateurs font partie d'une classe disciplinaire donnée et à laquelle correspond le dernier document consulté. En effet, le prototype envisagé s'appuiera sur les derniers documents consultés dans les différentes classes Dewey de nos 500 utilisateurs.

L'étape suivante étant de calculer une note générale qui prend en compte les notations produites. Après des discussions avec le service du système d'information, nous avons considéré que l'affichage du document dans les résultats aura une pondération élevée de 40%, puisqu'il exprime au mieux l'intérêt de l'utilisateur vis-à-vis du document, alors que les paramètres du temps passé sur l'OPAC et le nombre de rebonds de la requête traduisent plutôt les difficultés d'un utilisateur à trouver les documents qui répondent à ses besoins et auront une pondération de 30% chacun.

La formule proposée pour calculer cette note générale est la suivante :

Note générale = Note (le nombre de cas de rebonds de la requête) *30% +Note (l'affichage du document dans les résultats) *40% + Note (le temps passé sur l'OPAC) *30%

La formule présentée est certes un arbitraire, mais qui a été discutée et argumentée. Elle était nécessaire à poser pour calculer la moyenne finale à partir des trois notes obtenues dans l'opération de transformation des valeurs. IL est à noter que lors de l'évaluation de notre prototype, des paramètres de confiance seront vérifiés sur l'algorithme, pour mesurer la pertinence des ces données d'entrée du modèle.

Si nous reprenons toujours l'exemple de l'utilisateur ID 1, la note de 4/10 pour le taux de rebond de requêtes, la note de 4/10 pour l'affichage des notices et la note de 3/10 pour le temps passé sur l'OPAC correspondent à la formule suivante :

Note générale = Note 4/10 (Nombres de rebond de la requête) *30% +Note 4/10 (Affichage du document dans les résultats) *40% + Note 3/10 (Le temps passé sur l'OPAC) *30%= **3.7/10.**

Ainsi, le fichier de sortie ressemble au tableau 22, où nous ne retenons que l'ID de l'utilisateur, le dernier document consulté et la moyenne des paramètres des données *log* extraites et transformées :

ID de l'utilisateur	Dernier Document consulté	Note globale estimée
1	Doc 1	Note globale 1
2	Doc 2	Note globale 2
3	Doc 3	Note globale 3

Tableau 22: La présentation du fichier final

Chaque utilisateur de notre corpus est donc associé à une note globale calculée à partir des valeurs des données *log* extraites (affichage des documents dans la notice, le rebond de requête et le temps passé que l'OPAC) pondérées et transformées.

Chaque utilisateur se voit aussi associé à la classe CDD à laquelle appartient le dernier document consulté, faisant l'hypothèse que ce document correspond à ses préférences.

12. Le scénario de recommandation

Le scénario de recommandation se différencie selon deux cas distincts :

- **Si l'utilisateur est authentifié :**

Il se connecte par son login et mot de passe que l'administrateur lui a déjà défini au moment de son inscription. Le système charge son profil qu'il a créé et lui présente des recommandations basées sur ses interactions antérieures avec le système à travers une analyse de similarité à base de profil.

- **Si l'utilisateur n'est pas authentifié :**

Le système associe à son identifiant un modèle qui contient des informations de navigation, tel que l'historique de navigation, et lui propose des recommandations basées sur les documents consultés.

Le diagramme de la figure 37 ci-dessous explique ce scénario :

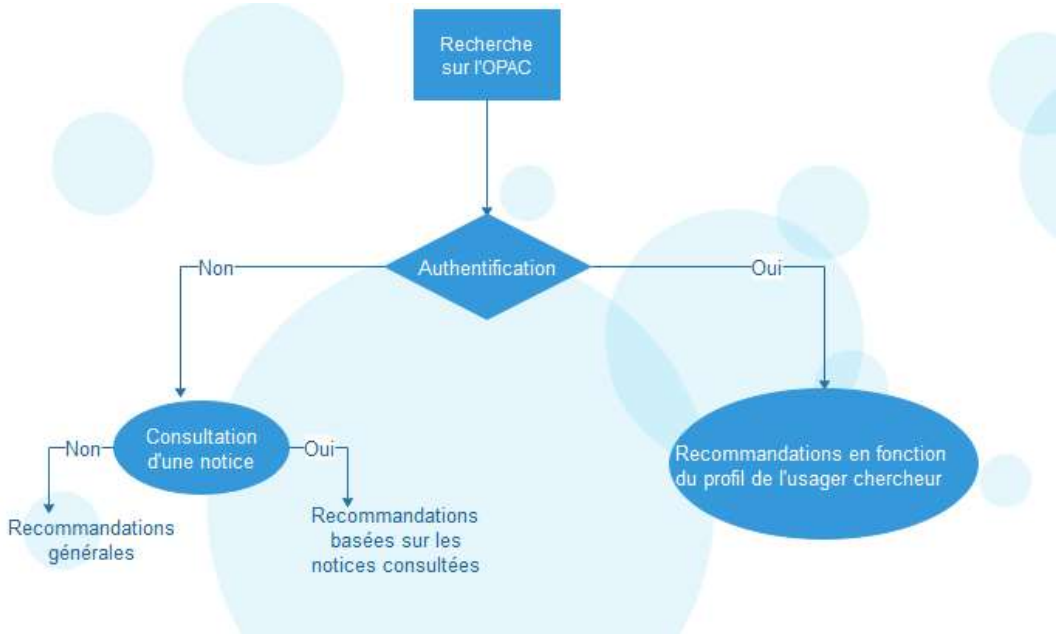

Figure 37: Le scénario de recommandation proposé à travers l'application

Ce schéma met en évidence la logique selon laquelle notre population d'utilisateurs chercheurs bénéficie des recommandations proposées par le système. Pour être précis, l'utilisateur authentifié reçoit des recommandations adaptées à son profil, et l'utilisateur anonyme aura des recommandations, en fonction des requêtes qu'il est en train d'effectuer. Ce fichier doit être converti en un fichier CSV que le modèle devrait exploiter.

A ce stade, il y a lieu d'établir une distinction entre deux cas de figure : un premier où l'utilisateur est abonné (authentifié) et le second où l'utilisateur est anonyme. Alors, un traitement spécifique est assuré pour chacun des deux cas :

- **Le cas d'un utilisateur authentifié**

Pour un utilisateur authentifié dont le modèle connait déjà le profil, le processus de recommandation s'opère selon le processus qui suit :

- **Le chargement du profil :**

Pour cet utilisateur le système dispose déjà du profil de l'utilisateur construit lors de sa connexion et il ne fait que le charger.

- **La création du modèle**

A travers le calcul de similarité entre les profils des utilisateurs, en s'appuyant sur le modèle de l'algorithme ALS ou des moindres carrés alternés qui consiste à estimer la notation de chaque utilisateur pour tous les documents, en se basant sur les notations existantes, pour prédire l'intérêt que pourrait porter l'utilisateur aux documents appréciés par des profils similaires au sien.

- **L'affichage de recommandations**

Une fois les documents à recommander déterminés, il reste à les présenter. Une page sera mise en place, pour interfacer les résultats de recommandations et permettre à l'utilisateur de naviguer et faire ses recherches.

- Le cas d'un utilisateur anonyme

Un utilisateur anonyme peut faire des recherches bibliographiques. Mais, il ne profite pas des recommandations dès le début, puisque son profil reste inconnu pour le modèle, qui lui crée alors un profil intermédiaire et prend en compte les notices qu'il consulte, afin de de lui fournir des documents similaires.

Pour un utilisateur anonyme, le processus de recommandation se déroule selon le scénario suivant :

- **La construction du profil**

Dans le cas d'un utilisateur qui se connecte à la première fois ou de façon anonyme, le système lui crée un profil instantané. Ce profil est lié aux informations de navigation de cet utilisateur, tel que l'historique des livres consultés. En bref, les recommandations se matérialisent par une barre de suggestion de documents que des chercheurs similaires ont consultés.

- **Le calcul de similarité à base de documents**

A travers le calcul de similarité à base de livres consultés, le système prédit et génère une liste de livres similaires à ceux que l'utilisateur a vus lors de la navigation.

Le calcul de similarité d'une notice consiste à chercher les documents les plus proches à cette notice, cette recherche dépend du modèle employé. Dans la plupart des cas, la similitude est calculée, en comparant la représentation vectorielle de deux items, et en utilisant une mesure de similarité. Les mesures de similarité communes incluent la corrélation de Pearson et la similitude du cosinus pour les vecteurs à valeur réelle et la similitude Jaccard pour les vecteurs binaires.

- **L'affichage de recommandations**

Une fois les documents déterminés, il faut les présenter et rafraichir le profil. Une page sera mise en place pour interfacer les résultats de recommandations, qui permet ainsi à l'usager de naviguer et effectuer ses requêtes. Au total, les recommandations se matérialisent par une barre, présentant les suggestions de documents que des chercheurs similaires ont consultés.

13. Les objectifs du modèle de recommandation proposé

Nous avons fixé plusieurs objectifs principaux pour notre modèle :

- Exploiter les donnes implicites des utilisateurs et les convertir en données explicites ;
- Dresser des profils d'utilisateurs ;
- Créer un modèle de recommandation et l'implémenter ;

- Interfacer les résultats de recommandations.

Ces objectifs constituent des modules qui composent notre prototype. Pour ce faire, nous allons développer une interface pour visualiser ces recommandations et tester le modèle. Une fois testé, le service pourra s'occuper de son intégration dans le SIGB de l'IMIST.

14. La description des acteurs intervenant dans le modèle proposé

Le modèle fait appel à plusieurs acteurs, soit de manière directe, soit indirecte, nous pouvons distinguer :

- **L'administrateur**

C'est un acteur qui intervient de façon indirecte, il gère le système et peut ajouter un nouvel utilisateur, modifier ou mettre à jour ses informations, et le supprimer.

- **Le professionnel (acteur métier)**

C'est un acteur qui intervient aussi de manière indirecte dans notre système, le professionnel s'occupe de la gestion des notices bibliographiques : l'ajout, la modification et la suppression. C'est lui qui met à jour la base de données de livres.

- **L'utilisateur**

C'est l'acteur principal du système. Ainsi, nous définissons, dans notre cas, deux types d'utilisateurs :

- L'utilisateur abonné : c'est un utilisateur identifié par le système, il peut se connecter et accéder à son compte, se déconnecter, mettre à jour son profil. Il peut également faire des recherches, visualiser les résultats de ses recherches sur l'OPAC et recevoir des recommandations d'articles, selon son profil ;

- L'utilisateur anonyme : il se connecte de manière anonyme, il peut également faire des recherches sur l'OPAC, consulte les résultats de ses recherches et reçoit des recommandations de documents, selon les notices qu'il a consultées.

La figure 38 synthétise les cas d'usages :

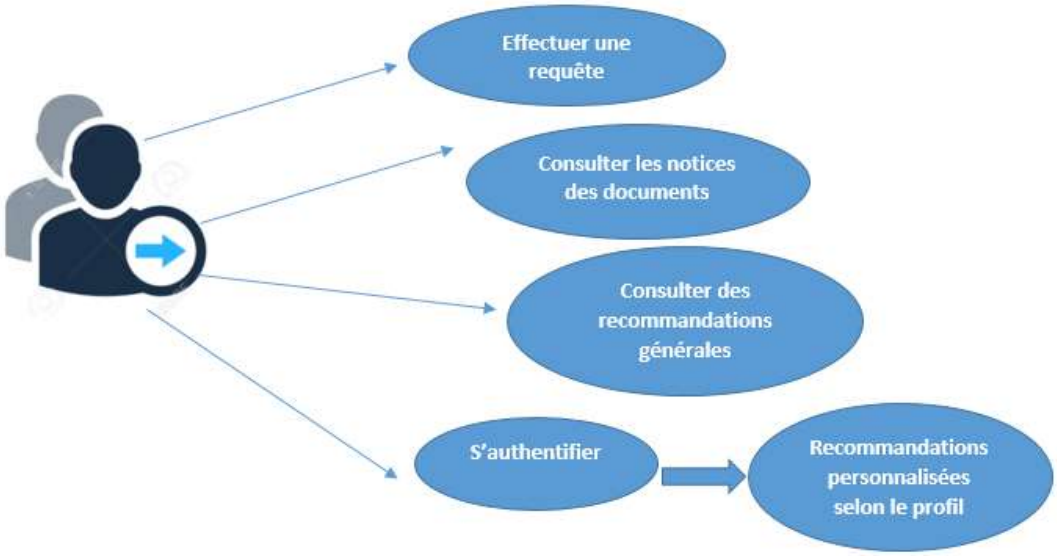

Figure 38: le diagramme de cas d'utilisation pour un utilisateur selon Unified Modeling Language (UML)

La différence entre un utilisateur abonné (qui s'est authentifié) et un autre utilisateur anonyme, réside dans le fait que le système propose à un utilisateur anonyme des recommandations basées sur la similarité entre des documents et sa recherche courante, puisque son profil reste inconnu, tandis qu'un utilisateur authentifié bénéficie de recommandations pertinentes liées à son profil enregistré dans le modèle.

Conclusion

Le modèle de recommandation que nous proposons part des centres d'intérêt communs en termes de la thématique du document recherché (classes Dewey), pour générer des groupes d'utilisateurs homogènes (catégories par affinités communes et obtenues à l'issue de l'opération de la classification supervisée).

Ainsi, les utilisateurs d'une même catégorie devraient pour toute recherche effectuée recevoir des recommandations de documents qu'ils n'ont pas consultés auparavant à travers l'approche du filtrage collaboratif passif à base de modèle, mais ceci dans le cadre de leur discipline de spécialisation correspondant à une classe de la CDD.

Pour concevoir notre modèle de recommandation, nous avons déployé deux algorithmes d'apprentissage automatique, à savoir la classification supervisée de nos utilisateurs par classe disciplinaire ou thématique, permettant d'obtenir des groupes homogènes d'utilisateurs et la factorisation matricielle par le biais de l'algorithme des moindres carrés alternés (ALS), pour dresser un modèle de nos utilisateurs authentifiés, en déterminant la valeur des matrices des utilisateurs et des documents dans le but de réduire la dimension de la matrice de base et diminuer la densité des données.

Le même algorithme a été utilisé pour proposer des recommandations correspondantes aux documents similaires à ceux retournés par le système pour donner suite à une première requête effectuée par un utilisateur anonyme.

Pour y parvenir, il a été décidé de transformer les données implicites, portant sur les documents consultés, le taux de rebond des requêtes et le temps passé sur l'OPAC en données explicitées par des notes après les avoir chargées et nettoyées, selon le cas de l'utilisateur anonyme et celui d'un utilisateur authentifié. Une note globale par utilisateur a été calculée, la pondération la plus importante a été conférée à l'affichage de la notice du document sur le résultat.

Le chapitre suivant présente l'architecture technique de notre modèle pour mettre en place le prototype envisagé selon l'algorithme de recommandation choisi.

Chapitre 4 : La mise en œuvre du prototype du système de recommandation

Introduction

Après avoir conçu le modèle de notre solution, nous allons présenter l'architecture technique du prototype de système de recommandation attendu et les différentes simulations, en fonction des deux scénarii d'usage déjà présentés : utilisateur authentifié et utilisateur anonyme.

Du point de vue technique, nous allons procéder à l'installation de l'environnement big data à travers le dispositif de stockage Hadoop, qui prendrait en charge le stockage distribué des données et lui adjoindre Spark pour traiter ces données, en les chargeant et créant le modèle de recommandation par le biais de l'algorithme ALS.

1. L'architecture technique du système

Si les systèmes de recommandation sont en plein essor dans les domaines du commerce en ligne et l'industrie, notamment en début des années 90, c'est en grande partie grâce au déploiement des technologies du big data et la puissance des algorithmes d'apprentissage, qui en ont considérablement amélioré les performances.

Dans notre cas, le filtrage collaboratif pose le défi du calcul des similitudes entre les utilisateurs, la classification des utilisateurs en groupes cohérents et leur actualisation en temps réel, selon des préférences enregistrées dans le modèle de prédiction établi. Un tel environnement justifie le recours à une architecture big data, qui devrait tenir compte de l'évolution des données du point de vue du stockage (nombre des utilisateurs et des documents) et du traitement (l'accélération de la courbe d'apprentissage, en vue de traiter les données et recommander des documents en temps réel).

Notre jeu de données actuel ne représente qu'un petit fragment des données enregistrées par le système, soient des données de 2015, comme nous l'avons signalé au moment de la présentation de notre corpus avec 500 utilisateurs, qui certes ne représente pas un volume considérable. Cependant, même dans ce contexte, les données *log* extraites ne peuvent être traitées

efficacement en temps réel pour construire le modèle de prédiction qu'avec une infrastructure big data.

Ainsi, l'historique de la base de données et son évolution ont été considérés lors de la phase de conception et de modélisation du prototype de recommandation, pour pouvoir disposer d'outils performants de stockage et de traitement des données. Les cas d'usage d'IA appliquée consistent justement en la spécification du modèle statistique par l'apprentissage sur les données, faisant varier un ensemble de paramètres.

L'environnement technique choisi concerne en premier l'installation d'Oracle VM VirtualBox, qui est un logiciel libre permettant la virtualisation. En d'autres termes, il s'agit d'un hyperviseur qui prend en charge la création d'une ou plusieurs machines virtuelles, à partir d'une autre machine physique sur un ou plusieurs systèmes d'exploitation invités (Linux Ubunto dans notre cas), comme le montre la figure 39 :

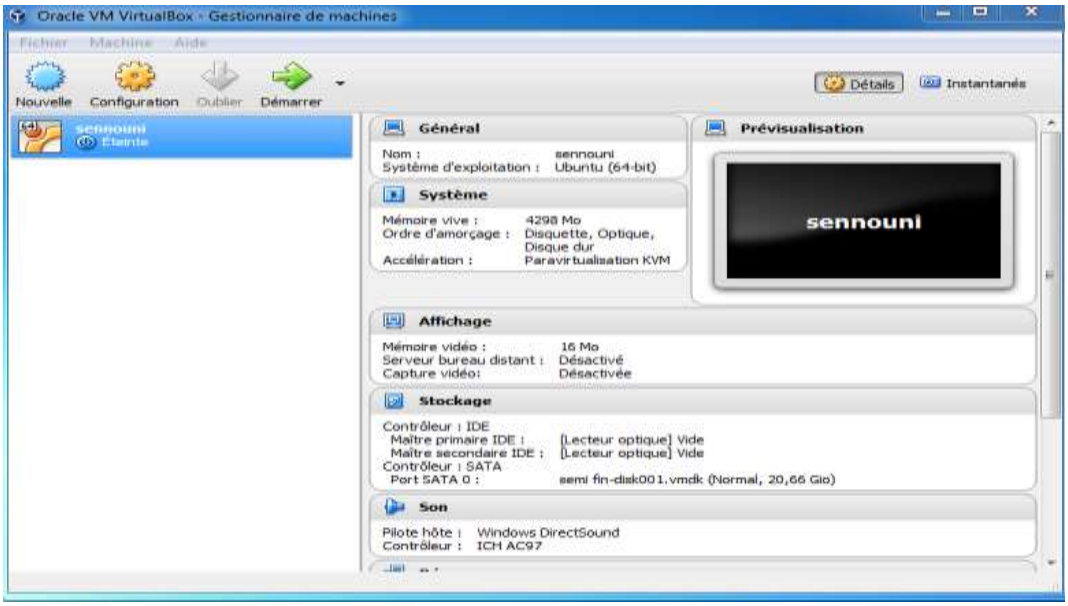

Figure 39: La machine virtuelle créée sous le logiciel « Oracle VM VirtualBox »

Sur notre machine virtuelle, nous procédons à l'installation de Hadoop, plateforme libre et open source en Java, destinée à faciliter la création d'applications distribuées (au niveau du stockage des données et de leur traitement) et échelonnables (scalables), permettant aux applications de travailler avec des milliers de nœuds et des péta-octets de données.

Son apport dans notre cas est notamment le système de gestion et de stockage de fichiers qui lui est propre connu sous l'acronyme HDFS (Hadoop Distributed File System) et sur lequel prendra appui le prototype.

Notre second cadre technique pour le traitement des données est Spark[156], comme indiqué dans la figure 40.

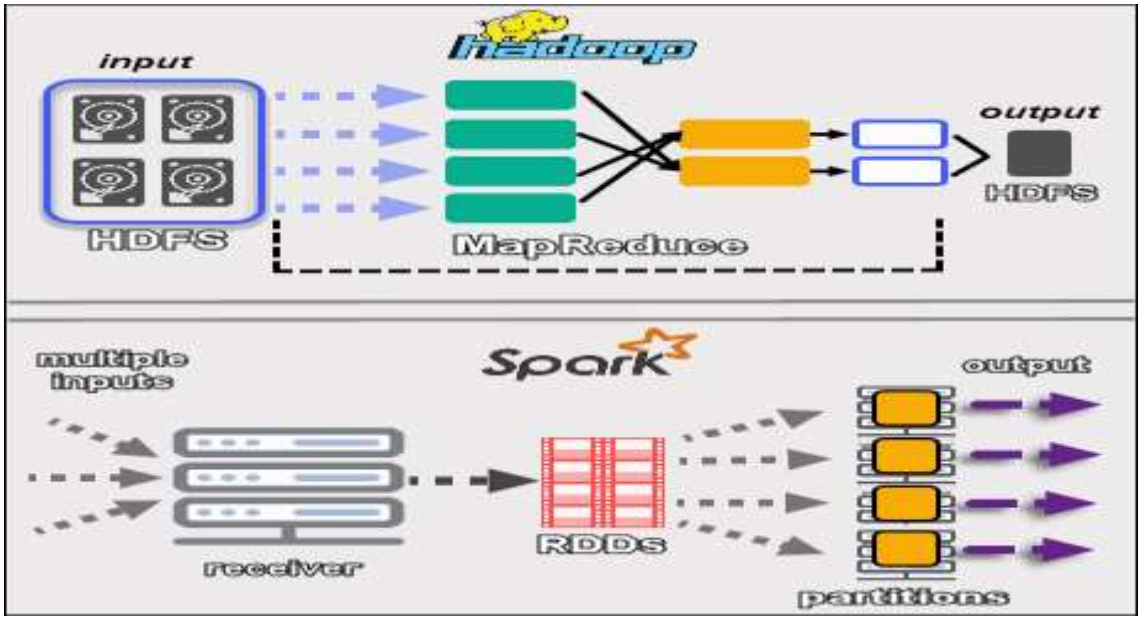

Figure 40: L'architecture de stockage (HDFS de Hadoop) et de traitement des données (Spark) (http://spark.apache.org/)

Figure 41 : Les fichiers d'installation de Hadoop et Spark sur la machine virtuelle

[156]Apache Spark™ - Lightning-Fast Cluster Computing [En ligne]. Disponible sur http://spark.apache.org/. [Consulté le 12 juin 2019].

Apache Spark est un moteur de traitement de données de manière distribuée et sur lequel sera utilisé l'algorithme ALS de la librairie Mllib. Il aide le traitement en temps réel palliant ainsi la latence importante, qui marque le composant traitement « MapReduce » de « Hadoop ».

L'architecture technique de notre prototype est résumée dans la figure 42.

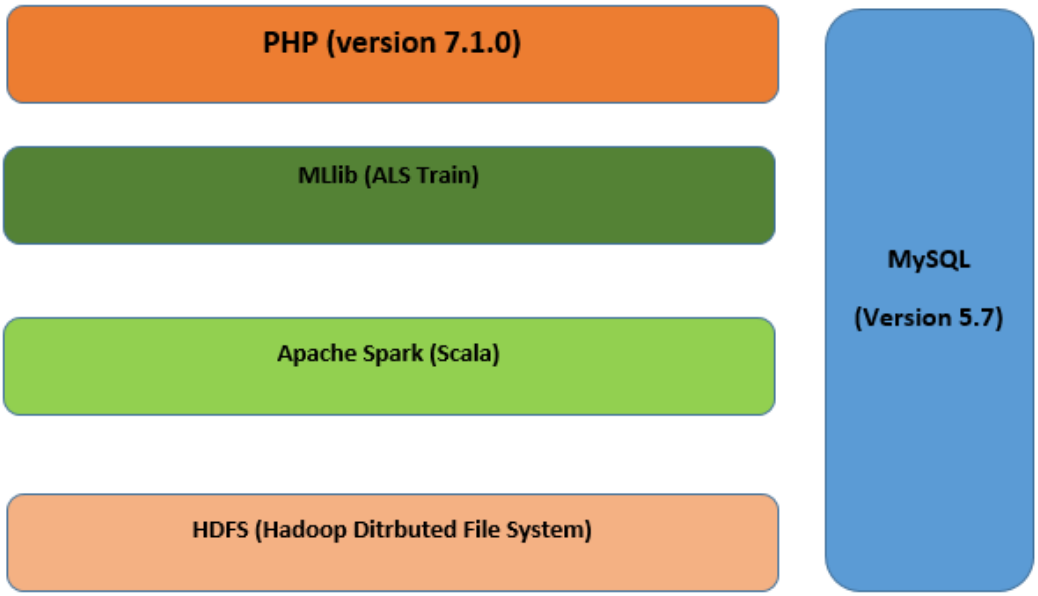

Figure 42: L'architecture technique de la solution

Notre solution préconise l'usage de plusieurs outils et composantes :

- **Application PHP** : cette application permet de collecter les données et présenter les recommandations à l'utilisateur (interfaçage des recommandations). Nous notons qu'il s'agit de la version 7.1.0 du PHP ;
- **MLlib :** il s'agit d'une librairie de Machine Learning, offrant la plupart des algorithmes et utilitaires d'apprentissage automatique. Elle nous a permis de mettre en place l'algorithme de recommandation ALS ;
- **Apache Spark :** c'est l'environnement d'analyse et de programmation utilisé ;
- **File system** : il sert à stocker les fichiers qui assurent l'intermédiation entre PHP et Scala, ces fichiers contiennent les recommandations, ainsi que les identifiants de l'utilisateur et de la notice ;
- **MySQL :** système qui gère la base de données de documents sur laquelle l'application PHP effectue la recherche, et récupère les métadonnées des livres à recommander.

2. Un aperçu sur Spark

Apache Spark est un moteur de traitement de données big data, open source, qui a été conçu pour effectuer des traitements sophistiqués sur de grands volumes de données, et se caractérise par la rapidité et la facilité d'utilisation.

Dans la palette des technologies big data, Spark présente plusieurs atouts dont :

- Un Framework complet qui permet d'effectuer des traitements Big data sur divers types de jeux de données ;
- Le traitement en streaming (en temps réel et avec une vitesse importante) ;
- Trois langages de programmation pour écrire un code : Scala, Java, ou Python ;
- L'exécution des applications sur les clusters de Hadoop ;
- Les traitements parallèles de données ;
- Les fonctionnalités de Machine Learning et les traitements orientés graphes[157].

L'écosystème Spark[158] peut être décrit par la figure 43.

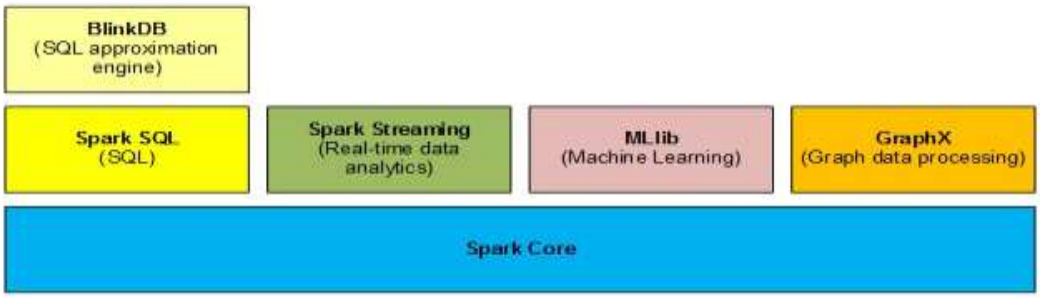

Figure 43: L'écosystème de Spark (http://spark.apache.org/)

L'écosystème de Spark se compose de plusieurs services de traitement de données :

- **Graph X** : il aide à calculer les graphes ;

[157] *Traitements Big data avec Apache Spark - 1ère partie : Introduction* [En ligne]. Disponible sur https://www.infoq.com/fr/articles/apache-spark-introduction. [Consulté le 12 juin 2019].
[158] Apache Spark™ - Lightning-Fast Cluster Computing [En ligne]. Disponible sur http://spark.apache.org/. [Consulté le 12 juin 2019].

- **MLlib** : il représente une librairie de Machine Learning (d'apprentissage automatique), il contient la plupart des algorithmes et utilitaires d'apprentissage automatique, tels que la classification, la régression, le clustering et la recommandation fondée sur le filtrage collaboratif. Nous allons utiliser cette librairie d'algorithmes de Spark pour utiliser le modèle ALS ;
- **Spark SQL** : il permet l'interrogation des données à travers une implémentation SQL ;
- **Spark streaming** : il est utilisé pour traiter en continu (temps-réel) les flux des données ;
- **Blink DB :** il s'agit d'un moteur de requêtes déployé pour exécuter des requêtes SQL interactives sur des volumes de données importants.

3. La mise en place du prototype du système de recommandation

Rappelons que l'approche de recommandation retenue dans notre modèle est celle du filtrage collaboratif passif à base de modèle, qui effectue une exploration sur les données d'entrée, en vue de générer des modèles de prédiction, comme nous l'avons déjà présenté dans le chapitre 1 de la partie III traitant de la modélisation des systèmes de recommandation. En fait, ce choix dépend principalement de la nature des données dont nous disposons et qui sont liées aux utilisateurs et à leurs interactions avec le système. Il s'agit de définir l'utilisateur dans le contexte des autres utilisateurs, ce qui est au cœur de l'approche de recommandation par filtrage collaboratif.

Le filtrage collaboratif déployé est à base de modèle et non à base de mémoire. C'est-à-dire que dans ce cas, la création d'un modèle de prédiction dépend des données d'entrée, sans avoir besoin de les solliciter par la suite, alors que dans le cas d'un système à base de mémoire, il dépend des données stockées dans la mémoire du système pour prédire la notation sur un objet cible et générer les recommandations, en s'appuyant sur les profils utilisateurs.

En principe, **l'approche de filtrage collaboratif à base de modèle s'opère par le calcul de la similarité entre les utilisateurs et/ou les objets selon le calcul de coefficients (Cosinus, Pearson ou Spearman). Les proches voisins sont ensuite identifiés, puis la prédiction et le classement des items sont opérés pour présenter les recommandations.**

L'algorithme retenu pour le filtrage collaboratif à base de modèle est celui de la factorisation matricielle, dont le choix trouve sa pertinence dans la nature des données *log* recueillies et présentées dans une base de connaissance et que nous avons transformées en notations des documents consultés, en fonction du temps passé sur l'OPAC, du taux de rebond et de l'affichage de la notice des documents en question.

Ainsi, ces notations dérivées de notre processus de conversion, seront rassemblées dans une matrice composée des vecteurs d'utilisateurs et des items, et qui a pour objectif de mettre en œuvre la factorisation matricielle, qui consiste à réduire la dimension de la matrice et prédire les valeurs manquantes, vu que les utilisateurs n'arriveront pas à exprimer un usage à l'égard de tous les documents sur le système.

Dans notre cas de figure, la factorisation matricielle est déployée par le seul modèle d'apprentissage présent sur Spark dédié à la recommandation, à savoir Alternating Least Squares (ALS) ou le modèle des moindres carrés alternés, sur lequel la prédiction des valeurs manquantes de la matrice creuse prendra appui.

Notre système de recommandation est réparti en deux composantes, dont chacune repose sur un traitement (algorithme) précis. Nous distinguons entre deux architectures selon la nature de l'utilisateur connecté (authentifié ou anonyme).

4. Le cas de l'utilisateur authentifié

Dans le cas de figure d'un utilisateur authentifié, le fonctionnement du prototype de recommandation se présente de la manière ci-après :

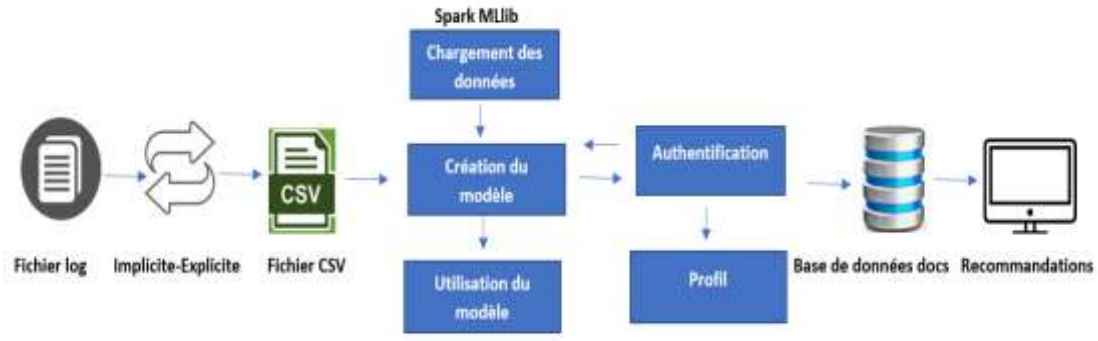

Figure 44: Le fonctionnement du système de recommandation envisagé : cas d'un utilisateur authentifié

L'utilisateur peut s'authentifier depuis le catalogue en ligne, représenté par une application PHP que nous avons conçue pour jouer le rôle du catalogue en ligne de notre prototype. Le couple identifiant et mot de passe, permettent à Spark de solliciter l'algorithme de recommandation, qui charge à son tour les données des utilisateurs contenues dans le fichier CSV, données sous forme de valeurs séparées par des virgules ou des points virgules.

Dans notre cas, le fichier CSV représente la sortie de la transformation du fichier log d'entrée, c'est l'input du calcul de similarité entre les profils des utilisateurs, pour générer des recommandations, ces dernières seront communiquées à l'usager via l'application s'apparentant dans notre prototype au catalogue en ligne.

Pour tester notre modèle, nous prenons quatre utilisateurs dont le comportement de navigation sur l'OPAC est similaire, et qui font partie du groupe de la classe 600, selon notre classification supervisée inspirée de la CDD (groupe des technologies et des sciences appliquées). Nous vérifions si le système recommande bien à l'un d'eux les documents appréciés par les autres utilisateurs. Le tableau 23 présente les données associées aux utilisateurs :

ID de l'utilisateur	ID du document	Intitulé du dernier document consulté	Nombre de rebonds de la requête	Affichage du document dans les résultats	Temps passé sur l'OPAC	Note globale estimée
403	**201**	**12 Simple Steps To A Winning Marketing Plan**	**8**	**10**	**9**	**9,1**
405	**205**	**Les nouveaux tableaux de bord des managers: le projet Business Intelligence clés en main**	**7**	**9**	**8**	**8,1**

Tableau 23: Les données relatives aux profils servant au test du prototype

Notre prototype recommande à un utilisateur les documents que d'autres utilisateurs du même profil ont apprécié. De cette manière, si nous effectuons un test avec l'utilisateur 403, le système

devrait proposer les documents que l'utilisateur 405 a consultés et inversement, dans la mesure où la note globale explicite exprimant l'intérêt d'un utilisateur à certains documents dans leur classe disciplinaire est quasi-similaire pour ces deux utilisateurs avec une note globale estimé de 9.1 pour l'utilisateur 403 et 8.1 pour l'utilisateur 405, sachant qu'il s'agit de deux documents différents.

Il est à souligner que la moyenne est calculée sur la base des paramètres de la base de connaissance dérivée des données *log*. Il s'agit de la moyenne des valeurs attribuées aux paramètres du taux de rebond de la requête, le nombre d'affichage des notices des documents dans les résultats selon la pondération qui y correspond.

Pour le cas d'un utilisateur abonné (authentifié), le processus de recommandation se déroule comme suit :

- L'authentification de l'utilisateur

L'utilisateur s'authentifie à travers le catalogue en ligne matérialisé par l'application PHP conçue et ceci à l'aide de son identifiant et son mot de passe. L'interface d'accueil du catalogue se présente comme suit :

Figure 45: L'interface de connexion à l'application (catalogue en ligne)

L'utilisateur s'authentifie via son identifiant et son mot de passe. Une fois authentifié, l'identifiant de cet utilisateur est transmis à Spark où se poursuivront les traitements.

- **Le calcul de recommandations sur Spark**

Spark offre un langage de programmation par défaut appelé Scala. A l'aide de ce langage, nous avons élaboré un modèle ou algorithme qui génère des recommandations suivant trois phases de traitement :

1ère étape : Le chargement des données

L'algorithme procède au chargement des données nécessaires pour le traitement en scala, à savoir les bibliothèques à utiliser, la base de connaissance et l'identifiant de l'utilisateur Chargement des librairies à utiliser, à savoir :

  - org.apache.spark.mllib.recommendation.ALS
  - org.apache.spark.mllib.recommendation.Rating
  - org.apache.spark.mllib.recommendation.matrixFactorizationModel
- Chargement des données de la base (fichier de notations), (/home/sennouni/Bureau/bases/rating.csv)
- Chargement de l'identifiant de l'utilisateur depuis un fichier texte « user.txt », et nous le convertissons en entier, ce fichier nous garantit l'échange de données entre le PHP et Scala.

Ci-dessous un extrait du script Scala correspondant :

```
// chargement des données
Val rawData = sc.textfile(" /home/sennouni/bureau/bases/rating.csv)
Val rawRatings = rawData.map (_.split (" ;").take(3))
//chargement des librairies
import org.apache.spark.mllib. recommendation .ALS
import org.apache.spark.mllib. recommendation .Rating
import scala.to.source
//chargement imputs
Val ratings = rawRatings . map {case Array (user, book, rating) => Rating (user. to Int, book. toInt, rating.toDouble)}
//chargement de l'utilisateur
Val reader = source. FromFile (" /var/www/html/rec/test.tst " ). getLines.mkString
Val i = reader. ToInt
Println(i)
```

Figure 46: Le chargement des données par le langage Scala

2ème étape : la création du modèle

Après avoir chargé les données, l'étape suivante est de créer le modèle avec la méthode de la factorisation matricielle à travers l'algorithme ALS de la librairie MLLib. Cette méthode dispose d'une variante dédiée aux données implicites appelée « ALS.trainImplicit », qui s'appuie sur des notations calculées sur la base des données implicites. Cette méthode prend en charge quatre variables :

- **Le rang** : il désigne le nombre de facteurs dans notre modèle ALS. C'est-à-dire, le nombre de caractéristiques cachées dans les matrices d'approximation de bas rang ;
- **Le nombre des itérations à exécuter** : le modèles ALS convergera vers une solution raisonnablement bonne après peu d'itérations ;
- **Le paramètre « lambda »** : il établit un contrôle sur la régularisation de notre modèle, dans la mesure où plus la valeur de lambda est élevée, plus la régularisation est appliquée ;
- **Le paramètre « alpha »** : il est lié aux données implicites, qui contrôle le niveau de référence de pondération de confiance appliqué.

Il y a aussi lieu de désigner un paramètre de confiance, puisque les notations que nous possédons ne sont pas exprimées par les utilisateurs, mais nous les avons calculées à partir des paramètres reliés à leurs comportements de navigation, nous introduisons alors un degré de confiance du système basé sur les quatre variables citées ci-dessus.

La création du modèle dans le langage Scala, est illustrée dans la figure 47 :

Figure 47: La création du modèle par le langage Scala

Après plusieurs tests et en agissant sur les paramètres du modèle à travers la confrontation de nos fichiers CSV relatifs aux utilisateurs et aux documents avec le modèle ALS.trainImplicit. Ainsi, nous avons abouti à un modèle optimal dont les valeurs des paramètres sont : le rang=50, itérations=10, lambda=0.01 et alpha=1.

3ème étape : l'utilisation du modèle et génération de recommandations

Le modèle est, ensuite, sollicité par le système, pour générer des recommandations pour l'utilisateur, ces dernières seront conservées dans un fichier (/home/sennouni/bureau/recommandations.txt) pour les exploiter plus tard. La figure 48 rend compte du code représentant le moteur du modèle ALS train implicit :

```
// affichage des recommandation
println(topKRecs.mkString("\n"))
//inspection des titre
val books  = sc.textFile("/home/sennouni/Bureau/bases/books.csv")
val titles =       .map(line => line.split(";").take(2)).map(array=> (array(0).toInt, array(1))).collectAsMap()
//---------------
val booksForUser = ratings.keyBy(_.user).lookup(100)
println(      ForUser.size)
// recom for user pas sur als
booksForUser.sortBy(-_.rating).take(10).map(rating => (titles(rating.product), rating.rating)).foreach(println)
//affichage des reco base sur als
//topKRecs.map(rating => (titles(rating.product), rating.rating)).foreach(println)
//val reco = topKRecs.map(rating => (titles(rating.product), rating.rating))
//val reco = topKRecs.map(rating => (titles(rating.product)))
//reco.(_.mkString(";")).saveAsTextFile("/home/ sennouni/Bureau/rec.txt")
```

Figure 48: L'utilisation du modèle et la génération de recommandations

A travers ce script, nous effectuons les opérations suivantes :

- Le calcul des prédictions à travers l'utilisation du modèle ;
- Le chargement du fichier contenant les livres, pour pouvoir retrouver les titres et les ISBN : (/home/sennouni/Bureau/bases/books.csv) ;
- L'enregistrement des recommandations sur un fichier texte, ce sont les ISBN des livres à recommander que l'application PHP exploitera par la suite.

Le script ci-dessus une fois exécuté en langage Scala génère des recommandations de la manière suivante :

Figure 49: Les recommandations générées depuis Spark

Le script génère les top K recommandations pour l'utilisateur courant dont l'identifiant est 503, et qui s'est vu recommander les documents dont les identifiants sont : 201, 200, 212, 203, 202 sur la base de la note globale associée à son profil. Finalement, le système vérifie les ISBN des documents, qui seront transmis à l'interface PHP, pour afficher les notices.

- **L'affichage des recommandations**

Au niveau de l'application de recherche PHP que nous avons développée et qui joue le rôle du catalogue en ligne (OPAC), lorsque l'utilisateur effectue une recherche, il reçoit en plus des résultats de recherche, des recommandations liées à son profil.

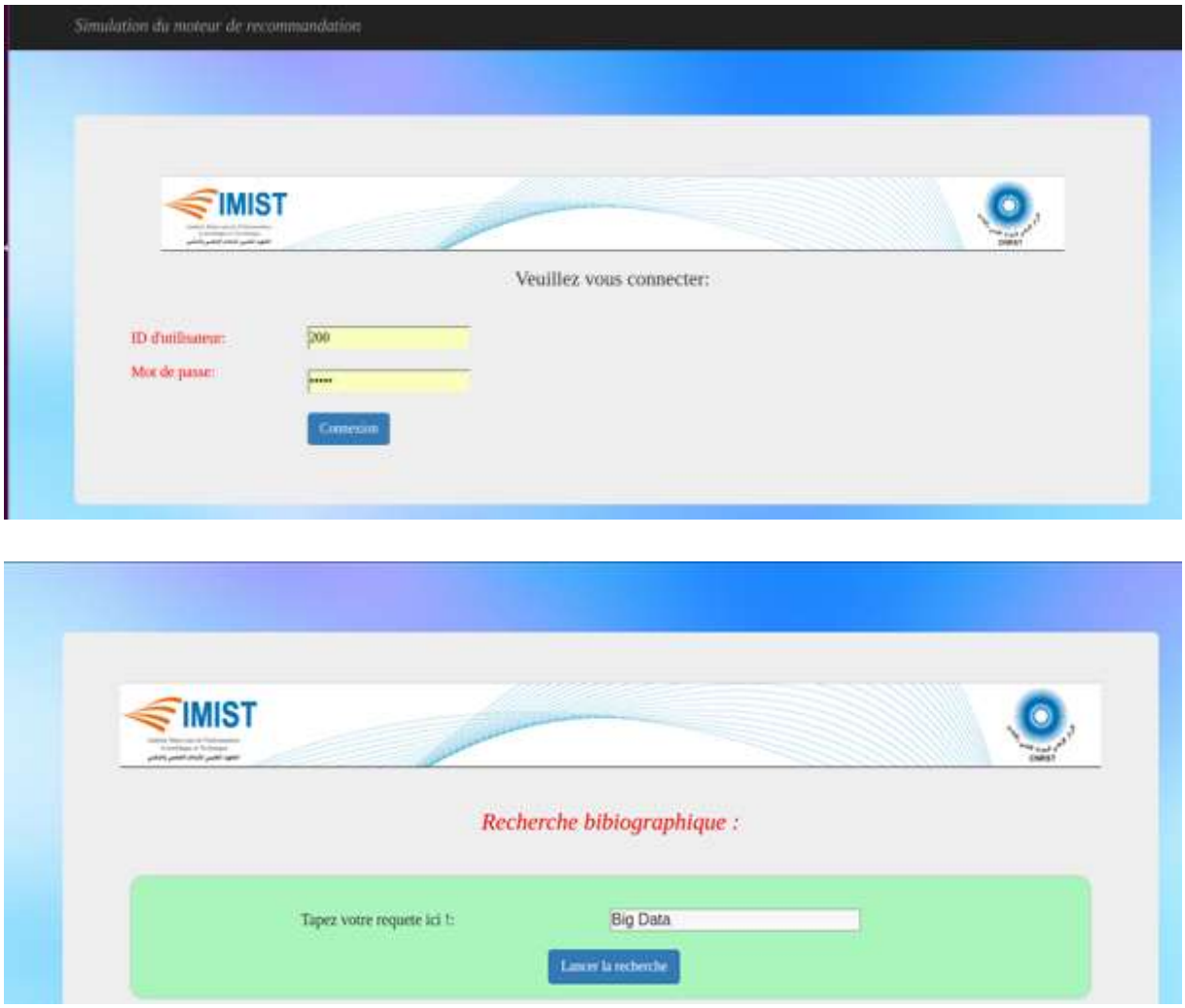

Figure 50: Le lancement de la requête par un utilisateur authentifié

Une fois sa requête lancée, l'utilisateur reçoit des recommandations qui lui sont appropriées après identification du groupe CDD auquel il fait partie à travers des recommandations associées au champ disciplinaire du dernier document consulté. La figure 51 montre les résultats de recherche suivis par des recommandations :

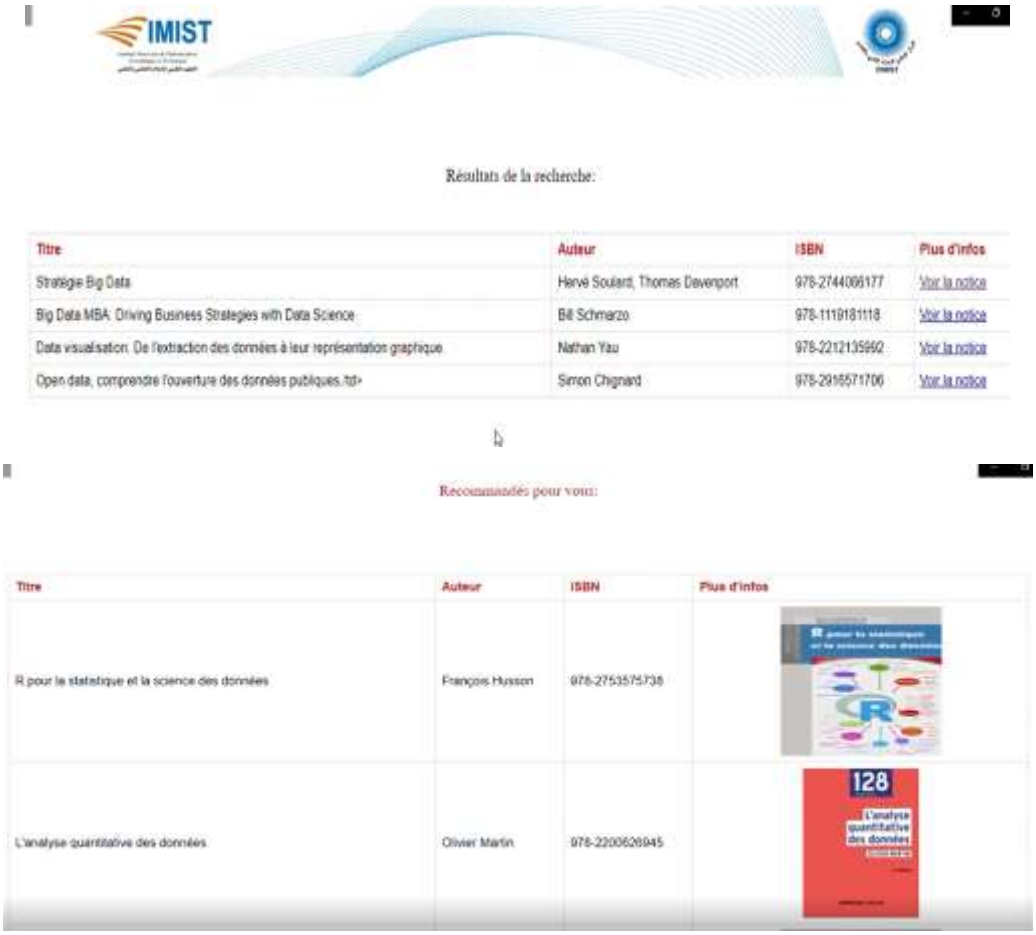

Figure 51: La présentation des recommandations à un utilisateur authentifié

Les top cinq recommandations sont affichés directement après les résultats de recherche. Ces recommandations contiennent le titre du livre, l'auteur, L'ISBN, et sa couverture.

5. Le cas de l'utilisateur anonyme

Un utilisateur anonyme peut accéder à l'interface de recherche du catalogue en ligne et consulter les notices bibliographiques.

Le prototype s'appuie alors sur les identifiants des notices pour lui proposer des livres similaires à la notice du document visualisé lors d'une requête. La figure 52 explicite le schéma de fonctionnement de la recommandation dans ce cas :

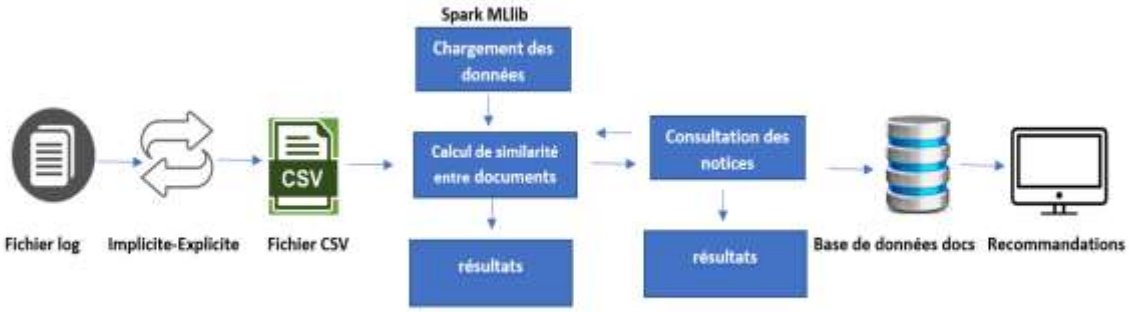

Figure 52 : Fonctionnement du système de recommandation : cas d'un utilisateur anonyme

Le fonctionnement du système de recommandation pour un utilisateur anonyme (Figure 52) diffère de celui d'un utilisateur authentifié (Figure 44), par le fait que dans le second cas le modèle est établi sur la base des données de notations obtenues alors que dans le premier cas la similarité est calculée avec les documents qui existent dans la base de données.

Pour le cas d'un utilisateur anonyme, le processus de recommandation se déroule comme suit:

- **La navigation sur le catalogue en ligne**

Un chercheur anonyme peut lancer des recherches bibliographiques, mais ne profite pas des recommandations dès le début. La figure 53 montre l'affichage des résultats de recherche à propos de « Big data » pour un utilisateur anonyme :

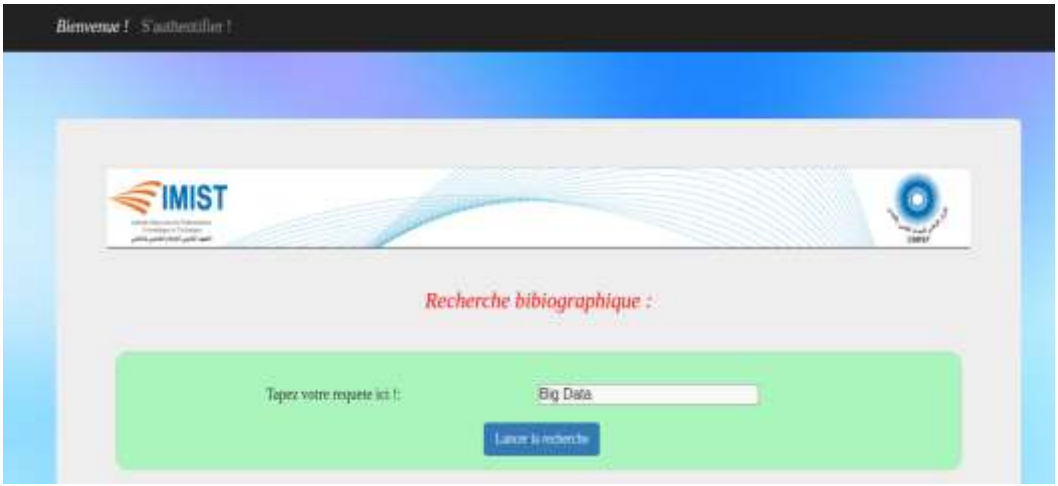

Figure 53 : Les résultats de recherche pour le cas d'un utilisateur anonyme sur «big data »

Lors d'une recherche bibliographique avec le mot clé « Big data », l'application simulant le catalogue en ligne a retourné trois résultats. Les recommandations sont absentes, puisque le système ne connait pas encore le besoin de l'utilisateur. Si l'utilisateur choisit de consulter une notice, l'identifiant de celle-ci sera transmis à Spark, qui calcule la similarité avec les livres de la base de données et lui propose les top K livres similaires comme des recommandations.

- **Le calcul de recommandations sur Spark**

1ère étape : Le chargement des données

L'algorithme procède au chargement des données nécessaires pour le traitement en Scala, à savoir les utilisateurs à utiliser et la base de connaissance et l'identifiant de la notice.

Notre algorithme développé en Scala sur l'environnement Spark, procède aux opérations suivantes :

- Le chargement des librairies à utiliser, à savoir :
  - org.apache.spark.mllib.recommendation.ALS ;
  - org.apache.spark.mllib.recommendation.Rating ;
  - org.apache.spark.mllib.recommendation.matrixFactorizationModel.
- Le chargement du fichier de notations ;
- Le chargement de l'identifiant de la notice pour laquelle Spark calculera la similarité.

Ci-dessous un extrait du code de chargement des données écrit en Scala :

```
//chargement des library
import org.apache.spark.mllib.recommendation.ALS
import org.apache.spark.mllib.recommendation.Rating
// chargement de inputs
val ratings = rawRatings.map { case Array(user,books, rating) => Rating(user.toInt,books.toInt, rating.toDouble) }
//chargement du model
val model = ALS.trainImplicit(ratings, 50, 10, 0.01,1)
//calcul  de predictions
//val predictedRating = model.predict(100, 123)
val userId = 496
val K = 50
val topKRecs = model.recommendProducts(userId, K)
```

Figure 54 : Chargement des données par Scala pour calculer la similarité

2ème étape : La création du modèle et calcul de similarité

Après avoir chargé les données, l'étape suivante est de créer le modèle comme celui créé dans le cas de l'utilisateur authentifié. Nous utilisons le même algorithme « ALStrainimplicit », pour calculer la similarité. Pour ce faire, nous avons, dans ce cas, opté pour la mesure de similarité de cosinus, qui calcule l'angle résultant de deux vecteurs dans un espace de « n » dimensions. Le choix de cette mesure est expliqué par le fait que nous ne disposons pas de données sur l'utilisateur pour implémenter une factorisation matricielle, qui requiert des notations affectées à des documents.

Le code du modèle, qui calcule la similarité est présenté dans la figure 55 :

```
//creation fonction cosineSimilarity
def cosineSimilarity(vec1: DoubleMatrix, vec2: DoubleMatrix): Double = {  vec1.dot(vec2) / (vec1.norm2() * vec2.norm2()) }

//---------------------creation de la matrice de similarité
//utilisateur
//val reader = Source.fromFile("/var/www/html/rec/test.txt").getLines.mkString
//val i = reader.toInt
//println(i)

val itemId = 12
val itemFactor = model.productFeatures.lookup(itemId).head
val itemVector = new DoubleMatrix(itemFactor)
cosineSimilarity(itemVector, itemVector)

//application
val sims = model.productFeatures.map{ case (id, factor) =>   val factorVector = new DoubleMatrix(factor)
val sim = cosineSimilarity(factorVector, itemVector)  (id, sim) }

//calcul
// recall we defined K = 10 earlier
val sortedSims = sims.top(K)(Ordering.by[(Int, Double), Double] { case (id, similarity) => similarity })

//affichage
println(sortedSims.take(10).mkString("\n"))
```

Figure 55: Le script de calcul de similarité

Le script ci-dessus procède à la création d'une fonction de calcul de la similarité Cosinus, qui prend en charge deux vecteurs. Cette fonction sera implémentée, pour calculer les similarités avec l'ensemble des livres de la base de données, puis générer les 10 top notices similaires (proches) à notre notice de départ. Finalement, le résultat obtenu sera stocké dans le fichier relatif aux similarités.

3ème étape : Présentation des recommandations

Après le stockage des résultats du calcul de similarité, l'étape suivante est de présenter ces résultats sous formes de recommandations à l'utilisateur.

Figure 56 : La présentation des recommandations pour un utilisateur anonyme

L'utilisateur de notre structure (IMIST) bénéficie des top 5 recommandations que le système lui présente, en se basant sur les notices qu'il consulte.

Nous soulignons que l'intégration du prototype de la recommandation dans le Système Intégré de Gestion de Bibliothèque (SIGB) de l'IMIST n'a pas été possible en dimension réelle pour des raisons administratives relatives à l'aval des structures décisionnelles.

Conclusion

Au terme de ce chapitre, nous avons présenté l'architecture de notre solution composée d'un dispositif de stockage distribué des données, en l'occurrence Hadoop et d'un dispositif de traitement déployé dans une librairie des algorithmes de Machine Learning, soit la librairie Mllib de Spark, pour charger les données et construire un modèle de recommandation.

Le prototype de recommandation repose sur la méthode de factorisation matricielle du filtrage collaboratif, qui consiste à travers le modèle ALS, soit de partir des résultats retournés par le système et suggérer des notices de documents similaires pour le cas d'un utilisateur anonyme, soit de construire le profil d'un utilisateur, qui prend la forme d'un modèle que le système sollicite à chaque requête pour le cas d'un utilisateur authentifié.

Chapitre 5 : Evaluation du prototype de SR proposé

Introduction

Après avoir mis en œuvre notre prototype, nous avons envisagé son évaluation par différentes méthodes, par le biais de fonctions mesurant la pertinence et la précision de ces systèmes au moyen du calcul de l'écart moyen entre les prévisions (du système) et les observations correspondantes.

Puis, nous allons évaluer notre modèle et son prototype en mode hors ligne à travers le calcul des fonctions d'erreur et de mesure de pertinence et en ligne par la comparaison des données relatives aux utilisateurs avant et après le test du prototype.

Finalement, une évaluation explicite à travers un questionnaire couvrant un échantillon des utilisateurs disponibles et ayant accepté de collaborer pour ce test. Ils seront appelés à répondre à deux questions explicites sur la performance globale du prototype.

1. L'évaluation hors ligne : le calcul des fonctions d'erreur (RMSE, MSE)

Pour savoir si notre modèle est pertinent en termes des recommandations faites à l'échantillon des utilisateurs chercheurs de l'IMIST, nous devons évaluer et juger ses performances prédictives et son alignement sur les besoins.

Les fonctions d'erreur représentent les métriques les plus utilisées en matière d'évaluation des systèmes de recommandation. Ils permettent de mesurer l'écart moyen entre les prévisions estimées et les observations correspondantes (voir le chapitre 1 : modélisation des systèmes de recommandation de la troisième partie). De cette manière, une prédiction est jugée idoine, du moment où la valeur de la fonction est proche de 0.

Les fonctions les plus courantes dans l'évaluation des systèmes de recommandations et que nous pouvons vérifier sur Spark sont :

- **La fonction d'erreur moyenne quadratique (MSE) « Mean Squared Error »** : elle calcule la moyenne arithmétique des écarts et des carrés entre les prévisions et les observations ;
- **La fonction RMSE « Root Mean Squared Error » :** elle représente la racine de l'erreur moyenne quadratique.

Il est à rappeler que L'Erreur Quadratique Moyenne (RMSE : Root Mean Square Error) met l'accent sur les plus grands écarts entre les scores prédits et les scores réels, elle sera calculée automatiquement par notre algorithme sur Spark. Elle est définie ainsi :

$$|\bar{E}| = \frac{\sum_{(c,s)\in W} |u^p(c,s) - u(c,s)|}{|W|}$$

Sachant que :

- U (c, s) est le score réel ;
- Up (c, s) est le score prédit par le SR ;
- W = {(c, s)} est la paire *usager-objet* avec laquelle le SR fait la prédiction.

Pour calculer ces fonctions pour notre solution, nous avons exécuté un algorithme qui charge, de prime abord, l'ensemble des données et calcule, ensuite, les 10 top de recommandations pour chaque utilisateur, puis il procède à des comparaisons, en vue d'estimer les valeurs de ces fonctions d'évaluation à partir de l'historique des tests, soient les requêtes déjà formulées et pour lesquelles des recommandations ont déjà été retournées.

Ci-dessous, un aperçu de l'algorithme qui calcule les fonctions d'erreur :

```
// La précision moyenne en K
println(s"Mean average precision = ${metrics.meanAveragePrecision}")
// Gain cumulatif
Array(1, 10).foreach { k =>
  println(s"NDCG at $k = ${metrics.ndcgAt(k)}")
}

// Calcul de prédictions pour chaque point
val allPredictions = model.predict(ratings.map(r => (r.user, r.product))).map(r => ((r.user,
  r.product), r.rating))
val allRatings = ratings.map(r => ((r.user, r.product), r.rating))
val predictionsAndLabels = allPredictions.join(allRatings).map { case ((user, product),
(predicted, actual)) => (predicted, actual)}

 // RMSE
val regressionMetrics = new RegressionMetrics(predictionsAndLabels)
println(s"RMSE = ${regressionMetrics.rootMeanSquaredError}")
// R-squared
println(s"R-squared = ${regressionMetrics.r2}")
```

Figure 57: L'algorithme d'évaluation, lors du chargement des données et du modèle

La figure 57 présente le chargement des bibliothèques à utiliser, le fichier des données et le modèle utilisé. De fait, il effectue des comparaisons entre les 10 top prédictions pour chaque utilisateur, pour pouvoir calculer les fonctions d'erreur et afficher les résultats.

Après plusieurs tests, les valeurs des paramètres de ce modèle optimal, ainsi que le résultat de la racine de l'erreur moyenne quadratique (RMSE) correspondant sont présentés dans le tableau 24 :

Paramètres du modèle	RMSE
Itérations=110 Rang=10 Lambda=0.0.1 Alpha=1	2.712

Tableau 24: Les valeurs des fonctions d'évaluation

Ci-dessous la valeur de la fonction d'erreur RMSE obtenue après le calcul de la précision moyenne qui mesure les scores de pertinence moyenne d'un ensemble de documents top-K recommandés pour une requête de test :

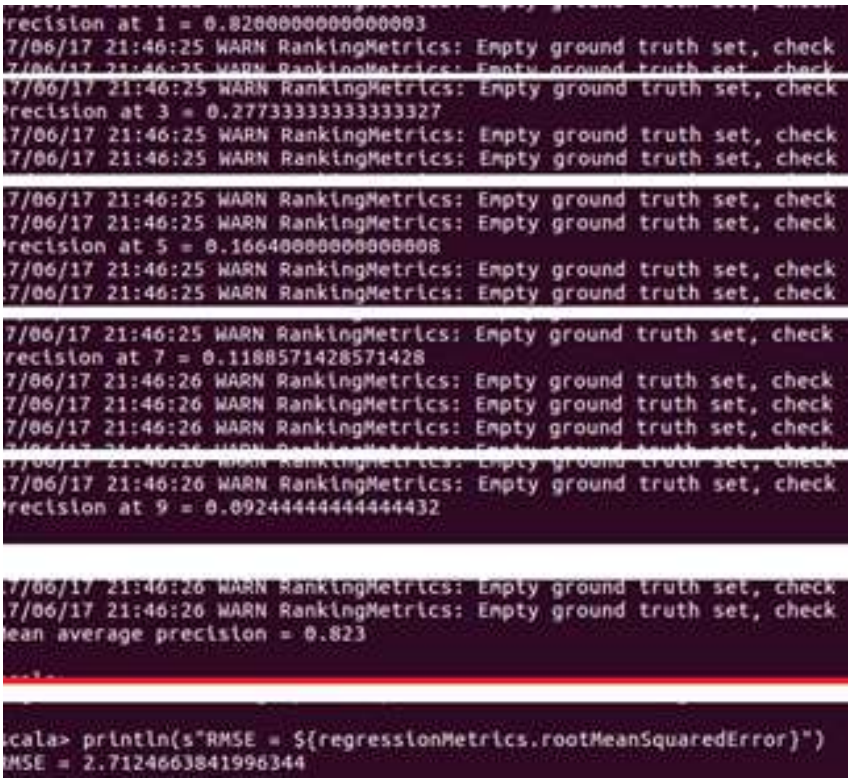

Figure 58: Les valeurs des mesures d'évaluation du système

La fonction RMSE (en bas du code) n'est pas significative dans le cas des données implicites liées au comportement en ligne de l'utilisateur et non pas ses notations explicites et la marge d'erreur de notre système est minime.

Nous pouvons en conclure que le système retourne des recommandations pertinentes à nos utilisateurs chercheurs de la structure documentaire de l'IMIST.

2. Le calcul du temps d'exécution du programme

Un autre critère qui peut nous servir à évaluer la performance de notre système est de connaître le temps d'exécution du script relatif aux recommandations.

A ce titre, nous avons procédé, par plusieurs tests d'exécution d'une requête sanctionnée par des recommandations sur le catalogue en ligne, au calcul du temps d'exécution et nous avons obtenu les résultats suivants :

Test N°	Durée d'exécution (en secondes)
1	6.64
2	6.40
3	6.23
4	6.12
5	6.37
6	5.83
7	5.48
8	5.59
9	5.71
10	5.20

Tableau 25: Les valeurs des tests de durées d'exécution du script de recommandation sur Spark

La moyenne des valeurs générées par le test est égale à 5.95 secondes, donc notre système fournit les recommandations dans un laps de temps estimé environ à 6 secondes. Cette durée peut être jugée comme acceptable pour un utilisateur.

3. L'évaluation en ligne implicite

En coordination avec quelques responsables de la structure de l'IMIST, nous avons pu adresser un message à certains utilisateurs joignables, pour les inviter à prendre part à un test du prototype de recommandation en leur précisant qu'il n'était pas hébergé sur un serveur donné, mais installé en local mais qu'il pourrait être implémenté par la suite sur leur catalogue en ligne.

Ainsi, un test a été effectué en ligne durant 40 minutes à travers des requêtes formulées par 102 utilisateurs. Les données *log* recueillis, avant et après le test du prototype sont présentées dans le tableau 26 :

Les données logs extraites avant le test de notre protype de recommandation pour 13 usagers			
ID de l'usager	Nombre de rebond de la requête	le temps passé sur l'OPAC	Clics pour affichage de la notice des documents
144	3 fois	15 Min	9 fois
201	3 fois	16 Min	8 fois
112	2 fois	65 Min	12 fois
467	2 fois	18 Min	6 fois
123	2 fois	29 Min	9 fois
134	4 fois	15 Min	4 fois
377	1 fois	12 Min	5 fois
441	1 fois	9 Min	1 fois
53	2 fois	23 Min	8 fois
15	2 fois	13 Min	12 fois
392	8 fois	25 Min	2 fois
117	4 fois	23 Min	3 fois
215	5 fois	15 Min	5 fois
520	2 fois	17 Min	3 fois

Les données logs extraites après le test de notre protype de recommandation pour 13 usagers			
ID de l'usager	Nombre de rebond de la requête	le temps passé sur l'OPAC	Clics pour affichage de la notice des documents
144	2 fois	2 Min	2 fois
201	1 fois	5 Min	2 fois
112	2 fois	23 Min	5 fois
467	3 fois	12 Min	3 fois
123	1 fois	9 Min	4 fois
134	2 fois	15 Min	4 fois
377	3 fois	17 Min	6 fois
441	2 fois	9 Min	2 fois
53	3 fois	14 Min	3 fois
15	2 fois	7 Min	2 fois
392	3 fois	12 Min	2 fois
117	2 fois	4 Min	1 fois
215	2 fois	7 Min	3 fois
520	3 fois	3 Min	2 fois

Tableau 26 : Comparaison entre les données log d'usage des ressources documentaires avant et après le test du prototype de recommandation

Le tableau 26 montre le changement des données relatives aux critères d'usage retenus dans notre modèle (le taux de rebond de la requête, le temps passé sur l'OPAC et le nombre d'affichage des détails d'une notice relative à un document).

Les valeurs sont plus grandes avant d'intégrer le prototype de notre solution. A titre d'exemple, avant le test de notre modèle de recommandation, les valeurs extrêmes maximales des trois paramètres de données, correspondent respectivement et dans l'ordre à (8 fois, 65 min et 12 fois). Alors qu'après le test du prototype, les nouvelles valeurs maximales de ces paramètres sont dans l'ordre (3 fois, 23 min et 6 fois).

S’agissant des moyennes des paramètres relatifs aux données *log*, la moyenne du taux de rebond est passée de 2.92 fois sur le catalogue en ligne à 2.21 fois après avoir testé notre solution, le temps passé sur l’OPAC s’est réduit, quant à lui, de 21,07 minutes à 9,92 minutes, et le nombre de clics pour afficher la notice des documents figurant dans les résultats a reculé à son tour, en passant de 6,21 fois à seulement 2.92 fois.

Nous déduisons que le prototype de notre solution de recommandation pourrait faire face au problème de la surcharge informationnelle et cognitive des utilisateurs d’une structure documentaire dédiée à la recherche, ce qui devrait leur faciliter l’accès, la sélection et l’exploitation optimale des ressources documentaires.

4. L’évaluation explicite des utilisateurs par un questionnaire

Afin de recueillir les feedbacks directs des utilisateurs ayant pris part à cette séance de test, et mesurer le degré de leur satisfaction vis-à-vis du prototype de la solution proposée, nous avons aussi administré un petit questionnaire auprès d’eux **(voir Annexe 2 : Mini-questionnaire sur la satisfaction d’un échantillon des utilisateurs du prototype de recommandation).**

En effet, 102 utilisateurs ont fait un retour au message adressé et ont, par conséquent, répondu présent à cette séance d’évaluation, où ils ont effectué des requêtes d’interrogation du catalogue en ligne avec le prototype de recommandations, un petit-questionnaire sur la satisfaction leur a été soumis à l’issue de ce test.

Le mini-questionnaire a mis en évidence l’apport de notre modèle et prototype de recommandation au catalogue en ligne selon une échelle de notation, allant de faible à excellent, en passant par moyen et très bien.

La figure 59 illustre les retours de nos utilisateurs :

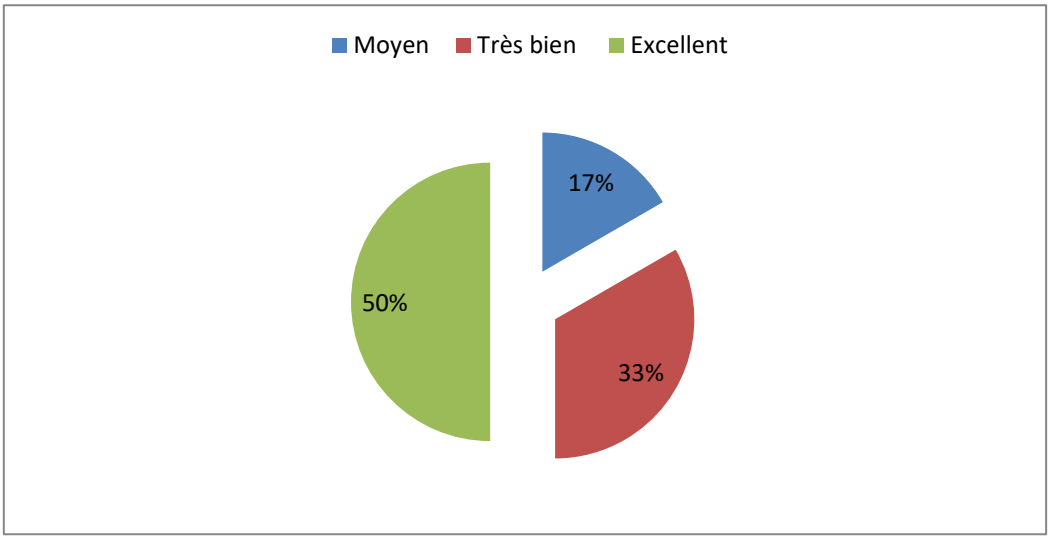

Figure 59: Appréciation des utilisateurs sur l'apport général du prototype

La majorité des utilisateurs interrogée, soit 83%, a jugé que notre prototype de recommandation est excellent (50%) ou très bien (33%). Tandis que, 17% seulement ont jugé que la solution était moyenne, alors qu'aucun utilisateur n'a estimé qu'elle était faible.

Les résultats satisfaisants obtenus plaident en faveur de l'incorporation de la solution dans le catalogue en ligne, ce qui aidera à l'automatisation du processus de génération de recommandations.

Quant à la satisfaction de nos utilisateurs par rapport aux recommandations retournées à l'issue de leurs requêtes, les feedbacks sont résumés dans la figure 60 :

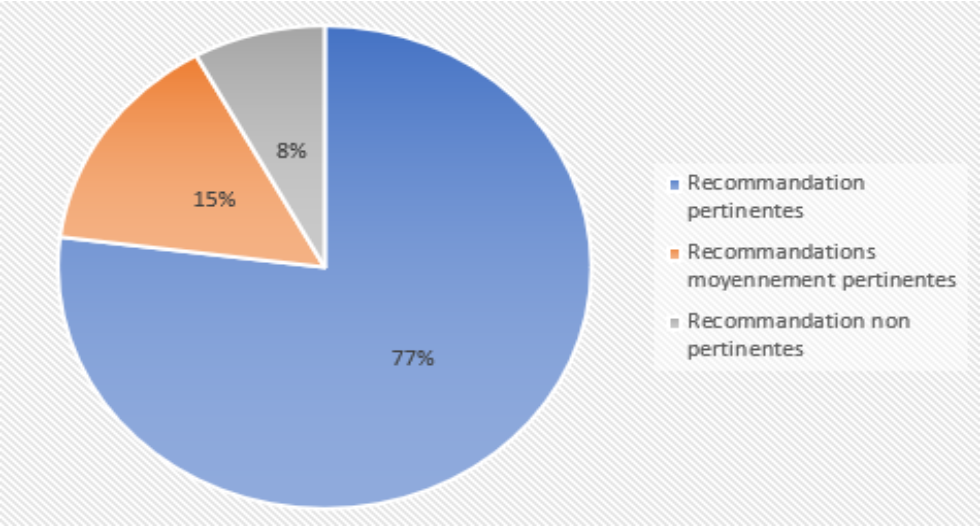

Figure 60 : Les feedbacks des utilisateurs par rapport à la pertinence des recommandations

Les résultats témoignent de la pertinence de la solution de recommandation proposée, puisque 77% des utilisateurs considèrent que leurs requêtes ont été sanctionnées par des recommandations fructueuses, et qui répondent à leur besoin initial. Seulement 15% jugent que les recommandations sont moyennement pertinentes et 8% estiment que les résultats retournés ne sont pas pertinents.

Il est à noter que, la recommandation dans un catalogue en ligne est tributaire des ressources disponibles dans le service d'information documentaire. Ainsi la pénurie de documents dans une thématique donnée risque inévitablement de conduire l'utilisateur au silence documentaire.

En somme, plus les documents tiennent compte de manière évolutive des besoins des utilisateurs, notamment des plus actifs et, plus la solution proposée pourra s'appuyer sur des données d'usage, plus elle sera en mesure de proposer des ressources documentaires pertinentes.

Conclusion

L'évaluation de notre prototype de recommandation a revêtu trois formes principales :

- Une évaluation hors ligne par le calcul de l'écart moyen entre les prévisions (du système) et les observations correspondantes ;
- Une évaluation en ligne dont l'objectif est d'observer, en temps réel, le comportement des utilisateurs et comparer ces données *log* avant et après le test du prototype ;
- Une évaluation explicite à travers un questionnaire de satisfaction, couvrant un échantillon des utilisateurs disponibles et ayant accepté de collaborer pour ce test.

Au terme de l'évaluation, le modèle de recommandation s'est avéré assez performant, aussi bien par l'intermédiaire des résultats du calcul des fonctions d'erreur que par le biais des tests d'usage et des retours de satisfaction des utilisateurs.

Limites du modèle et prototype proposés

Notre modèle de recommandation proposé présente certaines limitations que nous voudrions désormais pointer :

- Il s'agit d'un modèle de recommandation orienté vers les utilisateurs d'un groupe homogène, soient les chercheurs regroupés par classe disciplinaire de la CDD, ce qui risque d'appauvrir le champ de découverte des utilisateurs, les cantonnant dans leur spécialisation ;

- Ce modèle s'appuie sur des algorithmes intégrés dans des libraires Open Source et déployés localement sur les machines, ce qui laisse penser au risque de latence (temps de réponse) et conduire à des problèmes techniques de surcharge (*scalability* ou la mise à échelle) avec l'évolution du nombre des utilisateurs et des documents sur lesquels porte le modèle ;

- Ce modèle se focalise sur l'approche de recommandation de filtrage collaboratif, qui pourrait être enrichie par l'approche axée sur le contenu, notamment dans le cas d'autres types de ressources que les ouvrages ;

- Ce modèle exploite des données *log* implicites, qui ne sont pas exploitables directement par les algorithmes de recommandation. Ainsi nous étions dans l'obligation de les convertir en données explicites à travers une formule générale exprimant l'intérêt d'un utilisateur et calculée à partir des paramètres disponibles dans la base de connaissance disponible et auxquels une pondération a été attribuée suivant leur lien avec l'intérêt de l'utilisateur. Toutefois, nous avons utilisé les paramètres de confiance de l'algorithme, pour mesurer la pertinence de ces données d'entrée du modèle ;

- Notre modèle s'est appuyé sur un l'algorithme ALS de factorisation matricielle fourni par la librairie Mllib de Spark disponible en open source et s'inscrivant dans un environnement informatique stable ;

- Ce modèle fonctionne uniquement sur les ressources documentaires et informationnelles de type monographies ou ouvrages, sans s'étendre aux articles des revues scientifiques.

Conclusion générale et perspectives

Les systèmes de recommandation revêtent une grande importance dans les différentes plateformes et applications web, pour aider les utilisateurs à sélectionner les items et produits qui correspondent à leurs besoins.

Au terme de notre thèse, nous rappelons que notre recherche s'est inscrite dans l'optique d'apporter une réponse au problème de la surcharge cognitive, qui a pour corollaire la difficulté de prendre une décision en un temps opportun, quant au choix des ressources documentaires offertes à une communauté d'utilisateurs et pour lesquels les références ne sont qu'une première étape d'une démarche de recherche. Ainsi, notre terrain empirique était l'Institut Marocain de l'Information Scientifique et Technique (IMIST).

A cet égard, nous avons été conduits, tout d'abord, à dresser un état de l'art sur la place de la data dans les industries culturelles et du contenu, et plus particulièrement pour les Services d'Information Documentaires (SID), pour dessiner le contexte de notre thèse, avant de traiter des approches, des algorithmes et du fonctionnement des systèmes de recommandation, pour concevoir, modéliser et mettre en œuvre un prototype de recommandation. Le système proposé exploite les données *log* implicites des utilisateurs de la structure documentaire précitée, afin de personnaliser leur navigation et leur exploration des ressources documentaires sur le catalogue en ligne, en sélectionnant que les documents susceptibles de les intéresser selon leurs profils et leurs centres d'intérêt scientifiques.

Nous avons mis en relief le rôle de la data dans l'optimisation des services d'information documentaire, aussi bien au niveau interne impliquant les professionnels de l'information et de la documentation, qu'au niveau externe relatif aux services pour les utilisateurs. En considérant un certain périmètre des industries culturelles, nous avons mis en évidence certains usages des données dans les principaux secteurs, ainsi que les apports pour les SID des données structurées et non structurées concernant les services aux utilisateurs des unités documentaires.

Nous avons relaté le cadre théorique relatif aux systèmes de recommandation. Nous avons également analysé le contexte des besoins, avant de concevoir le modèle et le prototype de

recommandation sur le terrain de recherche choisi. Le processus de recommandation, ses approches, ses algorithmes et ses fonctions d'évaluation ont été investigués pour éclairer les choix finalement opérés pour la solution que nous avons proposée.

D'un point de vue technique, nous avons exploré quelques apports de l'apprentissage automatique (*Machine learning*) aux services documentaires. Nous avons également approfondi nos connaissances techniques et théoriques dans le champ du big data et l'analyse des données.

Nous avons vérifié au cours de notre thèse plusieurs constats relevés du terrain, en commençant par les limites des modèles de recherche classiques pour le repérage des ressources documentaires sur le catalogue d'une entité documentaire, et celui de l'utilité de générer des profils des utilisateurs basés sur leurs préférences et centres d'intérêt partagés avec ceux qui leur sont similaires. Nous avons testé l'apport du modèle de recommandation conçu à partir de la classification supervisée et la factorisation matricielle par le biais du modèle des moindres carrés alterné (ALS).

Ce modèle commence par le prétraitement et le nettoyage du jeu de données *log* implicites des utilisateurs, puis la classification des utilisateurs en apprentissage supervisé avec le cadre de la classification décimale de Dewey. Une scénarisation des cas de connexion de l'utilisateur authentifié et anonyme a été distinguée. Une transformation des données implicites d'usage (taux de rebond de requête, le nombre de fois que la notice d'un document a été affichée et le temps passé sur l'OPAC) en une notation globale a été proposée pour mettre en œuvre le prototype de recommandations.

Ce système a été implémenté dans un environnement big data, motivé par le volume qui croit au fil des années, et par l'exigence de vitesse de traitement et de génération des recommandations en temps réel. L'environnement s'est composé d'Hadoop pour le stockage distribué des données et d'Apache Spark pour la création du modèle de recommandation basé sur l'algorithme de recommandation (ALS) de factorisation matricielle de la librairie (Mllib) écrit en langage Scala, qui ont permis de procéder à la génération de recommandations en temps réel sur une interface d'interrogation similaire au catalogue en ligne.

Finalement, ce prototype a fait l'objet d'une évaluation en ligne, hors ligne et par expérimentation sur un échantillon d'utilisateurs.

Si notre thèse nous a permis d'obtenir des résultats satisfaisants, plusieurs perspectives et améliorations restent envisageables :

- Il serait important d'étendre le modèle de recommandation développé pour qu'il concerne aussi d'autres dimensions de l'information scientifique et technique, en l'occurrence les articles de revues scientifiques et ceci en trouvant un accord avec les fournisseurs des abonnements des périodiques, pour fédérer l'accès aux différents numéros de revues à partir du catalogue en ligne. Le service d'information documentaire pourrait utiliser notre modèle pour recommander les informations bibliographiques des articles de revues ;

- Il serait également possible d'automatiser une partie du processus de transformation des données implicites en données explicites ;

- Il serait aussi envisageable d'intégrer un dispositif, qui permettrait de tirer parti des données accumulées par le système de recommandation, en vue de générer des statistiques sur la consultation des ressources documentaires, connaitre les thématiques les plus privilégiées par les utilisateurs et éclairer ainsi la politique d'acquisition ;

- Analyser les effets d'enfermement des utilisateurs dans leurs spécialisations initiales sans ouverture réelle à cause des algorithmes de similarité, formant une bulle de filtre dans laquelle ils sont isolés les uns des autres avec la proposition des contenus qui les intéressent déjà, ce qui conduit à l'appauvrissement de leur champ de découverte culturelle[159] ;

[159] GUEBELS, Gilles. *Le phénomène des bulles de filtres sur Internet : le moteur de recherche Google nous oriente-t-il à notre insu à cause de son algorithme de personnalisation ?* [En ligne]. Faculté des sciences économiques, sociales, politiques et de communication, Université catholique de Louvain, 2018. Disponible sur : file:///C:/Users/hp/Downloads/GUEBELS_20671300_2018.pdf [consulté le 30 Mars 2021].

- Pour terminer, les services d'information documentaire numériques peuvent s'ouvrir également sur d'autres opportunités relatives à la fouille des données (*data et text mining*) et leur apport aux chercheurs reste à exploiter pour l'exploration de la littérature scientifique et technique.

L'emploi des techniques de fouille de données sur les collections commence à offrir de nouveaux instruments de recherche pour l'analyse documentaire. Toutefois, les questions juridiques sont aussi soulevées, quant au croisement de la data avec les services d'information documentaire, notamment en rapport avec le droit de la propriété intellectuelle et le droit de protection des données personnelles.

Les algorithmes de l'intelligence artificielle rythment nos interactions quotidiennes, ce qui interroge de plus en plus leur respect aux valeurs humaines et la confiance des utilisateurs. La question est vive et animera très certainement les prochaines années, ainsi le récent rapport de la commission européenne intitulé « De nouvelles règles et actions en faveur de l'excellence et de la confiance dans l'intelligence artificielle » [160] pointe les risques inhérents à l'intelligence artificielle et les obligations strictes auxquelles devraient être soumis les systèmes de l'IA, notamment en termes de qualité des données et de contrôle humain approprié.

Ainsi, ce rapport de la commission européenne prône l'excellence dans le domaine de l'intelligence artificielle, centrée sur l'humain, durable, sûre, inclusive et digne de confiance, tout en encourageant l'innovation et l'utilisation des technologies de l'IA dans l'ensemble des secteurs économiques et dans les États membres.

Cette ligne directrice affiche une identité particulière de l'IA européenne par rapport à d'autres périmètres géopolitiques au niveau mondial qu'il faudra particulièrement suivre dans les prochaines années.

[160] SITE DE L'UNION EUROPEENNE. *De nouvelles règles et actions en faveur de l'excellence et de la confiance dans l'intelligence artificielle* [En ligne]. Disponible sur : https://ec.europa.eu/france/news/20210421/nouvelles_regles_europeennes_intelligence_artificielle_fr [consulté le 26 Avril 2021].

Publications relatives à la thèse

Notre thèse a donné lieu à différentes publications (articles, communications, posters, etc.), que nous présentons ci-dessous :

- **Articles dans des revues internationales ou nationales avec comité de lecture**

  - SENNOUNI, Amine et CHERIF, Walid. *Adaptive user-product recommendation system using supervised and unsupervised classifications models*. In: Proceedings of the 3rd International Conference on Information Technology & Electrical Engineering (Scopus-ACM) - El Jadida (Maroc), November 25-26, 2019.
  - SENNOUNI, Amine. La Data au service de l'innovation dans les Services d'Information Documentaires (SID) universitaires nationaux. *Revue française des sciences de l'information et de la communication* [En ligne], 10 | 2016, mis en ligne le 01 janvier 2017. Disponible sur : http://rfsic.revues.org/2759 [consulté le 29 janvier 2017].
  - SENNOUNI, Amine ; SBIHI, Boubker [et al.]. The Role of Open Data in Making Strategic Decisions in the Telecommunication Sector in Morocco. *International Journal of Research Studies in Science, Engineering and Technology*. 2015, Vol.2, Issue 6, pp 42-51.

- **Articles publiés dans les actes des colloques / congrès scientifiques**

- SENNOUNI, Amine et BACHR, Ahmed Abdelilah. *Accès aux données ouvertes de la recherche : vers l'application d'un système de recommandation optimisant le repérage de jeux de données libres* In : Actes du 3e colloque international sur le Libre Accès – Rabat (Maroc), 28-30 novembre 2018, p.342.
- SENNOUNI, Amine. *The Data supporting innovation in the national academic documentary information services*. In: International conference on computing, wireless and communication systems (ICCWCS 2016), 15-16 November 2016, 5p.
- SENNOUNI, Amine. *Towards the integration of open data in decision making in the big enterprises: the case of the telecom sector in Morocco.* In: Conférence Internationale sous le thème "Beacons of Hope in the Quest for the Next Einstein in the MENA region"- la session "Big data, Business Intelligence, Data Mining and analytics", Fès 3-6 Mars 2015, 11p.

- **Des communications et posters en marge des séminaires, des congrès et des colloques**

- SENNOUNI, Amine. *Data Sciences : l'art de prédire ses activités clés : l'exemple des systèmes de recommandation*. In : Developers Festival Google Maroc-Rabat 2017 (DEVFEST), INPT-Rabat, 09 Décembre 2017,14 p.
- SENNOUNI, Amine. *The modeling of a recommender System intended for researchers using a documentary service. the case of Moroccan Institute of Scientific and Technical Information (IMIST).* In : 61 ISI World Statistics Congress, Marrakech, 16-21 Juillet 2017, 10 p.
- SENNOUNI, Amine. *Big data and Digital Economy: What Role for Data Scientist? (Poster).* In : Congrès Méditerranéen des Télécommunications, Tetouan 12-13 Mai 2016, 8p. ;
- SENNOUNI, Amine. *Towards the validation of the theoretical pedagogical approaches through Experimentation of Big data systems*. In: Conférence Internationale sous le thème "Beacons of Hope in the Quest for the Next Einstein in the MENA region"- la session "Big data, Business Intelligence, Data Mining and analytics", Fés 3-6 Mars 2015, 14p.

Bibliographie

ACM Digital Library. *Rubrique Journals* [En ligne]. Disponible sur : https://dl.acm.org/journals [consulté le 06 Avril 2020].

ADOMAVICIUS, Gediminas et KWON, Youngok. New recommendation techniques for multicriteria rating systems. *IEEE Intelligent Systems*, 2007.

AGGARWAL, Charu C. *Recommender Systems*, Springer International Publishing, 2016. 498 p.

ALA-FOSSI, marko., et al. The impact of the Internet on business models in the media industries. *The Internet and the Mass Media*. Los Angeles/London: SAGE, 2008, p.149–169.

ALEXANDER, Peter J., Peer-to-Peer File Sharing: The Case of the Music Recording Industry. *Review of Industrial Organization*. 2002, Vol. 20, n°. 2, p. 151-161.

AMATRIAIN, Xavier et al. *Data mining methods for recommender systems*. Ricci. Chapter 2, p. 39-71.

ANAND, S. et B, MOBASHER. Intelligent techniques for web personalization. In: Intelligent Techniques for Web Personalization. *Springer*. 2005, p. 1-36.

Apache Spark™ - Lightning-Fast Cluster Computing [En ligne]. Disponible sur http://spark.apache.org/ [Consulté le 12 juin 2017].

BAKER, Thomas et al. *Library Linked Data Incubator Group Final Report* [En ligne]. W3C. Disponible sur : http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/ [Consulté le 19 juin 2014].

BARNIER, Gerard. *Théories de l'apprentissage et pratiques d'enseignement*. Rapport de conférence. Nice, 2002.

BEER, David. Making Friends with Jarvis Cocker: Music Culture in the Context of Web 2.0. *Cultural Sociology*. 2008, vol. 2, n° 2, p. 222-241.

BERMES, Emmanuelle. *Le web sémantique en bibliothèque*. Ed. Du Cercle de la librairie, 2011. Collection bibliothèques. 176 p.

BESSE, Philippe et VIALANEIX, Nathalie. *Statistique et Big Data Analytics ; Volumétrie, L'Attaque des Clones* [en ligne], 2014. Disponible sur : https://hal.archives-ouvertes.fr/hal-00995801v3 [consulté le 06 Octobre 2016].

BEYERS, William, el al. *The Economic Impact of Seattle's Music Industry [En ligne]*. Office of Economic Development, 2004. Disponible sur : https://www.seattle.gov/Documents/Departments/FilmAndMusic/Seattle_Music_EIS_2008.pdf [Consulté le 23 Juin 2016].

BISAILLON, Jean-Robert. *Métadonnées et répertoire musical québécois : un essai de mobilisation des connaissances dans le nouvel environnement numérique* [En ligne]. Mémoire de maîtrise. Montréal : Institut national de la recherche scientifique, 2013. Disponible sur : http://espace.inrs.ca/id/eprint/1678/ [consulté le 22 Avril 2016].

BOLING, Elizabeth et al. Perspectives in Transition. *Educational Technology*. Vol. 51, n° 5, 2011, p. 34-38.

BOUKACEM-ZEGHMOURI, Chérifa, SCHÖPFEL, Joachim. Statistiques d'utilisation des ressources électroniques. *Bulletin des bibliothèques de France* [en ligne]. 2005, n° 4. Disponible sur : http://bbf.enssib.fr/consulter/bbf-2005-04-0062-001 [Consulté le 25 Janvier 2015].

BOUQUILLON, Philippe. Industries, économie créatives et technologies d'information et de communication. TIC *& Société*. 2010, Vol. 4, n°2, p. 7-40.

BOURREAU, Marc et GENSOLLEN, Michel. « L'impact d'Internet et des Technologies de l'Information et de la Communication sur l'industrie de la musique enregistrée ». *Revue d'économie industrielle* [en ligne], n°116, 2006/4, p.2. Disponible sur : https://www.cairn.info/revue-d-economie-industrielle-2006-4-page-2.htm Cconsulté le 23 Août 2020].

BOUSTANY, Joumana. « Données massives, science et bibliothèques », Michel Netzer éd., *Les sciences en bibliothèque* [en ligne]. Éditions du Cercle de la Librairie, 2017, pp. 191-200. Disponible sur : https://www.cairn.info/les-sciences-en-bibliotheque--9782765415251-page-191.htm [consulté le 12 Août 2019].

BRASSEUR, Christophe. *Enjeux et usages du big data : technologies, méthodes et mises en œuvre*. Hermès sciences publications, 2013. 203 p.

BROUDOUX, Evelyne et SCOPSI, Claire. Métadonnées sur le web : les enjeux autour des techniques d'enrichissement des contenus. *Etudes de communication*, n°. 36, 2011.

BURKE, Robin. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 2002, vol. 12, no 4, p. 331-370.

BUSTAMENTE, E. Cultural Industries in the Digital Age: Some Provisional Conclusions. Media, Culture and Society.2004, Vol. 26, n°6, p. 803-820.

BYRNE, Gillian et Lisa, GODDARD.The Strongest Link: Libraries and Linked Data. *D-Lib Magazine* [en ligne]. Vol. 16, n° 11/12, novembre 2010. Disponible sur http://www.dlib.org/dlib/november10/byrne/11byrne.html [Consulté le 19 juin 2014].

CALDERAN, Lisette et LAURENT, Pascale. *Big data : nouvelles partitions de l'information / Actes du séminaire IST Inria, octobre 2014.* Bruxelles : de Boeck, 2015.

CALLAN, Jamie et al. Personalisation and Recommender Systems in Digital Libraries Joint NSF-EU DELOS Working Group Report, 2003.

Capgemini. Service Amazon Personalize [En ligne]. Disponible sur : https://www.capgemini.com/service/amazon-personalize/ [Consulté le 05 Mai 2021].

CHANTEPIE, Philippe et, LE DIBERDER, Alain. *Révolution numérique et industries culturelles*. 2ed., Paris : La Découverte, 2010.

CHANTEPIE, Philippe ; LE DIBERDER, Alain. « IV. L'exploitation numérique : tout change », Philippe Chantepie éd., *Révolution numérique et industries culturelles.* La Découverte, 2010, pp. 55-72.

CHARTRON, Ghislaine et REBILLARD, Franck. *Modèles de publication sur le web*. Rapport d'activités ASCNRS 103, Jul 2004.

CHARTRON, Ghislaine. Cours l'économie de l'information, les différentes filières des industries culturelles, transformation numérique, 2019.

CHARTRON, Ghislaine. Edition et publication des contenus : regard transversal sur la transformation des modèles. Lisette Calderan, Pascale Laurent, Hélène Lowinger, Jacques Millet. Publier, éditer, éditorialiser, nouveaux enjeux de la production numérique, De Boeck Supérieur, pp.9-36, 2016, Information & Stratégie. Diponible sur https://halshs.archives-ouvertes.fr/halshs-01522295/document [Consulté le 06 Août 2020].

CHARTRON, Ghislaine. Tendances lourdes et tensions pour les filières du document numérique. *Le "Document" à l'ère de la différenciation numérique*, Dec 2011, Maroc.

CHLOE, L. L'impact de la data sur les métiers des industries culturelles et créatives. Retour sur l'étude de Cap Digital [en ligne], 2017. Disponible surhttps://mbamci.com/data-metiers-culturels-et-creatifs/[Consulté le 24 Août 2019].

CNIL (Commission Nationale de l'Informatique et des Libertés). Industries créatives, contenus numériques et données : les contenus culturels vus au travers du prisme des donnéesLe graal de la recommandation et de la personnalisation. *CAHIERS IPINNOVATION & PROSPECTIVE* [en ligne], n°3 ,2015. Disponible sur : https://www.cnil.fr/fr/innovation-prospective/publications [Consulté le 25 Juin 2016].

COLLET-THIREAU, Katell et THOMAS, Jean-Paul, « Big Data et Open Data : quel impact pour les professionnels de l'information ? », *I2D – Information, données & documents* [en ligne], 2015/4 (Volume 53), p. 9-10. Disponible sur : https://www.cairn-int.info/revue-i2d-information-donnees-et-documents-2015-4-page-9.htm [Consulté le 24 Septembre 2020].

COMMISSION NATIONALE DE L'INFORMATIQUE ET DES LIBERTES. *Délibération n°99-27 du 22 avril 1999 concernant les traitements automatisés d'informations nominatives relatifs à la gestion des prêts de livres, de supports audiovisuels et d'œuvres artistiques et à la gestion des consultations de documents d'archives publiques* [En ligne], 1999. Disponible sur : https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000197641 [consulté le 15 Septembre 2019].

CORINUS, Mickael ; DERY, Thomas. *Accès rapport d'étude sur le big data [en ligne]. Rapport d'étude sur le big data*. SRS Day, 2012. Disponible sur :<https://srs.epita.fr/wp-content/uploads/2014/02/big-data-rapport.pdf> [Consulté le 15 Septembre 2019].

CORINUS, Mickael et Thomas, DERY. Rapport d'étude sur le big data [en ligne]. SRS Day, 2012. Disponible sur :<https://srs.epita.fr/wp-content/uploads/2014/02/big-data-rapport.pdf> [Consulté le 15 Septembre 2014].

COUTAZ, Joelle. Essai sur l'interaction Homme-Machine et son évolution. *Bulletin de la société informatique de France*. Septembre 2013, n°1, p. 15-33.

CULTURE RP. Les enjeux du data-journalisme pour la Presse : Part II. Data Driven Journalism Roundtable [en ligne]. Octobre 2013. Disponible sur http://culture-rp.com/2013/10/17/les-enjeux-du-data-journalisme-pour-la-presse-part-ii/ [consulté le 20 Juin 2015].

DA SYLVA, Lyne. Les données et leurs impacts théoriques et pratiques sur les professionnels de l'information. *Documentation et bibliothèques*, Octobre–Décembre 2017, p. 5–34.

DE NART, Dario et TASSO, Carlo. A personalized concept-driven recommender system for scientific libraries. *Procedia Computer Science*, 2014, vol. 38, p. 84-91.

Définition du datamining [En ligne]. Disponible sur : https://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es [Consulté le 22/09/2020].

DELESPIERRE, Louis. *Les services personnalisés aux publics en bibliothèque universitaire, une exigence d'innovation et de transformation : l'exemple des services aux chercheurs*. Mémoire d'étude. ENSSIB, Mars 2019.

DELORT, Pierre. *« Big data : concepts et définition »*, éd., *Le Big data*. Presses Universitaires de France, 2015, pp. 29-47.

DUJOL, Lionel. *CRÉER DES SERVICES INNOVANTS*. Stratégies et répertoires d'action pour les bibliothèques, Partie II. Impact du numérique sur l'offre de la bibliothèque : nouvelles approches professionnelles, 2011.pp 70-80.

DUNSIRE, Gordon el al. Linked Data vocabulary management: infrastructure support, data Integration, and interoperability. *Information standards quarterly*, 2012.Vol. 24, n° 2/3, p. 4 13.

DYMYTROVA, Valentyna. Data journalisme, entre pratique créative innovante et nouvelle médiation experte ? Une analyse conjointe des discours et des productions journalistiques. *XXI Congrès de la SFSIC Création, créativité et médiations*, Juin 2018, Paris, France.

E.LINK, Forrest el al. *Mining and Analyzing Circulation and ILL Data for Informed Collection Devlopment*. [en ligne]. Preprint à paraître dans College & Research Libraries, 2015. Disponible sur : http://crl.acrl.org/content/early/2014/10/20/crl14632.full.pdf [Consulté le 10 Octobre 2015].

EKBI, Hamid ; MATTIOLI, Michael et al**.** Big data, bigger dilemnas: a critical review. *Journal of the Association for Information Science and Technology,* Août 2015, p. 1523-1545.

EKSTRAND, Michael D.; KLUVER, Daniel; HARPER, F. Maxwell et al. Letting users choose recommender algorithms: An experimental study. *Proceedings of the 9th ACM Conference on Recommender Systems*, 2015, p.11-18.

ELAHI, Mehdi, RICCI, Francesco, et RUBENS, Neil. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*, 2016, vol. 20, p. 29-50.

ERDMANN,Christopher. Teaching librarians to be data scientists. *Information outlook* [en ligne]. Mai-Juin 2014. Vol.18, n° 3.

Ernst & Young. *Les secteurs culturels et créatifs européens, générateurs de croissance* [En ligne]. Disponible sur : http://www.creatingeurope.eu/en/wp-content/uploads/2014/11/study-full-fr.pdf [Consulté le 15 Avril 2021].

EVANS, Christophe (dir). *Mener l'enquête : guide des études de publics en bibliothèque*, 2011. Collection la boîte à outils. 160 p.

EX LIBRIS GROUP. *BX Recommender* [en ligne]. Disponible sur :https://www.exlibrisgroup.com/fr/produits/bx-recommender/. [Consulté le 01/09/2019].

EY FRANCE. *Données personnelles et Big data*. Disponible sur :http://www.ey.com/FR/fr/Industries/Media---Entertainment/Comportements culturels-et-donnees-personnelles-au-coeur-du-Big-data [consulté le 22 Avril 2015].

FARIS, Robert etHeacock, REBEKAH.Measuring Internet Activity: a (selective) Review of Methods and Metrics. *Internet Monitor Special Report Series*. 2013, n°.2, 39 p.

FRANCO, Daniele. « Exploiter les données d'usages en bibliothèque : pourquoi faire ? : journée d'étude Enssib – Lyon, 14 janvier 2016 » [en ligne], *Bulletin des bibliothèques de France (BBF)*, 2016, n° 7. Disponible sur : https://bbf.enssib.fr/tour-d-horizon/exploiter-les-donnees-d-usages-en-bibliotheque-pourquoi-faire_65839 [Consulté le 31 Août 2020].

FRANKE, Markus, GEYER-SCHULZ, Andreas, et NEUMANN, Andreas W. Recommender services in scientific digital libraries. *Multimedia services in intelligent environments*. Springer Berlin Heidelberg, 2008. p. 377-417.

FRAOUA, K E., LEBLANC J-M., CHARRAIRE S., CHAMPALLE O., *Information and Communication Science Challenges for Modeling Multifaceted Online Courses*, Human Computer Interaction Orlando, 2019. pp. 141-154.

Figoblog. *Big Data et bibliothèques* [en ligne]. Disponible sur : https://figoblog.org/2015/01/13/big-data-et-bibliotheques/ [Consulté le 19 Juin 2020].

GAILLARD, Julien. *Systèmes de recommendation: adaptation Dynamique et Argumentation* [en ligne]. Thèse de doctorat. Université d'Avignon, 2014. Disponible sur : https://hal.archives-ouvertes.fr/tel-01168476/ [consulté le 12 Mars 2017].

GALLOWAY, S.et S. DUNLOP. A critique of definitions of the cultural and creative industries in public policy. *International Journal of Cultural Policy13*. 2007,17- 31P.

GANDON, Fabien ; FARON-ZUCKER, Catherine et CORBY, Olivier. *Le web sémantique : Comment lier les données et les schémas sur le web?*. Dunod, 2012. 224 p.

GILLIUM, Johann. *Big data et bibliothèques : traitement et analyse informatiques des collections numériques* [en ligne]. Mémoire d'étude. Enssib, 2016. Disponible sur : http://microblogging.infodocs.eu/wp-content/uploads/2016/04/66017-big-data-et-bibliotheques-traitement-et-analyse-informatiques-des-collections-numeriques.pdf [consulté le 22 Septembre 2020].

GOMEZ-URIBE, Carlos A. and HUNT, Neil. 2015. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manage. Inf. Syst*. 6, 4, Article 13 (December 2015). 19 p. DOI: http://dx.doi.org/10.1145/2843948.

GOMEZ-URIBE, Carlos A. et HUNT, Neil. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 2016, vol. 6, no 4, p. 13.

GREENBER, Jane. Metadata Generation: Processes, People and Tools, Bulletin of the American Society for Information Science and Technology. *Music and the Internet*. Décembre/Janvier 2003, Vol. 19, n° 2, p. 217-230.

GUEBELS, Gilles. *Le phénomène des bulles de filtres sur Internet : le moteur de recherche Google nous oriente-t-il à notre insu à cause de son algorithme de personnalisation ?* [en ligne]. Faculté des sciences économiques, sociales, politiques et de communication, Université catholique de Louvain, 2018. Disponible sur : file:///C:/Users/hp/Downloads/GUEBELS_20671300_2018.pdf [consulté le 30 Mars 2021].

GUIBERT, Gérôme ; REBILLARD, Franck et al. *Médias, culture et numérique. Approches socioéconomiques*, Paris, Armand Colin, coll. « Cursus », 2016, 237 p., ISBN : 978-2-200-61454-6.

GUY, Shani and GUNAWARDANA, Asela. Evaluating recommendation systems. Recommender Systems Handbook, *Springer US*, January 2011.

HERLOCKER, Jonathan L., KONSTAN, Joseph A., et RIEDL, John. Explaining collaborative filtering recommendations. *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 2000. p. 241-250.

HOUGARDY, Anne et OGER, Laurence. Une méthode en 4 x 4 pour l'analyse des besoins et la régulation en FAD. *Revue internationale des technologies en pédagogie universitaire*, 2006, vol. 3, no 1, p. 40-48.

HU, Yifan, KOREN, Yehuda, et VOLINSKY, Chris. Collaborative filtering for implicit feedback datasets. In: Data Mining, 2008. *ICDM'08. Eighth IEEE International Conference*, 2008. p. 263-272.

HÜGI, Jasmin et PRONGUÉ, Nicolas. *Le virage Linked Open Data en bibliothèque : étude des pratiques, mise en œuvre, compétences des professionnels* [en ligne]. Haute école de gestion de Genève, 2014. Disponible sur : http://www.ressi.ch/num15/article_100 [Consulté le 24 Août 2020].

HÜGI, Jasmin et PRONGUÉ, Nicolas. *Les bibliothèques face aux Linked Open Data* [en ligne]. Haute école de gestion de Genève, 2014. Disponible sur : http://doc.rero.ch/record/209598/files/M7-2014_memoire_HUGI-PRONGUE.pdf [Consulté le 24 Août 2020].

HUGI, Jasmin. *Développement d'une formation e-learning sur les Linked Open Data dans les bibliothèques* [en ligne]. Genève : Haute école de Gestion, 2003. 105 p. Disponible sur : http://doc.rero.ch/record/232855 [Consulté le 24 Avril 2016].

Information Technology Gartner Glossary. *Definition of big data* [En ligne]. Disponible sur : https://www.gartner.com/en/information-technology/glossary/big-data [Consulté le 25 Mai 2015].

INSTITUT MAROCAIN DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE. *IMIST* [en ligne]. Disponible sur : http://www.imist.ma [consulté le 10 juin 2020].

JAYASHRI, Sonawane ; SHIVANI, Gore ; LAXMI, Duggineni, et al. A Survey on Recommendation System Used in E Commerce. *International Journal of Advanced Research in Computer and Communication Engineering*. 2016, Vol. 5.

KAENEL DE, Isabelle et IRIARTE, Pablo. Les catalogues des bibliothèques : du web invisible au web social (I), *Revue électronique Suisse des sciences de l'information* [en ligne], 2007. Disponible sur : http://www.ressi.ch/num05/article_028 [consulté le 27 Août 2019].

KATZ-GERRO, Tally. Cultural Consumption Research: Review of Methodology, Theory, and Consequence. *Internationational Review of Sociology : revue internationale de sociologie*. 2004, Vol.14, n° 1, pp 11-29.

KELLAM, Lyndaet et Peter, KATHARIN. *Numeric data services and sources for the general reference librarian*. Oxford : Chandos Publishing, 2011.

KEMBELLEC, Gérald ; CHEVALIER, Max et DUDOGNON, Damien. Systèmes de recommandation. *Techniques de l'ingénieur*. Novembre 2015, pp 1-22

KEMBELLEC, Gérald ; CHARTRON, Ghislaine et SALEH, Imad. *Les moteurs et systèmes de recommandation*. Paris : ISTE éditions, 2014, 228 p.

KESSOUS, Emmanuel. *L'attention au monde : Sociologie des données personnelles à l'ère numérique*. Armand Colin, 2012. 320 p.

KOBO. Ebook [en ligne]. Disponible sur : https://www.kobo.com/ [Consulté le 20 Juin 2015].

KONSTAN, Joseph A., MILLER, Bradley N., MALTZ, David, et al. Groupb Lens: applying collaborative filtering to Usenet news. *Communications of the ACM*. 1997, vol. 40, no 3, p. 77-87.

L, Bastien. Édition : le Big data au secours du marché du livre. *Le big data* [en ligne]. *2016.* Disponible sur : https://www.lebigdata.fr/edition-big-data [Consulté le 23 Août 2019].

LAFORCE, P. ; RATTE, S. 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). 2018.

LANGE, André. Télévision, cinéma, vidéo et services audiovisuels à la demande : Le paysage paneuropéen. *Observatoire européen de l'audiovisuel*. 2012, Vol.2, p. 167-180.

LE COADIC, Yves-François. Le besoin d'information. *Bulletin des bibliothèques de France*.1999, n° 3.

Le data journalisme : entre retour du journalisme d'investigation et fétichisation de la donnée. Entretien avec Sylvain Lapoix », *Mouvements*.2014, vol. 79, n. 3, pp. 74-80.

LE HONG, H. et D, LANEY. Toolkit: Board-Ready Slides on Big data Trends and - Opportunities. *Gartner*. March 2013.

LECHANI, Tamine Lynda et BOUGHANEM, Mohand. *Accès personnalisé à l'information : Apporches et techniques*. Rapport interne. RIT, 2005. Disponible sur ftp://ftp.irit.fr/IRIT/SIG/rapport_Perso_0904_VF.pdf [Consulté le 6 Octobre 2014].

LEMBERGER, Pirmin et BATTY, Marc. *Big data et Machine Learning : Les concepts et les outils de la data science.* 3e édition. Paris: Dunod, 2019. 272 p.

LEONAR, Peter. *Mining large datasets for the humanities.* IFLA, 2014. Disponible sur : http://library.ifla.org/930/1/119-leonard-en.pdf] [Consulté le 19 Juin 2020].

LONE, Tariq Ahmad et Ahmad Khan, RAFI. Data Mining: Competitive Tool to Digital Library. DESIDOC. *Journal of Library & Information Technology*. Vol. 34, n°. 5, 2014, p. 401-406.

MAATALLAH, M.; SERIDI-BOUCHELAGHEM, H.A Fuzzy Hybrid Approach to Enhance the Diversity in TOP-N Recommendations. *International Journal of Business Information Systems* (IJBIS). 2015, vol. 19, No. 4. p. 505-530.

SHANG, Xuedong. *Recommandation personnalisée [en ligne]*, 2015. Disponible sur : https://xuedong.github.io/static/documents/recommender.pdf [Consulté le 22 Février 2016].

MANOUSELIS, Nikos et CONSTANTINA, Costopoulou. Experimental analysis of design choices in multi attribute utility collaborative filtering. *International Journal of Pattern Recognition and Artificial Intelligence*. 2007.

MILOT, Francis. « Big data Index » IDC et EMC In : Vertu-desk [en ligne] (mis en ligne en 2014). Disponible sur : http://www.virtu-desk.fr/blog/le-cloud/big-data-index-emc-et-idc.html [Consulté le 05 Mai 2015].

MOREAU, François. The Disruptive Nature of Digitization: The Case of the Recorded Music Industry. *International Journal of Arts Management*. 2012.

NAAK, Amine. Papyres : U*n système de gestion et de recommandation d'articles de recherche. Mémoire d'obtention de maitrise en sciences informatiques* [en ligne]. Mémoire, Université de Montréal : Montréal, 2009. Disponible sur : https://core.ac.uk/download/pdf/151544366.pdf [Consulté le 21 Mars 2016]

NARDUCCI, fedulcio; CATALDOMUSTO, Giovansemeraro. Exploited big data for enhanced representations in content-based recommender systems. *International Conference on Electronic Commerce and Web Technologies [en ligne]*. 2013, vol.152. Disponible sur: https://link.springer.com/chapter/10.1007/978-3-642-39878-0_17 [Consulté le 22 Janvier 2017].

OARD, Douglas W., KIM, Jinmook, et al. Implicit feedback for recommender systems. *Proceedings of the AAAI workshop on recommender systems*. 1998. p. 81-83.

OCLC. *Produits et services OCLC* [En ligne], 2020. Disponible sur : https://www.oclc.org/fr/services.html [Consulté le 25 Août 2020].

PARK, Jung-ran et al. Cataloging professionals in the digital environment: A content analysis of job descriptions. *Journal of the American Society for Information Science and Technology*, 2009.Vol. 60, n° 4, p. 844–857.

PARRA, Denis ; KARATZOGLOU, Alexandros ; AMATRIAIN, Xavier, et al. Implicit feedback recommendation via implicit-to-explicit ordinal logistic regression mapping. *Proceedings of the CARS-2011*, 2011.

PAVLOVIC RIVAS, Marina. *Médias et données : une influence sur la diffusion et la qualité de l'information ?* HEC Montréal | « Gestion », 2017/1, Vol. 4, p. 80-81.

PECUNE, F., MURALI, S., TSAI [et al.]. *A Model of Social Explanations for a Conversational Movie Recommendation System*. In Proceedings of the 7th International Conference on Human-Agent Interaction, July 2019. Pp. 135-143. ACM.

PORCEL, Carlos, MORENO, Juan Manuel, et HERRERA-VIEDMA, Enrique. A multi-disciplinar recommender system to advice research resources in University Digital Libraries. *Expert Systems with Applications* [En ligne]. 2009, vol. 36, n. 10. Disponible sur: https://www.sciencedirect.com/science/article/abs/pii/S0957417409003698?via%3Dihub [Consulté le 12 Janvier 2018].

PROJET COUNTER [en ligne]. Disponible sur : https://www.projectcounter.org/. [Consulté le 28 Août 2019].

PROJET LOD-B 1. *Processus de transformation des données en RDF : analyse des best practices* [en ligne], 2015. Disponible sur : http://campus.hesge.ch/id_bilingue/projekte/lodz/doc/processus_transformation.pdf [Consulté le 24 Août 2020].

Recommendation Engines: How Amazon and Netflix Are Winning the Personalization Battle MarTech Advisor [En ligne]. Disponible sur : https://www.martechadvisor.com/articles/customer-experience/recommendation-engines-how-amazon-and-netflix-are-winning-the-personalization-battle/. [Consulté le 01 Septembre 2019].

RENAUD-DEPUTTER, Simon. *Système de recommandations utilisant une combinaison de filtrage collaboratif et de segmentation pour des données implicites* [en ligne], 2013. Disponible sur : https://core.ac.uk/download/pdf/51339447.pdf [consulté le 05 Novembre 2017].

RENDA, M. Elena et STRACCIA, Umberto. A personalized collaborative digital library environment: a model and an application. *Information processing & management*. 2005, vol. 41, no 1, p. 5-21.

RESNICK, Paul et VARIAN, Hal R. *Recommender systems. Communications of the ACM*. 1997, vol. 40, no 3, p. 56-58.

ROLLAND, Sylvain. Comment les algorithmes révolutionnent l'industrie culturelle. *La Tribune* [en ligne].2015. Disponible sur https://www.latribune.fr/technos-medias/comment-les-algorithmes-revolutionnent-l-industrie-culturelle-523168.html. [Consulté le 22 Août 2019].

SANES, Julien. *La magie des séries et des grands nombres : ce que l'on peut faire avec le Big Data et les archives.* Actes du séminaire IST Inria, octobre 2014.

SHARMA, Richa et SINGH, Rahul. Evolution of recommender systems from ancient times to modern era: A survey. *Indian Journal of Science and Technology*. 2016, vol. 9, no 20.

SHORT EDITION. *Editeur communautaire* [en ligne]. Disponible sur http://short-edition.com/fr/p/distributeurs-d-histoires-courtes [Consulté le 20/06/2015].

SIMONNOT, Brigitte. *L'accès à l'information en ligne : Moteurs, dispositifs et médiations*. Hermès Lavoisier, 2012, pp.249. Systèmes d'information et organisations documentaires. ISBN 978-2-7462-3829-9.

SIMONNOT, Brigitte. *Le besoin d'information : principes et compétences*. Thémat'IC, 2006.

SITE DE L'UNION EUROPEENNE. *De nouvelles règles et actions en faveur de l'excellence et de la confiance dans l'intelligence artificielle* [En ligne]. Disponible sur : https://ec.europa.eu/france/news/20210421/nouvelles_regles_europeennes_intelligence_artificielle_fr [consulté le 26 Avril 2021].

SOUISSI, Amen. *Modélisation centrée sur les processus métier pour la génération complète de portails collaboratifs* [en ligne]. Thèse de doctorat. Université des Sciences et Technologie de Lille - Lille I, 2013. Disponible sur : https://tel.archives-ouvertes.fr/tel-00935324/PDF/ThA_se_Amen-SOUISSI.pdf [Consulté le 12 Septembre 2018].

STENGER, Thomas. La prescription dans le commerce en ligne : proposition d'un cadre conceptuel issu de la vente de vin par Internet. *Revue Française du Marketing*, Paris : A.D.T. d'exécution et de l'exploitation des études de marché, 2006.

SUGIMOTO, Cassidy R. et Ying,DING.Library and information science in the big data era: Funding, projects, and future.*Proceedings of the American Society for Information Science and Technology*.2012,vol.49, issue 1, p.1-3.

Syndicat National de l'Edition. *Métadonnées et livres numériques* [en ligne]. Disponible sur https://www.sne.fr/numerique-2/metadonnees-et-livres-numeriques/ [consulté le 23 Août 2019].

TEETS, Michael et Matthew, GOLDNER. Library's role in curating and exposing big data. *Future Internet*. 2013, n°5, p.429-438.

TENNANT, Roy. *Managing the Digital Library*. New York: Reed Press, 2004. Print.

Traitements Big data avec Apache Spark - 1ère partie : Introduction [En ligne]. Disponible sur https://www.infoq.com/fr/articles/apache-spark-introduction. [Consulté le 12 juin 2017].

TRICOT, André. *Recherche d'information et apprentissage avec documents électroniques. Hypermédias et apprentissage* [en ligne]. Disponible sur http://andre.tricot.pagesperso-orange.fr/TricotLECAI.pdf [Consulté le 15 Octobre 2014].

UTARD, Jean-Claude. « L'élu, le directeur et la bibliothèque ». *Bulletin des bibliothèques de France (BBF)* [en ligne]. 2003, n° 1, p. 38-44. Disponible sur : https://bbf.enssib.fr/consulter/bbf-2003-01-0038-007 [Consulté le 19 juin 2015].

VELLINO, Andre. Recommending research articles using citation data, *Library Hi Tech*, Vol. 33, No. 4, 2015, pp. 597-609.

VERLAET, Lise et DILLAERTS, Hans. « L'enjeu du web de données pour l'édition scientifique », *I2D – Information, données & documents.2016*, vol. 53, no. 2, pp. 49-49.

WEINBERGER, D. *Everything is Miscellaneous: The Power of the New Digital Disorder*. Times Books: New York, NY, USA, 2007.

WENZ, Romain. Linked Open Library Data in practice: lessons learned and opportunities for data. SWIB 2012 [en ligne]. Cologne, 27 novembre 2012. Disponible sur :http://swib.org/swib12/slides/Wenz_SWIB12_108.ppt [Consulté le 20 Octobre 2015].

WHITMAN, B., "How music recommendation works – and doesn't work". *Variogram* [en ligne]. December 11, 2012. Disponible sur: https://notes.variogr.am/2012/12/11/how-music-recommendation-works-and-doesnt-work/ [consulté le 15 Avril 2017].

YANG, Zhe, WU, Bing, ZHENG, Kan, et al. A Survey of Collaborative Filtering-Based Recommender Systems for Mobile Internet Applications. *IEEE Access*. 2016, vol. 4, p. 3273-3287.

YAU, Nathan. *Data visualisation : De l'extraction des données à leur présentation graphique*. Editions Eyrolles, 2013. 365 p.

ZDNET-Social Media Club. *La data : un levier d'innovation pour les groupes de presse* [en ligne]. Disponible sur : https://www.zdnet.fr/blogs/social-media-club/la-data-un-levier-d-innovation-pour-les-groupes-de-presse-39835362.htm [Consulté le 12 Juin 2016].

Annexes

Annexe 1 Questionnaire sur l'usage de la data dans l'offre documentaire destinée aux chercheurs a l'IMIST

Il s'agit d'un guide destiné aux professionnels d'information de l'Institut Marocain de l'Information Scientifique et Technique (IMIST), s'inscrivant dans le cadre d'une thèse de doctorat dans le domaine des sciences de l'information et de la communication, réalisé au Conservatoire National des Arts et des Métiers (CNAM-Paris), en co-encadrement avec l'Ecole des Sciences de l'Information (ESI-Rabat), traitant de la problématique de la modélisation d'un système de recommandation à la communauté des utilisateurs. L'objectif étant de mettre en relief l'usage fait des données de toute sorte dans l'amélioration des produits et services fournis aux usagers.

1) Quels sont les types de documents auxquels donnent accès à la bibliothèque ?

☐ Monographies ☐ Périodiques ☐ Ressources électroniques

Autres (à préciser) :

……

……

2) Quels sont les produits et les services de la bibliothèque ?

……

……

……

3) Quel Système Intégré de Gestion de Bibliothèque (SIGB) est utilisé par la bibliothèque ?

……

4) Ce SIGB permet un accès au catalogue public ?

☐ Oui ☐ Non

5) Si « Oui », l'accès à ce catalogue public se fait-il ?

☐ En local ☐ Sur interface web

6) Quels sont les différentes données internes et externes produites par les bibliothèques sur les collections et les utilisateurs ?

……

……

……

……

7) Ces données portent-elles sur ?

☐ Les utilisateurs ☐ Les éditeurs ☐ Les collections

☐ Les produits et services documentaires ☐ L'environnement externe

8) Les utilisateurs préfèrent-ils avoir accès au fonds par le biais des produits et des services documentaires sous forme ?

☐ Papier ☐ Electroniques

9) Combien de documents dispose-t-elle la bibliothèque dans son fonds ?

……

10) A quoi servent les différentes données produites par les bibliothèques ?

……

……

……

11) Avez-vous une visibilité sur le taux de satisfaction des usagers quant à l'accès et l'exploitation du fonds de la bibliothèque ?

……

12) Quel est le taux d'accroissement de la collection de la bibliothèque ?

…….

13) Comment les usagers trouvent les produits et les services d'accès aux collections par rapport à leurs besoins ?

☐ Adaptés parfaitement ☐ Adaptés moyennement ☐ Inadaptés

14) Sur combien d'années les données de la bibliothèque ont été accumulées ?

……

15) Ces données sont-t-elles d'ordre ?

☐ Quantitatif ☐ Qualitatif ☐ Quantitatif et qualitatif

16) Ces données sont –elles recueillies de manière ?

☐ Directe (la source/détenteur) ☐ Indirecte (statistiques, enregistrements de données)

17) Les usagers ne demandent-t-ils pas des orientations, même s'ils trouvent les documents dont ils ont besoin ?

☐ Oui ☐ Non

18) Quelles sont les données produites au niveau interne et externe de la bibliothèque dans les différents maillons de la chaine documentaire ?

- Acquisition …….
- Traitement ……..
- Diffusion …….

19) Que pensez-vous des systèmes de recommandation en vue de permettre aux usagers de personnaliser et filtrer leur accès aux ressources documentaires de la bibliothèque ?

……

……

……

20) L’utilisateur a-t-il besoin à des recommandations de documents au cours de leur recherche dans le fonds en se basant sur ?

☐ Le contenu des documents ☐ Les intérêts partagés avec les autres usagers

Annexe 2 mini- questionnaire sur la satisfaction d'un échantillon des utilisateurs du prototype de recommandation

Il s'agit d'un guide destiné à un échantillon des usagers chercheurs de l'Institut Marocain de l'Information Scientifique et Technique (IMIST), s'inscrivant dans le cadre d'une thèse de doctorat dans le domaine des sciences de l'information et de la communication, réalisé au Conservatoire National des Arts et des Métiers (CNAM-Paris), en co-encadrement avec l'Ecole des Sciences de l'Information (ESI-Rabat), traitant de la problématique de la modélisation d'un système de recommandation à la communauté des utilisateurs. L'objectif étant de mesure la satisfaction globale par rapport au prototype de recommandation proposé.

1) Quel est l'apport général de la solution de recommandation proposée au catalogue en ligne, à l'issue de votre test ?

☐ Faible ☐ Moyen ☐ Très bien ☐ Excellent

Autres (à préciser) :

……

……

2) Les recommandations affichées à l'issue de vos requêtes, sont-t-elles ?

☐ Pertinentes ☐ Moyennement pertinente ☐ Non pertinentes

Autres (à préciser) :

……

……

HESAM UNIVERSITÉ

Amine SENNOUNI

le cnam

DONNEES ET INDUSTRIES DES CONTENUS, MODELISATION ET PROTOTYPAGE D'UN SYSTEME DE RECOMMANDATION POUR LE CATALOGUE D'UN SERVICE D'INFORMATION DOCUMENTAIRE

Résumé

Les systèmes de recommandations font partie des modèles d'apprentissage automatique qui transforment la recherche de l'information.

C'est dans ce cadre que s'inscrit notre thèse qui a pour objectif principal de concevoir et mettre en œuvre un modèle de système de recommandation basé sur les données implicites d'un échantillon des utilisateurs de l'Institut Marocain de la Recherche Scientifique et Technique (IMIST), en s'appuyant sur l'approche de filtrage collaboratif.

Notre contribution est ensuite la mise en place d'un prototype de système de recommandation adapté, reposant sur les données implicites des utilisateurs, afin de leur fournir des recommandations pertinentes de documents

L'implémentation du système a été effectuée dans un environnement Spark dédié à l'apprentissage automatique à l'aide du langage Scala et moyennant le modèle des moindres carrés alternés (ALS) de factorisation matricielle, appartenant aux méthodes de filtrage collaboratif à base de modèle.

Mots-clés : système de recommandation, apprentissage automatique, données implicites, filtrage collaboratif, service d'information documentaire, catalogue d'ouvrages, OPAC, industries des contenus, valorisation des données, Maroc.

Résumé en anglais

Recommender systems are among the promising fields of machine learning, which have revolutionized the information retrieval.

The context of our thesis is the Moroccan Institute for Scientific and Technical Information (IMIST), having as an objective to design and implement a recommender system based on log data, according to the collaborative filtering approach.

To achieve these objectives, we first began to identify the IMIST's need for a recommender system, then we design and carry out a prototype based on user's implicit data to provide recommendations.

It should be noted that it was necessary to convert the implicit data of the users into explicit data in the form of scores, in order to exploit them.

This prototype distinguishes between the anonymous user and the subscriber user.

The implementation of the system was done in a Spark environment, using the Scala language and the ALS Train Implicit model of the matrix factorization.

Key words: recommender system, machine learning, implicit data, collaborative filtering, documentary services, books catalog, OPAC, content industry, data valuation, Morocco.