



MASTER 2 MIAGE

Mémoire de fin d'études

---

Comment le système de  
recommandation de Netflix a  
rendu accros ses utilisateurs ?

---

*Auteur :*  
Fabien MICHEL

*Tuteur :*  
Lom messan HILLAH

Promotion 2018-2019



## Remerciements

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma gratitude.

Je tiens à exprimer toute ma reconnaissance à mon tuteur de mémoire, Monsieur Lom messan HILLAH. Je le remercie de m'avoir encadré, orienté, aidé et conseillé.

J'adresse mes sincères remerciements à tous les professeurs, les intervenants et à toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques ont guidé mes réflexions et qui ont accepté de me rencontrer et de répondre à mes questions durant mes recherches.

Je tiens tout particulièrement à remercier Marjolaine LEVEAU pour ses conseils et ses remarques qui m'ont guidé et m'ont été d'une très grande aide durant la rédaction de ce mémoire.



## Résumé

Dans un avenir très proche, les systèmes de recommandation seront devenus incontournables, non seulement dans l'audiovisuel, mais globalement sur internet et peu importe leurs types de contenus. Demain, il n'est pas impossible que des sites tels que LinkedIn recommandent à des employeurs le parfait candidat pour être embauché.

Actuellement, une entreprise mise beaucoup sur les systèmes de recommandation : Netflix. Ce service, utilisé par des millions d'utilisateurs dans le monde met en avant son système de recommandation dans son plan de communication mais estime aussi qu'il s'agit d'un élément primordial pour garder un de ses utilisateurs sur sa plateforme.

En premier lieu, nous étudierons les systèmes de recommandation de manière général et l'approche que Netflix a appliqué sur sa plateforme.

Dans un second temps, nous répondrons aux questions suivantes : Quelles sont les différentes contraintes des systèmes de recommandation ? Quelles sont leurs limites ? Comment Netflix a limité les contraintes ? Et quelle est l'importance de la recommandation pour Netflix ?

Dans une troisième partie, nous dresserons un état de l'art en se basant sur le Netflix Prize permettant d'en apprendre d'avantage sur le fonctionnement technique de la recommandation par Netflix.

Finalement, je présenterai ma propre interprétation du Netflix Prize en y proposant une solution. Cette dernière sera schématisée et expliquée afin de déterminer sa pertinence.



## Mes motivations

Avant de commencer mon stage de Master 1 MIAGE au sein du Laboratoire de Recherche en Informatique, je n'avais pas pris conscience de l'importance de la recommandation et de son impact sur notre quotidien.

Durant ce stage, j'ai appris à découvrir le domaine de la recherche autour des systèmes de recommandation et ainsi j'ai pu m'initier à leurs fonctionnements et à la mise en place d'un système de recommandation complexe.

Il existe énormément de possibilités pour mettre en place une recommandation efficace, après une initiation dans le monde de la recherche, j'ai voulu en apprendre plus et découvrir comment une entreprise multinationale de l'ordre de Netflix, mondialement reconnue pour sa recommandation à mis en place un système efficace. De plus, cette entreprise a déjà démontré par le passé que sa recommandation a énormément d'importance autant sur le plan technique que dans sa stratégie marketing. C'est donc tout naturellement que pour l'ensemble de ses raisons, j'ai décidé de mettre en lien les systèmes de recommandation avec la plateforme Netflix.





# TABLE DES MATIÈRES

<b>Introduction</b>	<b>1</b>
<b>1 Contexte et concepts</b>	<b>3</b>
1.1 Généralités . . . . .	3
1.2 Les différentes approches . . . . .	4
1.2.1 Le filtrage collaboratif . . . . .	4
1.2.2 Le filtrage basé sur le contenu . . . . .	6
1.2.3 Le filtrage hybride . . . . .	7
1.3 Le machine learning . . . . .	8
1.4 L'approche de netflix . . . . .	9
<b>2 Problématique et contraintes</b>	<b>11</b>
2.1 Généralités . . . . .	11
2.2 Les contraintes et problématiques de Netflix . . . . .	12
2.3 L'importance de la recommandation pour Netflix . . . . .	12
<b>3 État de l'art</b>	<b>16</b>
3.1 Analyse de l'existant . . . . .	16
3.2 Comparaison des différentes approches . . . . .	19
3.3 Critique . . . . .	21
<b>4 Mon apport</b>	<b>23</b>
4.1 Recherche des plus proches voisins . . . . .	25
4.2 Trier les voisins selon le film . . . . .	26
4.3 Prédiction . . . . .	27
<b>5 Évaluation de mon apport</b>	<b>28</b>
<b>6 Évolutions possibles</b>	<b>30</b>
<b>Conclusion</b>	<b>32</b>



# INTRODUCTION

Le "Dayli mix" de Spotify, les vidéos recommandées sur la page d'accueil YouTube ou encore les suggestions d'abonnement sur Instagram, ces recommandations sont exclusivement basées sur vos "j'aime" ou interactions précédentes. Ces recommandations permanentes sont issues d'un système de recommandation, aujourd'hui très présent dans notre quotidien et dans beaucoup de services que nous utilisons.

Aujourd'hui, une entreprise mise énormément sur les systèmes de recommandation : **Netflix**. L'idée système de recommandation performant est rapidement devenue un besoin pour l'entreprise et ce dès la création de Netflix. Wilmot Reed Hastings est un entrepreneur américain et fondateur de Netflix. En 1997, la location de film est courante et comme un grand nombre de personnes à cette époque il oublie de rendre la copie d'un film. Au bout de six semaines, la boutique lui demande de payer une quarantaine de dollars de frais de retard<sup>22</sup>. Ainsi, l'idée de Netflix est née : une formule par abonnement avec la possibilité de conserver la copie d'un film aussi longtemps que souhaité le tout par internet. Aucune boutique physique mais un catalogue complet sur internet, plus complet que n'importe quelle boutique existante aussi grande soit-elle et la copie du film choisie est envoyée par la poste. L'idée est brillante mais devant l'immensité du catalogue, les clients choisissent les mêmes films : les gros succès du cinéma. L'idée est alors pour l'entreprise de mieux connaître les goûts de ses utilisateurs à partir de leurs statistiques de consommation. Les titres loués sont utilisés pour proposer, des titres moins populaires permettant à Netflix de faire tourner les blockbusters et d'envoyer les titres les moins populaires. Ce système de recommandation est appelé **CineMatch** et il est présenté comme l'un des atouts de Netflix.

En 2018, plus de 30 millions de nouveaux abonnés ont rejoint le service de SVOD<sup>a</sup> de Netflix, portant à environ 140 millions le nombre de comptes

---

a. Subscription Video On Demand : Permet d'accéder avec un abonnement payant

payants<sup>b</sup> soit plus de deux fois la population Française. D'après les chiffres du Figaro<sup>5</sup> en 5 ans, Netflix comptabilise en avril 2019 déjà 5 millions d'abonnés et a donc dépassé en terme d'abonnés payants, l'acteur historique CANAL+ qui comptabilise sur la même période 4,757 millions d'abonnés individuels en France.

Premièrement, il est nécessaire de connaître qui en France consomme Netflix, selon une enquête d'Harris Interactive<sup>c</sup> et NPA<sup>d</sup> qui dresse le profil type d'un utilisateur de SVOD en France<sup>17</sup>. En moyenne, 25% des utilisateurs d'un service de SVOD utilisent ce service quotidiennement pour une session durant en moyenne 1 à 3h. La tranche horaire la plus utilisée est celle du "prime time" (de 21h à 23h). Dans cette enquête, deux faits sont marquants, premièrement, 63% des utilisateurs d'un service de SVOD en France sont âgés de moins de 35 ans tandis que 64% des spectateurs de la "télévision classique" sont âgés de plus de 50 ans. Malgré cette différence d'âge, les utilisateurs de SVOD privilégient tout de même majoritairement le téléviseur pour regarder leurs contenus.

Autrefois, sans réels concurrents,<sup>4</sup> de nombreux acteurs en voyant le succès de Netflix ont décidés de rentrer dans la course, aujourd'hui NetFlix se trouve face à Amazon Prime Video et Hulu mais prochainement de nombreux géants vont rejoindre le marché comme Disney+, Paramount+ et bien plus encore.

Chaque utilisateur donne à son insu des informations à Netflix sur ses habitudes et les exploites. Ce qu'on regarde, sur quel support et à quel moment de la journée, permet par la suite de recommander du contenu qui serait susceptible de nous plaire. Aujourd'hui, la majorité des contenus consommés sur la plateforme de Netflix sont issus de son système de recommandation.

---

à un catalogue de vidéo à la demande généralement sans publicité, en illimité et sans engagement.

b. Le premier mois de souscription étant gratuit, ils ne sont pas comptabilisés

c. Harris Interactive est un institut d'études marketing et de sondages d'opinion (Politique - Digital - Corporate).

d. NPA Conseil est cabinet de conseil stratégique et opérationnel au service de la transformation numérique.

# Chapitre 1

## CONTEXTE ET CONCEPTS

Aujourd'hui, sur internet des entreprises mettent à disposition de leurs utilisateurs un nombre colossal de données comme YouTube, Amazon ou encore Netflix. Ainsi, il est devenu difficile pour l'utilisateur de trouver des informations pertinentes rapidement. Diverses techniques informatiques se sont développées pour contourner ce problème et ce bien avant les géants d'internet que nous connaissons. Dans le cadre de ce mémoire, nous nous intéresseront essentiellement aux systèmes de recommandation.

En 1967<sup>14</sup>, James McQueen applique un algorithme K-moyennes (K-means) permettant de construire des portions de populations homogènes selon plusieurs critères définis à l'avance. Actuellement, cette première étape est considérée comme le début de la recommandation personnalisée. En 1979, est crée Grundy<sup>20</sup>, il s'agit d'une tentative de recommandation automatique appliquée au métier de bibliothécaire. A l'aide d'une interview, le système classe les utilisateurs en "stéréotypes" permettant ainsi de recommander différents livres en fonctions de leurs "stéréotypes". Nous pouvons considérer qu'il s'agit de la première réelle application d'un système de recommandation.

### 1.1 Généralités

Un système de recommandation est une forme spécifique de filtrage de l'information qui a pour but de présenter à un utilisateur des éléments qui sont susceptibles de l'intéresser. Ce système fonctionne en se basant sur les préférences et le comportement d'un utilisateur ou d'un groupe d'utilisateur. In fine, un système de recommandation tente donc de prédire si un élément suggéré est susceptible d'être apprécié par un utilisateur donné.

Aujourd'hui, les systèmes de recommandations sont devenus extrêmement populaires, ils sont utilisés dans de nombreuses webapp et appliqués à de nombreux domaines comme la publicité, la vente de masse ou encore la consommation de contenu.

Actuellement, le nombre de données collectées par des entreprises comme Google par exemple, augmente de façon exponentielle, en raison de cette croissance l'importance des systèmes de recommandation augmente tout autant. Depuis les années 90, les systèmes de recommandation sont devenus un domaine de recherche autonome et en recherche constante d'évolution<sup>9</sup>.

## 1.2 Les différentes approches

Il existe une multitude de type de données avec chacune des spécificités différentes. Pour être le plus efficace possible, il existe différentes implémentations des systèmes de recommandation afin de traiter au mieux les données utilisateurs permettant d'apporter l'information la plus pertinente possible.

En me basant sur les travaux d'Elsa NEGRE<sup>16</sup>(Maître de Conférences à l'Université Paris - Dauphine), j'ai pu déterminer l'existence de 3 grands types de systèmes de recommandation : basé sur un filtrage collaboratif, filtré sur le contenu et enfin les systèmes hybrides. Chacun de ses types de systèmes de recommandation permet de s'adapter au mieux en fonction des différents besoins.

### 1.2.1 Le filtrage collaboratif

Les systèmes basés sur le filtrage collaboratif construisent des recommandations en observant la similarité entre les préférences d'un utilisateur et celles d'autres utilisateurs. Ce système ne permet pas d'analyser ou d'étudier les éléments à recommander mais à faire des prévisions basées sur les intérêts d'un utilisateur comparé à un ensemble d'utilisateurs ayant un profil similaire. Dans ce système, on suppose que plus le profil d'un utilisateur est proche d'un ensemble/groupe d'utilisateur alors ils auront tendance à aimer les mêmes éléments.

Lorsqu'il y a des milliards de produits et un nombre conséquent de clients, il faut une énorme puissance de calcul pour calculer les recommandations. Il existe donc un problème sur l'évolutivité de ce système de recommanda-

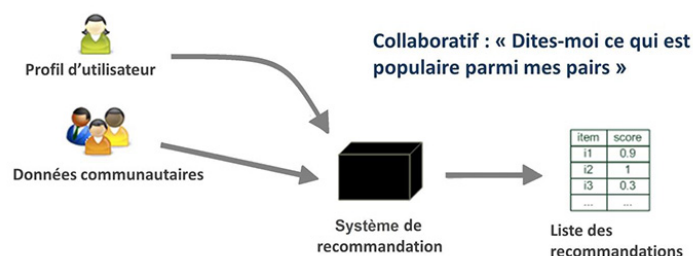


FIGURE 1.1 – Un système de recommandation collaboratif<sup>7</sup>.

tion. Lors de la première inscription d'un utilisateur, il est impossible de recommander des produits à cet utilisateur. Ce problème du démarrage à froid est contournable en demandant les préférences d'un utilisateur après son inscription.

Aujourd'hui, Amazon a un algorithme de recommandation pour sa boutique. C'est un algorithme utilisant le filtrage collaboratif et plus spécifiquement "Item-To-Item". Lors de la lecture de la fiche "produit d'un article", Amazon suggère une liste d'autres produits en fonctions des clients ayant déjà acheté ce produit. Amazon personnalise aussi sa page d'accueil en fonction des achats précédents ou des articles dans le panier du client.

« L'algorithme d'Amazon.com est basé sur le filtrage collaboratif appliqué aux éléments (Item-based collaborative filtering ou Item-to-Item). Le calcul en temps réel de cet algorithme s'adapte à la fois au nombre de clients et au nombre de produits dans le catalogue. L'algorithme construit une matrice de produits similaires en trouvant les produits que les clients ont tendance à acheter ensemble. »<sup>16</sup>



FIGURE 1.2 – Amazon et la recommandation collaborative.

### 1.2.2 Le filtrage basé sur le contenu

Pour les algorithmes de recommandation basés sur le contenu, le travail consiste à faire concorder les éléments du catalogue qui coïncident le mieux avec les préférences d'un utilisateur. Contrairement au filtrage collaboratif, ce type d'algorithme ne demande pas d'avoir un grand nombre d'utilisateur et peut s'appliquer à des structures plus petites.

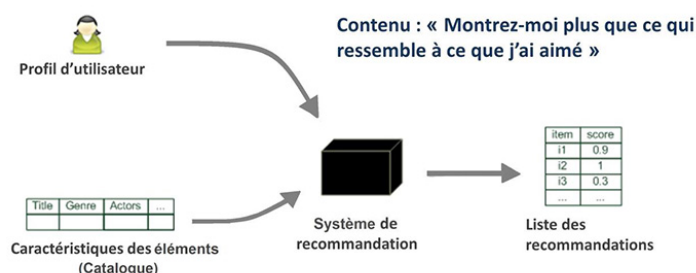


FIGURE 1.3 – Un système de recommandation basé sur le contenu<sup>7</sup>.

Premièrement, pour mettre en place ce type de recommandation, chaque éléments ou produits sont décrits par un nombre de caractéristiques connus et finis. Pour chaque utilisateur, il faut exprimer sous forme d'une liste ses intérêts et ses caractéristiques à nos produits ou nos éléments.

La seconde étape pour la mise en place d'un algorithme de recommandation basé sur le contenu est de faire coïncider les caractéristiques des éléments et le profil de l'utilisateur. Ainsi, cela peut être mesuré de différentes manières :

- **La mesure de similarité.**
- **Le TF-IDF** (Term Frequency-Inverse Document Frequency) : consiste à déterminer un score de pertinence.
- **«Les techniques basées sur la similarité des espaces vectoriels** (les approches bayésiennes, les arbres de décision, etc.) couplées avec des techniques statistiques, lorsqu'il y a trop de mots-clés.»<sup>16</sup>

Les systèmes de recommandation basés sur le contenu présentent de nombreux avantages dont le démarrage à froid. Si un nouvel élément est ajouté au catalogue, il est simple de le recommander directement aux utilisateurs



en faisant correspondre les caractéristiques et les centres d'intérêts d'un utilisateur. Avec ce système de recommandation, une personne avec des goûts atypiques différents de ceux de la base d'utilisateurs habituels n'est pas un problème. Pareil pour les éléments à recommander, un produit peu vendu ou atypique peu coïncider avec les goûts d'un nombre réduit d'utilisateur et donc être recommandé.

Malgrès les nombreux avantages que représente ce système de recommandation, il ne peut pas être applicable à tous les utilisateurs. Les utilisateurs ayant consulté un grand nombre d'éléments posent un problème (une énorme masse d'information dans le profil de l'utilisateur à faire correspondre avec les propriétés des différents éléments) et lorsqu'il n'existe pas d'historique (dans le cas d'un utilisateur qui commence à utiliser le système) par conséquent la recommandation pour ces deux cas peut-être difficile. Concernant les éléments à recommander, il est difficile de différencier deux éléments avec des caractéristiques similaires. Une étape difficile est la définition des caractéristiques de chaque utilisateur ainsi que la prise en compte de l'évolution des goûts et par conséquent des caractéristiques des utilisateurs.

### 1.2.3 Le filtrage hybride

Un système de recommandation hybride combine les approches collaboratives et basées sur le contenu. *«Ce système de recommandation, peut utiliser à la fois des connaissances extérieures et les caractéristiques des éléments, combinant ainsi des approches collaboratives et basées sur le contenu.»*<sup>16</sup>

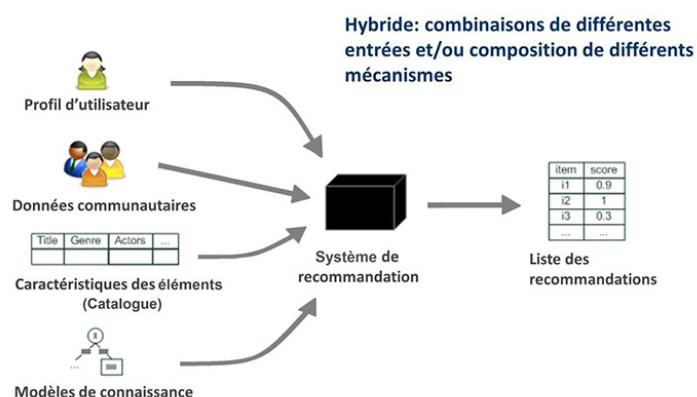


FIGURE 1.4 – Un système de recommandation hybride<sup>7</sup>.

L'approche hybride est une combinaison de différentes approches, par conséquent son but étant de présenter les avantages des approches qui la composent tout en limitant leurs inconvénients respectifs. Actuellement, il existe différentes catégories de combinaisons de systèmes de recommandation pour la création d'un système de recommandation hybride<sup>3,7</sup> :

- **La combinaison monolithique** (monolithic hybridization design)<sup>16</sup> : décrit une conception de l'approche hybride intégrant les aspects de différentes stratégies de recommandation en un seul algorithme.
- **La combinaison parallèle** (parallelized hybridization design)<sup>16</sup> : décrit une conception de l'approche hybride où sur la base de données d'entrée commune, les systèmes de recommandation fonctionnent de façon parallèles et indépendantes. Chacun produit une liste de recommandations distincte. Une étape ultérieure d'hybridation est nécessaire permettant de combiner le résultat en un ensemble final de recommandations.
- **La combinaison tubulaire** (pipelined hybridization design)<sup>16</sup> : décrit une conception de l'approche hybride où plusieurs systèmes de recommandation sont joints. Les données de sortie d'un système de recommandation devient une partie des données d'entrée du système de recommandation suivant.

## 1.3 Le machine learning

*« Les systèmes de recommandation sont couramment utilisés en lien avec l'intelligence artificielle. Leurs capacités à fournir un aperçu global, à prédire les événements et à mettre en évidence des corrélations contribuent à expliquer leurs utilisations en IA. Par ailleurs, les techniques de Machine Learning sont fréquemment utilisées pour créer les algorithmes de recommandation. Par exemple, chez Arcbees, un système de prédiction des préférences de films a été construit, il utilise un réseau de neurones et les données de IMDb. Les réseaux de neurones peuvent effectuer rapidement des tâches complexes et manipuler facilement des données massives. En fournissant une liste de films comme entrées et en comparant la sortie avec la note de l'utilisateur, le réseau peut apprendre par lui-même la règle permettant de prédire les évaluations futures d'un utilisateur spécifique. »<sup>8</sup>*

## 1.4 L'approche de netflix

Chez Netflix 80%<sup>6</sup> du contenu consommé par les utilisateurs est issu d'une recommandation. Fort de ce constat, le rôle du système de recommandation occupe une place centrale autant pour Netflix que pour ses utilisateurs. Ce système repose sur les utilisateurs, les caractéristiques des contenus et un algorithme.

Pour comprendre le besoins de ses utilisateurs, Netflix fait une analyse comportementale précise de chaque utilisateur. Ainsi, un contenu consommé sera utilisé pour recommander des series ou des films. Dans ce cas, Netflix utilise le machine learning pour dresser le portrait le plus précis possible de chaque utilisateur dont notamment :

- **La navigation précise** : combien de temps l'utilisateur met à trouver un contenu et par quel moyen (strates, catégories,...).
- **La temps de lecture d'un contenu** : permettant de déterminer si un contenu lancé est regardé en entier, à moitié ou partiellement
- **La notion de temps** : en fonction de la période de l'année, un type de contenu peut-être plus consommé par exemple les films de Noël. Mais aussi en fonction de l'heure, par exemple en début de soirée peut-être que les utilisateurs consomment plus de films que de séries.
- **L'avis donné par l'utilisateur** : L'utilisateur peut donner son avis sur un contenu avec un icône "pouce vers le haut" ou "pouce vers le bas".

Une analyse du comportement seule ne permet pas de déterminer les séries et films à recommander à un utilisateur, il faut analyser le contenu et faire concorder le portrait de chaque utilisateur avec le contenu. L'ensemble du catalogue disponible sur Netflix est indexé pour décrire au mieux un contenu selon une immense bibliothèque de mots-clés. Le rôle de l'algorithme de machine learning est d'ordonnancer les données des utilisateurs et les contenus indexés. En pratique, le système de recommandation de Netflix appelé Cinematch analyse les scores cumulés de chaque contenu en utilisant une variante du coefficient de corrélation de Pearson avec tous les autres films afin de déterminer une liste de films « semblables » qui sont susceptibles de plaire à l'utilisateur. Puis, la partie en ligne et en temps réel du système calcule une régression multivariée<sup>a</sup> basée sur ces corrélations pour déterminer une

---

a. Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement

prédiction unique et personnalisée pour chaque film recommandable fondée sur ces scores.

---

mis en œuvre pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple.

---

## Chapitre 2

# PROBLÉMATIQUE ET CONSTRAINTES

### 2.1 Généralités

Afin de fonctionner correctement et d'être efficace pour une population donnée, un algorithme de recommandation fait face à un certain nombre de contraintes selon son type :

- **Item cold start** : C'est lorsqu'un item est nouveau ou simplement lorsque l'item n'a pas encore été « cliqué » ou « acheté ». Cet item est difficilement recommandable via des algorithmes de recommandation de type collaborative filtering. En employant des méthodes de type content-based un nouvel item est facilement recommandable même si cet item est nouveau.
- **User cold start** : Si un utilisateur est nouveau, il existe aucun historique d'interaction avec le système. Il faut alors avoir recours à des données extérieures pour pouvoir commencer à recommander des items pertinents comme des données extraites des réseaux sociaux ou une demande explicite des goûts des nouveaux utilisateurs.
- **Diversity** : Ce problème survient souvent lors de l'utilisation de méthodes de type "contentbased filtering", le système recommande toujours les mêmes items à un utilisateur, ce qui l'enferme sans possibilité d'exploration en dehors des intérêts détectés par le système.
- **Grey sheep** : certains utilisateurs ne font rien comme les autres, on ne peut donc pas leur recommander d'item via les méthodes de type collaborative filtering.

- **Quality** : le content based filtering analyse le contenu, mais la qualité du produit n'est pas forcément déterminé par son contenu (A survey on recommendation system de Singh et al.).
- **Trust** : les utilisateurs doivent avoir confiance et croire en la performance du système pour continuer à l'utiliser.

Il existe un grand nombre d'autres contraintes comme **Scalability**, **Privacy**, et bien d'autres.

## 2.2 Les contraintes et problématiques de Netflix

A la création d'un compte Netflix, pour contourner le problème de l'**User Cold Start** le système demande à l'utilisateur de choisir un ensemble de séries ou de films qu'il aime déjà. Ainsi le système de recommandation de Netflix peut commencer à recommander d'autres contenus. Les contenus recommandés doivent être de bonne qualité (contrainte **Quality**) sinon l'utilisateur lassé n'utilisera plus le service.

Chaque mois, Netflix met à disposition de nouveaux programmes originaux sur sa plateforme et donc se retrouve face à la contrainte **Item-Cold-Start**. Ce contenu étant nouveau, il manque d'avis permettant de déterminer sa qualité. Les programmes sont ainsi mis en avant par la plateforme pour faire leurs promotions mais aussi dans les strates de recommandation en fonction des mots-clés associés.

Les abonnements étant mensuels et sans engagement, chaque mois Netflix doit convaincre les utilisateurs de rester sur la plateforme. Ainsi de façon mensuelle, la plateforme subie les contraintes **Diversity** et **Trust**. En offrant de nouveaux contenus chaque mois, la plateforme doit convaincre ses utilisateurs que la recommandation est pertinente vis à vis de leurs besoins.

## 2.3 L'importance de la recommandation pour Netflix

Avec un catalogue composé de milliers de films et de séries par pays ainsi qu'une base d'utilisateurs grandissante de plusieurs millions d'utilisateurs à

travers le monde, il est difficile pour Netflix d'utiliser un unique algorithme pour recommander des contenus.

Dans un article publié par Netflix<sup>10</sup>, les auteurs (responsables du machine learning chez Netflix) font l'état de six algorithmes présents sur la page d'accueil afin d'afficher une recommandation unique à chaque utilisateur :

- « Le "**Personalized Video Ranker**" opère le classement personnalisé des vidéos et ordonne les 40 rangées de 75 titres qui composent la page d'accueil.
- Le "**Top-N Video Ranker**" sélectionne, parmi les contenus les plus populaires dans l'ensemble du catalogue, ceux susceptibles de plaire à l'utilisateur.
- Le "**Trending Now**" détermine les tendances à court terme chez les consommateurs : les comédies romantiques de la Saint-Valentin ou les films de Noël, par exemple.
- Le "**Continue Watching**" sélectionne les vidéos que l'utilisateur a commencé à regarder et dont il souhaite probablement reprendre la lecture.
- Le "**Video-Video Similarity**" choisit les vidéos susceptibles de plaire à un utilisateur, compte tenu des similitudes existantes entre elles et celles qu'il a regardées.
- Le "**Page Generation : Row Selection and Ranking**" détermine les rangées à faire figurer sur la page d'accueil et leur ordre d'apparition, en tenant compte des résultats des algorithmes précédents.»<sup>2</sup>

Selon leurs études, un utilisateur de Netflix doit mettre au maximum 90 secondes à choisir un film, au-delà il perd de l'intérêt pour la plateforme. Pour Netflix, cela correspond entre 10 et 20 films proposés ou recommandés. Dans ce cas, si un utilisateur ne lance pas de film, il quitte le service. Le risque à terme est que l'expérience se renouvelle pour un utilisateur et qu'il ne soit plus satisfait de la proposition du contenu et se désabonne.

Personnaliser la page d'accueil afin de la rendre unique pour chaque utilisateur est une première étape dans la stratégie de recommandation de Netflix. Lorsqu'un utilisateur navigue dans l'interface, il faut le convaincre de consommer un contenu en moins de 90 secondes. Pour rendre sa recommandation plus attractive et personnelle à chaque utilisateur, Netflix a commencé

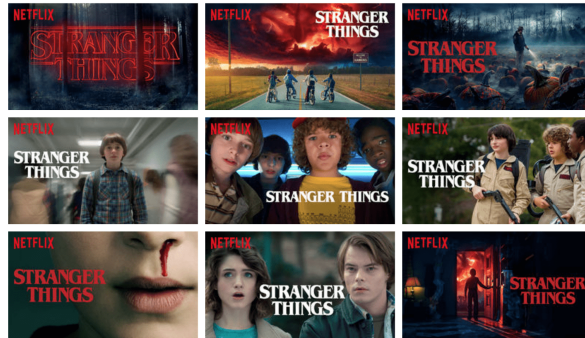


FIGURE 2.1 – Exemple de la diversité des illustrations.<sup>2</sup>

à produire plusieurs illustrations pour chaque contenu. Chaque illustration doit faire ressentir une émotion différente permettant de faire correspondre la vignette du contenu mis en avant avec les goûts de l'abonné.

Profile Type	Score Image A	Score Image B
Comedy	5.7	6.3
Romance	7.2	6.5



Image A



Image B

FIGURE 2.2 – Exemple de score et profil pour le film "Good Will Hunting" .

Si nous prenons l'exemple du film, Good Will Hunting qui est une comédie romantique. Un profil qui est déterminé comme regardant principalement des comédies, l'algorithme va choisir la vignette du film mettant en avant Robin Williams un acteur célèbre pour des comédies comme Madame Doubfire par exemple. A l'inverse, si un abonné est déterminé comme préférant les contenus romantiques, c'est alors l'image d'un couple s'embrassant qui est choisie. Pour chaque vignette, un ensemble de méthode détermine la correspondance sur 10 de chaque illustration avec chaque type (comédie, romance, horreur, science fiction,...). L'algorithme choisit pour chaque profil d'utili-



sateur, l'illustration qui correspond le mieux à ses goûts. Cette étape est déterminée par le type qui correspond à l'utilisateur et le score de l'illustration le plus élevé pour ce type de profil.

En plus de cibler ses abonnés et grâce à ses multiples vignettes pour une seule série, Netflix peut tester la vignette la plus efficace. Permettant ainsi de proposer les vignettes les plus efficaces pour un nouvel abonné mais aussi d'agrandir la fanbase d'une série. Un fan sur cinq<sup>11</sup> de *Stranger Things* n'avait jamais regardé de contenus horribles sur Netflix.

Début 2019, Netflix a dévoilé ses résultats trimestriels. Dans ce communiqué<sup>18</sup>, Netflix affirme que ses principaux rivaux sont désormais Fortnite et YouTube. L'entreprise explique que sa présence globale sur les écrans est supérieure à HBO (chaîne qui possède *Game of Thrones*) mais inférieure à celle de Fortnite le phénomène incontesté du jeu vidéo en 2018. Le temps utilisé pour jouer à Fortnite n'est pas utilisé pour regarder des séries, des films ou des documentaires sur Netflix. On parle ici d'une bataille de l'attention et finalement d'argent. En effet : l'argent dépensé pour acheter des tenues dans Fortnite n'est pas de l'argent investi dans un abonnement mensuel pour Netflix. Ainsi, plus que jamais, Netflix doit s'appuyer sur son contenu et sur l'ensemble de son système de recommandation pour convaincre ses utilisateurs de consommer sur leur site et donc de passer du temps sur leur plateforme plutôt que de jouer à Fortnite pour gagner la bataille de l'attention.

## Chapitre 3

# ÉTAT DE L'ART

En 2006, la recherche sur les algorithmes de recommandation a suscité beaucoup d'intérêt lorsque Netflix a lancé une compétition le « **Netflix Prize** » pour améliorer son approche et sa recommandation de films.

A cette époque, l'entreprise est encore un service de location de DVD en ligne mais chaque utilisateur pouvait laisser son avis et attribuer une note entre un et cinq au film. Avant le lancement de cette compétition, Netflix avait déjà son système de recommandation « **CineMatch** », permettant de suggérer aux clients un certain nombre de films qu'ils seraient susceptibles d'aimer. L'intérêt était double pour Netflix, une bonne recommandation fidélise son audience et une récompense d'un million de dollars permet de faire un joli coup de pub au service permettant d'augmenter par la suite son chiffre d'affaires.

L'objectif du concours était de construire un algorithme de recommandation qui pourrait surpasser CineMatch de 10 pourcent. Le concours a suscité beaucoup d'intérêt, tant dans le milieu de la recherche que dans celui des amateurs de films surement grâce à la somme mise en jeu.

### 3.1 Analyse de l'existant

Le Netflix Prize était une compétition ouverte au public avec la possibilité de participer en équipe ou de façon individuelle. Il n'était pas possible de soumettre son résultat plus d'une fois par jours. Les gagnants du concours ont eu l'obligation de publier leurs résultats ainsi que le code et les licences sur Netflix. Permettant ainsi à Netflix d'exploiter le code de l'équipe gagnante.

Pour la compétition, Netflix a mis à disposition un ensemble de données pour les participants à la compétition. L'ensemble des données contient au total :

- **480,189 Utilisateurs** : Ce sont des abonnés réels mais ils sont anonymisés.
- **17 770 films**.
- **Des évaluations de films** allant de 1 à 5.

Le répertoire "training set" contenant 17770 fichiers soit un par film. La première ligne de chaque fichier contient l'identifiant du film ensuite chaque ligne suivante du fichier correspond à une note donnée par un client. Le format de donnée est le suivant :

Film X:

```
Identifiant client, évaluation, date
Identifiant client, évaluation, date
...
```

Le fichier "qualifying.txt", contient les données éligibles pour la compétition. Cet immense fichier est composé d'un identifiant de film et un ensemble de couple <IdentifiantClient, Date> associé au film. L'algorithme rendu doit prévoir toutes les évaluations des clients donnés sur les films dans le jeu de données de qualification en se basant sur les informations disponibles sur chaque utilisateur dans les données "training set". Le format des données est le suivant :

```
Identifiant Film 3 :
Identifiant Client 5, date11
Identifiant Client 12, date12
```

```
Identifiant Film 4:
Identifiant Client 45, date 11
Identifiant Client 54, date 12
```

Concernant le format du rendu, il faut remplacer <IdentifiantClient, Date> par la prédiction de note pour le film. Par exemple, si les données du fichier "qualify" sont les suivantes :

111:  
3245,2005-12-19  
5666,2005-12-23  
6789,2005-03-14  
225:  
1234,2005-05-26  
3456,2005-11-07

Alors le fichier contenant les prédictions doit ressembler au résultat suivant :

111:  
3.0  
3.4  
4.0  
225:  
1,0  
2.0

Le résultat 3.0 signifie que le client 3245 évalue le film 111 le 19 décembre 2005 avec la note de 3,0 étoiles.

Le dernier fichier donné par Netflix pour réaliser cette compétition est le "probe.txt". Sa structure est la même que le fichier permettant de se qualifier à l'exception qu'il ne contient aucune date (simplement des identifiants de film et de client). Les données contenues dans ce fichier font aussi références aux données dans le répertoire "training set".

Identifiant Film 3 :  
Identifiant Client 5  
Identifiant Client 12

Identifiant Film 4:  
Identifiant Client 45  
Identifiant Client 54

Netflix fournit aussi un morceau de code permettant de calculer le RMSE<sup>a</sup> des prévisions calculées et surtout de comparer cette valeur à celui de Cinematch sur les mêmes données. Pour information, le RMSE de Cinematch est de 0,9514 pour les données fournies.

---

a. Le Root Mean Square Error est l'écart type des résidus (erreurs de prédiction).

## 3.2 Comparaison des différentes approches

Au cours de la compétition, beaucoup d'équipes sont parties sur des approches différentes. Trois années après le lancement du « Netflix Prize », le challenge est réussi et remporté par l'équipe "BellKor's Pragmatic Chaos". Leur solution est un agrégat de plusieurs algorithmes et de différents modèles mais aussi d'équipes ( BellKor<sup>13</sup>, The Pragmatic Theory<sup>19</sup> et The BigChaos<sup>21</sup>). Cette solution fut la première à avoir un meilleur RMSE que cinematch et par conséquent à donc remporté la compétition.

Durant la compétition, certains algorithmes étaient plus populaires que d'autres et sont apparus régulièrement. Parmi eux, l'algorithme du plus proche voisin et la factorisation matricielle se sont montrés particulièrement efficace<sup>12</sup>.

Une approche via l'algorithme du plus proche voisin (Nearest neighbors) permet de déterminer si deux utilisateurs ont des préférences identiques, proches ou éloignés. Dans le cas du Netflix Prize, les notes attribuées aux différents films permettent de calculer le degré de préférence. Des préférences éloignées ne permettent pas de déterminer les goûts ou notes d'un autre utilisateur. Au contraire des préférences identiques indiquent une relation forte entre deux utilisateurs permettant ainsi de déterminer plus facilement les films à recommander et prédire leur note. Une relation modérée obtenue avec des préférence similaires permet aussi de calculer la note d'un film en fonction d'un delta plus ou moins précis.

	movi e 1	movi e 2	movi e 3	movi e 4	movi e 5	movi e 6	movi e 7	movi e 8	movi e 9	movi e 10	...	movi e 17770
user r 1			1		2							3
user r 2	2		3	3				4		2		
user r 3							5	3				
user r 4	2				3			2				2
user r 5		2		3				4		2		4
user r 6			2			5						
user r 7			2					4	2			
user r 8	3	1			3	4		5		4		
user r 9									3			
user r 10			1		2							2
...												
use r 480189	4				3			3				

Préférences identiques = poids fort

Préférences similaires = poids modéré

FIGURE 3.1 – Exemple algorithme du plus proche voisin.

La factorisation matricielle (Matrix factorization) fait partie des algo-

rithmes de filtrage collaboratif utilisé dans les systèmes de recommandation. Cette approche a prouvé son efficacité au cours du Netflix Prize. Les algorithmes de factorisation matricielle fonctionnent en décomposant la matrice d'interaction "utilisateur-élément" en un produit de deux matrices rectangulaires de plus faibles dimensions. Pour déterminer la note d'un film, il faut alors multiplier et additionner les valeurs pour obtenir la prédiction.

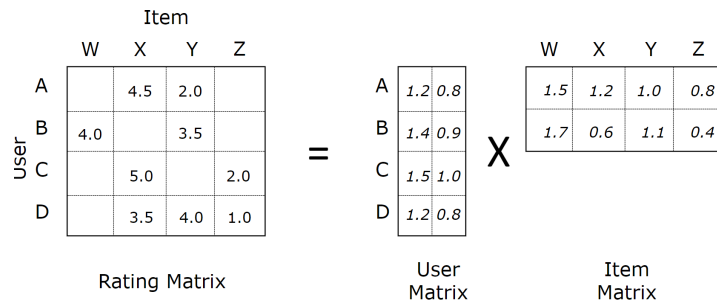


FIGURE 3.2 – Exemple de factorisation matricielle.

Ainsi dans l'exemple ci-dessus, la prédiction de l'utilisateur A avec l'item W est calculée de la manière suivante :

$$1,2 \times 1,5 + 0,8 \times 1,7 = 3,16$$

L'approche proposée par l'équipe gagnante est articulée autour de 3 axes. Le premier axe modélise les données selon différents niveaux (global, régional et locale). C'est la combinaison de l'extraction des schémas locaux qui sont ensuite factorisés au niveau régional puis affectent globalement les données. Le second axe, est la qualité du modèle, c'est un axe essentiel pour la robustesse permettant la dérivation et les itérations mais aussi d'éviter les effets de débordements lors des factorisations et de l'application d'algorithmes locaux. Afin d'être considéré comme une donnée de qualité et ainsi respecter les critères énoncés il faut que la donnée soit la plus simple possible. Le dernier axe, est une représentation binaire des données selon si les utilisateurs ont donné leur avis de façon implicite ou explicite. Les données sont caractérisées en fonction de leurs évaluations et de la manière dont elles ont été évaluées. Le comportement d'un utilisateur est considéré comme implicite s'il est facile à collecter par son historique de navigation, son historique de location, ses recherches etc... Le retour implicite permet de prédire les évaluations pour

les utilisateurs qui n'ont pas encore évalué un film. Quant au retour explicite de l'utilisateur, c'est lorsque ce dernier donne son avis sur le film ou la série par un système de notation (notation du film par une quantité d'étoiles ou bien un symbole de pouce vers le haut ou bien vers le bas).

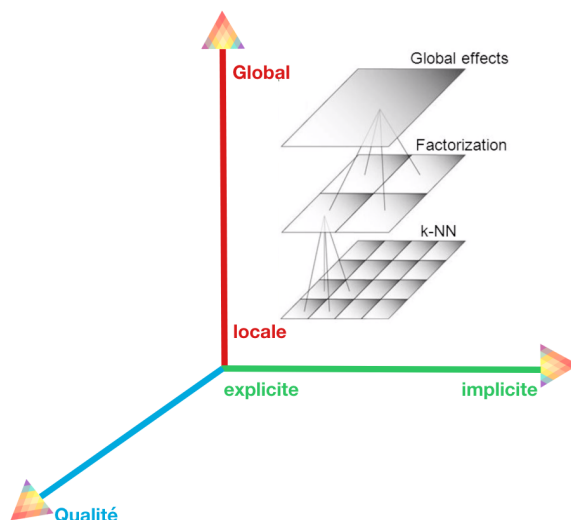


FIGURE 3.3 – Représentation simplifiée de la solution BellKor's Pragmatic Chaos .

### 3.3 Critique

Les données mises à disposition pour le Netflix Prize sont issues de réels utilisateurs de Netflix et anonymisées pour des soucis de vie privé. Cependant en 2007, deux chercheurs de l'Université du Texas<sup>15</sup> ont réussi à identifier des utilisateurs de façon individuelle par correspondance avec les ensembles de données issues de IMDb<sup>b</sup>.

Après une simple étude des données fournies pour réaliser le Netflix Prize, plusieurs choses apparaissent immédiatement comme évidentes. Premièrement, l'ensemble des données fournies est immense avec des variations de

---

b. L'Internet Movie Database (littéralement « Base de données cinématographiques d'Internet »), abrégé en IMDb, est une base de données en ligne sur le cinéma mondial, sur la télévision, et plus secondairement les jeux vidéos.

notes extrêmes entre les différents utilisateurs. Les variations de note ne permettent pas de déterminer un groupe d'utilisateur semblable facilement. Et deuxièmement, le fichier permettant de s'entraîner et de se qualifier a des propriétés différentes ce qui rajoute une complexité supplémentaire.

Concernant l'utilisation du code de l'équipe gagnante du Netflix Prize, Xavier Amatriain et Justin Basilico travaillant sur la Science de personnalisation chez Netflix déclarent<sup>1</sup> dans Medium<sup>c</sup> :

«A year into the competition, the Korbell team won the first Progress Prize with an 8.43% improvement. They reported more than 2000 hours of work in order to come up with the final combination of 107 algorithms that gave them this prize. And, they gave us the source code. We looked at the two underlying algorithms with the best performance in the ensemble : Matrix Factorization (which the community generally called SVD, Singular Value Decomposition) and Restricted Boltzmann Machines (RBM). SVD by itself provided a 0.8914 RMSE, while RBM alone provided a competitive but slightly worse 0.8990 RMSE. A linear blend of these two reduced the error to 0.88. To put these algorithms to use, we had to work to overcome some limitations, for instance that they were built to handle 100 million ratings, instead of the more than 5 billion that we have, and that they were not built to adapt as members added more ratings. But once we overcame those challenges, we put the two algorithms into production, where they are still used as part of our recommendation engine.».

Finalement, même si l'équipe «BellKor's Pragmatic Chaos» a remporté la compétition par un mélange de différents modèles, l'agrégat de plusieurs modèles s'est avéré beaucoup trop coûteux en temps et en énergie pour être mis en place. Dans le reste de l'article, les ingénieurs ont mis en avant qu'entre 2006 (date du début de la compétition) et 2009 (fin du concours) les problématiques de Netflix ont changé. Il se pose donc la question de la pertinence d'avoir laissé perdurer cette compétition aussi longtemps. Cependant, selon la citation, les ingénieurs ont déterminé des sous-algorithmes pertinents au sein du travail de l'équipe gagnante. Ce qui a permis d'améliorer les algorithmes de Netflix existants et qui actuellement sont toujours utilisés.

---

c. Medium est une plateforme web de blog créée en août 2012 par Evan Williams et Biz Stone, les fondateurs de Twitter et Blogger.



## Chapitre 4

# MON APPORT

L'équipe qui à remporté le « Netflix Prize » regroupe plusieurs dizaines de personnes au sein de différentes équipes initiales accumulant plusieurs centaines d'heures de travail et de recherche. Il est évident, que seul et ayant des connaissances limitées des algorithmes et des systèmes de recommandation sur le plan technique, je ne peux rivaliser avec Cinematch ou encore l'équipe gagnante du « Netflix Prize » .

En étudiant la structure de données fournie par Netflix, je me suis posé la question de l'application à une plus grande échelle. Devant le nombre grandissant des plateformes permettant de regarder du contenu mais aussi le suivi sur une chaine de télévision ou via des téléchargements illégaux. Je pense qu'il est possible de générer un algorithme de recommandation pour une plateforme généraliste permettant ainsi de suivre des séries et des films et d'en découvrir des nouveaux.

Ainsi, lors de l'analyse des données, je me suis rapidement aperçu qu'un manque d'informations sur les films était présent comme notamment le genre associé (Dramatique, Romantique, Comédie, ...). Cette information est pourtant essentielle et permet de déterminer avec plus de précision si un utilisateur à une plus forte appétence pour un genre. Je pense que cette information est primordiale lors de la recherche du plus proche voisin, permettant ainsi de rajouter une étape tout en confirmant ou non si la relation entre deux voisins est proche ou extrêmement proche. Il en découle ainsi une augmentation de la précision lors de l'estimation d'un avis, d'une note ou simplement d'une recommandation.

Ainsi, pour démontrer la pertinence de mon raisonnement je me suis basé uniquement sur deux types de données. Premièrement, un fichier contenant

la liste des 25 plus gros succès du box-office Nord-américain, structuré de la façon suivante :

...  
 Movie2,1977,Star Wars episode IV : Un nouvel espoir,Science-Fiction  
 ...

Ici, la donnée importante n'est pas le film en lui-même mais le genre principal associé. Par exemple, "Avatar" est un film de Science-Fiction alors que "Titanic" est un film de Romance. Secondement, il me fallait une structure de données pour les utilisateurs du système. Afin de me simplifier la tâche, un utilisateur est représenté comme une liste de note. Les notes sont comprises de 0 et 5, je considère donc que chaque film vu est obligatoirement noté entre 1 et 5 (1 étant un film non apprécié et 5 un film coup de cœur). Évidemment, un film ayant une note égale à 0 est considéré comme un film non regardé et par conséquent potentiellement recommandable.

Au cours du développement de ma solution, initialement j'ai décidé de créer une liste d'utilisateur aléatoirement. Cependant, une génération aléatoire ne permet pas à chaque fois de tomber sur un cas suffisamment pertinent pour démontrer l'efficacité de ma solution dans certains cas précis. Ainsi, j'ai donc décidé de m'orienter vers une liste d'utilisateur défini et d'appliquer mon système de recommandation sur un unique utilisateur permettant de reproduire le cas concret d'un utilisateur qui se connecte sur Netflix. Connaissant les données, je connais aussi le résultat à atteindre vérifiant ainsi la véracité du système. Ci-dessous un schéma des données utilisateurs utilisé :

	Movie1	Movie2	Movie3	Movie4	...	Movie20	...	Movie25
Fabien	0	3	4	5		3		0
Mathieu	0	0	4	0		3		1
Marjolaine	0	3	0	5		0		4
Croc	0	3	0	5		0		1

Comédie	
SF	

FIGURE 4.1 – Schéma des données utilisateurs utilisé.

Hormis le profil de « Fabien », nous avons 3 profils utilisateurs. L'utilisateur « Mathieu » est très client des comédies, il a consommé deux films de types Comédie et un film de Science-Fiction le film « movie25 » qu'il

n'a malheureusement pas apprécié. Les profils « Marjolaine » et « Croc » sont friands des films de Science-Fiction, seul leurs notes attribuées au film « movie25 » diffèrent.

Hormis son manque de notation pour « movie25 », le profil « Fabien » est similaire aux autres profils consommant ainsi des films de Science-fiction et de Comédie. Mon système de recommandation crée en Python aura pour but de prédire la note que Fabien attribuera au film numéro 25 correspondant à "Jurassic World".

## 4.1 Recherche des plus proches voisins

Rechercher les plus proches voisins est une méthode très efficace pour créer un groupe d'utilisateur avec des consommations identiques permettant ainsi une recommandation rapide pour un utilisateur donné. Pour rappel, cette méthode a démontré son efficacité lors du Netflix Prize.

Concrètement, je prends le « profil Fabien » et je calcule la distance euclidienne entre les notes attribuées par Fabien et les autres utilisateurs. Mon algorithme des plus proches voisins à une condition supplémentaire, la distance entre Fabien et un autre utilisateur est calculée pour chaque film individuellement et seulement si les deux personnes ont regardé le film sélectionné. La valeur cumulée de la distance euclidienne pour chaque film permet d'obtenir une distance entre Fabien et les autres utilisateurs. A la fin de cette étape, j'ai une liste de profil organisée de la façon suivante :

**Distance : données de l'utilisateur Mathieu**

Ensuite, je trie cette liste pour obtenir une liste ordonnée de façon croissante sur la distance. Mais avec ce jeu de données réfléchi, la distance entre Fabien et les autres profils est identique avec une valeur égale à 0, ce qui implique des profils similaires. Au préalable, j'ai défini une limite d'acceptation, dans le cadre de ce mémoire, j'ai choisi de garder les 3 profils les plus proches de l'utilisateur Fabien.

Ainsi, à la fin de cette étape, les profils : « Marjolaine, Croc et Mathieu » sont considérés comme des voisins proches sur lequel une recommandation est possible pour Fabien.

	Movie1	Movie2	Movie3	Movie4	...	Movie20	...	Movie25	Distance
Fabien	0	3	4	5		3		0	-
Mathieu	0	0	4	0		3		1	0
Marjolaine	0	3	0	5		0		4	0
Croc	0	3	0	5		0		1	0

Comédie	
SF	

FIGURE 4.2 – Recherche des plus proches voisins.

## 4.2 Trier les voisins selon le film

En partant des données pour Fabien et sa liste des plus proches voisins, mon système de recommandation va essayer de recommander un ou plusieurs films pour Fabien. Techniquement, pour chaque film non regardé par Fabien, je vais rechercher si parmi ses voisins les plus proches au moins un voisin a regardé ce film. Dans ce cas précis, seul le film "Jurassic World" (« Movie25 ») est éligible.

Premièrement, je vais enlever de la liste des voisins les plus proches ceux qui n'ont pas regardé ce film. Dans notre cas très précis, aucun voisin n'est évincé. Deuxièmement, je vais appliquer une variante de l'algorithme des plus proches voisins mais basé sur le type du film. Concrètement, "Jurassic World" est un film de Science-Fiction et je vais calculer la note moyenne attribué par Fabien sur les films de type Science-Fiction. Cette moyenne est aussi calculée pour chaque voisin proche restant.

Finalement, il faut classer les voisins restant selon leur moyenne attribuée aux films de Science-fiction et non pas par ordre croissant mais selon si leur moyenne est proche ou non de celle de Fabien. Donnant le résultat suivant :

	Movie1	Movie2	Movie3	Movie4	...	Movie20	...	Movie25	Classement
Fabien	0	3	4	5		3		0	-
Mathieu	0	0	4	0		3		1	3
Marjolaine	0	3	0	5		0		4	1
Croc	0	3	0	5		0		1	2

Comédie	
SF	

FIGURE 4.3 – Classement des voisins.

### 4.3 Prédiction

Dans cette section, il faut prendre en compte plusieurs paramètres. Notamment si le genre du film sur lequel le système travaille à déjà au moins une évaluation par Fabien. On distingue deux cas :

Le premier cas, si Fabien a déjà évalué au moins un film de Science-fiction, alors il faut calculer la différence entre la moyenne attribuée par Fabien au genre Science-fiction et la moyenne du voisin sélectionné pour ce même genre. Dans le cadre de mon développement, j'ai transformé la différence entre ces deux moyennes en un multiplicateur appelé « delta ». Le delta vaut exactement 1 si les deux utilisateurs ont une moyenne identique, le delta est négatif si Fabien note plus faiblement que le voisin sélectionné et inversement le delta est positif si Fabien note plus généreusement. Finalement, pour prédire la note du film, il faut multiplier la note donnée par le voisin par le delta.

Le second cas, si Fabien n'a jamais attribué une note aux films de Science-fiction. Une moyenne donnée par Fabien pour le genre Science-fiction est impossible. J'ai choisi d'appliquer la même méthodologie que pour le premier cas mais sur les notes données de façon globale. Ainsi, je vais calculer la moyenne de toutes les notes données par Fabien et le voisin sélectionné. A partir des deux moyennes comme dans le premier cas, je vais créer un delta entre les deux moyennes. Finalement, pour prédire la note du film, il faut multiplier la note donnée par le voisin par le delta.

Dans notre cas, Fabien a déjà attribué des notes à des films de Science-fiction avec une moyenne de 4 pour ce genre de film. Le profil voisin le plus proche est Marjolaine avec une moyenne de 4 attribuée sur les films de Science-fiction. Ici, le delta entre les deux moyennes a pour valeur 1 et Marjolaine a attribué la note de 4 au film "Jurassic World". Par conséquent, le système prédit que Fabien va attribuer une note de 4 pour le film "Jurassic World".

Predict for Fabien

```

- - - - -
[0, 3, 4, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 4]
new rating : 4
['Movie25', '2015', 'Jurassic World', 'Science-Fiction']
- - - - -

```

## Chapitre 5

# ÉVALUATION DE MON APPORT

En choisissant de modifier les données fournies pour la réalisation du « Netflix Prize », je change un peu le paradigme de base. Ainsi, il ne faut plus uniquement se baser sur la création d'un groupe d'utilisateurs avec des notes similaires mais bien de faire coïncider les préférences des utilisateurs selon leurs genres préférés.

Une comparaison de mon approche avec celle obtenue par Netflix ou les vainqueurs du concours est compliquée au vu des différences entre les données utilisées. L'ajout d'un genre pour les films entraîne une complexité supplémentaire à la problématique de base. Ainsi, si les données utilisées pour la réalisation du « Netflix Prize » avaient contenu le genre de chaque film cela aurait ajouté une difficulté supérieure, des calculs supplémentaires via les algorithmes mais aussi des temps d'accès supplémentaires à la donnée pour délivrer le résultat. Dans ces conditions, il est facile de dire que les algorithmes utilisés et les méthodes appliquées pour résoudre la problématique du Netflix Prize auraient été drastiquement différentes, impactant par la même occasion le résultat obtenu.

Evaluer un système de recommandation permet de déterminer si ses performances sont reconnues vis-à-vis de l'objectif de départ, de la concurrence ou de l'existant (un système de recommandation déjà en place). Ici, le système appliqué repose sur des données différentes et ne permet pas d'évaluer sa performance vis-à-vis de la concurrence et de l'existant.

Mon implémentation reprend en partie la problématique du « Netflix Prize » avec une donnée supplémentaire, le genre du film associé. Mon im-

plémentation se présente comme une solution dans un cas très spécifique. Il s'agit du cas où plusieurs voisins ont une distance identique et permet donc d'apporter une information permettant de choisir le voisin le plus fiable pour créer une recommandation pertinente. Il s'agit d'une situation extrêmement spécifique, reproductible que dans de rares cas. Malgré le degré de complexité supplémentaire, ma solution permet de mettre en évidence une recommandation plus pertinente dans le cas des voisins similaires mais avec des goûts différents. Cette approche permet de répondre positivement à cette nouvelle problématique. Cependant, ma réponse à ce problème entraîne un coût supplémentaire en termes de temps et de complexité. Il serait important de se poser la question du coût de cette implémentation à plus grande échelle, sur un nombre de données beaucoup plus important et couplé à d'autres algorithmes comme le fait Netflix ou les différents participants du « Netflix Prize ». L'efficacité de ma solution peut être considéré comme néfaste et négligeable par rapport au nombre de cas où dans la réalité ma solution pourrait permettre une meilleure recommandation. De plus, les données fournies par Netflix son issues d'un choix réfléchi permettant de remettre en cause la pertinence d'avoir un genre associé à chaque film pour en faire une recommandation efficace.

## Chapitre 6

# ÉVOLUTIONS POSSIBLES

Dans ce chapitre, je vais décrire les évolutions possibles concernant le « Netflix Prize » puis celles de Netflix et enfin celles de mon apport.

Concernant le « Netflix Prize », l'équipe BellKor's Pragmatic Chaos est la première équipe à avoir remporté le Netflix Prize challenge avec un RMSE meilleur que l'algorithme Cinematch. Ce concours a permis une avancée considérable pour des systèmes de recommandation mettant en avant la puissance de la factorisation matricielle pour le filtrage collaboratif mais aussi l'émergence de nouveaux algorithmes pour le machine learning. Netflix a autorisé l'utilisation des données mise à disposition pour la recherche, ce qui permet encore actuellement des avancées et des recherches sur l'approche des systèmes de recommandation et leur évolutivité. De nombreuses personnes déposent leurs solutions sur des outils de partages collaboratifs comme GitHub par exemple. Le Netflix Prize ne permettra plus de générer autant d'enthousiasme de la part des chercheurs et passionnés qu'entre 2006 à 2009. Cependant, il n'est pas impossible qu'un nouvel algorithme issu de ce challenge bouscule le monde de la recommandation.

Au sujet de Netflix, avec l'arrivée de nouveaux concurrents, la plateforme doit perpétuellement se renouveler. Une grande partie de son budget est utilisé pour l'acquisition de droit sur des contenus, comme pour la série "Friend" mais aussi pour la création de contenus originaux. La création de contenus originaux est un choix logique et raisonné. En effet, de nombreux acteurs prennent leur indépendance comme Disney, ce qui induit une perte des droits pour Netflix de beaucoup de licence forte comme l'ensemble des films PIXAR. A court terme, il est logique d'imaginer que Netflix va utiliser ses points forts pour garder ses utilisateurs et en acquérir de nouveaux. Comme démontré tout au long de ce mémoire, la grande force de Netflix



est représentée par son système de recommandation à partir duquel 80% du contenu consommé provient. Il est logique de penser que Netflix va utiliser son système de recommandation pour mettre en avant des contenus exclusifs à la plateforme. Logiquement, les futures évolutions de Netflix côté software se baseront sur l'interface mais principalement sur l'amélioration continue de leur algorithme de recommandation permettant ainsi de diminuer petit à petit la marge d'erreur du système de recommandation bien quelle soit aujourd'hui déjà négligeable.

Pour finir, concernant mon apport, il existe plusieurs axes d'améliorations. Premièrement, une amélioration de la fiabilité et la qualité de mes algorithmes permettrait de pousser la prédiction plus loin. Une prédiction plus fiable permet de produire une note pratiquement aussi fiable que celle donnée explicitement par l'utilisateur et donc d'être utilisée pour prédire d'autres films. Par exemple, si un utilisateur n'a jamais regardé de Comédie et qu'une première prédiction a mis en évidence une prédiction fiable pour un film Comique, une seconde étape pourrait être l'exploitation de cette note pour recommander d'autres films dans le genre Comédie.

Secondement, j'ai essayé de démontrer que prendre en compte une seconde donnée tel que le genre du film permettait d'apporter une information pertinente et ce malgré la complexité supplémentaire. Je suis convaincu que l'utilisation de sous-genre comme les Comédies Romantiques permettrait d'augmenter la pertinence de la recommandation. L'utilisation de l'année du film serait sûrement une autre voie à explorer, permettant ainsi de déterminer si l'utilisateur préfère les films anciens ou plus récents.

Rétrospectivement en analysant mon travail, j'ai appliqué deux fois l'algorithme des plus proches voisins. Il serait sûrement plus pertinent d'utiliser plusieurs algorithmes sur plusieurs dimensions comme l'équipe gagnante du « Netflix Prize ». Je pense que l'agrégat des recommandations issu de plusieurs algorithmes offre une force supplémentaire afin d'obtenir une recommandation plus qualitative et plus pertinente.

# CONCLUSION

Aujourd'hui, les systèmes de recommandation sont déjà extrêmement présents dans notre quotidien.

Il semblerait que l'une des approches à envisager pour un grand nombre d'acteur du web soit une adaptation aux systèmes de recommandation. Un grand nombre de personnes ont déjà l'habitude d'utiliser ces systèmes aux quotidiens. Pour une entreprise, c'est un moyen efficace d'adapter son contenu face à chaque client permettant ainsi d'apporter une expérience unique et satisfaisante. Un internaute satisfait est un client potentiel pour chaque acteur du web désirant faire fructifier son business et son chiffre d'affaire.

C'est dans ce but que Netflix a introduit Cinematch, avec un nombre limité de DVD pour le dernier blockbuster à la mode, il ne fallait pas frustrer les abonnés et toujours avoir une recommandation suffisamment intéressante pour pallier cette pénurie de DVD. Au moment de sa digitalisation, les besoins de Netflix en matière de recommandation ont évolué. Aujourd'hui, 80% des contenus consommés sur la plateforme sont issus d'une recommandation, preuve de son importance au sein de son business. Avec un catalogue grossissant de mois en mois, un nombre d'utilisateurs toujours en forte hausse et des concurrents toujours plus nombreux, il est devenu primordial pour Netflix d'être toujours plus efficace et c'est pour cela que la plateforme a décidé d'utiliser la recommandation à outrance. C'est dans cette logique que la plateforme a lancé en 2007 la compétition « Netflix Prize » essayant avec le minimum d'information d'apporter la recommandation la plus pertinente possible. Cette compétition a permis une avancée formidable pour le monde de la recherche en terme de recommandation en mettant en évidence la force de certains algorithmes. Le rayonnement du « Netflix Prize » a été une grande publicité positive pour l'entreprise avec un gain double : le nombre d'abonné à la plateforme a drastiquement augmenté tandis que sa recommandation s'est affiné grâce aux travaux des différents participants.

## *Conclusion*

---

Pour finir, ce mémoire m'a permis d'étudier les systèmes de recommandation de Netflix sous différents aspects. Me permettant ainsi de comprendre fonctionnement de la recommandation autant sur le plan marketing que technique, son but et son importance pour l'entreprise et ses abonnés. J'ai moi-même essayé de mettre en place un algorithme efficace permettant de faire une recommandation pertinente dans un cas extrême. Preuve qu'il existe encore beaucoup d'améliorations et de recherches possibles autour de la recommandation afin de rendre les utilisateurs toujours plus accros aux différents services.

# BIBLIOGRAPHIE

- [1] Xavier AMATRIAIN et Netflix inc JUSTIN BASILICO. *Netflix Recommendations : Beyond the 5 stars (Part 1)*. URL : <https://medium.com/netflix-techblog/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429>.
- [2] Elodie Drouard et BENOÎT ZAGDOUN. *ENQUETE FRANCEINFO. COMMENT NETFLIX S'Y PREND POUR NOUS RENDRE ACCROS*. 22 mai 2019. URL : [https://www.francetvinfo.fr/culture/series/netflix/enquete-franceinfo-comment-netflix-sy-prend-pour-nous-rendre-accros\\_3189939.html](https://www.francetvinfo.fr/culture/series/netflix/enquete-franceinfo-comment-netflix-sy-prend-pour-nous-rendre-accros_3189939.html).
- [3] Robin BURKE. « Hybrid recommender systems : Survey and experiments. User Modeling and User-Adapted Interaction ». Thèse de doct. California State University, Fullerton, 2002.
- [4] Julien CADOT. *Netflix, un géant seul au monde*. 13 avr. 2016. URL : <http://www.numerama.com/pop-culture/162823-netflix-un-geant-seul-au-monde.html>.
- [5] Enguérand Renault CAROLINE SALLÉ. *Netflix dépasse la barre des 5 millions d'abonnés en France*. 24 avr. 2016. URL : <http://www.lefigaro.fr/medias/2019/02/13/20004-20190213ARTFIG00119-netflix-depasse-la-barre-des-5-millions-d-abonnes-en-france.php>.
- [6] Sameer CHHABRA. *Netflix says 80 percent of watched content is based on algorithmic recommendations*. 22 juin 2017. URL : <https://mobilesyrup.com/2017/08/22/80-percent-netflix-shows-discovered-recommendation/>.
- [7] A. Felfernig D. JANNACH M. Zanker. *Recommender Systems : An Introduction*. 2010.

- [8] Toby DAIGLE. *Introduction aux systèmes de recommandation*. 8 fév. 2017. URL : [http://penseeartificielle.fr/difference-intelligence-artificielle-machine-learning-deep-learning/#Le\\_machine\\_learning](http://penseeartificielle.fr/difference-intelligence-artificielle-machine-learning-deep-learning/#Le_machine_learning).
- [9] Alexander Tuzhilin GEDIMINAS ADOMAVICIUS. « Toward the Next Generation of Recommender Systems : A Survey of the State-of-the-Art and Possible Extensions ». In : *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* 17.6 (2005), p. 734–749. DOI : <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.695.6428&rep=rep1&type=pdf>.
- [10] Carlos A. GOMEZ-URIBE et Neil HUNT. « The Netflix Recommender System : Algorithms, Business Value, and Innovation. » In : *ACM Trans. Management Inf. Syst.* 6.4 (2016), 13 :1–13 :19. URL : <http://dblp.uni-trier.de/db/journals/tmis/tmis6.html#Gomez-UribeH16>.
- [11] Netflix INC. *Decoding the Defenders : Netflix Unveils the Gateway Shows That Lead to a Heroic Binge*. URL : <https://media.netflix.com/en/press-releases/decoding-the-defenders-netflix-unveils-the-gateway-shows-that-lead-to-a-heroic-binge>.
- [12] Yehuda KOREN. *Lessons from the Netflix Prize*. 8 fév. 2015. URL : <https://slideplayer.com/slide/4969876/>.
- [13] Yehuda KOREN. « The BellKor Solution to the Netflix Grand Prize ». In : *Netflix inc* (2009). DOI : [https://www.netflixprize.com/assets/GrandPrize2009\\_BPC\\_BellKor.pdf](https://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf).
- [14] James MACQUEEN. « SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS ». Thèse de doct. University of California, Los Angeles, 1967.
- [15] GARvind NARAYANAN et Vitaly SHMATIKOV. « How To Break Anonymity of the Netflix Prize Dataset ». In : *The University of Texas at Austin* (2007). DOI : <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.100.3581&rep=rep1&type=pdf>.
- [16] elsa NEGRE. *Systèmes de recommandation : Introduction*. 2015.
- [17] Harris Interactive NPA. *SURFER EN TOUTE LIBERTÉ-LE BAROMÈTRE SVOD*. 24 avr. 2016. URL : <https://harris-interactive.fr/wp-content/uploads/sites/6/2017/11/Barometre-Conso-SVoD-Harris-NPA-resultats.pdf>.

## BIBLIOGRAPHIE

---

- [18] Matt PATCHES. *Netflix says Fortnite is bigger competition than HBO or Hulu*. 17 jan. 2019. URL : <https://www.polygon.com/2019/1/17/18187400/netflix-vs-fortnite-hbo-hulu-competition>.
- [19] Martin PIOTTE et Martin CHABBERT. « The Pragmatic Theory solution to the Netflix Grand Prize ». In : *Netflix inc* (2009). DOI : [https://www.netflixprize.com/assets/GrandPrize2009\\_BPC\\_PragmaticTheory.pdf](https://www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf).
- [20] ELAINE RICH. « User Modeling via Stereotypes\* ». Thèse de doct. University of Texas, Austin, 1979.
- [21] Andreas TÖSCHER et Michael JAHNER. « The BigChaos Solution to the Netflix Grand Prize ». In : *Netflix inc* (2009). DOI : [https://www.netflixprize.com/assets/GrandPrize2009\\_BPC\\_BigChaos.pdf](https://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf).
- [22] Amy ZIPKIN. *Out of Africa, Onto the Web*. 17 déc. 2006. URL : <http://www.nytimes.com/2006/12/17/jobs/17boss.html>.

# TABLE DES FIGURES

1.1	Un système de recommandation collaboratif <sup>7</sup> . . . . .	5
1.2	Amazon et la recommandation collaborative. . . . .	5
1.3	Un système de recommandation basé sur le contenu <sup>7</sup> . . . . .	6
1.4	Un système de recommandation hybride <sup>7</sup> . . . . .	7
2.1	Exemple de la diversité des illustrations. <sup>2</sup> . . . . .	14
2.2	Exemple de score et profil pour le film "Good Will Hunting" .	14
3.1	Exemple algorithme du plus proche voisin. . . . .	19
3.2	Exemple de factorisation matricielle. . . . .	20
3.3	Représentation simplifié de la solution BellKor's Pragmatic Chaos . . . . .	21
4.1	Schéma des données utilisateurs utilisé. . . . .	24
4.2	Recherche des plus proches voisins. . . . .	26
4.3	Classement des voisins. . . . .	26

Université Paris-Nanterre  
200 Avenue de la République  
92000 Nanterre