République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahlab Blida

N° D'ordre :……

Faculté des sciences

Département d'informatique

Mémoire Présenté par :

Aggoune Yassine Boudani Abdelhamid

En vue d'obtenir le diplôme de master

Domaine : Mathématique et informatique

Filière : Informatique
Spécialité : Informatique
Option : Ingénierie de logiciel

Sujet :

Développement d'un système de recommandation pour voyageurs dans un contexte Big data

Soutenu le :

Mr. N.Chikhi	Président
Mr. S.Ferfira	Examinateur
Mme. N.Farah	Examinateur
Mme Zahra Fatima Zahra	Promotrice
Mme Hedjazi Badiâa Dellal	Encadrante

Promotion
2014 / 2015

Remerciements

*Avant toute chose, Nous remercierons **ALLAH** le tout puissant, de nous avoir donnée la force et la patience pour mener à terme ce travail*

*Nous présentons nos sincères remerciements à notre promotrice Mme **Zahra Fatima Zahra** et notre encadreur **Hedjazi Badiâa Dellal** pour avoir assuré le suivi de ce projet mais aussi pour leur soutien, leurs conseils et leur disponibilité.*

Nous tenons à remercier aussi tous nos enseignants qui ont participé à notre formation tout au long de notre cursus.

Nous tenons à remercier nos parents, nos familles et amis pour leur soutien tout au long de ce mémoire.

Nous remercions également toutes ces personnes sur internet qui ont partagé leurs connaissances et expériences et répondu à nos questions, ils nous ont été d'une grande aide.

Enfin, nous remercions tous ceux qui ont aidé de près ou de loin à l'élaboration de ce mémoire et à la réussite de ce projet.

Dédicaces Abdelhamid

Je dédie ce modeste travail d'abords à mes très chers parents

Aucun mot ne saurait exprimer mon grand amour, mon respect et ma reconnaissance pour tout ce que vous avez fait pour ma formation et ma réussite. Uniques et irremplaçables, aucun hommage ne pourrait être à la hauteur de vos sacrifices et de l'affection dont vous n'avez jamais cessé de m'entourer.

Puisse Dieu vous accorder santé et longue vie.

A Mes chers frères et sœurs

En leurs espérant le plein succès dans leur vie.

A toute ma famille.

A tous mes amis, En souvenir d'agréables moments passés ensemble en témoignage de notre amitié.

A tous mes professeurs.

A toute notre promotion.

A toute personne qui m'est chère.

« Que dieu vous garde Inchallah »

Avec tous mes sentiments de reconnaissance et de gratitude.

Boudani Abdelhamid

Dédicaces Yassine

C'est avec profonde gratitude et s'incères mots, que je dédie ce modeste travail de fin d'étude a mes chers parents, qui ont sacrifié leur vie pour ma réussite et éclairé le chemin par leurs conseils judicieux que dieu les protège.

Je dédie aussi ce travail à mon frère, a toute la famille et a tous mes professeurs qui m'ont enseigné, et a tous ceux qui nous sont chers

A mes meilleurs amis Hamdi Hamadi, Boudani Abdelhamid, Lakrib Ali, Riali Ishaak et à tous mes camarades de la promotion 2010.

A ma famille et toutes les personnes que j'aime.

Aggoune Yassine

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Contexte général

Le voyageur s'expose fréquemment à faire des choix, pour voyager dans des bonnes conditions et cela sans connaitre les bons endroits à visiter, les hôtels pour hébergement et aussi les compagnies aériennes à prendre. Il est donc recommandé de se renseigner avant d'embarquer dans un vol ou bien de réserver un hôtel.

C'est dans ce contexte que nous intervenons pour mettre un système de recommandation basé sur le filtrage collaboratif (est ce que tous les systèmes de recommandation utilisent le filtrage collaboratifs ? sinon vous dites pourquoi vous avez choisi le filtrage collaboratif). Les systèmes de recommandations tentent d'anticiper les besoins des utilisateurs et leur proposent des items qu'ils sont susceptibles d'apprécier contrairement aux systèmes de recherche d'information qui attendent la requête de l'utilisateur pour agir. En effet, notre objectif est de recommander à des utilisateurs des items (vol, hôtel) sélectionnés parmi un large choix et censés être appréciés par eux. Notre système tente de prédire si un utilisateur donné appréciera ou non un item. Pour parvenir à un tel but, un système de recommandation a besoin d'accumuler les offres disponibles et aussi des données sur les utilisateurs pour lui recommander des offres selon son profil.

Problématique

Les utilisateurs subissent une surcharge informationnelle et des recommandations à tort et à travers liée à la multitude de ressources présentes dans notre système. C'est sur cet aspect particulier que nous nous focaliserons dans ce mémoire : améliorer la qualité des recommandations.

Pour qu'un système de recommandation soit efficace, il doit prendre en considération le contexte temporel. C'est une notion très vaste qui peut désigner par exemple l'emploi du temps de l'utilisateur, ou encore la périodicité dans son comportement. Exemple, les appréciations d'un utilisateur ce différencient selon le temps, donc il faut recommander des offres intéressantes sinon l'utilisateur peut abandonner s'il n'est pas convaincu du résultat à court ou à moyen terme.

Objectif

L'objectif de ce projet est de développer une application de prédiction des besoins des voyageurs et de recommandation dans un contexte BigData. Pour pallier les problèmes de la surcharge informationnelle de l'utilisateur et aussi le contexte temporel, nous proposons de ne recommander aux utilisateurs que des ressources pertinentes en se basant sur le filtrage collaboratif et aussi en temps réel.

Pour présenter au mieux notre travail, nous avons structuré ce document en quatre parties. Avant d'entamer ces quatre parties nous avons introduit le contexte de notre étude et fixé la problématique ainsi que les objectifs du projet.

Le contenu des quatre parties peut être résumé comme suit :

- **Partie 1 Etat de l'art BigData:** Il s'agit d'une partie théorique où nous présentons les notions du BigData dans les domaines des systèmes décisionnels et son impact.
- **Partie 2 Etat de l'art sur les systèmes de recommandation:**

Il s'agit d'une partie théorique où nous présentons les systèmes de recommandation.

- **Partie 3 Etat de l'art sur analyse prédictive:**

Il s'agit d'une partie théorique où nous présentons l'analyse prédictive .

- **Partie 4 Conception de la solution :** Dans cette partie, nous présentons les différentes étapes de la conception de la solution qui consiste la modélisation, présenter l'architecture conceptuelle qui permettra de faire des recommandations sur des offres de vol et d'hôtel pour voyageur aérien
- **Partie 5 Implémentation :** C'est la dernière partie de notre projet où nous procédons à l'implémentation de la solution conçue. Nous débutons avec la présentation de l'environnement technique et fonctionnel et des outils utilisés. Nous clôturons notre document avec une conclusion générale où nous synthétisons notre travail.

Chapitre I

Big Data

Introduction:

Le Big Data est un terme que l'on entend partout depuis quelques années. Il est en train d'envahir tous les secteurs (entreprises, universités, administrations publiques,…). Derrière ce terme se cache en réalité une myriade de technologies et dont le but est la manipulation de gros volumes de données.

La raison pour laquelle le Big Data commence à prendre de l'ampleur est l'augmentation de la quantité de données, due entre autres à l'augmentation des sources de données (blogs, médias sociaux, recherches sur internet, réseaux de capteurs, etc.).

En effet, la vraie question réside dans l'utilité du volume de données. Ces dernières n'ont pas forcement de valeur en elles-mêmes, mais quand on regroupe les unes avec les autres il y'a une mine d'informations utilisées pour des fins d'analyses. En entreprise, la mise en place d'outils Big Data répond généralement à plusieurs objectifs :

- Améliorer l'expérience client.
- Optimiser ses processus et sa performance opérationnelle.
- Renforcer ou diversifier son business model.

1. Présentation du « Big Data » :

1.1. L'historique du « Big Data » :

L'expression "Big Data" (ou grosse donnée, ou données volumineuses) est apparue pour la première fois en 2008 : elle a émergé car la quantité de données à traiter ces dernières années est sans équivalent avec ce qui se passait il y a seulement 10 ans et augmente de manière explosive.

Sans que tous les chiffres avancés soient aussi spectaculaires, les observateurs s'accordent à constater une croissance exponentielle des volumes de données, liée à un besoin de numérisation à tout crin des documents en tous genres. Les entreprises capturent désormais quotidiennement des milliards de milliards d'octets dans tous les domaines, depuis des données clients ou fournisseurs jusqu'aux données opérationnelles ou contractuelles, sans oublier les millions de capteurs disséminés à travers tous les réseaux, dans des unités embarquées dans les véhicules ou les téléphones mobiles, qui eux-mêmes recueillent, transforment, créent et communiquent des données.

1.2. Définition du « Big Data » :

Il *y a plusieurs définitions existantes sur le « Big Data ». La définition la plus précise donnée par* Claude Bernard [14] , Directeur R&D Technologies et Innovation chez Sage, explique le « Big Data » comme suit:

Le terme vient d'un anglicisme. Il caractérise l'ensemble des données produites à partir de diverses sources se multipliant à l'infini. Internautes, entreprises, secteur public, associations… tous concourent à ce phénomène au travers des réseaux sociaux et des blogs. L'humain produit donc de l'information. En parallèle, il existe une autre source d'alimentation massive: les machines comme les satellites, les capteurs divers ou la vidéo protection par exemple. C'est ce qu'on appelle la source machine to machine, ou M2M, qui recueille des masses informatives pour les envoyer à d'autres machines à des fins d'analyse.

Toutes ces informations représentent un véritable trésor qui permet, une fois analysées, dē répondre à différentes questions. En effet, les entreprises, de toutes tailles, les conservent et s'en servent en vue d'en faire des déductions. Il est possible aujourd'hui d'analyser des comportements et des usages au travers de l'historique des données comme la consommation en fonction des profils d'utilisateurs et de créer ensuite des produits et services adaptés aux différents usages. On peut juxtaposer des informations qui jusqu'alors n'étaient pas associées, voire créer des lois de coïncidence pour affiner et disposer d'une visibilité plus juste des activités que ce soit pour une entreprise ou une administration[14] .

1.3. Caractéristiques du « Big data » :

Les trois V du « Big data » :

- **Vitesse** : la vitesse à laquelle les données sont générées, capturées puis traitées simultanément. Les entreprises génèrent plus de données dans des temps beaucoup plus courts et doivent les capturer et traiter en temps réel sinon ces données n'ont alors déjà plus aucune valeur puisque le cycle de génération de nouvelles données a déjà commencé.
- **Variété** : l'origine de la variation des données c'est qu'elles proviennent des sources de données non structurées (formats, codes, langages différents...), comme les médias sociaux, les interactions Machine to Machine et les terminaux mobiles. Ils créent une très grande diversité au-delà des données transactionnelles traditionnelles.
- **Volume** : la quantité de données générées par des entreprises ou des personnes peut dépasser aisément les téraoctets de données.

Il est donc nécessaire pour les entreprises de comprendre ces caractéristiques. Gérer des données de plus en plus nombreuses, de plus en plus diverses, et de plus en plus rapides nécessite d'obtenir une meilleure représentation de l'interaction des clients avec l'entreprise. Avoir une meilleure compréhension de ce que les clients aimeraient réaliser à chaque point de contact permet de minimiser le risque de perdre ces clients lors du passage d'un point de contact vers un autre et garantir la pertinence de l'information qui leur est délivrée. Ainsi, pour améliorer à la fois la qualité de service, aspect clé pour les clients, et le taux de transformation des informations sur ces clients, il est important pour l'entreprise de ne pas perdre de vue les 3 V du Big Data.

1.4. Les revenus du Big Data

Le marché Big Data, mesuré par les recettes des fournisseurs provenant de la vente de matériel connexe, logiciels et services a atteint $18,6 milliards pour l'année civile 2013 dans le monde, Ce qui représente un taux de croissance de 58% par rapport à l'année précédente.

Ventilés par type, les revenus des services Big Data liés constituent 40% du marché total, suivi par le matériel à 38% et à 22% des logiciels. Cette ventilation est due en partie à la nature open source de beaucoup de logiciels Big Data et modèles d'affaires connexes des fournisseurs BigData . Ainsi, il est nécessaire, pour les services professionnels d'aider les entreprises à identifier les cas Big Data et les solutions pour maintenir la performance [35] .

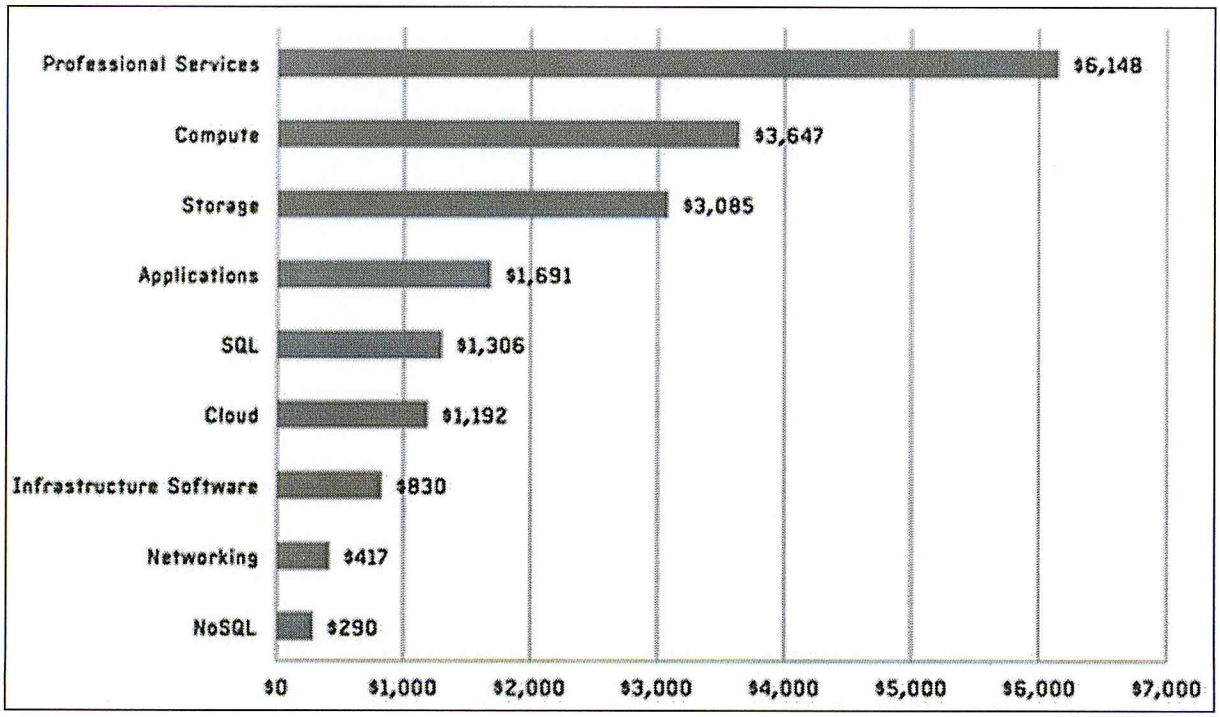

Figure 1: Revenue big data par type en 2013 (US -Billion $)[35]

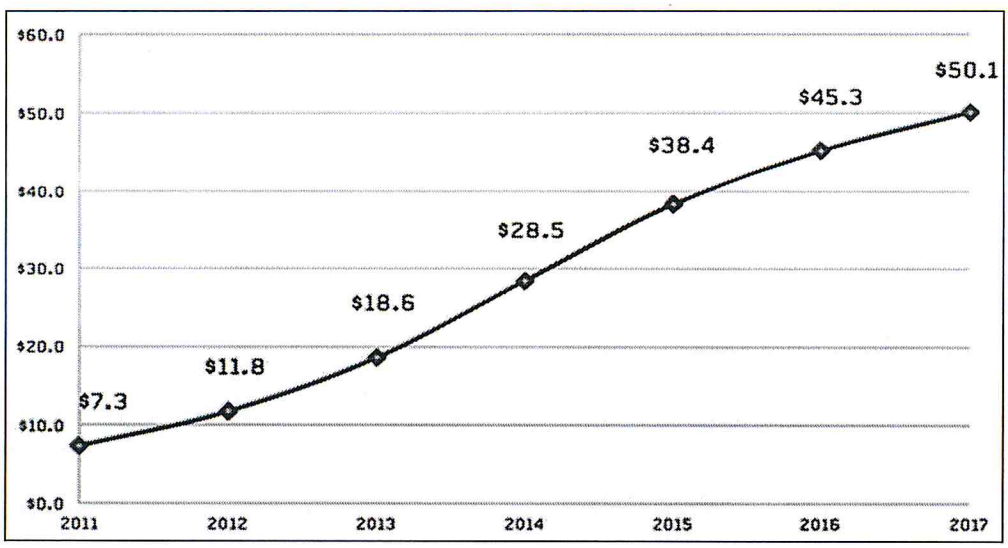

Figure 2: Prévisions des marchés big data (US-Billion$) 2011-2017 [35]

1.5. Les secteurs qui s'intéressent au Big Data

Beaucoup estiment que le Big Data n'est qu'un passage à l'échelle des traitements traditionnels alors que tous les secteurs sont concernés. Les premiers à s'intéresser au Big Data sont:

- **Dans la distribution et les télécoms** : Le Big Data permet de bien connaître les clients à la fois par leur comportement en boutique, mais aussi en analysant leur activité sur internet, y compris sur les réseaux sociaux. Anticiper leurs besoins pour cibler des offres personnalisées est devenu le *«must do »* du marketing tiré par les données [49] .
- **Dans le secteur de la santé**: Les perspectives de la recherche fondamentale et du ciblage des médicaments sont importantes. Les données sont essentielles à l'analyse des médicaments avant leur mise sur le marché, en phase de tests, ou pour mesurer leur efficacité une fois sur le marché.
- Les nouveaux appareils connectés qui mesurent en permanence notre rythme cardiaque, notre niveau de glycémie, les calories brulées, etc., génèrent des flux d'information qui vont améliorer la prévention et réduire les coûts d'hospitalisation, en effectuant les mesures en ligne. [49]
- **Le secteur banque et finance:** est un consommateur de modèles mathématiques permettant de mieux cibler les produits financiers et surtout de suivre l'analyse du risque. Une banque de détail pourra affiner par exemple les points des clients pour les conditions d'octroi de prêt, optimiser ses actions commerciales ciblées, mais surtout mieux lutter contre la fraude. Les assureurs vont également tenter de réduire la fraude en détectant des signaux faibles, mais aussi par exemple optimiser leurs tarifs auto en utilisant des données de capteurs situés dans le véhicule [49] .

1.6. Les technologies derrière la révolution du Big Data

L'univers technologique du Big Data s'appuie sur des outils bien identifiés qui constituent la base innovante de ce mode de traitement. A eux seuls, ces outils résument le vocabulaire technologique du Big Data et en constituent la référence.

1.6.1. Hadoop :

Hadoop est aujourd'hui la plateforme de référence permettant l'écriture d'applications de stockage et de traitement de données distribuées en mode batch. L'idée principale derrière Hadoop est que les données sont automatiquement distribuées dans le cluster par HDFS. Les traitements doivent s'effectuer au plus près de la donnée (ce que permet Mapreduce). Les transferts de données sont ainsi réduits au minimum.

Hadoop est écrit en Java et soutenu par plusieurs startups américaines. Il est en outre devenu une sorte de standard de fait pour l'écriture d'applications de traitement de données ralliant l'ensemble des acteurs majeurs du secteur [45] .

Comme la plupart des Frameworks, Hadoop comporte plusieurs composants:

- **Hadoopdistributed File System (HDFS)**: le HDFS est un système de fichiers distribué, extensible et portable développé par Hadoop à partir du Google File System (GFS). Écrit en Java, il a été conçu pour stocker de très gros volumes de données sur un grand nombre de machines équipées de disques durs banalisés. Il permet l'abstraction de l'architecture physique de stockage, afin de manipuler un système de fichiers distribué comme s'il s'agissait d'un disque dur unique[4] .
- **Hadoopmapreduce**: Un modèle de programmation pour le traitement de données à grande échelle.Tous les modules de Hadoop sont conçus avec l'hypothèse fondamentale que les pannes sont fréquentes, soit d'une machine individuelle ou d'un ensemble de machines. L'architecture Hadoop corrige automatiquement ces situations**[4]** .

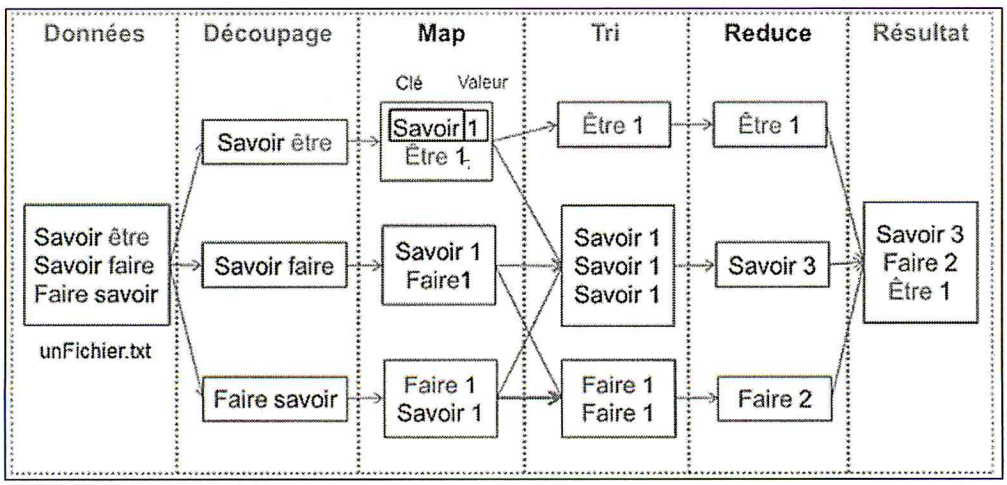

Figure 3 : Processus d'un traitement MapReduce – exemple d'un compteur de mots[52]

Les caractéristiques de Hadoop:

- **Evolutif**, car pensé pour utiliser plus de ressources physiques, selon les besoins, et de manière transparente.
- **Rentable**, car il optimise les coûts via une meilleure utilisation des ressources présentes.
- **Souple**, car il répond à la caractéristique de variété des données en étant capable de traiter différents types de données.
- Et enfin, **Résilient**, car pensé pour ne pas perdre d'information et être capable de poursuivre le traitement si un nœud du système tombe en panne.

1.6.2. Spark :

Spark Apache est un open source, le cadre de traitement de données parallèle qui complète Apache Hadoop pour le rendre facile de développer rapidement, unifiées applications Big-Data combinant lot, en streaming, et d'analyse interactifs sur toutes vos données. Spark dispose d'un moteur d'exécution DAG avancée qui prend en charge les flux de données cyclique et en mémoire informatique.

Les caractéristiques de spark:

- **Facilité d'utilisation**

Écrire des applications rapidement en Java, Scala ou Python. Spark propose plus de 80 opérateurs de haut niveau qui font qu'il est facile de construire des applications parallèles. Il est utilisé de manière interactive à partir de coquilles Scala et Python.

- **Généralité**

Combine SQL, streaming, et des analyses complexes. Spark est une pile d'outils de haut niveau, contenant SQL Spark, MLLib pour l'apprentissage de la machine, graphx et Spark streaming. Il est possible de combiner ces bibliothèques de façon transparente dans la même application.

- **Court partout**

Spark fonctionne sur Hadoop, Mesos, autonome ou dans le nuage (cloud). Il peut accéder à diverses sources de données, y compris HDFS, Cassandra, Hbase, S3.Spark s'exécute facilement en utilisant son mode cluster autonome, sur EC2, ou l'exécuter sur Hadoop fil ou Mesos Apache. Il peut lire des HDFS, Hbase, Cassandra, et toute source de données Hadoop.

1.7. L'impact du Big Data sur les entreprises publiques

1.7.1. Les enjeux du Big Data

Par son ampleur et par ses nombreuses promesses le BIG DATA a rapidement attiré l'attention des entreprises grâce à ses enjeux.

Le Big Data permet de valoriser les Péta-octets de données, ou à explorer la valeur cachée dans l'immensité du contenu non structuré comme les fichiers, les emails, ou les pages web.

- **En marketing :**

C'est tout le secteur qui se trouve renouvelé. Le Big Data permet en effet aux professionnels du secteur de connaître leur client « à 360° », c'est-à-dire à la fois par son parcours internet mais également par ses achats en magasin ou ses préférences affichées sur les réseaux sociaux. Anticiper les besoins de celui-ci et cibler des offres personnalisées ou encore l'analyse de sentiment pour la détection de comportements sur les réseaux sociaux. Le marketing se fait de plus en plus prédictif avec le Big Data, et l'on assiste à une éclosion de nouveaux modèles statistiques davantage inductifs.

- **Dans le domaine du pilotage de l'entreprise :**

Sur cet aspect, les usages sont également nombreux et porteurs d'innovation. En assurant une circulation immédiatement généralisée de l'information sur l'activité, le Big Data laisse entrevoir une optimisation complète des processus et des ressources métiers. Il réduit de facto le temps de réaction face à des erreurs ou des pannes et permet d'ajuster en permanence les équilibres offre-demande et temps-ressource. C'est une promesse importante dans des secteurs comme ceux de l'énergie ou des transports qui sont constamment portés par la logique de flux. Outre une réduction importante des coûts, le Big Data permet ici d'identifier au plus près les moteurs de l'activité, ce qui n'était pas possible avec les indicateurs traditionnels, soumis à des délais de latence bien plus importants.

- **Pour la Recherche :**

Domaine d'application originel du Big Data, l'apport de celui-ci est assez évident. En autorisant le traitement de multitudes de données, le Big Data permet à la science de réaliser des avancées importantes, lorsqu'il s'agit d'explorer l'infiniment petit (ex : exploration géologique), de croiser des données complexes (ex : imagerie) ou d'effectuer des simulations (ex : domaine spatial). C'est d'ailleurs en génétique que le Big Data a fait ses premières armes car ce secteur réclamait une approche à la fois quantitative et qualitative avancée.

- **Dans le domaine de l'Information :**

Le traitement des Big Data a profondément modifié la donne. Pour une requête donnée, il est désormais possible d'accéder à un croisement d'informations très disparates, issues de sources jusque-là négligées. L'instantanéité des réseaux sociaux est à ce titre une innovation de taille. L'analyse des tweets est devenue une source de renseignements courante pour comprendre les comportements ou les goûts de populations segmentées. De plus, au-delà de la compréhension de phénomènes, ne pas s'intéresser au BIG DATA aujourd'hui, c'est peut être risqué demain de perdre en compétence et d'être en retard sur son marché. Le BIG DATA comme toute avancée technologique, peut comporter des risques, qu'il ne faut surtout pas ignorer. En effet, le BIG DATA repose sur la confiance du consommateur et toute rupture dans cette confiance entrainerait automatiquement un retour en arrière. De la même façon, on craint que le Cloud ne soit pas assez protecteur. Il est donc urgent de maitriser ces risques pour garder la confiance des consommateurs. Cela nécessite d'avoir les compétences nécessaires par le recrutement de personnel qualifié.

1.7.2. Modèles d'organisation « Big Data » dans l'entreprise:

Les modèles d'organisation privilégiés par l'entreprise sont soit centraliser les données ou bien disposer d'architectures réparties au sein des directions métiers.

Trois modes d'organisation sont envisageables :

- **Une option « centralisée » :** Dans laquelle toutes les compétences sont regroupées au sein d'une entité transverse, sorte de Centre de Services Big Data au service des Métiers. En centralisant les ressources, on mutualise les coûts et on évite a priori la duplication des efforts, des données, et des budgets.
- **Une vision « décentralisée »:** Où ce sont les Métiers qui gardent la main en gérant leurs projets, leurs compétences, pour satisfaire au plus près leurs objectifs.
- **Une vision « externalisée » :** Dans laquelle l'entreprise confie à un prestataire spécialisé la gestion des données et des traitements associés.

2. Conception d'un entrepôt de données

Le « Big Data » permet d'enrichir les informations stockées dans l'entrepôt de données (Data-Warehouse). Le DW est une base de données utilisée dans le cadre décisionnel. Une plateforme Big-Data n'est pas de modifier à l'entrepôt les données, mais bien de le compléter et plus exactement de compléter les sources de données dont il se nourrit (s'alimente).

2.1. Phases de construction d'un DW:

Il y'a trois parties interdépendante qui relève la construction d'un Data Warehouse[2] :

- L'étude préalable qui va définir les objectifs, la démarche à suivre, le retour sur investissement,…)
- L'étude du modèle de données qui représente le DW conceptuellement et logiquement.
- L'étude de l'alimentation du Data Warehouse.

2.1.1. Etude préalable

- ✓ Etude des besoins:
  - Définir les objectifs du DW.
  - Déterminer le contenu du DW et son organisation, d'après:
    - Les résultats attendus par les utilisateurs.
    - Les requêtes qu'ils formuleront.
    - Les projets qui ont été définie.
  - Recenser les données nécessaires à un bon fonctionnement du DW:
    - Recenser les données disponibles dans les bases de production.
    - Identifier les données supplémentaires requises.
  - Choisir les dimensions
    - Typiquement: le temps, le client, le produit, le magasin...
  - Choisir les mesures de fait
    - De préférences de quantités numériques additives.
  - Choisir la granularité des faits
    - Niveau de détails des dimensions.
- ✓ Coûts de déploiement:
  - Nécessite des machines puissantes, souvent une machine parallèle
  - Capacité de stockage très importante (historisation des données)
    - Evaluer la capacité de stockage.
  - Equipes de maintenance et d'administration
  - Les coûts des logiciels
    - Les logiciels d'administration du DW.
    - Les outils ETL (Extract-Transform-Loading).
    - Les outils d'interrogation et de visualisation.
    - Les outils de Datamining.

2.1.2. Modélisation multidimensionnelle

- ✓ Niveau conceptuel:
  - ➢ Un DW est basé sur une modélisation multidimensionnelle qui représente les données dans un cube.
  - ➢ Un cube permet de voir les données suivant plusieurs dimensions:
    - Tables de dimensions.
    - La table des faits contient les mesures et les clés des dimensions.
- ✓ Niveau Logique:
  - ➢ Plusieurs schémas types sont proposés pour représenter un DW: Schéma en étoile, et Schéma en flocon.
- ✓ Schéma en étoile
  - ➢ Une (ou plusieurs) table(s) de faits : identifiants des tables de dimension ; une ou plusieurs mesures.
  - ➢ Plusieurs tables de dimension : descripteurs des dimensions.
- ✓ Schéma en flocons
  - ➢ Raffinement du schéma étoile avec des tables normalisées par dimensions.

2.1.3. Alimentation

L'alimentation de l'entrepôt de données est une étape importante dans le projet décisionnel, car elle garantit la pertinence et la qualité des données que contient l'entrepôt et assure au décideur l'accès à la bonne information. L'alimentation se fait grâce à l'opération d'ETL. Concrètement on dispose de sources souvent hétérogènes à partir desquelles sont extraites des données pour alimenter un entrepôt pour des fins d'analyse. Les sources peuvent être des bases de données, des fichiers (Csv, Excel, Txt, Xml, ...) Et voir d'autre format.

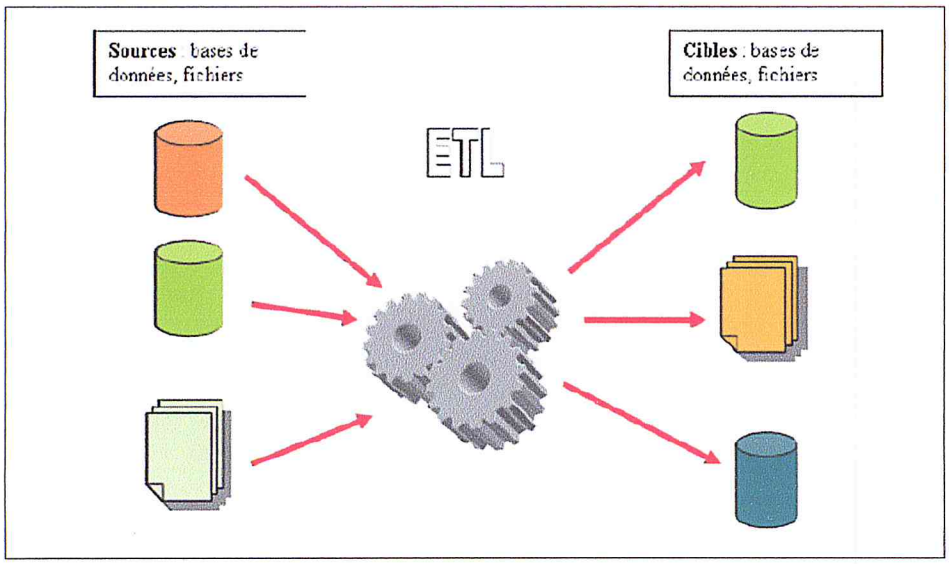

Figure 4 Processus ETL [22]

Nous présentons les différentes étapes à suivre pour l'alimentation d'un entrepôt de données.

2.2. Extract, Transform and Load (ETL):

Un ETL permet l'extraction, la transformation et le chargement de données depuis des sources diverses vers un entrepôt de données. Il permet la consolidation des données à l'aide des trois opérations suivantes : Extraction, Transformation et Chargement.

2.2.1. Tâches et étapes ETL

Pour bien concevoir un entrepôt de données qui répond àtous nos besoins, on doit d'abord connaitre les données nécessaires et leurs localisations, définir des règles d'extraction, transformation et chargement, comme le schéma de la figure 4 le montre.

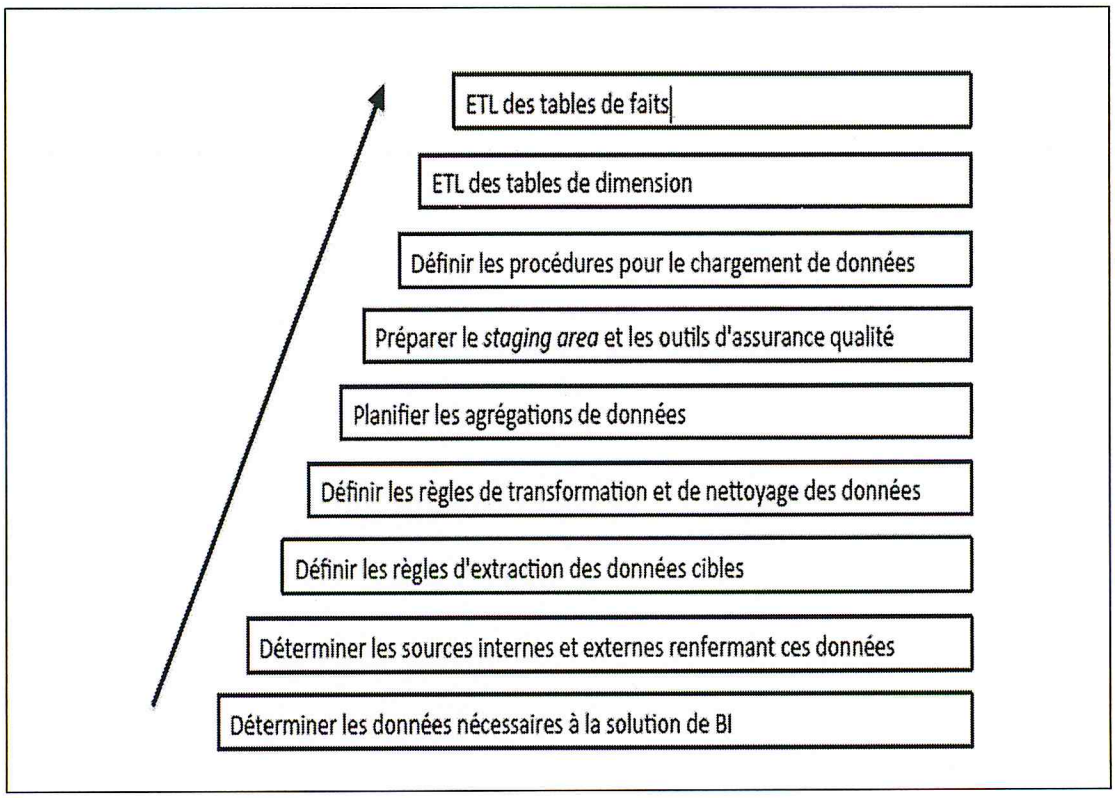

Figure 5 : Taches et étapes ETL [54]

2.2.1.1. Extractions

C'est la première opération ayant pour but d'extraire les donnes homogènes ou hétérogènes provenant de diverses sources, Pour sélectionner les données les plus utiles à notre system on doit suivre les étapes suivantes [22] :

- Enumérer les items cibles nécessaires à l'entrepôt de données : Mesures et attributs de dimension.
- Pour chaque item cible : Trouver la source et l'item correspondant de cette source.
- Si plusieurs sources sont trouvées, choisir la plus pertinente.

- Si l'item cible exige des données de plusieurs sources : Former des règles de consolidation.
- Si l'item source referme plusieurs items cibles : Définir des règles de découpage.
- Inspecter les sources pour des valeurs manquantes.

- **Type d'extraction :**

1) Extraction complète :

Capture l'ensemble des données à un certain instant. Qui est utilisé dans deux situations :

- Chargement initial des données.
- Rafraichissement complet des données.

Peut être très coûteuse en temps : Plusieurs jours.

2) Extraction Incrémentale :

Capture uniquement les données qui ont changées et aussi ajoutées depuis la dernière extraction. Qui peuvent normalement employée dans deux situations:

1. Chargement initial des données;
2. Rafraîchissement complet de données (ex: modification d'une source). [54]

3) Extraction temps-réel :

S'effectue au moment où les transactions surviennent dans les systèmes sources

a) Capture à l'aide du journal des transactions :

- Utilise les logs de transactions de la BD.
- Aucune modification requise à la BD ou aux sources.
- Doit être faite avant le rafraichissement périodique du journal.
- N'est possible qu'avec les BD journalisées : Sources à base de fichiers

b) Capture à l'aide de triggers :

- Des objets de la base de données.
- Attachés à une table.
- Vont déclencher l'exécution d'une instruction, ou d'un bloc d'instructions.

Lorsqu'une, ou plusieurs lignes sont insérées, supprimées ou modifiées dans la table à laquelle ils sont attachés.

- Evènements déclenchant le trigger.
- Une fois le trigger déclenché, ses instructions peuvent être exécutées soit juste avant l'exécution de l'événement déclencheur, soit juste après

c) Capture à l'aide des applications sources :

- Les applications sources sont modifiées pour écrire chaque ajout et modification de données dans un fichier d'extraction.
- Exige des modifications aux applications existantes.
- Entraîne des coûts additionnels de développement et de maintenance.

Peut être employée sur des systèmes à base de fichiers

4) Extraction différée :

Extrait tous les changements survenus durant une période donnée :

Exemple : Heure, Jour, Semaine, Mois.

a) Capture basée sur les timestamp :

- L'horodatage (timestamping) est un mécanisme qui consiste à associer une date et une heure à un événement, une information ou une donnée informatique.
- Il a généralement pour but d'enregistrer l'instant auquel une opération a été effectuée.
- La valeur représentant la date et l'heure est appelée timestamp.

b) Capture par comparaison de fichiers :

- Compare deux snapshots successifs des données sources.
- Extrait seulement les différences (ajouts, modifications, suppressions) entre les deux snapshots.
- Peut être employée sur des systèmes à base de fichiers, sans aucune modification.
- Exige de conserver une copie de l'état des données sources. Approche relativement coûteuse

2.2.1.2. Transformation

Cette seconde étape a pour objectif la transformation des données. Elle est bien évidemment indispensable si l'on veut obtenir des cibles différentes des sources.

C'est cette étape qui va permettre de joindre les différentes sources selon les clés précédemment spécifiées. Elle va aussi permettre de filtrer les données. Le filtrage est bien différent de l'extraction puisque l'on filtre selon des critères à définir, par exemple on va filtrer les produits dont le prix est supérieur à 1000 euro.

Une partie importante de l'étape de transformation est de pouvoir effectuer des calculs. Ils peuvent être simples comme une addition ou multiplication, mais peuvent être aussi plus complexes. Disposer d'un outil ETL proposant de nombreuses opérations par défaut

Est donc un plus. La transformation doit aussi s'occuper des différentes agrégations : effectuer les commandes SQL classiques tels que SUM (somme), COUNT (comptage) ou AVG (moyenne) [22] .

- **Types de transformations**
  - **a) Révision de format**

    Changer le type ou la langueur du champ exemple : « femme » « Homme »vs « Male » « Femelle »vs « 1 » « 2 »
  - **b)** Décodage de champs :

    Traduire les valeurs cryptes des champs
  - **c) Pré-calcul des valeurs dérivées:**

    Exemple Profit calculé à partir de ventes et coûts
  - **d) Découpage de Champs complexes :**

    Extraire les valeurs prénom, second prénom et nom famille à partir d'une seule chaîne de caractère non-complet.
  - **e) Fusion de plusieurs champs :**

    Information d'un voyage :

    Source 1 : Code et description

    Source 2 : coûts du billet du voyage
  - **f) Conversion de jeu de caractères:**

    Ex: EBCDIC (IBM) vers ASCII.
  - **g)** Conversion des unités de Mesure:

    Ex: impérial à métrique.
  - **h) Agrégations:**

    Ex: ventes par produit par semaine par région.
  - **i)** *Déduplication*

    Eliminer les redondances. Exemple: plusieurs enregistrements pour un même client.
  - **j) Conversion de Dates :**

    Convertir les dates à un seul format unique. Exemple : plusieurs formats de dates 10 jan 2015' vs '10/01/2015' vs '10/01/2015'

2.2.1.3. Chargement des Données

La dernière étape, est le chargement des données, après extractions puis transformation, dans des cibles hétérogènes.

a) Chargement Initial

Se fait lors de la première réalisation de l'entrepôt de données. Il peut prendre une longue durée.

b) Chargement Incrémental

Se fait une fois où le chargement initial est fait. Il peut être fait en temps réel ou en lot.

c) Rafraîchissement Complet

Est employé lorsque le nombre de changements rend le chargement incrémental complexe.

Conclusion

Nous avons présenté dans ce chapitre l'historique du big data, son impact sur les entreprises, ses enjeux, et le changement dû à l'arrivée du Big Data. Ce dernier est la formalisation de l'évolution des volumes, de la vitesse et de la variété des données, qui crée de la valeur ajoutée.

Jusqu'ici, il n'existe pas encore sur le marché un logiciel Big Data prêt à l'emploi que l'on puisse installer dans une entreprise. Le Big Data est avant tout une démarche stratégique, il faut penser à la stratégie pour donner de la valeur aux données. C'est grâce à cette stratégie que le Big Data fonctionnera à l'intérieur d'une entreprise.

Chapitre II

Etat de l'art sur les systèmes de recommandation

Nous présenterons dans cette partie une synthèse bibliographique sur les systèmes de recommandation.

Introduction :

Vu l'accroissement de la quantité d'information et le nombre d'utilisateurs sur internet il est devenu difficile de trouver des données que l'on souhaite. Même les outils de recherche d'information classiques ne fournissent pas toujours des résultats pertinents. Néanmoins, pendant les dix dernières années, les systèmes de recommandation se sont imposés comme moyen efficace pour réduire la complexité dans la recherche d'information. Dans ce chapitre nous allons présenter les techniques existantes pour la production des systèmes de recommandation, et des exemples de systèmes de recommandation sur le web.

1. Définition des systèmes de recommandation :

Un système de recommandation a pour objectif de fournir à un utilisateur des ressources pertinentes en fonction de ses préférences. Ce dernier voit ainsi réduit son temps de recherche mais reçoit également des suggestions de la part du système auxquelles il n'aurait pas spontanément prêtées attention. L'essor du Web et sa popularité ont notamment contribué à la mise en place de tels systèmes comme dans le domaine du e-commerce.

Les systèmes de recommandation peuvent être vus initialement comme une réponse donnée aux utilisateurs ayant des difficultés à prendre une décision dans le cadre d'utilisation d'un système de recherche d'information "classique".

« *Système capable de fournir des recommandations personnalisées ou permettant de guider l'utilisateur vers des ressources intéressantes ou utiles au sein d'un espace de données important.* »[28] .

2. Techniques des systèmes de recommandation:

Les systèmes de recommandation utilisent plusieurs techniques de filtrage d'information qui sont catégorisées en trois grandes approches :

- Filtrage à base de contenu.
- Filtrage collaboratif.
- Filtrage hybride.

Dans la suite, nous présentons ces approches de filtrage, en particulier le filtrage collaboratif qui est au centre de notre travail.

2.1 Filtrage à base de contenu :

Les systèmes de filtrage à base de contenu recommandent des documents similaires à ceux que l'utilisateur a déjà appréciés. Ceci est calculé en comparant les centres d'intérêt des utilisateurs (exprimés implicitement à travers la surveillance de son comportement ou explicitement par exemple à travers un questionnaire) avec les métadonnées ou les caractéristiques utilisées pour représenter les ressources ou les produits, sans prendre en compte les avis et les informations concernant d'autres utilisateurs.[48]

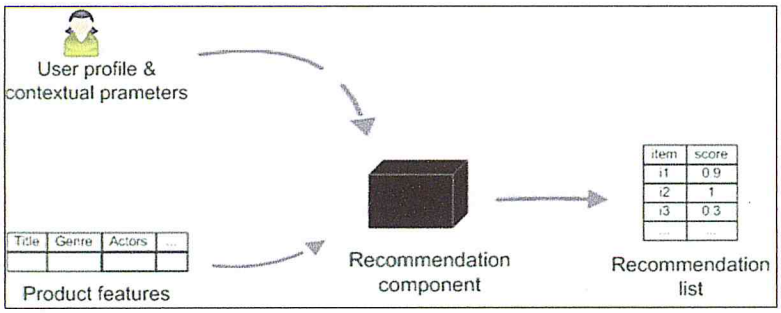

Figure 6 : Filtrage à base de contenu[43]

Deux fonctionnalités centrales ressortent de ce type de systèmes:

- La première étant la sélection de ressources correspondant au profil de l'utilisateur.
- La deuxième étant la mise à jour du profil de l'utilisateur après retour de pertinence des résultats.

Cependant, ce type de systèmes présente certaines limitations:

- L'effet « entonnoir » : La spécification des besoins de l'utilisateur empêche la diversité des sujets.
- Le filtrage des documents basé sur le contenu ne permet pas d'intégrer d'autres facteurs de pertinence (la qualité scientifique, le public visé, l'intérêt porté par l'utilisateur) que le facteur thématique.
- La difficulté d'indexation de documents multimédia cause une difficulté de recommander ce type de document.
- Problème de démarrage à froid : Un nouvel utilisateur du système éprouve des difficultés à exprimer son profil en spécifiant des thèmes qui l'intéressent.[69]

2.2 Filtrage collaboratif :

A l'opposé du filtrage basé sur le contenu, le filtrage collaboratif (Collaborative Filtering CF) prend en compte les « évaluations » que les utilisateurs ont données à certains des documents, pour recommander ces mêmes documents à d'autres utilisateurs. Ainsi, la fonction de prédiction F utilise la matrice des votes (utilisateur × item) U × I → [1,10].[69] .

	S1	S2	S3
C1		6	7
C2	5	4	
C3	8		3
C4	5		
C5		2	

Table 1 : Matrice (utilisateur, item)

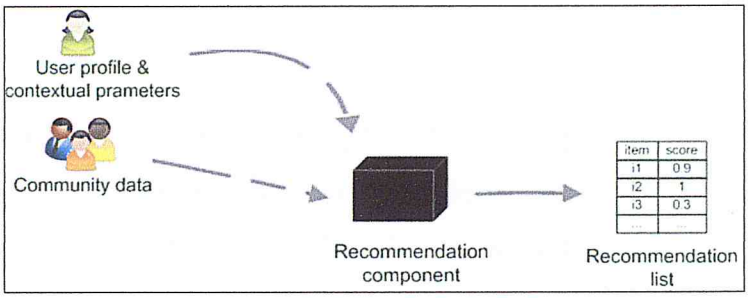

Figure 7 : Filtrage collaboratif [43]

Un système de filtrage collaboratif est organisé comme suit :

- Collecter les appréciations des utilisateurs. En général les évaluations sont exprimées sous forme de notes.
- Intégrer ces informations dans les profils utilisateurs.
- Et utiliser ceux-ci ensuite pour aider les utilisateurs dans leurs prochaines recherches.

Avantage de ce type de filtrage:

- Utilise des recommandations d'autres utilisateurs (score) pour évaluer l'utilité des items.
- L'essentiel est de trouver des utilisateurs ou groupes d'utilisateurs dont les intérêts correspondent à l'utilisateur courant.
- Plus il y a d'utilisateurs, plus il y a de scores: meilleurs sont les résultats.

Inconvénients de ce type de filtrage:

- Trouver des utilisateurs ou groupes d'utilisateurs similaires est difficile.
- **Démarrage à froid**
  - Nouvel utilisateur = Pas de préférences
  - Nouvel item = Pas de score
  - Faible densité de la matrice utilisateur/score

(a)Pondération (Weighted) : Une méthode hybride qui combine la sortie d'approches distinctes, utilisant, par exemple, une combinaison linéaire des scores de chaque technique de recommandation.

(b) Commutation (Switching) : C'est une technique qui permet de faire le choix d'un modèle de recommandation parmi plusieurs, en se basant sur plusieurs critères. La détermination de la technique appropriée dépend de la situation. Lesystème se doit alors de définir les critères de commutation, ou les cas où l'utilisation d'une autre technique est recommandée. Ceci permet au système de connaître les points forts et les points faibles des techniques de recommandation qui le constituent.

(c) Technique mixte (Mixed) : Dans cette approche, le recommandeur ne combine pas, mais augmente la description des ensembles de données, en prenant en considération les estimations des utilisateurs et la description des articles. La nouvelle fonction de prédiction doit faire face aux deux types de descriptions et permet d'éviter les problèmes posés par le filtrage collaboratif, à savoir, le démarrage à froid.

(d) Combinaison de caractéristiques (Featurescombination): Dans un hybride basé sur la combinaison de caractéristiques, les données provenant de techniques collaboratives sont traitées comme une caractéristique, et une approche basée sur le contenu est utilisée sur ces données.

(e) Cascade : La cascade implique un processus étape par étape. Dans ce cas, une technique de recommandation est appliquée en premier, produisant un ensemble de candidats potentiels. Puis, une deuxième technique raffine les résultats obtenus dans la première étape. Cette méthode a pour avantage que si la première technique génère peu de recommandations, ou si ces recommandations sont ordonnées afin de permettre une sélection rapide, la deuxième technique ne sera plus utilisée.

(f) Augmentation de caractéristiques (Features augmentation): L'augmentation de caractéristiques est semblable à la cascade, mais dans ce cas-là les résultats obtenus (le classement ou la classification) de la première technique sont utilisées par le deuxième comme une caractéristique ajoutée.

(g) Méta niveau (Meta-level) : Dans un hybride basé sur méta niveau, une première technique est utilisée, mais différemment que la précédente méthode (augmentation de caractéristiques), non pas pour produire de nouvelles caractéristiques, mais pour produire un modèle. Et dans la deuxième étape, c'est le modèle entier qui servira d'entrée pour la deuxième technique.

Méthode	Description
Pondération (a)	Les résultats (ou votes) des différents techniques de recommandation sont combinés pour produire une seule recommandation
Commutation (b)	Le système commute entre les techniques de recommandation selon la situation actuelle
Technique mixte (c)	Les recommandations de différents recommandeurs sont présentées en même temps
Combinaison des caractéristiques (d)	Les données provenant d'une technique sont traitées comme une caractéristique, et une approche basée sur une autre technique est utilisée sur ces données
Cascade (e)	Un recommandeur raffine les recommandations données par un autre
Augmentation des caractéristiques (f)	La sortie d'une technique est utilisée comme une caractéristique d'entré à l'autre
Méta niveau (g)	Le modèle appris d'un recommandeur est employé comme entrée à l'autre

Table 2 : Les méthodes hybrides[44]

3. Exemples des systèmes de recommandation existant:

Tapestry:

Goldberg et al.[16] représente l'un des premiers systèmes de recommandation. Il a été développé en 1992 par le centre de recherche de "Xerox" aux Etats Unis. Il s'agit d'un système de recommandation intégré à une application de mail électronique, permettant de recommander des listes de diffusion aux utilisateurs. Tapestry est fondé sur le Filtrage Collaboratif (FC) exploitant les annotations (les tags) des utilisateurs attribués aux listes de diffusion. L'analyse de ces annotations par le système de FC permet de déterminer et de proposer les listes de diffusion qui sont pertinentes pour chaque utilisateur.

Tapestry a aussi introduit la prise en compte de la confiance dans la source de l'information. Le système a souffert de deux problèmes [18] . Le premier est la taille de sa base d'utilisateurs. Puisque Tapestry est basée sur un système commercial de base de données, il ne peut être fourni librement. De plus, il n'a pas été conçu pour l'usage d'un grand nombre de personnes géographiquement distribuées. Ces deux facteurs se combinent pour limiter la population d'utilisateurs potentiels aux chercheurs à Xerox Parc. Cependant, cette population ne semblait pas assez grande pour constituer une masse critique d'utilisateurs et la grande majorité des documents passaient sans annotations (un ensemble de mots décrivant le contenu d'un document). Ainsi le système souffrait d'un manque d'informations pour pouvoir fonctionner normalement.

Le deuxième problème avec Tapestry est le moyen par lequel les utilisateurs interagissent avec les filtres. Une interface commune exigeait des utilisateurs d'indiquer des requêtes en un

langage dérivé de SQL. Cette forme d'interface a été un obstacle à l'exploration de nouveaux secteurs et a rendu difficile la visualisation de l'information disponible.

Il n'en demeure pas moins que Tapestry fut un des premiers systèmes de filtrage existants.[13]

GroupLens:

GroupLens est un système expérimental de l'université du Minnesota[46] . Il fait des recommandations de messages de groupes de discussion. Group-Lens a travaillé sur un principe similaire à Tapestry: un groupe d'utilisateurs annotent des documents, de manière simplifiée: d'abord, des annotations ont été réduites à une note d'intérêt pour chaque document.

Deuxièmement, le langage de requête : la principale difficulté de Tapestry, a été remplacée par une fonction de prédiction automatique des scores, calculée à partir de la corrélation des évaluations des utilisateurs.

GroupLens a été évalué pour sept semaines avec 250 utilisateurs, avec seulement un rapport qualitatif [19] .

GroupLens a validé le principe de filtrage collaboratif à grande échelle, mais aussi a confirmé ses faiblesses comme le problème de démarrage à froid [19] e système n'était pas capable de faire de bonnes prédictions avant une certaine masse critique d'utilisateurs. Cela pourrait dissuader les premiers utilisateurs à s'impliquer dans le système.

GroupLens était le premier article introduisant le terme "filtrage collaboratif". Le système de GroupLens a été basé sur une méthode de KPPV(K plus proches voisins) sur les utilisateurs. Le coefficient de Pearson a été utilisé comme mesure de la similitude entre les utilisateurs.

GroupLens est souvent considéré comme le fondateur de l'approche de filtrage collaboratif.[13]

Fab:

Fab[41] est un système de recommandation hybride du contenu Web qui essaye de combiner les deux approches : l'approche basée sur le contenu sémantique et l'approche collaborative du filtrage pour en récupérer les avantages et en réduire les inconvénients. La notion de profil basée sur l'analyse du contenu y est maintenue et les profils y sont systématiquement comparés pour identifier les similarités entre utilisateurs. Un utilisateur reçoit un document

soit parce qu'il correspond à son profil soit parce qu'il a été apprécié par un autre utilisateur ayant un profil ressemblant.

Le processus de recommandation peut être reparti en deux phases : une phase de collecte de ressources pour constituer une base ou un index et une phase de sélection de ressources de cette base pour des utilisateurs particuliers. La phase de collecte peut être triviale dans le cas général mais pose un vrai problème dans le cas du web, pour le concepteur du système. Dans Fab, cette phase consiste à rassembler des pages pertinentes pour un nombre réduit de sujets, et qui sont regroupées automatiquement suivant les domaines d'intérêt des utilisateurs. Ces pages sont ensuite diffusées à un large nombre d'utilisateurs dans la phase de sélection. Un sujet peut intéresser plusieurs personnes et une personne peut être intéressée par plusieurs sujets. Pour l'implémentation, des agents sont utilisés. Les pages retrouvées par l'agent de collecte sont envoyées à un routeur central qui se charge de les transférer aux utilisateurs dont les profils correspondent, à partir d'un certain seuil. D'autres fonctionnalités sont assurées par les agents personnels de chaque utilisateur. Les pages déjà consultées sont éliminées, et sur les pages présentées, ils assurent qu'il y a au plus une page d'un même site. Une fois que l'utilisateur a envoyé une requête, reçu et consulté des recommandations, il lui est demandé de fournir une note de 0 à 7. Ces notes servent d'une part à mettre à jour les profils personnels et à informer l'agent de collection. De plus, toute page très bien notée est automatiquement passée aux utilisateurs estimés les plus proches.

Fab est un exemple-type de la combinaison des approches basées sur le contenu sémantique et collaborative dans les systèmes de recommandation.[13]

Amazon:

Amazon est un exemple typique du succès de la technologie des moteurs de recommandation. Cette technologie est à la base de la stratégie de marketing du site. La fonction principale utilisée est basée sur recommandations contextuelles collaborative item-item. Cela a été introduit très tôt sur le site (fin des années 90). Il est basé sur les journaux d'achats et correspond au calcul d'une matrice de similarité des items avec un algorithme optimisé à l'échelle avec les volumes traités par Amazon [26] . Amazon a popularisé la célèbre fonction de recommandation que nous qualifions d'item-to-items, le fameux : " les personnes qui ont vu/acheté cet article ont aussi vu/acheté ces articles " [23]

Un exemple de l'interface fournie par Amazon pour la recommandation contextuelle item-to-item est donnée dans cette figure :

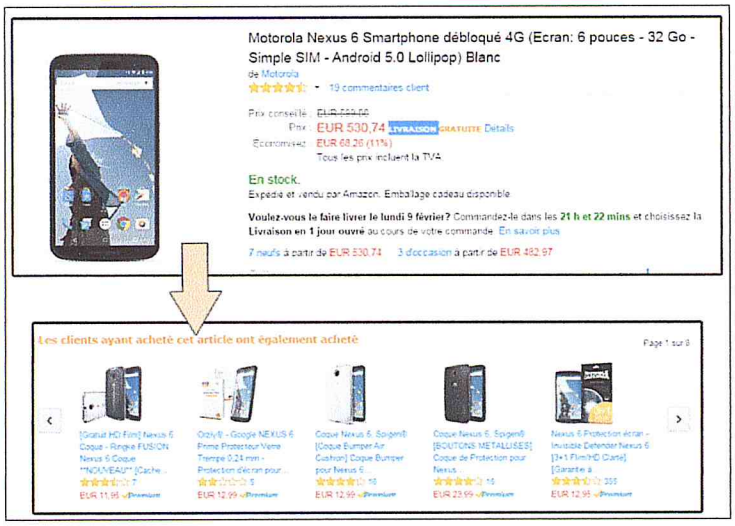

Figure 9 : Un exemple de l'interface fournie par Amazon

Amazon peut désormais être considéré comme un système hybride comme les fonctionnalités de recommandation sont basées sur des métadonnées (genre, auteur).[23]

Kayak [Kayak.com]:

Kayak.com est un moteur de recherche de voyage américain créé en 2004. Il permet de trouver et de réserver des billets d'avion, des chambres d'hôtel, des séjours et des voitures à bas prix grâce à un site web dont le moteur de recherche permet de comparer les prix et de recommander le meilleur aux clients.

Il existe plusieurs autres systèmes de recommandation online dans le domaine des voyages tel que Booking.com, Tripadvisor.com, Travelocity.com et Orbitz.com.

Conclusion

Dans ce chapitre nous avons présenté les différentes techniques utilisées dans la production des systèmes de recommandation. Nous avons ainsi évoqué deux types d'approches ainsi que la combinaison de ces deux approches. Chaque type de système possède des avantages et inconvénients. Les méthodes collaboratives donnent de meilleurs résultats, Par rapport aux méthodes basées sur le contenu. Néanmoins, leur principal problème réside dans le fait qu'il faut une base d'utilisateurs ayant déjà fait des choix pour faire des recommandations. Néanmoins, le contenu à proposer n'a pas besoin d'être défini de manière complexe. Les méthodes basées sur le contenu permettent de faire des propositions dès l'initialisation du système. En contrepartie, il faut bien définir le contenu à proposer pour avoir de bonnes propositions.

Dans la prochaine section nous définissons les trois approches d'analyse des données qui sont la descriptive, la prédictive et la perspective. L'approche prescriptive réside au sommet du modèle de maturité analytique établi par Gartner. Ce modèle va de l'analytique descriptif à l'analytique prescriptif en passant par l'analytique prédictif. Les trois approches sont complémentaires car pour prédire le futur, l'approche prédictive doit connaitre le passé et réciproque pour prendre des décisions efficaces (perspective), il faut aussi passer par les deux dernières méthodes (prédictive et prescriptive). Bien que l'analyse prescriptive ouvre des perspectives exceptionnelles, elle peut rapidement devenir fastidieuse et complexe avec la croissance exponentielle des données (Big Data). C'est en partie pour ces raisons qu'elle reste largement inexploitée. Selon Gartner, 3 % seulement des entreprises font appel à un logiciel d'analyse prescriptive, contre 30 % qui utilisent activement des outils d'analyse prédictive qui est de plus en plus répandue.

L'analyse prédictive reste malgré tout une technique parfois difficile à appréhender mais efficace. Nous choisissons donc l'approche prédictive par ce que notre objectif c'est de prédire des recommandations susceptibles d'intéresser le voyageur aérien.

Chapitre III

Etat de l'art sur l'analyse prédictive

Nous présenterons dans cette partie une synthèse bibliographique sur l'analyse prédictive.

Introduction

Depuis l'utilisation de statistiques dans le sport de haut niveau jusqu'aux algorithmes de recommandation d'Amazon, en passant par le programme de surveillance PRISM de la NSA « Agence nationale de la sécurité »et la médecine analytique, le Big Data et les Analytics (Analyse de données) se sont construit une place de premier plan dans tous les domaines de la société.

Le terme Analytics recouvre l'utilisation intensive de Data, de techniques d'analyses statistiques et quantitatives, de modèles explicatifs et prédictifs pour influencer la prise de décision, transformer les méthodes de management et revoir l'approche de la création de valeur en entreprise.[63]

La montée en puissance de l'analyse quantitative dans le processus de prise de décision est une tendance de fond dans toutes les industries[63] :

- **Dans le sport de haut niveau**, l'utilisation de statistiques et de données de performance quantitatives, aura bientôt définitivement remplacé le recrutement « au feeling » des scouts et des chercheurs de talents.
- **Chez Netflix et Amazon**, les ***systèmes de recommandation*** analysent des volumes gigantesques de données transactionnelles et comportementales pour personnaliser les suggestions à l'extrême et créer des segments de marché individuels.
- **Dans le domaine médical**, un nombre croissant de professionnels commence à accepter que les outils analytiques peuvent fournir une aide précieuse dans l'aide au diagnostic (il existe 11000 maladies répertoriées) et la personnalisation des traitements

La typologie la plus fréquemment utilisée consiste à distinguer trois familles d'Analytics, en fonction des méthodes d'analyse employées et des objectifs recherchés :

- **Analyse descriptive :** *Un aperçu du passé*

Comme son nom l'indique, l'analyse descriptive facilite l'organisation, le classement des données, la visualisation (leur représentation graphique) des données ; mais aussi la synthèse des données. Dans cette description, l'extraction de données du logiciel située en amont est essentielle. On va « piocher» des données qui jouent un rôle important et contribuent à la mise en forme des résultats.[60]

Le but de l'analytique descriptif est de résumer les informations émanant des données récoltées de divers objets et équipements communicants. Par exemple, à l'aide d'un boîtier

connecté au véhicule, on est capable de remonter le nombre de kilomètres parcourus par celui-ci sur une distance donnée.[57]

Les technologies les plus utilisées dans l'analyse :

- La modélisation des données
- La Régression
- Visualisation de données statistiques

- **Analyse prédictive :** *Comprendre l'avenir*

L'analyse prédictive permet de mieux identifier les caractéristiques fondamentales des clients afin de les modéliser et d'anticiper au mieux les comportements.

L'analyse prédictive n'est pas un outil en soi, c'est plutôt une pratique qui s'appuie sur les outils statistiques bien sûr, mais aussi le data mining et la recherche de corrélation et la théorie des jeux.[5]

« L'analyse prédictive est définie par l'éditeur SPSS comme l'analyse des données historiques et actuelles disponibles sur le client afin de créer des prévisions sur ses comportements, préférences et besoins futurs ».[11]

Les technologies les plus utilisées dans l'analyse prédictive :

- Data-mining
- La modélisation prédictive

- **Analyse prescriptive :** *Prendre la bonne voie*

L'émergence d'un analytique prescriptif va bien au-delà de ces deux précédents concepts puisqu'il consiste en la prescription d'une action, d'une décision. Le modèle prescriptif est en quelque sorte capable de prédire les possibles conséquences des actions qui sont choisies d'être faites et/ou recommander la meilleure action à adopter.[57]

Les technologies les plus utilisées dans l'analyse prescriptive:

- Optimisation
- Simulation

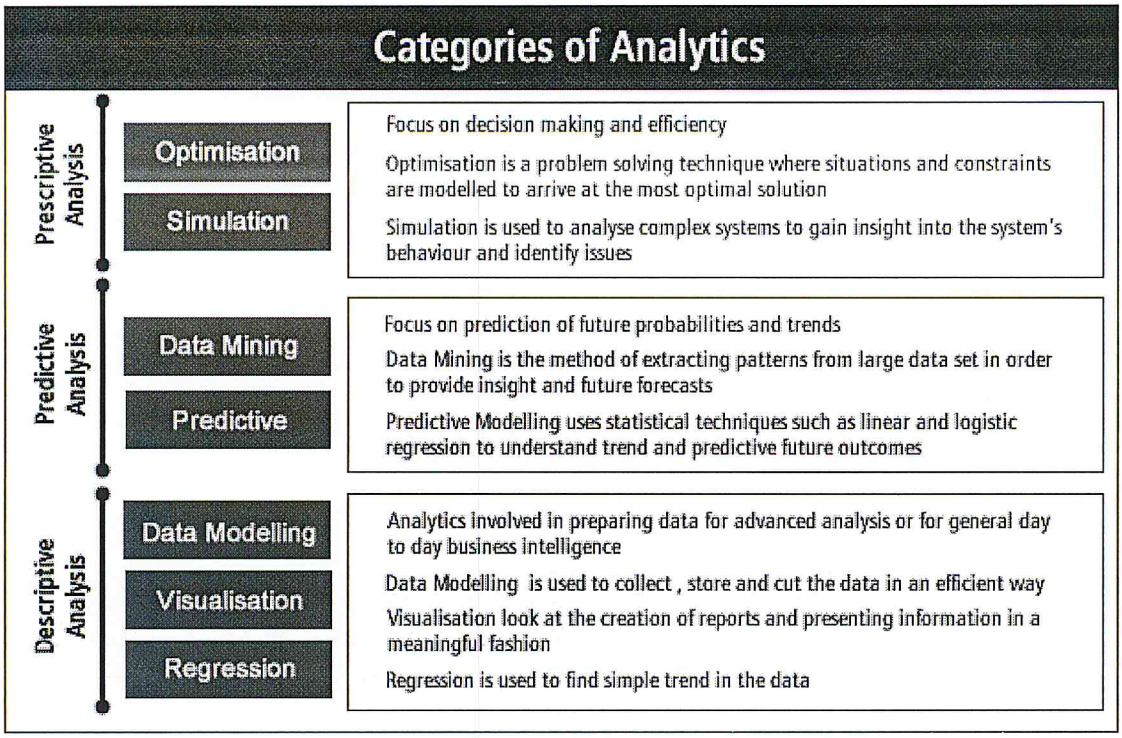

Figure 10 : Les trois familles d'Analytics[27]

1. Les technologies les plus utilisées dans l'analyse prédictive:

2.1 Data-Mining :

Le terme Data-mining est souvent employé pour désigner un ensemble d'outils permettant aux utilisateurs d'accéder aux données de l'entreprise et des analyses. Les outils d'aide à la décision, qu'ils soient relationnels ou OLAP, laissent l'initiative à l'utilisateur de choisir les éléments qu'il veut observer ou analyser. Au contraire, dans le cas du data mining, le système a l'initiative et découvre lui-même les associations entre les données, sans que l'utilisateur ait à lui dire de rechercher plutôt dans telle ou telle direction ou à poser des hypothèses. Les modèles classiques de recherche d'informations ne sont pas adaptés pour traiter des masses gigantesques de données, souvent hétérogènes. C'est ce constat qui a permis au data mining d'émerger et vulgariser les méthodes d'analyse.

Le Data-mining (ou la fouille de données) a pour objet l'extraction d'un savoir à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. La fouille de données repose sur un ensemble de fonctions mais aussi sur une méthodologie de travail.

Le terme de Data-mining signifie littéralement exploitation des données. Comme dans toute exploitation, le but recherché est de pouvoir extraire de la richesse. Ici, la richesse est la connaissance de l'entreprise. Fort du constat qu'il existe au sein des bases de données de

chaque entreprise une ressource de données cachées et surtout inexploitée, le Data Mining permet de faire les apparaître, et cela grâce à un certain nombre de techniques spécifiques. Nous appellerons Data mining l'ensemble des techniques qui permettent de transformer les données en connaissances. Le périmètre d'exploitation du Data mining ne se limite pas à l'exploitation des Data warehouse. Il veut d'être capable d'exploiter toutes bases de données contenant de grandes quantités de données telles que des bases relationnelles, des entrepôts de données mais également des sources plus ou moins structurées comme internet. Dans ces cas, il faut néanmoins construire une base de données ou un entrepôt de données qui sera dédié à l'analyse.[56]

On distingue généralement deux méthodes dans l'étape d'analyse dans un problème Data-Mining : **prédictive** et **descriptive**, Les méthodes prédictives utilisent un ensemble de variables observées pour prédire les valeurs futures ou inconnues d'autres variables, les méthodes de prédiction comprennent la classification, la régression et la détection de déviation. Les méthodes descriptives se concentrent sur la recherche de motifs significatifs qui aident à comprendre et à interpréter les données. Il s'agit notamment du regroupement, la découverte de règles d'association et la découverte de motifs. Dans notre cas on va étudier la technique prédictive afin de l'utilisé dans le contexte d'un système de recommandation.[64]

Les méthodes de Data-Mining

a) Méthodes descriptive :

Le principe de ces méthodes et de pouvoir mettre en évidence des informations présentes dans le data warehouse mais cachées par le volume des données et il n'y a pas de variable cible «prédire ».[40]

Parmi les techniques et algorithmes utilisés dans l'analyse descriptive, on cite[56] :

- Analyse factorielle (ACP, AFC et ACM)
- Méthode des centres mobiles
- Classification hiérarchique
- Classification neuronale (réseau de Kohonen)
- Recherche d'association

b) Méthodes prédictive :

Contrairement à l'analyse descriptive, cette technique fait appels à de l'intelligence artificielle. L'analyse prédictive, est comme son nom l'indique une technique qui va essayer de prévoir une évolution des événements en se basant sur l'exploitation de données stockés dans le Data warehouse. [56]

L'analyse se fait sur l'ensemble des variables de la base de données utilisant une classification a apprentissage supervisée (la classe est connue) basée sur l'apprentissage automatique. Ces techniques visent à extrapoler de nouvelles informations à partir des informations cachées (c'est le cas des scoring), dans ce cas il y a une variable cible à prédire.[40]

En effet, l'observation et l'historisation des événements peuvent permettre de prédire une suite logique. Le meilleur exemple est celui des prévisions météorologiques qui se base sur des études des évolutions météorologiques passées. En marketing, l'objectif est par exemple de déterminer les profils d'individus présentant une probabilité importante d'achat ou encore de prévoir à partir de quel moment un client deviendra infidèle.[56]

Parmi les techniques et algorithmes utilisés dans l'analyse prédictive, on cite [56] :

- Arbre de décision
- Réseaux de neurones
- Régression linéaire
- Analyse discriminante de Fisher
- Analyse probabiliste
- La classification supervisée
- Machines à vecteurs de support

Les taches du data mining :

Le Data Mining est un domaine multidisciplinaire, il regroupe plusieurs tâches:

1. Description
2. Classement (classification dans l'école anglo-saxonne)
3. Association
4. Estimation
5. Segmentation
6. Prévision

Ces types d'analyse se répartissent dans les techniques descriptives et prédictives [6] :

Techniques descriptives		**Techniques prédictives**		
Corrélation simple	**Corrélation complexe**	**Présent**		**Futur**
		Variable cible numérique	**Variable cible catégorielle**	
Description	Segmentation Association	Estimation	Classification	Prévision

Table 3 : Les tâches du data mining [12]

c) La description (technique descriptive) :

C'est souvent l'une des premières tâches demandées à un outil de Data Mining. On lui demande de décrire les données d'une base complexe. Cela engendre souvent une exploitation supplémentaire en vue de fournir des explications.[56]

La technique la plus appropriée à la description est : L'analyse du panier de la ménagère (Ou règles d'association. Cette tâche peut être effectuée pour identifier des opportunités de vente croisée et concevoir des groupements attractifs de produit).

d) Classification (technique prédictive) :

La classification se fait naturellement depuis déjà bien longtemps pour comprendre et communiquer notre vision du monde (par exemple les espèces animales, minérales ou végétales). Elle consiste à examiner des caractéristiques d'un élément nouvellement présenté afin de l'affecter à une classe d'un ensemble prédéfini. Dans le cadre informatique, les éléments sont représentés par un enregistrement et le résultat de la classification viendra alimenter un champ supplémentaire.

La classification permet de créer des classes d'individus (terme à prendre dans son acception statistique). Celles-ci sont discrètes : homme / femme, oui / non, rouge / vert / bleu, ... Les techniques les plus appropriées à la classification sont : les arbres de décision, Méthode des k plus proches voisins, Réseau de neurones, le raisonnement basé sur la mémoire, éventuellement l'analyse des liens.[56]

e) Association (technique descriptive) :

Consiste à déterminer les valeurs qui sont associées, c.à.d. les objets qui vont ensemble.

eg : la principale application est la recherche d'associations pour trouver les articles achetés ensemble.

Les techniques les plus appropriées sont : Les règles d'association, l'algorithme Apriori.

f) Estimation (technique prédictive) :

Contrairement à la classification, le résultat d'une estimation permet d'obtenir une variable continue. Celle-ci est obtenue par une ou plusieurs fonctions combinant les données en entrée. Le résultat d'une estimation permet de procéder aux classifications grâce à un barème. Par exemple, on peut estimer le revenu d'un ménage selon divers critères (type de véhicule et nombre, profession ou catégorie socioprofessionnelle, type d'habitation, etc.).

Il sera ensuite possible de définir des tranches de revenus pour classifier les individus.

Un des intérêts de l'estimation est de pouvoir ordonner les résultats pour ne retenir si on le désire que les n meilleures valeurs. Cette technique sera souvent utilisée en marketing, combinée à d'autres, pour proposer des offres aux meilleurs clients potentiels. Enfin, il est facile de mesurer la position d'un élément dans sa classe si celui-ci a été estimé, ce qui peut être particulièrement important pour les cas limitrophes.[56]

La technique la plus appropriée à l'estimation est : le réseau de neurones.

g) Segmentation (technique descriptive) :

Consiste à segmenter une population hétérogène en sous populations homogènes. Contrairement à la classification, les sous populations ne sont pas préétablis. La technique la plus appropriée à la clusterisation est l'analyse des clusters.

Les techniques les plus appropriées à la segmentation sont : la classification hiérarchique et la classification des K moyennes[56]

h) La prévision (technique prédictive) :

La prévision est similaire à l'estimation et à la segmentation mise à part que pour la prévision, elle s'appuie sur le passé et le présent mais son résultat se situe dans un futur généralement précisé.[12]

Les techniques les plus appropriées à la prédiction sont :

- L'analyse du panier de la ménagère
- Le raisonnement basé sur la mémoire
- Les arbres de décision
- les réseaux de neurones

2.2 La modélisation prédictive :

La modélisation prédictive est un processus par lequel on cherche à identifier le meilleur modèle qui va permettre d'estimer la probabilité de survenance d'un événement ou d'un comportement. En ce sens, elle est un outil d'aide à la décision[20] .La modélisation prédictive utilise une variété de techniques issues des statistiques telles que la régression linéaire et logistique pour prédire les résultats futurs[27] .

2. Méthodes utilisées dans l'analyse prédictive:

a) Les méthodes de régression :

- **La régression linéaire :**

La régression linéaire est l'une des techniques statistiques les plus utiles et l'une de celles qu'on emploie de plus en plus couramment dans le cas d'une variable dépendante continue. De plus, parce qu'on peut l'étendre au-delà des données bi-variées en l'appliquant à une situation multi-variée, la régression se révèle un outil très utile de la recherche sociale. L'analyse de régression multiple permet de construire une équation explicative d'un phénomène donné. On identifie alors les variables indépendantes les plus significatives, ce qui permet de «prédire» les comportements non mesurés directement[29]

- **La régression logistique :**

La régression logistique permet d'estimer la force de l'association entre une variable qualitative dichotomique (binaire) dépendante et des variables qualitatives ou quantitatives indépendantes. La régression logistique peut être uni-variée mais son intérêt réside dans son utilisation multi-variée. La régression logistique est un outil qui permet de mettre en relation des variables explicatives à une variable réponse dichotomique, c'est-à-dire qui ne peut prendre que deux valeurs, le cas classique étant celui d'une variable réponse (dépendante) binaire. Cette situation est fréquente dans divers champs d'application, particulièrement dans les sciences sociales[29]

Cette technique est utilisée pour des études ayant pour but de vérifier si des variables indépendantes peuvent prédire une variable dépendante dichotomique.

b) La méthode des k-plus proches voisins :

C'est une approche très simple et directe. Elle ne nécessite pas d'apprentissage mais simplement le stockage des données d'apprentissage. Son principe est le suivant. Une donnée de classe inconnue est comparée à toutes les données stockées. On choisit pour la nouvelle

donnée la classe majoritaire parmi ses K plus proches voisins (Elle peut donc être lourde pour des grandes bases de données) au sens d'une distance choisie. Afin de trouver les K plus proches d'une donnée à classer, on peut choisir la distance euclidienne[21] . Soient deux données représentées par deux vecteurs x_i et x_j, la distance entre ces deux données est donnée par:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{ik})^2}$$

- **La méthode des k-plus proches voisins dans les Système de recommandation :**

La méthode de k-plus proches voisins est l'une des approches les plus courantes pour le filtrage collaboratif (et donc pour concevoir un système de recommandation). En réalité, tout aperçu sur les systèmes de recommandations (comme celui d'Adomavicius et Tuzhilin[25]) va comporter une introduction à l'utilisation de la méthode des k-plus proches voisins dans ce contexte.

L'un des avantages de ce classificateur c'est qu'il est théoriquement très lié à la notion de filtrage collaboratif, Trouver les utilisateurs partageant les mêmes goûts (ou des produits similaires) est équivalent à trouver des voisins pour un utilisateur ou un produits (item) donné.[64]

Frank Meyer [23] , présente Reperio, un moteur hybride utilisant une technique de K-Plus-Proches Voisins (KPPV). Il a étudié les performances des méthodes KPPV, notamment l'impact des fonctions de similarités utilisées. Et les performances de Reperio dans le cas critique du démarrage à froid. Frank Meyer à montrer que les méthodes de type K-plus-proches voisins sont très compétitives à condition de bien spécifier les mesures de similarité utilisées pour définir les voisinages.

c) Arbre de décision (DecisionTrees)

Ensemble de règles de classification basant leur décision sur des tests associes aux attributs (ou classes), organisés de manière arborescente. Les éléments à classer sont composés d'attributs et leur valeur cible, Les nœuds de l'arbre peuvent être :

a) des nœuds de décision, dans ces nœuds une seule valeur d'attribut est testée pour déterminer à quelle branche de la sous-arborescence s'applique.
b) Ou nœuds feuilles qui indiquent la valeur de l'attribut cible.

Les Branches de l'arbre correspondent à une valeur d'attribut.

« Un arbre de décision est un outil pour déterminer l'appartenance d'un objet à une classe en fonction de ses caractéristiques/attributs. »[7]

On donne un ensemble X de N dont les éléments sont notés xi et dont les P attributs sont quantitatifs. Chaque élément de X est étiqueté, c'est-à-dire qu'il lui est associé une classe ou un attribut cible que l'on note y appartenant à Y.

A partir de ce qui précède, on construit un arbre dit « de décision » tel que :

- chaque noeud correspond à un test sur la valeur d'un ou plusieurs attributs.
- chaque branche partant d'un noeud correspond à une ou plusieurs valeurs de ce test.

Les arbres de décisions ont pour objectif la classification et la prédiction. Leur fonctionnement est basé sur un enchaînement hiérarchique de règles exprimées en langage courant.

Un arbre de décision est une structure qui permet de déduire un résultat à partir de décisions successives. Pour parcourir un arbre de décision et trouver une solution il faut partir de la racine. Chaque noeud est une décision atomique. Chaque réponse possible est prise en compte et permet de se diriger vers un des fils du noeud. De proche en proche, on descend dans l'arbre jusqu'à tomber sur une feuille. La feuille représente la réponse qu'apporte l'arbre au cas que l'on vient de tester.

- Débuter à la racine de l'arbre
- Descendre dans l'arbre en passant par les nœuds de test
- La feuille atteinte à la fin permet de classer l'instance testée.

Très souvent on considère qu'un nœud pose une question sur une variable, la valeur de cette variable permet de savoir sur quels fils descendre. Pour les variables énumérées il est parfois possible d'avoir un fils par valeur, on peut aussi décider que plusieurs variables différentes mènent au même sous arbre. Pour les variables continues il n'est pas imaginable de créer un nœud qui aurait potentiellement un nombre de fils infini, on doit discrétiser le domaine continu (arrondis, approximation), donc décider de segmenter le domaine en sous-ensembles. Plus l'arbre est simple, et plus il semble techniquement rapide à utiliser. En fait, il est plus intéressant d'obtenir un arbre qui est adapté aux probabilités des variables à tester. La plupart du temps un arbre équilibré sera un bon résultat. Si un sous arbre ne peut mener qu'à une solution unique, alors tout ce sous-arbre peut être réduit à sa simple conclusion, cela simplifie le traitement et ne change rien au résultat final.[56]

Il existe de nombreux algorithmes pour les arbres de décision tel que[64] :

- Algorithme de Hunt (1966)
- CART (Classification And Regression Trees) (1984)
- ID3 (1986), C4.5 (1993)
- SLIQ, SPRINT

Nous décrivons brièvement **l'algorithme Hunt [7]** :

Soit $\mathbf{D_t}$ l'ensemble des données qui a été associé au nœud **t**

- Si $\mathbf{D_t}$ contient que des nœuds appartenant à la classe $\mathbf{y_t}$, alors le nœud **t** est une feuille étiquetée $\mathbf{y_t}$
- Si $\mathbf{D_t} = \emptyset$, alors **t** est une feuille étiquetée par la classe de défaut $\mathbf{y_d}$
- Si $\mathbf{D_t}$ contient des données appartenant à plus d'une classe alors
  a) utilisez un attribut diviseur pour créer des nœuds fils à **t**. Ces nœuds contiendront les données de **t** en fonction de la valeur de l'attribut choisi.
  b) appliquer les étapes précédentes aux nœuds créés

Un exemple simple d'arbre de décision :

Sports fan	Marital Status	Annualincome	Likes Pizza
Yes	Divorced	90K	Yes
No	Single	125K	No
Yes	Married	100K	No
Yes	Married	60K	No
Yes	Married	75K	No
Yes	Single	105K	No
Yes	Single	85K	Yes
Yes	Single	90K	Yes
No	Divorced	220K	No
No	Married	120K	No

Table 4 : Attributs et attribut cible à partir d'observations[64]

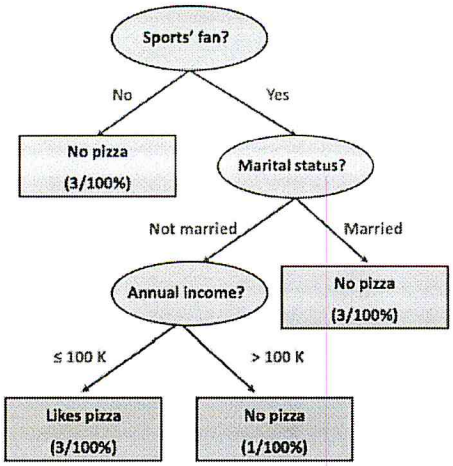

Figure 11: Exemple d'un arbre de décision pour les données résumées dans la Table 4[64]

- **Les arbres de décisions dans les Système de recommandation :**

Les arbres de décision peuvent être utilisés dans une approche basée sur un modèle pour un système de recommandation. L'une des possibilités est d'utiliser les caractéristiques de contenu pour construire un arbre de décision qui modélise toutes les variables impliquées dans les préférences de l'utilisateur.

Bouza et al. [1] utilisent l'idée de construire un arbre de décision à l'aide d'informations sémantiques disponibles pour les items. L'arbre est construit après que l'utilisateur a évalué seulement deux items. Les caractéristiques de chaque item sont utilisées pour construire un modèle qui explique les évaluations d'utilisateurs. Ils utilisent le gain d'information de toutes les caractéristiques comme critères de séparation. Il faut noter que même si cette approche est intéressante d'un point de vue théorique, la précision qu'ils rapportent sur leur système est pire que celle de recommander la note moyenne.

Comme on pouvait s'attendre, il est très difficile et peu pratique de construire un arbre de décision qui tente d'expliquer toutes les variables impliquées dans le processus de prise de décision. Cependant, Les arbres de décision peuvent également être utilisés pour modéliser une partie particulière du système. Cho et al [65] , par exemple, présentent un système de recommandation pour les achats en ligne qui combine l'utilisation de règles d'association et les arbres de décision. L'arbre de décision est utilisé comme un filtre pour sélectionner les utilisateurs qui doivent être ciblées avec des recommandations. Afin de construire le modèle, ils créent un utilisateur candidat mis en sélectionnant les utilisateurs qui ont choisi des produits d'une catégorie donnée au cours d'une période donnée. Dans leur cas, pour la construction de l'arbre de décision, la variable dépendante est choisie comme si le client est susceptible d'acheter de nouveaux produits dans la même catégorie[64]

d) Classificateur bayésien naïf (Réseaux bayésiens)

Le classificateur bayésien naïf est un outil largement utilisé dans les problèmes de classification supervisée. Il a pour avantage de se montrer efficace pour de nombreux jeux de données réels [17] . Cependant, l'hypothèse naïve d'indépendance des variables peut, dans certains cas, dégrader les performances du classificateur[53]

Comme son nom l'indique, ce classificateur se base sur le théorème de Bayes permettant de calculer les probabilités conditionnelles. Dans un contexte général, ce théorème fournit une façon de calculer la probabilité conditionnelle d'une cause sachant la présence d'un effet, à partir de la probabilité conditionnelle de l'effet sachant la présence de la cause ainsi que des probabilités a priori de la cause et de l'effet.[67]

- **Les Classificateurs bayésien dans les Système de recommandation :**

Classificateurs bayésiens sont particulièrement populaires pour les systèmes de recommandation à base de modèles. Ils sont souvent utilisés pour obtenir un modèle pour les systèmes de recommandation à base de contenu. Cependant, ils ont également été utilisés dans le cadre de filtrage collaboratif.

Ghani et Fano [51] , par exemple, utilisent un classificateur bayésien naïf pour mettre en œuvre un système de recommandation basé sur le contenu. L'utilisation de ce modèle permet de recommander des produits de catégories indépendants dans le contexte d'un grand magasin.

Miyahara et Pazzani[38] mettent en œuvre un système de recommandation collaboratif basé sur un classificateur bayésien naïf. Pour ce faire, ils définissent deux classes: 'aime' et 'n'aime pas'. Dans ce contexte, ils proposent deux façons d'utiliser le classificateur bayésien naïf: Le modèle de données Transformé suppose que toutes les caractéristiques sont complètement indépendantes, et la sélection de caractéristique est implémentée comme une étape de prétraitement. D'autre part, le modèle de données Sparse suppose que seulement les caractéristiques connues qui sont informative pour la classification.

En outre, il ne fait que l'utilisation des données dont les deux utilisateurs ont noté en commun lors de l'estimation des probabilités. Les expériences montrent que ces deux modèles sont plus performants qu'un filtrage collaboratif basé sur une corrélation.

Pronk et al. [61] utilisent un classificateur bayésien naïf comme base pour intégrer le contrôle de l'utilisateur et améliorer la performance, en particulier dans les situations de démarrage à froid. Pour ce faire, ils proposent de maintenir deux profils pour chaque utilisateur:

a) Tirés de l'historique d'évaluations (rating en Anglais).

b) Explicitement créés par l'utilisateur.

Le mélange des deux classificateurs peut être contrôlé de manière à ce que le profil défini par l'utilisateur (b) soit favorisée aux premières étapes, quand l'historique d'évaluations est insuffisant, et le classificateur (a) succède lors d'étapes ultérieures.

Les Réseaux bayésiens hiérarchiques ont également été utilisés dans plusieurs cadres comme un moyen d'ajouter des connaissances de domaine pour le filtrage de l'information [39] . L'un des problèmes avec les réseaux bayésiens hiérarchiques, cependant, c'est qu'il est très coûteux de faire de l'apprentissage et de mettre à jour le modèle quand il y'a de nombreux utilisateurs en elle. Zhang et Koren[66] proposent une variation sur le modèle standard Espérance-Maximisation (EM) pour accélérer ce processus dans le scénario d'un système de recommandation basé sur le contenu.[64]

e) Machine à vecteurs de support (SVM)

Les machines à vecteurs de support (Support Vector Machine, SVM) appelés aussi séparateurs à vaste marge sont des techniques d'apprentissage supervisées destinées à résoudre des problèmes de classification. Les machines à vecteurs supports exploitent les concepts relatifs à la théorie de l'apprentissage statistique et à la théorie des bornes de Vapnik et Chervonenkis[62] . La justification intuitive de cette méthode d'apprentissage est la suivante : si l'échantillon d'apprentissage est linéairement séparable, il semble naturel de séparer parfaitement les éléments des deux classes de telle sorte qu'ils soient le plus loin possibles de la frontière choisie. Ces fameuses machines ont été inventées en 1992 par Boser et al[8] , mais leur dénomination par SVM n'est apparue qu'en 1995 avec Cortes et al. [15] . Depuis lors, de nombreux développements ont été réalisés pour proposer des variantes traitant le cas non-linéaire. Le succès de cette méthode est justifié par les solides bases théoriques qui la soutiennent. Elles permettent d'aborder des problèmes très divers dont la classification. SVM est une méthode particulièrement bien adaptée pour traiter des données de très haute dimension[32] .

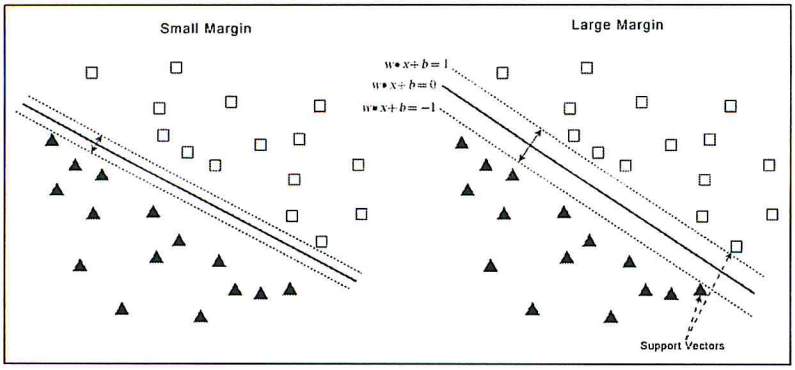

Figure 12 : Machine à vecteurs de support [64]

- **Les Machine à vecteurs de support dans les Système de recommandation :**

Les Machines à vecteurs de support ont récemment gagné en popularité pour leur performance et efficacité dans de nombreux contextes. Les SVM's ont également montré des résultats récents prometteurs dans les systèmes de recommandation.

Kang et Yoo[31] , par exemple, Effectuent une étude expérimentale qui vise à sélectionner la meilleure technique de prétraitement pour prédire les valeurs manquantes pour un système de recommandation basé sur SVM. En particulier, ils utilisent SVD (Décomposition en valeurs singulières) et la régression par les machines à vecteurs de support (SVR). Le système de recommandation basé sur SVM est construit d'abord par la binarisation des 80 niveaux de données disponibles sur les préférences d'utilisateur. Ils expérimentent avec plusieurs contextes et rapportent de meilleurs résultats pour un seuil de 32, C'est à dire une valeur de 32 et moins est classé comme 'préférer' et à une valeur plus élevée 'ne préfère pas'. L'id d'utilisateur est utilisé comme étiquette de classe et les valeurs positives et négatives sont exprimées en valeurs préférence 1 et 2.

Xia et al. [68] présentent différentes approches à l'aide de SVM pour les systèmes de recommandation dans le cadre de filtrage collaboratif. Ils étudient l'utilisation Machine à vecteurs de support avec lissage (Smooth Support Vector Machine, SSVM). Ils introduisent également une heuristique basée sur SSVM (SSVM-based heuristic, SSVMBH) pour estimer de manière itérative des éléments manquants dans la matrice user-item. Ils calculent les prévisions en créant un classificateur pour chaque utilisateur. Leurs résultats expérimentaux indiquent de meilleurs résultats pour le SSVMBH par rapport aux deux SSVM et le filtrage collaboratif traditionnel à base d'item et à base d'utilisateur.[64]

f) Réseau de neurones artificiels (ANN)

Les réseaux de neurones représentent la technique de data mining la plus utilisée. Pour certains utilisateurs, elle en est même synonyme. C'est une transposition simplifiée des neurones du cerveau humain. Dans leur variante la plus courante, les réseaux de neurones apprennent sur une population d'origine puis sont capables d'exprimer des résultats sur des données inconnues. Ils sont utilisés dans la prédiction et la classification dans le cadre de découverte de connaissances dirigée. Certaines variantes permettent l'exploration des séries temporelles et des analyses non dirigées (réseaux de Kohonen). Le champ d'application est très vaste et l'offre logicielle importante.

Cependant, on leur reproche souvent d'être une "boite noire" : il est difficile de savoir comment les résultats sont produits, ce qui rend les explications délicates, même si les résultats sont bons.

Donc, Utiliser des technologies d'intelligence artificielle afin de découvrir par l'apprentissage du moteur des liens non procéduraux. Ces deux dernières techniques s'appuient sur des algorithmes mathématiques et tentent à travers des méthodes d'apprentissage de constituer des logiques non procédurales[56] .

Dans un réseau de neurones, un neurone est simplement une fonction non linéaire, de variables réelles et bornée. Cette fonction est généralement définie comme suit :

$$f(x_1, \ldots, x_k;\ w_1, \ldots, w_k) = \left| \sum_{i=1}^{k} w_i x_i \right|$$

où les variables w_1, ..., w_k correspondent à des poids à associer aux variables x_1, ..., x_k, qui sont déterminés à partir d'un corpus d'apprentissage. La fonction tangente hyperbolique est une fonction sigmoïde qui a certaines propriétés particulièrement appropriées pour l'apprentissage de réseaux de neurones [9] . De tels neurones sont associés en réseau selon deux types d'architecture : les réseaux bouclés qui correspondent à des graphes orientés avec circuit et les réseaux non-bouclés qui correspondent à des graphes orientés sans circuit.

Dans le cadre de la recommandation basée sur le contenu, les variables x_1, ..., x_k correspondent à la fréquence des termes utilisés pour caractériser les ressources (qui peut être normalisée par rapport à la longueur du texte). L'architecture la plus fréquemment adoptée est l'architecture en réseaux non bouclés avec une structure de perceptron multicouche [37] . Plus précisément, cette structure consiste en général en k entrées (les k attributs d'une ressource), une couche d'un certain nombre de neurones cachés, et un certain nombre de neurones de sortie. Chaque neurone de sortie indique un score permettant de déterminer si une ressource

appartient à la classe du niveau d'appréciation à laquelle il est associé. Un algorithme répandu pour effectuer l'apprentissage des poids est l'algorithme PLA (Perceptron Learning Algorithm). Il consiste à initialiser les variables de façon aléatoire et à les ajuster itérativement de façon à minimiser le nombre de ressources disposées dans de mauvaises classes.

En plus de permettre un apprentissage rapide, l'utilisation de réseaux de neurones a l'avantage de permettre un ajustement particulièrement fin grâce à l'utilisation de la fonction sigmoïde. Selon le domaine d'application il peut s'avérer plus ou moins efficace que ses alternatives [47] .

- **Les réseaux de neurones dans les Système de recommandation :**

Les réseaux de neurones peuvent être utilisés de la même manière que les réseaux bayésiens pour construire des systèmes de recommandation à base de modèles. Cependant, il n'existe aucune étude concluante à savoir si les réseaux de neurones introduisent des gains de performance.[64]

Les réseaux de neurones ne sont pas souvent utilisés dans les systèmes de recommandation. Il pourrait y avoir plusieurs raisons [23] :

- Pour les tâches de classification dans les systèmes basés sur le contenu, il n y'a pas besoin d'utilisé des classificateurs non linéaires complexes, Comme l'expérience de Pazzani et Billsus[47] semble démontrer.
- L'apprentissage dans les grands réseaux de neurones classiques est très lent.
- Les réseaux de neurones ont un effet d'une 'boite noire', ce qui rend l'interprétation des résultats obtenus très difficile. Même si certaines méthodes d'explications ont été proposées, par exemple [24] ,la complexité perçue des réseaux de neurones est toujours un problème dans les applications e-commerce.

3. Étude comparative:

	Arbre de décision	Réseau de neurones	Bayésien naïf	KPPV	SVM
Précision en général	Bon	Très Bon	Moyen	Bon	Excellent
Vitesse d'apprentissage	Très Bon	Moyen	Excellent	Excellent	Moyen
Vitesse de la classification	Excellent	Excellent	Excellent	Moyen	Excellent
La tolérance aux valeurs manquantes	Très Bon	Moyen	Excellent	Moyen	Bon
Tolérance aux attributs non pertinents	Très Bon	Moyen	Bon	Bon	Excellent
Tolérance aux attributs redondants	Bon	Bon	Moyen	Bon	Très Bon
Tolérance aux attributs hautement interdépendants	Bon	Très Bon	Moyen	Moyen	Très Bon
Traitement des attributs discrets / binaire / continue	Tous -	Non Discrète	Non Continu	Tous	Non Discrète
Tolérance au bruit	Bon	Bon	Très Bon	Moyen	Bon
Traitement des risques de sur-apprentissage	Bon	Moyen	Très Bon	Très Bon	Bon
Tentatives d'apprentissage progressif	Bon	Très Bon	Excellent	Excellent	Bon
La capacité d'explication / transparence des connaissances / classification	Excellent	Moyen	Excellent	Bon	Moyen
Support de classification multiple	Excellent	Naturellement Etendue	Naturellement Etendue	Excellent	Classificateur Binaire

Table 5 : Étude comparative des techniques de classification couramment utilisées[33]

- **SVM** est parmi les algorithmes connu pour sa et précision et dispose d'une base théorique solide et très rapide dans la classification qui nécessite seulement une dizaine d'exemples d'entraînement et aussi insensibles au nombre de dimension. SVM est moins sensible au sur-apprentissage (Où apprentissage par cœur : le modèle apprend par-cœur l'ensemble de la base d'apprentissage, sans pour autant être capable de généraliser.)que les autres méthodes. Les SVM garantissent de trouver la meilleure solution (en termes de la marge de séparation), ce qui peut être fait de manière efficace. L'un des points forts de l'algorithme SVM est la vitesse de classification.

Les points faibles de SVM qu'il est coûteux en calcul et en mémoire, comme les méthodes de résolution de programmes quadratiques qui nécessitent de grandes opérations matricielles

ainsi que du temps des calculs numériques .Les SVM sont extrêmement lents à l'apprentissage [33]

- **L'algorithme Bayésien Naïf** est simple à mettre en œuvre, il a une grande efficacité de calcul et un taux de classement élevé. Il prédit des résultats précis pour la plupart des problèmes de classification et de prédiction, Par contre la précision de l'algorithme diminue si la quantité de données est faible. Pour obtenir de bons résultats, il nécessite un très grand nombre d'enregistrements[55] .L'algorithme bayésien naïf est moins précis par rapport aux autres classificateur[33] .
- **L'algorithme KPPV** est facile à comprendre et facile à mettre en œuvre la technique de classification, Il donne de bons résultats sur l'application dans laquelle un échantillon peut avoir plusieurs étiquettes de class[33] . L'une des points forts de l'algorithme KPPV est la vitesse d'apprentissage. Et il est robuste à des données d'entraînement bruyant et efficace si les données d'entraînement sont grandes. L'autre avantage est que le KPPV est un classificateur paresseux, il ne nécessite pas d'apprendre et de maintenir un modèle donné. Par conséquent, le système peut s'adapter à des changements rapides dans la matrice des évaluations de l'utilisateur, Malheureusement, cela conduit à recalculer les voisinages et donc la matrice de similarité[64] .

Parmi les inconvénients, La performance de l'algorithme qui dépend de la valeur du paramètre k utilisé, et l'autre inconvénient est qu'il a une vitesse de classification moyenne par rapport aux autres algorithmes[33] .

- **Les arbres de décision** sont très simples et rapides. Un arbre de décision ne nécessite pas de connaissances de domaine ou paramétrage et il est capable de manipuler des données de haute dimension. Les Arbres de décision ont une bonne précision (peut dépendre de données à portée de main).

Les arbres de décision peuvent être beaucoup plus complexes pour la représentation de quelques concepts en raison du problème de réplication[33] .D'autres inconvénients des arbres de décision :

- Coûteux en mémoire, lorsqu'il s'agit de grandes bases de données.
- Sensibilité au bruit et points aberrants.
- Instabilité (effet papillon): La modification d'une variable dans l'arbre transforme l'arbre complètement.

Comme nous l'avons déjà cité, Il est très difficile et peu pratique de construire un arbre de décision qui tente d'expliquer toutes les variables impliquées dans le processus de prise de décision.
Les expériences dans[59] montrent que les arbres de décision ne fournissent pas les meilleures performances en tant que système de recommandation.

- **Les réseaux de neurones artificiels** ont une résistance naturelle aux données bruitées lors de l'apprentissage. En effet, si la base d'exemples est assez grande, une erreur ne faussera pas beaucoup la mise à jour des poids [10] . Ils sont bien adaptés pour les entrées/sorties numériques continues. Les réseaux de neurones artificiels sont connus pour être des modèles intrinsèquement parallèles. Les techniques de parallélisation peuvent être utilisées pour accélérer le processus de calcul[33] .
Un réseau de neurones a une structure statique. Une fois que la structure est fixée (nombre de neurones d'entrée, sortie, neurones cachés) et que le réseau a appris, il est impossible de lui faire apprendre de nouvelles données sans recommencer son apprentissage du début, contrairement aux algorithmes de type paresseux k-plus proche voisins, classificateur naïf de Bayes[10] . Les réseaux de neurones ont un effet d'une 'boite noire', ce qui rend l'interprétation des résultats obtenus très difficile.
Comme nous l'avons déjà cité précédemment, Les réseaux de neurones ne sont pas souvent utilisés dans les systèmes de recommandation[23] .

L'étude comparative a montré que chaque algorithme a ses propres avantages et inconvénients ainsi que son propre domaine de mise en œuvre. Aucun algorithme ne peut satisfaire tous les critères. D'après nos recherches le classificateur KPPV est le plus souvent utilisé pour son efficacité dans la recommandation.

Conclusion :

Dans ce chapitre, nous avons présenté un état de l'art sur l'analyse prédictive. Aujourd'hui, cette dernière est plus en plus utilisé dans le web (site e-Commerce, réseaux sociaux, Plateforme multimédia) et nul doute que chacun d'entre nous a déjà été confronté à ce genre de fonctionnalités.

Chapitre IV

Conception du système

Nous présentons dans cette partie la conception de notre système et les différentes fonctionnalités offertes pour les utilisateurs.

Introduction

Dans ce chapitre nous présentons l'architecture conceptuelle de notre système qui permet de faire des recommandations sur des offres de vol et d'hôtel pour voyageur aérien. Il est donc nécessaire d'enregistrer les villes, les horaires des vols, les places disponibles sur chaque vol, le prix, les hôtels, etc. Le programme doit pouvoir afficher les vols et hôtels disponibles, selon les désirs des utilisateurs, et permettre aux utilisateurs de réserver leurs places.

1. Conception du système de réservation :

1.1 Diagramme des cas d'utilisation:

Les diagrammes de cas d'utilisation décrivent les services les plus importants rendus par un système. Partant des acteurs, participants externes qui interagissent avec le système, ils représentent les cas les plus importants du système en cours d'utilisation. Un cas d'utilisation peut être divisé en diagrammes de séquence, qui détaillent les différentes fonctions du cas d'utilisation

Identification des acteurs

Les acteurs sont des entités externes qui interagissent avec le système, comme une personne humaine, un autre système ou un robot. Les acteurs sont représentés par un pictogramme sous-titré par le nom. Les principaux acteurs qui auront à utiliser le système sont les suivants :

Administrateur	est l'administrateur du système, il a le droit de faire toutes les actions affectées aux autres acteurs, plus la gestion des comptes des utilisateurs.
Client	Cet acteur peut consulter les recommandations, réserver vol ou chambre dans un hôtel, et suivre le déroulement d'une réservation.
Visiteur	Il a le droit de consulter les offres et les prix des vols et d'hôtels, et d'inscrire à partir du portail.
Moteur de recommandation	Recommander des offres selon le profil de l'utilisateur

Table 6 : Identification des acteurs

Description détaillée des cas d'utilisation :

- Gestion des réservations hôtels et vols
- Recommandation

Diagramme de cas d'utilisation global

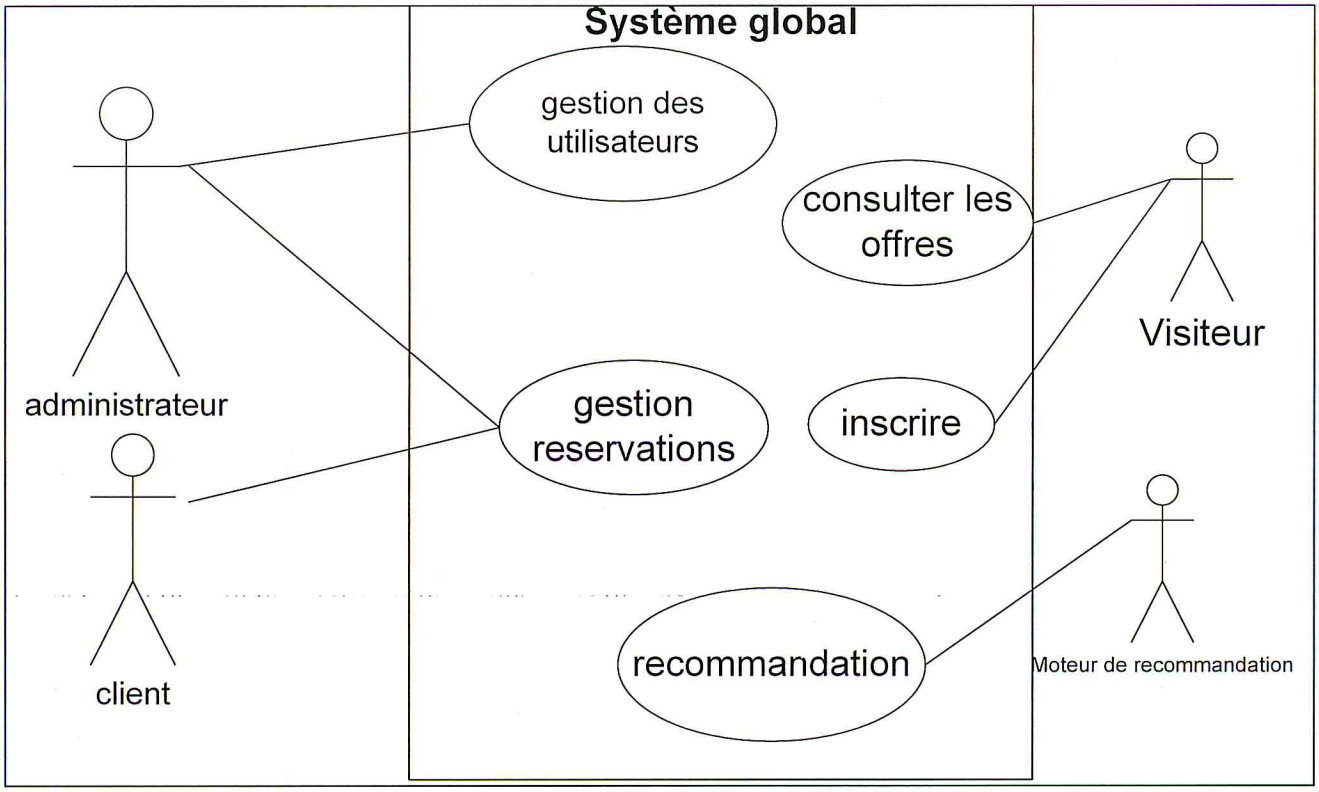

Figure 13 : Diagramme de cas d'utilisation global

Gestion des utilisateurs

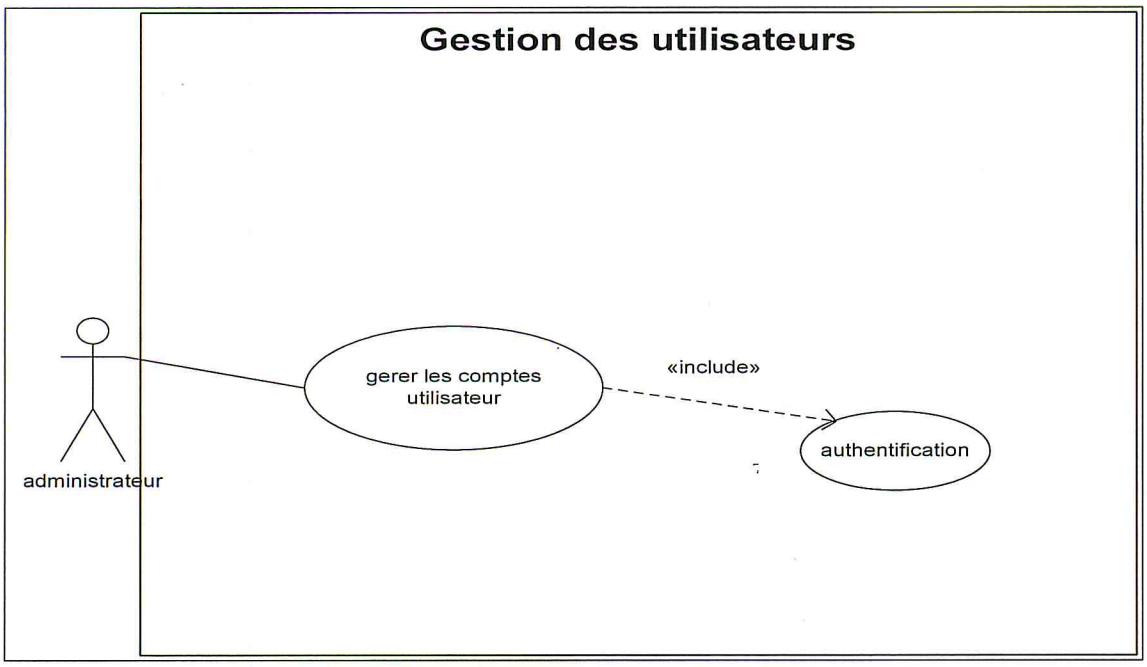

Figure 14: Diagramme de cas d'utilisation: gestion utilisateur

Sommaire d'identification :

Titre : Système de gestion des utilisateurs.

Objectifs : Gérer les comptes des utilisateurs.

Résumé : Cette fonctionnalité permet :

1 - Aux administrateurs de gérer complètement les comptes des utilisateurs.

Acteurs : Administrateur.

Description détaillée :

Pré conditions : L'administrateur doit s'authentifier pour avoir accès aux fonctionnalités du système.

Description du traitement nominal :

L'administrateur peut :

1. Gérer les comptes des utilisateurs.

Exceptions :

[Exception 1 : Champs Obligatoires] : Message d'erreur si l'un des champs obligatoires n'est pas rempli.

Gestion des réservations

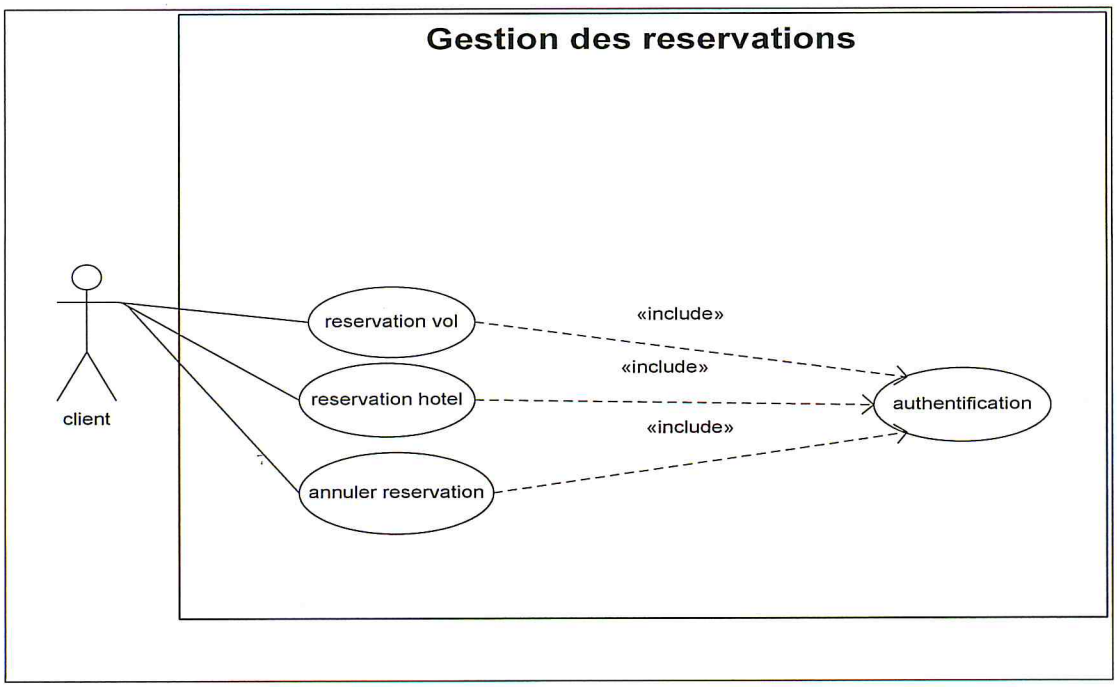

Figure 15: Diagramme de cas d'utilisation gestion des réservations

Sommaire d'identification :

Titre : Système de gestion des réservations.

Objectifs : Gestion des réservations.

Résumé : Cette fonctionnalité permet :

1 - Aux Client de Réserver sur un vol ou bien un hôtel, annuler réservation.

2 - Aux Visiteur de Consulter l'horaire des vols, les prix des offres vol et hôtel, et de s'inscrire comme client d'après le portail de l'agence.

Acteurs : Client.

Description détaillée :

Pré conditions : L'agent doit s'authentifier pour avoir accès aux fonctionnalités du système.

Description du traitement nominal :

Le client peut :

1. Réserver un vol. Réserver dans un hôtel
2. Annuler une réservation.

Exceptions :

[Exception 1 : Champs Obligatoires] : Message d'erreur si l'un des champs obligatoires n'est pas rempli.

Système de recommandation

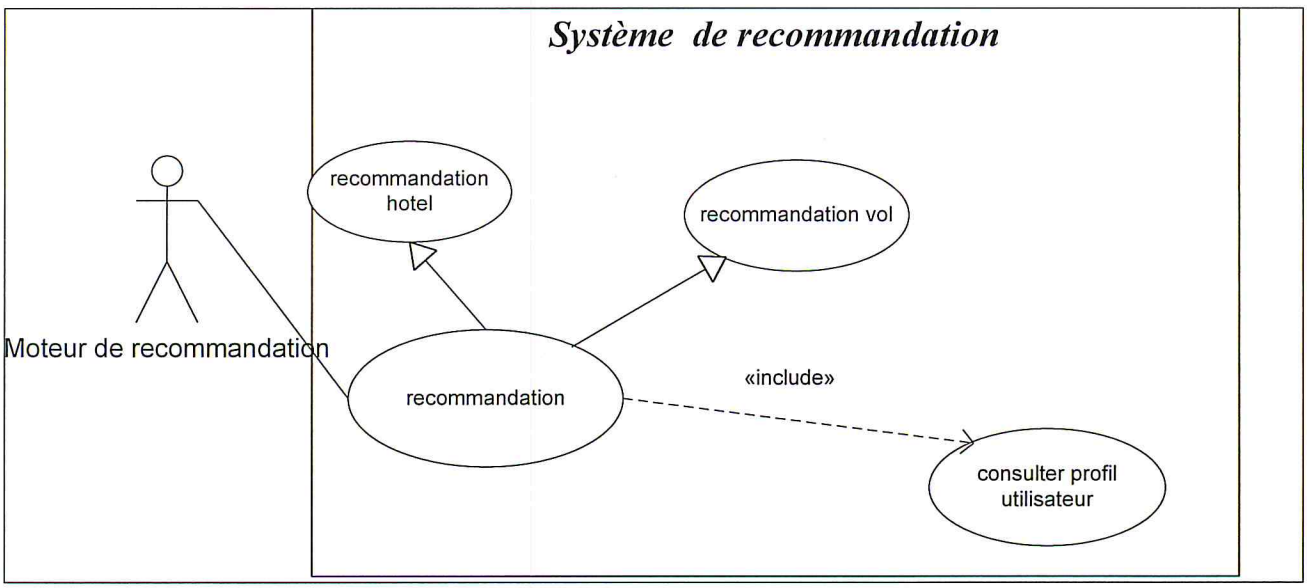

Figure 16: Diagramme de cas d'utilisation système de recommandation

Sommaire d'identification :

Titre : Système de recommandation.

Objectifs : Recommander des vols et des hôtels.

Résumé : Cette fonctionnalité permet : de recommander aux utilisateurs des offres susceptibles de les intéresser.

Acteurs : Moteur de recommandation.

Description détaillée :

Pré conditions : le Moteur de recommandation doit accéder au profil des utilisateurs.

Description du traitement nominal : Moteur de recommandation peut recommander des vols et des hôtels.

1.2 Diagramme de séquence :

Le diagramme de séquence décrit la dynamique du système. À moins de modéliser un très petit système, il est difficile de représenter toute la dynamique d'un système sur un seul diagramme. Aussi la dynamique globale sera représentée par un ensemble de diagrammes de séquence, chacun étant généralement lié à une sous fonction du système.

Le diagramme de séquence décrit les interactions entre un groupe d'objets en montrant, de façon séquentielle, les envois de message qui interviennent entre les objets. Le diagramme peut également montrer les flux de données échangées lors des envois de message.

Description détaillée des diagrammes de séquence :

Réserver dans un hôtel.

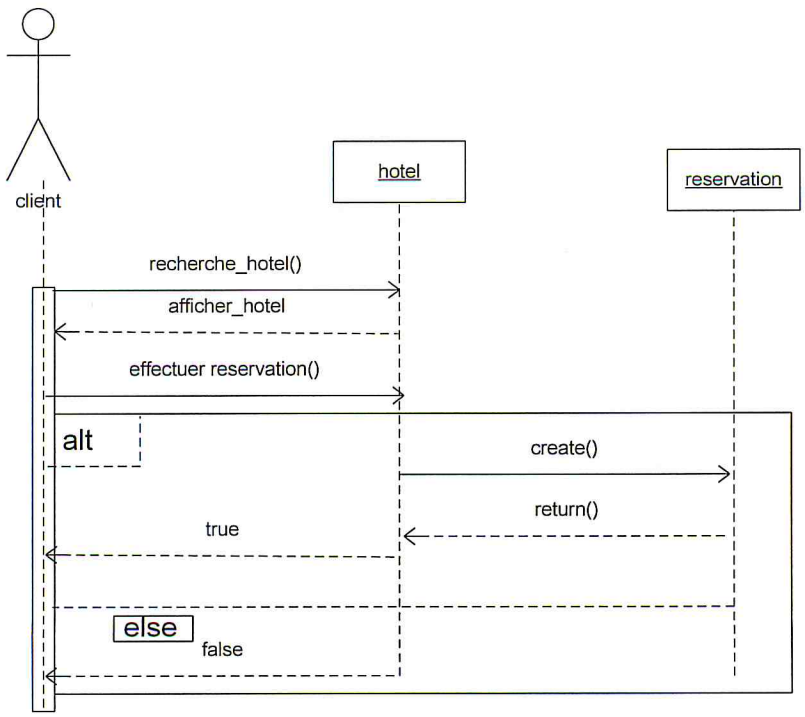

Figure 17: Diagramme de séquence réservation hôtel

Pré conditions :

Le Client s'est authentifié sur le système.

Description : Le client recherche un hôtel si le client demande une réservation il remplis le formulaire ex: nombre de chambre, nombre de personne …) - Si le nombre de places réservés est inférieur au nombre de place disponible dans un hôtel la création de la réservation se lance, si non une erreur s'affichera au client.

Réserver un vol.

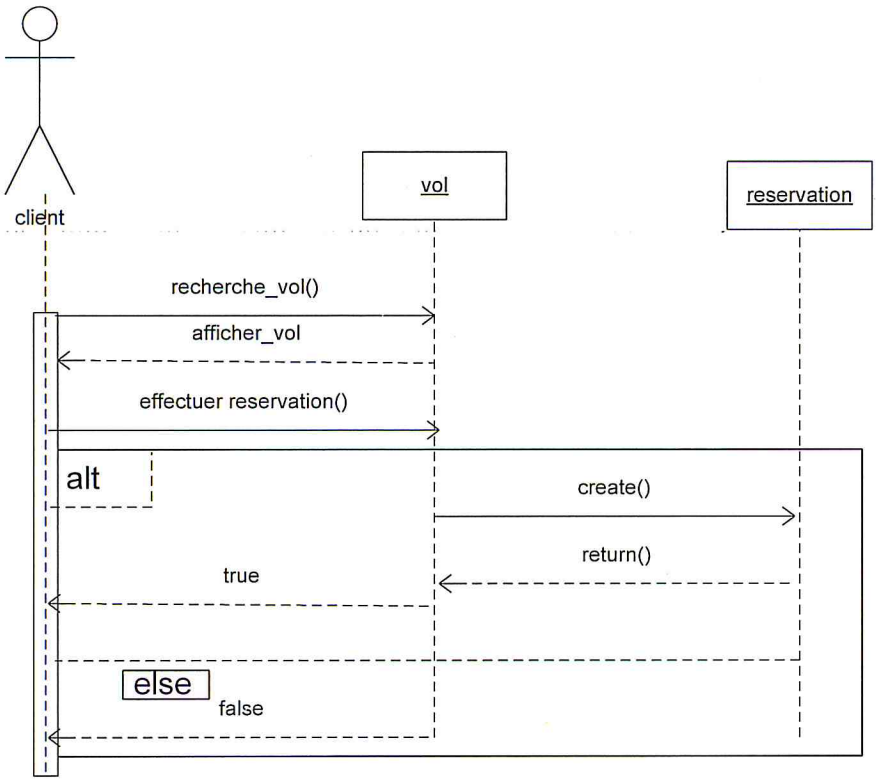

Figure 18: Diagramme de séquence réservation vol

Pré conditions :

Le Client s'est authentifié sur le système.

Description : Le client sélectionne un vol, si le client demande une réservation il remplit le formulaire nombre de personne, la classe voulu…), Si le nombre de places réservés est inférieur au nombre de place disponible dans un vol la création de la réservation se lance, sinon une erreur s'affichera au client.

1.3 Diagramme de classes :

Le diagramme de classe de conception représente bien la structure statique du code, par le biais des attributs et des relations entre classes C'est un diagramme principal qui est la vue de plus haut niveau avec l'ensemble des classes de l'application, Il montre les briques de base statiques : classes, sassociations, interfaces, attributs, opérations, généralisations, etc.

Règles de gestion

La réalisation du diagramme de classe se base sur le dictionnaire de données et les règles de gestion. L'analyse sémantique des données du dictionnaire permet de les regrouper dans des entités à part. Les liens qui les relient tiennent compte des règles de gestion. Les règles de gestion relatives à notre activité sont :

1. Une compagnie est composée d'un ou plusieurs vols.
2. Un vol appartient à une seule compagnie.
3. Un utilisateur peut donner un seul avis sur un hôtel dans une date donnée.
4. Un hôtel à plusieurs avis.
5. Un utilisateur peut donner un seul avis sur une compagnie dans une date donnée.
6. Une compagnie à plusieurs avis.
7. Un utilisateur peut effectuer une seule réservation dans un vol dans une date donnée.
8. Un vol à plusieurs réservations.
9. Un utilisateur peut effectuer une seule réservation dans un hôtel dans une date donnée.
10. Un hôtel à plusieurs réservations.

Dictionnaire de données

- **Le dictionnaire des données :**
  - Permet de décrire le cadre de l'utilisation des classes dans le domaine d'application. Il doit inclure les associations, les attributs et les opérations.
  - La description de chaque classe objet est donnée par :
    - Description des classes.
    - Description des associations.
    - Description des méthodes de classes.
- ***Une classe :***
  - une classe définit la structure commune d'un ensemble d'objets et permet la construction d'objets instances de cette classe. Une classe est identifiée par son nom.

- ***Une association :***
  - Une association exprime une connexion sémantique bidirectionnelle entre deux classes qui décrit les connexions structurelles entre leurs instances. Une association indique donc qu'il peut y avoir des liens entre des instances des classes associées.
- ***Multiplicité ou cardinalité :***
  - La multiplicité associée à une terminaison d'association, d'agrégation ou de composition déclare le nombre d'objets susceptibles d'occuper la position définie par la terminaison d'association.

Description des classes

CLASSE	*ATTRIBUTS*	*RUBRIQUES (Type, Taille)*	*METHODES*
Utilisateur	Code Utilisateur Nom Prenom Email Tel Pays Adresse Date de naissance Sexe Nom d'utilisateur Mot de passe	id_utilisateur (N,8) nom(C,25) prenom(C,25) email(C,50) tel(N,10) pays(C,25) adr(C,150) date_n(C,8) sexe(C,5) nom_utilisateur(C,25) mot_de_pass(AN,25)	Ajouter () Modifier () Supprimer () Consulter () ReserverHotel( ) ReserverVol() EcrireAvis()
Compagnie	Code compagnie Nom compagnie adresse pays	id_compagnie(C, 2) nom_compagnie(C, 40) adresse(C, 150) pays(C, 25)	Ajouter () Modifier () Supprimer () Consulter ()
Hotel	Code hôtel Nom hôtel pays adresse nombre d'étoiles	id_hotel (N,8) nom_hotel(C, 50) pays(C, 25) adresse(C, 150) valeur (N, 5)	Ajouter () Consulter () Modifier () Supprimer ()
Vol	Code vol Date départ Date d'arrivé Heure départ Heure d'arrivé Prix Durée Nombre de places	id_vol(C,6) date_depart(C,8) date_arrive(C,8) heure_depart(C, 5) heure_arrive(C, 5) prix(N,6) duree(C, 5) nb_places(N,4)	Ajouter () Modifier () Supprimer () Consulter ()
ReservationHotel	Code reservation Code Utilisateur Date debut Date fin Nombre de chambre Nombre d'adulte Nombre d'enfant	id_reservation(N,6) id_utilisateur (N,8) date_debut(C,8) date_fin(C,8) nb_chambre(N, 2) nb_adulte(N, 2) nb_enfant(N,2)	Ajouter () Modifier () Supprimer () Consulter ()

ReservationVol	Code reservation Code Utilisateur Date Type classe Nombre d'adulte Nombre d'enfant Nombre de bébé	id_reservation(N,6) id_utilisateur (N,8) date (C,8) type_classe(C,1) nb_adulte(N, 2) nb_enfant(N, 2) nb_bebe(N,2)	Ajouter () Modifier () Supprimer () Consulter ()
AvisHotel	Code avis Code Utilisateur Code hotel Note Date	id_avis(N,8) id_utilisateur (N,8) id_hotel(N,8) note(N,1) date(C,8)	Ajouter () Modifier () Supprimer () Consulter ()
AvisVol	Code avis Code Utilisateur Code compagnie Note Date	id_avis(N,8) id_utilisateur (N,8) id_compagnie(C,2) note(N,1) date(C,8)	Ajouter () Modifier () Supprimer () Consulter ()

Table 7 : Description des classes

Description des associations

Association	***Classes impliquées***	***Cardinalité***
Attribue	Utilisateur Compagnie	1 *
Attribue1	Utilisateur Hotel	1 *
Effectue	Utilisateur Vol	1 *
effectue1	Utilisateur Hotel	1 *

Table 8 : Description des associations

Diagramme de classes pour le système de réservation

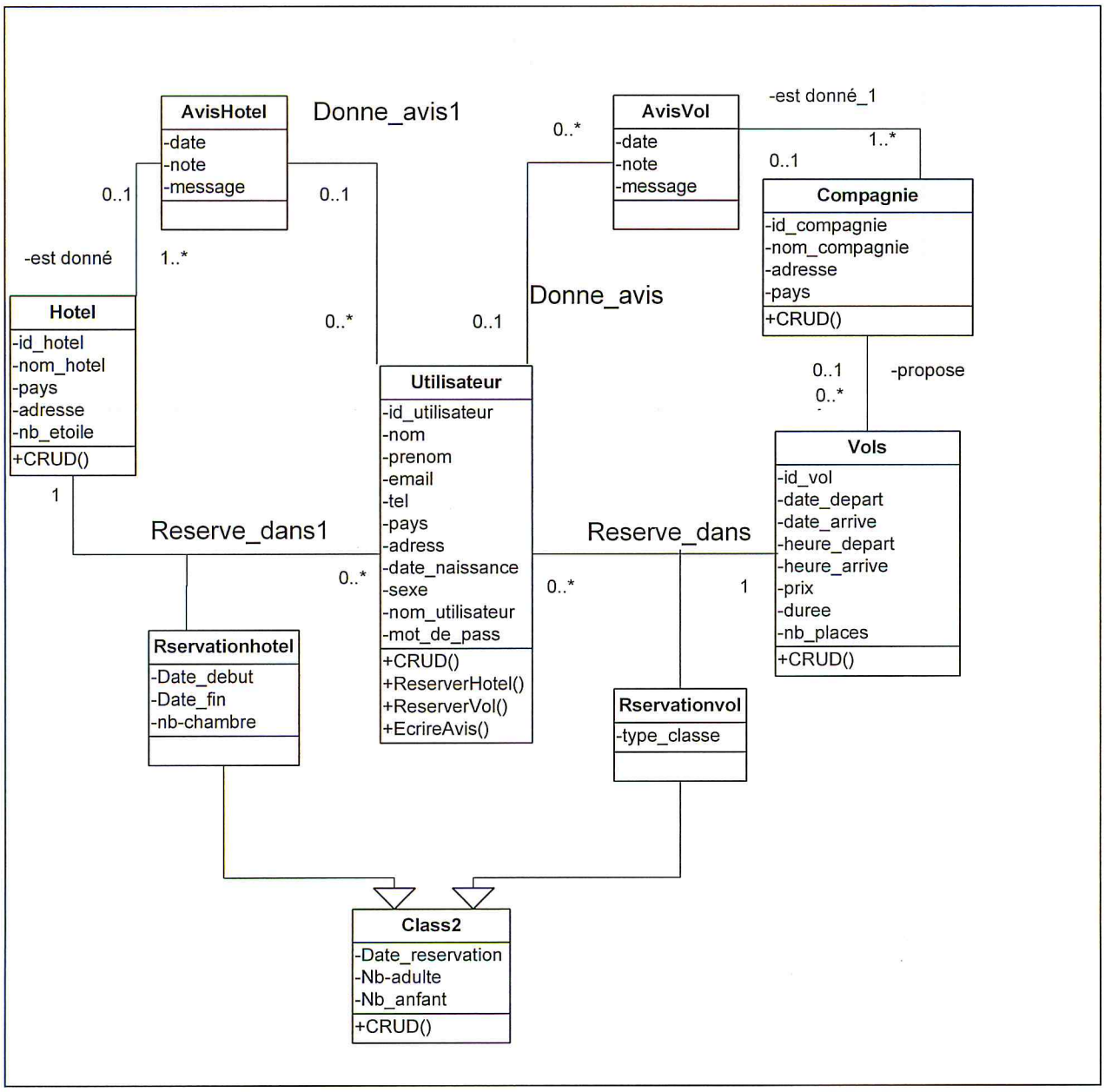

Figure 19: Diagramme de classes

2. Conception du système de recommandation :

Pour la recommandation, la solution envisagée consiste à s'appuyer sur les historiques des clients correspondant aux avis des utilisateurs sur des vols ou bien sur des hôtels. Nous obtenons ces informations via les données classiques (structurées) depuis notre base de données relationnelle, mais également depuis des sources de données externe à notre système comme TripAdisor.com, L'analyse de ces données volumineuses permet la prise de décision, et de recommander à l'utilisateur des offres susceptibles de l'intéresser.

Afin d'arriver à la recommandation, on doit passer par certains étapes(voir ***Figure 20***) :

1. Chargement des offres de vols, hôtels et les avis sur ces offres, ces données sont volumineuses, Les SGBD traditionnels ne peuvent pas traiter des volumes de téraoctets de données. Seuls Les systèmes de fichiers distribués peuvent affronter des volumes de type Big Data. Les données sont chargées à partir de notre base de données et aussi des sources des données externes.

2. Appliquer les algorithmes pour la recommandation, Il y'a deux types de filtrage collaboratif, Ceux basés sur les utilisateurs, et ceux se basés uniquement sur les items et les similarités entre ceux-ci.
   - Le Filtrage collaboratif basé sur l'utilisateur(UserSimilarity Method) [42] principe prendre le voisinage d'un utilisateur u , Il choisit les plus similaires et prédit la note qui est normalement attribuée par l'utilisateur u (voir *Table 11 : Ficher de sortie 2* -).

   - Le Filtrage collaboratif basé sur l'item (ItemSimilarity) [30] utilise l'algorithme des k plus proches voisins. Le principe est de calculer d'abord la distance entre les items (vol ou hôtel). Le calcul est appliqué sur un fichier d'entré qui contient les utilisateurs, l'item (vol ou bien hôtel) et la note comprise entre 1 et 5 (voir *Table 9 : Fichier d'entré*), On obtient les distances entre les items (voir *Table 10 : Ficher de sortie 1 - Similarité entre les items*). On prend après les k plus proches items.

Ficher d'entré		
Id_utilisateur	Id_item	Note
User_1 User_1 . . User_i	Item_1 item_2 . . item_j	3.0 5.0 . . Note (i,j)

Table 9 : Fichier d'entré

Ficher de sortie 1 : Similarité entre les items		
Id_item_a	Id_item_b	Similarité
Item_1	Item_2	0.48
Item_1	item_3	0.86
.	.	.
.	.	.
Item_i	Item_j	Similarité (i,j)

Table 10 : Ficher de sortie 1 - Similarité entre les items

Ficher de sortie 2 : Notes prédites		
Id_utilisateur	Id_item	Note prédit
User_1	Item_2	3.48
User_1	item_3	4.86
.	.	.
.	.	.
User_i	Item_j	Prédiction (i,j)

Table 11 : Ficher de sortie 2 - Notes prédites

3. Chargement des résultats dans la base de données pour la mise à jour des appréciations des utilisateurs pour des recommandations nouvelles et pertinentes.

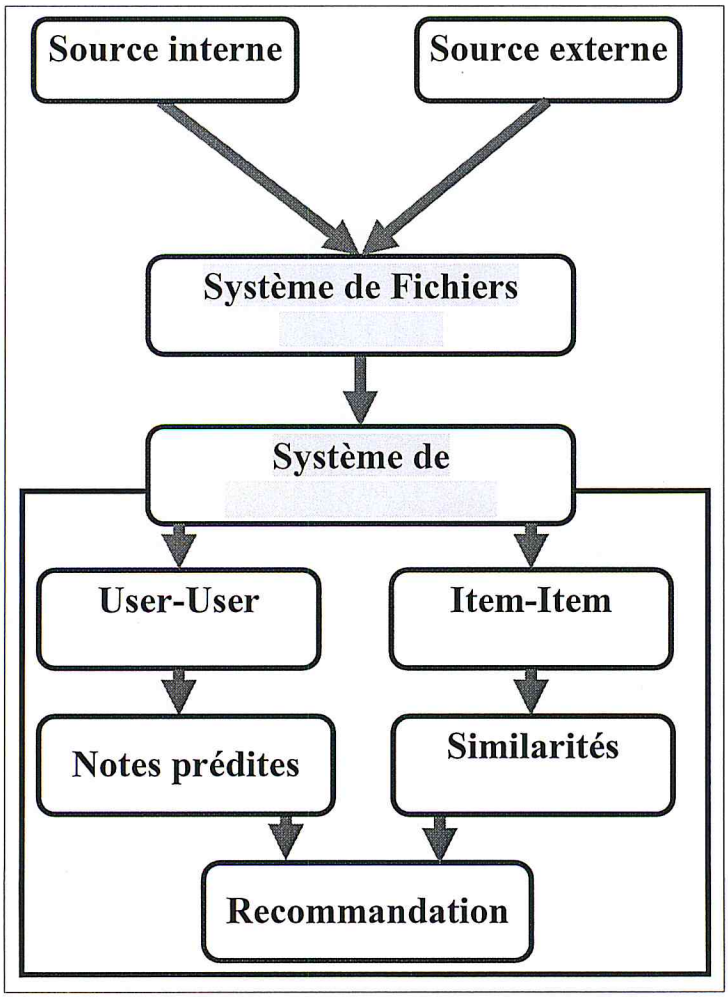

Figure 20 : Architecture du système de recommandation

2.1 Filtrage collaboratif basé sur item :

L'algorithme peut être décrit dans le *pseudo-code* suivant [30] :

Pour chaque item **i** dans la liste des items

  Pour chaque client **C** qui a noté **i**

    Pour chaque item **j** qui a été noté par **C**

      Enregistrez que le client **C** a noté **i** et **j**

  Pour chaque item **j**

    Calculer la similarité entre **i** et **j**

$$SIM_{i,j} = \frac{\sum_{u \in U}(r_{u,i} - \overline{r_i}) \times (r_{u,j} - \overline{r_j})}{\sqrt{\sum_{u \in U}(r_{u,i} - \overline{r_i})^2 \times \sum_{u \in U}(r_{u,j} - \overline{r_j})^2}}$$

U : L'ensemble des utilisateurs qui ont notés à la fois l'item i et j.
$\overline{r_i}$: La moyenne des notes de l'itemi
$r_{u,i}$: La note de l'utilisateur u pour l'item i

Table 12 : Algorithme de calcule de similarité entre les items

Exemple :

	Item 1	**Item 2**	**Item 3**	**Item 4**	…
Ali	-	2	7	8	…
Hamdi	4	1	-	7	…
Ishak	3	8	-	4	…

Table 13: Table de correspondance utilisateur - item

$$SIM_{1,4} = \frac{\sum_{u \in Ali,Hamdi,Ishak}(r_{u,i} - \overline{r_a}) \times (r_{u,j} - \overline{r_j})}{\sqrt{\sum_{u \in Ali,Hamdi,Ishak}(r_{u,i} - \overline{r_i})^2} \times \sqrt{\sum_{u \in Ali,Hamdi,Ishak}(r_{u,j} - \overline{r_j})^2}}$$

$$SIM_{1,4} = \frac{(4 - 3.5) \times (7 - 5.5) + (3 - 3.5) \times (4 - 5.5)}{\sqrt{((4 - 3.5)^2 + (3 - 3.5)^2)} \times \sqrt{((7 - 5.5)^2 + (4 - 5.5)^2)}} = 1$$

Donc la similarité entre l'item 1 et 4 est égale à 1.

- **La distribution de cet algorithme avec la logique mapreduce :**

NB : On va représenter les entrée et les sorties des méthodes Map et Reduce comme suit :

Méthode> entrée/sortie : <**clé**>, <**valeur**>

Chaque étape représente un Job MapReduce.

Etape 1 : Transformation du fichier d'entrée (Table 9)

Map > entrée : <**numero_ligne**>,<**contenue_ligne**>

Pour chaque ligne on prend l'identifiant d'utilisateurs comme clé, et l'identifiant de litem avec la note attribué par l'utilisateur comme valeur.

Map> sortie : <**id_utilisateur**>,<**note_utilisateur**>

Reduce > entrée : <**id_utilisateur**>, <**liste <note_utilisateur>>**

Pour chaque utilisateur on collecte ces notes. Et on prend l'identifiant d'utilisateurs comme clé et ses notes comme valeur

Reduce > sortie : <**id_utilisateur**>, <**notes_utilisateur**>

Etape 2 : Calcule des similarités

Map > entrée : <**id_utilisateur**>, <**notes_utilisateur**>

Pour chaque utilisateur on collecte chaque paire d'items noté par l'utilisateur et on prend l'identifiant des deux comme clé <item1, item2>, et leur notes comme valeur <note_item1, note_item2>

Map> sortie : <**item1, item2**>, <**note_item1, note_item2**>

Reduce > entrée : <**item1, item2**>, <**liste<note_item1, note_item2>>**

Pour chaque paire d'items on calcule la similarité entre ces deux items avec les vecteurs des notes de chaque item (liste <note_item1, note_item2>), enfin on retourne l'identifiant des deux items comme clé <item1, item2> et la similarité entre eux comme valeur <similarité>.

Reduce > sortie :<item1, item2>, <similarité>

Le résultat final de cet algorithme est représenté dans la (*Table 10*).

2.2 Filtrage collaboratif basé sur l'utilisateur :

L'algorithme peut être décrit dans le *pseudo-code* suivant [42] :

Etape 1 : Calculer la similarité entre chaque couple d'utilisateur **u**,**v** , pour notre cas on a choisi le coefficient de corrélation de Pearson pour calculer la similarité entre les utilisateurs :

$$SIM_{a,u} = \frac{\sum_{i \in N}(r_{a,i} - \overline{r_a}) \times (r_{u,i} - \overline{r_u})}{\sqrt{\sum_{i \in N}(r_{a,i} - \overline{r_a})^2} \times \sqrt{\sum_{i \in N}(r_{u,i} - \overline{r_u})^2}}$$

N: L'ensemble des items notés à la fois par a et b.
$\overline{r_a}$: La moyenne des notes de l'utilisateur a
$r_{a,i}$: La note de l'utilisateur a pour l'item i

Etape 2 : pour chaque utilisateur **u**, sélectionnez les k-plus proches voisins
Etape 3 :
pour chaque utilisateur **u**
pour chaque item **i** dans la liste des items non-noté de l'utilisateur **u**
prédire la note de l'utilisateur **u** pour l'item **i**

$$P_{a,i} = \overline{r_a} + \frac{\sum_{u=1}^{n}(r_{u,i} - \overline{r_u}) \times SIM_{a,u}}{\sum_{u=1}^{n} SIM_{a,u}}$$

n : Le nombre d'utilisateurs dans le voisinage
$\overline{r_a}$: La moyenne des notes de l'utilisateur a
$r_{a,i}$: La note de l'utilisateur a pour l'item i
$SIM_{a,u}$: similarité entre l'utilisateur a et u

retourner les k-item qui ont les notes les plus grandes pour l'utilisateur **u**

Table 14 : Algorithme de prédiction des notes d'utilisateurs

Exemple :
Supposons maintenant que nous souhaitions prédire les notes que donnerait Hamdi à l'item 3. Nous utilisons alors la formule suivante :

Supposons que nous voulions prédire la note que donnerait Hamdi à l'item 3 et que nous prenions Ali et Ishak comme voisins. Nous utilisons alors la formule suivante :

$$\boldsymbol{P_{Hamdi,3}} = \overline{r_{Hamdi}} + \frac{\sum_{u=Ali,Ishak}(r_{u,i} - \overline{r_u}) \times SIM_{Hamdi,u}}{\sum_{u=Ali,Ishak} SIM_{Hamdi,u}}$$

$$\boldsymbol{SIM_{Hamdi,Ali}} = 1 \text{et} \boldsymbol{SIM_{Hamdi,Ishak}} = -1$$

$$\boldsymbol{P_{Hamdi,3}} = \frac{17}{3} + \frac{(4-4)\times 1 + (3-5)\times(-1)}{1+1} = 5.67 + 1 = 6.67$$

Donc la note prédit de Hamdi pour l'item 1 est 6.67

- **La distribution de cet algorithme avec la logique MapReduce :**

On résume la distribution de cet algorithme avec les 5 étapes suivantes:

NB : Chaque étape représente un Job MapReduce.

Etape 1 : Transformation du fichier d'entrée (*Table 9*)

Etape 2 : Calcule des similarités entre utilisateurs

Etape 3 : Générer la liste des k-plus proche voisins pour chaque utilisateur

Etape 4 : Générer la liste des items non-noté pour chaque utilisateur

Etape 5 : Prédire la note de chaque item

Le résultat final de cet algorithme est représenté dans la (*Table 11*)

Conclusion

Dans ce chapitre, nous avons modélisé le système de recommandation de vol et hôtel tout en appliquant le formalisme du langage objet UML.et aussi nous avons modélisé l'architecture système pour la recommandation des offres en adéquation avec l'envie des utilisateurs en utilisant le filtrage collaboratif à base d'utilisateur et d'item.

Dans le chapitre suivant, nous présenterons l'environnement de développement, les différentes éléments de structures utilisés pour la réalisation de notre projet, ainsi que la description de l'application conçue ; d'une manière générale, tout en évoquant ses différentes interfaces.

Chapitre V

Implémentation

La dernière phase du projet, à savoir la réalisation de la solution conçue. En utilisant les outils appropriés conformément à notre conception, nous aboutissons à un système de recommandations pour les voyageurs.

Introduction

Cette partie présente le dernier volet de ce rapport. Elle a pour objet d'exposer le travail réalisé. D'abord, nous commençons par la présentation de l'environnement logiciel et technologie utilisé. Ensuite nous illustrons quelques aperçus d'écrans montrant les différentes fonctionnalités mises en place.

1. Environnement de développement

L'environnement de travail est un choix décisif pour l'implémentation d'une application, notre choix s'est porté sur plusieurs critères dont nous citerons :

- Hadoop, qui permet de traiter plusieurs TeraOctets de données.
- Mahout qui permet d'exécuter des algorithmes de machines learning sur des clusters Hadoop.
- Un serveur Tomcat ayant peu de failles connues.
- Une base de données Mysql qui permet l'interaction simple et efficace des données.
- Des pages web JSP basée sur Java qui simplifie le processus de développement de sites web dynamiques.

1.1.Ecosystème Hadoop

Hadoop est un Framework open source écrit en Java et géré par la fondation Apache. Il a été conçu pour réaliser des traitements de volumes de données en masse. Il fonctionne sur le principe des grilles de calcul consistant à répartir l'exécution d'un traitement intensif de données sur plusieurs nœuds ou grappes de serveurs.

Hadoop repose sur le système de fichiers HDFS et MapReduce pour distribuer et gérer les calculs.

HDFS(Hadoop Distributed File System) est utilisé pour stocker des fichiers de manière efficace dans le cluster. Quand un fichier est placé dans HDFS il est décomposé en blocs, 64 Mo de taille pour chaque bloc. Ils sont ensuite répliquées sur les différentsnœuds(DataNodes)dans le cluster. La valeur de réplication par défaut est 3, Soit 3 copies du même bloc dans le cluster comme l'illustre la figure 52 qui montre les différentes composantes d'un cluster Hadoop.

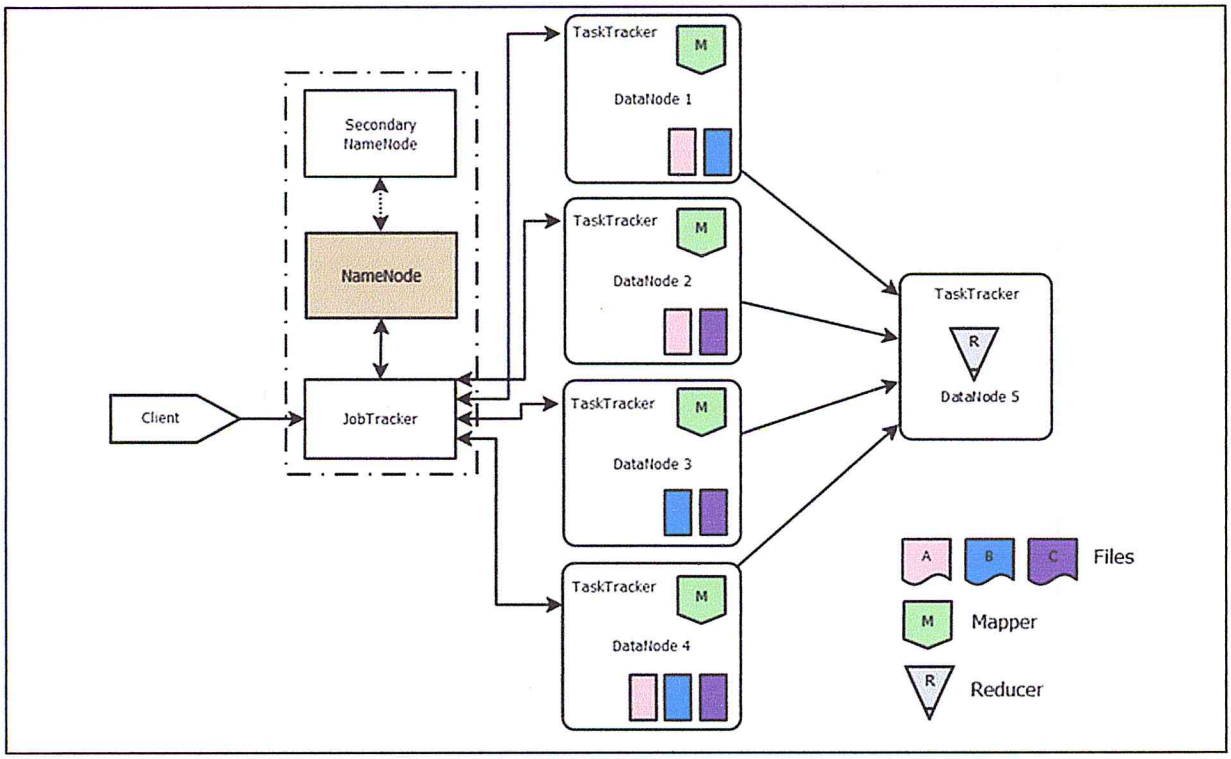

Figure 21 : Les composants d'un cluster Hadoop[52]

Le diagramme ci-dessus montre un cluster Hadoopà 6 Nœuds

NameNode (Master) - NameNode, NameNodesecondaire, JobTracker

DataNode 1 (esclave) - TaskTracker, DataNode

DataNode 2 (esclave) - TaskTracker, DataNode

DataNode 3 (esclave) - TaskTracker, DataNode

DataNode 4 (esclave) - TaskTracker, DataNode

DataNode 5 (esclave) - TaskTracker, DataNode

Hadoop suit une architecture maître-esclave. Comme mentionné précédemment, Un fichier dans HDFS est divisé en blocs et répliquée dans *Datanodes* dans un cluster Hadoop. Les trois fichiers A, B et C ont été répliqués trois fois à travers les différentes *Datanodes.*

NameNode : Le NameNode dans Hadoop est le nœud où Hadoop stocke toutes les informations de localisation des fichiers dans HDFS. En d'autres termes, il contient les métadonnées pour les HDFS. Chaque fois qu'un fichier est placé dans le cluster une entrée correspondante de l'emplacement, il est maintenu par le NameNode. Donc, pour les fichiers A, B et C, nous aurions quelque chose comme suit dans le NameNode:

Fichier A - DataNode1, DataNode2, DataNode4

Fichier B - DataNode1, DataNode3, DataNode4

Fichier C - DataNode2, DataNode3, DataNode4

SecondaryNameNode : Ce rôle intervient pour la redondance du NameNode. Normalement, il doit être assuré par une autre machine physique autre que le NameNode car il permet en cas de panne de ce dernier, d'assurer la continuité de fonctionnement du cluster.

DataNode : Le DataNode est chargé de stocker les fichiers dans HDFS. Il gère les blocs de fichiers dans le noeud. Il envoie les informations au NameNode sur les fichiers et les blocs stockés dans ce nœud et répond au NameNode pour toutes les opérations du système de fichiers.

JobTracker : JobTracker est chargé de prendre des demandes du client et attribuer les tâches à exécuter a TaskTrackers. Le JobTracker tente d'assigner des tâches à TaskTracker sur le *DataNode* où les données sont présents localement .Si cela est impossible, il essaye d'assigner des tâches à TaskTrackers où la réplique des données existe à travers les DataNodes. Cela garantit que le travail s'effectue même si un nœud échoue au sein du cluster.

TaskTracker : TaskTracker: ce rôle permet à un esclave d'exécuter une tâche MapReduce sur les données qu'elle héberge. Le TaskTracker est piloté par JobTracker d'une machine maître qui lui envoie la tâche à exécuter.
Maintenant que nous avons vu globalement l'architecture Hadoop. Dans la suite nous expliquons le principe de MapReduce.

1.1.1. HDFS

HDFS est le système de fichiers Java, permettant de gérer le stockage des données sur des machines d'une architecture Hadoop. Il s'appuie sur le système de fichier natif de l'OS (Unix) pour présenter un système de stockage unifié reposant sur un ensemble de disques et de systèmes de fichiers.
La consistance des données réside sur la redondance, une donnée est stockée sur au moins n volumes différents. Le NameNode rendrait le cluster inaccessible s'il venait à tomber en panne, il représente le SPOF (maillon faible) du cluster Hadoop. Actuellement, la version 2.0 introduit le Fail-over automatisé (capacité d'un équipement à basculer automatiquement vers un équipement alternatif, en cas de panne). Bien qu'il y ait plusieurs Name-Nodes, la promotion d'un Name-Node se fait manuellement sur la version 1.0.
Dans un cluster les données sont découpées et distribuées en blocks selon la taille unitaire de stockage (généralement 64 ou 128 Mo) et le facteur de réplication (nombre de copie d'une

donnée, qui est de 3 par défaut).Un principe important de HDFS est que les fichiers sont de type « write-one », ceci est lié au fait que lors des opérations analytiques, la lecture des données est beaucoup plus utilisée que l'écriture.

1.1.2. MapReduce

Mapreduce qui est le deuxième composant du noyau Hadoop permet d'effectuer des traitements distribués sur les nœuds du cluster. Il décompose un job (Unité de traitement mettant en œuvre un jeu de données en entrée. Un programme MapReduce (Packagé dans un JAR (Java Archive : fichier d'archive, utilisé pour distribuer un ensemble de classes Java)) et des éléments de configuration) en un ensemble de tâche plus petites qui vont produire chacune un sous ensemble du résultat final ; ce au moyen de la fonction **Map**.L'ensemble des résultats intermédiaires est traité (par agrégation, filtrage), ce au moyen de la fonction **Reduce**.

1.1.3. Apache mahout

Mahout est un projet de la fondation Apache visant à créer des implémentations d'algorithmes d'apprentissage automatique et de datamining :Clustering, Classification et Filtrage collaboratif.

Nous utilisant Mahout pour mettre en place notre moteur de recommandation,Le moteur de recommandation Mahout s'appuie sur un modèle de donnée qui s'appuie sur différentes données:

- Les utilisateurs ayant consulté ou acheté via notre site.
- Les Items qui correspondent aux différents produits hôtel et vol.
- Les avis sont les notations, des visualisations de page produit, des achats concrets…

Ces données sont stockées dans HDFS.

Pour réaliser ses recommandations, Apache Mahout s'appuie sur :

- Des algorithmes de similarité pour déterminer les utilisateurs les plus proches en utilisant par exemple la distance euclidienne, la corrélation de Pearson, la similarité cosinus…
- Des algorithmes de voisinage pour déterminer un ensemble d'utilisateurs proches selon la règle de similarité choisie. On distingue des algorithmes de type *Nearest* (les X utilisateurs les plus similaires) ou *Threshold* (tous les utilisateurs dépassant un certain seuil de similarité).

Il y'a deux type de filtrage collaboratif basé sur les utilisateurs, Il est également possible de mettre en place des moteurs de recommandation se basant uniquement sur les items et les similarités entre ceux-ci.

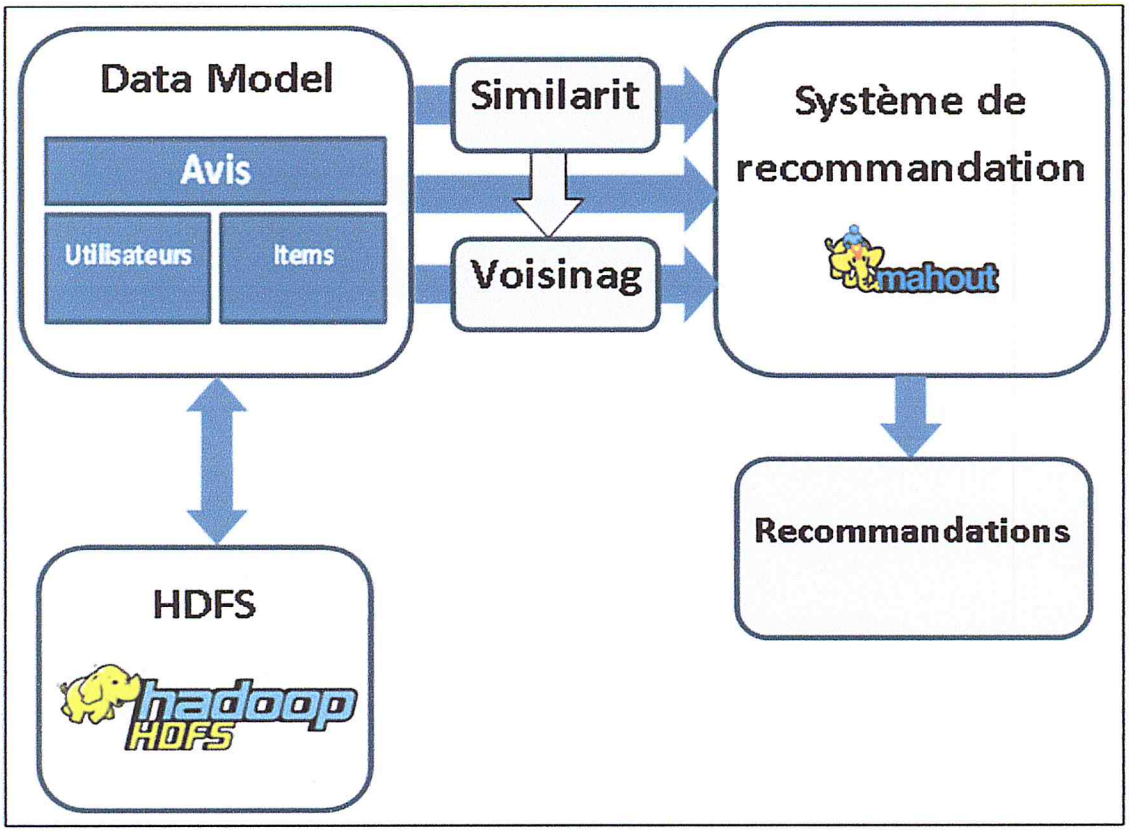

Figure 22 : Recommandation à base d'utilisateurs

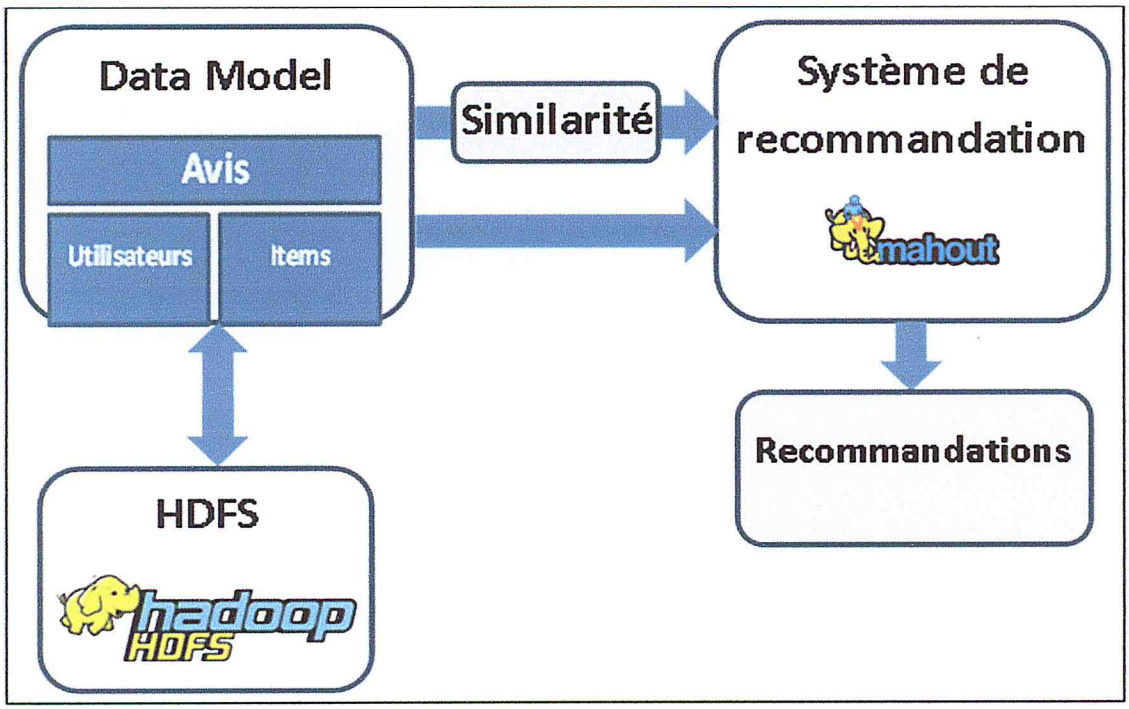

Figure 23 : Recommandation à base d'items

Idéalement, ces recommandations personnalisées peuvent être générées en temps réel. Il est à noter une fois recommandé, Le client va lui-même participé au processus d'enrichissement des données en fournissant ses propres évaluations.

2. Les technologies Web utilisées

2.1.Java Server Pages

Les JSP (*Java Server Pages*) sont une technologie Java qui permettent la génération de pages Web dynamiques. La technologie JSP permet de séparer la présentation sous forme de code XHTML et les traitements sous forme de classes Java. Les JSP permettent d'introduire du code Java dans des balises prédéfinies à l'intérieur d'une page XHTML. Ils mélangent la puissance de

Java côté serveur et la facilité de mise en page de XHTML côté client. Une JSP est habituellement constituée :

- de données et de balises XHTML;
- de balises JSP;
- de scriptlets (code Java intégré à la JSP).

Les JSP possèdent plusieurs avantages dont :

- l'utilisation de Java qui permet une indépendance de la plate-forme d'exécution mais aussi du serveur Web utilisé;
- la séparation des traitements de la présentation : la page Web peut être écrite par un designer et les balises JSP peuvent être ajoutées ensuite par le développeur. Les traitements peuvent être réalisés par des composants réutilisables (des Java beans, servlets);
- les JSP sont basées sur les servlets : tout ce qui est fait par une servlet pour la génération de pages dynamiques peut être fait avec une JSP.

Concrètement, au premier appel de la page JSP, le moteur de JSP génère et compile automatiquement une servlet qui permet la génération de la page Web. Le code XHTML est repris intégralement dans la servlet.

Dans le fonctionnement d'une application Web basée sur la technologie JSP, lorsqu'une requête demandant une page JSP est envoyée par un client http, le serveur Web http transmet la requête au moteur de JSP qui va l'interpréter puis compiler le code et générer la réponse sous forme d'une page XHTML statique. Donc pour exécuter une page JSP, il faut, en plus d'un serveur http comme Apache, un moteur de JSP comme Tomcat, Jetty ou GlassFish. Sur le serveur d'application il faut installer un *Java Development Kit* (JDK) qui contient une Machine Virtuelle Java (MVJ).

2.2.Serveur d'application Tomcat[34]

Le Serveur Tomcat est le moteur utilisé pour exécuter les pages JSP, Apache Tomcat est une implémentation open source d'un conteneur web qui permet donc d'exécuter des applications web reposant sur les technolgoies servlets et JSP.

Tomcat est diffusé en open source sous une licence Apache. C'est aussi l'implémentation de référence des spécifications servlets jusqu'à la vesion 2.4 et JSP jusqu'à la version 2.0 implémentées dans les différentes versions de Tomcat.

En tant qu'implémentation de référence de plusieurs versions des spécifications servlets/JSP, facile à mettre en œuvre et riche en fonctionnalités, Tomcat est quasi incontournable dans les environnements de développements. Les qualités de ses dernières versions lui permettent d'être fréquemment utilisé dans des environnements de production.

Depuis la version 4, Tomcat est composé de plusieurs éléments :

- ***Catalina*** : un conteneur servlet qui implémente les spécifications de Sun pour les servlets et les JSP;
- ***Coyote*** : un connecteur http qui écoute le trafic entrant, dirige les requêtes au moteur de Tomcat et renvoie la réponse au client;
- ***Jasper*** : un moteur JSP qui compile les fichiers JSP en tant que servlets et est capable de détecter les modifications des fichiers et de les recompiler à la volée.

3. Le Système de Gestion de Base de données

L'accès à une base de données par des utilisateurs ou des applications passe indirectement par un système connu par le système de gestion de base de données (SGBD). Le SGBD et la base de données forment le système de base de données.

Le SGBD est un ensemble de services (applications logicielles) permettant de gérer les bases de données, c'est-à-dire :

- Permettre l'accès aux données de façon simple.
- Autoriser un accès aux informations à de multiples utilisateurs.
- Manipuler les données présentes dans la base de données (insertion, suppression, modification).

Etant donné que notre choix s'est posé sur MySQL, voilà une petite description de ses caractéristiques.

MySQL

MySQL est un système de gestion de base de données (SGBD). Selon le type d'application, sa licence est libre ou propriétaire. Il fait partie des logiciels de gestion de base de données les plus utilisés au monde, autant par le grand public (applications web principalement) que par des professionnels, en concurrence avec Oracle et Microsoft SQL Server.

MySQL est un serveur de bases de données relationnelles SQL développé dans un souci de performances élevées en lecture, ce qui signifie qu'il est davantage orienté vers le service de données déjà en place que vers celui de mises à jour fréquentes et fortement sécurisées. Il est multi-thread et multi-utilisateur.

MySQL fait partie du quatuor LAMP: Linux, Apache, MySQL, PHP. Il appartient également à ses variantes WAMP (Windows) et MAMP(Mac).

4. Interface graphique

Nous allons essayer de présenter quelques pages web de notre application. Tout d'abord il y a plusieurs types d'utilisateurs de la plateforme, chacun avec des droits et des devoirs comme suite :

- Utilisateur de type visiteur si vous vous contentez de consulter notre plateforme sans vous enregistrer. Si vous n'êtes pas enregistré, vous ne disposez que d'un accès limité et aussi on ne peut pas vous recommander des offres.
- Un statut de Visiteur passe à celui « d'Abonné » dès l'enregistrement d'un compte d'utilisateur. Cet enregistrement est nécessaire pour consulter toutes nos offres et vous recommander des hôtels et des vols en adéquation avec le profil.

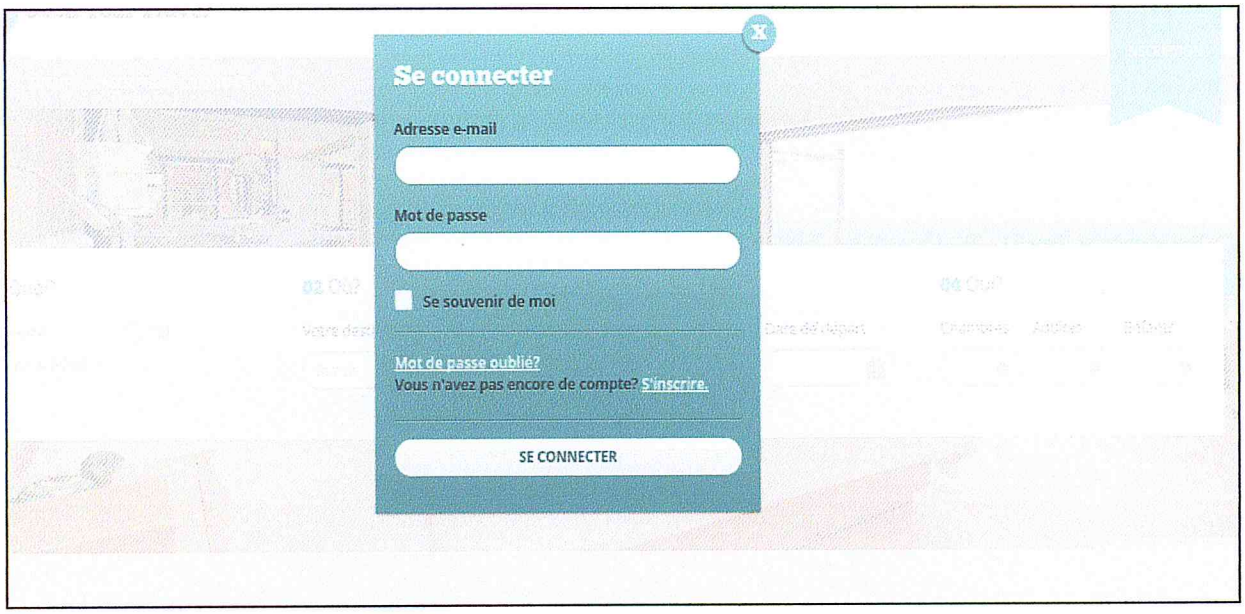

Figure 24 : page d'authentification

Une fois que l'abonné saisie son adresse email et son mot de passe, le système établie une vérification pour ensuite l'orienter vers son espace de travail

La page suivante est la page principale de l'abonné, elle émerge d'un menu en haut, qui lui permet de passer d'une page web à une autre (vol, hôtel ou bien contact).aussi un menu de recherche de vol ou bien d'hôtel et le bas de la page des recommandations vol et hôtel selon ses préférences.

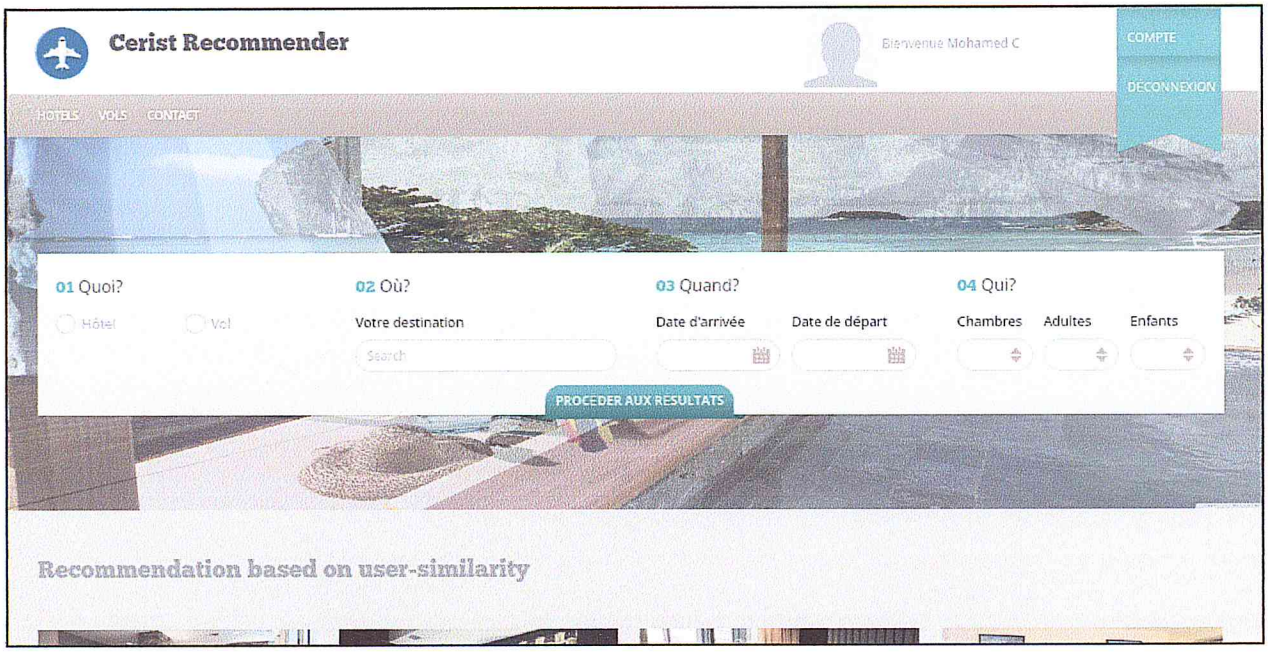

Figure 25 : Page accueille

Quand un client abonné effectue une recherche exemple sur un hôtel dont le nom est «Parc central » qui se trouve en Amérique, le système affiche l'hôtel recherché et aussi les hôtels similaires à « Parc central ».

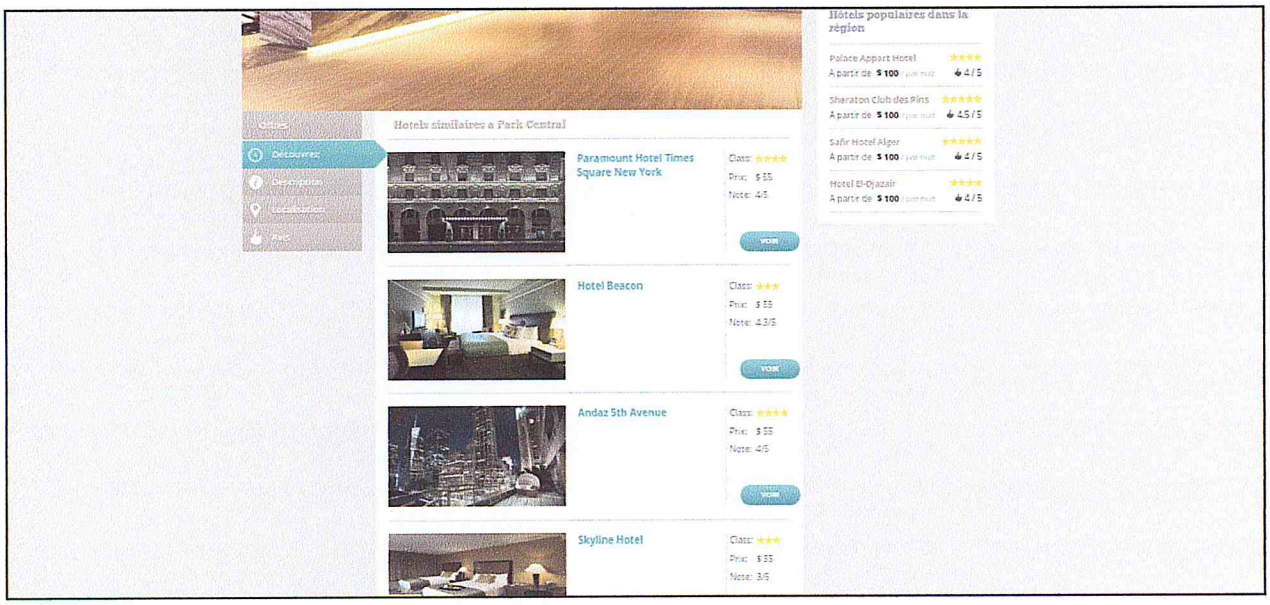

Figure 26:Offres similaire à un hôtel

L'utilisateur peut consulter les hôtels que nous publiant Ce dernier se voit alors proposer un ensemble d'offres affichées. Il peut voir de plus près l'hôtel, Exemple le nombre d'étoile, l'aménagement intérieur, les avis d'autres utilisateurs qui ont partagées leur expérience, aussi voir l'emplacement géographique avec Google Maps.

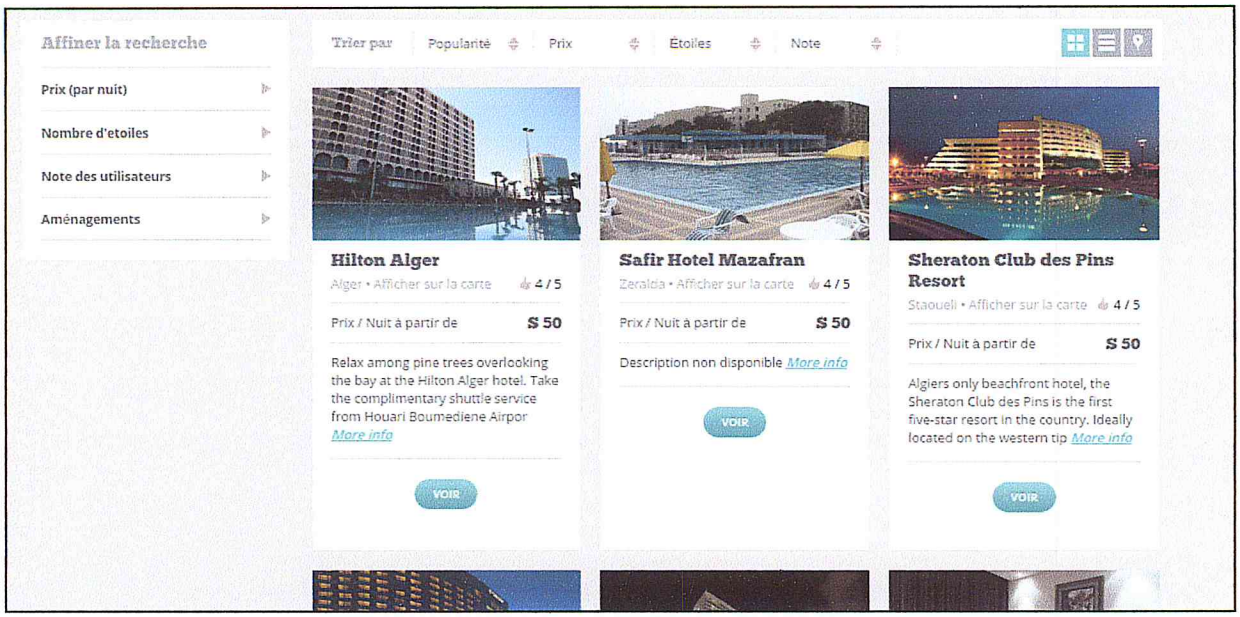

Figure 27: Consulter les offres hôtels

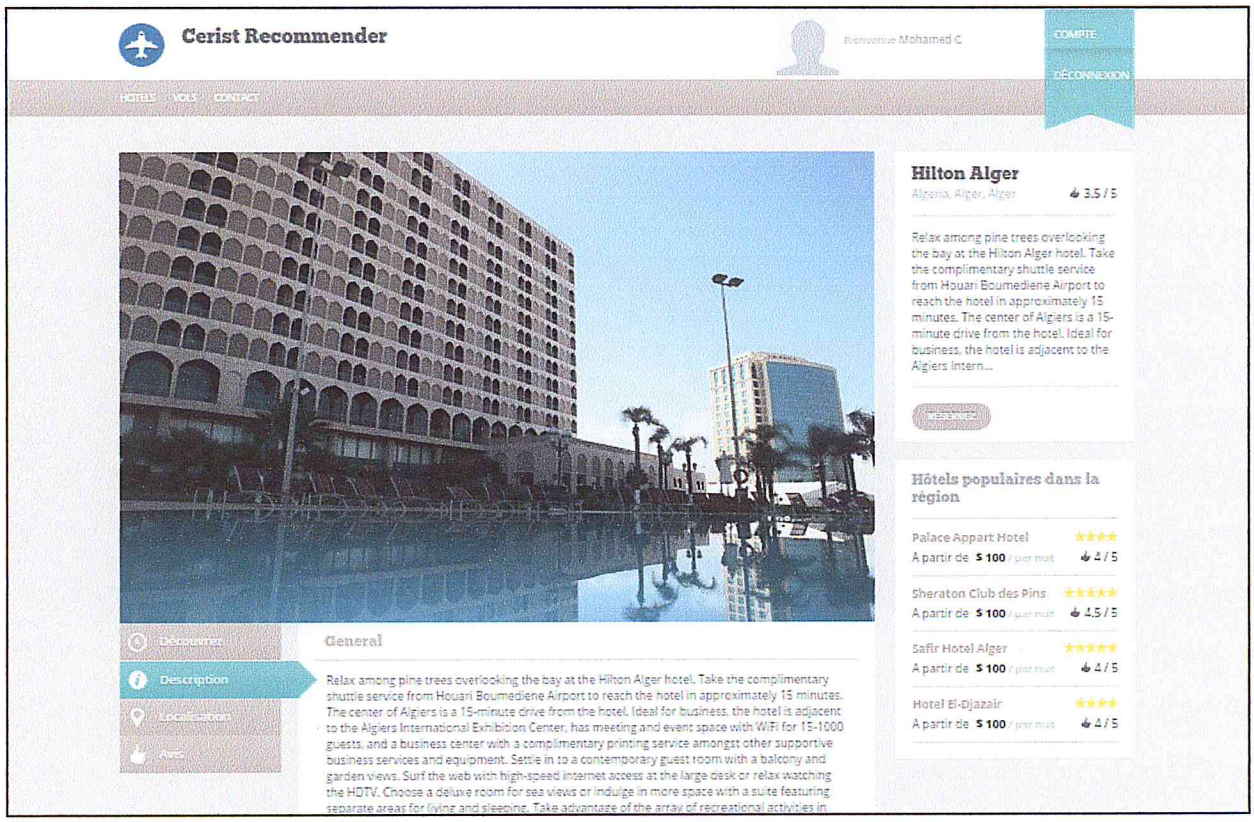

Figure 28:Hôtel avec plus de détails

Conclusion

Ce chapitre offre une description plus ou moins détaillée des interfaces que nous avons développées. Il étale un peu l'aspect réalisation de notre application. Toutefois, notre travail reste ouvert pour des extensions et des améliorations éventuelles.

Conclusion générale

Dans le cadre de cette thèse, nous nous sommes intéressés au système de recommandation, Le plus grand inconvénient que nous avons rencontré c'est que les utilisateurs subissent une surcharge informationnelle et des recommandations à tort et à travers liée à la multitude de ressources présentes dans notre système. C'est sur cet aspect particulier que nous nous focaliserons dans cette thèse : améliorer la qualité des recommandations en recommandant à l'utilisateur des offres en se basant sur ces appréciations et aussi sur l'utilisateur similaire à lui.

En premier plan nous avons parlé sur le Big Data car nous utilisons des données volumineuses provenant de notre système et aussi d'un site web externe afin de récolter les informations sur des vols, hôtels et aussi les avis des utilisateurs. C'est pour cela il faut faire une étude sur le Big Data pour comprendre l'utilité et faire sortir la valeur existante dans ces données.

Dans la deuxième partie nous avons fait une étude sur les différentes techniques des systèmes de recommandation, avec une comparaison entre les différentes approches pour les systèmes de recommandation. Nous avons choisi l'approche collaborative pour recommander aux voyageurs selon les appréciations des voyageurs similaire à lui.

Dans la troisième partie nous avons fait une étude sur les différentes techniques de l'analyse prédictif, avec une étude comparative. Nous avons choisi la méthode K plus proche voisin pour la méthode de prédiction.

Dans la quatrième partie nous avons conçu notre système pour les besoins des utilisateurs réservation vol, hôtel et aussi voir des avis d'autre utilisateurs et donner aussi ses propre avis. En explique par la suite la conception du système de recommandation basé sur utilisateur et item et les méthodes employé.

Dans la dernière partie nous présentons les outils qui nous ont aidés pour implémenter notre système exemple Apache Mahout et aussi l'écosystème Hadoop.

Parmi nos perspectives dans les versions prochaines :

- Création d'un moteur de recommandation autonome, accessible via une API, qui permettra de faire des recommandations pour d'autres systèmes.
- Utilisation des informations sociales présentes sur les réseaux sociaux publics tels que Facebook ou Twitter pour améliorer la recommandation.

Références bibliographique

[1] A. Bouza, G. Reif, A. Bernstein, and H. Gall. Semtree: ontology-based decision tree algorithm for recommender systems. In International Semantic Web Conference, 2008.

[2] A. Elouardighi.2009, « Support Cours & TD Datawarehouse », Consulté le 18 mars 2015.

[3] A. Naak, Un système de gestion et de recommandation d'articles de recherche. Papyres, Montréal, Canada, 2009.

[4] Abraham Gomez. 18 juin 2014 .En ligne sur le site de l'école de technologie supérieure université de Québec «substance.etsmtl.ca/hadoop-larchitecte-du-big-data » consulté le 04 décembre 2014.

[5] Alain Fernandez, Analyse prédictive et réseau de neurones « *www.piloter.org/business-intelligence/analyse-predictive.htm* », Consulté le 02-03-2015.

[6] Alain Fernandez, Le projet Data Warehouse, un processus continu, «www.*piloter.org/business-intelligence/projet-datawarehouse.htm* », Consulté le 12-02-2015.

[7] André Mayers,IFT 603 Techniques d'apprentissage, Thème 02, Arbre de décision, Partie I, 2012.

[8] B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages 144{152. ACM Press, 1992.

[9] B Kalmanet S. Kwasny, Why tanh : Choosing a Sigmoidal Function. In Neural Networks, IJCNN, International Joint Conference on, volume 4, pages 578–581, 1992.

[10] Belgacem Amar, Classification des signaux EGC avec un système-multi-agent neuronale, UNIVERSITE ABOU BAKR BELKAID-TLEMCEN, 2012.

[11] Bertrand Bathelot,Définition Analyse prédictive , Définitions marketing , lundi 29 décembre 2014 « *www.definitions-marketing.com/Definition-Analyse-predictive* », Consulté le 10-01-2015.

[12] B.Liaudet ,Modelisation Presentation Generale, 2008, « *bliaudet.free.fr/IMG/pdf/Cours_de_data_mining_3-Modelisation-EPF.pdf* », Consulté le 8-03-2015.

[13] Catherine Berrut et Nathalie Denos, Filtrage collaboratif, Assistance intelligente à la recherche d'informations, Chapitre 8, Hermes - Lavoisier, 2003, « *http://mrim.imag.fr/publications/2003/CB001/berrut03b.pdf* » Consulté le 23-01-2015.

[14] Claude Bernard, Le phénomène des big data touche toutes les entreprises, 3 juillet 2012, «*www.institut-sage.com/2012/07/le-phenomene-des-big-data-touche-toutes-les-entreprises*»Consulté le 23-01-2015.

[15] Cortes C. et Vapnik V. Support vector networks. Machine Learning, 20: 273–297,1995.

[16] D.Goldberg, D.Nichols, B.M.Oki, D.Terry, « Using collaborative filtering to weave an information Tapestry », Communications of the ACM, vol. 35, n° 12, p. 61- 70, Décembre 1992.

[17] D. J. Hand et K. Yu, Idiot's bayes - not so stupid after all? International StatisticalReview 69(3), 385–398, 2001.

[18] D.Maltz, K.Ehrlich, Pointing the way: active collaborative filtering, Proceedings of CHI'95 : Conference on Human Factors in Computing Systems, p. 7-11, Mai 1995.

[19] D.Miller, J.L.Maltz, L.R Herlocker, A.Gordan, J.A Riedl., B.N.Konstan, GroupLens: applying collaborative filtering to Usenet News , Communications of the ACM, vol. 40, n° 3, p. 77-87, Mars 1997.

[20] David Dubois, Damien Migout, Modèles prédictifs : quelles utilisations pour améliorer les processus de souscription en assurance de personnes ?, Bruxelles, 29 avril 2014.

[21] Faicel CHAMROUKHI, Classification supervisée : Les K-plus proches voisins, Université de Toulon, 2013-2014.

[22] Florian Francheteau, Rapport de stage Étude des ETL Open Source, 2007-2008, Consulté le 16.av.2015.

[23] Frank Meyer. Systèmes de recommandation dans des contextes industriels, Université de Grenoble, 2012.

[24] Fraud R, Clérot R, A methodology to explain neural network classification. Neural Networks 15(1), 2002.

[25] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6):734–749, 2005.

[26] G.Linden, B.SmithetJ.York, Amazon.com recommendations: Item-to-item collaborative filtering, IEEE Internet computing, 7(1), 7680, 2003.

[27] Gavin Blackett, Analytics Network - O.R. & Analytics, « *www.theorsociety.com/Pages/SpecialInterest/AnalyticsNetwork_analytics.aspx*» , Consulté le 02-03-2015.

[28] Geoffray Bonnin, Vers des systèmes de recommandation robustes pour la navigation Web : inspiration de la modélisation statistique du langage, Université Nancy 2, Avril 2011, p 9-10

[29] Gilles Valiquette, Analyse des données quantitatives, 25-01-2010, «*ciqss.umontreal.ca/docs/formations/ateliers/2010-01-25_adq-sas.pdf* »,Consulté le 15-03-2015.

[30] Greg Linden, Brent Smith, Jeremy York, Amazon.com Recommendations: Item-to-Item Collaborative Filtering, IEEE Internet Computing, v.7 n.1, p. 76-80, January 2003

[31] H. Kang and S. Yoo. Svm and collaborative filtering-based prediction of user preference for digital fashion recommendation systems. IEICE Transactions on Inf&Syst, 2007.

[32] H. CHERIF Ikram, Classification des tracés CardioTocoGraphiques (CTG) d'un fœtus À l'aide de classifieurs multiples, Université Aboubaker Belkaid Tlemcen, 2011.

[33] Hetal Bhavsar, Amit Ganatra, A Comparative Study of Training Algorithms for Supervised Machine Learning, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-4, September 2012.

[34] Jean-Michel doudoux, Développons en Java,Apache Tomcat, 19/05/2014, « *jmdoudoux.developpez.com/cours/developpons/java/chap-tomcat.php* », Consulté le 25 mai 2014.

[35] Jeff Kelly, 12 février 2014, « Big Data Vendor Revenue and Market Forecast 2013-2017 », « *wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017*» Consulté le 5 décembre 2014.

[36] Jen Underwood, Beginning Prescriptive Analytics with Optimization Modeling, b-eye-network.com, 19Novembre 2013.

[37] K. Hornik, Some New Results on Neural Network Approximation, Neural Networks, 6(8): 1069–1072, 1993.

[38] K. Miyahara and M. J. Pazzani. Collaborative filtering with the simple bayesian classifier. In Pacific Rim International Conference on Artificial Intelligence, 2000.

[39] K. Yu, V. Tresp, and S. Yu. A nonparametric hierarchical bayesian framework for information filtering. In SIGIR '04, 2004.

[40] KadousDjamila, Utilisation des réseaux de neurones comme outil du datamining,Universite Abou-BakrBelkaidde Tlemcen, 2012.

[41] M.Balabanovic, Y.Shoham, Fab: content-based, collaborative recommendation, Communications of the ACM, vol. 40, n° 3, p. 66-72, mars 1997.

[42] M.Surendra P.Babu, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2(3), 2011, 1283-1286

[43] NEGRE Elsa, Les systèmes de recommandation, Université Paris-Dauphine, 2012, «*irit.fr/recherches/ADRIA/Documents/Fargier/documentsBR4CP/E_Negre_Lens_juin_12.pps* », Consulté le 03.Février.2015.

[44] Octavian Rolland Arnautu, Système de recommandation de musique, Université de Montréal, 2012, p 28-30.

[45] Open Source, 2011, Hadoop, En ligne « open-source-guide.com/Solutions/Developpement-et-couches-intermediaires/Big-data/Hadoop » consulté le 05 décembre 2014.

[46] P.Resnick, N.Iacovou, M.Suchak, P. Bergstrom etJ.Riedl, GroupLens: An Open Architecture for Collaborative Filtering of Netnews. pp. 175–186, ACM Press, 1994.

[47] Pazzani M, Billsus D, Learning and revising user profiles: The identification of interesting web sites. In Machine Learning, 27, 1997.

[48] Peis Eduardo, Del Castillo JM Morales, et Delgado-López J. A., Semantic recommender systems analysis of the state of the topic. Hipertext.net, 2008, vol. 6, p. 1- 5.

[49] Philippe Roux.15 janvier 2014. Groupe de travail Big Data.Contribution éditoriale, Consulté le 11 decembre 2014.

[50] R. Burke, Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction, 12(4):331370, 2002.

[51] R. Ghani, A. Fano. Building recommender systems using a knowledge base of product semantics. In In 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, 2002.

[52] R.Menon, Introducing Hadoop, 4-01-2013, « *rohitmenon.com/index.php/introducing-hadoop-part-ii* »Consulté le 23.Mai.2015.

[53] Romain Guigourès, Marc Boullé, Optimisation directe des poids de modèles dans un prédicteur Bayésien naïf moyenné, « *www.marc-boulle.fr/publications/GuigouresEtAlEGC11.pdf* »,Consulté le 16.janvier.2015.

[54] S.chaei, C.Desrosiers.MTI820 Entrepôts de données et intelligence d'affaires Intégration des donnéesetETL, 2011.

[55] Sagar S. Nikam, A Comparative Study of Classification Techniques in Data Mining Algorithms, Oriental Journal of Computer Science and Technology, Volume 8, Number 1, Avril 2015.

[56] Séraphin LOHAMBA OMATOKO, Analyse et détection de l'attrition dans une entreprise de télécommunication, Université Notre Dame du Kasayi , 2011.

[57] SynoxGroup,Les technologies analytiques et big data , Mardi 09 septembre 2014 « *www.synox-group.com/sites/all/newsletter/10/Article_AnalyticBigData.pdf* »,Consulté le 13.Mars.2015.

[58] T. Med Ilyas, RAHALI Youssouf, Une application médicale de recommandation contextuelle des documents, Université Abou BakrBelkaid– Tlemcen, 2014.

[59] Tong Zhang et Vijay S. Iyengar, Recommender Systems Using Linear Classifiers, Journal of Machine Learning Research 2,313-334, 2002.

[60] Thierry Outrebon,Tableau Software démocratise l'analyse des données, 2014, « *www.informatiquenews.fr/tableau-software-democratise-lanalyse-donnees-27326* »,Consulté le 13.Mars.2015.

[61] V. Pronk, W. Verhaegh, A. Proidl, and M. Tiemann. Incorporating user control into recommender systems based on naive bayesian classification. In RecSys '07: Proceedings of the 2007 ACM conference on Recommender systems, pages 73–80, 2007.

[62] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[63] Vincent de Stoecklin,Business Analytics : Définition, Enjeux, Applications, « *www.data-business.fr/business-analytics*», Consulté le 05-03-2015

[64] Xavier Amatriain, Alejandro Jaimes, Nuria Oliver, and Josep M. Pujol , Data Mining Methods for Recommender Systems, Recommender Systems Handbook, 2010. « *xavier.amatriain.net/pubs/RecsysHandbookChapter.pdf*»

[65] Y. Cho, J. Kim, S. Kim. A personalized recommender system based on web usage mining and decision tree induction. Expert Systems with Applications, (23), 2002.

[66] Y. Zhang, J. Koren. Efficient bayesian hierarchical user modeling for recommendation system. In SIGIR 07, 2007.

[67] Y. Hanane Z. Mokhtar, « Algorithmes d'apprentissage pour la classification de documents », Mémoire de licence en ligne, Université de Mostaganem - Algérie, 2009.

[68] Z. Xia, Y. Dong, and G. Xing. Support vector machines for collaborative filtering. In ACM-SE 44: Proceedings of the 44th annual Southeast regional conference, pages 169–174, New York,NY, USA, 2006. ACM.

[69] Zouag Rafika, Khial Nabila, Filtrage collaboratif des objets pédagogiques, Université Tlemcen, 2013, p 9 – 14.

Résumé

L'industrie du voyage est face à un tournant majeur grâce aux grands volumes d'informations générés dans chaque transaction et chaque réservation. Les expériences de voyages sont de plus en plus discutées et partagées sur les réseaux sociaux, Ces raisons ont poussé les acteurs du voyage et du tourisme à changer leur façon de travail et leur manière de gérer leurs activités. Les compagnies de voyages peuvent prendre de meilleures décisions en utilisant les données collectées des clients afin de personnaliser leurs services. Ils utilisent des questionnaires sur les sites internet en demandant aux clients d'évaluer chaque service fournis, ce qui permet d'améliorer leurs expériences et de les rendre bien plus agréables en répondant à leurs besoins.

Il nous a été confié, dans le cadre de ce projet de fin d'étude, le développement d'un système de recommandation des besoins des voyageurs dans un contexte Big Data. Ce système a pour objectif de recommander a des utilisateurs des offres (vol et hôtel) sélectionnés parmi un large choix, et censés être appréciés par eux. Notre système tente de prédire si un utilisateur donné appréciera ou non un item. Pour parvenir à un tel but, un système de recommandation a besoin d'accumuler des données sur les utilisateurs et les offres disponibles. Il sauvegarde les traces de ses utilisateurs, avant d'appliquer des méthodes statistiques pour prédire leur comportement futur.

Mots-clés : Les systèmes de recommandation, filtrage collaboratif, Big Data

Abstract

Travel industry faces a major turning point thanks to big data generated from every transaction and reservation. Travel experiences are being more and more shared and discussed on social networks, This encourages travel and tourism actors to improve the way they work and manage their business. Travel companies can make better decisions based on the data aggregated from customers to personalize their services. They use questionnaires in their websites that asks customers to rate every service provided, which could enhance the business and make customer's travel a lot more pleasant than it is.

In this final year project, we have been asked to develop a travel recommendation system using Big Data. This system aims to suggest travel offerings (Airlines, Hotels) that match the customer's preferences. It tries to predict regarding whether a person prefers or not a service. In order to accomplish that, our recommendation system has to collect data about customers and available offerings. It saves users traces, and apply some statistic methods to determine offerings they may want in the future.

Keywords: Recommendation systems, collaborative filtering, Big Data

ملخص:

قطاع السياحة و السفر يشهد نقلة نوعية بفضل البيانات الضخمة المستخلصة من كل عملية تسجيل أو حجز. تجارب السفر يتم مشاركتها و مناقشتها بشكل متزايد عبر شبكات الانترنت. مما شجع الفاعلين في ميدان السفر و السياحة على تطوير طريقة عملهم و تسييرهم لنشاطاتهم. شركات السفر تستطيع اصدار قرارات أفضل لتحسين خدماتها بفضل المعلومات المجمعة . هذه الشركات تقوم بعمل استبيانات على مواقعها الالكترونية و تطلب من الزبائن تقييم الخدمات الموفرة لهم. كل هذا من شأنه تحسين القطاع و جعل تجربة الزبائن أكثر متعة مما كانت عليه في السابق من خلال الاستجابة لمتطلباتهم.

في مشروع تخرجنا، طُلب منا تطوير نظام توصية لحاجيات المسافرين باستعمال البيانات الضخمة. هذا النظام يهدف الى اقتراح عروض سفر (فنادق, شركات طيران) تتماشى مع رغبات الزبائن, و يحاول التنبؤ ما اذا كان الزبون يفضل خدمة ما أو لا. من أجل تحقيق هذا، يقوم نظامنا بتجميع المعلومات حول الزبائن و العروض الموفرة. يسجل و يتتبع خطى المستخدمين و يقوم بتطبيق أساليب احصائية لتحديد سلوكهم و العروض التي من الممكن ان ترضي رغباتهم مستقبلاً.

الكلمات المفتاحية : نظم التوصية، التصفية التعاونية، بيانات ضخمة.