Des textes communautaires à la recommandation

Damien Poirier

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE SCIENCES ET TECHNOLOGIE

LABORATOIRE D'INFORMATIQUE FONDAMENTALE D'ORLEANS

THÈSE présentée par :

Damien POIRIER

soutenue le : **11 février 2011**

pour obtenir le grade de : **Docteur de l'université d'Orléans**

Discipline/ Spécialité : Informatique

Des textes communautaires à la recommandation

THÈSE dirigée par :

Isabelle TELLIER	Professeur, Université d'Orléans
Patrick GALLINARI	Professeur, Université Pierre et Marie Curie

RAPPORTEURS :

Mohand BOUGHANEM	Professeur, Université Paul Sabatier
Patrice BELLOT	Maître de conférence, Université d'Avignon

JURY :

Brigitte GRAU	Professeur, ENSIIE, Présidente du jury
Françoise FESSANT	Ingénieur R&D, Orange Labs
Isabelle TELLIER	Professeur, Université d'Orléans
Patrick GALLINARI	Professeur, Université Pierre et Marie Curie
Patrice BELLOT	Maître de conférence, Université d'Avignon
Mohand BOUGHANEM	Professeur, Université Paul Sabatier
Cécile BOTHOREL	Ingénieur R&D, Telecom Bretagne
Nathalie DENOS	Maître de conférence, Université Pierre Mendès France

Remerciements

Je tiens tout d'abord à remercier tous les membres de mon jury. Je remercie Mohand Boughanem et Patrice Bellot pour m'avoir fait l'honneur d'être rapporteurs de cette thèse. Un gand merci également à Brigitte Grau pour avoir accepté de présider ce jury ainsi qu'à Nathalie Denos pour avoir été une des examinatrices. À vous quatre, merci également pour l'intérêt que vous avez porté à mes travaux et pour les retours souvent positifs (et agréables à entendre) concernant ce manuscrit.

Je remercie plus particulièrement les autres membres du jury qui ont tous participé à l'encadrement de cette thèse. Merci donc à Cécile Bothorel qui m'a proposé un sujet (qui a légèrement bifurqué en cours de route comme on peut s'en douter) et qui m'a permis de me lancer dans l'aventure ! Cécile n'ayant pas pu encadrer cette thèse jusqu'au bout pour cause de départ de chez Orange, je remercie également Françoise Fessant pour la prise de relais ainsi que pour sa forte implication dans cet encadrement, notamment lorsque les conseils et encouragements ont été nécessaires dans la dernière ligne droite. Je tiens également à remercier fortement Isabelle Tellier et Patrick Gallinari pour avoir accepté de diriger cette thèse et pour leur disponibilité et leurs conseils avisés.

Je tiens également à remercier tous les gens que j'ai pu croiser durant ces trois années, que ce soit à Orange Labs, au LIFO ou au LIP6. Je remercie tout particulièrement Bruno Guerraz, Romain Trinquart et Marc Boullé qui m'ont suivi du début à la fin et qui m'ont appris énormément. Merci également aux doctorants et post-doctorants d'Orange Labs Lannion que j'ai pu croiser et avec qui j'ai pu échanger lors des pauses café, à la Sodexo ou dans les différentes soirées lannionnaises.

Je tiens également à remercier de tout cœur tous mes amis, sans qui je n'aurais jamais réussi, en commençant par les Gerboisiens Buendon, Bobo, Boyan, Jérem (Docteur :D), Carole, Hélène, Hélène, July, Titi, Vianney, Émilie et Kiki, la team de l'impasse Berthelot Romain, Aurélie, Ben, Tom, Guigui, Vivi et Dom, ainsi que tous les autres Anaïs, Élé, Artem, David, Dédé, Chrousse, Aurel, Philippe, Stephan, Camille, Meriem, Lizzy, Tacha, Nico, Alex, Arianna, etc.

Pour finir, je remercie bien entendu toute ma famille, mes frères et sœur, et plus spécialement mes parents qui m'ont toujours fait confiance et soutenu au maximum. Ce diplôme, c'est à vous que je le dois.

Table des matières

Chapitre 1

Introduction

Sommaire

1.1 Contexte : les enjeux du Web 2.0

Durant les années 2000, l'Internet a subi une énorme transformation. Après l'explosion de la bulle Internet en 2000, et la prise de conscience du potentiel de cette technologie, les services se sont développés, devenant de plus en plus nombreux, mais surtout de plus en plus interactifs. L'Internaute d'aujourd'hui n'est plus simplement spectateur, il peut s'il le souhaite devenir acteur. Et tout est mis en œuvre pour que le simple spectateur devienne acteur du Web afin, entre autres choses, de le fidéliser. Cette révolution a donné naissance à ce que l'on nomme aujourd'hui le Web 2.0 [O'R05] appelé encore Web Social ou Web Participatif.

Bien que la définition du Web 2.0 ne soit pas encore parfaitement établie, deux grands aspects caractérisant ce Web nouvelle génération peuvent être mis en avant : le rôle central de l'utilisateur et la personnalisation.

1.1.1 L'utilisateur au centre des communications

En premier lieu, une caractéristique principale du Web 2.0 qui le distingue grandement du Web 1.0 est la prise de contrôle de l'information par les utilisateurs. N'importe quel internaute peut aujourd'hui apporter sa pierre à l'édifice. Il peut se faire une place sur la toile, collaborer, partager des informations, des outils, des fichiers multimédias, donner ses opinions, commenter, réagir, etc. et tout ceci sans connaissances spécifiques. En effet, quand auparavant il fallait un minimum de savoir faire en informatique et en programmation pour créer son espace sur le Net, aujourd'hui il suffit de savoir cliquer car de plus en plus d'outils sont mis à disposition de tout un chacun afin de faciliter toutes ces interactions. Parmi ces outils, nous pouvons bien entendu citer les réseaux sociaux, les blogs, les wikis, les boîtes à réactions, les sites de partage de vidéos, de photos, de musiques, etc. La grande majorité des sites présents sur le Net aujourd'hui offrent la possibilité à tous leurs visiteurs de laisser, au minimum, une trace textuelle et ainsi s'exprimer publiquement. Tout ce contenu, qu'il soit textuel ou autre, est appelé Contenu Généré par les Utilisateurs ou UGC (pour User Generated Content). Il représente une quantité de données de plus en plus importante sur la toile et est composé, en très grande partie, de données textuelles.

Une anecdote qui peut démontrer l'importance de la prise de contrôle du Web par les utilisateurs est que les internautes ont été nommés personnalité de l'année en 2006 par le *Time Magazine*[1] [Gro06]. Ce nouvel espace d'expression représente une grosse quantité d'informations, notamment en termes d'avis et d'opinions, susceptibles d'être exploitées à des fins diverses. Les données textuelles, notamment, peuvent être analysées dans différents buts. Par exemple, dans le domaine de la fouille d'opinion (Opinion Mining), les textes sont utilisés afin de permettre à des entreprises de connaître automatiquement l'image que les consommateurs ont d'eux (comme le propose *Nielson BuzzMetrics*[2]), de même pour les projets et les personnalités politiques, ou encore pour faire de la comparaison d'articles

1. Magazine d'information hebdomadaire américain
2. en-us.nielsen.com/tab/product_families/nielsen_buzzmetrics

de vente (comme sur le site *Vozavi*[3]), réaliser des sondages, détecter des rumeurs, etc. En effet, les textes rédigés par les internautes sont en général beaucoup plus subjectifs que les articles rédigés par des professionnels et donc beaucoup plus porteurs d'opinion. De plus, ils contiennent un éventail d'avis et de jugements souvent plus représentatifs du « consommateur lambda », étant rédigés par un panel d'individus varié parmi lesquels on retrouve tout autant de profanes que de connaisseurs du sujet abordé. Pour ce qui est du cas des films par exemple, il est possible de trouver des avis de spectateurs de tous âges et de tous horizons, ce que n'offrent pas les critiques journalistiques qui sont généralement rédigés par des personnes du même « milieu », soit dans le cas présent des journalistes cinéphiles appartenant à la population active et ayant fait un certain nombre d'années d'études.

1.1.2 Un Internet par utilisateur

Cet effet d'appropriation du Web par les internautes a évidemment des conséquences, notamment en terme de quantités de données disponibles en ligne. En offrant la possibilité de participer au développement du Web, et surtout en encourageant tout un chacun à le faire, la quantité de données présentes sur la toile s'en est trouvée multipliée et continue de l'être jour après jour. L'un des grands défis d'aujourd'hui concernant les technologies du Web est donc de proposer des solutions permettant de parcourir cette masse toujours grandissante de données et de contenus afin de trouver ce que l'on cherche le plus aisément et rapidement possible. Les domaines de recherche dédiés à ces problématiques sont la Recherche d'Information et le Filtrage d'Information. L'une des voies existantes, qui est apparue bien avant l'arrivée du Web 2.0, est le moteur de recherche tel que *Google*[4] qui en est l'exemple le plus célèbre. Une autre voie, qui elle s'est considérablement développée avec l'arrivée du Web 2.0, est la personnalisation. Par personnalisation, on entend ici l'adaptation des pages Web pour un utilisateur en particulier. Le but de la personnalisation est de moduler le Web afin d'aider les internautes à accéder, le plus simplement et rapidement possible, aux ressources qu'ils désirent. Cette adaptation peut être faite manuellement par l'utilisateur, comme sur le site *Netvibes*[5] par exemple. Ce site est un portail Web individuel et personnalisable permettant d'agréger une partie du contenu en provenance d'autres sites Web (flux RSS). Chaque utilisateur peut ainsi organiser sa page, en déplaçant, ajoutant, supprimant ces contenus et ainsi se simplifier l'accès vers les données qui l'intéressent le plus. La personnalisation peut également être automatisée. Les *cookies* (témoins) par exemple peuvent être utilisés en tant qu'outils pour la personnalisation afin de guider l'utilisateur suivant ses actions antérieures. Les systèmes de recommandation sont également une des solutions à cette personnalisation automatisée. Ils sont devenus, à l'instar des moteurs de recherche, un outil incontournable pour tout site Web focalisé sur un certain type d'articles disponibles dans un catalogue riche, que ces articles soient des objets, des produits culturels (livres, films, morceaux de musique, etc.), des éléments d'information (news) ou encore simplement des pages (liens hypertextes). L'objectif de

3. www.vozavi.com/
4. www.google.com
5. www.netvibes.com

ces systèmes est de sélectionner, dans un catalogue, les articles (ou items) les plus susceptibles d'intéresser un utilisateur particulier. Tandis que les moteurs de recherche ont un rôle générique et répondent à des requêtes, les moteurs de recommandation ont un rôle plus spécifique et personnalisent leurs réponses en fonction de l'utilisateur. Nageswara et Talwar [NRT08] ont répertorié un vaste ensemble de systèmes de recommandation pour différents domaines applicatifs, dans des contextes académiques et industriels. Différentes approches peuvent être mises en œuvre afin de faire de la recommandation personnalisée, mais toutes ont la particularité de requérir un minimum de données de départ sur lesquelles les algorithmes vont pouvoir s'appuyer.

Les enjeux de la personnalisation peuvent également s'étendre au delà du Web. On peut par exemple citer les services de VOD (Vidéo à la Demande) qui se développent via la télévision numérique et qui proposent une grande quantité de films, séries, documentaires, etc., les applications pour smartphones qui sont de plus en plus nombreuses, les documents disponibles pour les livres numériques, etc.

1.2 Problématique défendue et contributions

Nous avons donc d'un côté des textes riches d'informations en termes d'avis et d'opinions non encore exploitées et d'un autre côté, nous avons des outils nécessaires à l'organisation du Web et dont l'efficacité dépend d'une grande quantité d'informations. C'est dans ce contexte que se placent les travaux de cette thèse. La problématique réside dans l'exploitation des textes subjectifs produits par les internautes sur les sites communautaires dans le but de proposer de nouvelles perspectives à la recommandation personnalisée. L'idée dominante est de mettre à profit les textes en question afin de limiter les problèmes liés au manque d'informations. La grande majorité des systèmes de recommandation existants fonctionne uniquement sur des données « internes » au système. Le site Amazon [6] par exemple, qui est un site de vente par correspondance, construit ses recommandations uniquement à l'aide des informations apprises sur ses propres clients. Si l'on souhaite, dans le cas d'un service débutant qui possède très peu de clients et qui a des connaissances réduites sur ceux-ci, mettre à profit les mêmes outils que sur Amazon, il est alors nécessaire d'acquérir des informations complémentaires sur le contenu du catalogue. L'Internet étant devenu plus que jamais une source de connaissances, l'exploitation de la richesse de l'Internet *ouvert* au service d'un site web *fermé* apparaît naturellement comme une voie de recherche nouvelle et incontournable.

L'objectif principal de cette thèse est d'exploiter les données textuelles produites par les utilisateurs sur le Web pour alimenter un système de recommandation débutant et en manque d'informations. L'exploitation des textes provenant du contenu généré par les utilisateurs a déjà été étudiée dans la littérature. Shani et al. [SCM08] ont établi des recommandations basées sur les listes de films favoris rédigées par les internautes sur leur

6. www.amazon.com

blog Myspace [7]. En appliquant des méthodes de clustering basées sur les co-occurences des films dans les listes, ils sont alors capables d'établir des recommandations. La grande majorité des autres travaux répertoriés sur le sujet exploitent un moteur de recommandation *thématique*, dans lequel les articles à recommander sont décrits par un ensemble d'*attributs*. Prendre comme attributs les mots de textes qui parlent de ces articles est alors effectivement une solution possible. Mais les moteurs de recommandation les plus efficaces procèdent plutôt par *filtrage collaboratif*, une méthode qui se fonde non pas sur des attributs mais sur des *notes* attribuées par des utilisateurs à des articles.

La fouille d'opinion est un sous-domaine de la fouille de textes qui consiste à analyser des textes afin d'en extraire des informations liées aux opinions et sentiments. L'une des tâches de la fouille d'opinion, appelée classification d'opinion, a pour objectif de classer les textes suivant l'opinion qu'ils expriment. Cette classification peut se faire sur deux classes (positif ou négatif), sur trois classes (positif, négatif ou neutre) ou sur plus de classes encore. Ces classes sont ordonnées et peuvent donc être assimilées à des notes, données nécessaires à la recommandation par filtrage collaboratif. Enchaîner un système d'affectation de notes par fouille d'opinion et un système de recommandation par filtrage collaboratif sur des données textuelles réelles est une des principales contributions de cette thèse. En effet, l'idée de faire de la recommandation en combinant la classification d'opinion et une méthode de filtrage collaboratif a déjà été proposée dans la littérature [CSC06, DWW07] mais, à notre connaissance, cela est resté à l'état d'intuition.

Dans cette thèse, une chaîne complète de traitements est mise en œuvre en allant de l'acquisition des textes sur un site communautaire, à leur mise en forme pour les deux grands types de moteurs de recommandation (filtrage thématique et collaboratif) et jusqu'à l'évaluation de ces données dans un système de recommandation. Chacun des éléments stratégiques de la chaîne est choisi parmi les techniques existantes et argumenté afin de le positionner et de l'évaluer dans le contexte applicatif de la recommandation. Le domaine d'étude est celui du cinéma et des films en général, ces travaux entrant dans le cadre d'un service de recommandation prochainement disponible sur la plateforme de VOD de l'entreprise Orange.

Une autre contribution de cette thèse porte sur la nature des données textuelles étudiées. Les textes récupérés sont en effet très spécifiques. Ils sont en général très courts (une dizaine de mots) et le style dans lequel ils sont rédigés se rapproche de celui utilisé dans les SMS (Short Message Service) ou dans les systèmes de messagerie instantanée. Ils sont porteurs de caractéristiques sémantiques et syntaxiques très particulières comme des abréviations de mots, des fautes d'orthographes volontaires et involontaires, des onomatopées, des étirements de mots (multiplications volontaires de certaines lettres à l'intérieur d'un mot), des smileys, etc. Les linguistes commencent fortement à s'intéresser à cette forme d'écriture qu'ils considèrent comme un dialecte à part entière, avec ses propres règles, et qui possède sa propre communauté linguistique qui fait vivre et évoluer cette

7. www.myspace.com

écriture au fil du temps. Ce type de textes est toutefois très peu étudié dans le domaine de la fouille d'opinion. Dans cette thèse, différentes approches, traitements et outils couramment utilisés dans la littérature sont testés et évalués sur ces données particulières. Nos expériences visent à identifier quels sont les traitements pertinents pour ce type de textes et quelle méthode de classification est à privilégier.

1.3 Organisation du document

La figure 1.1 récapitule l'enchaînement des différentes tâches présentées dans cette thèse. Le manuscrit est ordonné suivant ces différentes tâches. Nous présentons tout d'abord l'objectif final des travaux, ainsi que de la chaîne de traitements, qui est la recommandation personnalisée. Après un état de l'art de la recommandation (Chapitre 2), nous décrivons le moteur de recommandation utilisé (Chapitre 3) puis les données textuelles employées pour les différentes expérimentations (Chapitre 4). Nous étudions ensuite chacun des chemins possibles permettant de rendre les données textuelles non structurées interprétables par le moteur de recommandation (Chapitres 5 à 7).

La suite de cette thèse est ainsi décomposée en six chapitres dont voici le plan détaillé :

- Le **Chapitre 2** fait l'état de l'art du domaine de la recommandation automatique.
  - La **première section** de ce chapitre présente les différentes formes de recommandation existantes. Elles peuvent être de deux formes. Il existe tout d'abord les *recommandations contextuelles* qui consistent à établir des recommandations basées sur la page Web courante visionnée par l'utilisateur. Il existe ensuite la *recommandation personnalisée* qui consiste à établir des recommandations basées sur le profil de l'utilisateur.
  - La **deuxième section** présente plus en détail les méthodes de recommandation personnalisée qui est la recommandation qui nous intéresse le plus. Les approches de la recommandation personnalisée sont de deux types appelés « filtrages ». Il existe tout d'abord le *filtrage thématique* qui consiste à trouver les items susceptibles d'intéresser l'utilisateur par le biais de descripteurs décrivant les articles du catalogue. Il y a ensuite le *filtrage collaboratif* qui consiste à comparer les goûts et préférences d'un grand nombre d'utilisateurs afin de retrouver les utilisateurs aux goûts similaires ou les items appréciés par les mêmes personnes.

- Le **Chapitre 3** présente le *moteur de recommandation* utilisé pour les différentes expériences. Ce moteur possède la particularité de pouvoir fonctionner soit en mode thématique soit en mode collaboratif.
  - La **première section** décrit le fonctionnement du moteur.
  - La **deuxième section** présente les résultats de différents étalonnages du moteur effectués à partir d'un corpus de notes en mode collaboratif, et d'un corpus de descripteurs en mode thématique.

- Le **Chapitre 4** présente les données textuelles utilisées pour les expérimentations. Il s'agit de commentaires d'internautes issus du site *Flixster*, un site communautaire réservé aux amateurs de cinéma.
  - La **première section** décrit tout d'abord les particularités générales que peuvent posséder les textes issus de sites communautaires.
  - La **deuxième section** présente le corpus construit dans le cadre de la thèse. Le processus d'extraction des textes est décrit et des informations statistiques ainsi que des exemples de commentaires sont énumérés.
  - Dans la **troisième section**, les résultats d'une analyse approfondie de la nature des commentaires, mettant en œuvre une méthode de co-clustering, sont présentés.

- Le **Chapitre 5** présente les premières expérimentations effectuées sur le corpus de textes. Dans cette partie, les textes sont employés afin d'extraire des descripteurs de films. Ces descripteurs permettent alors de faire de la recommandation personnalisée à l'aide du *filtrage thématique*.
  - La **première section** présente les différents choix effectués au niveau de la sélection des variables descriptives ainsi qu'au niveau des métriques employées pour mesurer les distances entre les documents.
  - La **deuxième section** présente les résultats en sortie de chaîne des évaluations et positionne ces résultats avec ceux obtenus lors de l'étalonnage.

- Le **Chapitre 6** fait l'état de l'art de la *fouille d'opinion* et se concentre plus spécifiquement sur la *classification d'opinion*. Trois approches dominantes existent pour la classification. Les *approches basées sur les lexiques*, les *approches basées sur l'apprentissage automatique* et les *approches hybrides*.
  - La **première section** aborde les objectifs, raison d'être et difficultés de la classification d'opinion.
  - La **deuxième section** fait l'état de l'art des approches basées sur les lexiques. Ce type d'approche consiste à répertorier les mots sémantiquement porteurs d'opinion et à classer les documents suivant qu'ils contiennent ou non ces mots.
  - La **troisième section** fait l'état de l'art des approches basées sur l'apprentissage automatique. Ces approches consistent à utiliser des méthodes issues du domaine de la classification supervisée afin de classer les documents.
  - La **quatrième section** fait l'état de l'art des approches hybrides qui consistent à assembler ou enchaîner les deux types d'approches précédents.
  - La **cinquième section** présente les différentes mesures d'évaluation utilisées dans le domaine. Ces mesures sont communes avec la fouille de textes.

- Le **Chapitre 7** présente toutes les expériences menées afin d'évaluer la chaîne de traitements en *mode collaboratif*, c'est-à-dire en passant par l'inférence de notes sur les commentaires à l'aide de la classification d'opinion.
  - La **première section** présente les résultats obtenus avec une approche basée sur les lexiques. Plusieurs lexiques sont comparés ainsi que différents essais de prise en compte de la négation qui est une des problématiques de la fouille d'opinion.

  - La **deuxième section** décrit les expériences menées avec des méthodes issues de l'apprentissage automatique. Les traitements et méthodes couramment employés dans le domaine sont comparés sur nos données particulières.
  - Pour finir, la **troisième section** présente l'évaluation des données d'usage obtenues à la suite de la classification d'opinion et les résultats sont discutés.

- Enfin, la dernière partie (**Chapitre 8**) propose un bilan des travaux menés au cours de cette thèse et ouvre sur des perspectives de travaux futurs.

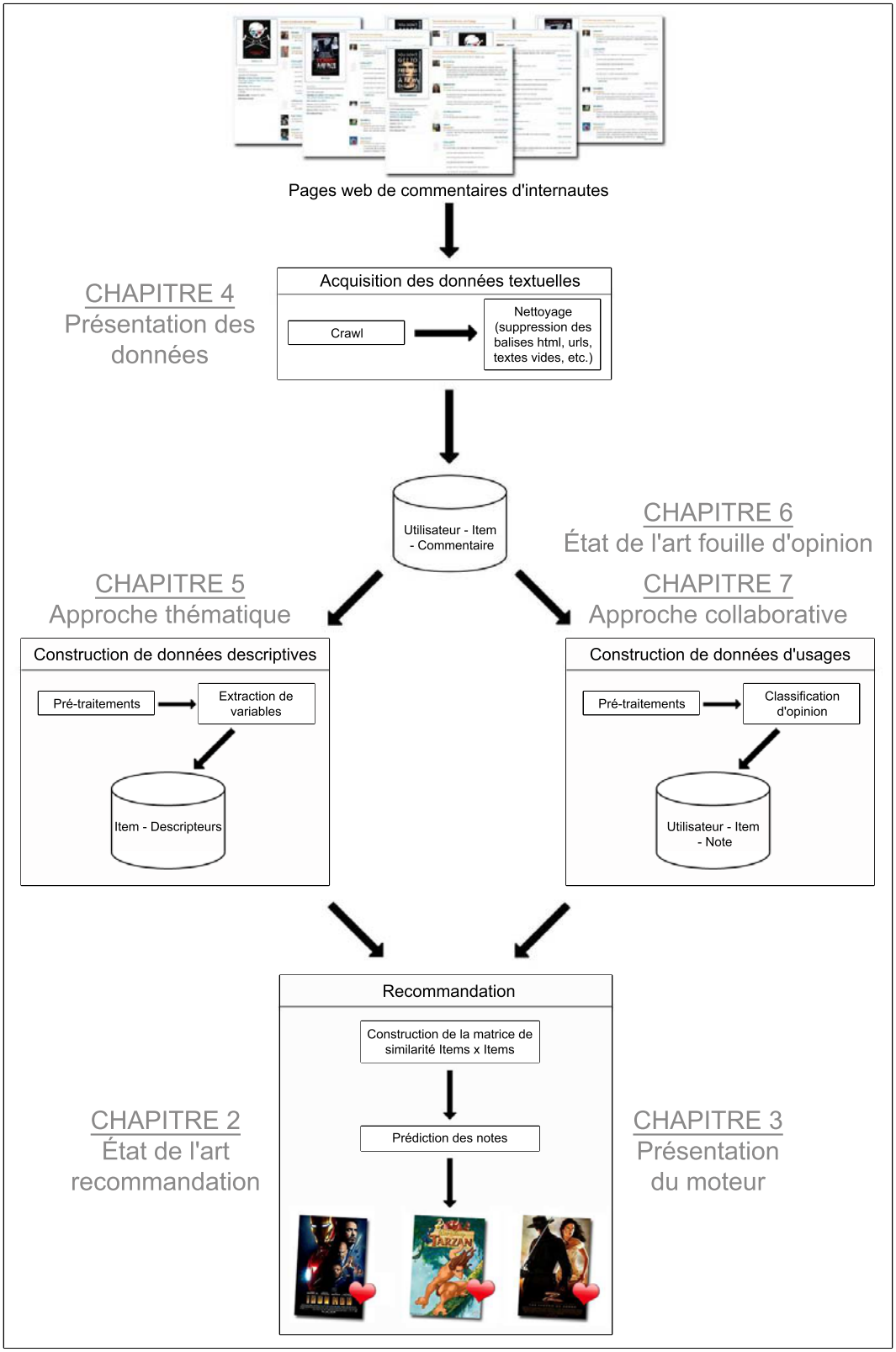


FIGURE 1.1 – Chaîne de traitement

Chapitre 2

Les systèmes de recommandation

Sommaire

Ce chapitre présente les différents principes de recommandation et décrit les techniques existantes les plus répandues dans le domaine. L'objectif principal de la recommandation est de gérer la surcharge d'information en la filtrant. Elle permet notamment de guider un utilisateur dans un catalogue de contenus, nommés plus couramment « items » par la communauté. Ces items peuvent être extrêmement variés allant des produits de consommation tels que les livres, CD ou DVD jusqu'aux pages Web, news, restaurants, vidéos, images, etc. Les moteurs de recommandation sont aujourd'hui de plus en plus présents sur la toile et vont certainement devenir indispensables dans le futur avec l'augmentation permanente des données disponibles en ligne (journaux, vidéos, jeux, musique, etc.) ainsi qu'avec l'avancée des technologies permettant d'accéder à tous ces contenus de n'importe où comme les nouvelles générations de téléphone mobiles, la VOD (Vidéo à la Demande) sur la télévision ou sur les baladeurs numériques, les livres électroniques, etc.

En plus de l'aide à la navigation, la recommandation a également pour objectif de promouvoir un catalogue de contenus. Un utilisateur est plus à même de parcourir un catalogue et d'aller plus loin dans ses recherches si les produits les plus susceptibles de l'intéresser sont mis en avant. C'est également un moyen d'attirer ou de fidéliser les clients ou les utilisateurs. La recommandation personnalisée par exemple, qui est une forme spécifique de recommandation, a pour objectif de découvrir des items qui plairont à un utilisateur en particulier. Un système de recommandation personnalisée joue donc, en quelque sorte, le rôle du vendeur de boutique de quartier. Il connaît ses clients, leurs goûts, leurs habitudes, et peut ainsi les conseiller et les guider dans leurs choix. Si le vendeur est agréable et donne de bons conseils, le client est plus susceptible de revenir.

Dans ce chapitre nous présentons tout d'abord les différentes formes de recommandation existantes, puis nous nous concentrons sur celle qui nous intéresse le plus, à savoir la recommandation personnalisée. Nous présentons notamment les deux méthodes les plus couramment utilisées en recommandation personnalisée, à savoir le filtrage collaboratif et le filtrage thématique ou filtrage basé sur le contenu. Pour finir, nous présentons les métriques d'évaluation couramment utilisées dans l'état de l'art.

2.1 Typologie de la recommandation

Il existe différentes formes de recommandation, suivant les données à recommander, suivant les informations disponibles et bien évidemment suivant l'objectif visé. Nous présentons ici les formes de recommandation les plus répandues et nous citons quelques exemples connus employant ces technologies.

2.1.1 La recommandation éditoriale

La recommandation éditoriale est généralement utilisée lorsqu'aucun autre système de recommandation n'est présent ou encore lorsque le système n'a aucune connaissance sur le visiteur du site. Cette forme de recommandation a pour principal objectif d'attirer rapidement l'œil de l'utilisateur novice afin de lui procurer l'envie de parcourir une partie du catalogue, comme le ferait la une d'un journal. Pour cela, on peut mettre en avant les produits les plus populaires, les nouveautés, les articles les mieux notés, les promotions, etc.

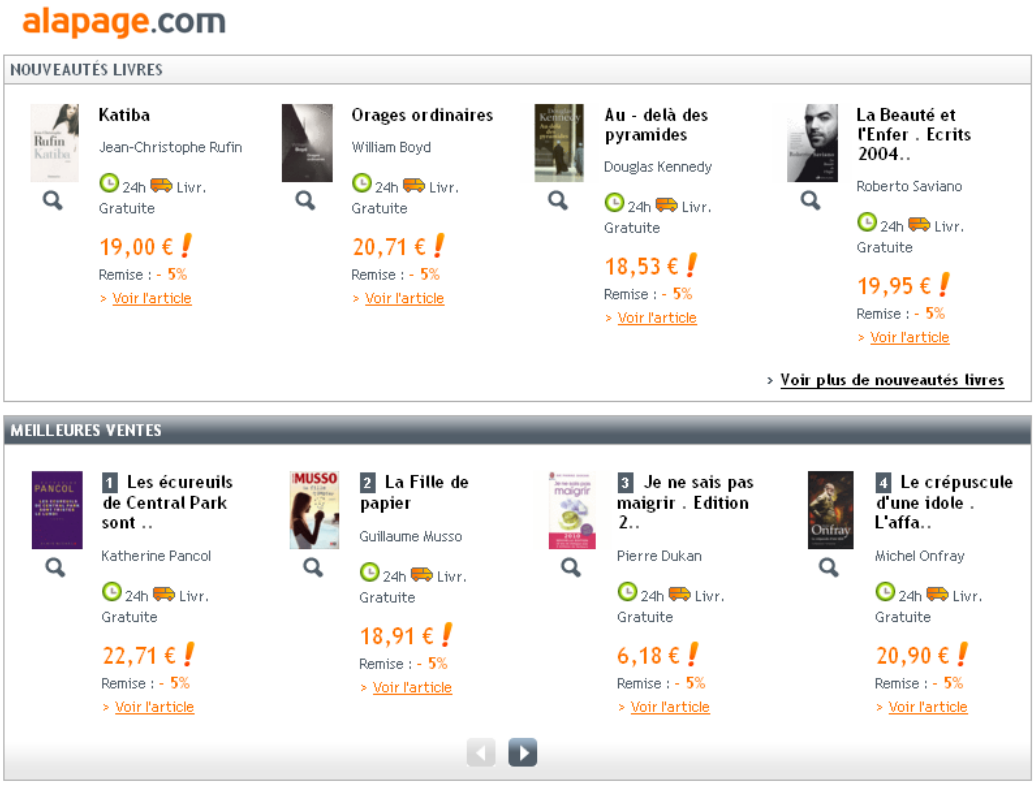


FIGURE 2.1 – Exemple de recommandation éditoriale présente sur le site *Alapage*

C'est ce que proposent tous les sites de vente par correspondance tels qu'*Alapage*[1]

1. http ://www.alapage.com

par exemple (voir figure 2.1), mais également tous les sites proposant du contenu en ligne comme les sites d'hébergement de vidéos ou de musiques.

2.1.2 La recommandation sociale

Avec cette méthode, les recommandations sont faites par des internautes, pour d'autres internautes. Il peut s'agir de simples utilisateurs, comme sur *Youtube*[2] ou *Flixster*[3], ou de consommateurs comme sur *Amazon*, *Priceminister*[4] ou encore le site de la *Fnac*[5]. Ce type de recommandations peut être basé sur le principe du bouche à oreille. Sur *Flixster* par exemple, les utilisateurs ont la possibilité de faire des recommandations à l'intérieur de leur réseau social en transmettant leurs commentaires et leurs appréciations à leurs amis. Une autre solution utilisée par certains sites est de permettre aux utilisateurs de créer des listes de « coups de cœur ». Ces listes accompagnent ensuite les profils des produits, comme sur *Amazon*, ou accompagnent les profils des utilisateurs, comme sur *Youtube*. La figure 2.2 présente ces deux exemples.

FIGURE 2.2 – Exemple de recommandations sociales proposées par *Amazon* et *Youtube*

Sur *Myspace*[6], le principe est légèrement différent. Ce sont les artistes, soit les producteurs de contenu, qui recommandent d'autres artistes. Un artiste conseille d'autres artistes dont il apprécie les œuvres et réciproquement, ce qui offre à chacun d'entre eux l'opportunité d'être vu par plus de personnes. Les auditeurs qui le désirent peuvent ainsi découvrir des nouveautés en se promenant de liens en liens.

2. http ://www.youtube.com
3. http ://www.flixster.com
4. http ://www.priceminister.com
5. http ://www.fnac.com
6. www.myspace.com

2.1.3 La recommandation contextuelle

Le principe de la recommandation contextuelle est de proposer des items proches de l'item consulté. Les techniques de rapprochement des items peuvent être simples. On peut par exemple sélectionner des items du même univers, du même auteur, du même réalisateur, du même compositeur, de même couleur, etc. Elles peuvent également être plus complexes, comme avec les méthodes basées sur les usages. *Youtube* ou *Amazon* sont des exemples parfaits de cette technique. Lorsqu'un item est consulté par un internaute, par exemple un disque de Georges Brassens, le système recommande une liste d'items qui ont été appréciés par les utilisateurs ayant également apprécié ce disque de Brassens. La figure 2.3 montre un exemple de recommandation contextuelle basée sur les usages.

FIGURE 2.3 – Exemple de recommandations contextuelles sur le site de la *Fnac*

On trouve également des méthodes basées sur le contenu, c'est-à-dire que les recommandations sont établies en fonction de la description des items. Ces descripteurs, fondés sur une analyse humaine ou automatisée, peuvent être très variés. *Flickr*[7] ou *IMDb*[8] par exemple utilisent les tags (ou étiquettes en français) pour rapprocher les items entre eux. *Pandora*[9] compare les morceaux musicaux de son catalogue en analysant plus de 400 caractéristiques sonores.

2.1.4 La recommandation personnalisée

La recommandation personnalisée a pour objectif de déterminer, pour un utilisateur particulier, les contenus ou services les plus susceptibles de l'intéresser. Les recommandations faites par ces systèmes sont dites personnalisées dans le sens où elles sont établies en fonction de l'utilisateur qui en est bénéficiaire et non pour l'ensemble des utilisateurs comme c'est le cas avec les moteurs de recherche par exemple ou encore avec les autres types de recommandation présentés auparavant. Les enjeux de la recommandation personnalisée sont divers et bénéficient aussi bien à l'utilisateur qu'au fournisseur du service.

7. http ://www.flickr.com
8. http ://www.imdb.com
9. http ://www.pandora.fm

Tout d'abord le système de recommandation peut être considéré comme un remplaçant du commerçant en chair et en os qui écoute les désirs de ses clients, au cas par cas, afin de les guider dans leurs choix d'achats ou de locations. Le client y gagne en temps car il n'a pas besoin de fouiller tout le catalogue et le système peut également lui permettre de découvrir de nouvelles choses. Quant au fournisseur, lorsque la recommandation est bien réalisée, cela peut entraîner un gain de confiance du client et ainsi favoriser la fidélité, ce à quoi aspire toute stratégie marketing. La recommandation permet également au fournisseur de valoriser son catalogue en guidant l'utilisateur vers des contenus autres que les plus populaires. La figure 2.4 présente l'exemple type de recommandations personnalisées que le site *Allociné*[10] propose à ses utilisateurs : le système prédit des notes pour des films ou séries que l'utilisateur enregistré n'a pas encore notés. Cette liste organisée d'items permet alors à l'utilisateur de faire ses choix sur une partie du catalogue et non sur le catalogue tout entier. Ce système en l'occurrence est un service offert à l'utilisateur ayant pour unique objectif de fidéliser l'utilisateur, le site *Allociné* ne distribuant pas directement les items présents dans le catalogue.

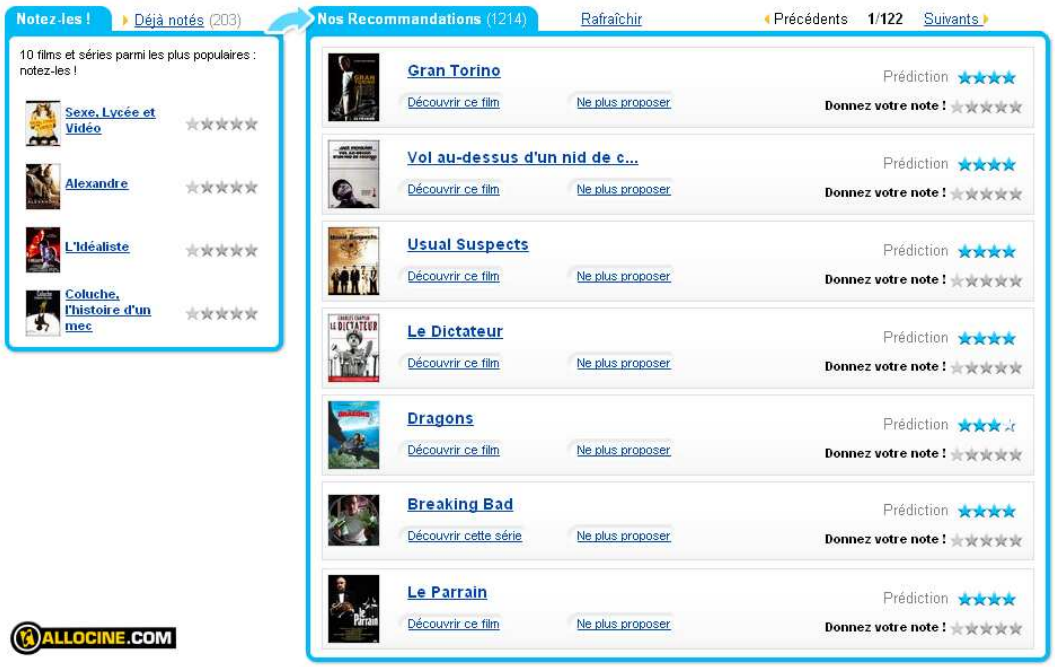


FIGURE 2.4 – Exemple de recommandations personnalisées proposées par *Allociné*

Parmi l'ensemble des types de recommandation, la recommandation personnalisée est certainement le domaine le plus actif du moment. Il intéresse autant les laboratoires de recherche pour toutes les problématiques qu'il entraîne que le monde industriel pour les possibilités marketing qu'offre ce type de systèmes.

10. http ://www.allocine.fr

2.2 État de l'art sur la recommandation personnalisée

Le premier moteur de recommandation personnalisée, nommé Tapestry, est apparu en 1992 [GNOT92]. Depuis lors, ces systèmes sont devenus le nerf de la guerre pour beaucoup d'entreprises. Les sites Web marchands ont tous adopté les recommandations, *Amazon* en tête, afin d'inciter les clients à acheter des produits auxquels ils n'auraient pas pensé. Les sites de vidéos ou de musique en ligne ont opté pour l'utilisation de ces outils afin de fidéliser leur clientèle en leur proposant les contenus les plus adaptés à leurs goûts sans qu'ils aient besoin de chercher par eux-mêmes. Les utilisateurs peuvent également trouver un intérêt à la recommandation personnalisée. La quantité de contenus disponibles aujourd'hui par le biais d'Internet est tellement considérable qu'il est impossible de tout voir ou tout connaître. Mettre à profit les machines afin de réaliser un premier filtrage sur ces nombreux contenus peut donc être très utile, voire nécessaire, afin de permettre aux utilisateurs de faire de nouvelles découvertes. Les moteurs de recommandation peuvent également guider l'utilisateur dans ses choix. Avant de voir un film au cinéma, de lire un livre ou même d'acheter un appareil électroménager, beaucoup de personnes ont l'habitude de se renseigner auprès de leur entourage afin de récolter des avis. Les intéressés analysent ensuite ces critiques afin de se faire une idée de ce qu'ils risquent de penser de l'objet en question. L'opinion prédite par les moteurs de recommandation peut alors servir d'information complémentaire et aider dans la prise de décision.

Plus concrètement, la recommandation personnalisée a pour objectif de filtrer des contenus ou items afin de ne conserver que les plus pertinents pour un utilisateur donné. Les items peuvent être des films, musiques, news, pages Web, livres, vidéos, images, etc. L'idée sous-jacente est de prédire l'opinion qu'un utilisateur portera sur les items qu'il ne connaît pas encore afin de ne lui proposer que ceux qu'il sera susceptible d'apprécier, ou tout du moins, qui auront une grande chance de l'intéresser. Une définition plus formelle de la recommandation est donnée par Adomavicius et Tuzhilin [AT05].

Définition 2.2.1. *Soit U l'ensemble de tous les utilisateurs, soit I l'ensemble de tous les items qui peuvent être recommandés, soit R un ensemble ordonné et soit $f : U \times I \rightarrow R$ une fonction qui prédit l'intérêt que portera l'utilisateur $u \in U$ à l'item $i \in I$. Alors pour chaque utilisateur $u \in U$, le système de recommandation sélectionne l'item $i' \in I$ qui maximise l'intérêt de u :*

$$\forall u \in U, i'_u = argmax_{i \in I} f(u, i)$$

L'intérêt d'un utilisateur pour un item (la fonction $f(u, i)$) est généralement représenté par une note indiquant l'appréciation que l'utilisateur porterait sur l'item. Afin de deviner cet intérêt, des connaissances sur l'utilisateur en question sont nécessaires. Les goûts connus des utilisateurs sont généralement caractérisés par leurs appréciations portées sur les contenus déjà consultés. Ces informations sont regroupées dans une matrice appelée « matrice d'usages ». Le tableau 2.1 présente un exemple fictif de matrice binaire contenant des informations de type « l'utilisateur u a apprécié/n'a pas apprécié l'item i ». Ces

informations peuvent également être « a acheté/n'a pas acheté », « a consulté/n'a pas consulté », etc. Elles peuvent également se mesurer sur un nombre plus élevé de classes : « a mis 1/2/3/4/5 étoiles », etc. Une fois la matrice d'usages construite, l'objectif du moteur de recommandation est de deviner les connexions utilisateur-item manquantes. En d'autres termes, on demande à l'outil de remplir les cases vides C_{ui} de la matrice en évaluant si l'item i intéressera l'utilisateur u ou non. Pour cela, trois types d'approches sont principalement utilisés [NRT08] : le filtrage basé sur le contenu, le filtrage collaboratif et le filtrage hybride.

	Item 1	Item 2	Item 3	Item 4	Item 5	...
User 1	😎			😎		...
User 2		😞	😎			...
User 3		😎				...
User 4			😞	😎	😞	...
...	...	...	...	...	...	...

TABLE 2.1 – Exemple de matrice d'usages

Le filtrage basé sur le contenu ou filtrage thématique, s'appuie sur le contenu des items (par le biais de descripteurs) afin de les comparer à d'autres profils eux-mêmes constitués de descripteurs [PB07]. Chaque utilisateur du système possède un profil qui décrit ses propres centres d'intérêt. Lors de l'arrivée d'un nouvel item, le système compare la représentation de l'item avec le profil utilisateur afin de prédire l'opinion que pourrait porter l'utilisateur sur cet item s'il le connaissait. Les items sont alors recommandés en fonction de leur proximité avec le profil de l'utilisateur.

Le filtrage collaboratif se base sur les appréciations données par un ensemble d'utilisateurs sur les items. Ces appréciations peuvent être des notes, des achats effectués, des pages consultées, etc. On distingue deux grandes approches de filtrage collaboratif. L'approche basée sur les utilisateurs [RIS+94a] consiste à comparer les utilisateurs entre eux et à retrouver ceux ayant des goûts en communs, les notes d'un utilisateur étant ensuite prédites selon son voisinage. L'approche basée sur les items [SKKR01] consiste à rapprocher les items appréciés par des personnes communes et à prédire les notes des utilisateurs en fonction des items les plus proches de ceux qu'ils ont déjà notés.

Le filtrage hybride consiste, comme son nom l'indique, à exploiter aussi bien les informations de type collaboratives que les descripteurs de contenus. Les systèmes hybrides peuvent également faire appel à des sources d'informations complémentaires telles que des données démographiques ou sociales [Paz99]. Différentes méthodes d'hybridation peuvent être envisagées afin de combiner les sources ou les modèles. On peut par exemple appliquer séparément le filtrage collaboratif et d'autres techniques de filtrage pour générer des recommandations candidates, et combiner ces ensembles de recommandations par pondération, cascade, bascule, etc. afin de produire les recommandations finales pour les utilisateurs

[Bur07].

Dans ce chapitre nous décrirons plus formellement les deux premières approches qui sont les approches qui nous intéressent le plus. Nous énoncerons ensuite les problèmes récurrents des systèmes de recommandation, notamment le cas du démarrage à froid qui est le plus connu d'entres eux. Pour finir, nous listerons les méthodes d'évaluations utilisées dans le domaine de la recommandation automatique.

2.2.1 Le filtrage collaboratif

Le terme « filtrage collaboratif » [SKKR01] désigne les systèmes de recommandation qui se basent sur les opinions et évaluations d'un groupe de personnes afin d'aider un individu particulier. Ce type de moteur utilise uniquement les informations contenues dans la matrice d'usages comme données d'entrée. La matrice peut être construite en surveillant les comportements des utilisateurs ou encore en proposant aux utilisateurs de déclarer eux-mêmes leurs avis sur les items qu'ils connaissent :

- On appelle filtrage collaboratif « passif » les systèmes de recommandation qui reposent sur l'analyse des comportements des utilisateurs (par exemple les achats effectués ou les pages visitées sur le site *Amazon*) ;
- On nomme filtrage collaboratif « actif » les systèmes de recommandations basés sur des données déclarées par les utilisateurs (comme des notes sur le site *Allociné*).

Afin de remplir les cases vides de la matrice, plusieurs options sont possibles. Deux grands axes se distinguent dans la littérature. Les approchés basées sur les plus proches voisins, appelées aussi approches basées sur la mémoire, et les approches basées sur les modèles. Des hybridations de ces approches existent également.

Les approches basées sur les modèles mettent en œuvre des méthodes issues de l'apprentissage automatique (Machine Learning) comme des modèles bayésiens ou des méthodes de clustering. Ces méthodes sont généralement performantes mais ont un coût de construction et de fonctionnement plus important que les méthodes basées sur les plus proches voisins [CMB07, SK09]. De plus, ces méthodes semblent plus efficaces que les approches basées sur les plus proches voisins uniquement dans le cas de données d'usages clairsemées. Dans cette thèse, nous nous plaçons dans un contexte industriel, par conséquent nous ne nous intéressons pas à ces méthodes coûteuses. De plus, l'objectif de ces travaux est de comparer l'information présente dans les différentes sources de données et non pas d'obtenir le meilleur système de recommandation qui soit. Les approches basées sur les plus proches voisins sont donc les méthodes qui nous intéressent le plus, de par leur simplicité mais également leurs résultats qui concurrencent les méthodes plus lourdes. Toutefois, pour le lecteur intéressé, une description précise des approches basées sur les modèles est proposée par Su et Khoshgoftaar [SK09].

Les approches basées sur les plus proches voisins consistent à estimer les similarités entre les lignes ou entre les colonnes de la matrice d'usages [AT05, SFHS07]. Dans le cas

des lignes, cela consiste plus précisément à retrouver les personnes ayant les mêmes comportements que la personne à qui l'on souhaite faire des recommandations. Dans le cas des colonnes, on recherche les items qui ont été appréciés par le même public. Ce type de recommandations se fait donc en deux étapes : une première étape où l'on calcule les similarités entre les lignes ou les colonnes de la matrice, et une deuxième étape où l'on remplit les cases vides de la matrice à l'aide d'une fonction de prédiction de notes.

Afin de présenter les différentes approches possibles pour chacune des deux étapes du filtrage collaboratif, notons :

- U un ensemble de N utilisateurs ;
- I un ensemble de M items ;
- R un ensemble de notes n_{ui} attribuées par l'utilisateur $u \in U$ sur l'item $i \in I$;
- $S_u \subseteq I$ l'ensemble des items notés par l'utilisateur u ;
- $S_i \subseteq U$ l'ensemble des utilisateurs ayant noté l'item i.

2.2.1.1 Le calcul des similarités

Une matrice peut être vue comme un ensemble de vecteurs. La décomposition de la matrice en vecteurs peut se faire selon les lignes ou selon les colonnes (voir figure 2.5). Le calcul d'une similarité consiste à mesurer la similitude entre deux de ces vecteurs. Le choix de la mesure de similarité utilisée dépend généralement de la nature des vecteurs. Si les vecteurs contiennent uniquement des données binaires par exemple, du type « a acheté/n'a pas acheté », la *distance de Jaccard* peut être utilisée (2.1). Cette distance mesure le recouvrement entre les attributs des deux vecteurs mais ne tient pas compte des différences de notes entre les deux vecteurs.

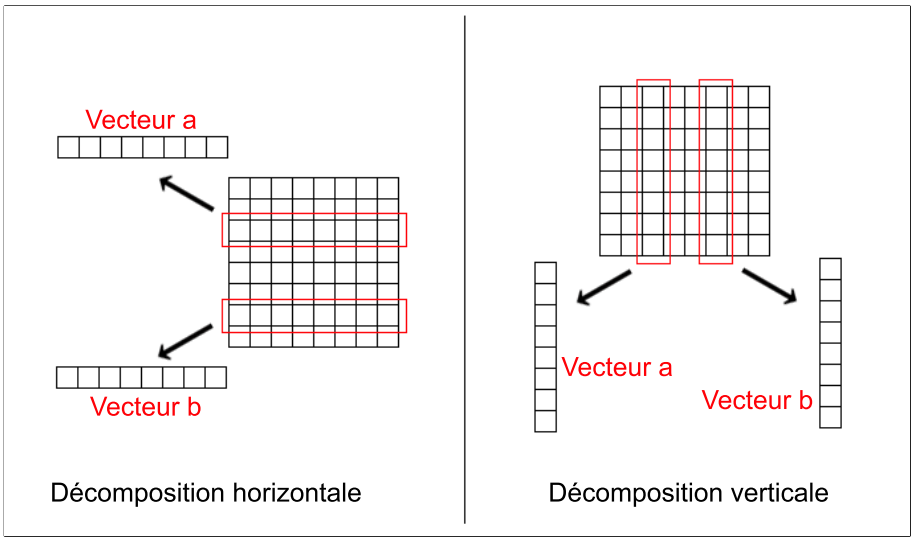


FIGURE 2.5 – Exemples de décompositions d'une matrice

$$Sim_{jaccard}(a,b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|} \tag{2.1}$$

Dans le cas où les données contenues dans les vecteurs sont des notes n et que l'on souhaite tenir compte de la valeur de ces notes, les deux mesures les plus utilisées sont la *similarité Cosinus* (2.2) et la *similarité de Pearson* (2.3).

$$Sim_{cosinus}(a,b) = \frac{\sum_{\{x \in S_a \cap S_b\}} n_{ax} \times n_{bx}}{\sqrt{\sum_{\{x \in S_a \cap S_b\}} n_{ax}^2 \sum_{\{x \in S_a \cap S_b\}} n_{bx}^2}} \tag{2.2}$$

$$Sim_{pearson}(a,b) = \frac{\sum_{\{x \in S_a \cap S_b\}} (n_{ax} - \overline{n_a}) \times (n_{bx} - \overline{n_b})}{\sqrt{\sum_{\{x \in S_a \cap S_b\}} (n_{ax} - \overline{n_a})^2 \sum_{\{x \in S_a \cap S_b\}} (n_{bx} - \overline{n_b})^2}} \tag{2.3}$$

Où $\overline{n_a}$ (respectivement $\overline{n_b}$) représente la moyenne des notes contenues dans le vecteur a (respectivement b).

Ces différentes mesures permettent de construire soit une matrice de similarités Items-Items (tableau 2.2), soit une matrice de similarités Utilisateurs-Utilisateurs (tableau 2.3) selon qu'on travaille sur les lignes ou les colonnes. Cette matrice de similarités est ensuite utilisée afin de prédire des notes.

	Item 1	Item 2	Item 3	Item 4	Item 5	...
Item 1	X	$Sim(1,2)$	$Sim(1,3)$	$Sim(1,4)$	$Sim(1,5)$	...
Item 2	$Sim(2,1)$	X	$Sim(2,3)$	$Sim(2,4)$	$Sim(2,5)$	...
Item 3	$Sim(3,1)$	$Sim(3,2)$	X	$Sim(3,4)$	$Sim(3,5)$	...
Item 4	$Sim(4,1)$	$Sim(4,2)$	$Sim(4,3)$	X	$Sim(4,5)$	...
...	...	...	...	...	...	...

TABLE 2.2 – Exemple de matrice de similarités Items-Items

2.2.1.2 La prédiction des notes

La prédiction des notes consiste à deviner l'intérêt qu'un utilisateur pourrait porter à des items qu'il ne connaît pas, ou plus précisément des items sur lesquels il n'a porté aucune opinion connue par le système. Concrètement, l'objectif de la prédiction de notes consiste à remplir les cases vides de la matrice d'usages. Cette tâche nécessite l'utilisation d'une matrice de similarités. Le fait de pouvoir construire différentes matrices de

	User 1	User 2	User 3	User 4	User 5	...
User 1	X	$Sim(1,2)$	$Sim(1,3)$	$Sim(1,4)$	$Sim(1,5)$	...
User 2	$Sim(2,1)$	X	$Sim(2,3)$	$Sim(2,4)$	$Sim(2,5)$	...
User 3	$Sim(3,1)$	$Sim(3,2)$	X	$Sim(3,4)$	$Sim(3,5)$	...
User 4	$Sim(4,1)$	$Sim(4,2)$	$Sim(4,3)$	X	$Sim(4,5)$	...
...	...	...	...	...	...	...

TABLE 2.3 – Exemple de matrice de similarités Utilisateurs-Utilisateurs

similarités, Items-Items ou Utilisateurs-Utilisateurs, entraîne différentes approches pour la prédiction des notes.

Pour les approches basées sur les utilisateurs [RIS+94b], le principe consiste à chercher des personnes ayant les mêmes comportements que la personne à qui l'on souhaite faire des recommandations. On établit alors les recommandations en fonction des notes d'utilisateurs similaires. On utilise ici une matrice de similarités construite selon les utilisateurs. Soit $sim(u, v)$ la fonction de similarité entre les utilisateurs $u \in U$ et $v \in U$ et soit U_u l'ensemble des utilisateurs au comportement proche de l'utilisateur u.

Une des façons possibles de calculer la prédiction de la note de l'utilisateur u sur l'item i consiste à utiliser la somme des notes des utilisateurs au comportement proche ayant déjà noté l'item i (2.4).

$$n_{ui} = \frac{\sum_{\{v \in U | i \in S_v\}} sim(u, v) \times n_{vi}}{\sum_{\{v \in U | i \in S_v\}} |sim(u, v)|} \tag{2.4}$$

L'un des principaux problèmes du filtrage collaboratif actif est la différence d'utilisation des systèmes de notes en fonction des utilisateurs. En effet un utilisateur peut, par exemple s'il considère que la perfection n'existe pas, ne jamais affecter la note maximale à un contenu et donc répartir ses notes de 1 à 4 (si les notes possibles vont de 1 à 5). À l'inverse, un utilisateur différent peut, s'il n'aime pas noter trop sévèrement ce qu'il n'a pas apprécié, répartir les notes qu'il attribue de 2 à 5. La solution la plus utilisée pour pallier cet inconvénient est l'utilisation de la moyenne des notes de l'utilisateur (2.5).

$$n_{ui} = \overline{n_u} + \frac{\sum_{\{v \in U | i \in S_v\}} sim(u, v) \times (n_{vi} - \overline{n_v})}{\sum_{\{v \in U | i \in S_v\}} |sim(u, v)|} \tag{2.5}$$

où $\overline{n_u}$ (respectivement $\overline{n_v}$) représente la moyenne des notes de l'utilisateur u (respectivement v) :

$$\overline{n_u} = \frac{\sum_{\{i \in S_u\}} n_{ui}}{|S_u|}$$

L'intérêt porté aux approches basées sur les items est plus récente que celles basées sur les utilisateurs [SKKR01, Kar01, LSY03, DK04]. Cette approche a été popularisée par le site *Amazon* avec un système qui consiste à construire une matrice de relations entre les items en se basant sur les achats des clients du site. Comme pour les approches basées sur les utilisateurs, les performances varient en fonction de la mesure de similarité utilisée et en fonction du nombre d'items proches considérés. La matrice de similarités utilisée est construite suivant les items. Comme pour l'approche basée sur les utilisateurs, une première façon de calculer la prédiction de la note d'un utilisateur u sur un item i ne prend pas en compte les moyennes de notes (2.6).

$$n_{ui} = \frac{\sum_{\{j \in S_u \cap I_i\}} sim(i,j) \times n_{uj}}{\sum_{\{j \in S_u \cap I_i\}} |sim(i,j)|} \tag{2.6}$$

Pour palier les différences d'utilisations des notes de la part des utilisateurs, une autre version utilise la moyenne des notes de chaque utilisateur (2.7).

$$n_{ui} = \overline{n_i} + \frac{\sum_{\{j \in S_u \cap I_i\}} sim(i,j) \times (n_{uj} - \overline{n_j})}{\sum_{\{j \in S_u \cap I_i\}} |sim(i,j)|} \tag{2.7}$$

où $\overline{n_i}$ (respectivement $\overline{n_j}$) représente la moyenne des notes reçues par l'item i (respectivement j) :

$$\overline{n_i} = \frac{\sum_{\{u \in U | i \in S_u\}} n_{ui}}{|\{u \in U | i \in S_u\}|}$$

Le filtrage collaboratif semble être la méthode de recommandation personnalisée qui garantit les meilleurs résultats. C'est par ailleurs la plus utilisée. Cependant, le bon fonctionnement de cette méthode nécessite une grosse quantité de données et donc d'utilisateurs. Elle nécessite également des items « durables », c'est-à-dire des items qui ont une actualité assez longue pour que les utilisateurs aient le temps de les noter et que l'algorithme ait le temps d'établir les recommandations. Concrètement, le filtrage collaboratif ne sera pas forcément très adapté pour un site de news. Une autre approche, non basée sur les utilisateurs, peut alors être utilisée afin de pallier le manque de données. Il s'agit du filtrage thématique.

2.2.2 Le filtrage thématique

Le filtrage thématique, ou filtrage basé sur les contenus [VMVS00, PB07], consiste à établir des recommandations à l'aide d' « attributs ». Ces attributs, parfois appelés « descripteurs », « caractéristiques », « propriétés » ou encore « variables » dans la

littérature, représentent les items. Plus formellement, les items sont représentés sur un vecteur $X = (x_1, x_2, ..., x_n)$ de n composantes. Chaque composante représente un attribut et peut contenir des valeurs binaires, numériques ou encore nominales. Dans le cas de la recommandation de films par exemple, les attributs peuvent être le genre, le réalisateur, l'année de production, le nombre de récompenses, etc. Ce type de vecteurs fait alors office de profil. Une fois les profils construits, l'objectif du moteur est d'évaluer leurs similarités.

Ce type de systèmes est généralement utilisé dans deux situations. Ils peuvent tout d'abord remplacer le filtrage collaboratif lorsque la quantité de données d'usages disponible est insuffisante pour obtenir de bons résultats. Ils sont également utilisés pour la recommandation d'items à courte durée de vie comme les news.

2.2.2.1 Construction des profils

Le filtrage thématique se base donc sur la description des items et l'enrichissement de profils utilisateurs afin de croiser les deux types d'informations connues. Comme pour le filtrage collaboratif, les profils utilisateurs peuvent être construits à partir d'informations collectées de deux manières :

- Ils peuvent tout d'abord être construits de manière *passive*. Dans ce cas, on considère les items sélectionnés par l'utilisateur ou en se basant sur son passif : les pages consultées, les produits achetés, etc.
- Ils peuvent également être construits de manière *active* en proposant aux utilisateurs de remplir des questionnaires par exemple, ou encore en permettant aux utilisateurs d'attribuer des notes aux items reflétant leur intérêt.

Selon la manière dont les informations ont été collectées, les profils utilisateurs peuvent contenir soit les items qu'ils ont appréciés ou non, soit des descripteurs. Ces descripteurs peuvent correspondre à ceux des items qu'ils ont notés ou consultés ou être déduits des réponses au questionnaire. Dans le premier cas, l'objectif du moteur de recommandation sera de retrouver les items du catalogue les plus proches des items appréciés par l'utilisateur, ainsi que de filtrer les items proches de ceux qu'il a détestés. Dans les deux autres cas, le moteur de recommandation cherchera des items ayant le plus grand nombre de descripteurs en commun avec l'utilisateur. Ce choix à faire a un impact direct sur les résultats, le deuxième cas permettant une recherche plus large. Considérons par exemple un utilisateur ayant acheté un pull rouge et un pantalon bleu et considérons deux descripteurs *type de vêtement* et *couleur*. Dans le premier cas, le moteur de recommandation pourra sélectionner dans le catalogue de nouveaux pulls rouges et de nouveaux pantalons bleus. Dans le deuxième cas, il sélectionnera ces mêmes produits mais également les pulls bleus et les pantalons rouges. Le choix de la solution dépend alors de l'objectif visé.

Les profils des items peuvent également être construits de plusieurs façons. Tout d'abord, on peut se baser sur des données concrètes comme le genre d'un film ou les plats servis dans un restaurant. Lorsque ces données ne sont pas disponibles ou peu informatives, on peut alors analyser l'item et en extraire des méta-données. C'est la méthode

utilisée par le site Pandora qui analyse les morceaux musicaux sur environ 400 critères. Ces critères forment le profil de chaque morceaux et des similarités peuvent ainsi être calculées entre les morceaux appréciés par l'utilisateur et les morceaux qu'il n'a pas encore écoutés. Une troisième méthode de construction des profils items consiste à se baser sur des informations apportées par les utilisateurs, généralement sous forme de textes. Il peut s'agir par exemple de données structurées, comme avec les systèmes de *Tags*, mais également de données non-structurées telles que des textes descriptifs (synopsis par exemple), des critiques journalistiques ou encore des commentaires utilisateurs. Ces textes peuvent être utilisés afin d'en extraire des caractéristiques. Les mots sont souvent pris comme descripteurs et subissent généralement des traitements linguistiques comme la lemmatisation, des corrections, des suppressions de mots, etc. (ces traitements sont abordés plus en détail dans le chapitre 6 qui fait l'état de l'art de la classification d'opinion).

Une fois les profils des utilisateurs et des items construits, des mesures de similarités sont appliquées afin de les comparer et trouver les items correspondant le plus au profil utilisateur.

2.2.2.2 Calcul des similarités

L'un des principaux objectifs, dans ce type de systèmes de recommandations, est de déterminer des similarités entre les attributs. En effet, la couverture de certains attributs par rapport à l'ensemble du catalogue peut être trop peu étendue afin de permettre la recommandation de certains items. L'union faisant la force, l'idée est donc de rapprocher certains attributs afin d'agrandir leur couverture et ainsi harmoniser les probabilités pour chaque item d'apparaître dans une recommandation. Dans les techniques de mesures de similarités des attributs, nous pouvons citer la *Distance Normalisée Google* [CV07] qui consiste à calculer le nombre de co-occurrences de termes textuels dans les pages répertoriées par *Google*. Deux termes ayant un fort taux d'apparitions sur des pages communes peuvent donc être considérés comme proches. Une autre façon de rapprocher les attributs est d'utiliser l'algorithmique des graphes. Bothorel et Bouklit [BB08] proposent une technique de clustering de graphes afin de détecter des communautés de nœuds (les attributs) grâce aux arêtes (une arête relie deux attributs présents sur un même item) afin de déterminer les similarités entre attributs.

Les similarités entre profils peuvent ensuite être calculées avec les distances de similarités traditionnelles comme la distance de *Jaccard*, la similarité *Cosinus*, la distance de la corrélation de *Pearson* ou encore la méthode des plus proches voisins. Cette tâche peut également être vue comme une tâche de classification. Des outils d'apprentissage automatique sont alors utilisés, comme des classifieurs naïfs bayésiens ou des systèmes à base de règles. Néanmoins, les distances de similarités sont plus régulièrement utilisées et semblent offrir les meilleurs résultats. Mais le choix de la méthode dépend généralement de la forme des profils construits.

2.2.3 Les problèmes récurrents

Les différents types de filtrage ont leurs forces et faiblesses et il est nécessaire pour le concepteur du système de choisir la stratégie la plus adaptée en fonction du problème qui est considéré. Nous tentons, dans cette partie, de répertorier et d'expliquer les problèmes les plus récurrents.

2.2.3.1 Le démarrage à froid

Le démarrage à froid est un phénomène qui se produit naturellement lorsque le système n'a pas assez de données pour procéder à un filtrage de bonne qualité. Ce problème peut survenir quelle que soit la méthode de filtrage utilisée à des degrés divers. Dans le cas du filtrage collaboratif, il n'existe pas de solutions à ce problème. Un système ne possédant pas ou très peu d'utilisateurs ne peut pas émettre de recommandations. Dans ce cas, si des informations descriptives sur les items peuvent être acquises, le filtrage thématique est alors considéré. On peut distinguer trois types distincts de démarrage à froid [Bur02] : les cas du système débutant, du nouvel utilisateur ou du nouvel item.

Le cas du système débutant provient lors du lancement d'un nouveau service de recommandation. Le système ne possède alors aucune information sur les utilisateurs et sur les items. Les méthodes de filtrage collaboratif ne peuvent pas fonctionner sur une matrice d'usages vide. La solution consiste en général à trouver des informations descriptives des items afin d'organiser le catalogue et inciter les utilisateurs à le parcourir jusqu'à ce que la matrice d'usages soit assez remplie et permette de passer en mode collaboratif.

Le cas du nouvel utilisateur provient lorsqu'un nouveau visiteur entre dans le système et que l'on n'a aucune information sur lui. Plusieurs solutions peuvent être employées. On peut le soumettre à des questionnaires ou à des listes d'items à noter ou encore faire de la recommandation éditoriale afin de l'inciter à parcourir le catalogue de contenus et ainsi acquérir des informations nécessaires au système.

Le cas du nouvel item provient lorsqu'un nouvel item est inséré dans le catalogue. Dans le cas du filtrage basé sur les contenus, il s'agit alors de trouver des descripteurs afin de le comparer aux profils existants. Dans le cas du filtrage collaboratif, un item n'ayant reçu aucune note ne peut pas être recommandé. Il s'agit alors de le rendre visible aux utilisateurs afin qu'il obtienne un certain nombre de notes et puisse être recommandé.

2.2.3.2 L'effet entonnoir

L'un des avantages du filtrage thématique est que l'utilisateur, dans un tel système, ne dépend absolument pas des autres. Ainsi, il recevra des recommandations du système même s'il est le seul inscrit, pour peu qu'il ait décrit son profil en donnant un ensemble de thèmes qui l'intéressent. En revanche, cette technique de filtrage est soumise à l'effet « entonnoir », car le profil évolue naturellement par restriction progressive sur les

thèmes recherchés. Ainsi, l'utilisateur ne reçoit que les recommandations relatives aux thèmes présentés dans son profil, une fois devenu stable. Par conséquent, il ne peut pas découvrir de nouveaux domaines potentiellement intéressants pour lui. À l'opposé du filtrage thématique, tous les utilisateurs d'un système basé sur le filtrage collaboratif peuvent tirer profit des évaluations des autres en recevant des recommandations pour lesquelles les utilisateurs les plus proches ont émis un jugement de valeur favorable, et cela sans que le système dispose d'un processus d'extraction du contenu des documents. Grâce à son indépendance vis-à-vis de la représentation des données, cette technique peut s'appliquer dans les contextes où le contenu est soit indisponible, soit difficile à analyser, et en particulier elle peut s'utiliser pour tout type de données : texte, image, audio et vidéo. De plus, l'utilisateur est capable de découvrir divers domaines intéressants, car le filtrage collaboratif ne se fondant pas sur la dimension thématique des profils, n'est pas soumis à l'effet « entonnoir ».

2.2.3.3 La longue traine

La longue traine [And06], appelée également longue queue, est un phénomène très connu en recommandation, et plus largement en statistiques. Dans le domaine de la recommandation, il concerne tous les items non populaires ou les items nouvellement apparus dans le catalogue qui sont bien souvent complètement ignorés des moteurs de recommandation basés sur le filtrage collaboratif. En effet, ces items étant moins consultés que les items populaires, les algorithmes de filtrage collaboratif ne les considèrent pas ou très peu. Dans le temps, la différence entre les items populaires et les items très peu consultés ne fait que s'accentuer, les items populaires étant beaucoup plus souvent recommandés, et donc consultés, que les items non populaires. De plus, il s'avère que la quantité d'items non populaires est beaucoup plus importante que celle des items populaires (voir la figure 2.6 extraite de Park et Tuzhilin [PT08]), leur perte d'audience peut alors être conséquente. Cette problématique est liée à un manque d'information sur les items non populaires. On peut alors imaginer qu'acquérir de l'information « ailleurs » sur le net pourrait être une solution à manque de connaissances.

2.2.4 Les différentes évaluations utilisées dans le domaine

La mesure la plus utilisée dans le domaine de la recommandation est sans conteste la Roor Mean Squared Error (RMSE). C'est elle qui a été sélectionnée pour le challenge *Netflix*. Ce challenge a été proposé en 2006 par la société *Netflix* dans le but d'améliorer les performances du moteur de recommandation utilisé sur son site de location de DVD. Ils ont alors mis à disposition de tous une grosse quantité de données d'usages qui sont encore souvent utilisées comme benchmark par la communauté, bien que le challenge ait déjà été remporté. La Mean Absolute Error (MAE) était, avant ce challenge, la métrique d'évaluation la plus utilisée. D'autres mesures existent également mais sont largement moins répandues.

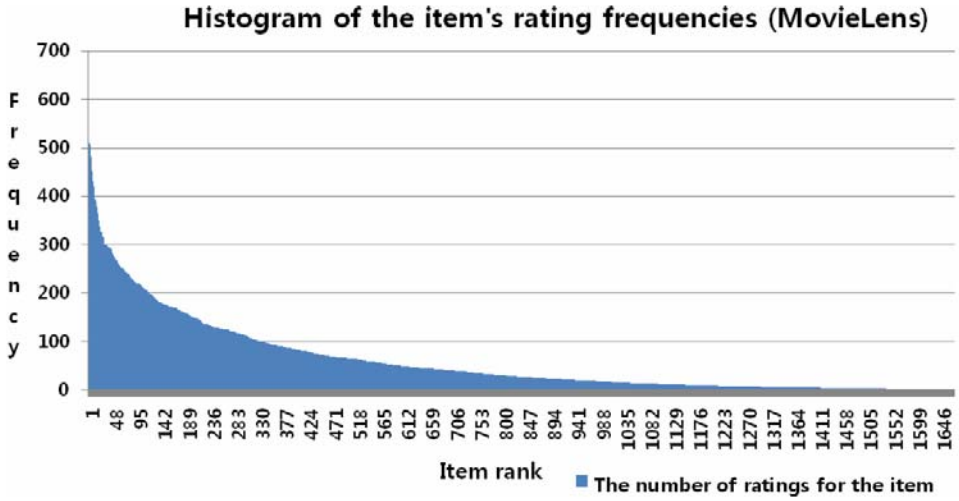


FIGURE 2.6 – Exemple de longue traine sur des données provenant du site MovieLens

2.2.4.1 La RMSE

La Root Mean Squared Error permet de mesurer l'erreur faite entre la note prédite et la note réelle donnée par l'utilisateur. Elle se calcule de la façon suivant :

$$RMSE = \sqrt{\frac{\sum_{u,i} (p_{ui} - n_{ui})^2}{n}}$$

Où n_{ui} représente la note réelle donnée par l'utilisateur u sur l'item i, p_{ui} la note prédite par le moteur de recommandation et n le nombre total de notes prédites.

La RMSE étant une mesure d'erreur, on attend d'elle qu'elle soit la plus basse possible. L'observation des résultats du challenge *Netflix* permet de se faire une idée de ce que représente une RMSE en recommandation automatique. Le moteur utilisé sur *Netflix* lors du lancement du challenge en 2006, appelé *Cinematch*, avait alors une RMSE 0,9525. L'objectif de ce challenge était d'améliorer ce score de 10%, soit obtenir une RMSE proche de 0,85. Ce score a été atteint en juillet 2009, soit trois ans après le lancement du concours, par une alliance de trois des meilleures équipes. La RMSE obtenue alors était de 0,8567, soit 10,06% de mieux que le score initial. L'obtention de ce score, et par conséquent le gain du million de dollars mis en jeu, a été obtenu à l'aide d'une combinaison de plus de 250 méthodes de prédiction, certaines d'entre elles étant paramétrées. On peut difficilement imaginer qu'une telle méthode puisse être déployée dans un contexte industriel, néanmoins, ce chiffre de 0,85 permet de se faire une idée des limites probables de la recommandation automatique.

Parallèlement à cela, l'évolution des résultats obtenus par les participants au cours des trois années d'existence du challenge est intéressante. Au bout d'un an, en 2007, la

meilleure RMSE obtenue était de 0,8723. La première année a donc permis l'obtention d'un gain assez important avec une baisse de la RMSE de 0,08. En 2008, le meilleur score obtenu était de 0,8627 et en 2009 il a donc atteint le seuil exigé avec un score de 0,8567. L'amélioration des performances a donc été beaucoup plus faible sur les deux dernières années que sur la première année et celle-ci a demandé la construction de systèmes très complexes et très gourmands en terme de puissance de calcul. Il y a donc un seuil, situé autour de 0,86-0,87, exigeant d'énormes efforts pour être franchi. Le tableau 2.4 récapitule les résultats du challenge *Netflix* (le tableau complet est disponible en annexe dans la section D).

Date	RMSE
2006	0,9525
2007	0,8723
2008	0,8627
2009	0,8567

TABLE 2.4 – Principaux résultats du challenge *Netflix*

Il est évident que les travaux de cette thèse n'ont pas pour objectif de se mesurer aux performances réalisées par les participants au challenge, mais leur connaissance permet tout de même de situer nos résultats. Il est également nécessaire de préciser que ces résultats ont été obtenus à l'aide de données d'apprentissage et de test en provenance d'une même source, les clients du site *Netflix*. On peut donc supposer qu'elles possèdent beaucoup de caractéristiques communes, ce qui est toujours un avantage dans une tâche de prédiction.

2.2.4.2 La MAE

La MAE est également une mesure d'erreur qui permet de calculer la moyenne des erreurs absolues entre les notes prédites et les vraies notes. Une erreur absolue mesure l'imprécision entre une note prédite et la note réelle correspondante par la formule suivante :

$$e_{ui} = |p_{ui} - n_{ui}|$$

Où n_{ui} est la n donnée par l'utilisateur u sur l'item i et p_{ui} la note prédite. La MAE se calcule alors de la manière suivante :

$$MAE = \frac{\sum_{u,i} |p_{ui} - n_{ui}|}{n} = \frac{\sum_{ui} e_{ui}}{n}$$

Où n représente le nombre total de notes prédites.

2.2.4.3 Les autres mesures

D'autres mesures peuvent également être utilisées bien que ce soit plus rare. Nous pouvons par exemple citer les mesures de précision et de rappel. La précision mesure la pertinence des recommandations émises. Une recommandation pertinente étant une recommandation faite sur un produit déjà noté positivement par l'utilisateur dans l'ensemble de test. La formule est la suivante :

$$Precision = \frac{|\{Recommandations\ pertinentes\} \cap \{Recommandations\ émises\}|}{|\{Recommandations\ émises\}|}$$

Le rappel mesure de son côté la capacité du système à faire des recommandations pertinentes à l'aide de la formule suivante :

$$Rappel = \frac{|\{Recommandations\ pertinentes\} \cap \{Recommandations\ émises\}|}{|\{Recommandations\ pertinentes\}|}$$

Il existe également des mesures basées sur les courbes de ROC [SPUP02]. La courbe de ROC (Receiver Operating Characteristic) est une mesure de la performance d'un classifieur binaire. Graphiquement, on représente souvent la mesure ROC sous la forme d'une courbe qui donne le taux de classifications correctes dans un groupe (dit taux de vrais positifs) en fonction du nombre de classifications incorrectes (taux de faux positifs) pour ce même groupe.

2.3 Conclusion

Dans ce chapitre, nous avons décrit les différentes méthodes de recommandation existantes. Elles se décomposent en deux catégories, les méthodes personnalisées, basées sur des profils utilisateurs, et des méthodes contextuelles basées sur les items consultés. Dans cette thèse nous nous intéressons uniquement à la recommandation personnalisée. Nous avons pu voir que les techniques existantes sont de deux natures différentes appelées « filtrages ». Il existe tout d'abord le filtrage collaboratif, qui consiste à recommander les items appréciés par des personnes aux goûts similaires, et le filtrage thématique qui consiste à trouver les items proches à l'aide de données descriptives. Le tableau 2.5 récapitule les principaux avantages et inconvénients de chacune de ces méthodes.

La recommandation personnalisée est un domaine très actif qui intéresse énormément de laboratoires de recherche mais également les industriels. Preuve en est, l'entreprise *Netflix*, premier loueur de DVD aux États-Unis, a lancé en 2006 un challenge international doté d'une prime au vainqueur d'un million de dollars afin d'améliorer son système de recommandation personnalisée de films d'au moins 10%. Notamment grâce à ce challenge, les performances des systèmes, en termes de taux de bonnes prédictions, ont atteint une telle qualité qu'il est devenu de plus en plus difficile de les améliorer, ne serait-ce que de quelques dixièmes de pourcentage. Cependant, un problème reste récurrent à tous

	Filtrage Thématique	Filtrage Collaboratif
Données nécessaires	Données descriptives	Données d'usages
Avantages	- Nécessite relativement peu de données - Pratique pour les items à courtes durées de vie	- Méthode la plus performante
Inconvénients	- Moins performant que le filtrage collaboratif - Sujet au problème de l' « entonnoir »	- Nécessite une grande quantité de données - Sujet au manque de connaissances (cold-start, longue traine)

TABLE 2.5 – Récapitulatif des avantages et inconvénients des deux types de filtrages

ces systèmes. Il s'agit du problème du démarrage à froid (Cold Start). Ce problème se présente par exemple lors du lancement d'un nouveau service, lorsque le système possède trop peu d'informations sur les produits et/ou les utilisateurs. Quel que soit le moteur de recommandation utilisé, s'il ne possède pas assez d'informations en entrée, il ne pourra pas produire de bons résultats en sortie. C'est sur l'apport de solutions à ce problème que cette thèse se positionne, l'objectif étant d'acquérir l'information manquante à l'aide de textes communautaires présents en grand nombre sur le Web.

Nous avons vu que les textes libres sont des données déjà utilisées en filtrage thématique. Une évaluation de nos données avec cette approche a donc été nécessaire. Avant de présenter les résultats obtenus, nous présentons l'outil de recommandation qui a servi aux expérimentations ainsi que les données textuelles utilisées. Nous n'avons par contre trouvé aucune trace dans la littérature d'expérimentations portant sur des méthodes de filtrage collaboratif basées sur des données textuelles. Les données d'entrée nécessaires au filtrage collaboratif sont des données d'usages de la forme utilisateur-item-note. Ces données peuvent être extraites de textes à l'aide de la classification d'opinion. Nous évaluerons également cette approche afin de vérifier si cette méthode peut concurrencer l'approche « classique »basée sur le filtrage thématique. Deux chemins s'offrent donc à nous afin d'établir des recommandations personnalisées à partir de textes communautaires.

Chapitre 3

Présentation du moteur

Sommaire

Ce chapitre présente le système de recommandation utilisé pour nos expérimentations. Ce moteur possède l'avantage de fonctionner en mode collaboratif ou en mode thématique, au choix de l'utilisateur. Les textes peuvent être pourvoyeurs de données descriptives mais peuvent également fournir des données d'usages à l'aide de la classification d'opinion. Ce moteur proposant les deux filtrages, collaboratif et thématique, va permettre d'évaluer les quantité et qualité de l'information présente dans les textes communautaires. Il permettra également d'évaluer quelle extraction de données il faut privilégier : les descripteurs ou les données d'usages.

Ce chapitre présente tout d'abord le fonctionnement du moteur puis énumère les résultats de différentes évaluations permettant de situer ses performances dans l'état de l'art et de fixer les limites des capacités du moteur.

3.1 Fonctionnement

Le moteur de recommandation qui est utilisé pour les expérimentations est issu de travaux menés à Orange Labs dans le cadre d'un projet de recherche appelé Reperio. L'architecture du moteur, son fonctionnement détaillé et son interface sont décrits par Humbert [Hum09]. Nous ne rappelons ici que quelques principes de fonctionnement : alimentation par les données d'usages et descriptives, utilisation pour la prédiction de la note.

3.1.1 Principe

Le moteur de recommandation utilisé pour les expérimentations possède l'avantage de fonctionner soit en mode collaboratif soit en mode thématique selon les besoins de l'utilisateur. Les recommandations se font en deux étapes. Une première étape consiste tout d'abord à calculer les similarités entre tous les items deux à deux. Ces similarités peuvent être calculées à partir d'une matrice d'usages en mode collaboratif, ou à partir de descripteurs en mode thématique. Une fois calculées, ces similarités forment une matrice dite Items-Items. À partir de cette matrice, une deuxième étape consiste à prédire les appréciations des utilisateurs sur les items sous forme de notes. Ces deux étapes peuvent être effectuées indépendamment ce qui offre la possibilité de produire des matrices Items-Items sans passer par la première étape du moteur (la construction de la matrice Items-Items). On peut alors appliquer uniquement la deuxième étape (la prédiction des notes) et ainsi évaluer la qualité des matrices de similarités produites. L'évaluation s'effectue après ces deux étapes à l'aide d'une matrice d'usages test. Les prédictions de notes sont faites sur des utilisateurs ayant déjà noté un certain nombre d'items. Une partie de ces notes (90%) est utilisée comme profils utilisateurs et permet de personnaliser les recommandations. L'autre partie (10%) sert à évaluer la qualité des notes prédites. La figure 3.1 décrit le moteur avec les deux étapes qui le composent ainsi que l'évaluation finale.

3.1.2 Première étape : construction des matrices de similarités

Le choix de construire des matrices de similarités Items-Items est basé sur plusieurs avantages qu'apportent ce type de méthodes, notamment dans un contexte industriel [Mey11] :

- Ce format permet tout d'abord de prédire la note qu'un utilisateur u pourrait attribuer à un item i ce qui offre la possibilité d'ordonnancer la liste d'items à recommander par ordre d'intérêt ou de pertinence ;
- La matrice de similarités permet également de faire de la recommandation non personnalisée d'items. Il s'agit de la recommandation item-to-item contextuelle classique, popularisée par le site de vente en ligne Amazon. La matrice de similarité permet de disposer immédiatement de cette fonctionnalité ;
- Enfin, les modèles de recommandation basés sur la construction de matrices Items-Items sont, à notre connaissance, les modèles procurant le maximum d'avantages en termes de couverture fonctionnelle (ces modèles n'étant pas du tout paramétrés,

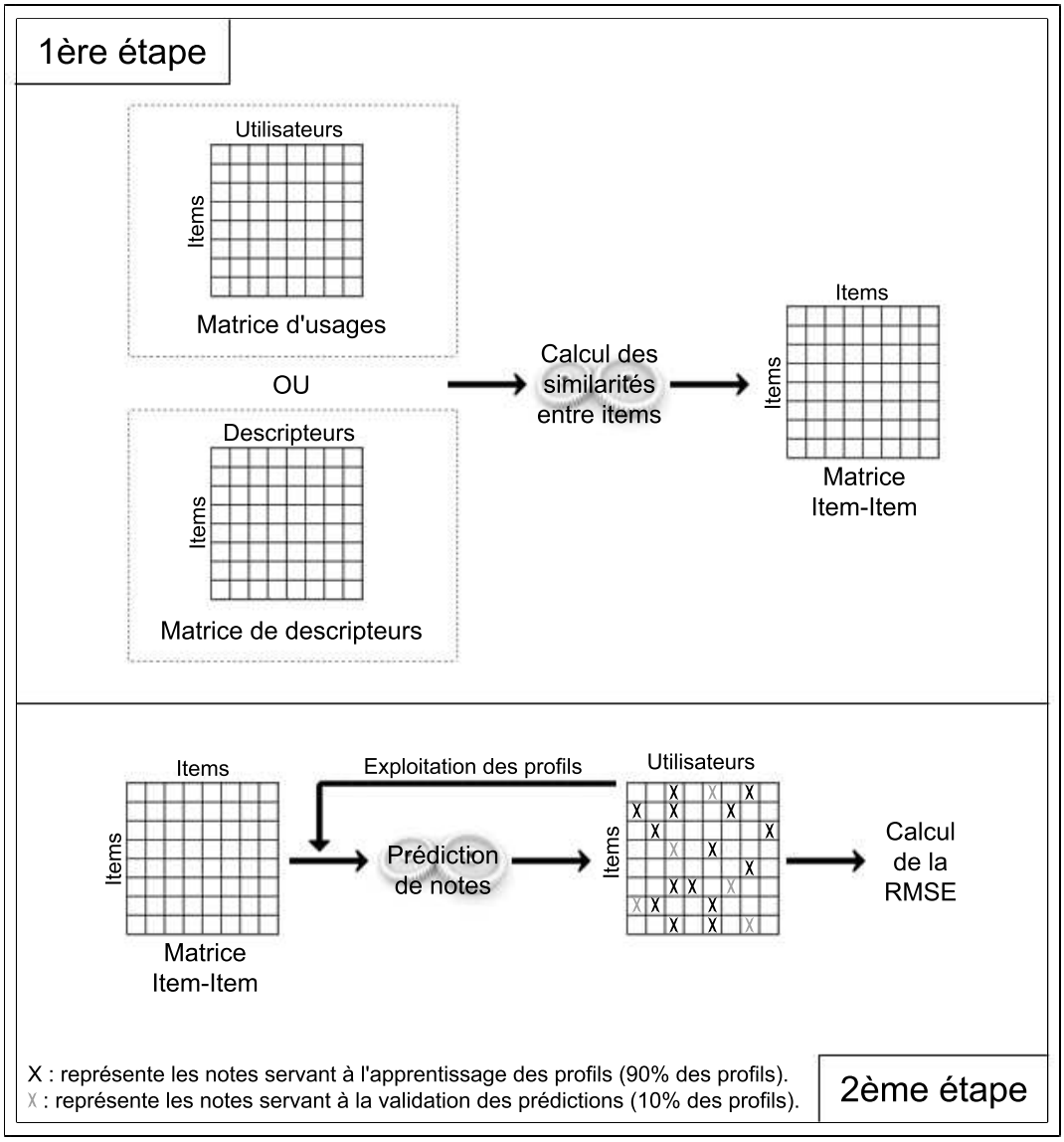


FIGURE 3.1 – Schéma de fonctionnement du moteur de recommandation

ils peuvent être appliqués sur tout type de données) et de qualité prédictive, mais également en termes de tenue de charge et de transparence (on sait que telle recommandation est faite car tels items ont été appréciés par l'utilisateur) [LSY03, Kor10].

Pour le moteur utilisé, les matrices Items-Items sont construites à l'aide de deux mesures de distance à choisir suivant les données d'entrée. Une extension de la fonction de similarité de Pearson est utilisée lorsque le moteur est lancé en mode collaboratif et que les données d'entrée sont des notations. L'extension en question concerne la prise en compte, au niveau du dénominateur, de l'intégralité des notes sur les items et non seulement leur

intersection comme c'est le cas dans la fonction originale. La formule utilisée est la suivante :

$$Pear(i,j) = \frac{\sum_{\{u \in S_i \cap S_j\}} (r_{iu} - \overline{r_i}) \times (r_{ju} - \overline{r_j})}{\sqrt{\sum_{\{u \in S_i \cup S_j\}} (r_{iu} - \overline{r_i})^2 \sum_{\{u \in S_i \cup S_j\}} (r_{ju} - \overline{r_j})^2}}$$

Où S_i (respectivement S_j) est l'ensemble des utilisateurs ayant noté l'item i (respectivement j), r_{iu} (respectivement r_{ju}) la note donnée par l'utilisateur u sur l'item i (respectivement j), et $\overline{r_i}$ (respectivement $\overline{r_j}$) la moyenne des notes obtenues par i (respectivement j).

Dans le cas du filtrage collaboratif avec des données binaires (a acheté/n'a pas acheté) ou dans le cas du filtrage thématique, c'est la distance de Jaccard qui est employée :

$$Jacc(i,j) = \frac{|\{S_i \cap S_j\}|}{|\{S_i \cup S_j\}|}$$

Où S_i (respectivement S_j) est l'ensemble des variables décrivant l'item i (respectivement j).

Ces mesures de similarité ont été préférées, après différentes évaluations, aux autres mesures populaires dans le domaine comme la distance du Cosinus [CMB07].

3.1.3 Deuxième étape : prédiction des notes

Une fois la matrice de similarités Items-Items construite, les recommandations sont établies à l'aide des notes connues des utilisateurs, ces notes constituant le profil. La fonction de prédiction de la note de l'item i pour l'utilisateur u est la suivante :

$$p_{ui} = \overline{r_i} + \frac{\sum_{\{j \in S_u\}} sim(i,j) \times (r_{uj} - \overline{r_j})}{\sum_{\{j \in S_u\}} sim(i,j)}$$

Où S_u est l'ensemble des notes connues données par l'utilisateur u. Il est à noter que cette fonction de prédiction suppose la connaissance des notes moyennes sur les items des autres utilisateurs ($\overline{r_i}$ et $\overline{r_j}$). Cette fonction, nommée Mean-Based est bien entendu exploitable sans la prise en compte de ces moyennes par la formule suivante, appelée Not-Mean-Based :

$$p_{ui} = \frac{\sum_{\{j \in S_u\}} sim(i,j) \times (r_{uj})}{\sum_{\{j \in S_u\}} sim(i,j)}$$

Les résultats sont toutefois bien meilleurs avec la fonction Mean-Based, c'est pourquoi celle-ci a été préférée.

3.1.4 Évalutation

La métrique d'évaluation utilisée et la RMSE. Afin de la calculer, les prédictions sont établies sur une matrice d'usages test contenant déjà des notes données par des utilisateurs réels. La RMSE permet de vérifier la qualité des notes prédites en mesurant l'écart moyen entre les notes prédites et les notes réelles données par les utilisateurs :

$$RMSE = \sqrt{\frac{\sum_{u,i} (p_{ui} - n_{ui})^2}{n}}$$

Cette mesure a été privilégiée afin de positionner les résultats avec ceux obtenus lors du challenge Netflix.

3.2 Étalonnage avec des données connues

La recommandation de films est un domaine très étudié qui sert souvent de référence en recommandation automatique. Les données disponibles sur ce genre d'items sont en effet très nombreuses et variées. Les données d'usages, les commentaires ou encore les forums portant sur les films sont très présents sur le Web. On trouve également facilement des données descriptives très riches telles que celles proposées par l'*Internet Movie Database*[1] qui contient des caractéristiques comme les acteurs, réalisateurs, sociétés de productions, etc. ainsi que de très nombreux « tags » (étiquettes). Nous avons donc étalonné le moteur avec des données connues et disponibles à tous :

- Pour le mode collaboratif, nous avons utilisé un corpus de notes proposé par la société *Netflix* ;
- Pour le mode thématique, nous avons extraits des descripteurs du site *IMDb*.

3.2.1 Évaluation du mode collaboratif

Afin de situer nos résultats dans l'état de l'art, nous avons sélectionné un corpus d'évaluation connu dans le domaine de la recommandation. Il s'agit du corpus d'apprentissage mis à disposition par la société *Netflix* lors du lancement de son challenge. Pour rappel, ce challenge lancé en 2006 avait comme objectif d'améliorer de 10% les résultats du moteur de recommandation utilisé jusqu'alors sur le site de location de DVD [BL07]. Une sélection de cette grande quantité de données a été faite afin d'obtenir le corpus de notes.

3.2.1.1 Description des données collaboratives

Le corpus initial est composé de 100 millions de logs qui concernent environ 480 000 utilisateurs anonymes et un peu moins de 18 000 films (17 770 exactement). Les dates

1. www.imdb.com

des logs sont également présentes. Chaque log peut donc être interprété de la manière suivante : sur le site de location de DVD *Netflix*, l'utilisateur u a attribué la note n à l'item i le jour j. Les notes en question représentent un nombre d'étoiles et sont comprises entre 1 et 5, 1 étoile signifiant que l'utilisateur a détesté le film, et 5 étoiles signifiant qu'il l'a énormément apprécié. La quantité de données étant conséquente, seule une partie du corpus a été conservée pour les expérimentations (environ 5%). La figure 3.2 décrit la façon dont le corpus a été découpé.

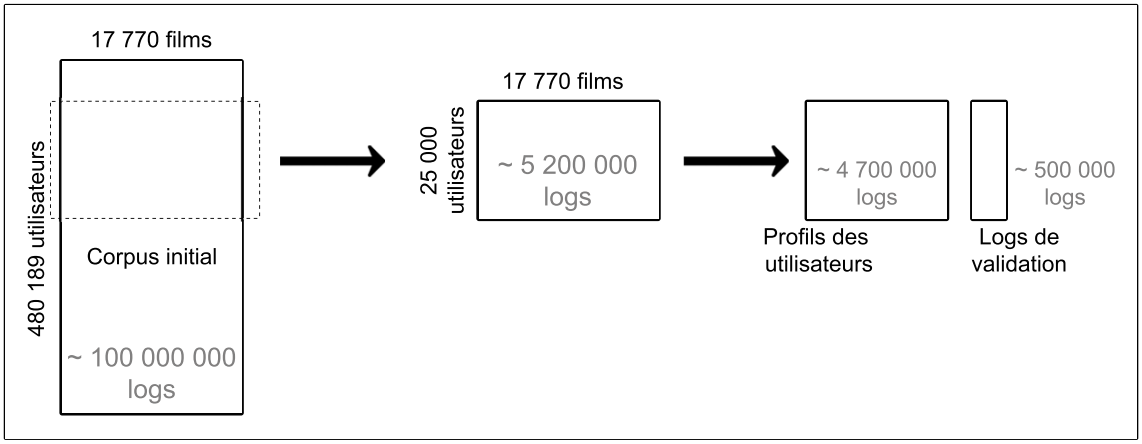


FIGURE 3.2 – Séléction des corpus extraits des données *Netflix*

25 000 utilisateurs ont été sélectionnés au hasard parmi les 480 000 présents dans la base. Cela représente environ 5 200 000 logs, soit une moyenne de logs par utilisateur supérieure à 200. La nouvelle base de données a été séparée en deux parties afin d'obtenir deux corpus distincts. Le premier corpus contient 90% du profil de chaque utilisateur, soit 4 700 000 logs, et est utilisé pour l'apprentissage des profils des utilisateurs à qui l'on souhaite faire des recommandations. C'est également ce corpus qui est utilisé pour construire la matrice Items-Items. Le deuxième corpus, composé des 10% restants de chaque profil utilisateur, soit environ 500 000 logs, a pour utilité de vérifier la qualité des recommandations émises à ces mêmes utilisateurs. L'échantillonnage 90/10 des logs de chaque utilisateur a été fait aléatoirement. Pour plus de facilité, nous appellerons le premier échantillon (les 90%) l'ensemble d'apprentissage des profils, et le deuxième échantillon (les 10%) l'ensemble de test des profils. Ces profils forment l'ensemble de validation de toutes les expérimentations portant sur la recommandation. Le tableau 3.1 récapitule les chiffres précédemment cités.

3.2.1.2 Résultat de l'étalonnage du filtrage collaboratif

Pour cet étalonnage, nous nous sommes placé dans les conditions réelles d'un déploiement industriel par filtrage collaboratif. Dans ce cas, les profils utilisateurs à partir desquels la matrice de similarités Items-Items est construite sont les mêmes utilisateurs qui reçoivent

Nombre d'utilisateurs	25 000 (5% du nombre original)
Nombre d'items	17 770 (100% du nombre original)
Nombre de logs	5 224 519
Ensemble d'apprentissage des profils	4 707 540 logs (soit 90%)
Ensemble de test des profils	516 979 logs (soit 10%)
Taille moyenne des profils en apprentissage	188 items notés
Taille moyenne des profils en test	20 items notés

TABLE 3.1 – Caractéristiques des données *Netflix*

les recommandations en sortie de chaîne, l'objectif étant de prédire les notes manquantes dans la matrice. L'apprentissage de la matrice Items-Items est donc fait à l'aide de l'ensemble d'apprentissage des profils. La RMSE obtenue sur l'ensemble de test des profils lors de cet étalonnage est de 0,862. Ce résultat correspond à ce qui se fait dans l'état de l'art à objectif équivalent.

3.2.2 Évaluation du mode thématique

Les descripteurs de films choisis pour cette deuxième évaluation sont des données textuelles structurées extraites du site *IMDb* (*Internet Movie Database*). *IMDb* est une base de données en ligne, accessible à tous, qui regroupe un grand nombre d'informations sur les films, les séries télévisées et les jeux vidéo. Il existe plusieurs bases de ce type sur le Web mais *IMDb* est connu pour être le site répertoriant la plus grosse quantité d'informations sur le plus grand nombre de contenus (plus d'un million et demi de titres sont répertoriés dans la base en 2010). Concernant les films, le site contient des informations telles que les synopsis, les acteurs, les réalisateurs, etc. mais également des données plus difficiles à trouver comme des informations budgétaires, les récompenses et palmarès, des informations techniques (durée, ratio d'écran, largeur de pellicule, procédé couleur, etc.), les dates de sorties en salle ou en DVD selon les pays, des informations sur toutes les entreprises intervenant dans l'élaboration du film, etc. On trouve également un grand nombre de mots clés (tags) qui sont attribués par les utilisateurs enregistrés. Les informations ajoutées par les utilisateurs sont vérifiées par une équipe de modérateurs avant d'être ajoutées publiquement sur le site afin d'être assuré que les données ne contiennent aucun bruit.

3.2.2.1 Description des données thématiques

Les identifiants des films présents dans les challenges *Netflix* et dans l'*IMDb* ne sont pas communs. Il a donc été nécessaire de réaliser une jointure entre les données communes connues pour chaque film, soit le titre et la date de sortie. Seuls 15 953 films parmi les 17 770 présents dans le corpus *Netflix* ont été retrouvés. Ceci peut s'expliquer par le fait que la base *Netflix* contient des DVD qui ne concernent pas toujours les films, comme des concerts ou des documentaires par exemple. La raison peut également être que les

dates ne correspondent pas (dû à plusieurs éditions du même film en DVD par exemple). Les titres ont été comparés à l'aide de la distance de Levenshtein après suppression des caractères non alphanumériques. Si A et B sont deux mots, la distance de Levenshtein $d(A, B)$ correspond au nombre minimal de remplacements, ajouts et suppressions de lettres pour passer du mot A au mot B. Une fois la jointure faite sur les titres, les dates ont été comparées afin de ne pas accepter les mêmes noms de films avec des dates différant de plus d'une année. Le taux d'erreur de la jointure a été estimé par sondage à 6%. Les données extraites d'*IMDb* décrivant les films sont répertoriées dans le tableau 3.2. Le nombre global de descripteurs s'élève à 861 507.

Types de descripteurs
Acteurs
Réalisateurs
Producteurs
Genres
Nationalités des films
Années de production
Scénaristes
Compagnies
Langue
Mots clés

TABLE 3.2 – Descripteurs extraits d'*IMDb*

3.2.2.2 Résultat de l'étalonnage du filtrage thématique

Les descripteurs ont tout d'abord été évalués indépendamment les uns des autres. Les mots-clés sont les descripteurs qui apportent les meilleurs résultats, suivis par les acteurs et le genre. Toutefois, la meilleure performance est obtenue par la sélection de tous les attributs avec une RMSE de 0,921.

3.3 Conclusion

Ce chapitre décrit le fonctionnement du moteur de recommandation qui va être utilisé lors des expérimentations à suivre. Ce moteur permet d'effectuer les deux types de filtrages, le collaboratif et le thématique, grâce à une approche basée sur la construction d'une matrice de similarités Items-Items. Du fait de cette méthode, le moteur offre la possibilité d'établir des recommandations à partir de matrices de similarités construites sans passer par le moteur. L'un des objectifs étant de prouver que l'apport de données textuelles non structurées provenant de sites communautaires peut être une alternative au filtrage thématique, nous avons mesuré les performances de cette technique sur le corpus d'évaluation. Pour cela, nous avons utilisé les données du site *IMDb*. Ces données sont composées, pour chaque film, des acteurs, réalisateurs, producteurs, genres, nationalité,

année de production, scénaristes, compagnies, langue et de centaines de mots clés. Il s'agit par conséquent de données très riches, qui ne pourraient pas nécessairement être obtenues pour des items autres que des films.

La RMSE obtenue par cette méthode est de 0,921. Nous avons également souhaité étalonner le système sur des données d'usages (filtrage collaboratif). Pour ce faire, nous avons utilisé une partie du corpus du challenge *Netflix*. Nous avons donc affaire à de véritables données d'usages du type de celles utilisées dans le monde industriel. La RMSE obtenue avec cette méthode est de 0,862, soit dans la moyenne de ce qu'on peut voir dans l'état de l'art. Le tableau 3.3 récapitule les résultats obtenus lors de cet étalonnage qui nous serviront de point de repère pour les prochaines expérimentations. Ce moteur va ainsi nous permettre d'évaluer les informations extraites des textes communautaires.

Filtrage collaboratif (Données *Netflix*)	Filtrage thématique (Données *IMDb*)
0,862	0,921

TABLE 3.3 – Évaluation du système de recommandation sur deux ensembles de données connus

Chapitre 4

Données expérimentales

Sommaire

Pour faire suite à la présentation du moteur de recommandation, ce chapitre présente maintenant les données expérimentales. L'évaluation de la chaîne de traitements mise en place requiert un corpus composé de textes subjectifs portant sur des items identifiés. C'est sur ces données d'entrée que seront apprises les informations nécessaires à la recommandation de contenus. Comme nous voulons évaluer à la fois le filtrage thématique et le filtrage collaboratif à partir des mêmes données d'entrée afin de comparer les performances des deux approches, le corpus doit être composé de triplets utilisateur-item-texte, l'utilisateur étant l'auteur du texte, et l'item le sujet de la discussion. Ces triplets permettront ainsi d'extraire aussi bien des données d'usages (triplets utilisateur-item-note) que des données descriptives (couples item-descripteur). Dans le chapitre précédent, le moteur de recommandation a été étalonné sur le domaine des films. Afin de pouvoir comparer tous les résultats entre eux, c'est donc ce domaine qui a été conservé pour le choix des données textuelles. Nous avons ainsi extrait des commentaires sur un site communautaire, nommé *Flixster*, regroupant un grand nombre de cinéphiles amateurs, anglophones pour la grande majorité. Ces commentaires sont des textes libres qui possèdent certaines particularités propres aux textes communautaires.

Le chapitre est organisé de la façon suivante. Dans un premier temps, nous décrivons ce qui rend les textes issus de sites communautaires particuliers par rapport aux textes « normaux » et en quoi ces caractéristiques peuvent compliquer les analyses. Nous présentons ensuite plus spécifiquement les données du corpus de textes ainsi que son acquisition. Pour finir, nous présentons des observations découlant d'une analyse exploratoire menée dans le but de mesurer la quantité d'information contenue dans les données.

4.1 Spécificités des textes communautaires

Ce que nous nommons ici « textes communautaires », sont les textes rédigés par les internautes par le biais d'espaces de communication tels que les réseaux sociaux, les blogs, les forums, etc. Les textes présents sur ces sites sont écrits par des amateurs pour des amateurs. En cela, ils peuvent répondre à des règles particulières à l'inverse des « textes professionnels » (écrits par des journalistes, chroniqueurs, etc.) qui respectent les règles fondamentales de grammaire et d'orthographe (ou qui sont, tout du moins, censés le faire). Ce type de textes a fait l'objet de travaux et d'études de la part de la communauté linguistique qui considère cette écriture comme un langage, ou plus précisément un dialecte, à part entière :

« Quand l'ordinateur est utilisé pour le courriel, les forums de discussion et les chats, en tant qu'outil permettant la communication entre individus, il devient un véritable *médiateur* ; son utilisation modifie notre discours et ainsi notre façon de communiquer avec autrui. Émerge alors un nouveau genre de discours, le *discours électronique médié* (DEM). Le DEM contient des marques linguistiques et extra-linguistiques qui lui sont propres et il entre dans le cadre plus global de la *communication médiée par ordinateur* (CMO). » [Pan06]

À l'instar de « discours électronique médié », plusieurs dénominations ont vu le jour afin de désigner ce nouveau type d'écriture : communication écrite médiatisée par ordinateur [Mar00], nouvelles formes de communication écrite [GDNV04], communication éléctronique [AdFF04], le parler Net (Netspeak) [Cry01], ou encore la cyberl@ngue ou cyberlangage [DM02]. Nous concernant, nous conserverons le sigle DEM pour désigner ce type d'écriture.

Les origines du DEM peuvent s'expliquer de différentes manières. Le Web communautaire, et la popularisation du Web en général, a été devancé de peu par la démocratisation des téléphones portables et l'utilisation massive d'un nouveau moyen de communication : le SMS [1]. Le SMS est un système de communication, surtout présent sur les téléphones portables, qui permet aux utilisateurs d'échanger des messages écrits de taille limitée (autour de 160 caractères, espaces compris). Du fait de cette taille limitée, les utilisateurs ont cherché (et trouvé) des astuces permettant d'insérer un maximum d'informations dans cet espace réduit telles que la suppression de la ponctuation, les abréviations, la suppression d'espaces, etc. Bien que les outils de communication disponibles sur Internet (blogs, forums, courriel, etc.) ne sont pas limités en taille, les habitudes prises par les utilisateurs du SMS ont perduré et sont maintenant lisibles (ou visibles seulement, parfois) sur la Toile. Le DEM peut également être dû à un désir de personnalisation du discours et surtout un désir de transmettre des informations telles que des émotions, des sensations, un état d'esprit, etc. qui ne sont pas toujours aisées à transcrire. Dans une lettre manuscrite, la personnalisation est plus facile. D'une part parce que chacun possède une écriture particulière, et

1. « Short Message Service » ou « Service de Messages Succints »

d'autre part, le support laisse beaucoup plus de libertés. Par le biais d'un ordinateur, les caractères sont limités et communs à tous. L'utilisateur va donc chercher des assemblages de ces caractères afin de personnaliser son discours et être ainsi reconnaissable. Cette personnalisation devant être comprise par tout le monde, on observe toutefois une certaine harmonisation des styles, et certaines règles, propres au DEM, semblent s'être établies au fil du temps. Quelques règles semblent donc être devenues universelles, comme certaines abréviations (« lol », « mdr »), alors que d'autres ne semblent pas figées. Elles peuvent évoluer dans le temps mais également selon les personnes ou, plus exactement, les groupes de personnes :

« Le cyberlangage a ceci de particulier qu'il n'est pas statique : il bouge, il vit au gré de l'imagination des internautes. L'écrit virtuel complète le français de façon subtile : une sorte de mélange entre l'oral et l'écrit, un style oratoire bousculé par la vitesse, chamboulé dans ses règles et ses conventions. Si la Toile semble signer le retour en force de la langue écrite, elle traduit surtout un phénomène social plus que linguistique. Une envie, voire un besoin de communiquer, mais autrement. » [DM02]

Malgré cet aspect changeant du DEM, nous pouvons citer plusieurs phénomènes non exhaustifs le caractérisant de manière significative [Pan09] :

- Les didascalies électroniques ;
- Les erreurs ou « ratages » ;
- La néographie.

4.1.1 Les didascalies électroniques

Les didascalies électroniques sont utilisées afin de traduire les émotions ressenties par l'auteur telles que la joie, l'excitation, la tristesse, etc. L'exemple le plus représentatif de ce phénomène est l'utilisation d'« émoticônes » ou de « smileys » (« ;-) », « :-(»). D'autres possibilités sont également utilisées comme les abus de majuscules (« NON »), les étirements de mots (« Géniaaaaaaaallll ») ou encore les répétitions de ponctuations (« ? ? ? ? ? ? ? »). Il peut également s'agir d'onomatopées comme « mouarf » ou « bof ».

4.1.2 Les erreurs

Les erreurs peuvent être orthographiques, typographiques ou grammaticales et sont également très fréquentes. Par « erreurs », on entend « maladresses involontaires » à ne pas confondre avec les fautes faites intentionnellement. Ces erreurs peuvent être de deux types. Il peut s'agir d'erreurs dues à l'utilisation d'un outil (fautes de frappe, accents manquants, etc.) ou d'erreurs résultant d'une méconnaissance des règles.

4.1.3 La néographie

La néographie est également très usitée. Celle-ci désigne une nouvelle manière d'écrire une langue à l'aide d'une orthographe qui diffère de la norme. Dans le cas du DEM, la néographie apparaît sous différents aspects dont voici les plus « simples » :

Les substitutions consistent à remplacer certaines lettres par de nouvelles tout en conservant la prononciation originale. Ces substitutions peuvent se faire sur plusieurs mots, comme quand on écrit « g » en lieu et place de « j'ai », ou sur un mot entier comme par exemple « o » pour « au » ou « 7 » pour « cette ». Elles peuvent également être inclues à l'intérieur d'un mot comme dans l'exemple « mwa ossi » (« moi aussi »). Les substitutions peuvent également se faire à l'aide de caractères autres qu'alphanumériques. C'est par exemple le cas avec le sigle « @+ » souvent utilisé pour signifier « à plus (tard) ».

Les réductions consistent à supprimer un certain nombre de caractères dans un mot sans en altérer la compréhension. Elles peuvent être utilisées lorsque l'utilisateur désire rédiger rapidement ou obtenir un gain de place, mais peuvent également permettre de simplifier le message. On retrouve dans cette catégorie les abréviations existant également à l'oral comme les apocopes (« télé » pour « télévision », « ciné » pour « cinéma ») et les aphérèses (« bus » pour « autobus », « ricain » pour « américain »). Il peut également s'agir de sigles comme « mdr » pour « mort de rire » ou « lol » pour « laughing out loud » ou « lots of laughs ». D'autres réductions peuvent être uniquement graphiques comme les suppressions de fin de mot (« je veu » pour « je veux »), les squelettes consonantiques (« dsl » pour « désolé »), les consonnes doubles (« poura » pour « pourra »), ou encore les aglutinations (« jattends » pour « j'attends »). Les réductions peuvent bien entendu être multiples (« jaten »).

Les suppressions consistent à supprimer des portions du texte. Il peut s'agir de la ponctuation (« salut ça va » pour « salut, ça va ? ») ou du caractère espace (« cavabien » pour « ça va bien »). Le dernier exemple contient également une suppression de signe diacritique (« ca » au lieu de « ça »). La suppression des accents entre également dans cette catégorie.

Les ajouts peuvent correspondre à des ajouts de lettres permettant de modifier la prononciation du mot (« oki » pour « ok »). Les ajouts sont généralement utilisés afin de renforcer l'aspect stylistique qui peut avoir son importance, notamment chez les jeunes utilisateurs (15-18 ans) [Fay07].

Ces différentes caractéristiques peuvent bien évidemment être combinées et entrainer des phénomènes beaucoup plus complexes à décrire et à analyser (comme par exemple « kestufé » en lieu et place de la phrase « qu'est-ce que tu fais ? »). Elles sont également associées à des phénomènes moins visibles mais qui ont leur importance en linguistique comme un pourcentage moins important de verbes, une raréfaction des pronoms personnels, etc. Panckhurst [Pan09], à qui nous avons emprunté certains des exemples précédents, propose une discussion très détaillée sur le sujet. Certains termes, non présents dans le dictionnaire, comme des mots appartenant au langage familier ou au verlan, peuvent également être employés.

4.1.4 Conséquences

Toutes ces caractéristiques propres au DEM ont bien entendu des conséquences sur les analyses de ce type de textes. Les caractéristiques de la langue bien écrite entraînent déjà un grand nombre de problématiques en analyse linguistique ou en fouille de textes, que ce soit pour les tâches de correction orthographique, d'analyses sémantique ou syntaxique ou encore pour la classification de textes. Tout d'abord, le caractère évolutif du DEM complique les traitements. Le fait que le DEM soit une langue changeante dans le temps complique par exemple la construction des dictionnaires comme ceux nécessaires à la correction orthographique. La construction de dictionnaires de synonymes, d'ontologies ou de thésaurus devient également plus difficile avec un nombre de termes à répertorier et à lier entre eux beaucoup plus élevé (conséquence direct des néographies notamment). De plus, le nombre de mots existants étant plus élevé, le nombre d'occurrences de chacun d'entre eux ne peut qu'être plus faible. Pour donner un exemple précis, la présence d'onomatopées est très informative, spécialement lorsqu'il est question d'étude des opinions, mais leur multiplicité, autant phonétique que graphique, rend leur analyse délicate. En effet, comme un grand nombre de choix est possible pour l'auteur du texte, chaque terme n'en devient que plus rare et cela complique l'interprétation par la machine qui a besoin soit d'exemples, soit de règles produites manuellement pour travailler.

4.2 Corpus de textes (*Flixster*)

4.2.1 Présentation

Les données textuelles utilisées pour nos expérimentations sont des commentaires en langue anglaise extraits du site communautaire *Flixster*[2]. Ce site est un site communautaire dédié aux amateurs de cinéma. Il permet à ses utilisateurs de créer et d'alimenter des pages personnelles (blogs) où ils peuvent partager leurs opinions via des boîtes à réactions, des forums, des listes de favoris, etc. Les utilisateurs peuvent également se déclarer « ami » avec d'autres utilisateurs et se créer ainsi un réseau social virtuel. Ils ont également la possibilité de commenter et de noter chaque film indépendamment, et ceci une fois seulement par film. Ainsi, chaque utilisateur actif présent sur le site est relié à un certain nombre de films par un unique commentaire. De plus, ce commentaire est généralement associé à une note postée par l'utilisateur lors de sa rédaction. Cette note, comprise entre 0,5 et 5 (soit 10 possibilités), est censée résumer l'opinion portée sur le film en question, le commentaire ayant pour objectif d'argumenter cette note. Nous pouvons donc logiquement considérer que ces commentaires sont des textes subjectifs et porteurs d'une opinion sur un film, tout du moins dans la grande majorité d'entre eux. C'est une partie de ces textes, ainsi que les notes associées, qui ont été récoltées afin de construire le corpus d'expérimentation. Des captures d'écran du site *Flixster* sont disponibles dans la partie **Annexes** (C).

2. www.flixster.com

4.2.2 Processus de crawl

Un aspirateur de site Web a été spécialement conçu afin d'extraire les données du site *Flixster*. Il intègre également un analyseur permettant de parcourir automatiquement les pages Web et d'extraire les informations désirées. Il requiert un point d'entrée sur le site sous la forme d'un ou d'une liste d'identifiants (il s'agira pour nous des identifiants de films) et permet ensuite d'aspirer les informations que l'on souhaite sur les pages correspondantes ainsi que les pages liées à celles-ci (par le biais de lien hypertextes). Le schéma 4.1 présente les différents liens entre les pages d'utilisateurs, de présentation de films et de présentation d'acteurs existants sur le site. Ce schéma est légèrement simplifié car les pages ne contiennent que des extraits de certains types de données comme les commentaires, les amis ou les favoris. En effet, ces données pouvant potentiellement être infinies (aucune limite n'est fixée), elles sont consultables sur de nouvelles pages accessibles via des liens hypertextes (≪ See all ... ≫).

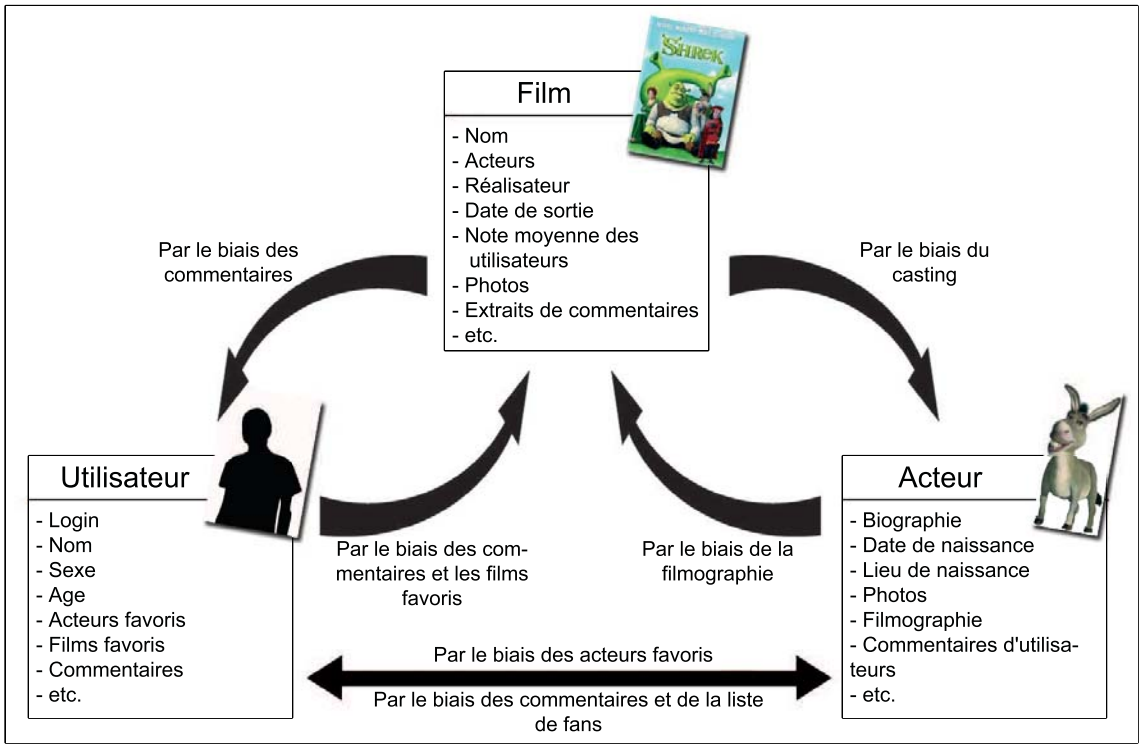


FIGURE 4.1 – Modélisation des liens existants entre les pages du site *Flixster*

Afin d'évaluer la chaîne de traitements, seuls les commentaires portant sur les films correspondants à ceux du challenge *Netflix* ont été pris en compte. Il a donc tout d'abord été nécessaire de retrouver, sur le site *Flixster*, le maximum de films présents dans la liste *Netflix*. À l'aide du moteur de recherche du site *Flixster*, environ 10 400 films de la liste ont pu être retrouvés. Une partie des commentaires rédigés à propos de ces 10 400 films

ont alors pu être aspirés et stockés dans une base de données. Ils forment le corpus de textes nommé *Corpus Flixster*.

4.2.3 Description statistique

Le corpus est composé de 3 326 485 commentaires contenant un nombre moyen de mots légèrement supérieur à 15. Cette taille moyenne indique que les textes sont généralement très courts, mais on peut observer que les tailles peuvent être très variables. Les commentaires les plus courts ne contiennent aucun mot (c'est-à-dire qu'ils ne contiennent que des caractères non alphanumériques) et le plus long commentaire contient plus de 10 000 mots. Toutefois, la figure 4.2 nous permet d'observer que les longs commentaires sont très minoritaires dans l'ensemble du corpus. On remarque notamment qu'environ 98% des commentaires contiennent moins de 100 mots, 90% en contiennent moins de 30 et 50% moins de 9.

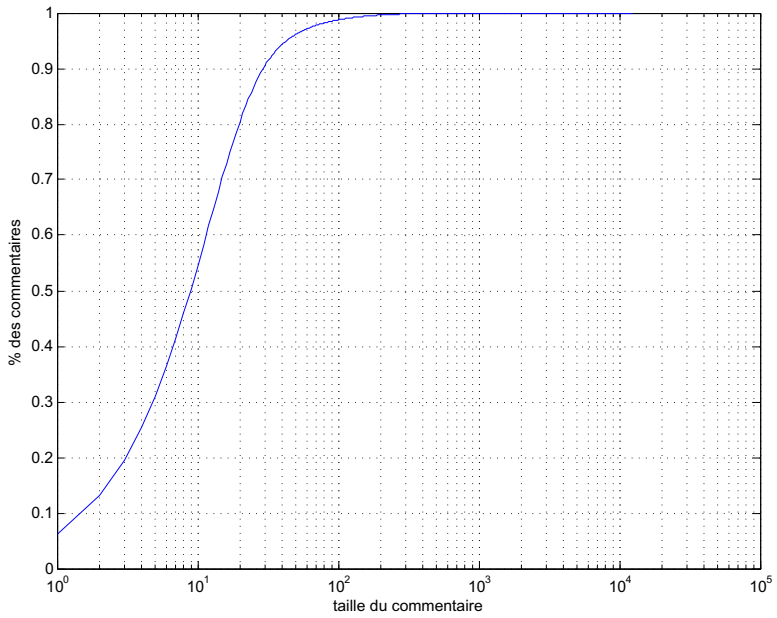


FIGURE 4.2 – Distribution des commentaires selon leur taille

Chacun de ces commentaires est associé à une note comprise entre 0,5 et 5 (0,5 signifiant qu'il s'agit d'un commentaire très négatif et 5 d'un commentaire très positif). La figure 4.3 représente la répartition des commentaires suivant la note qui leur est associée. On observe aisément que plus les commentaires sont positifs, plus ils sont nombreux et représentent un pourcentage important. La répartition des commentaires dans les classes n'est pas la seule statistique à être déséquilibrée. On remarque également que la taille moyenne des commentaires augmente en même temps que la note (figure 4.4). Le nombre

de mots moyen est supérieur à 17 pour les commentaires notés 4,5 alors qu'il n'est que de 13 pour les commentaires notés 0,5.

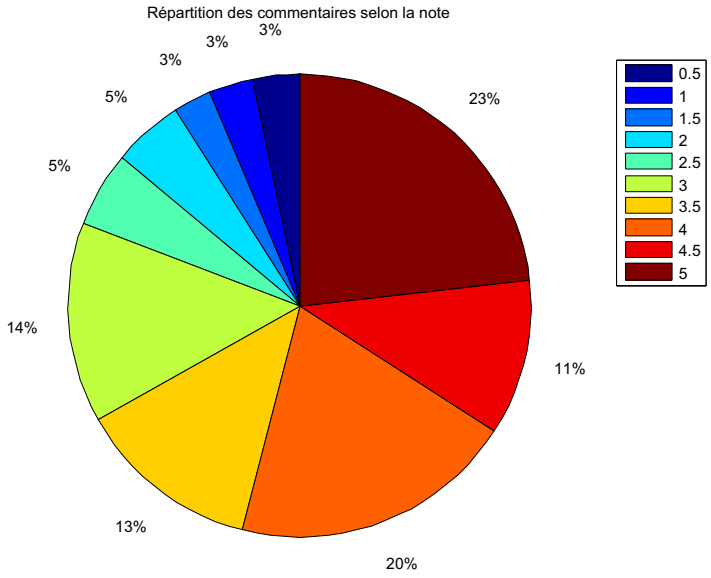


FIGURE 4.3 – Répartition des commentaires selon la note

Ces commentaires ont été rédigés par 97 746 utilisateurs différents et concernent exactement 10 411 films. Cela représente une moyenne de 34 commentaires par utilisateur et environ 318 commentaires par film. La répartition des commentaires selon les utilisateurs et les films n'est pas du tout équilibrée. Pour les utilisateurs tout d'abord, la figure 4.5 présente la distribution du nombre de commentaires par utilisateur. On observe que la médiane se trouve à 20 commentaires, soit un nombre inférieur à la moyenne. Les utilisateurs n'ayant rédigé qu'un seul commentaire ayant comme sujet l'un des films de la liste sont très peu nombreux (environ 900). Le nombre maximum de commentaires rédigés par le même utilisateur est de 3 912 mais il s'agit pratiquement d'un cas unique. En effet, le corpus ne contient que 60 utilisateurs ayant rédigé plus de 1 000 commentaires, dont deux utilisateurs ayant rédigés plus de 3 000 commentaires et deux utilisateurs qui en ont rédigé entre 2 000 et 3000. Ce sont ces quelques utilisateurs particuliers qui font que la moyenne est supérieure à la médiane.

Concernant les films, la courbe représentant leur distribution selon le nombre de commentaires les concernant (figure 4.6) est exponentielle, ce qui signifie que la distribution est encore plus déséquilibrée que celle concernant les utilisateurs. Le plus grand nombre de commentaires concernant un même film est de 51 635 (pour information, il s'agit du film

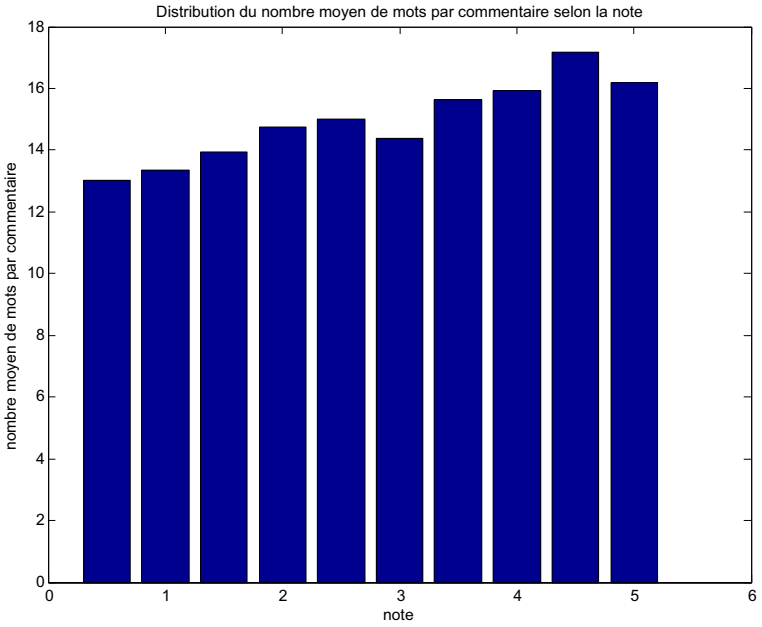


FIGURE 4.4 – Nombre moyen de mots par commentaire selon la note

≪ Titanic [3] ≫) alors que le plus petit nombre est de 1 et concerne 915 films (soit quasiment 10% de l'effectif total). On remarque également que la médiane se situe aux alentours de 20 commentaires, ce qui est très en deçà de la moyenne (318). La raison à cela est que très peu de films sont énormément commentés et qu'un grand nombre le sont peu. En effet, les 30 films les plus commentés se partagent à eux seuls 30% des commentaires, tandis que la moitié des commentaires (soit environ 1 500 000) ne concernent qu'1% des films. Ceci est lié au phénomène de la longue traine que nous avons présenté dans l'état de l'art (section 2.2.3).

4.2.4 Extraits et particularités

En parcourant manuellement le corpus, on peut observer que les caractéristiques spécifiques au DEM sont présentes dans un grand nombre de commentaires. Le tableau 4.1 présente 15 exemples pris totalement au hasard dans le corpus. Ils ont tous été rédigés par des utilisateurs différents et portent sur des films différents. Parmi ces exemples, une très grande majorité d'entre eux possède au moins l'une des caractéristiques décrites précédemment dans la section 4.1 :

– **Exemple 1 :** dans ce commentaire, la ponctuation est manquante (la majuscule en début de phrase et le point en fin). Du point de vue grammatical il manque également le sujet de la phrase ;

3. Film américain réalisé par James Cameron sorti en 1997.

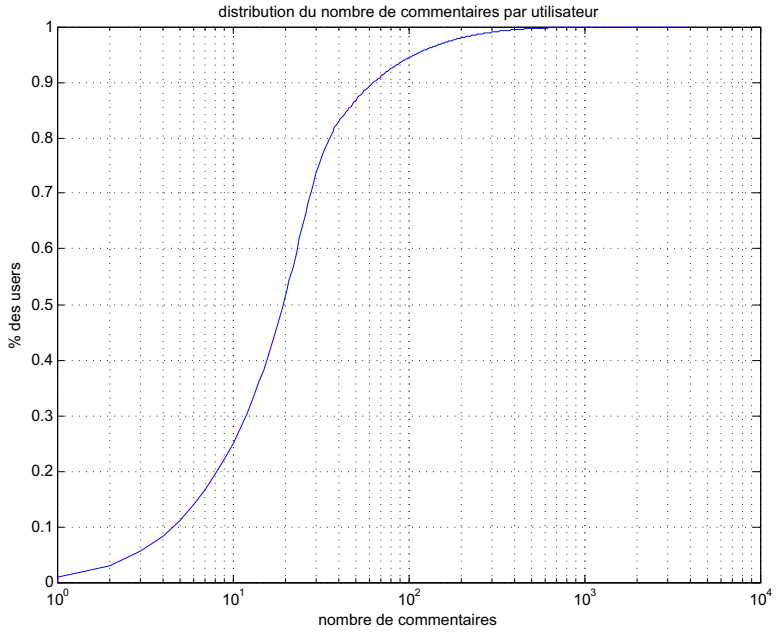


FIGURE 4.5 – Distribution des utilisateurs selon le nombre de commentaires qu'ils ont émis

- **Exemple 2 :** ici également, la ponctuation est manquante. Le commentaire contient de plus une onomatopée (« ugggg ») et il manque le sujet de la phrase ;
- **Exemple 3 :** orthographiquement parlant, le commentaire ne contient aucune erreur. On peut toutefois noter que la première phrase ne contient pas de verbe ;
- **Exemple 4 :** il ne s'agit pas du tout d'une phrase construite, la ponctuation est inexistante et l'un des mots est étiré volontairement ;
- **Exemple 5 :** graphiquement, le texte est correct, mais grammaticalement, la phrase ne contient ni verbe, ni sujet, ni complément ;
- **Exemple 6 :** ce commentaire ne contient aucun signe graphique distinctif du DEM, toutefois, il manque un verbe dans la première phrase ;
- **Exemple 7 :** il s'agit d'un commentaire rédigé dans une langue différente de l'anglais. Pour information, des commentaires rédigés en français, allemand, espagnol et italien ont également été aperçus dans le corpus ;
- **Exemple 8 :** on peut noter la présence du langage familier avec le mot « meh » utilisé afin de signifier l'indifférence [4] ;
- **Exemple 9 :** la majuscule est absente en début de phrase et l'abréviation « cuz » (signifiant « because ») est employée ;
- **Exemple 10 :** ce commentaire contient de multiples caractéristiques. Tout d'abord, un abus des majuscules ainsi que des multiplications de ponctuations sont présents. Il

4. "Meh" is an interjection, often an expression of apathy, indifference, or boredom. The word gained popularity as a result of its use on The Simpsons. *Wikipédia 2010*

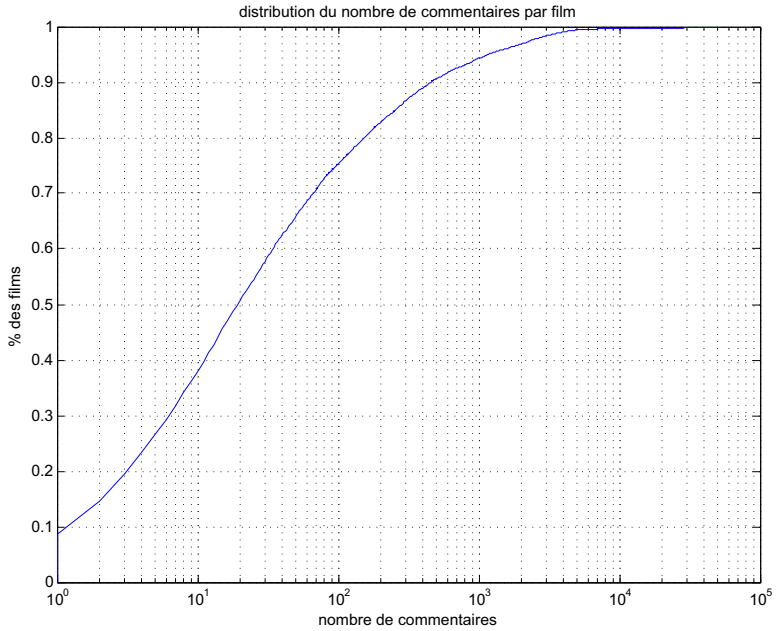


FIGURE 4.6 – Distribution des films selon le nombre de commentaires qui leur sont attribués

manque également des ponctuations en début de phrase. L'abréviation « OMG » (pour « Oh my God ») est utilisée. Il manque un caractère espace entre les mots « in » et « fact », on peut supposer qu'il s'agit d'une erreur involontaire. On observe également l'utilisation d'une didascalie électronique avec la présence du sigle « <3 » qui représente un cœur, symbole de l'amour (la multiplication des « 3 » étant censée en augmenter la taille et donc accentuer sa signification) ;

- **Exemple 11 :** l'utilisateur a utilisé le sigle « @ » à la place du mot « at » ;
- **Exemple 12 :** ce commentaire semble correctement rédigé. On notera juste qu'il s'agit d'un hors-sujet. On entend par là que le texte ne contient aucune information sur l'opinion que l'utilisateur peut avoir concernant le film commenté, ce qui est censé être le sujet principal du texte ;
- **Exemple 13 :** Une fois encore, le commentaire est correctement écrit, mais le lien avec l'opinion est difficilement compréhensible.
- **Exemple 14 :** ce commentaire contient un étirement de mot, un doublement du point d'exclamation, une apostrophe manquante entre les mots « it » et « s », une faute de frappe dans le mot « because » et pour finir, un smiley ;
- **Exemple 15 :** ce dernier exemple contient un étirement de mot doublé d'un abus de majuscules et une multiplication de la ponctuation.

Ces exemples extraits aléatoirement du corpus montrent que les caractéristiques décrites dans la section précédente ne font pas l'objet d'une utilisation marginale par les usagers

N°	Note	Commentaire
1	2.5	pretty scary when i saw it as a kid, corny as an adult but still ok until the end which was lame
2	1.5	hated it way too ugggg
3	3.5	Tim Burton's definitive work. You could watch it on mute and it would be just as powerful - the sets, the costumes and the performances are all works of art.
4	3	weeeeeeiiiiiiirrrrrdddddd movie
5	3	A little weird, but not bad...
6	3	Quite a bit dirtier than I remember from watching this as a kid, and the special effects really don't stand the test of time. Still a unique and entertaining film.
7	3	Hakket darligere enn eneren men underholdene, for all del
8	1.5	meh. neat effects, but that's really it for me.
9	4.5	i liked this cuz this shit can really happen.
10	5	OMG I LOVED THIS MOVIE!!!!!!!!!!!!!<333333333333333333333 infact i have the movie.... this is a must watch video!!!!!
11	1	I see it @ work everyday so this movie sucks!
12	5	You keep what you kill!
13	5	I feel smart.
14	3.5	I loved it, i love this movieeeeeeeee!! I guess its beacuse i love animals ^.^
15	4	YYYYYYYEEEEEEEEEEESSSSSSSSSSSSS!!!!!!!!!!!!!!!!!

TABLE 4.1 – Exemples de commentaires extraits du *Corpus Flixster*

des sites communautaires, tout du moins concernant ceux du site Flixster. On observe également que, bien qu'ils ne soient pas majoritaires, certains commentaires sont bien rédigés. Le fait qu'au moins deux styles d'écriture soient présents dans le corpus peut entraîner des complications pour l'application de tâches de fouille de textes.

4.3 Analyse exploratoire du corpus

Nous avons mené une analyse exploratoire visant à comprendre quel type d'informations peut être contenu dans les commentaires. Nous avons cherché à identifier quel est le vocabulaire révélateur d'opinions, mais également s'il existe un vocabulaire particulier permettant de regrouper des films. Nous montrons par cette analyse que certains mots sont caractéristiques d'un groupe de films précis, tandis que d'autres mots sont beaucoup mieux répartis dans les commentaires de l'ensemble du corpus textuels.

Une petite partie seulement du corpus a été utilisée pour cette étude. Nous avons sélectionné aléatoirement 50 000 commentaires qui portent indifféremment sur 7 114 films.

Nous partons du principe que nous n'avons aucun *a priori* sur les données. Nous avons donc fait le choix de n'appliquer que de très légers pré-traitements aux textes, uniquement dans le but de réduire la taille du vocabulaire. Les commentaires de films n'ont ainsi pas subi de traitements linguistiques lourds comme une lemmatisation ou stemmatisation. Les seuls traitements effectués sont une mise en minuscule de tous les caractères alphabétiques ainsi qu'une suppression de la ponctuation. Nous savons que la ponctuation joue un grand rôle sémantique dans ce type de textes mais nous souhaitions que l'analyse porte uniquement sur les mots. Les commentaires traités sont donc tous de la même forme que les exemples suivants :

– « *it was really good juss somethin u wouldnt expect from her* » ;
– « *my favourite horror film really good story line* » ;
– « *omg i love this film soooooooo soppy but sooooooooooo great* ».

Le vocabulaire est ainsi composé de 27 673 termes différents.

4.3.1 Technique utilisée

La méthode utilisée est une méthode de co-clustering. Un co-clustering [Har72] est défini comme le regroupement simultané des lignes et des colonnes d'une matrice. Dans le cas des jeux de données de faible densité, ayant de nombreux 0 dans le tableau croisé Individus-Variables, le co-clustering est une technique attractive pour identifier des corrélations entre groupes d'individus et groupes de variables [Har72, GN06, DMM03].

La méthode utilisée est une extension au cas non supervisé des modèles en grilles introduits dans le cadre de l'évaluation bivariée pour la classification supervisée [Bou07a]. On cherche à décrire conjointement les valeurs de deux variables catégorielles à expliquer Y_1 et Y_2, comme illustré sur la figure 4.7.

D	∅	•	∅	•
C	•	∅	•	∅
B	∅	•	∅	•
A	•	∅	•	∅
	a	b	c	d

{B, D}	∅	•
{A, C}	•	∅
	{a, c}	{b, d}

FIGURE 4.7 – Exemple de densité jointe pour deux variables catégorielles Y_1 ayant 4 valeurs a, b, c, d et Y_2 ayant 4 valeurs A, B, C, D. Le tableau de contingence sur la gauche ne contient des individus que sur la moitié des cases (marquées •), les autres cases étant vides. Suite au groupement de valeurs bivarié, le tableau de contingence sur la droite permet une description synthétique de la corrélation entre Y_1 et Y_2

Un modèle de groupement de valeurs bivarié non supervisé est défini par :
– Un nombre de groupes pour chaque variable à expliquer ;

– La partition de chaque variable à expliquer en groupes de valeurs ;
– La distribution des individus sur les cellules de la grille de données ainsi définie ;
– La distribution des individus de chaque groupe sur les valeurs du groupe, pour chaque variable à expliquer.

Le modèle est entièrement caractérisé par le choix des paramètres de partition des valeurs en groupes, des paramètres de distribution des individus sur les cellules de la grille et des paramètres de distribution des individus des groupes sur les valeurs des variables. Les nombres de valeurs par groupe sont déduits du choix des partitions des valeurs en groupes, et les effectifs des groupes par comptage des effectifs des cellules de la grille. Afin de rechercher le meilleur modèle, une approche Bayésienne est appliquée visant à maximiser la probabilité $P(M|D) = P(M)P(D|M)/P(D)$ du modèle sachant les données.

Considérons un jeu de données binaire de faible densité avec N individus, K variables et V valeurs non nulles. Un tel jeu de données peut être représenté sous la forme d'un tableau à V lignes et deux colonnes. Cela correspond a un nouveau *tableau de données* avec deux *variables* nommées « ID Individu » et « ID Variable » où chaque *individu* est un couple de valeurs (ID Individu, ID Variable), comme illustré sur la figure 4.8.

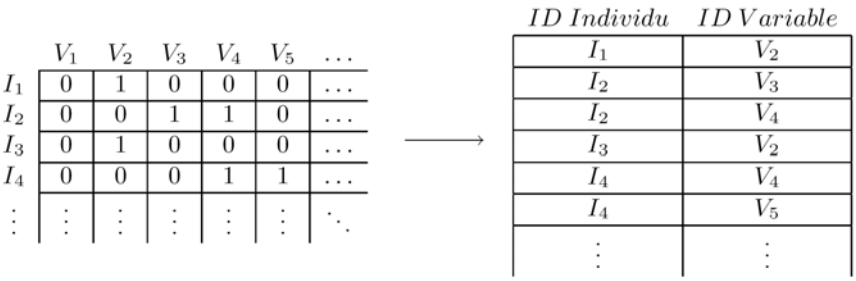


FIGURE 4.8 – Jeu de données binaire de faible densité : depuis la matrice creuse (*individus-variables*) au tableau bivarié dense

En appliquant la méthode de groupement de valeurs bivarié non supervisée à ce tableau bivarié dense, on obtient une grille bivariée basée sur le groupement des individus d'une part, le groupement des variables d'autre part. Il est à noter que les co-clusters sont ici les cellules de la grille, et qu'ils forment une partition sans recouvrement du jeu de données.

4.3.2 Analyse des résultats

À l'aide de cette méthode, et après une dizaine d'heures de traitements, 162 ensembles de films ont été créés, composés de 3 à 213 films, ainsi que 141 groupes de mots contenant entre 1 et 495 mots. On obtient ainsi environ 20 000 co-clusters, soit une réduction d'un facteur 10 000 de l'espace de données si l'on considère la matrice Films-Mots. On produit

ainsi un résumé de l'information que nous interprétons dans les paragraphes suivants.

4.3.2.1 Clustering de mots

Nous pouvons observer dans les résultats que les mots sont, en général, classés entre termes de même nature. On retrouve par exemple des groupes de mots :

- De liaison :
  - *in film with from he his has by who films most...* ;
  - *for all on are out more some can way make also...* ;
- Porteurs d'opinions :
  - *great story an well amazing brilliant acting cast fantastic excellent done...* ;
  - *cool fun awesome awsome entertaining hot flick enjoyable totally line lots...* ;
- Portant sur la typologie des films :
  - *action thriller packed adventure smart seat exciting ride suspense edge spectacular teriffic...* ;
  - *funny hilarious comedy laugh haha laughs funniest laughed jokes funnier laughing...* ;
- Portant sur la typologie d'acteurs ou personnages :
  - *bond james sean moore connery villain daniel roger spy franchise craig brosnan...* ;
  - *johnny depp orlando pirates bloom keira chocolate depps captain pirate knightley willy...* ;
- Etc.

D'autres caractéristiques ressortent de certains groupes. On retrouve par exemple les mots n'appartenant pas à la langue anglaise classés dans des ensembles communs. On retrouve ainsi des groupes de mots français d'une part et des groupes de mots espagnols d'autre part. On retrouve aussi des groupes de mots ne contenant qu'un seul terme comme « it », « a », « of », « to » ou encore « movie », termes qui n'apportent que peu de sens, voire pas du tout, et qui peuvent être employés quel que soit le contexte dans le langage courant. C'est pourquoi l'outil ne peut les rapprocher d'aucun autre mot.

4.3.2.2 Clustering de films

Parmi les résultats du clustering de films, on trouve des groupes sélectionnés selon différentes caractéristiques. On observe tout d'abord des ensembles où les films sont classés par genre.

- On retrouve par exemple les catégories de films pour enfants (*The Lion King, Sleeping Beauty, 101 Dalmatians...*), les « teens-movies » (films d'ado) (*Scary Movie, Big Mommas House, American Pie...*), les films d'horreur (*Saw, The Grudge, Final destination...*), etc. ;
- Certains films sont aussi classés en fonction de l'acteur principal ou du réalisateur. C'est le cas par exemple pour un groupe qui ne contient que des films ayant comme acteur Johnny Depp : *Pirates of the Caribbean, Charlie and the Chocolate Factory, Edward Scissorhands, Tim Burton Corpse Bride, Sleepy Hollow, Benny & Joon,*

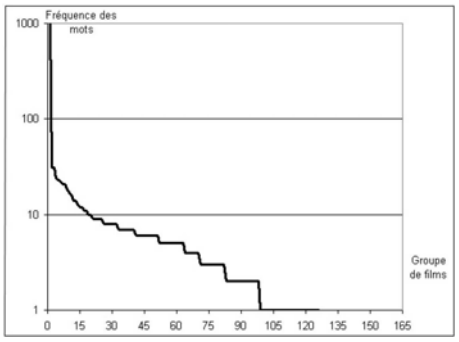


FIGURE 4.9 – Fréquence des mots du groupe « james bond » par groupe de films

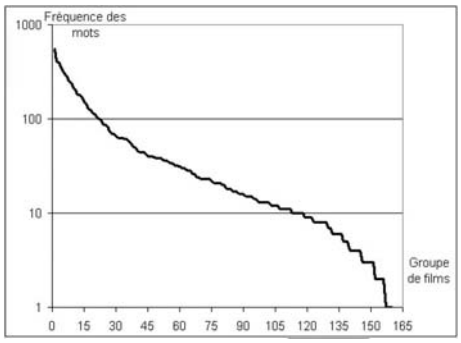


FIGURE 4.10 – Fréquence des mots du groupe « funny » par groupe de films

The Ninth Gate, etc. ;
- Enfin, on trouve également des groupes de films appartenant à la même saga comme les différents épisodes de *Star Wars* ou de *James Bond*.

4.3.2.3 Liens entre clusters de mots et clusters de films

Nous pouvons observer au moins trois tendances lorsque que l'on analyse les fréquences de chaque mot d'un même ensemble dans tous les clusters de films. Certains mots, comme ceux référents à un acteur ou personnage, ne sont présents que dans un très faible nombre de clusters de films (exemple avec l'ensemble de mots contenant *sean, connery, james, bond,* etc., voir figure 4.9). D'autres sont présents dans un plus grand nombre, comme les termes décrivant un type de films : comédie, action, policier, etc. (exemple avec l'ensemble de mots contenant *funny, hilarious, comedy,* etc., voir figure 4.10). Et enfin d'autres groupes de mots contiennent des termes présents quasi uniformément dans une grande majorité des clusters de films, c'est le cas pour les ensembles contenant des mots de liaison (*in, film, with, from, he, his,* etc.).

4.4 Conclusion

Dans ce chapitre nous avons présenté le corpus de textes extrait du site communautaire *Flixster*. Ce corpus est composé de plus de trois millions de triplets utilisateur-film-texte qui permettront par la suite d'apprendre les informations nécessaires à la production de recommandations que ce soit par le biais du filtrage thématique ou du filtrage collaboratif. Tous les textes sont des commentaires contenant des opinions d'internautes portant sur des films. Nous avons également présenté les caractéristiques que peuvent posséder les textes issus de sites communautaires (blogs, forums, etc.). Nous avons également pu observer, à l'aide d'exemples extraits aléatoirement du corpus de textes, que la plupart de ces caractéristiques étaient bel et bien présentes dans les données. Ces traits particuliers

différenciant les textes communautaires des autres textes jouent un rôle important lors des analyses et entraînent généralement des difficultés supplémentaires. C'est ce que nous pourrons constater dans les différents résultats d'expérimentations que nous allons maintenant présenter. Nous nous intéresserons tout d'abord à la recommandation par filtrage thématique qui est l'approche unanimement choisie lorsque l'on a affaire à des données textuelles comme seules données d'apprentissage. Nous verrons ensuite comment faire des recommandations basées sur du filtrage collaboratif par le biais de la classification d'opinion.

Chapitre 5

Expérimentations : Approche thématique classique

Sommaire

Dans ce chapitre, nous évaluons une approche thématique sur nos données textuelles issues du site Flixster. L'objectif est d'évaluer l'information présente dans les textes communautaires en considérant les mots comme des descripteurs de films. Chaque film est alors décrit par l'ensemble des commentaires rédigés à son propos. Ces descriptions permettent ensuite de mesurer des similarités entre films afin de prédire des notes et ainsi établir des recommandations. Cette méthode permettant de faire de la recommandation personnalisée à partir de données textuelles est la plus répandue, c'est pourquoi nous avons fait le choix de l'évaluer sur nos données.

Le chapitre est organisé de la manière suivante. La première section présente les choix effectués concernant la sélection des variables et les métriques permettant de mesurer les similarités entre documents. La deuxième partie présente les résultats obtenus en recommandation à l'aide de la méthode de filtrage thématique décrite dans la section 3.1.

5.1 Choix effectués

La méthode de filtrage thématique utilisée ici nécessite la construction d'une matrice Items-Items contenant des distances entre les items. Les prédictions de notes sont ensuite établies à l'aide des valeurs contenues dans la matrice (voir section 3.1). Les distances entre items sont calculées selon des variables décrivant chaque item. Pour ce faire, la quasi-totalité des méthodes existantes sont basées sur des représentations vectorielles dont chaque dimension représente une variable descriptive (voir section 2.2). Nous décrivons tout d'abord les choix effectués concernant la sélection des variables puis ceux concernant les distances de similarités. Pour finir, nous énonçons les résultats obtenus et nous les comparons à celui obtenu avec les données IMDb lors de l'étalonnage du moteur (RMSE = 0,921, voir section 3.2).

5.1.1 Sélection des variables et représentation

Les données textuelles utilisées n'étant pas structurées, à l'inverse des données de type IMDb, la première étape de cette approche consiste à segmenter les textes afin d'obtenir un ensemble de variables. Pour ce faire, les séparateurs choisis sont tous les caractères autres que les caractères alphanumériques. Tous ces caractères (ponctuations, symboles, etc.) sont toutefois conservés dans la représentation car nous estimons qu'ils sont porteurs d'information dans ce type de textes. Le tableau 5.1 montre un exemple de la segmentation réalisée.

Avant segmentation	5/5, the 2nd best film of 2008 ! ! !
Après segmentation	5 | / | 5 | , | the | 2nd | best | film | of | 2008 | ! | ! | !

Table 5.1 – Exemple de segmentation

Il est à noter que cette segmentation ne permet pas de conserver les chaînes de caractères non alphanumériques comme les smileys (« :-) »), chacun de ces caractères étant considéré indépendamment. Il aurait été possible de conserver ces suites de caractères en sélectionnant manuellement les motifs à conserver par exemple, ou en choisissant une taille minimale de suite caractères à conserver telle quelle, mais nous n'avons pas jugé cette complication nécessaire. Notons également que le seul pré-traitement appliqué sur les textes est une minusculisation de tous les caractères alphabétiques. Le corpus n'a subi aucun autre pré-traitement.

Le nombre total de variables distinctes s'élève à 370 889. Ce nombre est assez élevé pour une représentation thématique et entraine un ralentissement des traitements. De plus, toutes les variables ne sont pas informatives pour la description des items et génère du bruit dans les données. Ce bruit peut donc entraîner une baisse des performances en termes de temps de calcul mais également en termes de résultats. Il paraît donc nécessaire

de réduire le nombre de variables. Quatre méthodes, de complexités diverses, sont alors possibles :

- Il est possible de procéder à des regroupements de variables par des méthodes de clustering. Le clustering consiste à diviser un ensemble de données en différents « clusters » (groupes) partageant des caractéristiques communes. Quand il s'agit de mots, on essaye généralement de regrouper les mots synonymes, de même racine, de même famille ou de sens sémantique plus ou moins proche suivant l'application visée.
- On peut également chercher à projeter les vecteurs sur un nouvel espace vectoriel avec des méthodes d'Analyse en Composantes Principales (ACP) ou d'Analyse Sémantique Latente (LSA). Le but de l'ACP est de trouver une meilleure base de représentation des données obtenue par combinaison linéaire de la base originale. La LSA vise à construire un nouvel espace de « concepts » à partir de l'analyse statistique de l'ensemble des cooccurrences des mots dans le corpus de textes.
- Il est également possible d'appliquer des pré-traitements linguistiques aux textes comme des corrections orthographiques ou une lemmatisation afin de réduire l'ensemble des variables.
- La méthode la moins complexe consiste à supprimer les variables selon leur nombre d'occurrence dans le corpus.

Nous avons décidé, pour commencer, d'évaluer le modèle le plus simple de sélection des variables : la suppression par le nombre d'occurrences. L'objectif est donc d'éliminer les variables les moins informatives. Le graphique en échelle doublement logarithmique de la figure 5.1 représente la fréquence des 370 889 variables en fonction de leur rang. On observe que la courbe suit la loi de Zipf. Cette loi dit qu'en classant les mots d'un texte par fréquence décroissante, on observe que la fréquence d'utilisation d'un mot est inversement proportionnelle à son rang. La loi de Zipf peut alors s'exprimer de la manière suivante :

$$f(n) = \frac{K}{n}$$

avec $f(n)$ le nombre d'occurrences du mot de rang n et K une constante.

Cette loi a pour conséquence que dans un corpus de textes, les mots les plus informatifs ne sont pas parmi les plus fréquents et les moins fréquents. En effet, les mots qui apparaissent le plus dans le corpus sont, pour la plupart, des mots vides. Les mots vides sont une catégorie de mots jouant un faible rôle sémantique. Les déterminants, les pronoms ou encore les mots de liaison par exemple appartiennent à cette catégorie. Les mots vides s'opposent aux mots pleins dont le rôle sémantique est beaucoup plus important et permet de donner du sens au texte. Concernant les mots les moins fréquents, ils ne sont généralement pas non plus les plus informatifs car ils découlent souvent de fautes d'orthographes ou de l'utilisation d'un vocabulaire trop spécifique à un sujet ponctuel. Ceci signifie donc que les 370 889 variables issues de la segmentation ne sont pas toutes informatives et que les moins informatives d'entre elles sont certainement les plus fréquentes et les moins fréquentes.

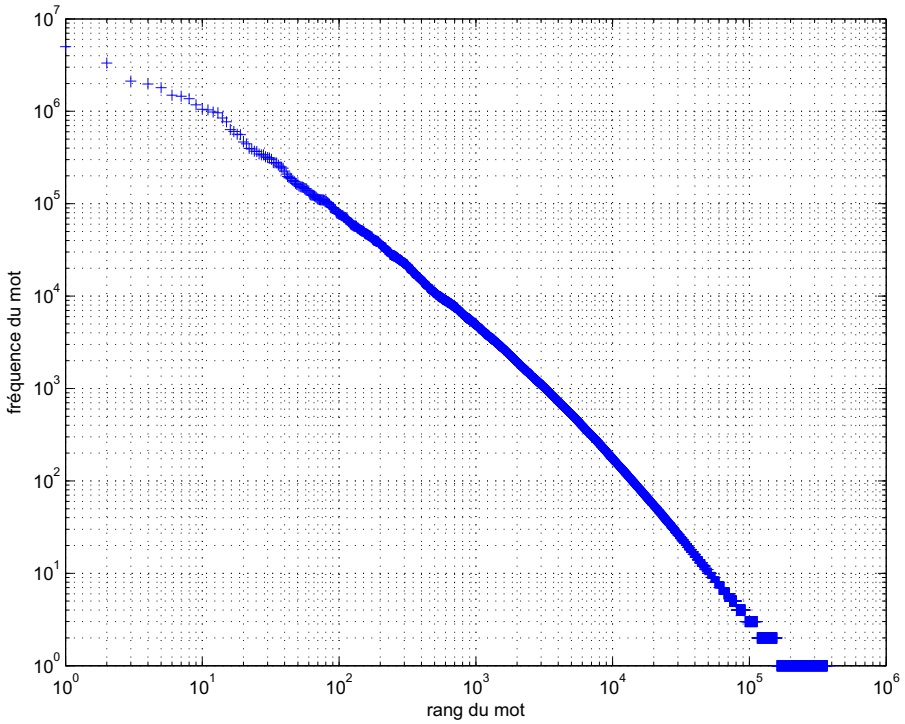


FIGURE 5.1 – Graphique log/log de la fréquence des mots par leur rang dans le *Corpus Flixster*

Nous avons donc modifié le vocabulaire en supprimant, de manière indépendante, les variables les plus fréquentes et les moins fréquentes. En modifiant le nombre de mots les plus fréquents à supprimer ou en modifiant le nombre minimal d'occurrences, nous avons construit plusieurs vecteurs et représenté chaque film sur ces vecteurs par le nombre d'occurrences des variables restantes qui apparaissent dans les textes correspondants. Le tableau 5.2 présente les choix faits concernant l'écrémage des variables, le nombre de variables restantes correspondant ainsi que la quantité de films pouvant être représentés par les variables restantes. En effet, lorsque l'on supprime certaines variables, notamment les plus fréquentes, certains films se trouvent représentés par le vecteur vide. Quant à la figure 5.2, elle présente deux exemples de vectorisation de textes. Une fois tous les textes vectorisés de cette manière, les similarités entre les films peuvent être calculées.

5.1.2 Métriques

Afin de mesurer les similarités entre les vecteurs deux à deux, nous avons sélectionné deux métriques faisant partie des plus utilisées dans le domaine. Il s'agit de l'*indice de Jaccard* et de la *similarité Cosinus*. Pour rappel, l'indice de Jaccard correspond à la cardinalité de l'intersection de deux ensembles rapportée à la cardinalité de l'union de ces deux

Méthodes d'écrémages (appliquées indépendamment)	Variables restantes	Films restants
Sans les 5 mots les plus fréquents	370 884	10 393
Sans les 20 mots les plus fréquents	370 869	10 267
Sans les mots dont la fréquence est inférieure à 4	93 028	10 411
Sans les mots dont la fréquence est inférieure à 100	14 005	10 411
Sans les mots dont la fréquence est inférieure à 1 000	3 215	10 411
Sans les mots dont la fréquence est inférieure à 10 000	531	10 411

TABLE 5.2 – Récapitulatif des différentes sélection de variables

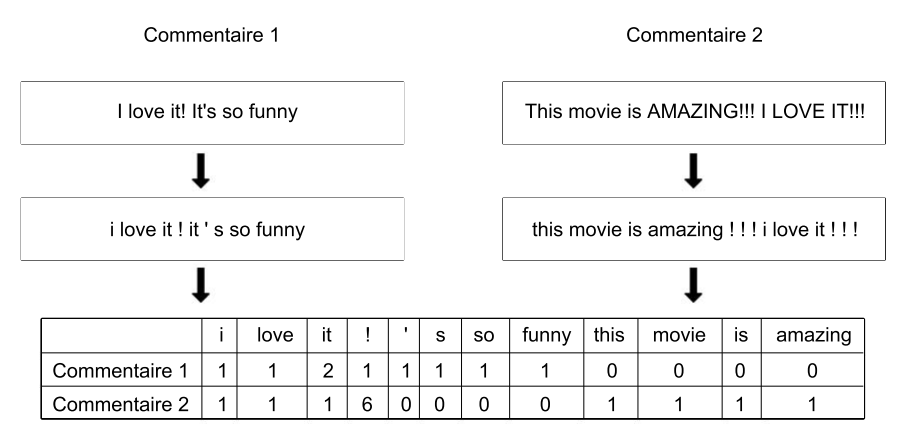


FIGURE 5.2 – Exemples de vectorisation de textes

ensembles (équation 5.1). Les ensembles en question sont constitués des variables reliées à une valeur non nulle dans le vecteur. Le résultat de l'indice de Jaccard est compris dans un intervalle [0,1]. La similarité Cosinus correspond quant à elle au rapport entre le produit scalaire et le produit de la norme de deux vecteurs (équation 5.2). Sa valeur est également comprise dans un intervalle [0,1]. Logiquement, la similarité Cosinus peut être comprise dans l'intervalle [-1,1] mais comme dans le cas présent les composantes sont toujours positives, nous ne pouvons pas obtenir de résultat inférieur à 0.

$$Sim_{jaccard}(i,j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \tag{5.1}$$

$$Sim_{cosinus}(i,j) = \frac{i \bullet j}{\| i \| \| j \|} \tag{5.2}$$

La grande différence entre ces deux mesures est que la distance de Jaccard considère uniquement la présence (> 0) ou l'absence ($= 0$) des variables dans les documents alors que la distance Cosinus prend en compte les nombres d'occurrences. Si l'on applique ces deux métriques sur les exemples précédents (figure 5.2), on obtient les distances suivantes :

$$Sim_{jaccard} = \frac{Nombre\ de\ composantes\ communes}{Nombre\ total\ de\ composantes} = \frac{4}{12} = 0,33$$

$$Sim_{cosinus} = \frac{1 \times 1 + 1 \times 1 + 2 \times 1 + 1 \times 6 + 1 \times 0 + 1 \times 0 + 1 \times 0 + 0 \times 1 + 0 \times 1 + 0 \times 1 + 0 \times 1}{\sqrt{1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2}\sqrt{1^2 + 1^2 + 1^2 + 6^2 + 1^2 + 1^2 + 1^2 + 1^2}}$$
$$= 0,46$$

5.2 Résultats des expérimentations

Pour les évaluations suivantes, nous utilisons uniquement la formule de prédiction de notes du système de recommandation décrite dans la section 3.1. Les évaluations sont faites sur les mêmes utilisateurs que lors de l'étalonnage du moteur (section 3.2). La formule de prédiction de notes, que nous rappelons ici (équation 5.3), est basée sur les similarités contenues dans les matrices construites à partir des textes collaboratifs et personnalise les recommandations à l'aide des profils des utilisateurs cibles.

$$p_{ui} = \overline{r_i} + \frac{\sum_{\{j \in S_u\}} sim(i,j) \times (r_{uj} - \overline{r_j})}{\sum_{\{j \in S_u\}} sim(i,j)} \tag{5.3}$$

L'objectif de ces évaluations est de comparer la quantité d'information présente dans les différentes matrices de similarités construites à partir de différentes sélections de variables. Nous souhaitons également observer la différence de performance par rapport au résultat obtenu avec les données structurées provenant d'*IMDb* (RMSE = 0,921). Rappelons que les données *IMDb* sont des données très riches qui sont, en règle générale, difficilement accessibles pour des domaines autres que celui des films. De plus, elles ne contiennent normalement aucun bruit, les descripteurs étant tous sélectionnés manuellement. On entend par là que tous les descripteurs présents sur *IMDb* sont *a priori* porteurs d'information. Tous les résultats obtenus à partir des données *Flixster* sont répertoriés dans le tableau 5.3.

Méthode d'écrémage	RMSE avec Jaccard	RMSE avec Cosinus
Sans les 5 mots les plus fréquents	0,931	0,928
Sans les 20 mots les plus fréquents	0,931	0,929
Sans les mots dont la fréquence est inférieure à 5	0,930	0,926
Sans les mots dont la fréquence est inférieure à 100	0,930	0,927
Sans les mots dont la fréquence est inférieure à 1 000	0,931	0,928
Sans les mots dont la fréquence est inférieure à 10 000	0,932	0,930

TABLE 5.3 – RMSE obtenues avec les différentes sélections de variables et les deux mesures de similarité

Les RMSE sont toutes très proches et se situent autour de 0,93. Les différences se font seulement au niveau de la troisième ou quatrième décimale. Les sélections de variables ne sont donc pas concluantes car aucune d'entre elles ne semble entraîner une amélioration nette des performances. On observe également qu'aucune des deux métriques, que ce soit Jaccard ou Cosinus, ne semble devoir être privilégiée. Ces deux mesures permettent en effet d'obtenir des résultats quasi-identiques. Ces observations n'encouragent pas à évaluer des sélections de variables plus évoluées, comme des variables issues de pré-traitements linguistiques ou encore de regroupements (clustering). On peut également observer que la meilleure RMSE obtenue ici est assez proche de la RMSE obtenue avec les descripteurs issus d'*IMDb* (0,926 contre 0,921). Ces résultats sont encourageants dans le fait qu'ils montrent que des données textuelles non structurées issues de sites communautaires peuvent être employés lorsque l'on ne possède pas de données descriptives riches comme c'est le cas dans le domaine des films. Ces résultats renforcent également l'idée que des traitements plus compliqués sont inutiles. En effet, au vu de la nature des données *IMDb*, nous ne nous attendions pas à obtenir de meilleurs scores. Les descripteurs d'*IMDb* sont structurés et sélectionnés avec sérieux par un grand nombre de véritables cinéphiles. Ceci n'a donc rien à voir avec les données provenant de *Flixster* qui sont produites par des gens de tous âges et de toutes provenances et qui sont des données personnelles et non collaboratives. Sur *IMDb*, des modérateurs vérifient toutes les informations apportées par les utilisateurs avant de les publier. S'ils sont en désaccord, si des fautes d'orthographes sont présentes, s'il y a des doublons, etc. les informations ne sont pas publiées. Concernant *Flixster*, personne ne peut modifier les textes publiés par les utilisateurs. L'≪ intelligence collective ≫ modérée permet donc aux données d'*IMDb* de se perfectionner sans ajouter de bruit alors qu'elle n'a pas du tout cours sur le site *Flixster*.

Les données *Flixster* semblent toutefois contenir des informations supplémentaires par rapport à celles d'*IMDb*. Il s'agit de leur aspect subjectif. Il semblerait même que beaucoup de commentaires ne contiennent que des informations subjectives comme nous avons pu le voir lors de la présentation des données (section 4.2). Les descripteurs *IMDb* résultent d'un travail collaboratif et reflètent donc beaucoup plus une énumération de faits que de sentiments ou d'opinions. Or, dans une tâche de recommandation, les informations subjectives comme les sentiments ou les avis divergeants ont également un rôle à jouer. Cet aspect des textes ne semble pas assez ressortir avec une approche thématique telle que nous l'avons étudiée ici. Une autre approche, où les opinions exprimées prendraient le dessus sur les aspect factuels, mérite par conséquent d'être explorée.

5.3 Conclusion

Dans ce chapitre, nous avons présenté des résultats issus d'une approche thématique appliquée aux textes communautaires. C'est le type d'approches classiquement utilisées en recommandation personnalisée lorsque les données permettant de distinguer le contenu sont des données textuelles. Les mots présents dans les commentaires sont considérés comme des variables descriptives. Les films sont représentés sur l'espace vectoriel décrit par ces mots et des similarités sont mesurées entre chacun d'entre eux deux à deux. Plusieurs sélections de mots ont été effectuées en supprimant un nombre variable de mots très fréquents et de mots peu fréquents. Les deux mesures de similarités les plus utilisées dans la littérature ont également été évaluées. Les recommandations sont ensuite établies et évaluées sur une base de test contenant des notes données par des utilisateurs sur des films. Les notes prédites par le moteur de recommandations sont comparées aux notes déjà présentes dans la base. Deux conclusions ressortent de ces évaluations. Tout d'abord, on observe que les résultats sont proches de ceux que l'on peut obtenir à l'aide de données structurées très riches et souvent difficiles à obtenir suivant la nature des items à recommander. On observe également que les résultats des différentes évaluations sont très proches, quelles que soient les variables ou la mesure de similarité sélectionnées. D'autres sélections de variables plus complexes et plus coûteuses, comme des pré-traitements linguistiques ou des méthodes de regroupements de variables, pourraient peut-être permettre d'améliorer les résultats mais ceci n'est pas assuré. De plus, il est connu que les méthodes de filtrage collaboratif sont plus efficaces que les méthodes de filtrage thématique. Cette information, associée aux résultats obtenus, nous a donc poussé à explorer une nouvelle voie qui, à notre connaissance, n'a encore pas été expérimentée. Elle consiste à extraire, à partir des commentaires utilisateurs, des données d'usages de type utilisateur-item-note. Cette approche nécessite l'emploi de méthodes appartenant au domaine de la fouille d'opinion qui est un sous-domaine de la fouille de texte. Avant de vérifier les capacités de cette approche, il est donc nécessaire de voir plus précisément en quoi consiste la fouille d'opinion.

Chapitre 6

État de l'art sur la fouille d'opinion

Sommaire

Avec l'émergence du Web 2.0, de plus en plus de documents textuels contenant des informations exprimant des opinions ou des sentiments sont disponibles sur la toile (boites à réactions, réseaux sociaux, forums, groupes de discussion, blogs, etc.). Du fait de cette prolifération de données textuelles porteuses d'avis divers, la détection ou l'extraction automatique d'opinions est un domaine de recherche qui s'est largement développé ces dernières années. Elle concerne de très nombreux domaines d'applications. Nous pouvons citer, par exemple, les clients qui souhaitent évaluer un produit avant de l'acheter, l'image que les clients peuvent se faire d'une entreprise, la détection de rumeurs sur le Web, la veille (technologique, marketing, concurrentielle, sociétale) qui peuvent se révéler cruciales pour les entreprises ou organisations.

Que cela soit dans les blogs, les forums ou dans les sites d'évaluations, les opinions sont nombreuses et sont le plus souvent écrites dans un langage libre. On se trouve donc dans un contexte où il faut appréhender de gros volumes de données, s'intéresser à la qualité de ces données et à la validation des résultats.

Les approches de détection d'opinions cherchent à déterminer, de la manière la plus automatique possible, les caractéristiques d'opinions positives ou négatives à partir d'ensembles d'apprentissage. Différents types d'algorithmes de classification sont utilisés. Ils se basent sur des techniques linguistiques ou statistiques.

Ce chapitre présente le domaine de la fouille d'opinions et ses problématiques. La problématique qui nous intéresse concerne la classification d'opinion que nous positionnons par rapport à la classification de textes plus classique. Nous discutons ensuite des grands types d'approches pour la tache de classification d'opinion et nous concluons par les méthodes d'évaluation des résultats couramment utilisées dans le domaine.

L'analyse de textes en terme d'étude des sentiments ou de l'opinion n'est pas récente [Car79, WB84]. Cependant le domaine de la fouille d'opinion, nommé encore analyse des sentiments, a pris de plus en plus d'importance et a commencé à intéresser de plus en plus de monde dès le début des années 2000 avec l'arrivée du Web communautaire et la multiplication des forums sur la toile. La première apparition du terme *Opinion Mining*, signifiant littéralement fouille d'opinion, date de 2003 [DLP03]. Depuis ce jour, le domaine est devenu très actif avec différents challenges qui se sont créés et des centaines d'articles sur le sujet qui ont été publiés.

D'après [PL08], la fouille d'opinion peut être divisée en trois sous-domaines :
- l'identification des textes d'opinion, qui peut consister à identifier dans une collection textuelle les textes porteurs d'opinion, ou encore à localiser les passages porteurs d'opinion dans un texte. Plus précisément, on parle ici de classer les textes ou les parties de textes selon qu'ils sont objectifs ou subjectifs ;
- le résumé d'opinion, qui consiste à rendre l'information rapidement et facilement accessible en mettant en avant les opinions exprimées et les cibles de ces opinions présentes dans un texte. Ce résumé peut être textuel (extraction des phrases ou expressions contenant les opinions), chiffré (pourcentage, note), graphique (histogramme) ou encore imagé (thermomètre, étoiles, pouce levé ou baissé...) ;
- la classification d'opinion, qui a pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime. On considère généralement les classes positive et négative, ou encore positive, négative et neutre.

Ce sont clairement l'identification des textes d'opinion et la classification d'opinion qui sont les plus étudiés dans la littérature et sont également sujets à différents challenges [OdRM+06, GAB+09]. En ce qui nous concerne, le corpus sélectionné étant composé de commentaires portant sur des films, nous supposons que tous les textes sont subjectifs et sont donc porteurs d'opinions. C'est pourquoi nous nous intéresserons uniquement à la tâche de classification d'opinion. Beaucoup de travaux ont été effectués sur le sujet, et trois grandes catégories de méthodes peuvent être mises en avant : les approches basées sur la linguistique, les approches basées sur l'apprentissage automatique et les approches hybrides.

Dans ce chapitre, nous décrivons tout d'abord plus précisément en quoi consiste la classification de données d'opinion, ses motivations et ses caractéristiques. Nous établissons ensuite un état de l'art des trois grandes approches les plus employées dans le domaine puis nous terminons par l'énumération des méthodes d'évaluations couramment utilisées.

6.1 Classification de textes et polarité

La fouille de textes est l'extraction de connaissances dans des textes en langage naturel. La classification d'opinion est une tâche particulière de la fouille de textes et répond à un problème de traitement automatique de textes qui s'intéresse au traitement des opinions

ou des sentiments et de la subjectivité dans les textes. On connait ce domaine sous le nom de opinion mining, sentiment analysis ou encore subjectivity analysis. Dans cette partie, nous décrivons plus précisément les objectifs de la classification d'opinion ainsi que ses intérêts, et nous finissons par les aspects la différenciant des autres tâches de classification de textes.

6.1.1 Objectifs de la classification d'opinion

L'opinion peut-être définie [Liu10] comme l'expression des sentiments d'une personne envers une entité. L'opinion est subjective et peut être décrite avec certains attributs. L'attribut d'opinion le plus étudié est la polarité (positive, négative, et éventuellement neutre) qui définit si l'opinion est favorable ou défavorable. D'autres attributs sont l'intensité de l'opinion et le degrés de subjectivité. La classification de polarité d'opinion consiste donc à déterminer si un texte exprime des opinions positives ou négatives, quel que soit le sujet du texte ou les caractéristiques qui y sont discutées. Cette tâche porte sur des textes subjectifs, c'est à dire des textes contenant des avis sur un sujet tel qu'un film, une voiture, un sèche-linge ou encore une personnalité. Une définition plus formelle de la classification d'opinion pourrait être la suivante :

Définition 6.1.1. *Soit C un ensemble de classes ordonnées représentant chacune un degré d'opinion et D un ensemble de textes subjectifs. La classification d'opinion consiste alors à trouver, pour tout $d \in D$, les couples (d, c) tel que $c \in C$.*

Ce type de classification se fait généralement sur deux classes, "contient une opinion positive" ou "contient une opinion négative". On parle alors de classification binaire. Elle peut également se faire sur trois classes en ajoutant une classe de neutralité destinée aux textes porteurs d'une opinion nuancée. On peut également envisager une classification sur un nombre de classes supérieur à trois afin de mieux préciser l'intensité de l'opinion, mais cela reste assez rare dans le domaine.

6.1.2 Raisons d'être et exemples d'applications

Le besoin de savoir ce que les autres pensent est présent chez beaucoup de monde et les conseils de tierces personnes ont souvent une incidence sur les choix que l'on fait, notamment en termes d'achats. Outre les conseils formulés par les proches, beaucoup de magazines ou journaux se sont spécialisés dans les "critiques professionnelles". Ces critiques sont faites par des personnes qualifiées de spécialistes d'un domaine particulier pouvant être l'art, la littérature, le théâtre, le cinéma, la musique ou encore la gastronomie. Ces personnes, nommées "critiques", émettent des avis, généralement par le biais de la presse écrite, qui ont souvent une influence sur le succès à venir de l'objet critiqué. Aujourd'hui, avec Internet, on se rend compte que le désir d'exprimer ses opinions au plus grand nombre prend de plus en plus d'importance dans notre société. Cela est notamment dû aux nouvelles technologies de communication disponibles sur le Web. La quantité de textes exprimant des avis divers et variés présents sur la toile ne fait donc qu'augmenter de jour en jour. Une étude faite sur des internautes américains en 2007 [gro07] montre que

81% des internautes (soit 60% des Américains) ont déjà effectué, au moins une fois, une recherche sur un produit en ligne et que pour 20% des sondés, c'est une pratique régulière. Elle montre également que pratiquement un internaute sur trois (30%) a déjà posté, au moins une fois, un commentaire sur un produit acquis auparavant. Une autre étude [Hor08] datant de 2008, également menée sur des consommateurs américains, annonce que parmi les lecteurs de commentaires postés sur le Net, 73 à 87% avouent avoir déjà été influencés par les différents avis lus (les chiffres dépendant du domaine : hôtels, restaurants, etc.). En France, d'après une enquête faite par le CRÉDOC (Centre de Recherche pour l'ÉtuDe et l'Observation des Conditions de vie) en mars 2009 [Leh09], 57% des internautes ont déjà recherché des avis de consommateurs sur Internet et 66% d'entre eux déclarent avoir confiance en ces commentaires.

L'importance de l'information contenue dans ces commentaires et leur quantité qui ne cesse de croître laissent donc penser que la compréhension automatique de ces textes peut apporter un plus à l'utilisateur, l'objectif étant de résumer l'information et de la rendre accessible plus rapidement et plus facilement. Dans ce but, [MYTF02] expliquent comment ils vérifient les réputations de produits ciblés en analysant les critiques des clients. Ils recherchent tout d'abord les pages Web parlant du produit concerné et extraient les phrases qui expriment de l'opinion. Ils classent ensuite les phrases selon qu'elles expriment une opinion négative ou une opinion positive et en déduisent la popularité du produit. Un simple coup d'œil de l'utilisateur sur ces résultats lui permet donc de se faire une rapide idée de la qualité du produit. Le même objectif est exprimé par [Tur02] qui classe les produits selon deux catégories : *Recommended* et *Not Recommended*.

La classification d'opinion peut également être utilisée afin de détecter des rumeurs, mener des sondages ou encore vérifier la notoriété de personnages publics comme les politiciens. Concernant les politiciens, cela peut également être utilisé dans le but d'analyser les différents avis portés sur des décisions qu'ils prennent et les lois qu'ils votent [KH07, MM06, CFB06, KSH06, SHCZ05].

La classification d'opinion semble également être une tâche adaptée à la problématique de la recommandation automatique personnalisée [CSC06, DWW07]. Elle peut en effet intervenir à différents stades de la recommandation. On peut l'utiliser afin de connaître les goûts et préférences d'un utilisateur et ainsi établir son profil nécessaire à la recommandation personnalisée. Elle peut également permettre de vérifier la notoriété d'un produit afin de mettre en avant les items ayant une meilleure réputation. Elle peut également être utilisée, comme c'est le cas dans cette thèse, pour créer ou enrichir une matrice d'usage nécessaire aux méthodes basées sur le filtrage collaboratif.

6.1.3 Complexité et caractéristiques de la tâche

La classification d'opinion pourrait se comparer à une classification de texte comme les autres : « étant donné les mots présents dans le texte, j'en déduis une classe ». Par exemple, pour la classification par thèmes, il suffit de trouver un certain nombre de mots

en lien avec le sport, à la politique ou encore au cinéma pour découvrir la classe du document. Seulement, en ce qui concerne les opinions, elles sont rarement exprimées par un seul mot. De plus, les discours contenant des opinions sont par définition des discours subjectifs et l'interprétation de la subjectivité est une tâche délicate. En effet, les opinions sont généralement exprimées par des verbes et des adjectifs qualificatifs qui possèdent une certaine hiérarchie suivant le degré d'opinion qu'ils expriment. Ces hiérarchies n'ont rien d' « officiel » et peuvent différer suivant le contexte et les personnes. De plus, il existe la possibilité d'ajouter des adverbes afin de modifier la « force » de l'adjectif ou du verbe.

Outre les problèmes d'intensité d'opinion liés aux adjectifs et verbes, les opinions et sentiments peuvent être exprimés de plusieurs façons. Ils peuvent par exemple être exprimés à l'aide de termes relevant de l'émotion. Ces termes liés à l'émotion tels que « triste », ou « effrayant » sont difficiles à interpréter du point de vue de l'opinion si l'on ne connait pas les goûts de l'auteur. Les goûts de l'auteur ont donc une importance mais le sujet également. Dans le cas d'un film par exemple, le commentaire « Ce film est à mourir de rire » peut exprimer un avis très positif si l'objet visé est une comédie mais pas forcément s'il s'agit d'un film d'horreur ou d'un drame. Les avis peuvent également être exprimés sans employer un seul mot directement lié à l'opinion comme dans le commentaire "courrez vite acheter le livre". Encore une fois, la signification est ambiguë. Si l'on parle du livre, l'avis est plutôt positif, alors que si l'on parle d'un film basé sur un livre, le commentaire semble beaucoup moins positif. Le sarcasme et l'ironie sont également des phénomènes très difficiles à détecter [TL03]. Le traitement des négations semble également être une grosse problématique de la classification d'opinion [PL08].

6.2 Les approches basées sur les lexiques

La principale tâche dans cette approche est la conception de lexiques (ou dictionnaires) d'opinion. L'objectif de ces lexiques est de répertorier le plus de mots possible qui sont porteurs d'opinion. Ces mots permettent ensuite de classer les textes en deux (positif et négatif) ou trois catégories (positif, négatif et neutre). Par exemple, [LHC05] décrivent un système, *Opinion Observer*, qui permet de comparer des produits concurrents en utilisant les commentaires écrits par les internautes. Pour cela, ils ont une liste prédéfinie de termes désignant des caractéristiques de produits. Lorsque l'une de ces caractéristiques est présente dans un texte, le système extrait les adjectifs proches dans la phrase. Ces adjectifs sont ensuite comparés aux adjectifs présents dans leur lexique d'opinion et ainsi, une polarité est attribuée à la caractéristique du produit.

6.2.1 Construction des lexiques d'opinion

Cette méthode nécessite donc la construction d'un ou de plusieurs lexiques d'opinion. Pour construire de tels lexiques, trois méthodes peuvent être appliquées [AB06, Liu09] :

- La méthode manuelle ;
- La méthode basée sur les corpus ;

– La méthode basée sur les dictionnaires.

La méthode manuelle, qui consiste à remplir le lexique de mots d'opinions sans l'aide d'outil particulier, demande un effort important en terme de temps. La sélection des mots porteurs d'opinion et le choix de leur polarité se fait donc uniquement par l'expertise humaine. On peut supposer qu'une part non négligeable de subjectivité peut alors entrer en jeu et peut entraîner certaines erreurs de classification. Quelle que soit la méthode de construction des lexiques, une première étape de classification manuelle est nécessaire. En effet, une première liste de mots et d'expressions doit être identifiée. Ces mots sont appelés les *germes* (*seed words* en anglais). Cet ensemble de mots est ensuite utilisé afin de découvrir, répertorier et classer d'autres mots et expressions porteurs d'opinions.

L'une des solutions permettant d'agrémenter cet ensemble de mots est donc l'utilisation de corpus de textes. [Tur02] propose la méthode suivante : afin de déterminer la polarité de mots ou expressions non classés, il compte le nombre de fois où ces mots ou expressions apparaissent dans le corpus à côté de mots ou expressions présents dans la liste de germes. C'est ce qu'on appelle chercher les "co-occurences" de mots. Un mot apparaissant plus souvent à côté de mots positifs sera donc classé dans la catégorie *positif* et inversement. Yu et Hatzivassiloglou [YH03] proposent une méthode similaire, mis à part qu'ils utilisent la probabilité qu'un mot non classé soit proche d'un mot classé afin de mesurer la force de l'orientation du premier nommé. D'autres méthodes [PTL93, Lin98] utilisent également cette hypothèse dans le but de compléter les lexiques d'opinion : deux mots ou groupes de mots ayant un fort taux d'apparitions communes sont supposés posséder une forte proximité sémantique.

Une autre méthode basée sur le corpus permettant de compléter le lexique d'opinion consiste à utiliser les conjonctions de coordination présentes entre un mot déjà classé et un mot non classé [HM97, KN06, DL07]. Par exemple, si la conjonction *AND* sépare un mot classé positif dans le lexique d'opinion et un mot non classé, alors le mot non classé sera considéré comme étant positif. À l'inverse, si la conjonction *BUT* sépare un mot classé positif et un mot non classé, alors le mot non classé sera considéré comme étant négatif. Les conjonctions utilisées sont les suivantes : *AND*, *OR*, *BUT*, *EITHER-OR*, et *NEITHER-NOR*.

La dernière solution la plus utilisée est une méthode basée sur les dictionnaires. Elle consiste à utiliser des dictionnaires de synonymes et antonymes existants tels que WordNet [MBF$^+$90] afin de déterminer l'orientation sémantique de nouveaux mots. Par exemple, Hu et Liu [HL04a] utilisent ce dictionnaire afin de prédire l'orientation sémantique des adjectifs. Dans WordNet, les mots sont organisés sous forme d'arbres (voir figure 6.1). Afin de déterminer la polarité d'un mot, ils traversent les arbres de synonymes et d'antonymes du mot et, s'ils trouvent un mot déjà classé parmi les synonymes, ils affectent la même polarité au mot étudié, ou bien la polarité opposée s'ils trouvent un mot déjà classé parmi les antonymes. S'ils ne croisent aucun mot déjà classé, ils réitèrent l'expérience en partant de tous les synonymes et antonymes, et ce jusqu'à rencontrer un mot d'orientation

sémantique connue. Cette méthode peut toutefois entrainer une certain nombre d'erreurs car un grand nombre de mots peuvent avoir différentes significations. Voici un exemple d'erreur possible qui a été trouvé à l'aide d'un dictionnaire français de synonymes en ligne [1] : l'un des synonymes du terme *affreux* est *terrible*, lequel est relié aux mots *extraordinaire* et *formidable*.

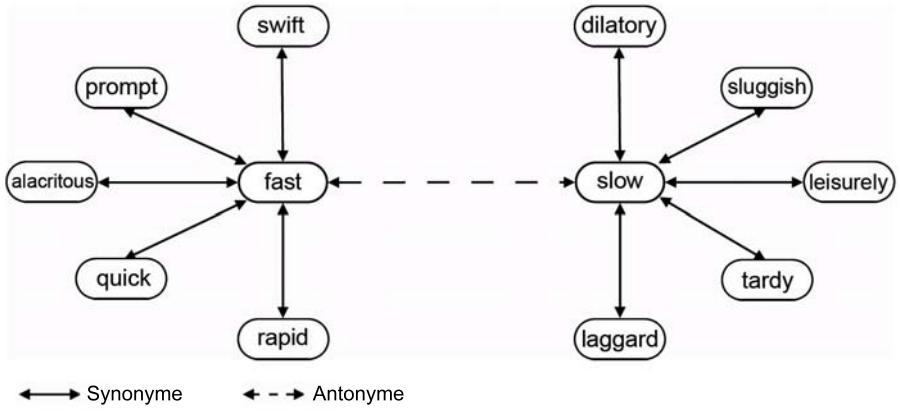


FIGURE 6.1 – Exemple d'arbre de synonymes et antonymes présents dans WordNet

Des analyses plus poussées sont également faites. Afin de mesurer plus précisément la force de l'opinion exprimée dans une phrase, un moyen utilisé est l'extraction des adverbes associés aux adjectifs. Pour ce faire, Benamara et al. [BCP+07] proposent une classification des adverbes en cinq catégories : les adverbes d'affirmation, de doute, de forte intensité, de faible intensité et les adverbes de négation et minimiseurs. Un système d'attribution de points en fonction de la catégorie de l'adverbe permet de calculer la force exprimée par le couple adverbe-adjectif. Le tableau 6.1 présente quelques exemples de mots classés dans les catégories définies par Benamara et al. [BCP+07].

Toutes ces catégories d'adverbes ne sont pas toujours prises en compte car elles n'ont pas la même importance au niveau de la prédiction de note. Les négations paraissent logiquement être des termes importants à détecter, en plus des adjectifs et des verbes, car ils permettent d'inverser la polarité d'une phrase. Pour traiter cet aspect, Das et Chen [DC01] proposent par exemple d'ajouter des mots dans le dictionnaire d'opinion comme "*like-NOT*" qui sont utilisés lors de la détection d'un couple like-négation. Les négations peuvent alors être *not, don't, didn't* ou *never* pour le cas de l'anglais et *ne pas* pour le français. Mais le problème de la détection de la négation reste un problème très ouvert, les méthodes existantes n'étant pas réellement convaincantes. En effet, l'expression de la négation peut être faite sans l'utilisation de ce type de mots. Les expressions telles que "*je nie*", "*je suis contre*" ou encore "*je m'oppose*" peuvent également permettre d'inverser la polarité du reste de la phrase. La difficulté est également due aux différentes façons d'uti-

1. www.synonymes.com

Classe	Mots
Adverbes d'affirmation	Absolutely, entirely, fully, certainly, fairly, exactly, totally, enough...
Adverbes de doutes	Even, possibly, roughly, apparently, seemingly...
Adverbes de forte intensité	So, really, very, pretty, highly, extremely, much, well, too, quite...
Adverbes de faible intensité	Only, a little, almost, a bit, little, rather, nearly, barely, scarcely, weakly, slightly...
Négations et minimiseurs	Not, never, less, no...

TABLE 6.1 – Exemple d'adverbes d'opinion

liser la négation comme le sarcasme ou l'ironie. L'interprétation de la négation nécessite alors une analyse syntaxique qui est un traitement très coûteux en temps de calcul et pas forcément très efficace suivant la qualité du texte analysé.

6.2.2 Classification des textes grâce aux lexiques

Une fois que les mots porteurs d'opinion sont répertoriés dans les lexiques, la dernière étape consiste à déterminer la polarité d'une phrase à l'aide de ces mots. La solution la plus simple consiste à compter le nombre de mots positifs et le nombre de mots négatifs présents. S'il y a une majorité de termes positifs, la phrase est déclarée positive. À l'inverse, si les mots négatifs sont les plus nombreux, la phrase est déclarée négative. Les phrases possédant autant de mots négatifs que de mots positifs peuvent être déclarées neutres [YH03], ou encore, la polarité de la phrase peut dépendre du dernier mot d'opinion parcouru [HL04a]. On peut encore extraire plusieurs opinions dans une même phrase et les associer aux caractéristiques discutées [HL04b].

6.3 Les approches basées sur l'apprentissage automatique

Dans ce type d'approches, les mots sont généralement considérés comme des variables équivalentes. L'aspect sémantique n'est alors pas pris en compte. Les méthodes les plus utilisées pour la classification d'opinion sont les méthodes de classification supervisée. Ce type de méthodes consiste à construire un modèle de classification à l'aide d'exemples. Des exemples sont des données dont on connait déjà la classe. On parle dans ce cas de données classées ou étiquetées. Voici une définition plus formelle du problème de la classification supervisée :

Définition 6.3.1. *Soit X un ensemble de données, Y un ensemble d'étiquettes (ou classes) et D un ensemble des représentations des données. Soit $d : X \longrightarrow D$, une fonction qui associe à chaque donnée $x \in X$ une représentation $d(x) \in D$ et $S \subset D \times Y$ un ensemble de données étiquetées $(d(x), y)$ avec $y \in Y$. La classification supervisée consiste à construire*

un classifieur en s'appuyant sur l'ensemble S, qui permette de prédire la classe de toute nouvelle donnée $x \in X$, représentée par $d(x) \in D$.

D'après cette définition, quatre éléments distincts entrent en jeu dans la classification supervisée :

- La représentation des données (l'ensemble D) ;
- Les étiquettes ou classes de prédiction (l'ensemble Y) ;
- Les données étiquetées, qui constituent le corpus d'apprentissage (l'ensemble S) ;
- Le classifieur ou prédicteur.

6.3.1 Le corpus d'apprentissage

Tout d'abord, il est nécessaire de posséder des exemples (données étiquetées) afin de construire le "corpus d'apprentissage". Ce corpus ayant un impact direct sur l'apprentissage du modèle et par conséquent, sur les résultats de la classification, il est nécessaire que les exemples soient les plus représentatifs possibles de l'ensemble des données. En classification d'opinion, ou plus généralement en classification de textes, les corpus étudiés sont assez restreints car l'étiquetage des exemples est souvent effectué à la main. Ceci entraîne un coût élevé et ne permet donc pas l'obtention d'un gros corpus d'apprentissage. Cependant, les données présentes sur les sites Web 2.0, qui peuvent souvent être étiquetées par les auteurs eux-mêmes ou par d'autres internautes, permettent aujourd'hui de traiter des corpus beaucoup plus conséquents.

6.3.2 Les classes de prédiction

Concernant le choix des classes, il est généralement imposé par le corpus d'apprentissage utilisé et la tâche visée. On parle de classification binaire lorsque le nombre de classes $|Y|$ est égal à 2. Il peut naturellement être supérieur, mais il ne faut pas oublier qu'un nombre de classe élevé augmente le taux d'erreurs. En classification d'opinion, le nombre de classes de prédiction choisi est généralement de deux (aime, n'aime pas) ou trois (aime, avis neutre, n'aime pas) [SKEK+07, SKEK+08, OdRM+06, MOS07, OMS08, GBEA+07].

6.3.3 La représentation

Les documents textuels sont des suites de caractères. Afin d'en permettre l'exploitation, il est nécessaire de représenter les textes numériquement. En fouille de textes, les documents sont généralement représentés sous leur forme vectorielle dite en "sac de mots". Les mots sont alors considérés comme des objets distincts non ordonnés. Cette représentation consiste tout d'abord à sélectionner des variables représentant les dimensions d'un vecteur puis à représenter chaque document sur ce vecteur. En classification de texte, la représentation implique donc une étape préliminaire appelée segmentation, qui consiste à découper le texte afin de sélectionner les variables. Une fois les variables acquises, plusieurs documents peuvent être représentés sur le même espace vectoriel. On obtient alors une matrice dite de Salton [SL65]. Nous présentons tout d'abord en quoi consiste exactement la segmentation, puis nous énumérons les types de représentations les

plus utilisées en recherche d'information qui sont également celles utilisées en classification d'opinion.

6.3.3.1 La segmentation

Un document est un ensemble de caractères. La segmentation consiste à découper cet espace de caractères afin d'obtenir un espace de variables. Cet espace de variables est appelé *vocabulaire* et différents choix concernant sa construction sont possibles suivant le délimiteur choisi. On peut considérer les marques de ponctuation, afin de découper le document en phrases, ou encore les espaces associés à la ponctuation afin d'obtenir des mots. Ce sont d'ailleurs les mots qui sont les variables les plus utilisées en classification d'opinion, mais d'autres choix existent. On peut par exemple limiter la taille des variables à un nombre n de caractères, on parle alors de *n-grammes* de lettres. Les n-grammes peuvent également être formés de mots. Ces n-grammes de mots peuvent être construits selon leur ordre dans la phrase, afin de conserver un sens sémantique. Par exemple, dans la phrase "je n'aime pas ce film", le bigrammes (ou 2-grammes) "aime pas" sera considéré. On peut également construire les n-grammes de mots selon qu'ils apparaissent dans la même phrase par exemple. Une fois le vocabulaire sélectionné, différents choix sont possibles concernant la représentation du document sur le vecteur.

6.3.3.2 La représentation binaire

Cette représentation est la moins coûteuse en temps de calcul. Elle consiste à indiquer, pour un document, quels mots du vocabulaire sont présents (valeur égale à 1) et quels mots sont absents (valeur égale à 0) [CGV07, NH06a, PL04].

6.3.3.3 La représentation fréquentielle

Cette représentation est une extension de la représentation binaire qui prend en compte le nombre d'occurrences des variables dans chaque document. Un texte est donc représenté par un vecteur dont chaque composante correspond à la fréquence des variables dans le texte [PRDP08, PLV02].

6.3.3.4 La représentation fréquentielle normalisée

Cette représentation consiste à normaliser les vecteurs de représentation des textes par la longueur des textes. C'est-à-dire que les fréquences des variables obtenues avec la représentation fréquentielle sont remplacées par la proportion des variables dans chaque document. La proportion s'obtient en divisant la fréquence de la variable par la taille du document.

6.3.3.5 La représentation TF-IDF

La mesure statistique TF-IDF permet l'évaluation d'une variable dans un document à la fois par sa fréquence dans le document concerné, mais également par sa présence

dans tous les autres documents du corpus. La valeur TF (Term Frequency) correspond à la fréquence de la variable v dans le document d normalisée par la taille de d. La valeur IDF (Inverse Document Frequency) mesure l'importance de la variable v dans l'ensemble du corpus en calculant le logarithme de l'inverse du nombre de documents contenant v [GS07, Tri07].

6.3.4 Le classifieur

Un classifieur est une procédure qui, à l'aide d'un ensemble d'exemples, produit une prédiction de la classe de toute donnée. Beaucoup de méthodes de classification supervisée existent et beaucoup d'entre elles ont été testées pour la classification d'opinion. On peut citer les arbres de décision, les réseaux de neurones, la régression logistique, les règles de décision ainsi que des méthodes combinant différents classifieurs comme les systèmes de votes ou les algorithmes de Boosting. Toutefois, les méthodes les plus présentes dans la littérature, et qui semblent également être les plus performantes sur les textes, sont les Machine à Vecteurs de Support (SVM) [PL04, KKB10, WWH04, NH06a, GS07, Tri07, CGV07, PRDP08] et les classifieurs Naïfs Bayésiens (NB) [PRDP08, MCD07, PL04, YH03].

6.3.4.1 Les Machines à Vecteurs de Support

Les Machines à Vecteurs de Support, appelés encore Séparateurs à Vaste Marge, sont des classifieurs binaires très populaires en classification de textes. Considérons tout d'abord le cas où les données sont linéairement séparables. Nommons *positif* et *négatif* les deux classes $y \in Y$. Si le problème est linéairement séparable, les individus positifs sont séparables des individus négatifs par un hyperplan H. Notons $H+$ l'hyperplan parallèle à H qui contient l'individu positif le plus proche de H, respectivement $H-$ pour l'individu négatif. Une machine à vecteurs de support linéaire recherche alors l'hyperplan qui sépare les données de manière à ce que la distance entre $H+$ et $H-$ soit la plus grande possible. Cet écart entre les deux hyperplans $H+$ et $H-$ est appelé la « marge » (voir figure 6.2). Intuitivement, le fait d'avoir une marge plus large procure plus de sécurité lorsque l'on classe un nouvel exemple.

Dans le cas où les données ne sont pas linéairement séparables, l'idée des SVM est de changer l'espace de représentation. La transformation non linéaire des données peut permettre une séparation linéaire des exemples dans un nouvel espace, généralement plus grand, appelé « espace de re-description ».

Dans le cas d'une classification multi-classes, plusieurs méthodes sont possibles, la plus connue étant le principe du *Un-contre-tous*. Il consiste à apprendre tout d'abord un modèle à l'aide de la première classe face à toutes les autres, puis de la deuxième face à toutes les autres et ainsi de suite. On obtient ainsi plusieurs classifieurs que l'on applique tous lors de la classification d'une nouvelle donnée. La classe attribuée à cette donnée est alors celle obtenant le meilleur score.

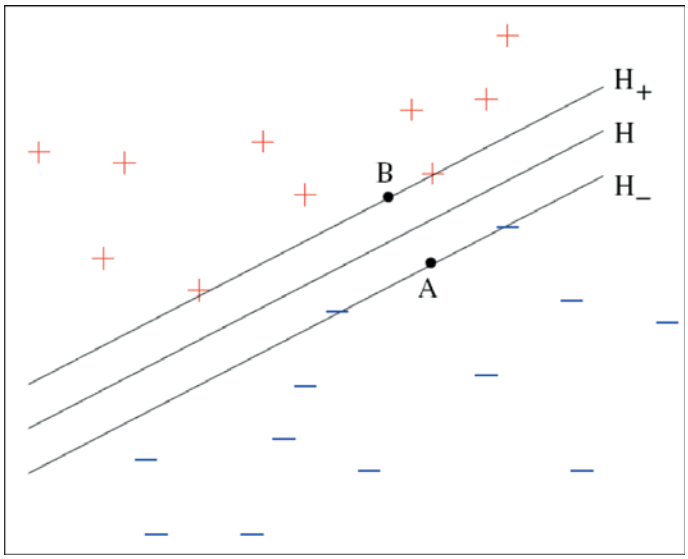


FIGURE 6.2 – Exemple d'hyperplan (H) séparant les individus appartenant à la classe $+$ et ceux appartenant à la classe $-$

6.3.4.2 Les classifeurs Naïfs Bayésiens

Le principe d'un classifieur naïf bayésien consiste à maximiser la probabilité $Pr(y|d)$, soit la probabilité d'occurrence de la classe de prédiction y connaissant la représentation de la nouvelle donnée x (on suppose donc ici $d = d(x) = (d_1, d_2, ..., d_n)$), et ce pour toutes les classes $y \in Y$ et toutes les composantes qui interviennent dans la définition de l'espace de représentation D. Pour cela, on fait appel la règle de Bayes.

Règle de Bayes Soient A et B deux évènements. Le règle de Bayes dit alors que la probabilité de l'évènement A sachant l'évènement B ($Pr(A|B)$) peut se calculer à l'aide des probabilités des évènements A et B ($Pr(A)$ et $Pr(B)$) et connaissant la probabilité de l'évènement B sachant l'évènement A ($Pr(B|A)$) par la formule suivante :

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

Application à la classification En appliquant la règle de Bayes à la problématique de la classification, on obtient l'équation suivante :

$$Pr(y|d) = \frac{Pr(d|y)Pr(y)}{Pr(d)}$$

Les probabilités de l'expression de droite doivent être estimées, à l'aide du corpus d'apprentissage S, afin de calculer la quantité qui nous intéresse, soit $P(y|d)$:

- $Pr(y)$ est la probabilité d'observer la classe y ;
- $Pr(d)$ est la probabilité d'observer la représentation d ;

– $Pr(d|y)$, la vraisemblance de l'évènement « observer la représentation d » si $s \in S$ est de classe y. Ce terme est plus difficile à estimer que le précédent.

En pratique, on ne s'intéresse qu'au numérateur, le dénominateur ne dépendant pas de y. Concernant $Pr(d|y)$, l'hypothèse habituellement faite dans ce type de classifieurs est que toutes les composantes d_i sont indépendantes, ce qui permet de calculer facilement la probabilité globale d'une classe connaissant une donnée. Cette non-dépendance des composantes correspond à « l'hypothèse de Bayes naïve ». On considère donc que $Pr(d|y) = \prod_i Pr(di|y)$. Maximiser $Pr(y|d)$ revient donc à maximiser $(\prod_i Pr(di|y))Pr(y)$. Les $Pr(di|y)$ sont évalués par les fréquences observées dans les exemples de l'ensemble S.

6.4 Les approches hybrides

Plusieurs types d'hybridation sont présents dans la littérature. Dans tous les cas, elles utilisent des éléments décrits dans les deux approches précédentes. On peut distinguer parmi ces approches trois méthodes distinctes :

– La linguistique au service de l'apprentissage automatique ;
– L'apprentissage automatique au service de la linguistique ;
– Une fusion *a posteriori* des résultats des deux approches.

6.4.1 La linguistique au service de l'apprentissage automatique

Une première approche hybride est donc d'utiliser les outils linguistiques afin de préparer le corpus avant de classer les textes à l'aide de l'apprentissage supervisé. Wilson et al. [WWH04] préparent les données à l'aide d'outils de Traitement Automatique des Langues afin de sélectionner un vocabulaire d'opinion. Ces mots pré-sélectionnés sont ensuite utilisés comme vecteurs de représentation des textes pour les outils d'apprentissage supervisé. Trois algorithmes d'apprentissage sont comparés : BoostTexter [SS00], Ripper [Coh96] et SVM^{light} [Joa99a]. Nigam et Hurst [NH06b] utilisent des techniques provenant du Traitement Automatique des Langues afin de détecter dans les textes les mots et expressions porteurs d'opinion et ajoutent des marques dans le texte (traits grammaticaux et + ou - pour opinion positive et opinion négative). Ils utilisent ensuite l'apprentissage automatique pour classer les textes selon leur opinion générale.

6.4.2 L'apprentissage automatique au service de la linguistique

Une autre façon de combiner les méthodes est d'utiliser les techniques d'apprentissage automatique dans le but de construire les dictionnaires d'opinion nécessaires à l'approche linguistique. Hatzivassiloglou et McKeown [HM97] présentent une méthode ayant pour objectif de définir l'orientation sémantique des adjectifs pour la construction du dictionnaire d'opinion. Ils extraient tout d'abord tous les adjectifs du corpus à l'aide d'un analyseur syntaxique, puis utilisent un algorithme de clustering afin de classer les adjectifs selon leur polarité. Riloff et Wiebe [RW03] combinent les deux approches afin de répertorier les

expressions porteuses d'opinion qui, selon eux, sont plus riches que des mots pris individuellement. Turney et Littman [TL03] utilisent une approche statistique pour classer un plus grand nombre de types de mots selon leur polarité : adjectifs, verbes, noms...

6.4.3 Une fusion *a posteriori* des résultats des deux approches

Une dernière façon d'utiliser conjointement les approches basées sur la linguistique et celles basées sur l'apprentissage automatique est de construire plusieurs types de classifieurs afin de combiner leurs résultats, soit par des systèmes de vote, soit par un algorithme d'apprentissage [DWW08]. Dans le cas d'un système de vote, on attribue généralement des poids suivant les performances des classifieurs utilisés. Si les résultats diffèrent, on conserve alors le résultat du classifieur ayant le poids le plus fort. Dans le cas de l'utilisation d'un algorithme d'apprentissage, les poids sont calculés automatiquement selon l'indice de confiance attribué par chaque classifieur à ses propres résultats.

6.5 Les différentes évaluations utilisées dans le domaine

La classification d'opinion est une tâche de classification dite supervisée car les classes sont déterminées à l'avance. Comme toute tâche de classification supervisée, les résultats obtenus peuvent alors être représentés sous la forme d'une matrice de confusion (voir 6.2). Les réponses du classifieur sont comptabilisées dans chaque ligne de la matrice et les réponses attendues sont comptabilisées dans les colonnes. À partir de cette matrice, plusieurs évaluations peuvent être calculées.

		Classes réelles	
		Classe 1	Classe 2
Classes Prédites	Classe 1	A	C
	Classe 2	B	D

TABLE 6.2 – Exemple de matrice de confusion pour une classification binaire

6.5.1 Le taux d'erreurs

Le taux d'erreurs représente le pourcentage de bonnes prédictions par rapport au nombre total de prédictions. La formule correspondant au calcul de ce taux est la suivante :

$$TE = \frac{Documents\ bien\ classées}{Nombre\ total\ de\ documents}$$

En appliquant la formule à l'exemple précédent, on obtient donc :

$$TE = \frac{A + D}{A + B + C + D}$$

6.5.2 F_{score}

Le F_{score} est le mesure la plus utilisée en classification d'opinion. Son évaluation permet de tenir compte à la fois de la précision ainsi que du rappel. Il se mesure à l'aide de la formule suivante :

$$F_{score} = 2 \times \frac{Précision \times Rappel}{Précision + Rappel}$$

6.5.2.1 Précision

La précision correspond au nombre de documents bien classés en rapport à la totalité des documents contenus dans le corpus. Cette mesure permet de vérifier la quantité de *bruit* présent dans une classe, le bruit représentant les documents non pertinents. Plus la précision est élevée, moins il y a de bruit. La précision est mesurée, pour chaque classe, de la façon suivante :

$$Précision_i = \frac{Documents\ correctement\ attribués\ à\ la\ classe\ i}{Nombre\ de\ documents\ attribués\ à\ la\ classe\ i}$$

Afin de calculer la précision totale, on effectue la moyenne des précisions de chaque classe :

$$Précision = \frac{\sum_{i=1}^{n} Précision_i}{n}$$

En appliquant cette formule à notre exemple, on obtient :

$$Précision = \frac{(\frac{A}{A+C}) + (\frac{D}{B+D})}{2}$$

6.5.2.2 Rappel

Le rappel est défini par le nombre de documents bien classés en rapport au nombre de documents pertinents contenus dans le corpus. Il s'oppose au *silence*, le silence représentant les documents pertinents non trouvés. Il se calcule de la façon suivante :

$$Rappel_i = \frac{Documents\ correctement\ attribués\ à\ la\ classe\ i}{Nombre\ de\ documents\ appartenant\ à\ la\ classe\ i}$$

Comme pour la précision, on effectue la moyenne des rappels de chaque classe afin de calculer le rappel total :

$$Rappel = \frac{\sum_{i=1}^{n} Rappel_i}{n}$$

En appliquant cette formule à notre exemple, on obtient :

$$Rappel = \frac{(\frac{A}{A+B}) + (\frac{D}{C+D})}{2}$$

6.6 Conclusion

Dans ce chapitre, nous avons tout d'abord introduit quelques notions appartenant au domaine de la fouille de texte. Cette introduction au domaine nous a ensuite permis d'aborder plus en détail une application de celui-ci qu'est la fouille de données d'opinion. La fouille d'opinion consiste à étudier les sentiments, jugements, avis exprimés dans des textes libres. La classification d'opinion est une sous tâche de la fouille. Cette tâche nous intéressant particulièrement, nous avons dressé l'état de l'art des méthodes existantes que sont les méthodes basées sur la construction de lexiques d'opinion, les méthodes basées sur l'apprentissage automatique et les méthodes hybrides qui mêlent les deux premières citées. Les méthodes basées sur les lexiques offrent une analyse fine mais sont plus gourmandes en intervention humaine que les méthodes statistiques et nécessitent des connaissances sur le langage qui peuvent être longues à paramétrer. Les méthodes hybrides permettent de trouver un compromis : minimisation de l'intervention humaine tout en permettant d'analyser les opinions de manière précise.

La plupart des études comparatives des deux approches principales (les approches hybrides étant un « assemblage » des deux) ont montré que les approches basées sur l'apprentissage supervisé sont les plus performantes. Ces études ont été menées sur des corpus très différents du corpus que nous possédons, c'est pourquoi nous n'avons aucun *a priori* sur l'approche à privilégier. Dans le chapitre suivant, nous évaluons par conséquent les deux types d'approches sur nos données extraites du Web communautaire.

Chapitre 7

Expérimentations : Classification d'opinion et approche collaborative

Sommaire

Le filtrage collaboratif nécessite des données d'usage en entrée. L'objectif de l'approche décrite dans ce chapitre est d'obtenir ces données d'usage nécessaires à partir des commentaires textuels récoltés sur le Net. La classification d'opinion est donc utilisée ici afin d'inférer des notes à partir de textes qui sont eux-même reliés à un sujet et à un auteur. On obtient par conséquent une matrice d'usages, c'est-à-dire des triplets utilisateur-item-note, en lieu et place des triplets utilisateur-item-texte. Les textes communautaires n'étant pas le type de textes le plus couramment traités dans le domaine de la fouille d'opinion, les méthodes les plus adaptées ne sont pas encore connues. La contribution apportée dans ce chapitre porte donc sur l'évaluation de différentes approches de classification ainsi que l'évaluation des différents paramètres de chacune des approches sur le corpus présenté dans le chapitre 4.

Ce chapitre présente toutes les expérimentations qui ont été menées concernant la classification d'opinion de textes communautaires. Dans un premier temps, les choix concernant les classes de prédiction et les corpus sont décrits. Dans un deuxième temps, les méthodes de classification d'opinion sont décrites et évaluées, en commençant par les méthodes basées sur les lexiques puis celles basées sur l'apprentissage automatique.

7.1 Construction du corpus de test et choix des classes de prédiction

Afin d'évaluer toutes les méthodes de classification d'opinion, nous avons construit un corpus de test avec des commentaires provenant du site *Flixster*. Ces commentaires ont été sélectionnés de manière à n'avoir aucun lien avec le *Corpus Flixster* dédié aux recommandations que nous avons présenté dans la section 4.2. La base de test ne contient en effet aucun commentaire relié à un utilisateur ou à un film présent dans le corpus *Flixster*. Le corpus a également été équilibré suivant les 10 classes d'origine (notes de 0,5 à 5) de manière à ne pas biaiser les résultats des expérimentations. Il est composé de 50 000 commentaires au total, soit 5 000 commentaires pour chaque classe. Il aurait été possible de réaliser la classification sur ces dix classes, tout du moins avec des approches basées sur l'apprentissage supervisé, mais l'augmentation du nombre de classes de projection rend l'interprétation des résultats d'autant plus difficile. Pour éviter ces problèmes d'interprétation et faciliter l'analyse des résultats, nous avons donc décidé de travailler sur un nombre réduit de classes. Ne sachant pas quel nombre serait le plus adapté à la problématique de la recommandation, nous avons volontairement conservé plusieurs possibilités de classification.

Afin de réduire le nombre de classes, nous avons effectué une classification binaire à l'aide d'un classifieur naïf bayésien sur des classes de projection nommées **NEG** pour négatif et **POS** pour positif. Nous sommes partis de l'hypothèse que les commentaires notés 0,5 ou 1 étaient tous négatifs, et que les commentaires notés 5 étaient tous positifs. Nous avons donc réalisé la phase d'apprentissage uniquement sur les commentaires appartenant à ces deux classes, puis nous avons réalisé la classification sur chacune des notes restantes. Les résultats obtenus sur le corpus de test sont présentés dans le tableau 7.1.

		Notes d'origine						
		1,5	2	2,5	3	3,5	4	4,5
Classes	NEG	**82,9**	**79,7**	**71,1**	**59,8**	**50,6**	32,1	20,5
Prédites	POS	17,1	20,3	28,9	40,2	49,4	**67,9**	**79,5**

TABLE 7.1 – Résultats de la projection sur deux classes

Nous pouvons observer que les résultats sont assez cohérents. Les résultats les plus ambigus sont dûs à la classification des commentaires notés 3 et 3,5. À partir de ce tableau, nous avons fait le choix de conserver une classification sur deux classes et une classification sur trois classes. Pour la classification binaire, la classe négative (*NEG*) est composée des commentaires ayant obtenus une note comprise entre 0,5 et 3,5 et la classe positive (*POS*) contient les textes restants, soit les commentaires notés 4, 4,5 ou 5. Concernant la classification sur trois classes, la classe *NEG* contient les commentaires notés de 0,5 à 2,5, la classe *NEUTRE* les commentaires notés 3 et 3,5 et enfin la classe *POS* est composée

des commentaires ayant reçu une note comprise entre 4 et 5.

Nous avons également souhaité évaluer une classification plus précise. Nous avons donc choisi, plus de manière intuitive cette fois, cinq nouvelles classes nommées *1*, *2*, *3*, *4* et *5*. La classe *1* contient tous les commentaires ayant une note comprise entre 0,5 et 1,5. Les classes *2*, *3* et *4* contiennent respectivement tous les commentaires notés 2 et 2,5, 3 et 3,5, 4 et 4,5. Enfin, la classe *5* contient les commentaires restants, soit tous ceux ayant obtenus la note 5, qui est la plus haute possible.

7.2 Approches basées sur les lexiques d'opinion

Ce type d'approches est basé sur la sémantique des mots. L'objectif est de répertorier des termes selon leur sens. Pour cela, des pré-traitements sont nécessaires afin de réduire le taille de vocabulaire en supprimant les fautes d'orthographe, formes conjuguées, etc. Cette section présente tout d'abord les traitements effectués sur le corpus afin de mener à bien la tâche de classification puis décrit la méthode de classification. Cette méthode est constituée d'une première étape de construction des lexiques d'opinion et d'une deuxième étape de classification. Pour finir, les résultats obtenus avec cette méthode sont présentés et discutés.

7.2.1 Pré-traitements appliqués aux textes

Le premier pré-traitement est un traitement linguistique qui consiste à appliquer un analyseur orthographique afin d'effectuer plusieurs types de corrections orthographiques sur les textes. Les corrections effectuées sont au nombre de cinq et sont appliquées dans l'ordre suivant : une correction phonétique, une correction par troncature, une tolérance aux répétitions, une correction typographique et enfin une correction morphologique. Deux nouveaux traitements linguistiques, nécessitant cette fois un analyseur morpho-syntaxique sont également appliqués. Il s'agit d'une lemmatisation suivie d'un étiquetage grammatical. L'outil employé pour effectuer tout ces pré-traitements, baptisé TiLT (pour Traitement Linguistique des Textes), est une plate-forme industrielle de TALN conçue par l'équipe Langues Naturelles de Orange Labs [1] [GdNBC+02, HSC+08]. C'est un outil modulaire et multilingue offrant un grand nombre de solutions liées au traitement automatique des langues naturelles.

La figure 7.1 rappelle l'ordre de ces différents pré-traitements que nous décrivons maintenant plus en détail. Les exemples que nous citons sont associés au français mais l'outil utilisé traite aussi bien les textes français que les textes anglais.

1. Anciennement France Télécom R& D

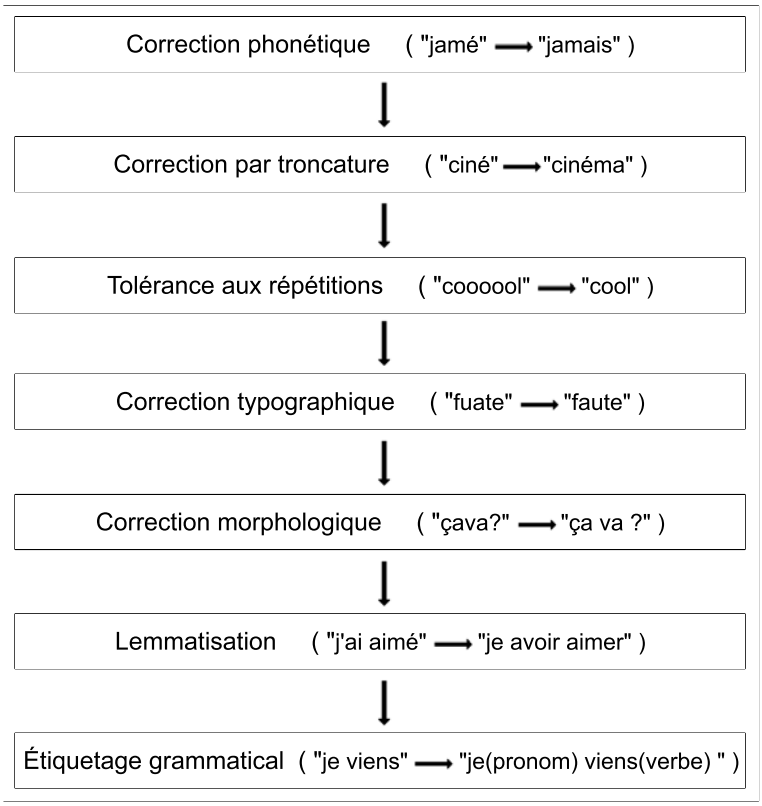


FIGURE 7.1 – Pré-traitements effectués sur les corpus de test et d'apprentissage

7.2.1.1 La correction phonétique

Cette correction permet de trouver, pour une forme donnée, les formes équivalentes du point de vue de la prononciation à l'oral. Concrètement, si l'analyseur rencontre un mot qu'il ne connait pas, il va chercher à le remplacer par un mot connu se prononçant de manière identique. Par exemple, le mot « farmacie » devient « pharmacie ». Cette correction est très utile dans le cas du traitement des textes communautaires car elle prend en compte la plupart des formes issues du langage de type SMS comme « koi », « jamé », « 2m1 », etc.

7.2.1.2 La correction par troncature

Comme son nom l'indique, la correction par troncature permet de corriger les troncations. Rappelons que les troncations sont des mots dont une ou plusieurs lettres de début ou de fin ont été volontairement supprimées afin d'aboutir à une réduction du mot initial. On peut par exemple citer les mots « ciné » et « télé » qui sont couramment employés pour désigner respectivement le « cinéma » et la « télévision ». On parle d'apocope lorsque la fin du mot est supprimée et d'aphérèse lorsqu'il s'agit du début.

7.2.1.3 La tolérance aux répétitions

Cette étape consiste à supprimer les étirements de mots dûs à des répétitions de lettres. Par exemple, le mot « suuuuuuupeeeeeeeeeeeer » sera remplacé par « super ». Cette correction est uniquement effectuée sur les caractères alphabétiques. La ponctuation, les caractères numériques et les divers symboles ne sont donc pas concernés et aucune modification ne leur est appliquée. Il faut également préciser que cette correction ne s'effectue que sur les mots existants, ou plus précisément sur les mots connus par l'outil.

7.2.1.4 La correction typographique

La correction typographique permet de corriger les fautes de frappe. Cela concerne donc les suppressions ou ajouts de lettres (« bazleine » → « baleine »), les remplacements d'une lettre par une autre (« facole » → « facile ») ainsi que les inversions de deux lettres (« gerbiose » → « gerboise »).

7.2.1.5 La correction morphologique

Cette étape consiste à corriger les oublis volontaires ou involontaires du caractère espace. Par exemple, la phrase « çava ? » sera corrigée de la manière suivante : « ça va ? ».

7.2.1.6 La lemmatisation

La lemmatisation consiste à remplacer chaque mot du texte par sa forme canonique conventionnelle, appelée « lemme ». Cette forme correspond aux entrées d'un dictionnaire. Dans la langue française par exemple, la lemmatisation consiste à mettre tous les verbes sous leur forme infinitive et à transformer les autres mots au masculin et au singulier.

7.2.1.7 L'étiquetage grammatical

L'étiquetage consiste à associer à chaque mot la catégorie grammaticale correspondante. Les catégories grammaticales choisies (et leur nombre) sont celles proposées par l'outil que nous utilisons. En français, les catégories grammaticales sont au nombre de neuf : les noms, les déterminants, les pronoms, les adjectifs qualificatifs, les verbes, les adverbes, les prépositions, les conjonctions et les interjections. En anglais, ces catégories ne sont qu'au nombre de huit, les déterminants étant considérés comme des adjectifs.

7.2.2 Méthode de classification

Comme nous l'avons vu dans l'état de l'art (6.2), deux étapes sont nécessaires à la classification. la première étape consiste à construire des lexiques qui contiennent les mots permettant d'exprimer des opinions. Une fonction de classification permet ensuite de déterminer la polarité des textes suivant la présence ou l'absence des mots porteurs d'opinion.

7.2.2.1 Construction des lexiques d'opinion

Pour les expérimentations, deux lexiques distincts ont été construits. Le premier d'entre eux contient des mots porteurs d'opinion positive et le second, des mots porteurs d'opinion négative. Pour trouver les mots exprimant une opinion et les classer, des commentaires ont été extraits du corpus *Flixster* puis classés selon la note qui leur a été attribuée par l'auteur. Dix sous-ensembles de commentaires, correspondant aux dix notes possibles (de 0,5 à 5), ont ainsi été obtenus. Nous nous sommes ensuite basé sur l'hypothèse couramment admise en classification d'opinion qui dit que les adjectifs qualificatifs et les verbes sont les deux traits grammaticaux les plus utilisés afin d'exprimer une opinion. Nous avons donc filtré les mots selon leur trait grammatical et leur fréquence dans chaque sous corpus, et conservé les adjectifs et les verbes ayant le plus d'occurrences. Les mots sélectionnés apparaissant souvent dans les ensembles de commentaires associés à une note comprise entre 4 et 5 ont été intégrés dans le lexique de mots positifs et inversement avec les mots apparaissant dans les commentaires associés à une note comprise entre 0,5 et 2. Les lexiques ont ensuite été nettoyés manuellement afin de supprimer les termes n'exprimant *a priori* aucune opinion, ou encore les termes ambigus. Quelques mots, très fréquents dans les commentaires et exprimant fortement une opinion mais n'appartenant pas aux catégories des adjectifs et verbes ont également été ajoutés manuellement aux lexiques. Le nombre de mots différents étant très élevé, seuls les mots fréquemment utilisés ont été intégrés dans les lexiques.

Au final, 183 mots *a priori* porteurs d'opinion ont été classés dans deux catégories. Le lexique de mots positifs contient 115 éléments et le lexique de mots négatifs en contient 68. Nous avons volontairement restreint la taille des lexiques afin de ne conserver qu'un vocabulaire vraiment spécifique au domaine. Le tableau 7.2 présente des exemples de termes contenus dans les deux lexiques (La totalité des lexiques est consultable en annexe E). Les mots contenus dans ces deux lexiques d'opinion, que nous réunirons sous le nom de « lexique *Maison* » pour la suite, doivent permettre de classer les commentaires selon leur polarité d'opinion.

Mots positifs	good, great, funny, awesome, cool, brilliant, hilarious, favourite, well, hot, excellent, beautiful, cute, sweet ...
Mots négatifs	bad, stupid, fake, wrong, poor, ugly, silly, suck, atrocious, abominable, awful, lamentable, crappy, incompetent ...

TABLE 7.2 – Sélection parmi les 183 mots contenus dans les lexiques d'opinion

7.2.2.2 Inférence de polarités

Cette dernière étape de l'approche basée sur les lexiques consiste à comptabiliser les mots porteurs d'opinion répertoriés dans le lexique d'opinion afin de déterminer la polarité de chaque commentaire. La méthode utilisée ici est la plus basique. On calcule un score pour chaque commentaire en ajoutant 1 lorsque l'on rencontre un mot positif et en soustrayant 1 lorsque le terme rencontré est négatif, le score ayant une valeur initiale nulle. Les commentaires obtenant alors un score strictement supérieur à 0 (majorité de mots positifs) sont classés dans la catégorie *Commentaires Positifs* alors que les commentaires obtenant un score strictement inférieur à 0 (majorité de mots négatifs) sont classés dans la catégorie *Commentaires Négatifs*. Les commentaires qui obtiennent un score nul ne sont pas interprétés (voir figure 7.2).

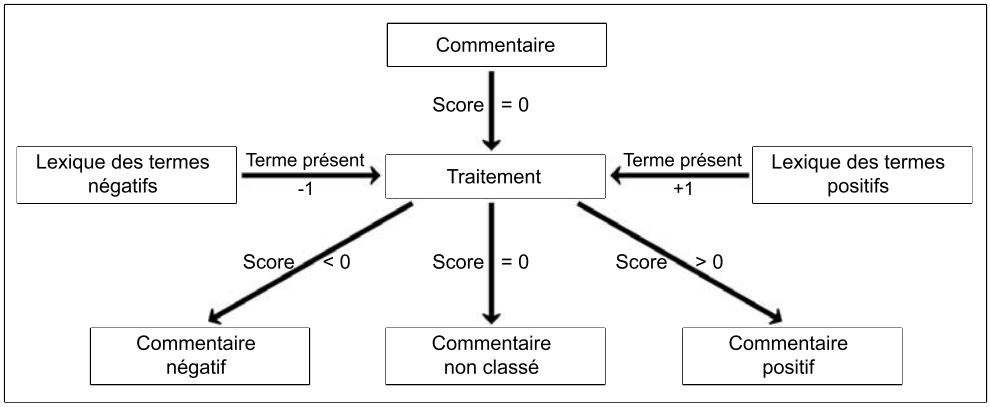


FIGURE 7.2 – Méthode de classification d'opinion

7.2.3 Résultats

Lors de la classification, seulement 66,7% des commentaires présents dans le corpus de test ont été interprétés, les commentaires restant ayant obtenu un score nul. Ce score nul survient lorsque le commentaire ne contient aucun des mots d'opinion répertoriés dans les lexiques ou lorsqu'il contient un nombre équivalent de mots positifs et de mots négatifs. En observant la répartition des commentaires classés suivant les notes originales (figure 7.3), on remarque que les commentaires bien notés sont plus souvent interprétés par le classifieur que ceux ayant une moins bonne note. En d'autres termes, plus le commentaire est positif, plus il est interprétable. Cela peut être dû au lexique d'opinion qui contient plus de mots porteurs d'opinion positive que de mots exprimant une opinion négative. Avant d'évaluer la qualité de la classification, quelques observations sur les commentaires non classés peuvent être faites.

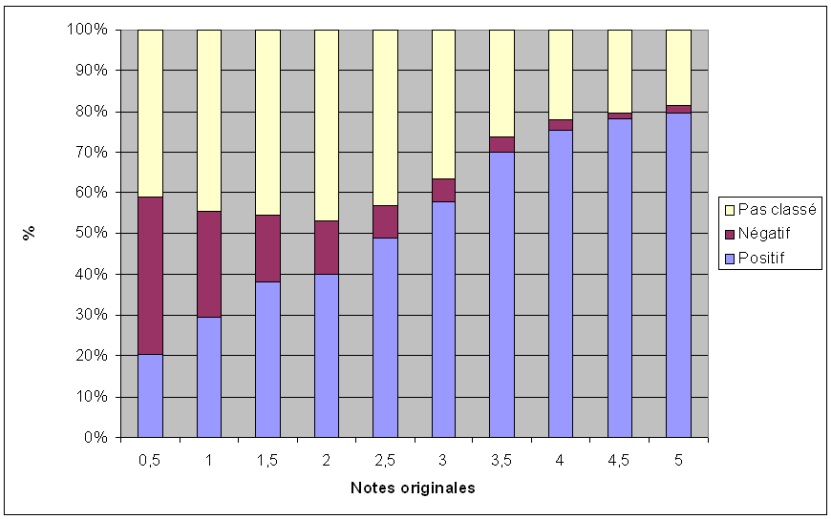


FIGURE 7.3 – Pourcentage de commentaires classés et bien classés à l'aide du lexique *Maison* selon leur note originale

7.2.3.1 Analyse des commentaires non classés

Différentes raisons peuvent expliquer ces « non-classifications » dont voici une liste non exhaustive.

Lexique trop pauvre En premier lieu, on trouve des commentaires qui expriment réellement une opinion mais qui ne contiennent que des mots non répertoriés dans les lexiques. Ces mots peuvent être de vrais mots comme dans « very overrate », « such a gross movie », « dowdy movie ». Dans ce cas, le problème provient des lexiques qui ne sont pas assez complets. Dans d'autres cas, les mots non reconnus sont des mots qui n'existent pas. Il s'agit généralement d'onomatopées ou de représentations de sons significatifs mais difficilement interprétables car ils ne sont pas toujours écrits de la même façon ; les exemples trouvés sont souvent des multiplications de lettres, de caractères ou de syllabes qui diffèrent par leur nombre comme dans « zzzzzzz », « ewwwwww » ou « ahahahahah ». Il peut également s'agir de mots inventés compréhensibles par un humain mais non connus par une machine comme par exemple le terme « awsomest ». Tous ces termes, que ce soit les onomatopées ou les mots inventés, sont restés inchangés suite aux pré-traitements car l'analyseur n'a pas su les rapprocher d'un mot connu. Les caractères de ponctuation sont également très utilisés dans les textes communautaires. Il arrive même que certains commentaires ne contiennent que ce type de signes. Il peut alors s'agir d'une suite de caractères exprimant un sentiment comme « ? ? ? ? » ou de smileys comme « :) ». Le lexique ne contenant aucun caractère de ponctuation, ces commentaires ne peuvent pas être interprétés.

Subjectivité On trouve également des commentaires n'exprimant pas explicitement l'opinion de l'auteur. Certains commentaires contiennent uniquement des émotions. Ces émotions sont difficilement interprétables car leur lien avec une polarité d'opinion est complètement subjectif. Un film « triste » parait comme tel pour la plupart des spectateurs, mais les avis sur le film en question peuvent être plus divergents du point de vue de l'appréciation. C'est par exemple le cas dans les commentaires « it's so old school », « so romantic », « very scary », « was very afraid ».

Autant de mots positifs et négatifs Une autre raison de non-classification est la présence, dans certains commentaires, d'un nombre identique de mots positifs et négatifs. « this is a *pretty bad* movie », « so *stupid* it's *funny* ! », « *awesome* movie ! ! The only thing that ruined it for me is Gwyneth Paltry she's so *annoyed* ! ! ! ». Les mots d'opinion répertoriés dans les lexiques ayant tous le même poids, la présence d'un mot de chaque classe ne permet pas au classifieur de faire son travail.

Fautes persistantes Malgré les nombreuses corrections orthographiques faites sur le texte, on trouve tout de même des fautes persistantes. Ces fautes, lorsqu'elles interviennent sur le ou les mots porteurs d'opinion, ne permettent pas d'interpréter le texte comme c'est par exemple le cas dans la phrase « this movie is one of my favories ! ! ! ». Le « t » manquant dans le mot « favorites » ne permet pas au classifieur de reconnaître le seul mot porteur d'opinion dans la phrase. Beaucoup de commentaires écrits de manière très éloignée des règles d'écriture conventionnelles sont également soit mal corrigés, soit non corrigés du tout comme « omfg I lurrrrvvveeee Johnny Dep ! ! SEXIIIIIII ... x ». Les commentaires de ce type (ceux qui ne possèdent quasiment aucun mot correctement écrit) n'ont aucune chance d'être interprétés par le classifieur à moins de répertorier toutes les fautes possibles et inimaginables dans le lexique.

Autres langues Parmi les commentaires non classés, on trouve également des commentaires rédigés dans des langues non anglophones. Le lexique ne contenant que des mots anglais ne permet pas de classer ces commentaires. On trouve par exemple du français (« Incroyablement bien pour un film d'animation. »), de l'italien (« Belissimo ! ») ou encore de l'espagnol (« Muy bien »).

Aucune opinion exprimée On trouve également des commentaires ne contenant aucune opinion et qui sont par conséquent extrêmement difficile, voire impossible, à interpréter même pour un être humain. On peut trouver des exemples ne contenant réellement aucune opinion tels que « I have the DVD but do not have the time to watch it » ou encore des textes complètement hors sujet comme « I'm a penguin, he's a penguin she's a penguin we are all penguins ! ».

Taille On remarque également que les commentaires non classés contiennent un nombre moyen de mots bien inférieur à la moyenne de tous les commentaires réunis (7 mots au

lieu de 13). Cette observation confirme que les commentaires courts sont plus difficiles à interpréter que les longs. En effet, les commentaires plus longs contiennent généralement plus de mots décrivant l'opinion, et un plus grand nombre de mots entraîne une plus grande probabilité de reconnaître et d'interpréter l'un d'eux et, par conséquent, d'interpréter le texte dans son ensemble. Nous avions pu observer dans la description des données (section 4.2) que les commentaires négatifs sont en moyenne plus courts que les commentaires positifs. Ceci peut être une explication au fait que les commentaires négatifs sont plus difficiles à interpréter que les commentaires positifs.

7.2.3.2 Analyse de la qualité de la classification

Afin de vérifier la qualité de la classification, on calcule le F_{score} sur les 66,7% de commentaires classés. Sur cette première expérimentation, il est de 63,5. On peut observer dans la matrice de confusion (tableau 7.3) que les commentaires positifs sont très bien classés (97,6% de bonnes classifications) alors que la classification des commentaires négatifs est très médiocre (seulement 21,8% de bonnes classifications). En observant visuellement les commentaires mal classés, on s'aperçoit qu'une grande partie des erreurs est due à la présence de négations qui ne sont pas du tout interprétées comme dans « this movie be not good », « sound good but be not », ou « not a great movie ». La courbe de la figure 7.3 semble confirmer cette explication. L'histogramme bleu représente le nombre de commentaires bien classés, parmi les commentaires classés, selon leur note d'origine. On remarque que les notes intermédiaires, soit les commentaires les plus ambigus, sont les plus mal classés. Une interprétation de ces résultats serait que l'utilisation de mots positifs associés à une négation, comme dans « je n'ai pas vraiment aimé », sont plus souvent utilisés afin d'exprimer un avis négatif tendant vers le neutre. D'un autre côté, les mots négatifs, comme dans « j'ai détesté » sont plus employés pour exprimer des avis vraiment négatifs. Le problème pourrait également venir du fait que les lexiques d'opinion contiennent presque deux fois plus de mots positifs que de mots négatifs, et que certains mots négatifs sont par conséquent ignorés par le classifieur.

		Vraies classes	
		NEG	POS
Classes Prédites	NEG	**21,8**	2,4
	POS	78,2	**97,6**

TABLE 7.3 – Matrice de confusion obtenue avec le lexique *Maison* (en pourcentage suivant les vraies classes)

Afin de vérifier la qualité des lexiques construits précédemment, nous avons évalué le classifieur avec un lexique d'opinion connu dans le domaine du langage naturel. Il s'agit du lexique *General Inquirer* construit par Stone et al. [SDSO66] et Kelly et Stone [KS75]

et disponible en ligne [2]. Ce lexique contient environ 13 000 lemmes étiquetés à l'aide de quelques 200 tags différents qui décrivent le cadre de l'utilisation des mots (par exemple : objet, religion, social, travail, etc.). Parmi ces nombreux tags, on trouve également les polarités (positif, négatif). Ces deux classes concernent 3 642 mots porteurs d'opinion (2 006 mots négatifs et 1 636 mots positifs), soit un nombre beaucoup plus important que le lexique *Maison* mais également une majorité de mots négatifs. L'utilisation de ce nouveau lexique a permis de classer 68,9% des commentaires du corpus de test, soit légèrement plus qu'avec le lexique *Maison* (1 115 commentaires supplémentaires ont été classés). Le F_{score} obtenu, qui est de 59,9, est inférieur à celui obtenu précédemment. Ce score peut être expliqué par le fait que le lexique *General Inquirer* a été construit pour analyser des corpus traitant de sujet généraux, et non juste des commentaires de films. On peut également constater deux changements sur la nouvelle matrice de confusion (figure 7.4). Tout d'abord, la mauvaise classification des commentaires négatifs, bien que légèrement atténuée, est toujours présente. Ensuite, une baisse de la qualité de la classification des commentaires positifs est également visible. On peut supposer que ces changements sont dûs à l'apport de nouveaux mots permettant de mieux détecter les critiques négatives. On peut également supposer que le grand nombre des erreurs est dû au problème de la négation qui n'est pas encore traité. Afin de palier à ce problème nous avons expérimenté plusieurs solutions de prise en compte des négations.

		Vraies classes	
		NEG	POS
Classes Prédites	NEG	**33,3**	14,6
	POS	66,7	**85,4**

TABLE 7.4 – Matrice de confusion obtenue avec le lexique *General Inquirer* (en pourcentage suivant les vraies classes)

7.2.3.3 Prise en compte des négations

Le traitement de la négation est une tâche généralement difficile, et elle l'est d'autant plus avec les textes communautaires qui ne respectent pas toujours les règles essentielles de grammaire et d'orthographe. Les deux lexiques contenant les mots d'opinion, le *Maison* et *General Inquirer*, ont été testés avec chacune des méthodes. Les termes de négation pris en compte pour les expérimentations qui vont suivre sont les suivant : « no », « not », « never » et « less ». « Less » joue plus souvent le rôle d'un *minimiseur* que d'une négation, mais il est tout de même généralement utilisé pour influer sur la polarité de l'adjectif qui le suit. C'est pourquoi il est ici considéré comme les négations.

2. http ://www.webuse.umd.edu :9090/tags/

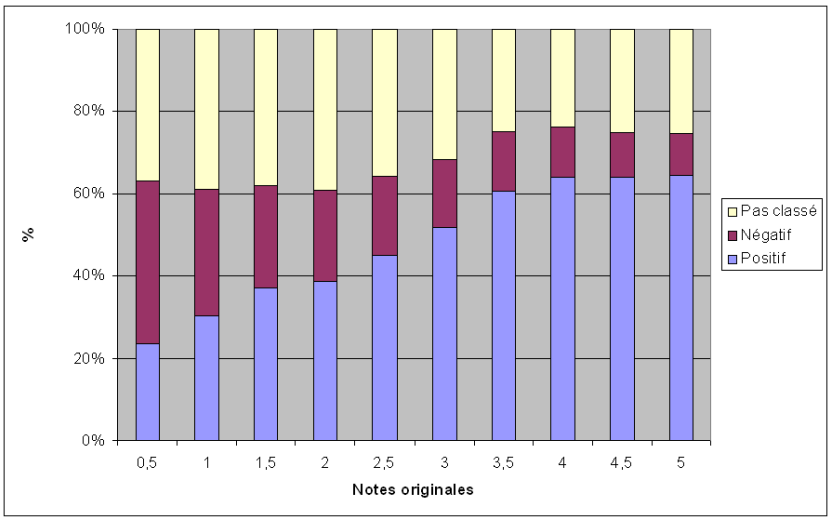


FIGURE 7.4 – Pourcentage de commentaires classés et bien classés à l'aide du lexique *General Inquirer* selon leur note originale

Inversion de la polarité du commentaire La première solution testée consiste à inverser la polarité du commentaire lorsqu'il contient une négation. En effet, les commentaires étant très courts, ils ne contiennent généralement qu'une seule phrase. On peut donc logiquement supposer que le nombre de termes d'opinion présents dans le commentaire sont d'une quantité très limitée et que la ou les quelques négations présentes jouent un rôle sémantique sur l'ensemble du texte. Avec cette méthode, le nombre de commentaires classés n'est pas modifié puisque les scores nuls obtenus précédemment restent nuls. Le pourcentage de commentaires classés est donc de 66,7% avec l'utilisation du lexique *Maison* et de 68,9% avec l'utilisation du lexique *General Inquirer*. Les F_{scores} obtenus avec cette solution sont de 62,3 en utilisant le lexique *Maison*, et de 59,7 en utilisant le lexique *General Inquirer*. Ils sont légèrement moins bons que ceux obtenus avec la méthode précédente, qui ne considérait pas les négations. Les matrices de confusion (figures 7.5 et 7.6) montrent une amélioration de la classification des commentaires négatifs (qui reste tout de même très mauvaise), mais également une baisse de performance de la classification des commentaires positifs.

		Vraies classes	
		NEG	POS
Classes Prédites	NEG	**38,5**	15,2
	POS	61,5	**84,8**

TABLE 7.5 – Matrice de confusion obtenue avec le lexique *Maison* et la Solution de prise en compte de la négation 1 (en pourcentage suivant les vraies classes)

		Vraies classes	
		NEG	POS
Classes Prédites	NEG	**44,1**	24,2
	POS	55,9	**75,8**

TABLE 7.6 – Matrice de confusion obtenue avec le lexique *General Inquirer* et la Solution de prise en compte de la négation 1 (en pourcentage suivant les vraies classes)

Inversion de la polarité du mot suivant la négation La première solution de prise en compte de la négation n'ayant pas convaincu, une solution plus complexe a été mise en place. Elle consiste à inverser uniquement la polarité du mot d'opinion qui suit la négation plutôt qu'inverser la polarité de tout le commentaire. Techniquement, une variable booléenne, initialisée à « false », est placée à « true » lorsque une négation est trouvée. Chaque fois qu'un mot appartenant au lexique d'opinion est repéré, la variable booléenne est consultée. Si elle est à « true », la polarité du mot d'opinion est inversée, si la variable est à « false » la polarité ne change pas. La quantité de commentaires classés est de 64,8% soit moindre qu'avec la méthode sans prise en compte de la négation. La qualité de classification est, elle, légèrement améliorée avec un F_{score} de 65,8 concernant le lexique *Maison* et de 62,2 avec le lexique *General Inquirer* (calculé sur les 68,3% de commentaires classés). Cette prise en compte de la négation, plus précise que la précédente, semble également être plus efficace. On peut toutefois observer, dans les matrices confusion (figures 7.7 et 7.8), que les commentaires négatifs sont toujours extrêmement mal classés.

		Vraies classes	
		NEG	POS
Classes Prédites	NEG	**30,7**	4,3
	POS	69,3	**95,7**

TABLE 7.7 – Matrice de confusion obtenue avec le lexique *Maison* et la Solution de prise en compte de la négation 2 (en pourcentage suivant les vraies classes)

		Vraies classes	
		NEG	POS
Classes Prédites	NEG	**40,1**	15,9
	POS	59,9	**84,1**

TABLE 7.8 – Matrice de confusion obtenue avec le lexique *General Inquirer* et la Solution de prise en compte de la négation 2 (en pourcentage suivant les vraies classes)

La négation comme un mot à polarité négative La considération des négations comme un cas particulier n'a pas apporté les résultats escomptés. Le fait que les mauvais résultats interviennent quasi uniquement sur les commentaires appartenant à la classe négative, on peut supposer que les négations ne sont uniquement utilisée que dans le cas de l'expression d'avis négatifs. Cette nouvelle solution de prise en compte de la négation consiste donc à considérer les négations non plus comme des cas particuliers mais comme des mots à connotation négative. Concrètement, les éléments présents dans le lexique des négations ont été ajoutés au lexique contenant les mots d'opinion négative. Les résultats obtenus avec cette solution sont légèrement meilleurs que ceux obtenus précédemment. Concernant le pourcentage de commentaires bien classés, il est de 65,7 avec le lexique *Maison* et de 67,7 avec le lexique *General Inquirer*. Les F_{scores} obtenus sur ces commentaires classés sont de 66,9 avec le lexique *Maison* et 62,7 avec le lexique *General Inquirer*. Les matrices de confusions 7.9 et 7.10 montrent qu'une nouvelle amélioration est apportée au niveau de la classification des commentaires négatifs.

		Vraies classes	
		NEG	POS
Classes Prédites	NEG	**35,6**	5,3
	POS	64,4	**94,7**

TABLE 7.9 – Matrice de confusion obtenue avec le lexique *Maison* et la Solution de prise en compte de la négation 3 (en pourcentage suivant les vraies classes)

		Vraies classes	
		NEG	POS
Classes Prédites	NEG	**44,1**	18,3
	POS	55,9	**81,7**

TABLE 7.10 – Matrice de confusion obtenue avec le lexique *General Inquirer* et la Solution de prise en compte de la négation 3 (en pourcentage suivant les vraies classes)

Pondération de la négation Le choix de cette dernière solution relève plus d'une intuition que d'une véritable logique. Les précédents résultats tendent à montrer que les négations jouent un rôle prépondérant dans l'expression des avis et sentiments. Nous avons donc souhaité renforcer leur impact sur la classification en leur accordant plus d'importance lors de l'analyse. Pour cela, nous avons doublé la valeur de leur poids. Lorsque le classifieur rencontre un mot positif ou négatif, il ajoute ou soustrait 1 point au score du texte, et lorsqu'il rencontre une négation, le score est abaissé de 2 points. Les résultats sont une nouvelle fois légèrement améliorés grâce à cette solution. Tout d'abord, concernant la quantité de commentaires classés, les pourcentages sont plus élevés que tous ceux obtenus jusqu'alors : 69,4% de commentaires classés avec le lexique *Maison* et 71,2% avec le lexique

General Inquirer. Les F_{scores} obtenus avec les deux lexiques sont respectivement de 67,2 et 63,6. Les matrices de confusion 7.11 et 7.11 montrent que la classification des commentaires négatifs s'est encore améliorée (bien qu'elle soit tout de même encore assez médiocre) mais également que la classification des commentaires positifs s'est légèrement détériorée, bien qu'ayant toujours de très bons résultats.

		Vraies classes	
		NEG	POS
Classes	NEG	**45,3**	11,2
Prédites	POS	54,7	**88,8**

TABLE 7.11 – Matrice de confusion obtenue avec le lexique *Maison* et la Solution de prise en compte de la négation 4 (en pourcentage suivant les vraies classes)

		Vraies classes	
		NEG	POS
Classes	NEG	**51,8**	23,1
Prédites	POS	48,2	**76,9**

TABLE 7.12 – Matrice de confusion obtenue avec le lexique *General Inquirer* et la Solution de prise en compte de la négation 4 (en pourcentage suivant les vraies classes)

7.2.4 Discussion

Concernant l'approche basée sur les lexiques, plusieurs méthodes de classification ont été comparées ainsi que différents lexiques d'opinion. Les lexiques contiennent des mots porteurs d'opinion ainsi que leur polarité, c'est-à-dire s'ils expriment une opinion positive ou une opinion négative. Ces listes de mots sont ensuite utilisées afin de classer les textes en fonction de poids attribués au mots d'opinion. Les meilleurs résultats de classification ont été obtenus à l'aide du lexique *Maison* et de la classification accordant plus d'importance aux négations qu'aux autres termes d'opinion (poids multipliés par deux). Deux niveaux d'observation des résultats sont possibles. Tout d'abord, l'observation des commentaires non classés apporte des informations sur la nature des textes. Ensuite, l'observation des commentaires mal classés apporte de nouvelles informations.

Concernant les commentaires non classés, la première information que l'on peut extraire concerne les deux lexiques utilisés pour les expérimentations qui semblent ne pas être assez complets. En effet, la taux de classification le plus élevé obtenu lors des expérimentations est seulement de 71%. Cela signifie que le classifieur n'arrive pas à interpréter plus d'un commentaire sur quatre. On remarque également que la quantité de commentaires classés n'est que très peu améliorée avec le lexique *General Inquirer* (score supérieur d'environ 2%) qui répertorie pourtant 20 fois plus de mots d'opinion que le lexique *Maison*

(3 642 contre 183). Les mots d'opinion seuls ne permettent donc pas d'interpréter tous les textes, notamment quand il s'agit de textes communautaires répondant à des règles spécifiques. Comme nous avons pu voir lors de la description des données (chapitre 4), les onomatopées sont très utilisées par les adeptes du DEM afin de partager leurs sentiments (« zzzzzzzzz », « beurk », etc.) alors qu'elles ne sont pas répertoriées dans les lexiques. Une autre problématique est la non interprétation des expressions. En effet, les lexiques utilisés ne contiennent que des termes uniques alors que parfois il est nécessaire d'interpréter plusieurs mots co-occurents afin de déduire l'opinion qui est exprimée : « I want my money back », « I fall asleep », etc. On peut également noter la présence dans les commentaires non classés d'abréviations ou de fautes d'orthographes non corrigées par l'analyseur linguistique (rappelons que cet outil n'a pas été conçu à l'origine pour traiter des corpus relevant du DEM). Ces observations laissent à penser que l'obtention d'un plus grand pourcentage de commentaires classés passe nécessairement par une amélioration des lexiques. Nous avons également pu observer qu'un lexique construit manuellement apporte de meilleurs résultats qu'un lexique construit automatiquement. En sachant que le corpus d'apprentissage contient environ 60 000 termes différents, l'amélioration du lexique *Maison* demanderait un travail extrêmement coûteux en terme de temps.

Concernant les différentes classifications, elles ont engendré des résultats assez homogènes. La figure 7.5 présente les pourcentages de bonnes réponses obtenus par chacune des classification suivant les dix classes originales. On peut observer, sans grande surprise, que les classes intermédiaires sont celles ayant obtenu les moins bons résultats. On observe également que la prise en compte des négations, quelle qu'en soit la manière, améliore grandement les résultats obtenus sur les classes négatives (0,5 à 3,5). La négation parait donc beaucoup plus importante pour les classes négatives que pour les classes positives et sa prise en compte est une problématique délicate. Une analyse syntaxique parait donc nécessaire pour interpréter convenablement les négations mais au vu de la nature du corpus (textes courts, phrases mal ou peu construites, etc.), l'utilisation d'un outil spécialement dédié à ce type de textes nous semble inévitable pour ce genre d'analyses et nous ne le possédons pas. Nous pouvons également noter la présence dans les textes de termes exprimant des émotions tels que « sad » ou « scary ». Ces termes sont fréquemment utilisés afin d'exprimer l'opinion mais, sans le contexte qui les accompagne, ils apportent peu d'information au vu de leur subjectivité.

Les expérimentations basées sur les lexiques et les observations des différents résultats montrent que l'utilisation de méthodes d'analyse beaucoup plus complexes et beaucoup plus coûteuses paraissent nécessaires si l'on souhaite améliorer les résultats. Les méthodes envisageables, qui sont l'amélioration manuelle des lexiques d'opinion ainsi que l'analyse syntaxique des textes, requièrent soit du temps, soit des connaissances que nous ne possédons pas. De plus, l'amélioration manuelle des lexiques irait à l'encontre de l'un des objectifs de ces travaux de thèse qui est l'automatisation des méthodes et de la chaîne de traitement. Pour toutes ces raisons, nous avons fait le choix d'explorer les méthodes existantes dans le domaine de l'apprentissage automatique.

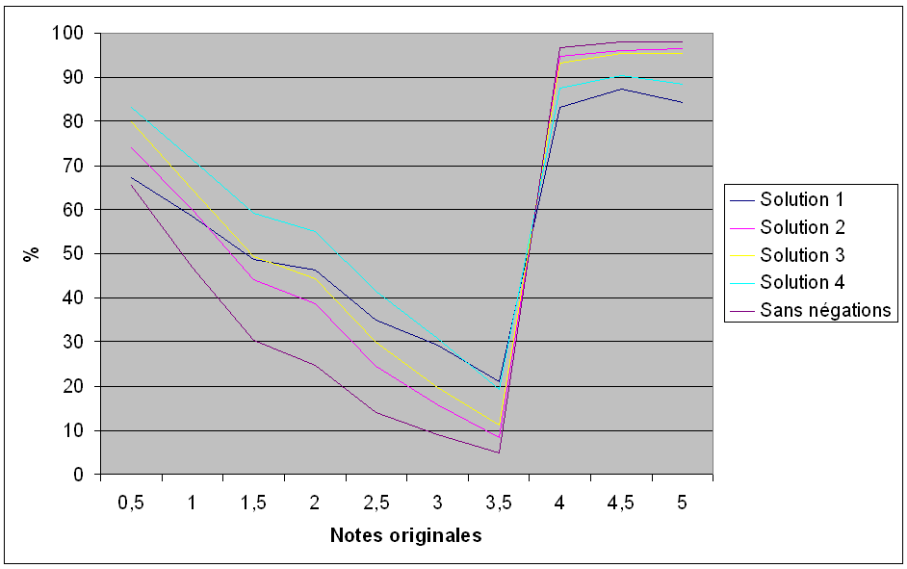


FIGURE 7.5 – Pourcentage de bonnes classifications selon les notes d'origines pour toutes les expérimentations avec le lexique *Maison*

7.3 Approches basées sur l'apprentissage supervisé

Dans cette partie, nous présentons les différentes classifications menées à l'aide de l'apprentissage supervisé. Comme il a été vu dans l'état de l'art (Section 6.3), la tâche de classification supervisée entraîne différents choix à faire concernant les classes de prédiction, les pré-traitements, la segmentation, la représentation et le classifieur. Dans une première section, nous énumérons les différents choix effectués concernant tous ces critères. Nous présentons ensuite les évaluations sur chacun d'entre eux. Les solutions ont été choisies, pour la plupart, en fonction des méthodes les plus répandues dans la littérature. L'objectif est alors d'évaluer si ces traitements habituellement réalisés sur des textes « normaux » (critiques professionnelles, débats politiques, avis de consommateurs convenablement rédigés, etc.) sont adaptés au traitement des textes communautaires et à leurs spécificités.

7.3.1 Solutions choisies

7.3.1.1 Corpus d'apprentissage

Afin de ne pas introduire de biais dans la comparaison des résultats des expérimentations, nous avons fait le choix d'équilibrer la représentation des notes dans les corpus suivant les cinq classes nommées *1* à *5*. Cet équilibrage a été réalisé en sélectionnant aléatoirement 35 000 commentaires par classe. Le corpus contient donc 175 000 commentaires. Les commentaires appartenant à ce corpus ont été sélectionnés de manière à n'avoir aucun lien avec le corpus dédié à la recommandation. C'est à dire que le corpus présenté dans la section 4.2 qui contient 3 300 000 triplets utilisateur-film-commentaire n'a aucun film ou utilisateur

en commun avec le corpus d'apprentissage. Pour finir, nous avons également fait en sorte d'avoir, au maximum, dix commentaires rédigés par le même auteur. Ce choix entraîne la présence d'un plus grand nombre d'utilisateurs dans le corpus d'apprentissage et donc une plus grande représentativité des différents styles d'écriture que l'on peut trouver sur le site.

7.3.1.2 Pré-traitements

Plusieurs pré-traitements ont été appliqués sur les textes afin d'évaluer leur impact sur la classification. Des pré-traitements minimaux sont appliqués à tous les textes puis sont ajoutés indépendamment des suppressions de mots vides à l'aide d'anti-dictionnaires et des pré-traitements linguistiques.

Pré-traitements minimaux Nous avons tout d'abord appliqué des pré-traitements que nous pouvons qualifier de « minimes » mais néanmoins nécessaires à la réduction du nombre de termes présents dans le corpus. La taille du vocabulaire étant conséquente (environ 150 000 termes si l'on considère uniquement le caractère espace comme séparateur), la première étape consiste à distinguer et séparer les caractères alphanumériques des autres. S'en suit une mise en minuscule de tous les caractères alphabétiques. Ces traitements mineurs ont permis de réduire la taille du vocabulaire à 60 000 termes environ. Après vérification du non-impact sur la classification, nous avons également supprimé tous les caractères ayant un nombre d'occurrences inférieur à trois dans l'ensemble du corpus d'apprentissage. La suppression des mots très peu fréquents permet une nouvelle réduction de la taille du vocabulaire d'un facteur 3, le nombre de mots restants étant de 19 255, ce qui est largement plus facile à exploiter par les outils d'apprentissage automatique que les quelques 150 000 présents à l'origine.

Utilisation d'un anti-dictionnaire Trois anti-dictionnaires disponibles sur Internet ont été comparés. Ils ont été sélectionnés de manière à représenter un panel de tailles assez variées. Le premier d'entre eux contient 671 mots vides [3], le deuxième en contient 319 [4] et pour finir nous avons pris la liste de mots vides utilisés par le moteur de recherche *Google* qui contient 31 termes. Ces anti-dictionnaires ont tous été appliqués aux textes ayant subi les pré-traitements minimaux.

Pré-traitements linguistiques Les pré-traitements linguistiques effectués ici sont identiques à ceux effectués dans l'approche basée sur les lexiques soit des corrections orthographiques (correction phonétique, correction par troncature, tolérance aux répétitions, correction typographique et correction morphologique) ainsi qu'une lemmatisation. L'analyseur syntaxique conserve tels quels les mots qu'il ne reconnait pas, c'est pourquoi, comme pour les pré-traitements minimaux, tous les termes ayant un nombre d'occurrences inférieur à trois dans le corpus d'apprentissage sont supprimés.

3. www.ranks.nl/resources/stopwords.html
4. http ://armandbrahaj.blog.al/2009/04/14/list-of-english-stop-words/

7.3.1.3 Segmentations

Deux solutions de segmentation ont été retenues.

Les mots Pour la première, les délimiteurs choisis pour la construction du vocabulaire sont l'espace et tous les caractères de ponctuation, y compris l'apostrophe. La ponctuation est conservée et fait ainsi partie du vocabulaire, contrairement à la majorité des méthodes rencontrées dans la littérature. En effet, nous avons pu voir dans la section décrivant la nature des textes communautaires (section 4.1) que les caractères de ponctuation semblent jouer un rôle très important, ne serait-ce que par la présence d'émoticônes.

Les tri-grammes de lettres La deuxième segmentation choisie est le découpage en tri-grammes de lettres qui, comme nous l'avons vu dans l'état de l'art, est une segmentation souvent employée en fouille de textes. Pour rappel, cette segmentation consiste à découper le texte selon une fenêtre glissante faisant une taille de 3 caractères. Ici, l'espace est considéré comme un caractère et non un délimiteur.

7.3.1.4 Représentations

Les représentations sélectionnées sont toutes des représentations vectorielles. Comme nous l'avons vu dans l'état de l'art (Chapitre 6), ce type de représentations est très largement privilégié dans le domaine de la fouille d'opinion. Les solutions choisies sont les trois représentations les plus populaires de la littérature, à savoir la représentation fréquentielle, la représentation fréquentielle normalisée et la représentation TF-IDF.

7.3.1.5 Classifieurs

Trois classifieurs, basés sur deux approches différentes (Machine à Vecteurs de Support et Naïf Bayésien), ont été évalués sur les données.

SVM (pour Support Vector Machine) Nous avons utilisé l'outil SVM^{light} [5]. Il s'agit d'une implémentation de la Machine à Vecteur Support de Vapnik [Vap95]. Les algorithmes d'optimisation utilisés sont décrits dans Joachims [Joa02] et Joachims [Joa99a].

SNB (pour Selective Naive Bayes) Nous avons utilisé l'outil KHIOPS [6] qui est un classifieur naïf bayésien avec sélection de variables [Bou05, Bou06, Bou07b]. Dans une première phase de préparation de données, les variables explicatives sont évaluées individuellement au moyen d'une méthode de discrétisation optimale dans le cas numérique et de groupement de valeurs optimal dans le cas catégoriel. Dans la phase de modélisation, un modèle de classification est construit, en moyennant efficacement un grand nombre de modèles basés sur des sélections de variables.

5. Outil téléchargeable en freeware sur http ://svmlight.joachims.org/
6. Outil téléchargable en shareware sur http ://perso.rd.francetelecom.fr/boulle/

NB (pour Naive Bayes) Nous avons également utilisé KHIOPS sans le mode de sélection de variables, ce qui revient à utiliser un classifieur naïf bayésien classique.

7.3.2 Protocole et résultats

Le tableau 7.13 récapitule les choix effectués. Les expérimentations ont été faites dans l'ordre suivant : nous avons tout d'abord évalué chaque classifieur sur une classification binaire, avec les pré-traitements minimaux et avec les trois représentations sélectionnées. Nous avons ensuite conservé le classifieur donnant le meilleur résultat puis nous l'avons évalué avec les autres pré-traitements. Afin de comparer toutes les expérimentations, nous calculons le F_{score} qui est l'évaluation la plus couramment utilisée en classification d'opinion. Pour finir nous avons conservé le meilleur modèle et nous avons effectué les classifications sur trois classes et sur cinq classes.

Pré-traitements	Segmentations	Représentations	Classifieurs
Minimaux	Mots	Fréquentielle	SVM
Anti-dictionnaires	Tri-grammes	Fréquentielle normalisée	SNB
Linguistiques		TF-IDF	NB

TABLE 7.13 – Récapitulatif des choix effectués pour les évaluation des classifications

7.3.2.1 Évaluation des classifieurs

Les trois classifieurs sélectionnés ont tous été évalués avec la segmentation en mots et les trois représentations possibles : la fréquentielle, la fréquentielle normalisée et le TF-IDF. Les pré-traitements minimaux ont été appliqués aux corpus afin de réduire le nombre de dimensions du vecteur de représentation des textes. Les attributs composant ce vecteur de représentation sont donc tous les mots du corpus d'apprentissage mis en minuscule excepté ceux apparaissant moins de trois fois dans le corpus d'apprentissage. Les résultats des différentes classifications sont présentés dans le tableau 7.14. C'est la méthode SVM associée à la représentation fréquentielle qui donne le meilleur résultat. Il est à noter que le classifieur SNB et le classifieur SVM obtiennent des résultats quasi-similaires avec les deux autres représentations alors que les résultats sont légèrement inférieurs avec le classifieur NB, et ce quelle que soit la représentation.

Le classifieur naïf bayésien est donc plus performant avec la sélection de variables (SNB). Il est d'ailleurs intéressant d'étudier ces variables sélectionnées automatiquement et jugées les plus informatives. Leur nombre varie de 635 à 741 selon la représentation choisie mais les plus informatives d'entre elles sont les mêmes dans tous les cas. Le tableau 8.3 présente les quarante variables les plus informatives ordonnées selon leur degré d'information. La variable la plus informative est le point d'exclamation, suivi par le mot

Classifieurs \ Représentations	Fréquentielle	Fréq. normalisée	TF-IDF
SVM	**0,741**	0,724	0,726
SNB	0,726	0,729	0,728
NB	0,695	0,708	0,708

TABLE 7.14 – F_{score} obtenu pour chaque représentation avec les trois classifieurs

« love », puis « loved » et ainsi de suite. L'analyse de la liste des variables donne des informations intéressantes sur le contenu des commentaires et on peut notamment extraire des renseignements sur l'impact que pourraient avoir certains pré-traitements couramment utilisés dans le domaine.

!	best	sucked	lame
love	but	excellent	horrible
loved	movie	was	okay
not	stupid	alright	too
ok	t	waste	luv
boring	worst	awsome	nothing
awesome	.	didn	fantastic
great	crap	terrible	is
amazing	brilliant	'	this
bad	hilarious	wasn	no

TABLE 7.15 – Variables les plus informatives trouvées par la méthode SNB

On remarque notamment que les caractères de ponctuation sont très représentés. Le point d'exclamation est d'ailleurs en tête de liste des variables informatives. Il est en effet un indicateur fort des commentaires positifs. Lorsqu'il apparait une fois dans un commentaire, il y a 52% de chance que ce commentaire soit positif, et lorsqu'il apparait plus de trois fois, ce score passe à 67%. Concernant le point, 66% des commentaires qui en contiennent deux ou plus sont des commentaires négatifs. Enfin, 74% des commentaires contenant un point d'interrogation sont des commentaires négatifs. On peut déduire de ces chiffres que la suppression de la ponctuation n'est pas un pré-traitement recommandé pour ce type de textes.

Le cas de la négation est souvent discuté dans la littérature et la prise en compte de celle-ci est une des problématiques de la classification d'opinion [PL08]. Les négations jouent en effet un grand rôle dans l'expression de l'opinion mais, en observant leur répartition dans les classes, on s'aperçoit qu'elles ne sont pas forcément un cas particulier. Par exemple, lorsque le mot « not » apparait dans un texte, il s'agit, dans 82% des cas, d'un commentaire négatif. À titre de comparaison, ce chiffre est de 86% pour le mot « bad ». Les commentaires contenant les mots « no », ou « didn't » ou encore « wasn't » sont également des commentaires plus souvent négatifs que positifs : 77% pour « no », 80% pour « didn't » et

83% pour « wasn't ». On peut donc en déduire que dans ce type de textes, un traitement particulier pour le cas de la négation n'est pas forcément nécessaire car les négations ont un comportement proche des mots à polarité négative.

On remarque également la présence de mots ayant la même racine, que ce soit dû à la conjugaison de ceux-ci ou à des fautes d'orthographe, volontaires ou involontaires. C'est par exemple le cas avec les mots « love », « loved », « luv » ou « awesome », « awsome » ou « great », « gr8 », ou encore « was », « is », « been », etc. Les exemples sont nombreux. L'union de ces différents mots en une seule variable pourrait apporter à la fois une réduction de l'espace des dimensions, et peut-être également un gain d'information pour la classification.

Les différents pré-traitements sont évalués dans cette perspective, à savoir la réduction de l'espace des dimensions. Nous comparons les résultats obtenus avec les meilleurs résultats obtenus jusqu'alors, soit ceux correspondant au classifieur SVM. Les variables informatives obtenues avec le classifieur SNB sont utilisées afin de mieux interpréter la nature des nouveaux résultats.

7.3.2.2 Évaluation des pré-traitements

L'impact des pré-traitements a été évalué à l'aide de la méthode SVM et des trois méthodes de représentations.

Anti-dictionnaire Les meilleurs résultats obtenus sont une nouvelle fois dûs à la représentation fréquentielle (tableau 7.16). Ces F_{scores} sont toutefois inférieurs au meilleur résultat obtenu sans la suppression des mots vides. On peut supposer que cette baisse de performance est due à la suppression de certains mots informatifs. En effet, les anti-dictionnaires contiennent un grand nombre des variables informatives extraites lors de l'évaluation des classifieurs.

Représentations / Pré-traitements	Fréquentielle	Fréq. normalisée	TF-IDF
Minimaux	**0,741**	0,724	0,726
Anti-dictionnaire (671)	0,711	0,694	0,697
Anti-dictionnaire (319)	0,720	0,700	0,703
Anti-dictionnaire (31)	0,737	0,722	0,725

TABLE 7.16 – F_{score} obtenus pour sans et avec les anti-dictionnaires et toutes les représentations avec le classifieur SVM

Le plus gros anti-dictionnaire (celui contenant les 671 mots vides) contient 145 termes présents dans les variables informatives sélectionnées précédemment par la méthode SNB. Dans ces variables, on retrouve notamment un grand nombre de négations qui, comme nous

l'avons vu, sont très importante en fouille d'opinion. Les négations qu'il contient sont par exemple « not », « no », « never », « didn't », « wasn't », « nothing », etc. Il contient également un grand nombre d'adverbes qui sont souvent utilisés pour accentuer l'opinion exprimée dans le texte comme par exemple « really », « very », « unfortunately », « so », etc. Il contient également des interjections, des déterminants ou des verbes tels que « but », « was », « is », « at », « to », « for », qui ne sont pas des mots au sens sémantique très fort mais dont la présence semble apporter une information non négligeable pour la détermination de l'opinion. Le deuxième anti-dictionnaire contient exactement 100 mots qui font également partie des variables informatives trouvées par KHIOPS. Enfin, le dernier dictionnaire, composé presque uniquement d'articles, contient tout de même 19 variables jugées informatives, ce qui représente pratiquement 60% de sa totalité. On peut donc déduire de ces expérimentations que l'utilisation d'anti-dictionnaires pour la classification de textes collaboratifs dégrade systématiquement les résultats.

Pré-traitements linguistiques Rappelons tout d'abord que les pré-traitements linguistiques appliqués sont des corrections orthographiques suivie d'une lemmatisation. Ces traitements permettent une réduction du nombre de dimensions d'environ 25%. Le nombre de variables est maintenant de 14 224. Ceci a donc un impact sur les temps de calcul qui ne sont pas négligeables pour la tâche de classification. Néanmoins, les pré-traitements sont plus importants et les deux tâches mises bout à bout sont plus lourdes que lors de l'expérimentation sans les pré-traitements linguistiques. De plus, les F_{score} obtenus ici sont très proches de ceux des premières expérimentations mais légèrement inférieurs (tableau 7.17). Cette très légère baisse des performances pourrait être expliquée par le fait que la lemmatisation annihile le degré d'information de certaines variables. Par exemple, on se rend compte, grâce aux variables informatives obtenues par la méthode SNB, que le temps utilisé pour la conjugaison des verbes joue un rôle non négligeable pour la prédiction de l'opinion. C'est le cas avec le verbe « to be » par exemple. En effet, 66% des commentaires contenant le mot « was » sont négatifs alors que c'est le cas pour seulement 51% des commentaires contenant « is ». La fusion de ces deux variables en une seule entraîne donc un nivellement de la quantité d'information contenue dans chacune d'entre elles.

Pré-traitements \ Représentations	Fréquentielle	Fréq. normalisée	TF-IDF
Minimaux	**0,741**	0,724	0,726
Linguistiques	0,738	0,722	0,724

TABLE 7.17 – F_{scores} obtenus avec les pré-traitements minimaux et linguistiques et toutes les représentations possibles avec le classifieur SVM

7.3.2.3 Évaluation des segmentations

Les pré-traitements testés n'améliorant pas les résultats de la classification, nous avons donc testé une autre segmentation dite en tri-grammes de lettres. Cette segmentation a

été appliquée aux textes n'ayant subit que les pré-traitements minimaux. Le nombre de variables s'élève à 49 136. Le tableau 7.18 montre qu'une nouvelle fois, les nouveaux résultats sont inférieurs aux premiers résultats obtenus.

Segmentations \ Représentations	Fréquentielle	Fréq. normalisée	TF-IDF
Mots	**0,741**	0,724	0,726
Tri-grammes	0,727	0,714	0,715

TABLE 7.18 – F_{scores} obtenus avec les segmentations en mots et en tri-grammes et toutes les représentations possibles avec le classifieur SVM

7.3.2.4 Évaluation des classes de prédiction

Les commentaires du corpus étant originellement répartis sur dix classes, nous avons jugé qu'il était nécessaire de réduire ce nombre en regroupant certaines classes entre elles. Nous avons donc testé plusieurs regroupements avec un nombre final de classes différent pour chacun. Nous avons évalué des classifications sur deux, trois et cinq classes à l'aide du modèle qui a obtenu les meilleurs résultats jusqu'à présent, c'est-à-dire le classifieur SVM, avec des pré-traitements minimaux, une représentation fréquentielle et une segmentation en mots. La comparaison des résultats de chaque classification s'avère délicate. En effet, les erreurs faites sur cinq classes n'ont pas nécessairement le même poids que les erreurs faites sur deux ou trois classes. Nous avons donc décidé de vérifier la qualité de ces classifications à l'aide de la tâche de recommandation (section 7.4). Les tableaux 7.19 et 7.20 sont les matrices de confusion obtenues pour les classification sur trois et cinq classes.

		Vraies classes		
		NEG	NEUTRE	POS
Classes Prédites	NEG	**71,7**	34,5	11
	NEUTRE	15,1	**34,5**	17,4
	POS	13,2	31	**71,6**

TABLE 7.19 – Résultats de la classification d'opinion sur 3 classes (en pourcentage suivant les vraies classes)

7.3.3 Discussion

Le meilleur résultat obtenu, concernant les classifications binaires, est donc dû au classifieur SVM associé aux pré-traitements minimaux et à une représentation fréquentielle en sac de mots. Le F_{score} obtenu est de 0,741. Nous avons souhaité observer les commentaires mal classés par cette méthode. Pour la grande majorité d'entre eux, il n'est pas difficile de comprendre pourquoi le modèle s'est trompé. Nous citons ici les cas d'erreurs les plus

		Vraies classes				
		1	2	3	4	5
Classes Prédites	1	**60,3**	29,2	14,9	7,3	5,4
	2	16,8	**30,7**	22,4	7,2	3,5
	3	7,1	15,2	**23,2**	14,6	6,8
	4	6	10,7	18,8	**24,9**	15
	5	9,8	14,2	20,7	46	**69,3**

TABLE 7.20 – Résultats de la classification d'opinion sur 5 classes (en pourcentage suivant les vraies classes)

fréquents, qui ne sont bien entendu pas exhaustifs.

Pour certains textes, la cause est due à une ou plusieurs fautes d'orthographe empêchant le modèle de repérer un mot important comme dans l'exemple suivant qui a été classé comme commentaire positif : « it wasent a very good movie ». D'autres peuvent être mal classés car ils ne contiennent que des informations subjectives comme le commentaire : « so sad » qui a été classé comme négatif alors que l'auteur a attribué une note de 4,5 au film ou encore « it was very crazy ! » qui a été classé positif alors que l'auteur n'a pas du tout apprécié le film. L'explication ici est donc simple, les utilisateurs ayant rédigé ces commentaires ont des goûts divergents par rapport à la majorité des utilisateurs présents dans le corpus d'apprentissage. On entend par là que la plupart des utilisateurs sondés semblent ne pas aimer les films tristes et apprécier les films « fous ». Enfin nous pouvons également citer les nombreux cas où l'auteur du commentaire est en cause car beaucoup d'entre eux se trompent en attribuant la note accompagnant son commentaire ou bien laisse la note attribuée par défaut qui est 0,5. Parfois, il n'y a donc aucune cohérence entre la note et le contenu du commentaire.

7.4 Évaluation de la classification par la recommandation

7.4.1 Résultats

Afin d'évaluer le système de recommandation avec les résultats de la classification d'opinion, nous faisons appel à la méthode collaborative. En effet, en sortie de la classification, les données obtenues sont des données d'usages de la forme utilisateur-item-note. Nous avons évalué les notes obtenues avec la meilleure classification correspondant à l'emploi des pré-traitements minimaux, de la segmentation en mots, de la représentation fréquentielle et du classifieur SVM. Le résultat obtenu en RMSE est de 0,898, ce qui est moins bon que les résultats obtenus lors de l'étalonnage du moteur en mode collaboratif (0,862) mais meilleur que le résultat obtenu en mode thématique avec les données IMDB (0,921) (Voir section 3.2). Afin de mesurer l'apport de la classification d'opinion, nous avons également entraîné le système sur les vraies notes données par les auteurs des commentaires et sur

des notes aléatoires (en conservant les relations entre utilisateurs et films). Les RMSE obtenues sont respectivement 0,897 et 0.989 (pour les notes aléatoires nous avons testé 3 tirages aléatoires). On constate donc que la classification d'opinion entraîne d'aussi bons résultats que l'utilisation des vraies notes et que les notes sont importantes car avec des notes aléatoires, les performances s'effondrent. Nous avons également souhaité évaluer l'importance des relations utilisateur-item. Pour cela nous avons construit 3 matrices d'usages avec un nombre identique de logs construits aléatoirement. Concrètement, nous avons distribué aléatoirement l'emplacement ainsi que la valeur des notes dans la matrice d'usages. Les RMSE obtenues tournent toutes autour de 1,023. Ce résultat signifie que le fait qu'un utilisateur ait vu un film contient déjà un certain degré d'information, que le film ait été apprécié ou non. L'explication pourrait être que les films notés par un utilisateur, même négativement, sont généralement des films qu'il a souhaité voir et qui font donc partie d'un genre ou type qu'il apprécie. Ceci est déjà une information en soit qui semble utile au moteur de recommandation. Le tableau 7.21 récapitule ces quatre résultats.

Filtrage collaboratif			
Vraies notes	Notes prédites	Notes aléatoires	Logs aléatoires
0.897	**0.898**	0.989	1,023

TABLE 7.21 – Résultats obtenus avec les données communautaires et les données aléatoires

Nous avons souhaité vérifier l'impact de la performance du classifieur sur la qualité des recommandations. Pour cela, nous avons réalisé la classification d'opinion avec un très mauvais classifieur. Nous avons effectué la classification sur les textes représentés uniquement à l'aide de la variable la plus informative trouvée par le classifieur SNB : le point d'exclamation. Le résultat obtenu en classification est très faible ($F_{score} \simeq 0,5$). La RMSE est, elle, étonnamment élevée avec un score de 0,903.

Nous avons également souhaité comparer le résultat obtenu avec la classification d'opinion binaire avec les résultats des classifications sur trois et cinq classes. Ceci afin de déterminer si une classification plus fine pouvait avoir une influence sur la qualité des recommandations. La RMSE obtenue à partir des résultats de la classification sur trois classes est de 0,907. Elle est de 0,913 dans le cas de la classification sur cinq classes. Il semble donc que le gain en précision obtenu avec les classifications multi-classes ne compense pas l'augmentation naturelle du nombre d'erreurs. Le tableau 7.22 récapitule ces derniers résultats.

Pour finir, nous avons souhaité évaluer la meilleure méthode de classification effectuée à partir d'un corpus d'apprentissage non-équilibré (tel qu'on l'a extrait du site *Flixster*). En effet, nous avons vu dans l'état de l'art (section 6.3) que les données d'apprentissage doivent être aussi représentatives que possible du corpus de test, ce qui n'était pas le cas dans les expérimentations précédentes. En effet, pour ne pas privilégier l'un ou l'autre des classifieurs, nous avions préféré travailler sur des corpus équilibrés. La RMSE obtenue

Filtrage collaboratif		
Avec la classification d'opinion sur 2 classes	Avec la classification d'opinion sur 3 classes	Avec la classification d'opinion sur 5 classes
0.898	0.907	0.913

TABLE 7.22 – Résultats obtenus avec les données communautaires et les différentes classifications d'opinion

avec un apprentissage sur un corpus déséquilibré est alors de 0,889, soit meilleure que le résultat précédent, et même meilleure que le résultat obtenu avec les notes réelles données par les utilisateurs (voir tableau 7.23).

Filtrage collaboratif		
Corpus d'apprentissage équilibré	Corpus d'apprentissage déséquilibré	Vraies notes
0.898	**0.889**	0,897

TABLE 7.23 – Résultats obtenus avec les données communautaires et les différentes corpus d'apprentissage équilibrés et déséquilibrés

7.4.2 Discussion

Les expérimentations précédentes ont permis de montrer que la classification d'opinion n'entraîne aucune perte d'information en comparaison à l'emploi des vraies notes données par les utilisateurs. La dernière expérimentation consistant à appliquer un classifieur entraîné sur un corpus de même nature que le corpus de test a même permis d'obtenir une meilleure RMSE que celle obtenue avec les vraies notes (0,889 contre 0,897). Ces résultats sont néanmoins inférieurs à ceux obtenus lors de l'étalonnage du moteur effectué à partir des logs *Netflix* (0,862 de RMSE, voir section 3.2). La raison provient certainement de la différence de richesse des deux corpus. En effet, les profils utilisateurs sont beaucoup plus riches dans le corpus *Netflix* que dans le corpus *Flixster*, chaque utilisateur de *Netflix* ayant noté 200 films en moyenne alors que ce chiffre n'est que de 30 pour les utilisateurs de *Flixster*. Toutefois, il est intéressant d'observer la courbe d'échantillonnage du nombre de commentaires et de la RMSE correspondante (voir figure 7.6). On peut observer que la RMSE s'améliore avec l'augmentation du nombre de commentaires et qu'il semblerait également qu'elle pourrait être améliorée avec un plus grand nombre de commentaires. Nous n'avons néanmoins pas eu la possibilité de vérifier cette théorie. Pour finir, nous avons également pu observer que la performance du classifeur d'opinion n'avait qu'une très faible influence sur la qualité des recommandations émises en sortie de chaîne. En effet, la différence entre les RMSE obtenues avec le meilleur classifieur et avec un très faible n'est que de 0,014.

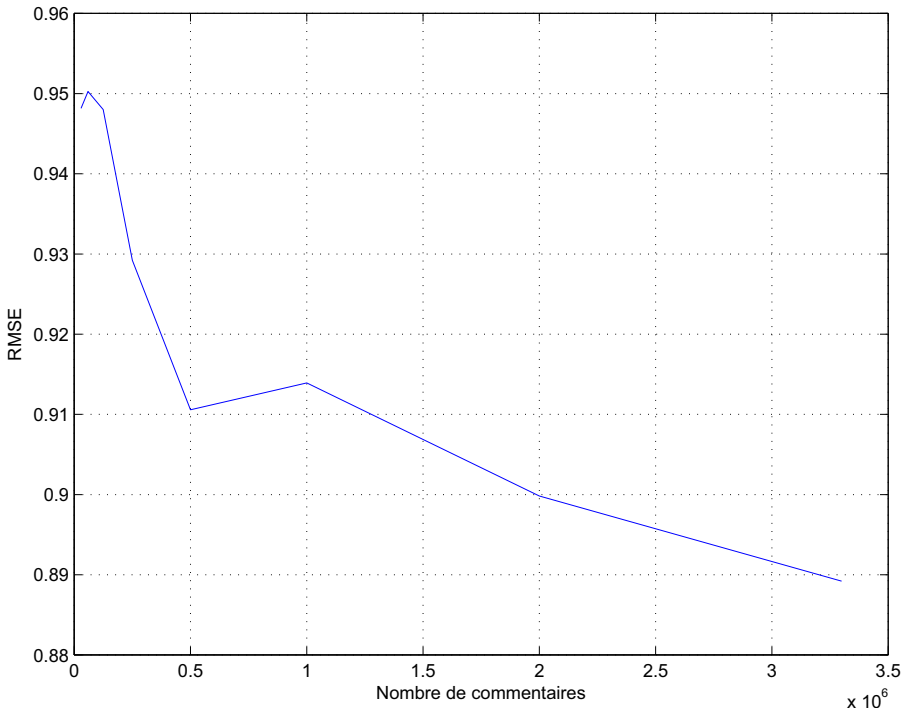


FIGURE 7.6 – RMSE suivant le nombre de commentaires classés

7.5 Conclusion

Dans ce chapitre, nous avons comparé différentes stratégies permettant d'associer une note à un texte. Les deux grands types d'approches de la classification d'opinion, à savoir les approches basées sur les lexiques et les approches basées sur l'apprentissage automatique, ont été testées. Comme l'ont déjà montré Pang et al. [PLV02], les résultats obtenus à l'aide des méthodes basées sur l'apprentissage supervisé sont supérieurs à ceux obtenus à partir des lexiques. De plus, ces méthodes permettent de classer la totalité des documents, à l'inverse des approches basées sur les lexiques qui ont une action restreinte. Pour la méthode basée sur l'apprentissage supervisé, nous avons évalué plusieurs classifieurs : un classifieur Naïf Bayésien, un classifieur Naïf Bayésien avec sélection de variables et une Machine à Vecteurs de Support. Nous avons également évalué différents pré-traitements et représentations. Les meilleures performances de classification ont été obtenues avec un SVM et ce, quelle que soit la représentation des textes choisie. On retrouve ici les résultats de Joachims [Joa99a] et de nombreux autres auteurs en classification d'opinion, les SVM étant très bien adaptés aux données décrites en grande dimension et aux données creuses. En outre, les différentes expérimentations menées afin de vérifier l'apport de pré-traitements, d'autres types de segmentations et de représentations plus complexes n'ont pas été convaincants. Il s'avère en effet que sur les données utilisées, les meilleurs résultats de classification ont été obtenus avec les données les moins pré-traitées (minusculisation, suppression des mots peu fréquents) et que l'utilisation d'un anti-dictionnaire ou de pré-

traitements linguistiques (correction orthographique et lemmatisation) n'ont pas permis d'améliorer les résultats. La complexification de la représentation n'a également pas apporté d'amélioration, la représentation fréquentielle ayant obtenu des meilleurs scores que les représentations fréquentielles normalisées et TF-IDF.

Le classifieur Naïf Bayesien avec sélection de variables est légèrement moins performant que le SVM en classification mais il nous a apporté des informations intéressantes sur le caractère informatif du vocabulaire. Il a ainsi notamment fait émerger des mots importants pour la classification d'opinion que l'on élimine avec les méthodes de pré-traitements classiques comme l'utilisation d'anti-dictionnaires, la suppression de la ponctuation ou l'application d'un lemmatiseur. Nous avons notamment pu observer l'importance de la conjugaison des verbes, l'emploi du passé ayant une connotation plutôt négative alors que l'emploi du présent de l'indicatif est plus souvent utilisé pour décrire des sentiments positifs. Nous avons également pu noter que les déterminants et les conjonctions de coordination ont une importance, ceci étant probablement dû au fait que les textes positifs sont en moyenne plus longs que les textes négatifs. Enfin, nous avons observé que les négations ont un rôle très proche des mots exprimant un avis négatif, celles-ci étant très peu utilisées dans les commentaires positifs.

Concernant les recommandations émises en sortie de chaîne, nous pouvons constater qu'elles sont meilleures que celles obtenues à l'aide de l'approche thématique (Chapitre 5) avec une RMSE de 0,889 contre 0,926. Il paraît donc intéressant, lors du démarrage d'un service de recommandation, de privilégier la construction de données d'usages plutôt que des données thématiques. Une autre information intéressante est que, quelle que soit la qualité du classifieur d'opinion, la qualité des recommandations émises est toujours meilleure qu'avec l'approche thématique. Ceci provient probablement du fait que les liens entre utilisateurs et items, quelle que soit la note les reliant, possède déjà un fort degré d'information. En effet, un item mal noté par un utilisateur implique tout de même que l'utilisateur s'est intéressé à l'item à un moment donné et qu'il est fort probable qu'il apprécie des items proches. Pour finir, nous avons montré que ces résultats obtenus avec des données extérieures sont moins bons qu'avec les notes initiales de *Netflix* ($RMSE = 0,862$) mais meilleurs qu'avec les informations provenant d'IMDB ($RMSE = 0,921$), ce qui montre qu'il peut être plus pertinent d'alimenter un moteur en données d'usages extraites de sites communautaires que de faire de la recommandation thématique.

Chapitre 8

Conclusion générale et perspectives

Sommaire

8.1 Bilan

Dans cette thèse, nous nous sommes intéressés à l'exploitation de textes produits par des internautes dans le contexte de la recommandation de contenu. Les moteurs de recommandation sont des systèmes permettant de personnaliser une interface, généralement une page Web, en filtrant les informations selon les besoins ou désirs des utilisateurs. Ces systèmes nécessitent de grandes quantités de données d'apprentissage afin de catégoriser les produits (ou items) à recommander. À côté de cela, le Web 2.0 est de plus en plus enclin à contenir des informations, avis, opinions exprimés par des consommateurs sur des produits de toutes sortes. L'objectif principal de ce manuscrit était de montrer que les commentaires d'utilisateurs présents sur le Web, et plus spécifiquement sur les sites communautaires, pouvaient permettre d'acquérir des informations susceptibles de pallier le manque de données d'apprentissage des systèmes de recommandation.

L'état de l'art sur la recommandation (Chapitre 2) a présenté les deux types d'approches dominantes en recommandation personnalisée : le filtrage thématique, basé sur des données descriptives, et le filtrage collaboratif, basé sur des données d'usages. Nous avons fait le choix d'explorer ces deux approches en extrayant tout d'abord des descripteurs des textes (Chapitre 5) puis en inférant des notes à l'aide d'une tâche de classification d'opinion (Chapitre 7). Afin de comparer tous les résultats, nous avons utilisé un moteur de recommandation basé sur la construction de matrices Items-Items capable de fonctionner soit en mode thématique, soit en mode collaboratif suivant les données d'entrée (Chapitre 3). Les expérimentations ont été faites à partir de commentaires issus du site communautaire *Flixster* (Chapitre 4). Ces commentaires possèdent des caractéristiques très proches du « discours électronique médié » : des abréviations, des fautes d'orthographes, des onomatopées, des erreurs grammaticales, etc.

Concernant l'approche thématique, nous avons représenté chaque film par tous les commentaires dont il est le sujet et nous avons effectué des sélections simples de variables parmi les mots composant ces commentaires. Nous avons ensuite mesuré des similarités entre chaque représentation de film deux à deux et effectué les recommandations en fonction de ces similarités. Cette méthode a permis d'obtenir des résultats proches de ceux que l'on peut obtenir à partir de données structurées très riches sélectionnées manuellement (les données références sont des descripteurs extraits de l'*Internet Movie DataBase* (*IMDB*) qui est une source connue comme étant la plus riche pour le domaine des films). Malgré tout, les différentes sélections de variables n'ont entrainé que de faibles variations des performances ce qui nous a poussé à expérimenter une nouvelle approche.

Les textes communautaires véhiculent l'opinion de leurs auteurs, nous avons donc cherché à transformer ces textes non structurés en données d'usages pour contribuer à l'alimentation du moteur de recommandation. Nous avons proposé une chaîne de traitements qui associe une étape de classification d'opinion, qui permet d'inférer des notes aux commentaires, et une étape de recommandation par filtrage collaboratif afin d'exploiter ces notes. Pour la partie classification d'opinion, nous avons comparé différentes

approches appartenant aux deux catégories dominantes du domaine : les approches basées sur les lexiques et les approches basées sur l'apprentissage automatique. Nous avons observé que sur ce type de textes, les approches basées sur l'apprentissage automatique sont à privilégier lorsque l'on possède des exemples d'apprentissage. Concernant la partie recommandation, nous avons montré que l'apport de données extérieures pour le calcul des matrices de similarités entre articles dans le moteur de recommandation est tout à fait pertinent. Nous avons confronté nos résultats à ceux obtenus sur le corpus de référence *Netflix* ainsi qu'à une méthode de filtrage thématique basée sur des descripteurs très riches provenant d'*IMDB*. Nos résultats avec des données extérieures sont moins bons qu'avec les notes initiales de *Netflix* mais bien meilleurs qu'avec les informations provenant d'*IMDB*, ce qui montre que l'alimentation d'un moteur de recommandation en données d'usages provenant de sites communautaires peut être plus pertinent que de faire de la recommandation thématique, d'autant plus que des données aussi riches que celles d'*IMDB* peuvent être difficiles à obtenir pour des domaines autres que celui des films.

Les travaux présents dans cette thèse ont fait l'objet de plusieurs publications. Un premier papier présente l'analyse exploratoire des commentaires de films extraits du site *Flixster* que nous avons présentée dans le chapitre 4 [PBB08]. Les deux articles qui ont suivi abordent la tâche de la classification d'opinion et comparent les approches basées sur les lexiques et celles basées sur l'apprentissage automatique sur nos données particulières [PBGDNB08, PFB+09]. Les dernières publications présentent la chaîne de traitements complète permettant de faire de la recommandation automatique à partir de données textuelles non structurées à l'aide de la classification d'opinion, de nouvelles expérimentations et résultats venant étayer chacune d'entre elles [Poi10, PTFS10, PFT10]. Enfin, une dernière publication plus complète synthétise l'ensemble des travaux faits en classification d'opinion par l'approche basée sur l'apprentissage automatique et présente les résultats obtenus en recommandation [PFT11].

Nous n'avons toutefois pas répondu à toutes les interrogations posées par le sujet. Nous aurions souhaité mieux évaluer les quantités respectives de notes initiales et de données externes nécessaires pour établir de bonnes recommandations. En effet, nous avons montré que les données extraites du Web contenaient de l'information utile aux moteurs de recommandation mais nous n'avons pas vérifié si cette information était compatible avec l'information déjà présente dans le moteur. Dans toutes nos expérimentations, nous avons considéré que le moteur ne possédait aucune information propre au système en apprentissage et nous n'avons évalué que l'apport d'informations extérieures. Dans le cas où c'est un système de classification automatique qui est utilisé sur des textes, nous souhaiterions également pouvoir mesurer plus précisément la corrélation entre la qualité d'une classification d'opinion et celle de la recommandation à laquelle elle contribue. Nous avons montré qu'un mauvais classifieur permet des recommandations quasiment d'aussi bonne qualité qu'un bon classifieur mais nous ne sommes pas encore en mesure de savoir pourquoi.

D'un point de vue personnel, cette thèse a permis l'acquisition de compétences variées du fait qu'elle recouvre plusieurs domaines de recherche tels que la fouille de textes, la

recherche d'information, la fouille d'opinion, le traitement automatique de la langue, l'apprentissage automatique ainsi que la recommandation automatique de contenus. Nous avons en effet eu l'occasion d'étudier l'existant de chacun de ces domaines afin de sélectionner les outils et techniques adaptées aux différentes problématiques et avoir ainsi une connaissance relativement approfondie de chaque domaine étudié.

8.2 Perspectives

Du côté des perspectives, nous souhaiterions tenter d'extraire des commentaires des informations autres que les opinions. Nous pensons en effet que le style d'écriture des utilisateurs peut avoir une influence sur leurs préférences. Des informations statistiques représentatives de la prolixité, de la variabilité du vocabulaire ou encore de la taille moyenne des mots pourraient permettre de « catégoriser » les films et les utilisateurs. Par exemple, pour un nombre identique de commentaires présents dans notre corpus, les utilisateurs rédigent cinq mots en moyenne pour le film *Le petit dinosaure et la vallée des merveilles*[1] alors que pour le film *2001, l'Odyssée de l'espace*[2], le nombre moyen de mots est de 90. Ces deux films ne sont pas destinés au même public et les styles d'écriture paraissent très différents. C'est un cas extrême bien entendu, mais nous pensons que l'analyse statistique des styles d'écriture est une voie à étudier, ne serait-ce que pour tenter de mesurer la portée de ce type d'informations.

Un point que nous aimerions également développer dans le futur est d'étudier comment enrichir un moteur de recommandation en prenant en compte de nouveaux types d'informations. La connectivité du réseau des « amis » déclarés dans les sites communautaires pourrait ainsi constituer une source de données nouvelles et précieuses à intégrer dans les modèles. Les réseaux sociaux peuvent être construits selon deux types de relations :

- Les relations explicites : déclarations « ami-ami » sur les réseaux sociaux ;
- Les relations implicites : discussions sur des forums, commentaires sur la page personnel d'un autre utilisateur, etc.

Nous pensons que ces deux types de relations contiennent de l'information qui pourrait être utile à la recommandation automatique. En effet, deux personnes qui se « rencontrent » sur le Web peuvent avoir des goûts proches dans le sens où ils se sont forcément croisés via une page commune. Sur *Flixster* ou *Allociné* par exemple, les gens peuvent se croiser sur les forums ou les boîtes à réactions qui concernent les films qu'ils connaissent, et généralement apprécient. Considérer les degrés de séparation des utilisateurs dans un réseau social pourrait peut-être alors permettre d'affiner des recommandations.

Une autre perspective pourrait également être d'intégrer les opinions des utilisateurs directement dans le réseau social. On entend par là que les nœuds des graphes pourraient

1. Dessin animé long métrage produit par Steven Speilberg et Georges Lucas en 1988.
2. Film de Stanley Kubrick sorti en 1968

être étiquetés à l'aide de couples item-opinion. Des algorithmes de la théorie des graphes pourraient alors permettre de retrouver des communautés d'utilisateurs aux goûts similaires et les recommandations pourraient être établies à l'aide de techniques de propagation d'étiquettes. Cette approche avait été envisagée en début de thèse mais des problématiques liées aux outils, aux données de validation ainsi qu'au temps imparti n'ont pas permis de l'exploiter.

Annexes

C Captures d'écran du site *Flixster*

Les figures C.a et C.b présentent respectivement la page d'accueil du site Flixster et un exemple type de page utilisateur contenant, entre autres choses, les commentaires portant sur les films. Ce sont de ces pages que les commentaires composant le corpus de textes ont été extraits (Chapitre 4).

FIGURE C.a – Page d'accueil du site Flixster en 2008

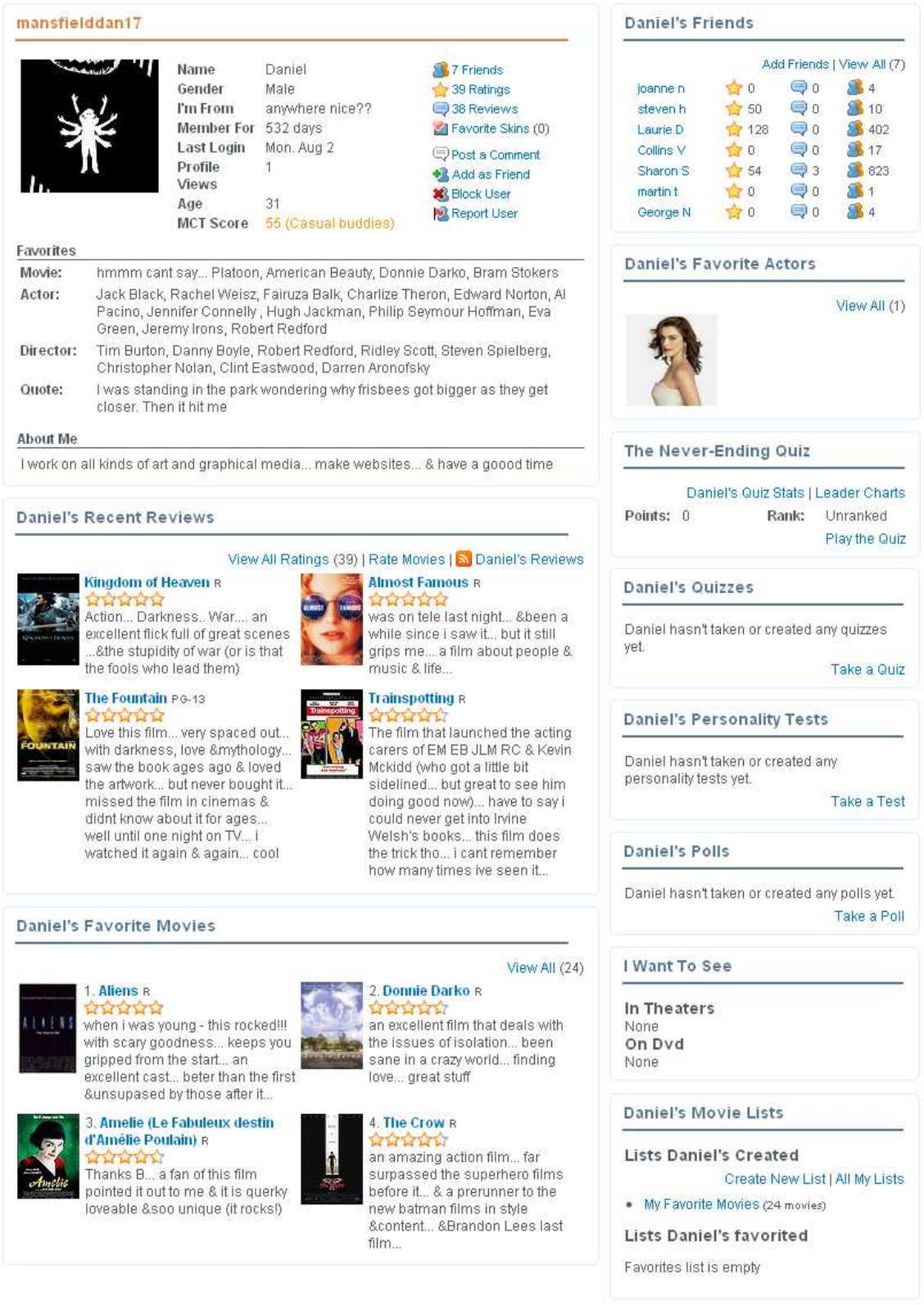
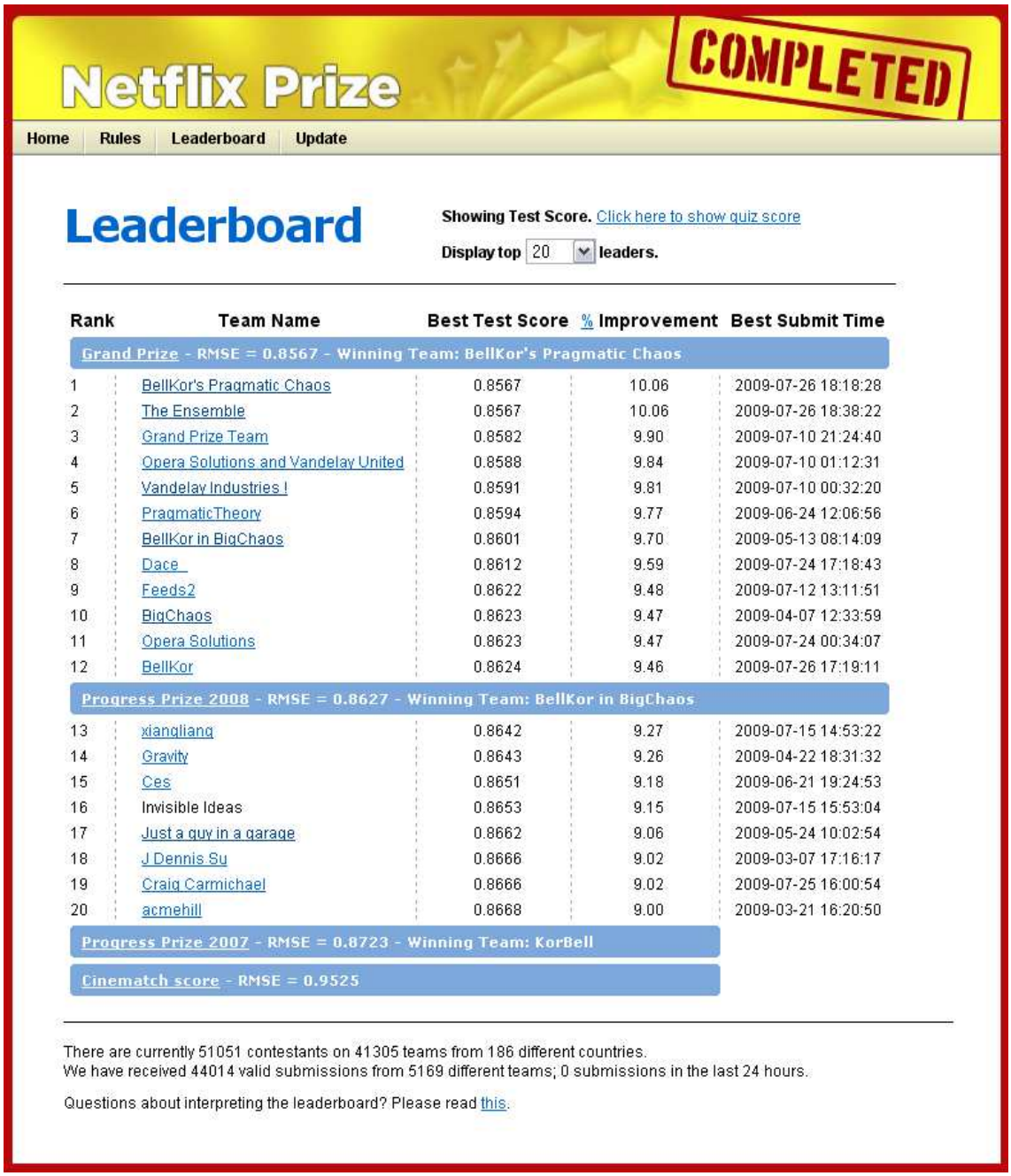


FIGURE C.b – Exemple de page utilisateur sur le site Flixster

D Leaderboard du challenge *Netflix*

La figure D.c présente le tableau des résultats obtenus par les participants au challenge Netflix (Netflix Prize).



Netflix Prize

COMPLETED

Home | Rules | Leaderboard | Update

Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11
Progress Prize 2008 - RMSE = 0.8627 - Winning Team: BellKor in BigChaos				
13	xiangliang	0.8642	9.27	2009-07-15 14:53:22
14	Gravity	0.8643	9.26	2009-04-22 18:31:32
15	Ces	0.8651	9.18	2009-06-21 19:24:53
16	Invisible Ideas	0.8653	9.15	2009-07-15 15:53:04
17	Just a guy in a garage	0.8662	9.06	2009-05-24 10:02:54
18	J Dennis Su	0.8666	9.02	2009-03-07 17:16:17
19	Craig Carmichael	0.8666	9.02	2009-07-25 16:00:54
20	acmehill	0.8668	9.00	2009-03-21 16:20:50
Progress Prize 2007 - RMSE = 0.8723 - Winning Team: KorBell				
Cinematch score - RMSE = 0.9525				

There are currently 51051 contestants on 41305 teams from 186 different countries.
We have received 44014 valid submissions from 5169 different teams; 0 submissions in the last 24 hours.

Questions about interpreting the leaderboard? Please read this.

FIGURE D.c – Page de résultats du challenge Netflix après qu'il eut été remporté

Ces résultats font office de références afin de placer les résultats de notre moteur de recommandation dans l'état de l'art (Chapitre 3).

E Lexique d'opinion *Maison*

Les tableaux 8.1 et 8.2 répertorient tous les mots positifs et négatifs répertoriés dans le lexique *Maison* utilisé pour la classification d'opinion (Chapitre 7).

good	interesting	magic	gr8	wondeful
great	superb	smart	satisfactory	inspiring
funny	sexy	pure	acceptable	magical
awesome	outstanding	pearly	suitable	favs
awsome	happy	legendary	superb	legend
cool	enjoyable	comical	love	fab
brilliant	emotional	adorable	like	loll
hilarious	creative	unbelievable	amaze	omg
favorite	talent	touching	enjoy	fun
favourite	gorgeous	spectacular	recommend	hilarity
well	stylish	telling	interest	original
hot	pretty	phenomenal	underrate	class
excellent	honest	immense	touch	thank
beautiful	famous	unforgettable	appreciate	satisfy
fantastic	epic	interested	inspire	cheesy
cute	entire	humorous	amazing	luv
sweet	super	fascinating	fav	fan
gd	memorable	dreamy	rock	excite
nice	lovely	visionary	wkd	exciting
classic	inspirational	universal	perfect	genius
powerful	huge	best	nicely	ok
intense	glad	top	funniest	okay
comic	extra	incredible	laugh	

TABLE 8.1 – Mots *positifs* répertoriés dans le lexique d'opinion *Maison*

bad	dreadful	shitty	unsightly	endure
wicked	painful	stinking	shit	depreciate
stupid	unspeakable	stinky	monstrous	worst
fucking	deplorable	incompetent	unnatural	waste
fake	lamentable	unskilled	hate	disappoint
wrong	pitiful	mediocre	disturb	crap
poor	fearful	mischievous	fuck	despise
ugly	frightful	naughty	overeat	bore
silly	horrid	negative	dislike	h8
suck	icky	rubber	annoy	creepy
atrocious	crappy	unsatisfactory	detest	hated
abominable	lousy	unsuitable	distressing	boring
awful	rotten	unlovely	suffer	kk
weak	freak	lame	horrible	

TABLE 8.2 – Mots *négatifs* répertoriés dans le lexique d'opinion *Maison*

F Liste des variables informatives

Le tableau 8.3 répertorie les 633 variables informatives trouvées par le classifieur naïf Bayésien avec sélection de variables KHIOPS. Elles ont été sélectionnées lors de la tâche de classification d'opinion binaire, portant sur un corpus ayant subi des pré-traitements minimaux (minusculisation et suppression des mots peu fréquents) et une représentation fréquentielle. Elles sont classées selon leur niveau d'information, le rang 1 désignant la variable la plus informative.

Rang	Mot	Rang	Mot	Rang	Mot
1	!	2	love	3	loved
4	not	5	ok	6	boring
7	awesome	8	great	9	amazing
10	bad	11	best	12	but
13	movie	14	stupid	15	t
16	worst	17	.	18	crap
19	brilliant	20	hilarious	21	sucked
22	excellent	23	was	24	alright
25	waste	26	awsome	27	didn
28	terrible	29	’	30	wasn
31	lame	32	horrible	33	okay
34	too	35	luv	36	nothing
37	fantastic	38	is	39	this
40	no	41	must	42	as

Rang	Mot	Rang	Mot	Rang	Mot
43	omg	44	dumb	45	film
46	awful	47	hate	48	kinda
49	been	50	?	51	disappointing
52	hated	53	better	54	predictable
55	like	56	pretty	57	asleep
58	half	59	sucks	60	wonderful
61	very	62	don	63	weird
64	wasnt	65	much	66	guess
67	ever	68	cried	69	wicked
70	favorite	71	decent	72	gay
73	be	74	see	75	didnt
76	:	77	kind	78	wow
79	dont	80	perfect	81	beautiful
82	absolutely	83	disappointed	84	meh
85	money	86	sooo	87	worse
88	some	89	so	90	only
91	funny	92	plot	93	funniest
94	wait	95	gr8	96	could
97	pointless	98	lt	99	bored
100	slow	101	cry	102	expected
103	just	104	poor	105	that
106	incredible	107	average	108	for
109	fell	110	rocks	111	enough
112	disappointment	113	parts	114	luved
115	original	116	would	117	moments
118	brill	119	amazin	120	favourite
121	or	122	exelente	123	annoying
124	rubbish	125	sad	126	bit
127	touching	128	minutes	129	and
130	shit	131	excelente	132	at
133	to	134	hot	135	eh
136	there	137	such	138	really
139	least	140	had	141	dull
142	sweet	143	soo	144	lol
145	nephews	146	have	147	were
148	fav	149	either	150	what
151	,	152	dis	153	retarded
154	why	155	seemed	156	=
157	my	158	poorly	159	i
160	ugh	161	idea	162	ridiculous
163	wasted	164	borin	165	heath
166	one	167	maybe	168	x
169	greatest	170	soooo	171	ledger

Rang	Mot	Rang	Mot	Rang	Mot
172	cheesy	173	point	174	hour
175	weak	176	channing	177	heart
178	mediocre	179	crappy	180	thought
181	classic	182	dissapointing	183	thing
184	scary	185	ace	186	seat
187	tried	188	over	189	johnny
190	let	191	alrite	192	k
193	old	194	genius	195	interesting
196	flat	197	blah	198	pathetic
199	^	200	true	201	romantic
202	fave	203	more	204	u
205	brillant	206	ruined	207	though
208	fab	209	care	210	wtf
211	edge	212	quite	213	save
214	few	215	)	216	cool
217	isn	218	mint	219	yawn
220	however	221	watchable	222	trying
223	instead	224	dissapointed	225	little
226	&	227	sorry	228	potential
229	hillarious	230	they	231	intense
232	down	233	never	234	first
235	off	236	hmmm	237	long
238	awsum	239	dragged	240	silly
241	couldn	242	packed	243	outstanding
244	felt	245	laughing	246	super
247	superb	248	rocked	249	nowhere
250	saw	251	overrated	252	seems
253	tatum	254	rather	255	jokes
256	enjoyed	257	supposed	258	avoid
259	garbage	260	else	261	anything
262	suck	263	failed	264	powerful
265	star	266	shite	267	disapointed
268	into	269	fuckin	270	action
271	cheap	272	kids	273	]
274	make	275	haha	276	cutest
277	expecting	278	interest	279	every
280	stupidest	281	yay	282	found
283	adorable	284	movies	285	hott
286	reason	287	fake	288	total
289	attempt	290	wanted	291	looks
292	less	293	trailer	294	sooooo
295	well	296	tries	297	suppose
298	nt	299	the	300	any

Rang	Mot	Rang	Mot	Rang	Mot
301	unless	302	bother	303	unfunny
304	forgettable	305	year	306	sadly
307	typical	308	job	309	depp
310	joker	311	2006	312	about
313	brilliantly	314	fairly	315	wierd
316	wkd	317	budget	318	trilogy
319	previews	320	lacking	321	recommend
322	same	323	sexy	324	gross
325	will	326	disgusting	327	freakin
328	lovely	329	doesn	330	cute
331	/	332	gotta	333	nah
334	fabulous	335	unfortunately	336	-
337	get	338	fucking	339	overall
340	eva	341	masterpiece	342	remember
343	disapointing	344	v	345	something
346	script	347	type	348	mildly
349	s	350	harry	351	p
352	daniel	353	tv	354	premise
355	truly	356	being	357	nope
358	lacked	359	d	360	near
361	did	362	hoped	363	lacks
364	horribly	365	utter	366	strange
367	hype	368	both	369	sense
370	ehh	371	ew	372	idk
373	amzain	374	if	375	rest
376	awww	377	even	378	sleep
379	everyone	380	overly	381	uninteresting
382	um	383	do	384	zac
385	seem	386	freaking	387	zero
388	rox	389	shia	390	orlando
391	;	392	terrific	393	gorgeous
394	entertaining	395	stopped	396	corny
397	holes	398	mostly	399	rated
400	aight	401	dumbest	402	compared
403	favourites	404	because	405	wonderfully
406	left	407	unoriginal	408	10
409	stick	410	exallent	411	holy
412	redeeming	413	started	414	sparrow
415	version	416	muy	417	hoping
418	barely	419	ummm	420	hmm
421	33	422	load	423	other
424	effort	425	rules	426	absolutly
427	bothered	428	30	429	shitty

Rang	Mot	Rang	Mot	Rang	Mot
430	own	431	ultimately	432	spent
433	wathching	434	rip	435	dvd
436	totally	437	you	438	2007
439	on	440	somewhat	441	give
442	kicks	443	dance	444	hours
445	favorites	446	unrealistic	447	funnier
448	awesum	449	laughable	450	low
451	films	452	inspiring	453	scenes
454	yuck	455	painful	456	favs
457	odd	458	a	459	lost
460	turned	461	then	462	coolest
463	music	464	where	465	keeps
466	needed	467	buy	468	downhill
469	potter	470	umm	471	lack
472	amusing	473	wouldn	474	burton
475	stunning	476	lov	477	luvd
478	bland	479	badly	480	captain
481	wiked	482	excuse	483	purpose
484	kick	485	fukin	486	life
487	mess	488	looove	489	dissapointment
490	whatsoever	491	trash	492	bleh
493	bravo	494	definately	495	”
496	damn	497	breathtaking	498	amazingly
499	wanna	500	finally	501	remake
502	fails	503	go	504	mins
505	emotional	506	going	507	lv
508	buena	509	joke	510	rock
511	disaster	512	gd	513	generic
514	hum	515	weren	516	sort
517	songs	518	unnecessary	519	good
520	video	521	craig	522	watching
523	stand	524	funni	525	terribly
526	horror	527	grade	528	laughed
529	childish	530	marvelous	531	oscar
532	allright	533	bestest	534	making
535	hmmmm	536	stupidity	537	pile
538	makes	539	n	540	two
541	bunch	542	efron	543	boaring
544	substance	545	which	546	cruise
547	ellen	548	big	549	skip
550	complete	551	moive	552	soooooo
553	retarted	554	highly	555	mean
556	taste	557	m	558	recomiendo

Rang	Mot	Rang	Mot	Rang	Mot
559	forced	560	bullshit	561	should
562	place	563	basically	564	out
565	tears	566	recommended	567	sux
568	tedious	569	cryed	570	me
571	understand	572	ages	573	plain
574	concept	575	special	576	dnt
577	ehhh	578	socks	579	adore
580	absolutley	581	tea	582	walked
583	bale	584	looked	585	dicaprio
586	batman	587	performance	588	paid
589	it	590	somehow	591	alryt
592	know	593	dreadful	594	gayest
595	now	596	drawn	597	bloom
598	painfully	599	spoof	600	cast
601	please	602	neither	603	itself
604	isnt	605	marvellous	606	slasher
607	bits	608	20	609	lousy
610	thumbs	611	next	612	beautifully
613	insult	614	boo	615	flawless
616	uh	617	guillermo	618	honest
619	tom	620	eastwood	621	words
622	aliens	623	hehe	624	disgrace
625	act	626	having	627	way
628	falls	629	cage	630	soundtrack
631	comment	632	rental	633	emotions

TABLE 8.3 – Variables informatives pour la classification d'opinion selon KHIOPS

Liste des figures

Références bibliographiques

[AB06] A. Andreevskaia and S. Bergler. Mining wordnet for fuzzy sentiment : Sentiment tag extraction from wordnet glosses, 2006.

[AdFF04] J. Anis, M. de Fornel, and B. Fraenkel. La communication électronique : Approches linguistiques et anthropologiques. *Colloque international, EHESS*, 2004.

[And06] C. Anderson. *The Long Tail*. Hyperion press edition, 2006.

[AT05] 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 :734–749, 2005.

[BB08] C. Bothorel and M. Bouklit. Détection de structures de communauté dans les hyper-réseaux d'interactions. In David Simplot-Ryl and Sebastien Tixeuil, editors, *10ème Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications (AlgoTel'08)*, pages 57–60, Saint-Malo France, 2008.

[BCP+07] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, and V. Subrahmanian. Sentiment analysis : Adjectives and adverbs are better than adjectives alone. In *International Conference on Weblogs and Social Media (ICWSM), Boulder, Colorado, U.S.A, 26/03/2007-28/03/2007*, pages 203–206, http ://www.aaai.org/Press/press.php, 2007. AAAI Press.

[BHMK98] J. S. Breese, D. Heckerman, and C. Myers Kadie. Empirical analysis of predictive algorithms for collaborative filtering, 1998.

[BL07] J. Bennett and S. Lanning. The netflix prize. San Jose, California, USA, 2007. ACM.

[Bou05] M. Boullé. A Bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research*, 6 :1431–1452, 2005.

[Bou06] M. Boullé. MODL : a Bayes optimal discretization method for continuous attributes. *Machine Learning*, 65(1) :131–165, 2006.

[Bou07a] M. Boullé. Une méthode optimale d'évaluation bivariée pour la classification supervisée. In *Extraction et gestion des connaissances (EGC'2007)*, pages 461–472, 2007.

[Bou07b] M. Boullé. Compression-based averaging of selective naive bayes classifiers. *J. Mach. Learn. Res.*, 8 :1659–1685, 2007.

[Bur02] R. Burke. Hybrid recommender systems : Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4) :331–370, 2002.

[Bur07] R. Burke. Hybrid web recommender systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, chapter 12, pages 377–408. Springer, 2007.

[Car79] J. Carbonell. *Subjective Understanding : Computer Models of Belief Systems*. PhD thesis, Yale, 1979.

[CFB06] C. Cardie, C. Farina, and T. Bruce. Using natural language processing to improve erulemaking : project highlight. In *dg.o '06 : Proceedings of the 2006 international conference on Digital government research*, pages 177–178, New York, NY, USA, 2006. ACM.

[CGV07] É. Crestan, S. Gigandet, and R. Vinot. Approches naïves à l'analyse d'opinion. In *Actes de l'atelier de clôture du 3ème DÉfi Fouille de Textes*, pages 45–56, Grenoble, France, 2007. AFIA.

[CH01] M. Connor and J. Herlocker. Clustering items for collaborative filtering, 2001.

[CH08] K. B. Cohen and L. Hunter. Getting started in text mining. *PLoS Comput Biol*, 4(1) :e20, 01 2008.

[CMB07] L. Candillier, F. Meyer, and M. Boullé. Comparing state-of-the-art collaborative filtering systems. In Petra Perner, editor, *MLDM*, volume 4571 of *Lecture Notes in Computer Science*, pages 548–562. Springer, 2007.

[Coh96] W. Cohen. Learning trees and rules with set-valued features. In *In Proceedings of the 13th National Conference on Artificial Intelligence*, pages 709–716. AAAI Press, 1996.

[Cry01] D. Crystal. *Language and the Internet*. KCambridge University Press, Cambridge, 2001.

[CSC06] Cane, Stephen, and Fu L. Chung. Integrating collaborative filtering and sentiment analysis : A rating inference approach. In *Proceedings of The ECAI 2006 Workshop on Recommender Systems*, pages 62–66, Riva del Garda, I., 2006.

[CV07] R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance, 2007.

[DC01] S. Das and M. Chen. Yahoo ! for amazon : Extracting market sentiment from stock message boards. *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, 2001.

[DK04] M. Deshpande and G. Karypis. Item-based top-n recommendation algorithms, 2004.

[DL07] X. Ding and B. Liu. The utility of linguistic rules in opinion mining. In *SIGIR '07 : Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 811–812, New York, NY, USA, 2007. ACM.

[DLP03] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery : opinion extraction and semantic classification of product reviews. In *WWW '03 : Proceedings of the 12th international conference on World Wide Web*, pages 519–528, New York, NY, USA, 2003. ACM.

[DM02] A. Dejond and J. Mercier. *La cyberlangue française*. Bruxelle : La Renaissance du livre, 2002.

[DMM03] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic coclustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 89–98, New York, NY, USA, 2003. ACM.

[DWW07] G. Dziczkowski and K. Wegrzyn-Wolska. Rrss - rating reviews support system purpose built for movies recommendation. In *AWIC*, pages 87–93, 2007.

[DWW08] G. Dziczkowski and K. Wegrzyn-Wolska. An autonomous system designed for automatic detection and rating of film reviews. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1 :847–850, 2008.

[Fay07] M. Fayada. *Le langage SMS des jeunes. Approche lexicale et morphosyntaxique*. PhD thesis, Mémoire de Master 1, Gestion des apprentissage et formation ouverte et à distance, Département Sciences du Langage, Université Paul-Valéry Montpellier 3, 2007.

[GAB+09] C. Grouin, B. Arnulphy, J.-B. Berthelin, S. El Ayari, A. García-Fernandez, A. Grappy, M. Hurault-Plantet, P. Paroubek, I. Robba, and P. Zweigenbaum. Présentation de l'édition 2009 du DÉfi Fouille de Textes (DEFT'09). In *Actes de l'atelier de clôture de la cinquième édition du DÉfi Fouille de Textes*, pages 35–50, Paris, 2009.

[GBEA+07] C. Grouin, J.-B. Berthelin, S. El Ayari, T. Heitz, M. Hurault-Plantet, M. Jardino, Z. Khalis, and M. Lastes. Présentation de deft'07. In *Actes de l'atelier de clôture du 3ème DÉfi Fouille de Textes*, pages 1–8, Grenoble, France, 2007. AFIA.

[GdNBC+02] É. Guimier de Neef, M. Boualem, C. Chardenon, P. Filoche, and J. Vinesse. Natural language processing software tools and linguistic data developped by france télécom rd. 2002.

[GDNV04] É. Guimier De Neef and J. Véronis. 1 pw1 sr la kestion ;-). *Communication, Journée d'Étude de L'ATALA : Le traitement automatique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.)*, 2004.

[GN06] G. Govaert and M. Nadif. Classification d'un tableau de contingence et modèle probabiliste. In *EGC*, pages 457–462, 2006.

[GNOT92] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12) :61–70, 1992.

[Gro06] L. Grossman. Time's person of the year : You. In *Time Magazine*, 2006.

[gro07] ComScore/The Kelsey group. Online consumer-generated reviews have significant impact on offline purchase behavior. www.oreilly.com, 2007.

[GS07] M. Généreux and M. Santini. Défi : Classification de textes français subjectifs. In *Actes de l'atelier de clôture du 3ème DÉfi Fouille de Textes*, pages 83–93, Grenoble, France, 2007. AFIA.

[Har72] J. A. Hartigan. Direct Clustering of a Data Matrix. *Journal of the American Statistical Association*, 67(337) :123–129, 1972.

[HL04a] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD '04 : Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA, 2004. ACM.

[HL04b] M. Hu and B. Liu. Mining opinion features in customer reviews. In Deborah L. Mcguinness, George Ferguson, Deborah L. Mcguinness, and George Ferguson, editors, *AAAI*, pages 755–760. AAAI Press / The MIT Press, 2004.

[HL05] J. Huang and C. X. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Trans. on Knowl. and Data Eng.*, 17(3) :299–310, 2005.

[HM97] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[Hor08] J. A. Horrigan. Online shopping. Pew Internet & American Lif Project Report, 2008.

[HSC+08] J. Heinecke, G. Smits, C. Chardenon, É. Guimier De Neef, E. Maillebuau, and M. Boualem. Tilt : plate-forme pour le traitement automatique des langues naturelles. *revue TAL*, 49 :17–41, 2008.

[Hum09] D. Humbert. Conception et développement de systèmes expérimentaux de recommandation automatique. In *Rapport de stage ENSAT*. 2009.

[HV09] W. Hersh and E. Voorhees. Trec genomics special issue overview. *Inf. Retr.*, 12(1) :1–15, 2009.

[HY01] D.J. Hand and K. Yu. Idiot bayes ? not so stupid after all ? *International Statistical Review*, 69(3) :385–399, 2001.

[Joa98] T. Joachims. Text categorization with support vector machines : Learning with many relevant features. In *Machine Learning : ECML-98*, pages 137–142. Springer, 1998.

[Joa99a] T. Joachims. Making large-scale support vector machine learning practical. *Advances in kernel methods : support vector learning*, pages 169–184, 1999.

[Joa99b] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML '99 : Proceedings of the Sixteenth International Conference on Machine Learning*, pages 200–209, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[Joa02] T. Joachims. *Learning to Classify Text Using Support Vector Machines : Methods, Theory and Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[Kar01] G Karypis. Evaluation of item-based top-n recommendation algorithms, 2001.

[KH07] G. King and D. Hopkins. Extracting systematic social science meaning from text. In *Working Paper 2008-0078, Weatherhead Center for International Affairs, Harvard University*, 2007.

[KKB10] C. Kaiser, J. Krockel, and F. Bodendorf. Swarm intelligence for analyzing opinions in online communities. *Hawaii International Conference on System Sciences*, 0 :1–9, 2010.

[KN06] H. Kanayama and T. Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. EMNLP, 2006.

[Kor10] Y. Koren. Factor in the neighbors : Scalable and accurate collaborative filtering. *ACM Trans. Knowl. Discov. Data*, 4(1) :1–24, 2010.

[KS75] E. F. Kelly and P. J. Stone. *Computer Recognition of English Word Senses.* North-Holland, Amsterdam, 1975.

[KSH06] N. Kwon, S. W. Shulman, and E. Hovy. Multidimensional text analysis for erulemaking. In *dg.o '06 : Proceedings of the 2006 international conference on Digital government research*, pages 157–166, New York, NY, USA, 2006. ACM.

[LAR02] W. Lin, S. A. Alvarez, and C. Ruiz. Efficient adaptive-support association rule mining for recommender systems, 2002.

[Leh09] F. Lehuédé. L'internet participatif redonne confiance aux consommateurs, 2009.

[LHC05] B. Liu, M. Hu, and J. Cheng. Opinion observer : analyzing and comparing opinions on the web. In *WWW '05 : Proceedings of the 14th international conference on World Wide Web*, pages 342–351, New York, NY, USA, 2005. ACM.

[LHTD02] H. Liu, F. Hussain, C.L. Tan, and M. Dash. Discretization : An enabling technique. *Data Mining and Knowledge Discovery*, 4(6) :393–423, 2002.

[Lin98] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[Liu09] B. Liu. *Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications).* Springer, 1st ed. 2007. corr. 2nd printing edition, January 2009.

[Liu10] B. Liu. Sentiment analysis and subjectivity. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition.* CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010. ISBN 978-1420085921.

[LSY03] G. Linden, B. Smith, and J. York. Amazon.com recommendations : Item-to-item collaborative filtering. *IEEE Internet Computing*, 7 :76–80, 2003.

[Mar00] M. Marcoccia. La représentation du non-verbal dans la communication écrite médiatisée par ordinateur. *Communication et Organisation*, 18 :265–274, 2000.

[MBF+90] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet : an on-line lexical database*. *Int J Lexicography*, 3(4) :235–244, January 1990.

[MCD07] S. Maurel, P. Curtoni, and L. Dini. Classification d'opinions par méthodes symbolique, statistique et hybride. In *Actes de l'atelier de clôture du 3ème DÉfi Fouille de Textes*, pages 109–116, Grenoble, France, 2007. AFIA.

[Mey11] F. Meyer. Apport des données thématiques dans les systèmes de recommandation : hybridation et démarrage à froid. In *Extraction et gestion des connaissances (EGC'2011)*, Brest, France, 2011.

[MM06] T. Mullen and R. Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[MOS07] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the trec-2007 blog track. In *Proceedings of The sixteenth Text REtrieval Conference (TREC 2007)*, pages 17–30. NIST, 2007.

[MYTF02] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the web. In *KDD '02 : Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 341–349, New York, NY, USA, 2002. ACM.

[NH06a] K. Nigam and M. Hurst. Towards a robust metric of polarity. In *Computing Attitude and Affect in Text : Theory and Applications*, Dordrecht, The Netherlands, 2006. Springer.

[NH06b] K. Nigam and M. Hurst. Towards a robust metric of polarity. In *Computing Attitude and Affect in Text : Theory and Applications*, pages 265–279. 2006.

[NRT08] K. Nageswara Rao and V. G. Talwar. Application domain and functional classification of recommender systems a survey. In *Desidoc journal of library and information technology*, volume 28, pages 17–36. ACM Press, 2008.

[OdRM+06] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the trec-2006 blog track. In *Proceedings of The Fifteenth Text REtrieval Conference (TREC 2006)*, pages 17–31. NIST, 2006.

[OMS08] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the trec-2008 blog track. In *Proceedings of The seventeenth Text REtrieval Conference (TREC 2008)*. NIST, 2008.

[O'R05] T. O'Reilly. What is web 2.0 ? design patterns and business models for the next generation of software. www.oreilly.com, September 2005.

[Pan06] R. Panckhurst. Le discours électronique médié : bilan et perspectives. *Lire, écrire, communiquer et apprendre avec Internet*, 2006.

[Pan09] R. Panckhurst. Short message service (sms) : typologie et problématiques futures. *in Arnavielle T. (coord.), Polyphonies, pour Michelle Lanvin*, 2009.

[Paz99] M. J. Pazzani. A framework for collaborative, content-based and demographic filtering. *Artif. Intell. Rev.*, 13(5-6) :393–408, 1999.

[PB07] M. Pazzani and D. Billsus. Content-based recommendation systems. pages 325–341. 2007.

[PBB08] D. Poirier, C. Bothorel, and M. Boullé. Analyse exploratoire d'opinions cinématographiques : co-clustering de corpus textuels communautaires. In *Extraction et gestion des connaissances (EGC'2008)*, pages 565–576, 2008.

[PBGDNB08] D. Poirier, C. Bothorel, É. Guimier De Neef, and M. Boullé. Automating opinion analysis in film reviews : the case of statistic versus linguistic approach. In *Proceedings of the LREC 2008 Workshop on Sentiment Analysis : Emotion, Metaphor, Ontology and Terminology*, pages 94–101, 2008.

[PFB+09] D. Poirier, F. Fessant, C. Bothorel, É. Guimier De Neef, and M. Boullé. Approches statistique et linguistique pour la classification de textes d'opinion portant sur les films. *Revue des Nouvelles Technologies de l'Information (RNTI) Fouille de Données d'opinion*, E17 :147–169, 2009.

[PFT10] D. Poirier, F. Fessant, and I. Tellier. Reducing the Cold-Start Problem in Content Recommendation Through Opinion Classification. In *Web Intelligence*, Toronto, Canada, 2010.

[PFT11] D. Poirier, F. Fessant, and I. Tellier. De la classification d'opinion à la recommandation : l'apport des textes communautaires. *Revue Traitement Automatique des Langues (TAL)*, 51, 2011.

[PL04] B. Pang and L. Lee. A sentimental education : sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL '04 : Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[PL08] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2) :1–135, 2008.

[PLV02] B Pang, L. Lee, and S. Vaithyanathan. Thumbs up ? : sentiment classification using machine learning techniques. In *EMNLP '02 : Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[Poi10] D. Poirier. La Classification d'Opinion comme préambule à la Recommandation Automatique de Contenus. In *COnférence en Recherche d'Information et Applications 2010, Proceedings CORIA-CIFED 2010*, 2010.

[PRDP08] M. Plantié, M. Roche, G. Dray, and P. Poncelet. Is a voting approach accurate for opinion mining ? In *DaWaK '08 : Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery*, pages 413–422, Berlin, Heidelberg, 2008. Springer-Verlag.

[PT08] Y. Park and A. Tuzhilin. The long tail of recommender systems and how to leverage it. In *RecSys '08 : Proceedings of the 2008 ACM conference on Recommender systems*, pages 11–18, New York, NY, USA, 2008. ACM.

[PTFS10] D. Poirier, I. Tellier, F. Fessant, and J. Schluth. Towards Text-Based Recommendations. In *RIAO 2010 : 9th international conference on Adaptivity, Personalization and Fusion of Heterogeneous Information, Proceedings RIAO 2010*, PARIS, France, 2010.

[PTL93] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 183–190, Morristown, NJ, USA, 1993. Association for Computational Linguistics.

[RIS+94a] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens : an open architecture for collaborative filtering of netnews. In *CSCW '94 : Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186, New York, NY, USA, 1994. ACM.

[RIS+94b] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens : an open architecture for collaborative filtering of netnews, 1994.

[RW03] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[SCM08] G. Shani, M. Chickering, and C. Meek. Mining recommendations from the web. In *RecSys '08 : Proceedings of the 2008 ACM conference on Recommender systems*, pages 35–42, New York, NY, USA, 2008. ACM.

[SDSO66] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. *The General Inquirer : A Computer Approach to Content Analysis*. MIT Press, 1966.

[SFHS07] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. Collaborative filtering recommender systems. pages 291–324, 2007.

[Sha51] E. Shannon. Prediction and entropy of printed english. *The Bell System Technical Journal*, pages 50–64, January 1951.

[SHCZ05] S. Shulman, E. Hovy, J. Callan, and S. Zavestoski. Language processing technologies for electronic rulemaking : a project highlight. In *dg.o 2005 : Proceedings of the 2005 national conference on Digital government research*, pages 87–88. Digital Government Society of North America, 2005.

[SK09] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009 :2–2, 2009.

[SKEK+07] Y. Seki, D. Kirk Evans, L.-W. Ku, H.-H. Chen, N. Kando, and C.-Y. Lin. Overview of opinion analysis pilot task at ntcir-6. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 265–278, Tokyo, Japan, 2007.

[SKEK+08] Y. Seki, D. Kirk Evans, L.-W. Ku, L. Sun, H.-H. Chen, and N. Kando. Overview of multilingual opinion analysis task at ntcir-7. In *Proceedings of NTCIR-7 Workshop Meeting*, pages 185–203, Tokyo, Japan, 2008.

[SKKR00] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce, 2000.

[SKKR01] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *WWW '01 : Proceedings of the 10th international conference on World Wide Web*, pages 285–295, New York, NY, USA, 2001. ACM.

[SL65] G. Salton and M. E. Lesk. The smart automatic document retrieval systems—an illustration. *Commun. ACM*, 8(6) :391–398, 1965.

[SMS05] D. Smith, S. Menon, and K. Sivakumar. Online peer and editorial recommendations, trust, and choice in virtual markets. *Journal of Interactive Marketing*, 19(3) :15–37, 2005.

[SPUP02] A. I. Schein, A. Popescul, Lyle H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *SIGIR '02 : Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260, New York, NY, USA, 2002. ACM.

[SS00] R. E. Schapire and Y. Singer. Boostexter : A boosting-based system for text categorization. *Machine Learning*, 39(2/3) :135–168, 2000.

[TL03] P. D. Turney and M. L. Littman. Measuring praise and criticism : Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4) :315–346, 2003.

[Tri07] A. Trinh. Classification de texte et estimation probabiliste par machine à vecteurs de support. In *Actes de l'atelier de clôture du 3ème DÉfi Fouille de Textes*, pages 69–82, Grenoble, France, 2007. AFIA.

[Tur02] P. D. Turney. Thumbs up or thumbs down ? : semantic orientation applied to unsupervised classification of reviews. In *ACL '02 : Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[UF98] L. Ungar and D. Foster. Clustering methods for collaborative filtering, 1998.

[Vap95] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[VMVS00] R. Van Meteren and M. Van Someren. *Using content-based filtering for recommendation*, volume 4203/2006, pages 312–321. Citeseer, 2000.

[WB84] Y. Wilks and J. Bien. Beliefs, points of view and multiple environments. In *Proc. of the international NATO symposium on Artificial and human intelligence*, pages 147–171, New York, NY, USA, 1984. Elsevier North-Holland, Inc.

[WWH04] T. Wilson, J. Wiebe, and R. Hwa. Just how mad are you ? finding strong and weak opinion clauses. In *In Proceedings of AAAI*, pages 761–769, 2004.

[YH03] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions : separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

Damien POIRIER

Des textes communautaires à la recommandation

La thèse concerne la transformation de données textuelles non structurées en données structurées et exploitables par des systèmes de recommandation. Deux grandes catégories d'informations sont utilisées dans le domaine des moteurs de recommandation : les données descriptives de contenus comme les méta-données ou les tags (filtrage thématique), et les données d'usages qui peuvent être des notes ou encore des pages Web visitées par exemple (filtrage collaboratif). D'autres données sont présentes sur le Web et ne sont pas encore réellement exploitées. Avec l'émergence du Web 2.0, les internautes sont de plus en plus amenés à partager leurs sentiments, opinions, expériences sur des produits, personnalités, films, musiques, etc. Les données textuelles produites par les utilisateurs représentent potentiellement des sources riches d'informations qui peuvent être complémentaires des données exploitées actuellement par les moteurs de recommandation et peuvent donc ouvrir de nouvelles voies d'études dans ce domaine en plein essor. Notre objectif dans le cadre de la thèse est de produire, à partir de commentaires issus de sites communautaires (blogs ou forums), des matrices d'entrées pertinentes pour les systèmes de recommandation. L'idée sous-jacente est de pouvoir enrichir un système pour un service débutant, qui possède encore peu d'utilisateurs propres, et donc peu de données d'usages, par des données issues d'autres utilisateurs. Nous faisons tout d'abord un état de l'art de la recommandation automatique. Nous présentons ensuite le moteur ainsi que les données utilisées pour les expérimentations. Le chapitre suivant décrit les premières expérimentations en mode thématique. Nous faisons ensuite un nouvel état de l'art sur la classification d'opinion. Pour finir, nous décrivons les expérimentations menées pour l'approche collaborative à l'aide de la classification d'opinion.

Mots clés : recommandation automatique, classification d'opinion, fouille de textes, Web 2.0.

From texts to recommendation

The thesis is about the transformation of unstructured textual data in structured data in order to be used by a recommender system. Recommender systems can operate on two main types of data: content descriptors as metadata or tags (content-based filtering), and usage data as rates or visited Web pages for example (collaborative filtering). Other data exist on the Web which are not used yet. With the emergence of the Web 2.0, users share their feelings, opinions, experiences on products, personalities, movies, music, etc. (through comments for example). This textual data generated by users potentially represent rich sources of information which can supplement data exploited by recommender systems. The exploitation of this kind of data could open new paths in this burgeoning field. Our objective in this thesis is to generate matrices relevant for recommender systems. The underlying idea is to enrich a system for a beginner service, which has still few own users, then too little usage data, by information on other users on the Web. The thesis begins with a state of the art of automatic recommendation. Then, we present the recommender systems and the textual corpus used for experiments. The next chapter presents first experiments with the content-based filtering approach. The next part contains the state of the art of opinion mining. Finally, we describe experiments done with collaborative filtering approach using opinion classification.

Keywords : recommender systems, opinion classification, texte mining, Web 2.0.

LIFO - Bâtiment IIIA
Rue Léonard de Vinci
B.P. 6759
F-45067 ORLEANS Cedex 2