UNIVERSITÉ DE MONTPELLIER II

ÉCOLE DOCTORALE I2S
INFORMATION, STRUCTURES, SYSTÈMES

T H È S E

pour obtenir le titre de

Docteur de l'Université de Montpellier II
Discipline : INFORMATIQUE

Présentée et soutenue par

Benjamin DUTHIL

De l'extraction des connaissances à la recommandation

Thèse préparée au LGI2P de Nîmes

soutenue le 3 décembre 2012

Jury

Rapporteurs :	Thierry Charnois	MCF-HDR, Université de Caen
	Éric Gaussier	Professeur, LIG Université Grenoble I
Examinateurs :	Patrice Bellot	Professeur, LSIS Université Marseille
	Mathieu Roche	MCF-HDR, Université Montpellier II
	François Trousset	Enseignant Chercheur, École des Mines d'Alès
Directeurs :	Jacky Montmain	Professeur, École des Mines d'Alès
	Pascal Poncelet	Professeur, Université Montpellier II

Résumé : Les technologies de l'information et le succès des services associés (forums, sites spécialisés, etc) ont ouvert la voie à un mode d'expression massive d'opinions sur les sujets les plus variés (e-commerce, critiques artistiques, etc). Cette profusion d'opinions constitue un véritable eldorado pour l'internaute, mais peut rapidement le conduire à une situation d'indécision car les avis déposés peuvent être fortement disparates voire contradictoires. Pour une gestion fiable et pertinente de l'information contenue dans ces avis, il est nécessaire de mettre en place des systèmes capables de traiter directement les opinions exprimées en langage naturel afin d'en contrôler la subjectivité et de gommer les effets de lissage des traitements statistiques. La plupart des systèmes dits de recommandation ne prennent pas en compte toute la richesse sémantique des critiques et leur associent souvent des systèmes d'évaluation qui nécessitent une implication conséquente et des compétences particulières chez l'internaute. Notre objectif est de minimiser l'intervention humaine dans le fonctionnement collaboratif des systèmes de recommandation en automatisant l'exploitation des données brutes que constituent les avis en langage naturel. Notre approche non supervisée de segmentation thématique extrait les sujets d'intérêt des critiques, puis notre technique d'analyse de sentiments calcule l'opinion exprimée sur ces critères. Ces méthodes d'extraction de connaissances combinées à des outils d'analyse multicritère adaptés à la fusion d'avis d'experts ouvrent la voie à des systèmes de recommandation pertinents, fiables et personnalisés.

Mots clés : Fouille de texte, Fouille de données, Extraction d'opinion, Extraction conceptuelle, Système de recommandation, analyse multicritère.

Abstract : Information Technology and the success of its related services (blogs, forums, etc.) have paved the way for a massive mode of opinion expression on the most varied subjects (e-commerce websites, art reviews, etc). This abundance of opinions could appear as a real gold mine for internet users, but it can also be a source of indecision because available opinions may be ill-assorted if not contradictory. A reliable and relevant information management of opinions bases requires systems able to directly analyze the content of opinions expressed in natural language. It allows controlling subjectivity in evaluation process and avoiding smoothing effects of statistical treatments. Most of the so-called recommender systems are unable to manage all the semantic richness of a review and prefer to associate to the review an assessment system that supposes a substantial implication and specific competences of the internet user. Our aim is minimizing user intervention in the collaborative functioning of recommender systems thanks to an automated processing of available reviews in natural language by the recommender system itself. Our topic segmentation method extracts the subjects of interest from the reviews, and then our sentiment analysis approach computes the opinion related to these criteria. These knowledge extraction methods are combined with multicriteria analysis techniques adapted to expert assessments fusion. This proposal should finally contribute to the coming of a new generation of more relevant, reliable and personalized recommender systems.

Keywords : Text-mining, Data-mining, Opinion-mining, Concept characterization, Recommender system, multicriteria analysis.

Remerciements

Je souhaite remercier ici toutes les personnes qui ont contribué à l'aboutissement de ces travaux de thèse.

Une thèse est un travail de longue haleine et une étape très importante dans la vie d'un jeune chercheur. Je remercie mes directeurs de thèse Pascal Poncelet et Jacky Montmain pour le soutien et la confiance qu'ils m'ont accordés. Leur disponibilité, leurs conseils avisés et leur bonne humeur ont rendu cette aventure possible, qu'ils trouvent dans ces quelques mots ma plus profonde gratitude.

J'adresse toute ma reconnaissance à François Trousset pour sa gentillesse et sa précieuse aide scientifique. Nos longues heures de discussions et sa grande ouverture scientifique ont renforcé mon goût pour la recherche et permis l'accomplissement de cette thèse.

Un grand merci également à Mr Éric Gaussier, Professeur à l'Université de Grenoble, ainsi que Mr Thierry Charnois, Maître conférence à l'Université de Caen d'avoir accepté d'être rapporteurs de ce mémoire de thèse. Leurs remarques et leurs questions pertinentes sur mon manuscrit me permettent d'envisager de nombreuses perspectives de recherches.

Je remercie Mr Patrice Bellot, Professeur à l'Université de Marseille ainsi que Mr Mathieu Roche, Maître de conférence à l'Université de Montpellier II, d'avoir accepté d'examiner ce travail de thèse. La richesse de leurs remarques m'encouragent à ouvrir de nouvelles pistes de recherche pour mes futurs travaux.

Enfin, je tiens à remercier chaleureusement les membres du LGI2P, pour leur aide et leurs encouragements dans les moments les plus difficiles. Plus particulièrement, je remercie Françoise Armand pour son soutien quotidien, sa joie de vivre, et son aide précieuse lors de la rédaction de ce manuscrit notamment. Je tiens aussi à remercier Sylvie Ranwez et Gérard Dray pour leur soutien durant ces trois années. Mes remerciements s'adressent maintenant aux thésards et postdocs du laboratoire avec j'ai partagé trois de camaraderie et de bonne humeur, merci à vous tous.

Valérie Roman et Claude Badiou m'ont aidé dans les tâches administratives avec

bienveillance et efficacité , qu'elles trouvent ici toute ma sympathie et ma gratitude.

Enfin, j'adresse mes pensées les plus sincères à ma famille, leur confiance, leurs encouragements de tous les instants, avec une mention particulière pour Adeline qui a été mon point d'ancrage au cours de ces trois années.

Table des matières

Table des figures

"La connaissance progresse en intégrant en elle l'incertitude, non en l'exorcisant."

Edgar Morin

Chapitre 1

Introduction générale

"***La connaissance des mots conduit à la connaissance des choses.***"
Platon

De tout temps les sociétés humaines ont produit, stocké et échangé des données, des informations dans le but de transmettre des connaissances. Aujourd'hui, ces données sont majoritairement numériques, et elles sont stockées sur des ordinateurs et échangées sur des réseaux de télécommunication informatiques comme Internet. Le nombre de personnes connectées à Internet couvre à ce jour plus de 30% de la population mondiale, dont 7% se trouve en Europe [1]. On enregistre en 2012 plus de 42 millions d'internautes en France, soit 65% de la population française [2]. La dernière décennie a connu une augmentation fulgurante des données engendrées par les nouvelles technologies de l'information. Actuellement la quantité d'informations stockées de manière numérique double tous les dix-huit mois. On estime qu'en 2012 il devrait circule un demi-zettaoctets (soit 2^{70} octets) de données hétérogènes sur Internet. Citons par exemple le réseau social *Facebook* [3] qui draine plus de 700 millions de comptes internet dans le monde. En février 2011, les statistiques parlent de 20,54 millions de comptes *Facebook* en France (+1,6 million en 6 mois), soit 31,8% de la population et 46% des abonnés français [4]. Cette croissance exponentielle des données se traduit par une difficulté à organiser et à analyser ces informations brutes ouvrant pourtant de nouvelles voies sur les chemins de la connaissance. La question

1. http://www.internetworldstats.com
2. http://www.mediametrie.fr/
3. http://www.facebook.com
4. http://fr.wikipedia.org/wiki/Facebook

n'est donc plus de disposer de l'information, mais de trouver l'information pertinente au bon moment. Accéder rapidement à l'information devient alors un réel challenge pour les internautes souvent désarmés face à cette masse d'informations. Le Web 2.0 a fortement contribué à cette ascension fulgurante des données en rendant le Web plus convivial et en offrant à chacun l'opportunité de s'exprimer par l'intermédiaire de blogs, de réseaux sociaux, de sites spécialisés, etc. Partager son expérience et profiter de celle des autres est devenu naturel lorsqu'il s'agit d'avoir des avis sur quelque sujet que ce soit. Le problème pour un internaute en quête d'une information pertinente n'est plus l'accès à cette information qui tend à être uniformément distribuée, mais de savoir la trouver au milieu d'imposants flux informationnels extrêmement "bruités". S'aventurer dans le monde de l'information sur des sujets, où ses connaissances sont limitées, est un pari risqué. Comment choisir un téléphone mobile ? Un hôtel ? un film ? Tout est sur la toile, à chacun de savoir consulter et de se faire son opinion. Où trouver la recommandation fiable et pertinente qui éclairera notre choix ? Cet eldorado de l'information plonge finalement l'utilisateur dans un abîme de perplexité, où il est souvent incapable de trouver, parmi la multitude d'avis publiés et parfois plus ou moins contradictoires ceux qui pourraient l'intéresser et lui permettraient de se faire un avis sur le produit ou le service. Ce constat sociétal explique le succès incroyable des sites de recommandation qui sont développés dans le but d'apporter une aide à l'utilisateur en lui proposant des produits ou services en adéquation avec ses préférences, ses attentes, ses goûts. Ces systèmes ont considérablement changé le comportement des internautes qui ne cherchent plus à l'aveugle l'information dont ils ont besoin, n'achètent plus sans "l'aval" de leur site de recommandation préféré. Cependant, mettre en place de tels systèmes reste une tâche délicate qui nécessite un travail préliminaire conséquent d'évaluation de chacun des éléments, appelés "items", à recommander, mais aussi d'identification des préférences de l'utilisateur. C'est la condition sinequanone pour offrir une recommandation pertinente et adaptée. Aujourd'hui, la qualité d'un système de recommandation est étroitement liée à sa capacité à prendre en compte et à traiter une grande quantité d'évaluations. En effet, plus le nombre d'items évalués est importante plus le système peut proposer un large et riche panel de recommandations. Ces évaluations ne se réduisent pas à un simple score attribué à un item, elles s'accompagnent d'un avis critique écrit en langage naturel, avis qui peut parfois sembler incohérent avec le score global. Les internautes préfèrent et de loin déposer un avis sous forme d'une critique plutôt que de noter des évaluations

sous une forme chiffrée, standardisée. Pour tenir compte de ce comportement, des différences d'expertise d'un individu ainsi que de la richesse sémantique du langage naturel, il faut pouvoir développer des systèmes de recommandation plus sophistiqués. Ceci constitue un défi majeur pour notre communauté scientifique : rendre les machines capables d'analyser et d'interpréter des données complexes comme les retours d'expérience écrits en langage naturel, afin de leur donner du sens, d'y détecter opinions et tendances et pouvoir ainsi limiter les tâches d'expertise demandées aux utilisateurs.

1.1 Motivations

Chacun d'entre nous semble de plus en plus dépendant de l'information : un cyberconsommateur souhaitant acheter n'importe quel bien ou service sur le net ne saurait s'affranchir de consulter l'ensemble des offres commerciales via un comparateur de prix (ex : Kelkoo), de s'enquérir des biens ou produits réputés équivalents (service proposé sur la quasi-totalité des e-retailers), de s'assurer de la fiabilité du site sur lequel il envisage de faire son achat via un e-recommender (ex : Ciao.com)... En retour, les "marques" mettent à la disposition du consommateur de plus en plus d'informations relatives à leurs produits ou services sur la toile. Sur cet exemple banal, on voit à quel point la société de l'information a une emprise sur les choix et décision d'un individu. Si la quantité d'informations disponible est phénoménale, elle n'est pas un gage de qualité et de fiabilité. Quelle crédibilité accorder à l'évaluation d'un produit lorsqu'elle est proposée par le site marchand lui-même (ex : Fnac, Amazon) ? Le cyberconsommateur se trouve vite démuni dans cet océan d'informations. Sa connaissance souvent très superficielle du site, du produit ou du service recherché, les informations souvent subjectives, parfois mêmes contradictoires qu'il consulte ne lui permettent pas d'évaluer et de comparer objectivement, rationnellement et exhaustivement une pléthore de sites tous susceptibles de satisfaire à son besoin [McNee *et al.* 2003, Terveen & Hill 2001]. Ainsi confronté à cet inépuisable espace des possibles, les clients sans a priori, ont tendance à se tourner naturellement vers les opinions et les expériences d'autres cyberconsommateurs [Montmain *et al.* 2005]. Ce comportement est à l'origine du concept de e-recommandation : les sites de e-recommandation sont dédiés à susciter, gérer et automatiser les partages d'opinions et de recommandations de communautés de cyber-consommateurs (ex : Ciao.com, Leguide.com). On en vient à consulter de l'information sur l'information,

à recourir aux statistiques pour estimer la fiabilité de ce qu'on peut lire [Denguir-Rekik *et al.* 2006, Denguir-Rekik *et al.* 2009] !

L'automatisation de systèmes de recommandation est une entreprise difficile car il ne s'agit pas simplement de rechercher et d'afficher de simples données pour aider l'Homme dans sa quête. En effet, pour [Brooking 1998], les données sont des faits, des images, des nombres présentés sans aucun contexte. La notion de donnée est donc perçue comme la couche de plus bas niveau dans la hiérarchie conceptuelle du savoir. C'est sur elle que s'élaborent les notions d'information et de connaissance. Pour [Baizet 2004] citant [Ermine *et al.* 1996, Tsuchiya 1995], les informations sont pour les uns des données triées, sélectionnées et organisées par un individu dans un but précis, pour les autres des données auxquelles sont associées des significations par la description de méthodes et procédures d'utilisation. La notion d'information présente donc un degré conceptuel plus élevé que la notion de donnée dans la valeur et la signification qu'elle occupe dans l'application où elle est utilisée. La notion d'information est de plus en plus associée à la possibilité de traitement informatique c'est-à-dire son stockage sous forme exploitable pour les applications. Pour ces raisons, une information doit donc être associée à une représentation formelle permettant sa traduction informatique et conceptuelle. [Tsuchiya 1995] formule sa conception de la notion de connaissance ainsi : "L'information ne devient connaissance que lorsqu'elle est comprise par le schéma d'interprétation du receveur qui lui donne un sens" (sense-read). Toute information inconsistante avec ce schéma d'interprétation n'est pas perçue dans la plupart des cas. Ainsi la "commensurabilité" des schémas d'interprétation des membres de l'organisation est indispensable pour que les connaissances individuelles soient partagées. Pour [Penalva & Montmain 2002], la connaissance, à l'inverse de l'information, repose sur un engagement, des systèmes de valeurs et de croyances, sur l'intention. La connaissance est bâtie à partir de l'information pour faire quelque chose, pour agir... L'automatisation de systèmes de recommandation correspond donc bien à l'idée qu'il s'agit d'extraire de simples données accessibles sur le web, des connaissances utiles au sens des objectifs et préférences d'un utilisateur.

Ce mémoire s'intéresse donc à la problématique d'extraction automatique de connaissances appliquée aux systèmes de recommandation, dont l'objectif est d'aider l'Homme à exploiter les données massives mises à sa disposition sur Internet, qu'il s'agisse de produits ou services. Cependant, un problème majeur se pose : l'Homme et la machine n'utilisent pas les mêmes modes cognitifs, la machine ma-

nipule des données binaires (octets), l'Homme utilise des mots auxquels il rattache des concepts. Il est donc difficile pour la machine d'accorder une valeur conceptuelle à des données dont elle ignore la sémantique. Les algorithmes actuels les plus sophistiqués en sont à leurs balbutiements en matière d'analyse conceptuelle et ils se basent généralement sur les simples occurrences de termes spécifiques. La principale difficulté est de réussir à faire interpréter par la machine ce que l'auteur d'un texte a voulu dire : quelles sont les idées exprimées ? Quelles émotions ou sentiments se dégagent ?, etc. C'est là l'une des principales ambitions du web sémantique qui considère au-delà du simple texte contenu dans des pages HTML (mots, syntaxe...) la sémantique contenue dans chacune de ces pages web. Le passage entre le texte brut et l'extraction des connaissances contenues dans le document est à ce jour un réel problème. C'est pourquoi, les tâches de segmentation thématique de texte et d'extraction d'opinion sont aujourd'hui un des nombreux enjeux des SIC.
Dans ce manuscrit, nous nous intéressons à trois problématiques distinctes : la segmentation thématique de texte d'une part, qui a pour objectif d'identifier les parties d'un texte qui traitent de concepts choisis, et l'extraction d'opinion d'autre part, dont l'objectif est de déterminer la polarité (positif/négatif) d'un texte, et enfin l'intégration de ces outils d'extraction des connaissances aux systèmes de recommandation pour réduire au maximum l'intervention humaine . Ainsi, la segmentation thématique permettra d'isoler dans un texte les thèmes liés aux attentes de l'utilisateur (ses critères de choix) alors que l'extraction d'opinion permettra d'exprimer un sentiment sur chacun de ces thèmes.

1.2 Cas d'application

Au cours de ce travail, nous nous focaliserons sur l'extraction des connaissances pour les systèmes de recommandation. Nous illustrerons notre propos en considérant comme domaine d'étude le **cinéma** et plus particulièrement des critiques cinématographiques écrites en langage naturel. Nous avons choisi d'utiliser des documents en langue anglaise pour toucher un plus large public, mais les méthodes et principes présentés ici ne sont pas spécifiques à l'anglais, et peuvent être appliqués à d'autres langues.

La figure 1.1 est un exemple de critique cinématographique qui pourrait être considérée par un système de recommandation. Elle concerne le film *Avatar* et est ex-

FIGURE 1.1 – Site IMDB : Exemple de critique cinématographique pour le film *Avatar*.

traite du site web IMDB [5] spécialisé dans le domaine cinématographique. L'*Internet Movie Database* (IMDB) est une base de donnée en ligne répertoriant la majeure partie de la production cinématographique mondiale. Ce système compte à ce jour plus de 57 millions d'usagers et offre à chacun la possibilité d'exprimer son avis sur les œuvres référencées. La critique présentée ici a été rédigée et évaluée par un internaute. Nous pouvons remarquer que l'auteur attribue une opinion très positive au film, avec la note globale de neuf étoiles sur dix. Cependant, en lisant la critique on se rend compte que même si l'opinion globale est majoritairement positive, l'auteur note malgré tout que l'histoire est pathétique (***The Avatar's story is pathetic***) ce qui est plutôt une critique négative. Ainsi, la note de *9/10* accordée par l'inter-

5. http://www.imdb.com

naute ne reflète pas ici la complexité et la richesse de son avis et encore moins de son système de préférences. Ce constat nous montre que l'opinion repose sur des critères de choix, et que visiblement, sur cet exemple, les critères "mise en scène" et "acteur" contribuent majoritairement à l'opinion globale émise par le critique. En revanche, le critère "scénario" ne semble pas être pris en compte dans son référentiel de préférence. Nous constatons alors que le score agrégé masque la divergence qui existe entre les critères "acteur" d'une part et "scénario" d'autre part et donc restitue mal la richesse sémantique du document. Il est donc nécessaire de considérer l'opinion relative à des critères pour pouvoir délivrer une recommandation plus précise.

À partir de cet exemple, nous pouvons identifier les trois principales difficultés qui apparaissent lorsqu'on veut automatiser l'évaluation de critiques :

1. La première difficulté est d'identifier dans une critique les différentes thématiques d'intérêt abordées et qui font référence à des concepts parfois complexes. Extraire les parties d'un document traitant d'un concept n'est pas une tâche triviale. En effet, la langue permet une grande flexibilité d'expression d'un concept, car même si l'idée (concept) à exprimer est commune à tous, son expression peut être multiple et dépend notamment de la culture du locuteur, de son point de vue, de son vocabulaire.
2. La seconde difficulté est d'extraire l'opinion par rapport à des critères identifiés. L'expression d'une opinion est un processus complexe spécifique au contexte dans lequel il est déployé. De plus, l'opinion est souvent nuancée par des quantificateurs qui compliquent notablement son extraction en lui ajoutant une subjectivité difficile à analyser par un processus automatisé.
3. La troisième et dernière difficulté concerne l'intégration d'un processus d'extraction automatique des connaissances à un système de recommandation capable de gérer des données imprécises parce que subjectives, incertaines parce que multiples et variées, via un processus d'évaluation multicritère.

1.3 Organisation du mémoire

Ce mémoire est organisé de la manière suivante : le chapitre 2 est un état de l'art sur l'évaluation dans les systèmes de recommandation multicritères et sur l'extraction des connaissances. Ce chapitre a pour but de présenter le contexte dans lequel se placent nos travaux. Le chapitre 3 présente notre approche de segmenta-

tion thématique de texte et propose les expériences qui en montrent la pertinence. Le chapitre 4 décrit notre approche d'extraction d'opinion ainsi que des expériences qui la valident. Le chapitre 5 illustre par des prototypes logiciels les approches et algorithmes présentés dans les chapitres 3 et 4. Le chapitre 6 montre comment nous développons un nouveau type de système de recommandation par l'intégration des techniques de segmentation thématique 3 et d'extraction d'opinion 4. Le chapitre 7 est une conclusion générale.

Chapitre 2

État de l'art

"***Toute connaissance dégénère en probabilité.***"

David Hume

Sommaire

2.1 Les systèmes de recommandation

Ces dernières années, de nombreuses entreprises et sites Web ont mis en place des systèmes permettant l'analyse des préférences de leurs utilisateurs afin d'améliorer la qualité de leurs services et produits. En analysant les préférences des utilisateurs, ils sont en mesure d'améliorer leurs performances, et ainsi, de mieux répondre aux attentes de leurs clients. À ce jour, les systèmes de recommandation sont présents dans différents domaines tels que le tourisme/les loisirs (hôtels, restaurants, les parcs, les plages, etc) [Loh *et al.* 2004], la publicité [Cheung *et al.* 2003], le commerce électronique [Ghani & Fano 2002], les films, la TV, la musique, les sites de rencontre, entre autres. En raison de l'explosion de la quantité de données diffusées sur l'Internet ces dernières années, rechercher et trouver des produits, des services ou des contenus pertinents est devenu une tâche difficile pour l'utilisateur qui est souvent perdu face à une telle masse d'informations. Ceci explique l'intérêt croissant porté aux systèmes de recommandation tant par les utilisateurs que par les sites commerciaux. Un système de recommandation doit permettre d'instrumenter la relation que l'homme entretient avec le monde de l'information pour faciliter sa décision et son action, pour contrôler et limiter les surcharges cognitives que le web peut générer. Le système de recommandation devient alors un Système Interactif d'Aide à la Décision (SIAD).

La problématique de recommandation a été identifiée comme étant le moyen d'aider les individus à trouver des informations, ou des éléments, qui sont susceptibles de les intéresser. Généralement, on considère un ensemble d'utilisateurs $Users$ et un ensemble d'éléments $Items$ (services ou produits) pouvant être recommandés à chacun des utilisateurs. Une fonction d'utilité permet de mesurer la pertinence de l'élément $item \in Items$ à recommander à un utilisateur. Elle est définie par $U : Users \times Items \longrightarrow E$, où E est l'espace d'évaluations (nombre d'étoiles, score dans [0 ;1], etc). L'objectif est alors de chercher les éléments qui maximisent la fonction d'utilité d'un utilisateur $user \in Users$ donné tels que : $item* = arg \max_{item \in Items} U(user, item)$.

Bien souvent l'évaluation d'un quelconque élément porte sur plusieurs facettes. L'évaluation est multidimensionnelle. Dans son état de l'art, Adomavicius [Adomavicius & Kwon 2007, Adomavicius *et al.* 2011, Manouselis & Costopoulou 2007] parle de recommandation multicritère. La définition de critère reste pourtant vague dans les systèmes de recommandation puisqu'au critère peut tout aussi bien être associée la valeur qui caractérise un attribut ou un item (cylindrée du moteur : 1.9l)

que l'utilité accordée par un utilisateur (cylindrée du moteur : *** ou 8.5/10). En pratique, les **critères** permettent d'évaluer un item sur un ensemble de caractéristiques élémentaires (ex : la cylindrée de la voiture) ou complexes (ex : la puissance ou le confort de la voiture) d'un item, mais l'évaluation partielle sur chacune de ces dimensions n'est pas forcément explicitée dans tous les systèmes de recommandation. Autrement dit, l'utilisateur dispose certes souvent des données sur plusieurs caractéristiques de l'item, mais les évaluations fournies sont en général globales (ex : excellente voiture) même si implicitement elles correspondent à une évaluation sur le produit cartésien $\prod_{i=1}^{n} E_i$ où les E_i sont les échelles d'évaluation associées aux attributs de l'évaluation multidimensionnelle. Par conséquent, la plupart du temps, il s'agit d'une évaluation certes multicritère mais implicite, et pour laquelle on ne dispose que d'un critère de synthèse unique au final. Il n'est donc généralement pas possible de savoir quels critères ont principalement contribué à l'impression générale restituée pour un item donné, ce qui rend le résultat de l'évaluation moins pertinent pour la recommandation. Dans de récents travaux, cette restriction a été considérée comme une limite à la fiabilité de la recommandation [Adomavicius & Kwon 2007, Adomavicius *et al.* 2011, Manouselis & Costopoulou 2007]. Dans son état de l'art très complet sur les systèmes de recommandation, Adomavicius réserve donc une place toute particulière à la recommandation multicritère car elle semblerait pouvoir s'imposer dans les années à venir puisqu'elle offre une recommandation plus pertinente et plus fiable. L'espace d'évaluation multicritère des items permet en effet une analyse plus précise et l'explicitation de ces dimensions permet à l'utilisateur de comprendre si l'évaluation globale d'un item est due à un critère qui l'intéresse lui personnellement ou pas, ce qui confère à rendre le système de recommandation plus fiable. Ainsi,le multicritère contribue à améliorer la qualité des recommandations car il permet en fait une représentation plus précise du système de préférences d'un utilisateur.

L'objectif théorique dans les systèmes de recommandation doit donc être d'identifier au mieux le système de préférences de l'utilisateur sur l'ensemble des critères considérés, et ainsi de pouvoir restituer une recommandation pertinente parce que personnalisée sur la base de l'évaluation multicritère. Cependant, dans la pratique, cet objectif est quelque peu restreint en fonction des informations effectivement recueillies auprès des utilisateurs. En effet, les utilisateurs de sites de recommandation n'ont pas toujours le temps, la patience, le recul ou les compétences pour se lancer dans une véritable évaluation multicritère des items. L'évaluation des items va donc

être plus ou moins riche d'un système de recommandation à l'autre et l'identification des systèmes de préférences des utilisateurs s'en ressentira. Afin, de préciser ce diagnostic et d'envisager des améliorations des systèmes de recommandation, nous allons revisiter formellement la classification des systèmes de recommandation multicritères proposée par Adomavicius. [Adomavicius *et al.* 2011] identifie trois modèles pour la recommandation multicritère qui ont pour objectif commun de déterminer au mieux le système de préférences de l'utilisateur pour lui proposer un ensemble d'items correspondant parfaitement à ses attentes :

1. *Modèle multicritère de préférences de contenu (MMPC).* Ce modèle s'appuie sur une description multicritère des items : le système de recommandation fournit les valeurs (numériques ou symboliques) de n caractéristiques des items (ex : "type du film : polar"). Les utilisateurs fournissent uniquement une évaluation globale des items. Pour la recommandation, le système de recommandation va alors chercher des items ayant des caractéristiques voisines des items que l'utilisateur a appréciés par le passé. La notion de voisinage se réfère donc à une distance sur les caractéristiques des items (et non pas à une distance sur $\prod_{i=1}^{n} E_i$). La recommandation repose donc sur une distance décorrélée des préférences de l'utilisateur : elle ne tient pas compte des éventuelles importances relatives que l'utilisateur accorde à tel ou tel critère, des échelles d'évaluation subjectives E_i qui lui sont propres, etc. Le système de préférences de l'utilisateur est déduit de manière simpliste en considérant pour chaque critère l'ensemble des valeurs associées aux items qu'il a appréciés dans le passé sans tenir compte de ce que l'utilisateur avait réellement trouvé pertinent lorsqu'il avait émis son avis (ex : il a au final mis une excellente note à une "escapade gourmande" parce qu'il garde un souvenir ému du repas aux chandelles bien que la décoration de la chambre laissât à désirer, la justification de l'évaluation globale est perdue). Par exemple, dans un système de recommandation de films, les critères peuvent être "le genre de film" (comédie, thriller, film d'auteur, etc.), le "nom des acteurs", etc. Le système de recommandation va donc chercher des films du même genre que les films réputés préférés par l'utilisateur avec les mêmes acteurs, le même réalisateur, etc. Ces systèmes n'encouragent pas la "découverte" de nouveaux films, puisqu'ils recommandent uniquement des films ayant les mêmes descriptions que ceux réputés appréciés sans aucune visibilité sur les critères qui ont fait que tel ou tel film ait été apprécié par l'utilisateur. Puisqu'ils ne peuvent prendre en compte un grand nombre de critères

pour fournir une note globale, ces systèmes de recommandation ne peuvent se fonder que sur un modèle rudimentaire des préférences de l'utilisateur ce qui nuit à la pertinence et à la fiabilité des recommandations. Plus formellement, considérons deux items x et y et l'évaluation globale $u(x)$ de *user* sur l'item x. Le système de recommandation calcule la distance $d(x, y)$ entre les deux items sur la base de leurs caractéristiques stockées par le système. Ensuite, on fait l'hypothèse suivante : si $d(x, y) \leq \varepsilon$, alors on peut supposer que l'évaluation de y sera telle $u(y) \simeq u(x)$, autrement dit que $\mid u(y) - u(x) \mid \leq \varepsilon'$. Par conséquent, si l'item x a été apprécié, le système recommandera simplement les items tels que $d(x, y) \leq \varepsilon$.

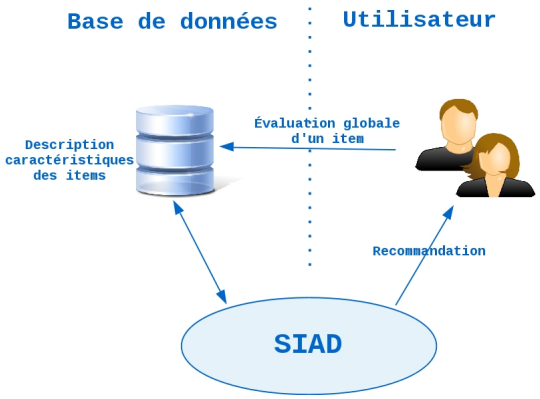

FIGURE 2.1 – Schéma global de fonctionnement du modèle MMPC

La figure 2.1 illustre le fonctionnement global du modèle de type *MMPC*. L'utilisateur fournit uniquement une évaluation globale sur un (ou plusieurs) item(s), au système qui va ensuite calculer une distance $d(x, y) \leq \varepsilon$ entre l'item préféré et les autres items de la base de données. Le SIAD recommande ensuite les items les plus proches de l'item préféré par l'utilisateur, c'est-à-dire les items qui ont simplement des caractéristiques (et pas des évaluations partielles) similaires à l'item préféré.

Considérons pour l'exemple le site web *Allocine*[1] qui fournit des services et des informations cinématographiques en ligne et qui utilise un modèle de type *MMPC* pour recommander des films à ses utilisateurs en fonction de leurs préférences. À partir d'une évaluation sur un film, le système va recommander un ensemble de films susceptibles d'intéresser l'utilisateur. Considérons (c.f. figure 2.2) un utilisateur ayant évalué le film "Avatar" avec une note de 5

1. http://www.allocine.fr/

étoiles sur 5, il se voit alors proposer quatre films susceptibles de lui convenir : "John Carter", "Avengers", "E.T. l'extra-terrestre et "La Planète des singes : les origines". Considérons que l'utilisateur ait évalué le film "Avatar" comme un excellent film uniquement parce qu'il est un fan inconditionnel de "Sigourney Weaver" et qu'il l'a trouvée particulièrement remarquable dans ce film. La présence de "Sigourney Weaver" dans "Avatar" contribue donc vraisemblablement à l'évaluation à 5 étoiles pour cet utilisateur. Les quatre films proposés par le système partagent clairement plusieurs caractéristiques avec "Avatar" mais en tout cas pas le critère "Acteurs". Dans un tel cas, le modèle des préférences utilisateurs rudimentaire utilisé par le site web *Allocine* risque fort de décevoir le fan de Sigourney. Cet exemple illustre le fait que ce type de SIAD n'est pas forcement bien adapté à une telle utilisation et que le modèle des préférences y est trop rudimentaire. De plus, cela met en évidence que la distance sur les caractéristiques des items n'est pas forcément bien adaptée à la recommandation puisque les évaluations partielles de chacune des caractéristiques ne sont pas explicitées et que l'importance relative des critères propres à l'utilisateur n'entre pas dans le calcul de cette distance. Un fan de Sigourney et un passionné de science fiction vont partager un excellent avis sur "Avatar" et pourtant leurs goûts cinématographiques communs peuvent très bien s'arrêter là. Ce type de SIAD dispose d'un modèle des préférences trop simple pour éviter ces recommandations non pertinentes. Pire encore, il se peut très bien que le fan de Sigourney déteste la science fiction, et que par conséquent aucun des quatre films qui lui sont proposés ne correspond à ses attentes... ce sont simplement des films qui sont "objectivement" proches de "Avatar" qui lui ont été retournés. Pour pallier ce problème, les *modèles de recherche de contenu multi-attributs et de filtrage* sont apparus en offrant une recommandation qui s'appuie sur un modèle plus fin du système de préférences de l'utilisateur.

2. *Modèle de recherche de contenu multicritère et de filtrage (MMRC).* Ce modèle permet à un utilisateur de spécifier directement ses préférences sur chacun des critères relatifs aux items au travers de sa méthode derecherche et de filtrage (ex : l'utilisateur filtre selon le critère "Science Fiction (SF)" puis parmi les films de "SF", il cherche ceux où joue Sigourney Weaver ; ou bien il indique explicitement dans le système que le critère "Genre" est plus important pour lui que le critère "Scénario", ou encore il donne des poids aux critères, etc.). Cela revient à ne pas accorder la même importance à chacun

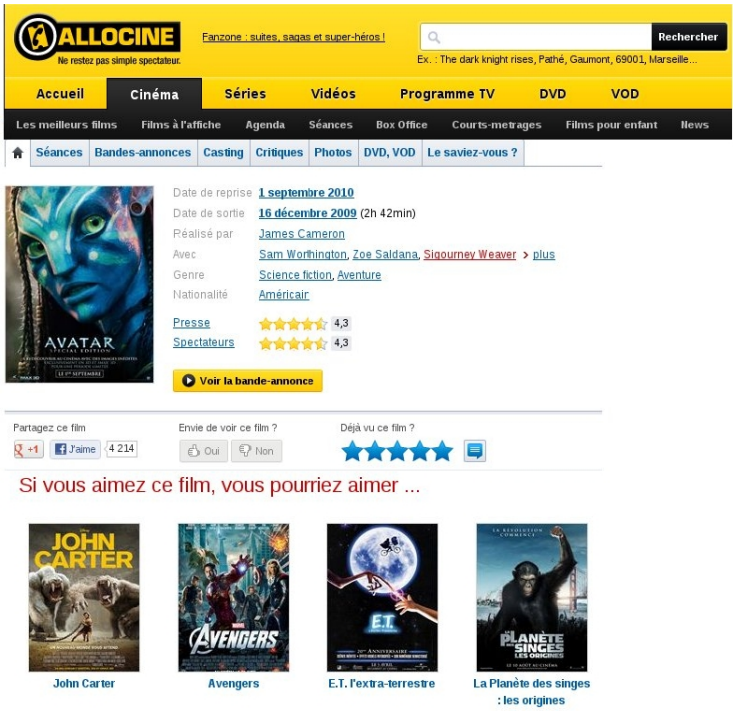

FIGURE 2.2 – Un exemple de modèle MMPC : le site *Allocine*

des critères de l'évaluation. Cette information renseignée par l'utilisateur dans le système permet d'établir une distance entre les items qui tienne compte des préférences de l'utilisateur contrairement aux *MMPC*. La distance $d(x, y)$ entre deux items reste certes toujours sur la base de leurs caractéristiques stockées par le système, mais ce sont les préférences de l'utilisateur qui permettent de la choisir en fonction du profil utilisateur (prise en compte de la subjectivité de l'utilisateur auquel s'adresse la recommandation). Plus formellement, l'idée est la suivante. A partir de la liste de critères $(C_1, \ldots, C_n)$, l'introduction d'un ordre partiel sur l'importance relative des critères permet de réécrire la liste sous la forme $(P_{(1)}, \ldots, P_{(n)})$ où (.) est une permutation telle que $\forall i \in 1 \ldots n, C_{(i)} \succcurlyeq C_{(i+1)}$($C_{(i)}$ est préféré à $C_{(i+1)}$ ou est plus important que $C_{(i+1)}$). Notons que l'affectation directe de poids aux critères n'est qu'un cas particulier. La distance $d(x, y)$ choisie peut alors être par exemple une distance lexicographique basée sur l'ordre partiel des critères pour l'utilisateur. On retient en premier lieu les items y tels que $d(x, y) =\mid x_{(1)} - y_{(1)} \mid \leq \varepsilon$,... puis parmi ceux-ci, ceux tels que $\mid x_{(i)} - y_{(i)} \mid \leq \varepsilon$, puis $\mid x_{(i+1)} - y_{(i+1)} \mid \leq \varepsilon$, etc. L'idée des distances qui intègrent l'ordre partiel des critères est que : plus deux

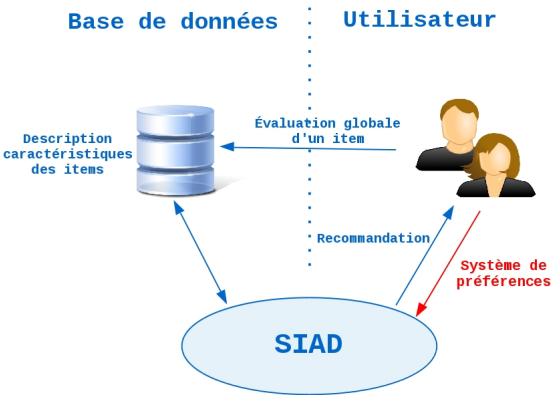

FIGURE 2.3 – Schéma global de fonctionnement du modèle MMRC

items partagent de caractéristiques sur des critères importants pour l'utlisateur, plus ils sont considérés proches. C'est une façon d'intégrer la subjectivité et les préférences de l'utilisateur dans le système de recommandation. L'idée reste la même que pour les systèmes de type *MMPC* : si $d(x, y) \leq \varepsilon$ alors $\mid u(y) - u(x) \mid \leq \varepsilon'$ et le système peut donc raisonnablement recommander y si x a été apprécié, seule la sémantique de la distance qui intègre la subjectivité de l'utilisateur diffère entre les *MMPC* et les *MMRC*. Le modèle de préférence de l'utilisateur est donc moins rudimentaire ici que dans les *MMPC*. Cependant, il reste encore incomplet en particulier parce que la distance entre les items est une distance sur la valeur des attributs et non pas sur l'évaluation de ceux-ci en terme de degré de satisfaction ou d'utilité élémentaire. Par exemple, sur la base des critères "Genre" et "Acteur", les deux films de SF où joue Sigourney Weaver que sont "Avatar" et "Alien" seront considérés proches l'un de l'autre quel que soit l'ordre induit sur les critères par l'utilisateur. Ce n'est pas pour autant que l'utilisateur apprécie les deux films : il peut très bien considérer l'un comme un "bon" film de SF et l'autre comme un "mauvais" film de SF. Sur cet exemple, on voit bien que ce n'est pas tant la valeur de la caractéristique "SF" elle-même qui importe, mais l'appréciation qui lui est associée. Il est donc nécessaire de considérer, en plus de l'ordre introduit par l'utilisateur sur les critères, une distance basée sur l'espace des évaluations (des scores, notes, etc.) et pas simplement sur l'espace des caractéristiques. Ainsi, la recommandation ne se base plus uniquement sur du filtrage de contenu mais sur une "réelle" évaluation multicritère comme le propose le *modèle Multicritère basé sur la découverte des préférences à partir d'évaluations (MMPE).*

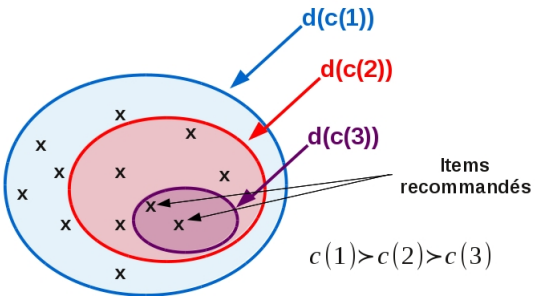

FIGURE 2.4 – Exemple sur le principe de clustering où trois critères sont considérés.

La figure 2.3 illustre le fonctionnement du modèle MMRC. L'utilisateur fournit, en plus d'une évaluation globale sur un ensemble d'items, son système de préférences au travers d'un ordre sur l'importance relative des critères. L'objectif ensuite est de trouver les items dont les caractéristiques sont proches des items préférés pour les critères qui lui importent le plus. La figure 2.4 illustre le principe de "clustering" associé à la sélection des items pertinents avec l'ordre sur les critères comme modèle de préférence de l'utilisateur. Si $c_{(1)} = c_2 \succ c_{(2)} = c_3 \succ c_{(3)} = c_1$, l'idée est de rechercher d'abord les films qui sont proches selon le critère prioritaire c_2, puis parmi ceux-ci ceux proches selon c_3 et en dernier lieu ceux proches selon c_1. L'exemple ici met en évidence deux items recommandés au sens d'une distance qui respecte toutes les préférences de l'utilisateur parmi tous les items de la base de données.

Prenons l'exemple du site web *CinéKiosk*[2] spécialisé dans la recommandation de films à partir des préférences de l'utilisateur (c.f. figure 2.5). Celui-ci, dans un premier temps répond à une enquête qui permet au système de déduire son système de préférences. L'enquête consiste à déterminer un ordre de préférences sur une base de films en demandant à l'utilisateur de faire une comparaison par paire. Un classement des films de la base est donc établi à partir duquel on en déduit un ordre sur l'importance des critères. Il s'agit d'une méthode d'identification indirecte. L'idée d'une telle méthode basée sur une comparaison par paire est la suivante : si tous les films préférés par l'utilisateur sont des films de SF alors le genre est un critère prioritaire pour l'utilisateur. À partir du système de préférences identifié, on propose une vingtaine de films à l'utilisateur qui sont susceptibles de lui convenir.

3. *Modèle Multicritère basé sur la découverte des préférences à partir d'évalua-*

2. http://beta.cinekiosk.tv/

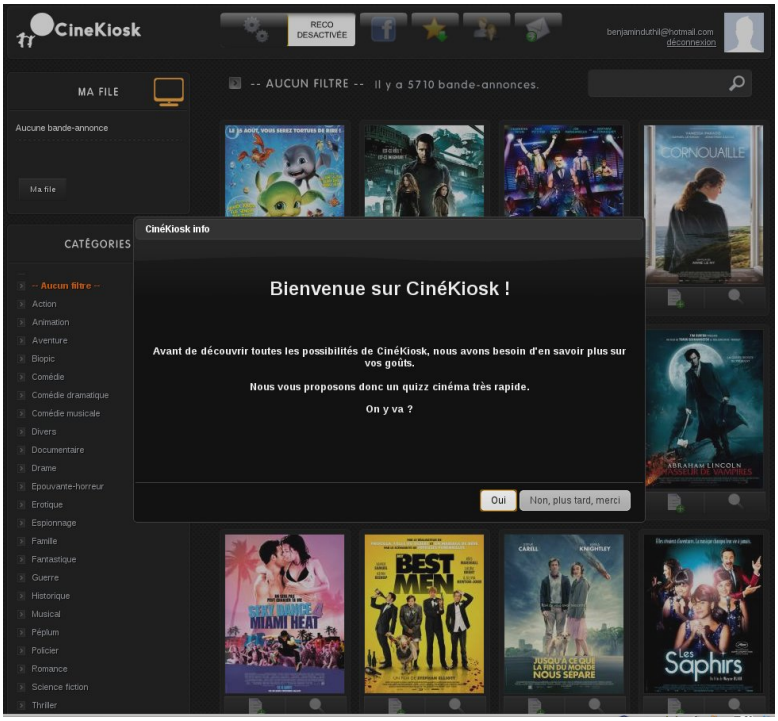

FIGURE 2.5 – Un exemple de modèle MMRC : le site *CinéKiosk*

tions (MMPE). Ces systèmes permettent à un utilisateur de spécifier ses préférences à travers des notes sur un espace de n critères (par exemple, le scénario du film a valu une note de 2 et les effets spéciaux une note de 5 pour un même film). Autrement dit, un item est décrit par un vecteur d'évaluations partielles sur l'ensemble des critères : $x = (u_1(x), .., u_n(x))$ avec $u_i(x) \in E_i$ l'évaluation partielle de x selon le critère i et l'échelle E_i. Cette fois-ci la distance entre les items peut être définie sur l'espace des évaluations partielles des items. Plus formellement, pour un item x, ce type de systèmes propose d'une part les évaluations partielles $u_i(x)$ et d'autre part $u(x)$, l'évaluation globale de l'item x. Considérons l'utilisateur *me* qui a donné son avis sur un ensemble d'items x. On peut alors chercher les utilisateurs *other* qui ont évalué ces items de la même façon : $\{other/\exists x \mid u^{me}(x) - u^{other}(x) \mid < \varepsilon\}$ où $u^{individu}(x)$ est l'évaluation de x par *individu*. Parmi ceux-ci, on peut sélectionner ceux qui ont attribué les mêmes évaluations partielles aux items x évalués à la fois par *other* et *me*. A partir de cette analyse, le système de recommandation identifie les utilisateurs *other* qui ont des analyses proches de *me* sur la base des items évalués en commun. La recommandation consiste alors à proposer à *me* les

films appréciés par les utilisateurs *other* sélectionnés. Plus le nombre d'items évalués en commun par *me* et *other* est grand, plus la recommandation est supposée fiable. Plus formellement, on introduit une distance sur $\prod_{i=1}^{n} E_i \times E$: $d((u_1^{me}(x), \ldots, u_n^{me}(x), u^{me}(x)), (u_1^{other}(x), \ldots, u_n^{other}(x), u^{other}(x))) \leq \varepsilon$ qui permet de sélectionner les individus qui font la même analyse de l'item x que *me*. Comme pour les deux types de systèmes de recommandation précédents, cette distance peut ou non intégrer un ordre partiel sur l'importance des critères. Si en plus des évaluations partielles, on dispose de cet ordre, alors le modèle des préférences de l'utilisateur intégré dans le système de recommandation est évidemment plus précis. Le modèle du système de préférence de *me* est identifié indirectement. Une autre façon de procéder avec un système de recommandation qui fournit les évaluations partielles sur les items, est de calculer directement un modèle du système de préférence de l'utilisateur *me*. Ainsi, sur l'ensemble des items x que *me* a évalué, on dispose d'une part des évaluations partielles $(u_1^{other}(x), \ldots, u_n^{other}(x))$ et d'autre part des évaluations globales $u^{me}(x)$. L'idée est d'identifier directement pour l'utilisateur *me* la relation qui existe entre ses évaluations globales et partielles : cette relation décrit le modèle du système de préférences de *me*, et peut dans certains cas être un modèle d'agrégation h : $\prod_{i=1}^{n} E_i \to E$ défini par $Argmin_h \sum_x \left(u(x) - \sum_{i=1}^{n} h(u_1(x), \ldots, u_n(x)) \right)^2$. Si de plus, h est linéaire, le modèle du système de préférence de *me* est donné par $Argmin_h \sum_{x \in X} \left(u(x) - \sum_{i=1}^{n} p_i.u_i(x) \right)^2$ et se résume à une distribution de poids sur les critères. On peut alors ensuite calculer le score global des items de la base avec l'opérateur h et recommander à *me* les items les mieux évalués au sens de h. On peut également envisager des stratégies mixtes (directes et indirectes) pour l'identification du système de préférences de *me*. Les systèmes du type *MMPE* proposent des recommandations plus pertinentes aux utilisateurs parce qu'ils intègrent explicitement un modèle plus précis du système de préférence des utilisateurs. Leur principal défaut est qu'ils requièrent un comportement collaboratif plus contraignant pour les utilisateurs qui doivent se plier à l'évaluation des items selon une base de critères pour compléter leur analyse. Si les *MMPE* sont clairement les plus à même de délivrer une recommandation fiable et pertinente, renseigner leur base est une tâche contraignante.

La figure 2.6 illustre le fonctionnement d'un SIAD basé sur le modèle *MMPE*. Le système impose à l'utilisateur une évaluation globale et des évaluations

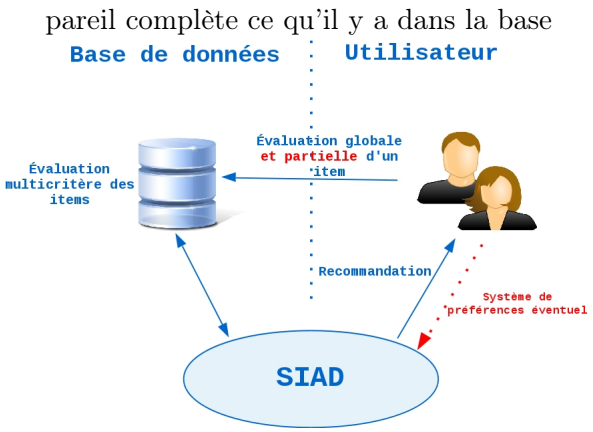

FIGURE 2.6 – Schéma global de fonctionnement du modèle MMPE

partielles des items par rapport aux critères considérés. L'utilisateur peut aussi spécifier un ordre sur l'importance relative des critères.

Considérons par exemple le site web *Circuit City*[3], site de e-commerce spécialisé dans le matériel électronique (TV, Audio, etc). Ce site utilise un système de recommandation de type MMPE qui considère les évaluations globales ainsi que les évaluations partielles sur quatre critères : "Value", "Features", "Quality", "Performance" fournis par les utilisateurs (c.f. figure 2.7) sur les produits proposés par le site. En fournissant ces évaluations (globales+partielles) le système de recommandation peut déduire, à partir de plusieurs items évalués par le même utilisateur, le système de préférence de cet utilisateur. Beaucoup d'autres sites basent leur système de recommandation sur ce type de modèle comme *Yahoo ! Cinéma*[4] où les utilisateurs évaluent les films selon quatre critères d'évaluation : "histoire", "interprétation", "réalisation", et "images". Un autre exemple est le site *Zagat* qui permet de recommander des restaurants évalués selon trois critères : "nourriture", "décor", et "service".

Nous pouvons remarquer que la différence principale entre les différents modèles réside dans la considération des critères. Ces critères peuvent correspondre aux caractéristiques d'un item (la durée d'un film, son type, par exemple), comme c'est le cas pour les modèles *MMPC* et *MMRC*, mais peuvent aussi correspondre à l'évaluation faite par l'utilisateur lui-même (le jeu des acteurs par exemple) comme c'est le cas pour le modèle *MMPE*. Dans les systèmes de recommandation multicritères le critère peut se rapporter aux dimensions de l'item (les caractéristiques) sur les-

3. http://www.circuitcity.com/
4. http://fr.cinema.yahoo.com/

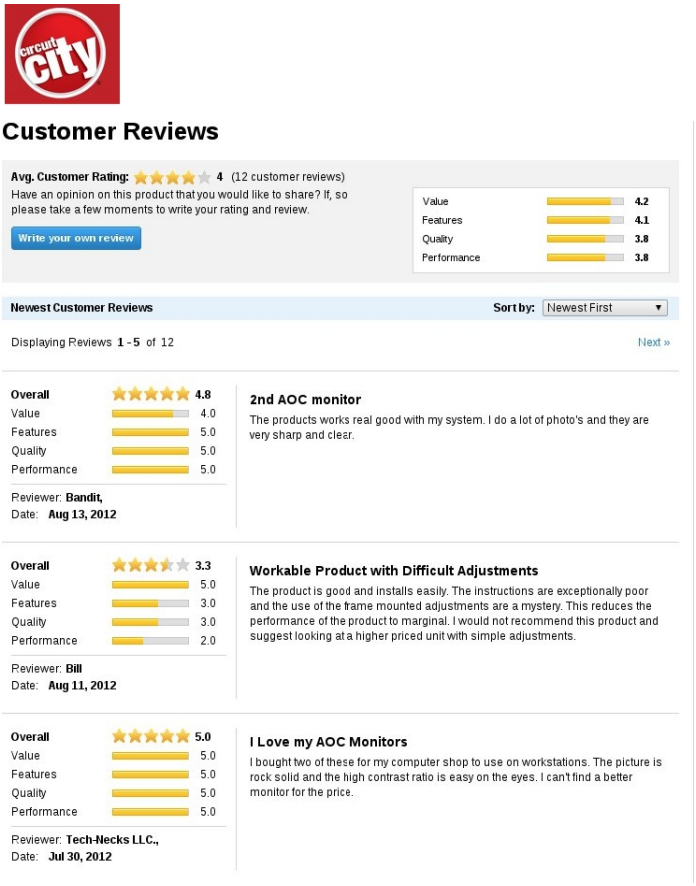

FIGURE 2.7 – Un exemple de modèle MMPE : le site *Circuit City*

quelles porte alors implicitement l'évaluation ou bien aux évaluations elles-mêmes (l'utilité des caractéristiques). Plus le SIAD intègre d'informations qui permettent de construire un modèle précis des préférences de l'utilisateur (évaluations partielles par critère, ordre sur l'importance relative des critères, etc.), plus la recommandation sera pertinente, mais en revanche plus l'expertise à fournir est importante.

D'après [Adomavicius & Kwon 2007] et l'interprétation formelle que nous en avons proposée en terme de modèle de préférence, le modèle *MMPE* est objectivement le plus performant , c'est pourquoi nous choisirons ce modèle comme référence. Cependant plus le système de recommandation s'appuie sur un modèle précis du système de préférences de l'utilisateur, plus les informations subjectives, dont l'acquisition automatique est difficile, sont nombreuses et par conséquent plus l'utilisateur est sollicité [Rekik 2007, Plantie *et al.* 2008]. Le modèle *MMPE* parait alors difficile à mettre en place sans une aide informatisée poussée pour en faciliter la construction. L'expertise à fournir pour scorer, sur une multitude de critères, les items de la base de données du système est un frein majeur car même si les sites de recommandation reposent sur un comportement collaboratif des utilisateurs, ceux-ci n'ont pas forcément l'envie, le temps ou les compétences pour se plier aux formats exigés par le site pour réaliser correctement ses analyses. D'autre part, les évaluations que fourniraient les utilisateurs, même multicritères, ne seraient pas nécessairement comparables les unes aux autres car la subjectivité impacte fortement l'évaluation. D'un utilisateur à l'autre quatre étoiles n'a pas forcément le même sens, il est courant de trouver des critiques dithyrambiques sur le confort d'un hôtel alors que le score attribué est finalement de 3* sur 5. Moyenner des scores qui ne se réfèrent même pas à la même échelle ne donne que l'illusion d'un résultat précis à l'utilisateur. Il est donc nécessaire de définir un SIAD multicritère dont le modèle permette d'évaluer automatiquement des items (produits ou services) selon des critères personnalisables par l'utilisateur, tout en gérant l'imprécision des évaluations et les divergences d'opinions exprimées par les internautes sur un même item. Nous avons montré que l'opinion (ici évaluation ou avis) repose sur des critères [Duthil *et al.* 2012a] et qu'il est nécessaire de considérer chacun d'eux pour avoir une évaluation précise et pertinente. C'est pourquoi dans un système de type *MMPE*, l'évaluation de chaque élément recommandable doit être effectuée sur tous les critères considérés. Cette contrainte est imposée par le modèle. Elle est très lourde pour l'utilisateur et il est donc essentiel de la rendre plus flexible pour faciliter l'utilisation d'un tel système de recommandation. De plus, l'évaluation automatique multicritère d'opinions doit

permettre d'évaluer de façon normalisée tous les items des différentes bases de données afin de proposer une comparaison objective de ceux-ci. Elle doit également, afin de garantir sa fiabilité, limiter l'impact de l'inexpérience de certains utilisateurs. Enfin, elle doit, à partir des textes critiques en langage naturel, homogénéiser la part de subjectif dans l'évaluation individuelle et limiter les erreurs d'interprétation ou tout au moins les harmoniser. La fusion d'avis imprécis de centaines d'internautes permet de contrôler la subjectivité de l'évaluation [Imoussaten 2011].
Les modèles de type *MMPE* ont à ce jour été largement étudiés et les outils mathématiques disponibles pour gérer le multicritère sont nombreux [Akharraz 2004, Plantie *et al.* 2008, Rekik 2007], c'est pourquoi nous ne les présenterons pas en détail dans ce manuscrit. En revanche, l'extraction des connaissances, et plus particulièrement la phase d'évaluation multicritère automatique des critiques nécessaires à la recommandation offrent encore, à ce jour, beaucoup de perspectives, notamment pour obtenir un système orienté utilisateur et personnalisable qui nécessite un niveau de supervision/expertise négligeable. Nous restreindrons cet état de l'art à l'**extraction des connaissances** à partir de données textuelles. En effet, dans la plupart des systèmes de recommandation aujourd'hui, les utilisateurs attribuent un score à un item et lui associent une critique/un avis. Cette critique est supposée expliquer en langage naturel le score attribué, le justifier par une analyse moins synthétique qu'un simple score. C'est seulement à travers le texte de la critique que l'on peut s'apercevoir que l'évaluation est vraiment multicritère. Le couple "évaluation globale - critique en langage naturel" semble néanmoins être le format le plus apprécié parce qu'il convient aussi bien au néophyte (qui n'a ainsi pas à évaluer des critères au-delà de ses compétences) qu'à l'expert (qui est libre d'argumenter sa notation avec les critères de son choix). Nous considérerons donc que la critique ou avis en langage naturel est le matériel de base des systèmes de recommandation collaboratifs et que l'objectif pour une informatisation poussée de ces outils de recommandation devrait reposer sur l'extraction de critères et la détection d'opinions automatiques à partie de ces textes critiques en langage naturel. Cette automatisation permettrait de transformer n'importe quel site basé sur un modèle de recommandation de type *MMPC* et *MMRC* en un *MMPE*.

Les problématiques d'acquisition et de traitement de l'information pour l'aide à la décision ont engendré de nombreux travaux au cours des dix dernières années, et notamment ceux de [Plantie *et al.* 2008]. Plantié propose un SIAD multicritère de type *MMPE* qui permet l'analyse automatique et la recommandation de critiques ci-

nématographiques. L'auteur propose un système de recommandation qui utilise des techniques de fouille de texte et d'aide à la décision multicritère. Son objectif est, dans un premier temps d'analyser les critiques fournies par les utilisateurs, c'est-à-dire d'extraire l'opinion qui y est exprimée sur différents critères, puis dans un second temps, d'utiliser ces opinions par critère pour recommander des items grâce à des outils mathématiques d'aide à la décision multicritère (agrégation, analyse de sensibilité, etc). Son système repose donc sur deux tâches principales : une première qui concerne l'extraction des connaissances nécessaires à l'évaluation multicritère des critiques, et une seconde qui concerne à proprement parler la recommandation des items à partir des évaluations multicritères obtenues lors de l'extraction des connaissances. Compte tenu de l'objectif de notre travail, nous ne détaillerons dans ce chapitre uniquement les approches utilisées pour l'**extraction des connaissances** (fouille de texte). Pour ce qui est des outils de recommandation basés sur l'analyse multicritère, nous intervenons en simple utilisateur et par conséquent, ils ne seront décrits précisément que dans le dernier chapitre 6 "Recommandation".
L'extraction des connaissances dans ce contexte est un processus complexe qui vise à extraire l'opinion présente dans des critiques cinématographiques. [Plantie *et al.* 2008] identifie que ce problème fait référence à deux sémantiques différentes et qu'il est nécessaire de le décomposer en deux sous-problèmes distincts : d'une part l'**extraction des critères**, et d'autre part l'**extraction d'opinion**.

- **Extraction des critères :** Ce processus a pour objectif d'identifier les parties de texte d'une critique qui parle d'un critère spécifique. Ce processus nécessite une phase d'apprentissage qui permet d'identifier le vocabulaire (mots) spécifique à chacun des critères désirés. Ce vocabulaire permet ensuite d'identifier les passages dans le texte qui parlent de chacun des critères. L'auteur a choisi d'utiliser un apprentissage supervisé basé sur un corpus d'apprentissage annoté constitué d'un ensemble de critiques cinématographiques. Chacune des critiques de la base d'apprentissage a été analysée "manuellement" selon chacun des critères considérés. Ainsi, les extraits de texte relatifs à chacun des critères sur l'ensemble des critiques de la base d'apprentissage sont connus : ils sont indexés par les critères. L'apprentissage consiste ensuite à apprendre à partir des extraits identifiés manuellement, le vocabulaire spécifique à chacun des critères considérés. L'apprentissage est ensuite confié à un classifieur de type "Naïve Bayes" (c.f. section A.4.1.1) qui utilise des méthodes statistiques. Les résultats obtenus sur les trois critères qu'il traite pour son application sur

le cinéma (*acteur, scénario, réalisation*) montrent une extraction correcte de 77% en moyenne des extraits à identifier sur les bases de test.

– **Extraction d'opinions :** Ce processus a pour objectif de déterminer l'orientation sémantique, la polarité, d'une phrase ou d'un groupe de phrases (extrait) selon quatre niveaux de distinction : très positif, positif, négatif et très négatif. Le principe est donc d'attribuer un score à chacun des extraits propres à un critère et issu de la phase précédente. Cette problématique nécessite, comme l'extraction des critères, une phase d'apprentissage préalable qui permet d'apprendre le vocabulaire d'opinion (mots) relatif à chacun des quatre degrés d'opinion. À partir de ce vocabulaire, il est alors possible de déterminer l'orientation sémantique d'une phrase ou d'un extrait de texte, en évaluant l'orientation sémantique de chacun des mots précédemment appris qui la composent. Comme précédemment, un apprentissage supervisé a été choisi par l'auteur, et un corpus d'apprentissage de critiques a été annoté selon les quatre niveaux d'opinion souhaités. Un classifieur de type "Naïve Bayes" est à nouveau utilisé pour réaliser la phase d'apprentissage des vocabulaires. Les résultats obtenus montrent une extraction d'opinion correcte, toutes classes confondues (très positif, positif, négatif, très négatif), de 75% des extraits en moyenne sur les bases de test.

En combinant le processus d'extraction de critères et d'extraction d'opinion, il devient donc possible d'extraire une opinion relative à chacun des critères considérés et de lui attribuer un score à partir de critiques en langage naturel. Les scores fournis par le processus d'extraction (critères + opinions) sont ensuite utilisés pour une évaluation globale quantitative des films. L'outil mathématique permettant l'obtention d'un classement des films est l'agrégation multicritère. L'auteur a retenu la moyenne pondérée comme opérateur d'agrégation : elle lui permet en particulier d'intégrer la notion d'importance relative d'un critère dans son analyse. Ainsi, le score global d'un film est la moyenne pondérée sur l'ensemble des scores partiels obtenus selon les critères d'évaluation. Une analyse de sensibilité de l'opérateur d'agrégation consiste à calculer la contribution moyenne de chacun des critères à l'évaluation globale des films. Les critères ayant le plus contribué sont utilisés pour la recommandation.

Le SIAD proposé par [Plantie *et al.* 2008] est une première étape vers une automatisation complète d'un système de recommandation. Cependant, la tâche manuelle de construction et d'annotation des corpus d'apprentissage (critères et opinion) nécessaire à l'extraction des critères et à la détection d'opinion pour systé-

matiser les *MMPE* semble à ce jour irréaliste si l'on veut s'attaquer aux bases du web. De plus, pour rendre le système personnalisable, il faut que l'utilisateur puisse définir ses propres critères (et non pas subir ceux qui lui sont imposés par le site), ce qui multiplie encore le nombre de critères à considérer lors de la phase manuelle d'annotation du corpus d'apprentissage. Nous pensons donc que l'expertise humaine nécessaire à la mise en place d'un tel système est un frein majeur à la généralisation des *MMPE* dans la recommandation et qu'il est nécessaire de développer de nouveaux algorithmes capables de s'affranchir de la tâche fastidieuse et rédhibitoire d'annotation des bases d'apprentissage. L'objectif devient alors de rendre la machine capable de remplacer l'homme dans cette tâche, c'est-à-dire de la rendre capable d'interpréter des textes en langage naturel et d'en détecter les différents sens en utilisant une supervision humaine minimale. Ce sont ces problématiques d'extraction des connaissances qui constituent le cœur de notre travail. Nous pouvons donc identifier deux problématiques émergentes pour lever les verrous à la généralisation des *MMPE* dans la recommandation :

- **Comment automatiser l'extraction des critères dans des textes écrits en langage naturel sachant que ceux-ci peuvent faire référence à des concepts complexes et ce en minimisant l'intervention humaine ?**
- **Comment évaluer qualitativement un avis écrit en langage naturel en minimisant l'intervention humaine ?**

La problématique d'extraction de critères pour les systèmes de recommandation est très proche de la problématique de segmentation thématique de texte puisque cette dernière consiste à identifier les changements thématiques dans le discours. La thématique peut, ici, être considérée comme l'expression d'un critère. De même, la problématique d'évaluation d'avis est proche de l'extraction d'opinions qui cherche à déterminer l'orientation sémantique, la polarité d'un texte. Ces problématiques sont directement liées à l'extraction des connaissances dans des données textuelles, et plus généralement à l'analyse du langage, qui permet d'exprimer de telles connaissances. Cette section sur la recommandation multicritère explique donc les deux sections suivantes de cet état de l'art, consacrées pour l'une à la segmentation thématique, et pour l'autre à la détection d'opinion ou l'analyse de sentiments.

2.2 L'extraction des connaissances dans le langage

L'un des challenges actuels des chercheurs dans le domaine de la fouille de données est d'analyser des données brutes dans l'objectif d'en extraire un savoir. Prenons l'exemple d'un capteur électronique qui recueille des observations, ce sont des **données**. Si nous arrivons à en extraire un signal , cela devient de l'**information**, et si nous arrivons à décoder ce signal et à lui donner du **sens**, cela devient de la **connaissance**. De tout temps, les sociétés humaines ont produit des données, des informations, des connaissances et elles les ont stockées et échangées. Aujourd'hui, ces données sont numériques et elles sont gérées par des ordinateurs et échangées sur des réseaux informatiques. Ces informations sont désormais filtrées, triées et organisées intelligemment par les ordinateurs pour, notamment, des besoins de rapidité d'accès à l'information. Les outils actuels ont déjà atteint un bon niveau de sophistication, mais lorsque l'on recherche de l'information, nous interrogeons le plus souvent des données textuelles (données web) car les gens raisonnent et s'expriment via le langage naturel. Les algorithmes actuelles les plus performants qui manipulent ce genre de données, et notamment sur le web (moteurs de recherche) se basent principalement sur la recherche de mots, spécifiés par l'utilisateur, dans des documents. L'inconvénient de ce type d'algorithme est qu'il faille consulter chacun des documents retournés par le système pour accéder à l'information souhaitée. Pour être en mesure de répondre plus précisément à la demande de l'utilisateur, il faut passer de l'**analyse de textes en langage naturel** à des bases de connaissances, c'est-à-dire d'extraire à partir d'un signal, de la **connaissance**. La grande difficulté réside justement dans l'extraction de telles connaissances par la machine : certaines tournures de phrases sont complexes et mettent en échec les logiciels, le mot peut avoir plusieurs sens, le texte peut être ambigu, le niveau de langage peut différer d'un support à l'autre, etc. De plus, ces données fournies par les internautes peuvent être fausses, volontairement ou non, de nouveaux mots peuvent apparaître, de nouveaux sens peuvent leur être rattachés. Cela revient donc à rendre la machine capable de raisonner et à décider à partir d'informations vagues. Il est donc nécessaire de rendre ces données interprétables par la machine pour mieux gérer l'information qu'elles contiennent en développant des algorithmes intelligents capables de traiter des données avec un niveau cognitif proche de celui de l'homme, c'est notamment un des challenges du web sémantique qui s'intéresse à la sémantique des données plutôt qu'à la syntaxe des données textuelles. Le passage entre des données textuelles à des données sémantiques (conceptuelles, d'opinion, etc) est à ce jour une des pré-

occupations principales des chercheurs en fouille de données. Un **concept** fait ici référence à une idée générale et abstraite commune, que se fait l'esprit humain d'un objet pensé concret ou abstrait, et qui lui permet de rattacher à ce même objet les diverses perceptions qu'il en a, et d'en organiser les connaissances. Pour être en mesure d'effectuer des tâches d'extraction conceptuelle, il est nécessaire de comprendre, de donner du sens, aux données, cela revient à étudier le langage. À partir du langage, il est nécessaire d'étudier les mots qui portent l'information, nous parlerons de **descripteur**. Des études montrent que le langage a deux fonctions principales : l'expression et la communication. Par expression, nous entendons expression d'idées et de sentiments. La communication, elle, est action : nous agissons sur l'autre au moyen du langage. La communication langagière est une communication d'idées car les mots sont porteurs d'idées. La pensée fonctionne comme la langue ; la grammaire est l'expression de la raison. Il n'y a pas de pensée sans mise en forme de la pensée ; il existe un parallélisme logico-grammatical. Le langage n'est pas utilisé dans une forme purement logique : notre pensée peut être *confuse* et le langage exprime toutes les difficultés du locuteur à dire ce qu'il a à dire, c'est pourquoi le langage est équivoque : les mots ont plusieurs sens. À chaque mot ne correspond pas un seul sens (univocité), et de nouveaux sens peuvent apparaître, c'est là toute la richesse de la langue. Considérons qu'un mot puisse exprimer un seul sens, le langage serait pauvre et toute entreprise poétique serait vouée à l'échec ! Le langage a également une fonction de symbolisation permettant de faire le lien entre un mot (**signifiant**) et une chose (**signifié**). Selon [Morris 1938], **au mot sont attachées trois dimensions** : la première est **sémantique**, elle regarde la relation entre le signe et ce qu'il signifie, la seconde dimension est **syntaxique** : elle s'intéresse aux relations entre les mots, la troisième dimension concerne le **pragmatisme** d'un mot : elle regarde la relation entre les signes et le locuteur, et lui permet notamment d'éprouver des sentiments, des émotions, mais aussi de pouvoir raisonner par rapport à son propre ressenti. La poésie accorde au langage la plus haute fonction au pragmatisme. De plus, la manière d'exprimer une idée est différente selon le niveau de langage employé : le vocabulaire et la grammaire étant spécifique à chacun des niveaux. Il existe en français une gradation descendante entre trois principaux registres de langue[5], ce qui multiplie les manières de s'exprimer. Par exemple :

- Registre soutenu : "J'ignore ce qu'il est advenu."
- Registre courant : "Je ne sais pas ce qu'il s'est passé."

5. source : http://fr.wikipedia.org/wiki/Registredelangue

– Registre familier : "J'sais pas c'qu'y a eu."

Le registre soutenu est surtout employé à l'écrit, notamment dans la lettre officielle et dans le texte littéraire. Le vocabulaire est recherché et les règles prescrites par la grammaire normative sont parfaitement respectées.
Le registre courant est employé avec un interlocuteur que l'on ne connaît pas intimement, avec lequel on a une certaine distance. Le vocabulaire est usuel et les règles prescrites par la grammaire normative sont habituellement respectées.
Le registre familier est employé avec des proches, des intimes. Le vocabulaire est relâché, il peut être abrégé. Toutes les syllabes ne sont pas nécessairement prononcées. Les règles de la grammaire normative ne sont pas systématiquement respectées.

Pour revenir à notre contexte plus spécifique d'extraction d'opinion multicritère, nous pensons qu'il est indispensable de tenir compte de chacune des trois dimensions énoncées par [Morris 1938] : le syntaxique, le sémantique et le pragmatisme, pour pouvoir prétendre à une caractérisation pertinente d'un critère ou d'une opinion. Nous proposons de voir l'interprétation d'un concept, ou d'une opinion, pour un texte donné, comme une fonction qui dépend de la syntaxe, de la sémantique, du pragmatisme et du texte tel que :

$$interpretation(texte) = f(Se, Sy, Pr, texte) \tag{2.1}$$

où Se relève de la dimension sémantique, Sy de la dimension syntaxique, Pr de la dimension pragmatique et $texte$ est le texte sur lequel l'interprétation est produite. L'interprétation est pertinente si dans le cas de l'extraction de critère, on parvient à identifier dans le texte les extraits qui évoquent le critère à l'étude (segmentation thématique), si dans le cas de l'extraction d'opinion, on parvient à déterminer l'opinion qui s'en dégage.

2.3 Segmentation thématique

La segmentation thématique de textes est généralement la première étape d'un processus de fouille de textes qui a pour objectif d'identifier les ruptures thématiques dans un document afin de le découper en extraits homogènes et ainsi accéder plus facilement et plus rapidement à l'information désirée. Ces extraits sont considérés comme des "morceaux de texte" comportant de forts liens sémantiques internes, tout en étant détachés des extraits adjacents.

Plusieurs travaux de recherche ont montré l'intérêt d'un tel processus, notam-

ment les travaux de Hearst [Hearst 1997]. Ces vingt dernières années, de nombreux algorithmes ont vu le jour. Nous distinguons deux catégories d'approches : les approches non-supervisées, et les approches supervisées.

2.3.1 Approches non-supervisées

Les approches non-supervisées sont, à ce jour, les plus répandues et se basent généralement sur des algorithmes issus du traitement du signal (images, etc). Nous identifions quatre classes de méthodes de segmentation : celles basées sur la répartition lexicale, celles basées sur la théorie des graphes, celles basées sur des algorithmes génétiques, et enfin, celles basées sur la pondération lexicale. Cette classification n'est cependant pas exhaustive, et ces méthodes peuvent être combinées.

2.3.1.1 Méthodes basées sur la répartition lexicale

De nombreux algorithmes d'identification thématique considèrent qu'un document traitant d'une thématique peut se décomposer en fragments de texte homogènes relatifs à la thématique considérée, et qu'ils adoptent tous une répartition lexicale homogène : ils utilisent un vocabulaire commun [D'hondt *et al.* 2011]. Dans la littérature plusieurs techniques ont été proposées [Morris & Hirst 1991, Hearst 1997, Galley *et al.* 2003, Sitbon & Bellot 2007, Choi 2000]. Leurs différences résident essentiellement sur la manière dont la répartition lexicale est décrite/considérée dans un document. Plusieurs méthodes existent pour mesurer cette répartition lexicale. Cette mesure est appelée *cohésion lexicale* et elle représente la similarité entre deux entités considérées (groupes de mots, phrases...). Nous pouvons identifier plusieurs définitions pour cette mesure, notamment celles basées sur la répétition des termes (champ lexical), les vecteurs de répétition de contexte (*context vectors entity repetition*), le calcul de similarité sémantique, la modélisation de distances entre les mots. La répétition de certains termes spécifiques indique généralement la présence d'une thématique commune [Hearst 1997].

2.3.1.2 Méthodes basées sur la pondération lexicale

Les méthodes basées sur la pondération lexicale sont les plus populaires [Morris & Hirst 1991, Hearst 1997, Galley *et al.* 2003, Sitbon & Bellot 2007, Choi 2000]. Les mots fréquents dans le texte sont souvent révélateurs de la thématique globale du texte. D'autre part, les mots qui sont moins fréquents et qui ont une distribution

plus uniforme sont considérés comme neutres et ne fournissent que peu d'information. Les mots fréquents et ayant une distribution non-uniforme sont dits d'intérêt pour l'algorithme *TextTiling*, nous pouvons dire qu'ils sont supposés être révélateurs de la structure thématique (sous-thématique) du texte [Hearst 1997]. L'idée est de découvrir la structure de sous-thématiques du texte en utilisant la répétition des termes en calculant un TF-IDF (c.f. annexe A), plutôt que de considérer les relations sémantiques entre les concepts. Cela se justifie par le fait que la répétition d'un seul terme peut être un indicateur très utile pour déterminer la structure de sous-thématiques comme il est démontré dans [Hearst 1997].
Ces algorithmes placent des limites à l'intérieur du texte lorsqu'un changement thématique est identifié. Ce processus d'identification utilise une fenêtre glissante, de taille fixe, pour examiner les variations lexicales locales. Cette variation se traduit généralement par une baisse de la mesure de similarité employée. Comme indiqué précédemment, de nombreux algorithmes peuvent être décrits de cette manière. Les plus célèbres sont *TextTiling* [Hearst 1997], *C99* [Choi 2000], *Dotplotting* [Reynar 2000] ou *Segmenter* [yen Kan *et al.* 1998]. Nous choisissons de détailler plus précisément les algorithmes de *TextTiling* et de *C99* qui sont à ce jour les plus utilisés.

TextTiling découpe le texte en plusieurs segments (séquences de 3 à 5 phrases en moyenne) relatifs à une thématique ou à une sous-thématique d'un domaine. Cet algorithme comporte trois étapes principales : découpage du texte en unités lexicales (tokenization), calcul du score lexical et détermination des frontières (ruptures sémantiques). La première étape consiste à diviser le texte en unités lexicales (mots). La ponctuation est supprimée et les mots sont lemmatisés. Le texte est alors composé de "pseudo-phrases" (groupe de lemmes successifs qui conservent l'ordre des mots pour rester cohérents avec la structure de la phrase complète) ayant une taille fixe contrôlée par une fenêtre. Cette technique de fenêtrage permet la comparaison des pseudo-phrases entre elles en les considérant comme des unités homogènes. La taille de la fenêtre est l'unique paramètre de l'algorithme. La seconde étape de l'algorithme consiste à calculer une mesure de similarité entre les pseudo-phrases. Cette mesure de similarité est calculée en comparant les pseudo-phrases de texte adjacentes afin de pouvoir déterminer leur similitude par rapport aux mots qu'elles ont en commun. Ainsi, si les pseudo-phrases adjacentes ont de nombreux termes en commun, ceux-ci

sont autant d'éléments de preuve qu'elles traitent bien de la même sous-thématique. Les ruptures sémantiques entre deux pseudo-phrases sont déterminées en attribuant un score à chacune d'elles. Ce score mesure l'appartenance d'une pseudo-phrase à la sous-thématique considérée. Enfin, les ruptures sémantiques sont déterminées par la localisation des valeurs minimales de cette fonction d'appartenance (endroit où la similarité est minimale) [Hearst 1997].

C99 Dans la continuité de *TexTiling*, [Choi 2000] a proposé un algorithme qui utilise les mêmes hypothèses que *TextTiling* en tenant compte du fait que de longs documents portent généralement sur plusieurs thématiques ou alors sur différentes sous-thématiques dérivées de la thématique globale du document. Le but d'une segmentation linéaire est toujours de découvrir les ruptures thématiques, mais à la différence de l'approche adoptée dans *TextTiling*, cette méthode est indépendante du domaine considéré. Cet algorithme de segmentation, appelé *C99*, comporte deux éléments clés : une mesure de similarité et une stratégie de regroupement. L'algorithme prend en entrée des phrases, les lemmatise, puis construit un dictionnaire contenant la fréquence de chacun des lemmes. Une phrase est vue comme vecteur de lemmes dont les composantes sont les fréquences des lemmes. La similitude entre chaque paire de phrases est calculée en utilisant une mesure basée sur le cosinus. À partir de ces mesures, une matrice de similarité est construite. En raison de la petite taille des segments de texte, cette valeur de similarité peut alors être non fiable. Un ordre de similarité entre les phrases est alors estimé au moyen d'une matrice de classement. La dernière étape est une étape de clustering qui permet de déterminer les ruptures sémantiques entre les différents segments (un segment de texte étant composé de deux phrases).

2.3.1.3 Méthodes basées sur un algorithme génétique

D'après Lamprier [Lamprier *et al.* 2007] les méthodes séquentielles présentées précédemment nécessitent d'introduire la notion de fenêtre dans laquelle des mesures de cohésion sont utilisées. Néanmoins, il est difficile de déterminer la taille de cette fenêtre : une fenêtre trop courte peut entraîner l'algorithme à ne pas considérer certains indices de cohésion et une trop grande fenêtre peut tenir compte des répétitions de termes alors qu'ils ne devraient pas être associés s'ils sont utilisés dans des contextes différents. À partir de cette remarque, Lamprier propose une approche où les frontières sont considérées globalement. Basée sur une évaluation des diffé-

rentes possibilités de segmentation potentielle, la méthode *SegGen* permet d'avoir une vision complète du texte et ainsi d'être en mesure de calculer la cohérence entre les phrases sans avoir recours à un système de fenêtrage. Le problème de segmentation est alors considéré comme un problème d'optimisation combinatoire. *SegGen* évalue les segmentations potentielles de l'ensemble du texte : il recherche celles qui offrent la meilleure similitude dans les segments et la dissemblance maximale entre segments plutôt que de fixer des limites incrémentales. Le manque de connaissances sur la structure du texte ou sur le nombre de segments potentiels induit un espace de recherche conséquent et amène donc à considérer l'utilisation d'un algorithme génétique pour faire face à cette complexité.

2.3.1.4 Méthodes basées sur la théorie des graphes

D'autres approches statistiques utilisent l'information globale du texte. Malioutov [Malioutov & Barzilay 2006] présente un outil basé sur la théorie des graphes (graph-theoretic). Le texte est converti en un graphe non orienté pondéré dans lequel les nœuds représentent les phrases et les arcs les relations quantifiées entre les phrases (fonctions de la similarité entre les phrases). La segmentation de texte est effectuée en appliquant le critère de coupure normalisée (normalized-cut) [Shi & Malik 2000]. En utilisant ce critère, la similitude au sein de chaque partition est maximisée et la dissemblance dans les partitions est minimisée. L'approche à base de graphes étend la gamme de cohésion locale de la fenêtre glissante en tenant compte de la cohésion lexicale sur le texte complet et la distribution dans un texte. Les techniques de calcul qui permettent de trouver la solution optimale pour la coupure minimale sont toutefois complexes. La minimisation de la coupure normalisée est un problème NP-complet, mais, en raison de la contrainte de linéarité sur ce type de segmentation, l'obtention d'une solution exacte est possible [Malioutov & Barzilay 2006]. [D'hondt *et al.* 2011] propose de représenter le texte sous forme d'un graphe de cohérence. Ce graphe, basé sur les chaînes lexicales, représente la cohésion locale du document, et est le principal outil dans le processus d'identification et de segmentation proposé. À partir du graphe construit, une représentation dite "matrice laplacienne" est construite. Cette matrice représente mathématiquement les relations entre les nœuds du graphe. En utilisant les propriétés spectrales du document, la méthode détermine les extraits de textes cohérents.

2.3.2 Approches supervisées

Le modèle de langage thématique unigramme (topic unigram language model) est la technique la plus fréquemment utilisée [Ponte & Croft 1998]. Sur une base d'apprentissage étiquetée par la thématique, l'idée est d'extraire un ensemble de mots-clés propres à la thématique à l'aide d'un classifieur de type modèle de Markov caché [Bigi *et al.* 2000]. Les textes de la base d'apprentissage associés à la thématique sont représentés par des vecteurs de mots-clés dont les composantes sont les fréquences d'occurrence (TF-IDF). La thématique est alors caractérisée par un vecteur de référence. La similitude entre une thématique et un document représenté dans le modèle de l'espace vectoriel est calculée par la mesure de similarité cosinus. La similarité la plus élevée indique le sujet de ce document.

Les approches de segmentation s'appuient souvent sur l'analyse des caractéristiques statistiques du texte. Une autre catégorie de techniques est basée sur le traitement du langage naturel. Les méthodes linguistiques s'appuient sur un ensemble de règles spécifiques sur la base de corpus d'apprentissage et sur l'utilisation d'informations sémantiques tels que les thésaurus et les ontologies, éventuellement combinés à une ou plusieurs méthodes statistiques [Moens & De Busser 2001]. Au lieu de faire de l'apprentissage sur les seuls mots, ces techniques s'intéressent à des séquences de mots contraintes par des règles de syntaxe du langage. Le principal inconvénient de ce type de techniques d'identification est que les résultats sont tributaires des ressources sémantiques disponibles pour un texte spécifique [Utiyama & Isahara 2001]. Les modèles de Markov cachés et les réseaux de neurones sont souvent utilisés dans le cadre du processus d'apprentissage. [Amini *et al.* 2000] propose un cadre probabiliste pour apprendre les séquences ou motifs de mots caractéristiques de la thématique. Caillet propose une technique d'apprentissage également guidée par des règles syntaxiques du langage pour regrouper les termes caractéristiques de la thématique (term clustering) [Caillet *et al.* 2004]. Dans un premier temps, il identifie les différentes instances des concepts présents dans le texte ; ces concepts sont ensuite définis comme des ensembles de termes représentatifs ; enfin, le partitionnement en paragraphes cohérents est réalisé avec une méthode de regroupement basé sur le critère de maximum de vraisemblance [Jain *et al.* 1999].

Notons que les approches de segmentation présentent, selon nous, une faiblesse majeure : elles ne sont pas en mesure d'identifier la thématique précise d'un extrait sans une expertise conséquente comme l'utilisation de thésaurus ou d'ontologie par exemple. Pour résoudre cette problématique d'étiquetage, des techniques issues du

résumé thématique de texte permettent d'identifier les parties d'un document en fonction de la thématique dominante [Chuang & Yang 2000]. D'autres méthodes visent à identifier des extraits relatifs au titre [Kupiec *et al.* 1995]. La majorité des techniques de résumé automatique s'appuie sur des méthodes d'apprentissage supervisé qui nécessitent une intervention humaine conséquente pour la constitution d'un corpus d'apprentissage.

2.4 Extraction d'opinion

L'extraction d'opinion plus connue sous le nom d'opinion mining (fouille d'opinion en français) est en passe de devenir une véritable industrie, tout aussi stratégique que celle des sondages, jusqu'ici seuls à fournir ce suivi des mouvements d'opinion. De plus, la puissance de calcul de l'informatique permettrait de suivre toutes ces évolutions en temps réel quel que soit leur volume sur le web. Et mieux encore, les capacités de traitement linguistique permettraient de détecter les tonalités de tous les verbatims recueillis, grâce aux méthodes de "sentiment analysis" [Boullier & Lohard 2012]. Ce nouvel eldorado de la communication spontanée et directe grâce aux techniques du web 2.0 suscite de nombreuses attentes. Pour les spécialistes des sondages, l'absence d'identification des émetteurs des opinions recueillies sur le web et sur les réseaux sociaux est rédhibitoire puisqu'elle ne permet pas d'étendre les propos recueillis à une population de référence, classifiée selon ses attributs socio-démographiques. Toute idée de représentativité devrait donc être abandonnée. Cependant, les méthodes numériques de l'opinion mining compensent cette approximation par le volume de données recueillies et par la capacité à traiter automatiquement une grande partie de ces corpus en multipliant les itérations ce qui peut permettre à minima de dégager des tendances. L'opinion mining est aujourd'hui utilisé à des fins très différentes et sur des corpus très variés. Par exemple, il est devenu essentiel pour suivre la réputation d'une marque, les impacts d'un placement de produit, pour récupérer des avis de consommateurs, comparer des produits ou services dans le e-commerce à des fins de recommandation [Bai 2011, Pang & Lee 2002, He & Zhou 2010], pour manager une communauté, etc. Il est encore utilisé dans des contextes de veille stratégique, d'intelligence économique, etc.

2.4.1 Détection d'opinion et traitement de l'information

L'objectif premier des nombreuses techniques d'opinion mining proposées est d'extraire l'opinion exprimée dans des critiques (textes d'opinion). Cette phase d'extraction est généralement une tâche préliminaire nécessaire aux systèmes de recommandation notamment. Les premiers travaux sur l'extraction d'opinion remontent à la fin des années 1990 [Argamon *et al.* 1998, Kessler *et al.* 1997, Spertus 1997], mais c'est seulement dans le début des années 2000 que cette problématique est introduite dans le traitement de l'information [Chaovalit & Zhou 2005, Dimitrova *et al.* 2002, Durbin *et al.* 2003]. Jusqu'au début des années 2000, les deux principales approches de la détection d'opinion étaient basées sur des techniques d'apprentissage machine et sur des techniques d'analyse sémantique. Par la suite, des techniques, plus fines, intégrant le traitement du langage naturel ont été largement utilisées, notamment dans la détection d'opinion dans des documents. La fouille d'opinion est à ce jour une discipline au carrefour du traitement du langage naturel et de la recherche d'informations, et en tant que telle, elle partage un certain nombre de techniques propres à chacune des deux disciplines, comme l'extraction d'informations et la fouille de textes.

Nous distinguons deux types de classification d'opinion :

- la classification binaire des documents en apposant une étiquette positif ou négatif sur chacun d'eux (polarité simple).
- la classification multiclasse qui offre plusieurs degrés d'opinion exprimés sur une échelle de valeur généralement finie et discrète (fortement positif, positif, neutre, négatif, fortement négatif).

Dans un contexte de e-commerce, on se rend compte que chaque produit ou service offre la plupart du temps plusieurs fonctionnalités ou avantages dont seulement une partie d'entre elles sont susceptibles d'intéresser le consommateur [Morinaga *et al.* 2002, Taboada *et al.* 2006]. L'opinion mining a pour objet dans ce cas de détecter une opinion personnalisée, en fonction des besoins effectifs de l'utilisateur. Cette personnalisation permet une détection d'opinion affinée, très appréciée dans les systèmes d'analyse et de comparaison des opinions de consommateurs sur des produits concurrents. C'est ce que propose [Liu *et al.* 2005] avec un système prototype appelé *Opinion Observer*. Il se décompose en deux étapes : la première est l'identification des caractéristiques du produit fondée sur des techniques de traitement automatique du langage et l'extraction de motifs, ces caractéristiques constituent alors la base des critères de comparaison ; la seconde étape consiste à émettre une

opinion binaire (positif/négatif) sur les produits ou services comparés relativement à chacune des caractéristiques identifiées au préalable.

Dans [Jin *et al.* 2009], les auteurs proposent le système *OpinionMiner* dont l'objectif est d'extraire des propriétés des produits, *i.e.* des fonctions spécifiques liées aux produits et les opinions qui y sont associées.

Une autre application de l'opinion-mining est le résumé d'article d'opinion [Ku *et al.* 2005, Beineke *et al.* 2004] qui a pour objectif d'analyser les tendances des utilisateurs [Bai 2011], de détecter des produits phares, d'effectuer des retours clients, etc. Ces techniques de résumé visent à identifier la polarité des articles (positif/négatif) et les événements corrélés susceptibles d'en expliquer la source. [Hu & Liu 2004] proposent une technique d'analyse de commentaires d'internautes relative à un produit dont le principe se résume en trois points :

1. Identifier les caractéristiques du produit où les clients ont exprimé leur opinion.
2. Pour chaque caractéristique, identifier les phrases ou avis émis (positifs ou négatifs).
3. Produire un résumé en utilisant les informations découvertes.

[Cardie *et al.* 2003, Clarke & Terra 2003, Gopal *et al.* 2011] proposent une détection d'opinion argumentée (Opinion reason mining) en appuyant leur jugement sur l'extraction des phrases (ou extrait) qui leur ont permis d'émettre ces jugements.

Dans un contexte politique, [Thomas *et al.* 2006] tente de déterminer, à partir de la transcription des débats du Congrès des États-Unis, si les discours représentent un appui ou une opposition à la législation proposée. [Mullen & Malouf 2006] décrivent une méthode statistique d'analyse de sentiment politique sur des discussions de groupes politiques pour déterminer s'ils sont en opposition avec le message original.

La plupart des travaux sur l'analyse d'opinions a surtout mis l'accent sur une classification binaire des documents, mais il est souvent utile d'avoir plus qu'une information binaire, surtout quand l'objectif final est un système de recommandation ou de confrontation d'opinions où la différence entre les items évalués est nuancée [Xu *et al.* 2011].

La phase d'apprentissage des classifieurs multiclasses est évidemment plus conséquente. La classification de documents selon le degré d'opinion implique une lourde construction manuelle ou semi-manuelle d'un lexique de mots porteurs d'opinion [Hatzivassiloglou & McKeown 1997, Lin 1998, Pereira *et al.* 1993].

Les techniques d'opinion mining binaire ou multiclasse reposent sur l'utilisation d'un corpus de mots porteurs d'opinion. Deux méthodes d'annotation automatique

de mots porteurs d'opinion se distinguent : les approches basées sur un corpus d'apprentissage et les approches basées sur l'utilisation d'un dictionnaire. A un texte où la densité et la polarité de ces mots sont fortes, une opinion pourra être affectée.

2.4.2 Techniques pour l'opinion mining

2.4.2.1 Basées sur un corpus d'apprentissage

La sélection de mots réputés être porteurs d'opinion peut reposer sur un ensemble présélectionné de mots germes fournis par un expert et à partir desquels il s'agit d'apprendre de proche en proche les mots que l'on retrouve systématiquement proches des mots germes initiaux [Duthil *et al.* 2012b], ou encore sur des heuristiques linguistiques (adjectifs et leurs modulateurs) (par exemple, [Lin 1998, Pereira *et al.* 1993]). D'autres techniques non supervisées introduisent la notion d'apprentissage des termes d'opinion à partir de mots germes pour constituer leur propre dictionnaire d'opinion [Harb *et al.* 2008, Duthil *et al.* 2012b, Duthil *et al.* 2012a, Oelke *et al.* 2009] de façon automatique. Certaines études ont montré que restreindre la classification aux seuls adjectifs améliore les performances [Andreevskaia & Bergler 2006, Turney & Littman 2002, Wiebe & Riloff 2005]. Outre le fait que cela se vérifie pour certains cas particuliers, la plupart des méthodes existantes admettent que cette restriction à la seule analyse des adjectifs comme mots porteurs d'opinion est insuffisante, et qu'il est nécessaire de considérer également les adverbes, quelques noms et les verbes car ils sont porteurs d'une orientation sémantique [Andreevskaia & Bergler 2006, Esuli & Sebastiani 2005]. Les méthodes basées sur un corpus annoté s'appuient généralement sur une analyse syntaxique et de co-occurrence des mots [Hatzivassiloglou & Wiebe 2000, Turney & Littman 2002, Hong & Hatzivassiloglou 2003].

2.4.2.2 Basées sur un dictionnaire

D'autres méthodes utilisent un dictionnaire qui confère un degré de polarité à chacune de ses entrées (par exemple *WordNet*[6]), et s'appuient sur les informations fournies par le dictionnaire pour obtenir l'orientation sémantique d'un mot puis d'un texte [Xu *et al.* 2011, Kim & Hovy 2004]. [Kamps *et al.* 2004]s'intéresse à la distance entre les mots dont la polarité est à évaluer et les mots porteurs d'opinion comme *bad* et *good*. Ces techniques utilisent pour la plupart des approches supervisées [Yi

6. http://wordnet.princeton.edu/

et al. 2003, Oelke *et al.* 2009, Duthil *et al.* 2011b] : les mots d'opinion sont soit fournis par des dictionnaires : *WordNet*, *General Inquirer*, *SenticNet*, *Dictionary of Affect of Language* (DAL), soit par des sélections manuelles.

2.5 Conclusion

Les approches présentées pour l'extraction des connaissances (segmentation et détection d'opinion) supposent, pour la majorité d'entre elles, une étape d'apprentissage nécessitant qu'un expert spécifie, dans un grand volume de documents, toutes les phrases correspondant aux entités. Nous pensons que cette intervention experte, qui réclame beaucoup de temps et de compétences, est une contrainte majeure qui nuit à une diffusion plus large des techniques d'opinion-mining.

Les méthodes de segmentation de textes et d'extraction d'opinion présentées utilisent des algorithmes et des techniques communes, issues de la **fouille de données**. Nous n'avons pas décrit chacun de ces points techniques dans ce chapitre afin de ne pas nuire à la compréhension globale de notre approche. Pour plus de détails le lecteur est invité à se reporter à l'annexe A.

Les approches souffrent donc majoritairement d'une intervention experte trop importante. À cet effet, Harb a proposé une méthode automatique d'apprentissage qui ne nécessite pas d'expertise particulière pour la construction du corpus d'apprentissage. Cependant, la phase d'extraction d'opinion qu'il suggère, qui cherche à affecter un degré d'opinion à un texte, reste selon nous très supervisée, puisqu'elle nécessite la connaissance d'un ensemble de règles conséquent entre les mots porteurs d'opinion et leurs modulateurs ; en outre cet ensemble de règles est spécifique à une langue.

Les systèmes de recommandation sont à ce jour des outils indispensables aux utilisateurs pour les aider à trouver rapidement les informations, produits et services utiles ainsi qu'aux entreprises pour mieux répondre aux attentes de leurs clients. Nous avons mis en évidence que les SIAD multicritères sont les plus pertinents, mais qu'ils nécessitent encore aujourd'hui une expertise très conséquente pour garantir une efficacité suffisante. L'un des challenges actuels dans ce domaine est l'automatisation de ces tâches cognitives complexes, notamment sur l'étape d'évaluation multicritère des items. Les techniques actuelles de fouille de textes permettent de s'affranchir d'une partie de cette expertise, mais nous estimons qu'elle est encore

insuffisante surtout dans un contexte web où la personnalisation de la recommandation doit garantir sa crédibilité. De plus, ces techniques ne sont généralement pas évolutives puisque les items sont évalués sur un jeu de critères bien déterminés. Dans le cas où de nouveaux critères viennent à apparaître, une nouvelle expertise conséquente est indispensable pour relancer l'apprentissage sur ces nouveaux critères pour pouvoir réévaluer les items. Il est donc nécessaire de développer de nouveaux algorithmes d'extraction des connaissances adaptés à ce contexte. Nous proposons des méthodes qui permettent de résoudre les problèmes d'extraction de critères et d'extraction d'opinion précédemment évoqués (chapitres 3 et 4). Ces dernières répondent aux problématiques évoquées en section 2.1 : elles permettent d'extraire les opinions contenues dans un texte en langage naturel selon plusieurs critères avec une expertise minimale, indépendamment de la langue, ce qui offre des perspectives nouvelles pour la recommandation personnalisée sur le web.

Chapitre 3

Extraction de critères

"Lire, c'est déjà poser l'équation de l'incertitude."

Jean-Michel Wyl

Sommaire

3.1 Introduction

Comme nous l'avons vu au chapitre 2 dans notre problématique de recommandation multicritère automatisée à partir de textes en langage naturel est un problème complexe, et demeure une des tâches préliminaires indispensables à la mise en place d'un SIAD multicritère. L'extraction automatique de critères est une extension de la segmentation thématique de textes qui permet d'identifier les extraits relatifs à un critère. Les approches existantes n'allient pas, à ce jour, la segmentation thématique et l'identification thématique des extraits : d'une part, les méthodes de segmentation thématique non-supervisées ne sont pas en mesure d'identifier les différentes thématiques exprimées dans chacun des extraits, et d'autre part, les méthodes basées sur un apprentissage supervisé permettent la segmentation de texte à partir des concepts identifiés, mais utilisent une ressource sémantique extérieure comme une ontologie ou un thésaurus par exemple. Nous pensons que l'expertise nécessaire à l'apprentissage et à la construction de ressources sémantiques est un frein majeur à la diffusion et à la flexibilité de ces approches supervisées. Nous proposons dans ce chapitre une méthode qui permet de s'affranchir de cette supervision trop contraignante, et nous proposons un compromis entre une méthode entièrement non-supervisée et une méthode supervisée. Ainsi notre approche utilise des principes issus du traitement du signal, ainsi que des techniques linguistiques. Cependant, étant donné que l'expertise nécessaire à l'implémentation de notre méthode est minime par rapport aux approches supervisées, nous considérons notre approche comme "très peu supervisée".

Dans la section 3.2 nous présentons notre approche de segmentation thématique appelée *Synopsis*. La section 3.3 est consacrée aux expérimentations montrant la pertinence de l'approche. Enfin, nous discutons de l'apport et des limites de l'approche, tout en donnant un certain nombre de perspectives en section 3.4

3.2 Segmentation thématique

La segmentation thématique consiste à identifier les changements thématiques dans le discours. Les approches existantes ne permettent pas d'identifier les différentes thématiques présentes dans le texte, mais seulement de détecter les changements de sujet. C'est pourquoi nous proposons de raffiner la segmentation en identifiant les passages, ou extraits, du texte qui traitent de thématiques particulières, c'est-à-dire d'identifier les passages qui parlent de la thématique intéressante pour

l'utilisateur. Une thématique est ici considérée comme l'expression d'un **concept** (c.f. section 2.2). Nous avons montré dans la section 2.2 qu'il existe un vocabulaire défini par un ensemble de descripteurs spécifiques à chaque concept, c'est-à-dire l'ensemble des termes qui constituent le champ lexical associé au concept (ex : au concept du "bonheur" peuvent être rattachés les mots "joie", "extase", "rire", etc.). Ce sont ces descripteurs qui sont utilisés dans le langage pour évoquer un concept, c'est pourquoi il est nécessaire d'apprendre un tel vocabulaire pour être en mesure d'extraire les passages relatifs au concept considéré. Autrement dit, les passages d'un texte relatifs à un concept ne contiennent pas nécessairement ce concept, mais ils l'évoquent à travers le lexique qui lui est associé. Considérons l'exemple suivant : "Lorsqu'Iseult rejoint Tristan dans la mort, un pied de vigne et un pied de lierre poussent enlacés sur les lieux du drame". Le concept "Amour" n'est pas présent dans cette phrase et pourtant chacun sait que le lierre symbolise un attachement indéfectible et est le symbole de l'amour fidèle. Un bon outil de segmentation devrait donc restituer cette phrase à l'utilisateur qui a sélectionné le concept "Amour" dans ses critères. Les techniques d'apprentissage actuelles telles qu'elles sont décrites en section 2.3 sont selon nous trop contraignantes à mettre en place dans notre contexte de segmentation et d'identification thématique, car elles nécessitent une expertise importante lors de la construction du corpus d'apprentissage notamment. En effet, notre approche de segmentation est multi concepts, c'est-à-dire qu'elle doit être en mesure d'identifier autant de concepts que l'utilisateur le souhaite, ce qui veut dire qu'il serait nécessaire de construire manuellement autant de corpus d'apprentissage qu'il y a de concepts dans une langue pour satisfaire chacun ! C'est pourquoi nous avons développé une méthode d'apprentissage qui permet de s'affranchir de cette phase de construction manuelle de corpus. L'apprentissage devient alors plus complexe, puisque les algorithmes classiques d'apprentissage ne sont plus applicables lorsqu'il s'agit de les utiliser sur des corpus non étiquetés. C'est pourquoi nous nous sommes intéressés au **langage** et aux mécanismes que l'Homme mettaient en place pour apprendre une langue : nous proposons d'argumenter nos propositions par des analogies avec l'**apprentissage** chez l'enfant.

3.2.1 L'apprentissage du langage

Les problématiques liées au langage telles qu'elles sont exprimées par [Morris 1938] en section 2.2 compliquent notablement l'apprentissage de la langue. Pour en identifier les principes, il convient d'identifier les moyens mis en œuvre pour

effectuer un tel apprentissage. Pour cela, nous nous intéressons au cas réel de l'apprentissage de la langue chez l'enfant.
L'apprentissage du langage par l'enfant se déroule en parallèle avec le développement de nombreuses autres aptitudes cognitives et notamment de l'intelligence symbolique[1]. Nous pouvons distinguer **trois phases principales**. Lors de la première phase, l'enfant acquiert la capacité à **comprendre certains phénomènes** qui l'entourent indépendamment de leur signification. Dans un second temps, l'enfant **fait le lien entre les phénomènes qu'il a observés et la langue**. Les sens utilisés pour signifier le mot ont beaucoup d'importance. Plus un enfant est confronté à des phénomènes variés, plus grande est sa capacité à définir un mot : la **dimension pragmatique** que l'enfant associe à un mot étant construite à partir de chacune des situations rencontrées au cours de son apprentissage. C'est la répétition de ces situations qui renforce l'apprentissage et donc la capacité à les interpréter par la suite. C'est notamment la symbolique qu'il a associée à un mot qui détermine sa capacité à ressentir les choses (émotions, sentiments, colère, etc), c'est ici la dernière phase de l'apprentissage. On retrouve ces principes dans la sémiotique de Piaget sur laquelle nous reviendrons dans le chapitre suivant : l'enfant observe un événement et ses répétitions : les voitures s'arrêtent au feu rouge (c'est la dimension syntaxique chez Morris [Morris 1938]) ; il remarque ensuite que, dualement les voitures passent au vert, l'opposition rouge-vert lui fait découvrir les règles de la circulation (c'est la dimension sémantique chez Morris) ; lorsqu'il s'intéressera à l'orange ou à l'orange clignotant, la pratique lui permettra de comprendre que freiner et accélérer ne sont que des généralisations de l'arrêt et du démarrage (c'est la dimension pragmatique chez Morris).

3.2.2 Proposition et hypothèse

Dans notre contexte de traitement automatique de la langue où l'on cherche à faire un apprentissage automatique d'un vocabulaire caractéristique d'un concept, nous pensons qu'il est nécessaire d'adopter un processus d'apprentissage similaire à celui présenté en section 3.2.1 et ainsi de pouvoir apprendre un vocabulaire pertinent tenant compte des trois dimensions énoncées par [Morris 1938] (c.f. chapitre 2 section 2.2).
Les techniques existantes basées sur l'"analyse lexicale" s'intéressent au mot dans son contexte (une phrase par exemple) et en étudient son comportement vis-à-vis

1. Construction de la pensée.

des mots de son entourage (position, influence, etc). L'"analyse sémantique" quant à elle s'intéresse au mot, et plus particulièrement à sa construction : à ses *sèmes*[2]. Une analyse purement sémantique s'intéressant à un mot isolé dans un contexte unique ne serait pas assez riche pour dégager l'intégralité des sens potentiels portés par ce mot. Il est alors inenvisageable de prétendre caractériser un concept, un sentiment, une opinion en considérant seulement la sémantique des mots. C'est pourquoi il est nécessaire de considérer, non seulement, la sémantique du mot, mais aussi sa syntaxe et sa "dimension pragmatique". Cette dernière étant propre à chaque locuteur, il est nécessaire d'effectuer un apprentissage sur des supports variés en étudiant la manière dont s'expriment les locuteurs sur un sujet commun. Pour exprimer le même concept, nous pouvons utiliser différents niveaux de langue correspondant aux différentes manières de s'exprimer.

L'approche présentée ici considère les relations sémantiques entre les mots (signifiés) en analysant leurs propriétés statistiques (entre signifiés). Notre approche ne se base pas sur des règles, mais sur la fréquence d'apparition de deux mots dans un même contexte, et donc à la probabilité qu'ils aient une relation directe évidente d'ordre sémantique ou syntaxique. De plus nous proposons d'identifier les différentes manières/modes d'expression d'un même concept propre à chaque locuteur en s'intéressant à différents supports d'expression (niveaux de langue). L'objectif est de pouvoir apprendre un vocabulaire relatif à un concept et ainsi être en mesure d'identifier les parties de texte relatives au concept considéré (c.f. section 2.2).

De plus, nous considérons qu'il existe un champ lexical spécifique à chaque concept à caractériser et que des termes génériques communs émergent entre les différents locuteurs parlant d'un même concept. Ces mots génériques seront par la suite appelés *mots germes* [Duthil *et al.* 2011b]. L'idée est que l'entendement d'un contexte est propre à chacun, selon ses connaissances, son niveau d'expertise, mais qu'en revanche il existe un ensemble minimal et objectif de descripteurs pour ce concept accepté de chacun. À partir de ces quelques mots fournis par le consensus, il est alors possible d'amorcer un apprentissage du champ lexical associé au concept. Cette expertise minimale est comparable au rôle des parents lors de l'apprentissage de la langue chez l'enfant, et plus particulièrement sur la manière de guider l'enfant sur le sens des mots. C'est ici l'idée de notre méthode d'apprentissage.

Dans un premier temps, l'expert analyse le contexte du concept à caractériser et propose une description succincte (c.f. section 3.2.4). Dans un second temps, le système

2. Un sème est la plus petite unité de sens.

construit automatiquement un corpus d'apprentissage (c.f. section 3.2.5). Ensuite, un processus d'apprentissage permet d'apprendre les descripteurs du concept en étudiant les relations entre la description de l'expert et le langage (c.f. section 3.2.6), tout en étudiant les différents modes d'expression (niveaux de langue) potentiels (c.f. section 3.2.7).

3.2.3 Présentation de l'approche *Synopsis*

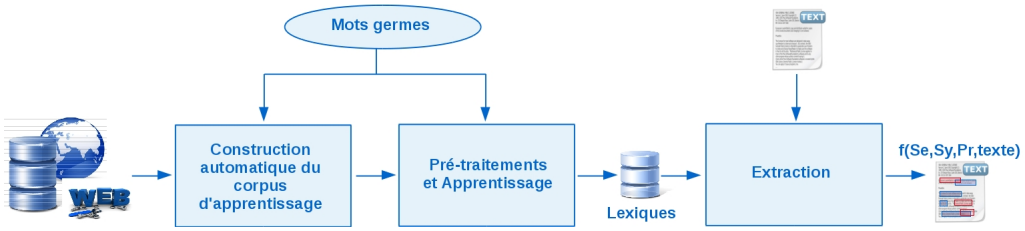

FIGURE 3.1 – Approche *Synopsis*

À partir des propositions et des hypothèses émises dans la section 3.2.2 nous proposons de décomposer notre approche en quatre étapes successives (c.f. figure 3.1) où chacune d'elles répond à un problème spécifique :

1. La première étape consiste à caractériser le concept à identifier en définissant un ensemble de termes sémantiquement proches de ce concept : les *mots germes* (c.f. section 3.2.4).
2. La seconde étape consiste à construire automatiquement un corpus d'apprentissage adapté à notre méthode d'apprentissage sur la base de ces mots germes (c.f. section 3.2.5).
3. La troisième étape consiste à apprendre le vocabulaire caractéristique du concept à identifier. Cette étape utilise le corpus précédemment construit et fournit un lexique de mots scorés selon leur appartenance au champ lexical du concept (fonction de la fréquence du mot candidat) (c.f. section 3.2.6).
4. La quatrième étape consiste à extraire les segments de textes relatifs au concept en utilisant le lexique précédemment construit (c.f. section 3.2.7).

3.2.4 Caractérisation d'un concept

Notre approche est basée sur l'étude de la répartition des termes autour des mots germes. La figure 3.2 montre comment sont répartis les k mots germes $G^q =$

$\{g_j^q, j = 1 \dots k_q\}$ autour d'un concept $C^q \in \{C^1 \dots C^p\}$ donné. Sur cet exemple, nous considérons sept mots germes caractéristiques du concept (c.f. exemple 1). A chaque germe est associé un champ lexical spécifique avec de possibles recouvrements avec les autres mots germes caractéristiques du concept dans un domaine donné. L'union des champs lexicaux des mots germes doit couvrir autant que possible celui du concept dans le domaine d'intérêt.

Exemple 1 *Reprenons notre exemple d'application dans le domaine du cinéma. L'utilisateur est intéressé par des informations concernant principalement le scénario ou les acteurs. Ainsi, nous pouvons définir les deux concepts suivants :* $C^1 = \{$"*actor*"$\}$ *et* $C^2 = \{$"*scenario*"$\}$*. L'utilisateur peut alors préciser pour chacun des concepts un ensemble de mots qui lui semblent caractéristiques. Par exemple, pour "actor", les mots "acting", "actor", "casting", "character", "interpretation", "role" and "star". De la même manière pour "scenario", les mots germes caractéristiques sont : "adaptation", "narrative", "original", "scenario", "scriptwriter", "story" et "synopsis".*

Chacun des mots germes sera utilisé pour apprendre un ensemble de mots, appelés descripteurs, qui sont sémantiquement proches du concept C^q. Cet ensemble noté $\mathcal{E}_j^q$ est représenté par une ellipse autour du j$^{\text{ième}}$ germe g_j^q du concept q. Le choix du nombre de mots germes n'est pas simple car il peut dépendre de la manière dont un concept apparaît dans les documents. Si l'occurrence de celui-ci est trop faible, cela implique que l'apprentissage des mots nécessitera de nombreux appels au moteur de recherche afin d'acquérir de nouveaux documents pour apprendre de nouveaux mots. Ceci est d'une part pénalisant en temps de réponse, mais surtout n'est techniquement pas simple à mettre en œuvre, en pratique, si trop de requêtes sont exécutées sur un moteur de recherche, ce dernier risque fort de bloquer la communication. Inversement, il est difficile pour l'utilisateur de spécifier un trop grand nombre de mots germes. Dans un contexte différent du nôtre, [Turney & Littman 2002] proposent d'utiliser 7 mots germes. Les expérimentations que nous avons menées ont confirmé que l'utilisation de 7 mots germes était également bien adaptée à notre contexte. En pratique, nous avons montré qu'il faut au minimum 4 mots germes pour avoir une caractérisation suffisante d'un concept C^q.

L'ensemble des mots germes de C^qe st noté G^q, l'union $\bigcup \mathcal{E}_j^q$ des ensembles $\mathcal{E}_j^q$ pour un concept C^q donné, est un ensemble de mots proches du concept C^q à caractériser. Cependant ce n'est pas suffisant pour déterminer la *frontière* entre l'ensemble

des mots caractéristiques du concept et l'ensemble des mots non-caractéristiques car il se peut que certains de ces descripteurs soient aussi caractéristiques de $\mathcal{D} \setminus X^q$. Pour pouvoir déterminer cette *frontière* nous étudions la répartition des mots dans l'ensemble $\bigcup \mathcal{E}_j^q$ appelé *classe* (X^q) du concept C^q et la répartition des mots dans l'ensemble $\mathcal{D} \setminus X^q$ appelé *anti-classe* ($\overline{X^q}$) du concept C^q.

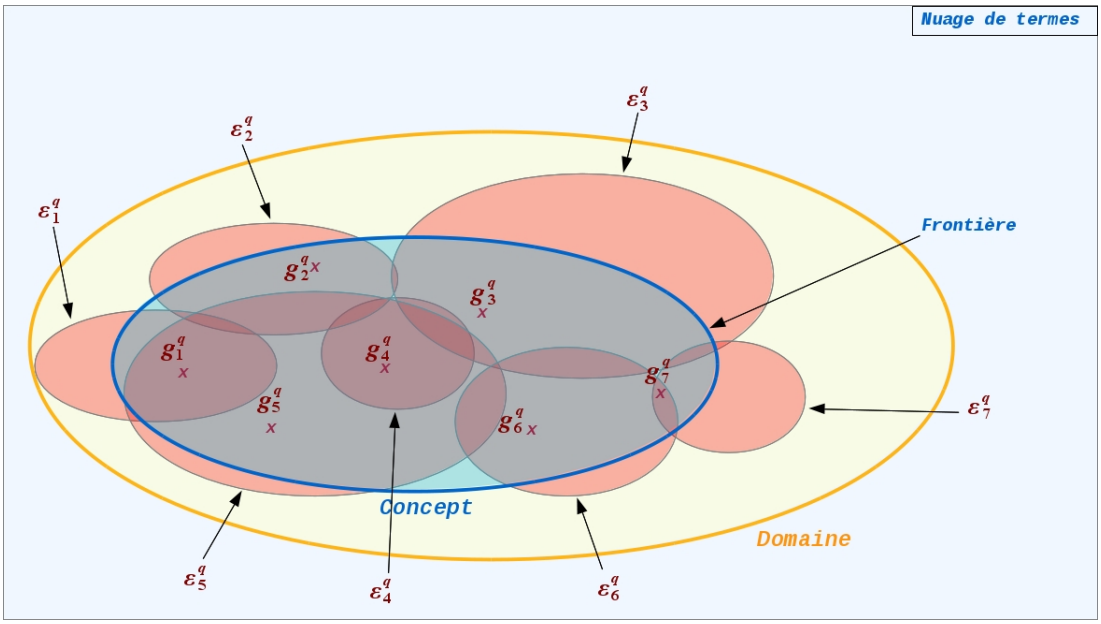

FIGURE 3.2 – Définition des *mot germes*.

La figure 3.3 illustre la répartition des termes pour le concept "actor" dans le domaine "movie". Comme nous pouvons le constater les différents mots germes sont bien caractéristiques du concept "actor". L'idée, est de déterminer l'ensemble des mots ($\mathcal{E}_j^{actor}$) sémantiquement proches de chacun des mots germes j. Considérons le mot germe "star", les mots qui pourront lui être rattachés seront par exemple des noms d'acteurs, le mot "Hollywood" qui symbolise le berceau des stars, etc. L'union de tous les ensembles de mots $\mathcal{E}_j^{actor}$ relatifs à chacun des germes constitue la classe du concept "actor".

Le fait de considérer une "anti-classe" est un élément clé de l'apprentissage. Les mots peuvent être polysémiques (avoir plusieurs sens), et peuvent avoir des sens radicalement différents selon le contexte dans lequel ils sont utilisés. C'est pourquoi il est nécessaire d'évaluer les mots dans le contexte dans lequel ils font référence au concept recherché (classe), mais aussi d'évaluer ces mots dans des contextes différents

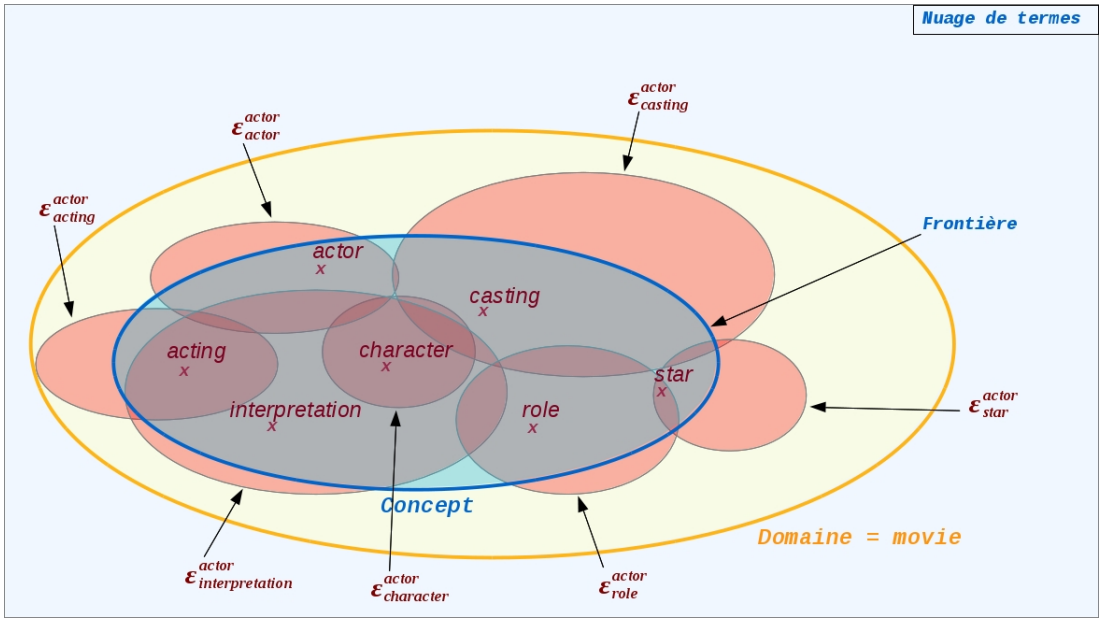

FIGURE 3.3 – Exemple de définition des *mots germes* pour le concept "actor" du domaine "movie".

de ceux du concept (anti-classe). Ainsi, il est possible de déterminer l'appartenance d'un mot à la classe ou à l'anti-classe. Cette technique permet aussi un filtrage de l'information, c'est-à-dire que les mots présents uniformément dans la classe et dans l'anti-classe ne sont pas discriminants, ils ne sont donc pas caractéristiques du concept d'intérêt. Ces mots n'apportent généralement pas d'information pertinente [Hearst 1997] et ils sont assimilés à du "bruit".

Les mots germes ont deux objectifs principaux. Le premier est de pouvoir amorcer l'apprentissage ; comme un enfant le fait en étudiant l'environnement qui l'entoure (c.f. section 3.2.1). Le second objectif est de pouvoir "diriger" l'apprentissage comme le font les parents avec leur enfant en donnant les relations entre signifié et signifiant. Ces mots germes seront utilisés dans les étapes suivantes de l'approche pour la construction du corpus d'apprentissage, mais aussi lors de la construction du lexique associé au concept.

3.2.5 Constitution du corpus d'apprentissage

Il est nécessaire pour pouvoir mettre en place l'apprentissage évoqué dans les sections précédentes de construire un corpus d'apprentissage adapté à la technique

utilisée. La constitution du corpus d'apprentissage de documents est donc une étape délicate dans un processus d'apprentissage. Deux difficultés sont à identifier :

1. Le corpus doit être assez riche : la richesse d'un corpus est définie par les différents *niveaux de langue* qu'il contient (soutenu, courant, familier). Notre objectif est d'identifier des concepts dans des textes de sources différentes. Qu'il s'agisse de textes littéraires ou de blogs, la manière d'exprimer un concept diffère puisque les champs lexicaux, le respect de la grammaire, etc. sont largement fonction des supports de communication : le style et les expressions employées, le vocabulaire technique de l'analyse, etc. ne seront pas les mêmes selon qu'il s'agisse d'une analyse du Masque et la Plume de France Inter ou bien du blog d'un cinéphile sur Overblog (c.f. chapitre 2). Pour que notre corpus d'apprentissage soit assez riche, nous proposons d'utiliser des textes web comme support d'apprentissage. Ces textes sont extrêmement variés et proviennent de sources différentes : blogs, textes scientifiques, etc. Ce qui répond à la question de légitimité et de représentativité des avis que les professionnels des sondages pointent lorsqu'ils parlent d'opinion mining comme nous l'avons évoqué dans le chapitre précédent. Cette diversité de supports est nécessaire, mais introduit des difficultés supplémentaires pour réussir à extraire l'information utile (les descripteurs) qui est souvent noyée dans le bruit (termes non pertinents). Le fait de considérer, d'un point de vue conceptuel, la classe X^q et l'anti-classe $\overline{X^q}$ nous permet d'assurer un apprentissage plus pertinent car plus robuste au bruit.
2. Le corpus doit être adapté à la technique d'apprentissage adoptée. Le corpus doit être en adéquation avec la proposition décrite en section 3.2.4 qui nécessite deux classes X^q et $\overline{X^q}$ pour caractériser un concept. Le corpus est alors composé de deux groupes de documents : les documents associés à la classe X^q appelés Doc^q et les documents associés à l'anti-classe $\overline{X^q}$ appelés $\overline{Doc}^q$ (pour un concept C^q d'un domaine $\mathcal{D}$).

   L'ensemble des n^q documents appartenant à la classe est noté :

   $$Doc^q = \{doc_n^q, n = 1 \dots n^q\} \tag{3.1}$$

   De même, l'ensemble des n'^q documents appartenant à l'anti-classe est noté :

   $$\overline{Doc}^q = \{doc_{n'}^q, n' = 1 \dots \overline{n'}^q\} \tag{3.2}$$

   Les documents Doc^q doivent contenir au moins une occurrence d'un mot germe de g_j^q et appartenir au domaine $\mathcal{D}$ pour assurer leur appartenance à $\bigcup \mathcal{E}_j^q$. De

la même manière, les documents $\overline{Doc}^q$ ne doivent contenir aucune occurrence des germes dans G^q et appartenir au domaine $\mathcal{D}$ pour assurer leur appartenance à $\mathcal{D} \setminus \bigcup \mathcal{E}_j^q$ (c.f. section 3.2.4). Pour obtenir de tels documents sur le web, nous utilisons un moteur de recherche et un ensemble de requêtes associées. L'exemple 2 donne l'exemple d'une requête qui permet d'obtenir des documents relatifs au concept "actor" en considérant le mot germe "casting". La figure 3.4 montre comment est construit le corpus d'apprentissage.

Exemple 2 *Considérons le concept "**actor**" du domaine "movie". La requête, en utilisant le moteur de recherche Google (en août 2010) qui nous permet d'obtenir des documents relatifs au mot germe **casting** est : "+movie +casting -acting -actor -character -interpretation -role -star". Le symbole + (resp. -) indique que ces mots doivent se retrouver (resp. ne pas se retrouver) dans les documents retournés.*

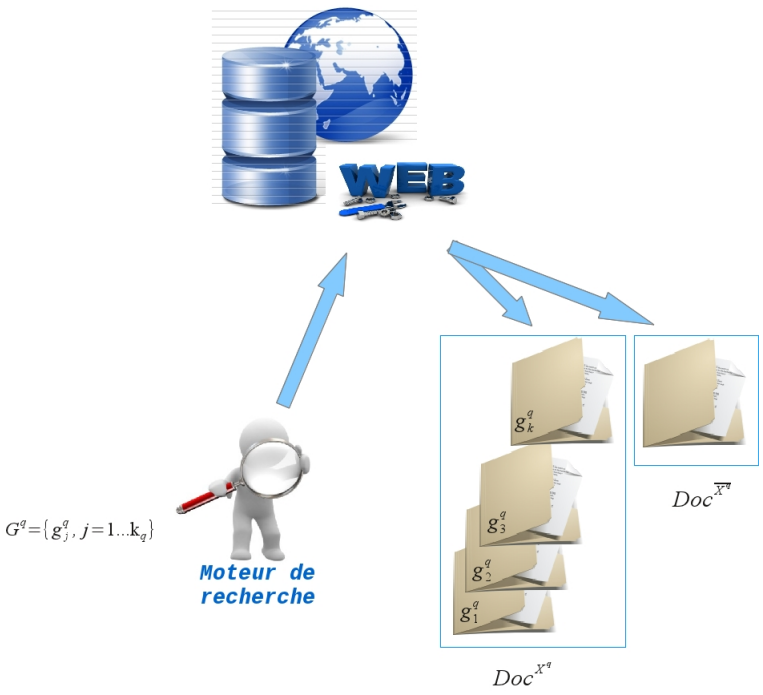

FIGURE 3.4 – Constitution du corpus d'apprentissage

Les textes constituant le corpus sont prétraités de la façon suivante. Tout d'abord, les éléments non pertinents (balises HTML, scripts, images, ...) sont supprimés. Les textes sont ensuite lemmatisés et chacun des mots fait l'objet d'une analyse morphosyntaxique pour déterminer sa classe grammaticale.

3.2.6 Apprentissage des descripteurs

Au cours de l'étape précédente nous avons construit un corpus de documents, cette section présente la technique d'apprentissage des descripteurs d'un concept C^q d'un domaine $\mathcal{D}$ qui peut être résumée en trois étapes :

1. **Compréhension de l'environnement :** Dans notre analogie avec l'apprentissage de la langue chez l'enfant, cette étape est la première du processus d'apprentissage. Elle consiste chez l'enfant à attacher une sémantique à son environnement. Dans notre contexte d'apprentissage automatique, l'approche est un peu différente puisque le système se restreint à un contexte particulier où un concept précis est exprimé. C'est pourquoi nous proposons d'utiliser des mots germes qui permettent d'assurer que l'apprentissage est bien effectué dans un contexte particulier. Cette phase de l'apprentissage fait référence à la dimension sémantique.

2. **Lien entre les phénomènes observés et la langue :** les nombreuses situations dans lesquelles est plongé l'enfant en phase d'apprentissage lui permettent de rattacher une sémantique observée entre une situation particulière et un mot. Par exemple, l'enfant peut rattacher le concept "voiture" au concept "garage" puisqu'à chaque fois qu'il est dans le garage on lui parle de voiture. Notre approche *Synopsis* utilise le même principe en observant la fréquence d'apparition des mots par rapport au concept spécifique : plus un mot est présent lorsque l'on parle d'un concept plus il est vraisemblable qu'il caractérise ce concept. Cette phase de l'apprentissage fait référence aux dimensions sémantique et syntaxique du langage.

3. **Capacité à interpréter le(s) sens d'un mot :** cette étape de l'apprentissage est la dernière étape du processus d'apprentissage chez l'enfant. En fonction de son vécu (nombre de situations dans lequel il s'est retrouvé lors de son apprentissage), l'enfant est plus ou moins capable d'identifier le vocabulaire qui s'y rattache, cela dépend directement du milieu culturel dans lequel son apprentissage a été effectué. C'est notamment une des raisons pour lesquelles les descripteurs d'un concept dépendent du domaine dans lequel ils sont utilisés. C'est pourquoi il est nécessaire d'apprendre ces descripteurs sur une grande variété de supports pour être en mesure d'identifier le concept considéré dans ses différents modes d'expression. Cette phase fait référence à la dimension pragmatique du langage.

Pour assurer un apprentissage des descripteurs qui tienne compte des dimensions *syntaxique*, *sémantique* et *pragmatique* [Morris 1938], ainsi que des principes d'apprentissage évoqués chez l'enfant, nous proposons d'utiliser pour chacune des dimensions les outils et principes suivants :

- **Sémantique** : des *mots germes* et la notion de proximité sémantique.
- **Syntaxique** : un lemmatiseur et un analyseur morphosyntaxique.
- **Pragmatique** : variété des supports d'apprentissage : plusieurs segmentations envisageables selon l'expertise souhaitée.

3.2.6.1 Notion de fenêtre et de proximité sémantique

Le fait d'utiliser des textes web introduit du bruit dans le processus d'apprentissage (c.f. section 3.2.5), il est donc nécessaire de développer une méthode pour supprimer ce bruit. Nous proposons d'introduire la notion de **fenêtre** centrée sur les mots germes. Cette technique classique en traitement du signal nous permet de faire un apprentissage local pour apprendre les descripteurs sémantiquement proches des mots germes considérés (notion de *proximité sémantique*).
De plus, nous proposons de considérer les relations sémantiques entre les mots en analysant leurs propriétés statistiques (entre signifiés). Notre approche ne se base pas sur des règles, mais sur la probabilité qu'un mot puisse apparaître dans le même contexte (dimension syntaxique et sémantique) qu'un mot germe sans qu'ils aient forcément une relation directe évidente. La fenêtre permet d'assurer un apprentissage dans un contexte.

Il est nécessaire, pour éviter de multiplier le nombre de descripteurs, de différencier les différentes classes grammaticales des mots qui ont chacune une fonction langagière particulière. C'est pourquoi, nous nous focalisons uniquement sur les noms qui sont reconnus comme porteurs de sens [Kleiber 1996]. Nous noterons l'ensemble de toutes les classes grammaticales par $*$, celle des noms communs sera notée par son abréviation NC, les adjectifs JJ... Pour chaque document doc du corpus relatif au concept C^q, nous utilisons une fenêtre $F(g_j^q, sz, doc)$ définie par :

$$F(g_j^q, sz, doc) = \{m \in doc / d_{NC}(g_j^q, m) \leq sz\} \tag{3.3}$$

où g_j^q est le mot germe j du concept q, sz la taille de la fenêtre et $d_{NC}(g_j^q, m)$ est la distance correspondant au nombre de noms communs (NC) séparant un mot m de g_j^q.

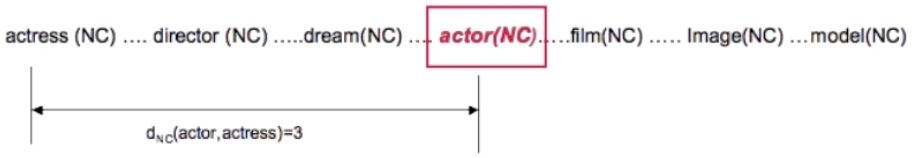

FIGURE 3.5 – Exemple pour une fenêtre de taille 3

Exemple 3 *La figure 3.5 montre un exemple de fenêtre de taille 3,* i.e.*, il y a trois noms communs à gauche du mot germe "actor" (actress, director, dream) et 3 noms communs à droite (film, image, model). Les pointillés illustrant les mots de nature grammaticale autres que "nom commun".*

3.2.6.2 Influence

L'utilisation d'une fenêtre $F(g_j^q, sz, doc)$ permet de filtrer et de limiter la portée de la recherche de distance sémantique entre le mots germes et les autres mots de la fenêtre. Nous considérons ensuite que plus un nom commun est loin d'un germe auquel il se rapporte, moins il influence ce germe, autrement dit, moins ce mot est "utile" à la compréhension du concept exprimé. Le filtrage avec la fenêtre ne suffit pas pour modéliser cette influence (c.f. section 3.3). A cet effet, il est nécessaire d'établir une distance entre le mot germe et les autres mots sur la fenêtre elle-même , indépendante de la nature grammaticale des mots de la fenêtre. Nous définissons alors l'influence d'un mot germe sur les autres mots de la fenêtre par :

$$I(m, g_j^q, sz, doc) = \left[\begin{array}{l} 0 \textbf{ si } m \notin F(g_j^q, sz, doc) \\ h(d_*^t(m, g_j^q)) \textbf{ si } m \in F(g_j^q, sz, doc) \end{array} \right. \tag{3.4}$$

où $d_*^t(m, g_j^q)$ est le nombre de mots, de nature grammaticale quelconque $(*)$ entre g_j^q et m. h est une fonction de lissage. Dans la pratique nous utilisons deux "semi-gaussiennes" centrées sur g_j^q (c.f. figure 3.6). L'intérêt de cette fonction est de pondérer l'apparition d'un mot dans la fenêtre par rapport à sa distance avec le mot germe. Considérons l et r les mots les plus éloignés à gauche et à droite de g_j^q dans une fenêtre $F(g_j^q, sz, doc)$, l'influence est la résultante d'un filtrage gaussien défini par :

$$gauss\left(\frac{d_*^t(g_j^q, m)}{d_*^t(g_j^q, l)}, \mu, \sigma\right) \tag{3.5}$$

pour un mot m à gauche de g_j^q et par :

$$gauss\left(\frac{d_*^t(g_j^q, m)}{d_*^t(g_j^q, r)}, \mu, \sigma\right) \tag{3.6}$$

pour un mot m à droite de g_j^q avec la fonction de $gauss$ définie par :

$$gauss(x, \mu, \sigma) = \exp - \frac{(x-\mu)^2}{2\sigma^2} \tag{3.7}$$

où $\mu = 0.0$ et $\sigma = 0.225$ (c.f. figure 3.7).

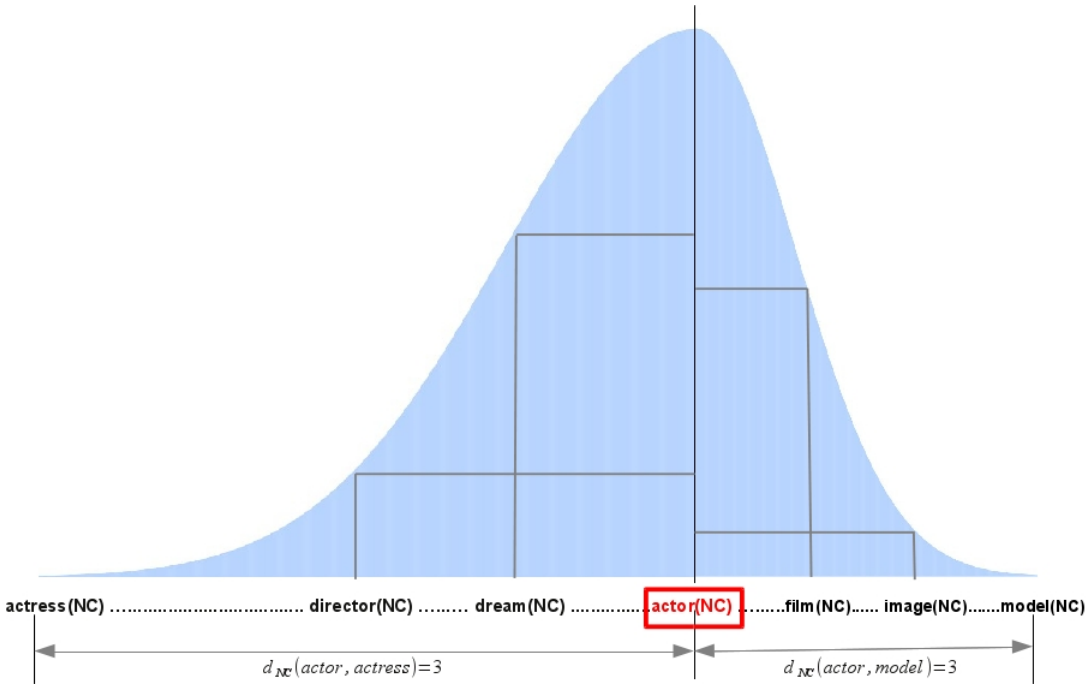

FIGURE 3.6 – Exemple de filtre gaussien pour une fenêtre de taille 3

L'idée d'un filtre gaussien est de pouvoir pondérer chaque occurrence d'un mot dans une fenêtre en fonction de sa distance par rapport au mot germe central. Nous ne considérons pas dans cette approche la structure des phrases pour pouvoir s'affranchir des traitements morphosyntaxiques. Le choix d'une gaussienne permet de tenir compte de la structure du discours : le découpage en phrases, en propositions (lieu, temps...).

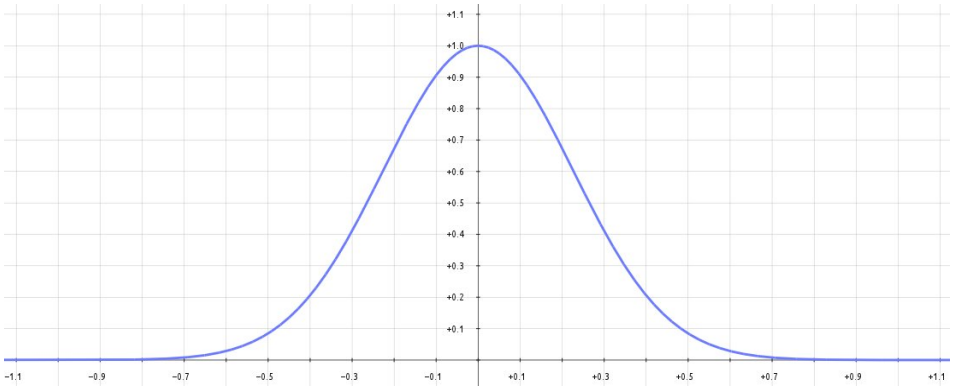

FIGURE 3.7 – Courbe de Gauss (c.f. équation 3.7) avec $\mu = 0.0$ et $\sigma = 0.225$

3.2.6.3 Représentativité

En utilisant la notion de fenêtre et un filtre gaussien nous assurons un apprentissage dans un contexte donné qui tient compte des dimensions sémantique et syntaxique ; la troisième dimension pragmatique est, elle, principalement liée au support d'apprentissage. Nous proposons d'étudier la répartition des mots dans chacune des classes X^q et $\overline{X^q}$, en utilisant cette notion contextuelle (fenêtre et filtre gaussien) : on parlera de *représentativité* d'un mot dans une classe. Pour déterminer la représentativité $\rho^q(M, sz)$ (resp. $\overline{\rho^q}(M, sz)$) d'un mot M nous étudions son nombre d'occurrence $\mathcal{O}(M, doc)$ dans chaque fenêtre $F(g_j^q, sz, doc)$ pondérée, comme expliqué en section 3.5, pour chaque document Doc^q (resp.$\overline{Doc^q}$) d'une classe. La représentativité d'un mot M dans la classe du concept q est définie par :

$$\rho^q(M, sz) = \sum_{g_j^q \in G^q} \sum_{doc \in Doc_j^q} \sum_{\gamma \in \mathcal{O}(g_j^q, doc)} \sum_{m \in \mathcal{O}(M, doc)} I(m, \gamma, sz, doc) \tag{3.8}$$

De même, nous calculons la représentativité $\overline{\rho}^q(M, sz)$ d'un mot M dans l'anti-classe par :

$$\overline{\rho}^q(M, sz) = \sum_{\overline{doc} \in \overline{Doc}^q} \sum_{\gamma \in \mathcal{O}(\mathcal{D}, \overline{doc})} \sum_{m \in \mathcal{O}(M, \overline{doc})} I(m, \mathcal{D}, sz, \overline{doc}) \tag{3.9}$$

Pour pouvoir calculer la représentativité de chaque mot M dans l'anti-classe $\overline{\rho}^q$ sur $\overline{Doc}^q$nous considérons comme mot germe unique le domaine $\mathcal{D}$ lui-même (c.f. Figure 3.2). En effet, l'anti-classe, de par la construction du corpus, ne peut contenir de mots germes.

La représentativité d'un mot dans chacune des deux classes nous donne déjà une idée sur son appartenance à l'une ou l'autre des classes, mais cette information n'est

pas suffisante pour pouvoir se prononcer puisqu'un mot peut être représentatif des deux classes. Par conséquent, on identifie quatre groupes de mots :

1. Les mots très fréquents dans la classe et peu fréquents dans l'anti-classe.
2. Les mots très fréquents dans l'anti-classe et peu fréquents dans la classe.
3. Les mots très fréquents dans la classe et très fréquents dans l'anti-classe.
4. Les mots peu fréquents dans la classe et peu fréquents dans l'anti-classe.

Les deux premiers groupes ne présentent pas de difficulté pour déterminer l'appartenance à l'une ou l'autre des classes. Les mots du troisième groupe ne sont pas discriminants, : ce sont des mots apparentés à du bruit ou des mots qui ne sont pas porteurs de sens par rapport au concept considéré. Pour les mots appartenant au quatrième groupe, nous ne pouvons pas nous prononcer. La représentativité des mots étant très faible dans les deux classes, nous manquons d'information pour avoir une représentativité significative. Ces mots sont appelés *mots candidats* et font l'objet d'un traitement particulier (c.f. section 3.2.6.5).

	classe	**anti-classe**
M	$\rho^q(M, sz)$	$\overline{\rho}^q(M, sz)$
film	1080	460
actress	170	1
theater	1	370
poster	700	700
Matt Vaughn	2	2
Sam Worthington	1	2
story	100	120

TABLE 3.1 – Exemple de mots appris (noms communs et noms propres) pour le concept "acteur" avec leurs représentativités dans la classe et dans l'anti-classe.

3.2.6.4 Calcul du score des descripteurs

Comme dans tous les problèmes de classification et de clustering, il est nécessaire de pouvoir déterminer l'appartenance d'une entité à une classe unique au final. Ici, nous cherchons à déterminer en fonction de la représentativité d'un mot son appartenance à une classe plutôt qu'à l'autre (classe/anti-classe). Nous proposons de déterminer une fonction score qui permettra de l'affecter à l'une des deux classes. Ce score sera stocké, pour chaque descripteur, dans un lexique L^q associé au concept

C^q. À partir de leurs représentativités respectives dans la classe et dans l'anti-classe, nous établissons une fonction de discrimination qui mesure la croyance attribuée à la classification. Pour un mot M, plus la différence entre sa représentativité dans la classe et sa représentativité dans l'anti-classe est importante, plus il est vraisemblable que M appartienne à l'une ou l'autre des deux classes, et inversement, plus cette différence est faible, moins il est certain que le mot appartienne à l'une ou l'autre des deux classes.

Pour effectuer cette discrimination, il est nécessaire de choisir un opérateur f qui respecte les besoins précédemment énoncés et que nous allons détailler plus formellement :

1. Arbitrairement, les valeurs positives seront attribuées pour l'appartenance à la classe et les valeurs négatives à l'anti-classe.
2. la fonction de discrimination f a pour argument $\rho^q(M, sz)$ et $\overline{\rho}^q(M, sz)$) et fournit une valeur représentant l'appartenance à la classe ou à l'anti-classe $f : \mathbb{R}^+ \times \mathbb{R}^+ \longrightarrow \mathbb{R}$, pour simplifier, nous noterons $f(\rho^q, \overline{\rho}^q)$ lorsqu'il n'y a pas d'ambiguïté sur le mot M et la taille de fenêtre sz .
3. La valeur retournée par l'opérateur de discrimination doit caractériser la différence d'occurrence entre les deux classes, $\mid f(\rho^q, \overline{\rho}^q) \mid$ peut donc être choisie comme étant homogène à une fréquence et vérifier que $f(0, \overline{\rho}^q) = -\overline{\rho}^q$, $f(\rho^q, 0) = \rho^q$ et $-\overline{\rho}^q \leqslant f(\rho^q, \overline{\rho}^q) \leqslant \rho^q$. Ceci signifie que le différentiel de fréquence n'est autre que la fréquence d'origine signée lorsque la classe (+) et l'anti-classe (-) sont d'intersection vide $(X^q \cap \overline{X}^q) = \emptyset$. Enfin, $-\overline{\rho}^q \leqslant f(\rho^q, \overline{\rho}^q) \leqslant \rho^q$ signifie que f doit être convexe.
4. L'opérateur doit être continu/dérivable sur $\mathbb{R}^{+2}_*$. Par extension, $f(0, 0) = 0$, mais ce cas ne peut jamais se présenter car il n'est pas possible d'être en présence d'un mot jamais rencontré (fréquence nulle dans la classe et dans l'anti-classe).
5. Lorsque la représentativité dans la classe et dans l'anti-classe est identique, alors le résultat est nul : $(f(\rho^q, \overline{\rho}^q) = 0$ si $\rho^q = \overline{\rho}^q$.
6. Même si la différence de représentativité entre la classe et dans l'anti-classe est faible, plus le module de ses représentativités croît, plus il est peu probable que ce mot soit représentatif de la classe ou de l'anti-classe et donc, le résultat doit tendre vers 0 avec l'augmentation de la représentativité. Par exemple, si $\rho =$

100 et $\overline{\rho} = 0$, le mot appartient clairement à la classe, alors que si $\rho = 100100$ et $\overline{\rho} = 100000$, il est peu probable que le mot soit discriminant. Ceci s'exprime plus formellement par : $\forall R > 0, \rho^q \neq \overline{\rho}^q, f(\rho^q, \overline{\rho}^q) > f(R + \rho^q, R + \overline{\rho}^q)$.

7. Plus la représentativité dans l'une des classes augmente, alors qu'elle reste constante dans l'autre, plus grande est la certitude que le mot appartienne à la première classe (celle où elle augmente). Cela signifie que toutes les coupes, pour un ρ^q fixé, sont décroissantes et que les coupes pour un $\overline{\rho}^q$ fixé sont croissantes. Ceci s'exprime par :

$$\underset{\forall R>0}{\rho^q \neq \overline{\rho}^q} \quad f(\rho^q, \overline{\rho}^q) < f(R + \rho^q, \overline{\rho}^q) \tag{3.10}$$

et

$$\underset{\forall R>0}{\rho^q \neq \overline{\rho}^q} \quad f(\rho^q, \overline{\rho}^q) > f(\rho^q, R + \overline{\rho}^q) \tag{3.11}$$

Les fonctions précédentes vérifient toutes les propriétés requises pour tout $n > 0$:

Si $n = 0$: $f^0(\rho^q, \overline{\rho}^q) = \rho^q - \overline{\rho}^q$ ne respecte pas la propriété 6.

En pratique, f^2 est le meilleur compromis. Le score $Sc(M, sz)$ d'un mot M est alors défini par :

$$Sc(M, sz) = f^2(\rho^q(M, sz), \overline{\rho}^q(M, sz)) \tag{3.12}$$

M	$Sc(M, sz)$
film	100.5
actress	165
theater	-365
poster	0
Matt Vaughn	515
Sam Worthington	771
story	-0.17

TABLE 3.2 – Exemple de scores de discrimination de mots candidats .

Chaque mot appris devient ensuite une entrée du lexique L^q relatif au concept C^q et est affecté du score $Sc(M, sz)$ dans L^q..

3.2.6.5 Mots candidats

Les mots candidats sont les mots ayant une représentativité faible dans la classe et dans l'anti-classe. Ce cas peut-être dû à plusieurs causes : I) Le corpus d'apprentissage n'est pas assez complet ; l'échantillon n'est pas assez représentatif ; II) La fréquence intrinsèque du mot est toujours faible par rapport aux autres mots, mais il est représentatif de la classe et de l'anti-classe ; III) Ces mots ne sont pas représentatifs du concept. Dans tous les cas, il est nécessaire d'enrichir le corpus pour pouvoir prendre une décision quant à ces mots.
Pour faire face à ce manque d'informations, nous construisons, pour chaque mot candidat un nouveau corpus de documents spécifiques, en adoptant le même principe de construction qu'exposé en section 3.2.5 mais en imposant, en plus des règles de construction définies précédemment, que les documents retenus par le moteur de recherche contiennent au moins une occurrence du mot candidat. Nous pouvons grâce à ce processus obtenir de l'information ciblée sur le mot candidat et ainsi pouvoir déterminer une représentativité plus fiable du mot candidat (c.f. section 3.2.6.3) . Le nouveau score $Sc(M, sz)$ attribué à un mot candidat pose un problème de commensurabilité vis-à-vis des autres mots du lexique qui ont été appris sur des supports de tailles différentes. Nous n'avons pas encore résolu ce problème à ce jour et nous avons considéré une fonction d'homogénéisation naïve qui donne de bons résultats en pratique.
Ce traitement particulier sur les mots candidats n'est pas uniquement utilisé lors de la phase d'apprentissage d'un concept, il est aussi utilisé lors du processus d'extraction thématique pour effectuer un apprentissage des mots inconnus par le système dans l'objectif de découvrir de nouveaux descripteurs. Lors du processus d'extraction, chaque mot ne faisant pas l'objet d'une entrée du lexique est considéré comme un mot candidat. Ainsi, ces nouveaux descripteurs sont intégrés au lexique relatif au concept considéré. Ce principe nous permet d'avoir un système évolutif qui est en mesure d'apprendre de nouveaux descripteurs et ainsi permettre de suivre l'évolution du langage au cours du temps (apparition d'un nouveau vocabulaire, etc). Nous pouvons donc prétendre à un apprentissage continu du langage au cours du temps.

3.2.7 Extraction thématique

L'objectif dans cette section est d'identifier les parties d'un document *doc* traitant d'un même concept en utilisant son lexique construit au préalable (c.f. figure

M	$\rho^q(M, sz)$	$\overline{\rho}^q(M, sz)$
Matt Vaughn	2	2
Sam Worthington	1	2

TABLE 3.3 – Exemple de représentativité de mots candidats **avant** le processus d'enrichissement.

3.8), c'est-à-dire déterminer la fonction d'interprétation $interpretation(texte) = f(Se, Sy, Pr, texte)$ introduite dans le chapitre précédent. En reprenant la définition d'un concept, on peut retenir deux points clés. Le premier fait référence à l'idée qu'il existe un niveau de description objectif du concept , c'est-à-dire partager par l'ensemble des individus : ils recherchent tous la même chose. Le second est lié à la perception subjective de l'individu. Les modes d'expression varient d'un individu à l'autre selon la culture et l'expertise de celui-ci. Notre système tient compte de cette idée en proposant d'identifier les parties de textes correspondant à l'acceptation commune tout en permettant de distinguer les différents points de vue ou modes d'expression potentiels.

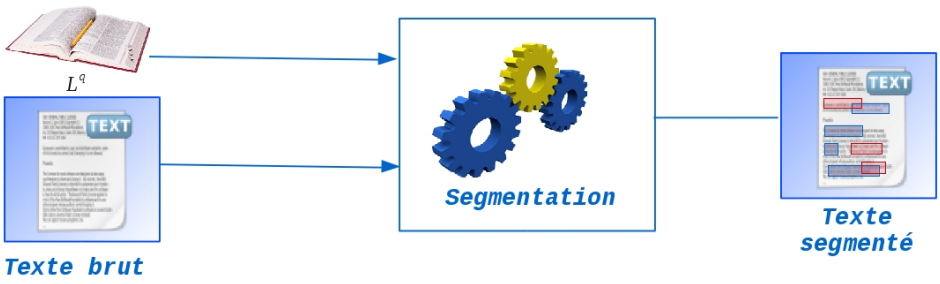

FIGURE 3.8 – Processus de segmentation.

Nous avons montré en section 3.2.6.1 qu'il est nécessaire de considérer les termes dans leur contexte pour qu'ils aient un sens. À la lecture d'un texte, le lecteur lit les mots les uns après les autres et réussit, grâce à un effet mémoire (il se souvient des mots précédemment lus), à construire l'idée générale exprimée. Cet effet mémoire se limite au nombre de mots nécessaires et suffisants à la compréhension du texte. La succession de chacune des idées et sentiments, liés à la sémantique des mots retenus, construit l'idée générale exprimée par le document pour l'individu. La gymnastique de l'esprit consiste donc à agréger les nuances des mots porteurs de sens pour dégager le sens général d'un paragraphe, d'un texte. Nous proposons de modéliser cette agrégation en utilisant une fenêtre (c.f.section 3.5) qui représentera la plus petite

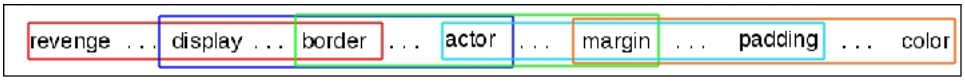

FIGURE 3.9 – Exemple de fenêtre glissante de taille 1 successivement centrée sur les noms communs.

entité sémantique pouvant évoquer le concept pour l'individu. Par ailleurs, pour être en mesure de détecter les ruptures sémantiques (conceptuelles), il est nécessaire d'adopter le principe de lecture continue consistant à considérer successivement des fenêtres qui se chevauchent partiellement, nous parlerons alors de fenêtre glissante Fs.

$$Fs(m_{NC}, sz, doc) = \{m \in doc / d_{NC}(m_{NC}, m) \leq sz\} \tag{3.13}$$

Fs est une fenêtre de taille sz appliquée sur un document doc. Contrairement à la section 3.2.6.1, la fenêtre n'est pas centrée sur des mots germes, mais elle est successivement centrée sur chacun des noms communs du document (c.f. figure 3.9). De plus, seuls les noms communs sont considérés pour la construction des fenêtres, il semblerait judicieux de considérer une fenêtre de taille identique à celle utilisée lors de l'apprentissage. En effet, lors de l'apprentissage, les mots sont appris dans un certain contexte (taille de la fenêtre. c.f. section 3.2.6.1), et le score qui leur est attribué en dépend, il semble donc naturel d'évaluer chacun des mots dans le même contexte que celui dans lequel ils ont été appris pour être cohérent (c.f. section 3.3).

À partir d'un lexique L^q (c.f. section 3.2.6.4) relatif au concept à identifier, nous sommes en mesure d'attribuer un score à chaque mot présent dans le document doc. Le score des mots nous permet de calculer un score moyen $Score(fs_j)$ pour chaque fenêtre fs_j afin de reproduire le principe d'agrégation précédemment évoqué. $Score(fs_j)$ est la moyenne arithmétique des scores des mc mots contenus dans fs_j, toutes classes grammaticales apprises confondues, $Score(fs_j)$ est défini par :

$$Score(fs_j) = \frac{1}{mc} \sum_{M \in fs} Sc(M, sz) \tag{3.14}$$

Le score $Score(fs_j)$ modélise deux choses : d'une part l'intensité du concept exprimé, et d'autre part la certitude que fs_j traite du concept C^q. L'intensité dépend de l'effet "moyennage" (propre à la combinaison des termes), et la croyance dépend du score de chacun des mots de la fenêtre (homogène à une fréquence).

3.2.7.1 Segmentation du texte

La segmentation de texte est un processus complexe dont l'objectif est de déterminer les ruptures thématiques d'un document. Ici, l'objectif est de déterminer les ruptures dans le discours, cela revient dans notre cas à identifier les passages traitant d'un même concept C^q. Cette problématique revient à considérer chacune des fenêtres fs précédemment construites, et d'étudier à partir de quel moment le locuteur ne parle plus du même concept. L'idée est de déterminer un seuil th à partir duquel les fenêtres fs_j de score $Score(fs_j)$ supérieur à ce seuil, seront considérées comme traitant du concept C^q. Pour déterminer th nous proposons de faire une analyse de sensibilité, au sens automatique du terme, du système de segmentation. Cela revient donc à étudier le comportement du système sur toute sa plage de fonctionnement, c'est-à-dire pour th strictement positif et inférieur ou égal au plus grand des scores obtenus par les fenêtres $fs_j : 0 < th \leq max_j(Score(fs_j))$. Le fait d'exclure $th < 0$ signifie qu'une fenêtre ayant un score négatif n'est pas en relation avec le concept C^q. L'idée, en faisant varier $th \in [0, max_j(Score(fs_j)]$ qui représente la croyance minimale autorisée en la présence de C^q pour classer un texte comme traitant de C^q est de regarder comment l'interprétation, c'est-à-dire les extraits identifiés, varient (même interprétation) en fonction de th. Lorsque les extraits identifiés varient peu sur une plage de th alors cela signifie qu'on a une interprétation stable sur cette plage (états stables du système).

La figure 3.10 est un exemple de résultat d'une analyse de sensibilité d'un texte pour un concept donné. Nous faisons varier th dans $]0, max(Score(fs_j)]$ et nous regardons le pourcentage de mots retenus en fonction de th. Cette fonction est strictement décroissante, puisque plus th est grand, autrement dit plus la contrainte de sélection sur le score est forte, moins les segments de texte jugés rattachés au concept sont nombreux. Si nous faisions varier th de $]-\infty, +\infty[$, lorsque $th = -\infty$ alors l'intégralité du texte serait retourné, ce qui revient à ne mettre aucune contrainte sur la croyance minimale en la présence de C^q. Lorsque $th = +\infty$, la contrainte est maximale, alors le système n'est pas en mesure de proposer une solution. Sur cet exemple, nous pouvons identifier trois "états stables" ou paliers, c'est-à-dire que la variation de th sur ces paliers de fonctionnement $([5; 24], [43; 102], [427; 508])$ n'affecte pratiquement pas le nombre de mots retenus (extraits). Ces "états stables" sont comparables à un processus cognitif où l'esprit, à la lecture d'un document, interprète une relation évidente entre les mots utilisés pour évoquer le concept avec un niveau de connaissance donné. Nous identifions dans cet exemple trois paliers si-

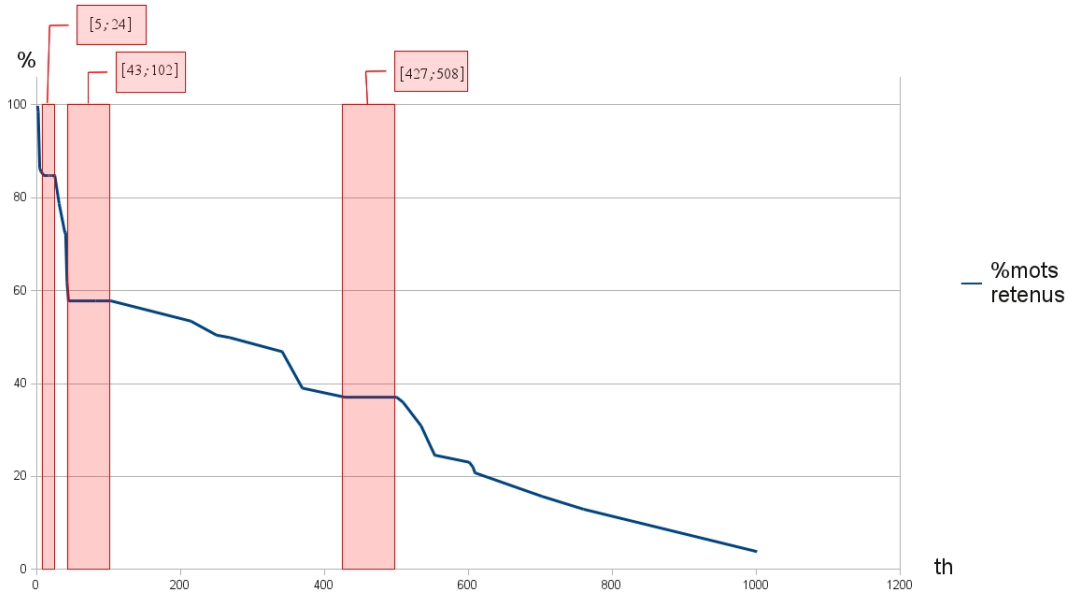

FIGURE 3.10 – Exemple d'analyse de sensibilité pour un document donné.

gnificatifs, ils correspondent chacun d'eux à des situations d'évidence où le concept considéré est clairement identifiable. Chacun des paliers correspond à des granularités sémantiques différentes. En effet, selon le vécu et l'expertise du lecteur, il y a plusieurs manières d'interpréter un document à sa lecture. Pour cet exemple nous identifions trois interprétations ou points de vue potentiels. Si l'analyse de sensibilité d'un document ne met pas en évidence ce type de motif, alors c'est que le document ne traite pas du concept considéré car aucune "évidence" n'a pu être construite. Plus formellement, la détection de ces paliers caractéristiques est réalisée par le calcul de la dérivée de la fonction qui donne le nombre de mots retenus en fonction des valeurs de th , c'est-à-dire la courbe d'analyse de sensibilité. Lorsque la dérivée s'annule, un palier est détecté. Ces "états stables" pourraient être associés à des "quantum de connaissances", la granularité de lecture est une fonction constante par morceaux. Un élève de cours élémentaire (CE1 et CE2) aura une lecture d'un texte qui différera de la lecture de ce texte quand il sera en cours moyen (CM1 et CM2) ou encore de celle qu'il en fera au collège car son vocabulaire se sera enrichi, son expérience sera plus grande, et son interprétation variera en conséquence. Lorsque l'on est dans l'un de ces quantums, il faut être beaucoup plus exigeant quant à la définition du concept

pour observer une variation significative des mots peuplant son lexique (pour sortir du "niveau d'énergie").
Il est alors possible de segmenter un document selon plusieurs niveaux de granularité propres à l'interprétation de chacun. Cette granularité permet de segmenter un texte selon un niveau d'expertise paramétrable et évoque la complexité de l'interprétation d'un texte. Cette notion fait référence à l'imprécision de la langue et des mots qu'elle véhicule [Mélès 1971]. Lors de la segmentation, seules les fenêtres sont considérées, les parties de textes résultantes en sont alors dépendantes (les phrases peuvent alors être coupées), c'est pourquoi, un processus purement syntaxique permet de reconstruire ces phrases incomplètes, soit en ajoutant un bout de phrase manquante, soit en supprimant le fragment de phrase considéré. Le choix entre les deux alternatives dépend de la proportion du texte extrait dans chaque phrase. Dans nos expérimentations, nous avons choisi à priori un seuil égal à la moitié de la taille de la phrase.

Si le fragment de phrase est supérieur en nombre de mots à la moitié de la phrase, le fragment manquant est ajouté pour avoir une phrase complète, sinon le fragment est supprimé.

À partie du choix des mots germes (qui donnent le contexte sémantique du concept), du choix de la taille de fenêtres (qui influent sur l'analyse syntaxique) et du choix de la croyance minimale autorisée par le biais du seuil th (qui modélise l'expertise de l'utilisateur), nous sommes en mesure de contrôler la fonction $interpretation(texte) = f(Se, Sy, Pr, texte)$.

Notre segmentation tient compte des trois dimensions du langage définies par Morris : syntaxe, sémantique et pragmatisme.

Les figures 3.11 et 3.12 sont des exemples de segmentations possibles selon la granularité choisie en considérant le concept "scenario". Sur cet exemple, notre outil *Synopsis* identifie deux segmentations possibles pour le texte considéré. La figure 3.11 présente (le texte retenu est en gras) le résultat de segmentation pour la granularité la plus élevée (point de vue le plus large), la figure 3.12 présente (le texte retenu est en gras) le résultat de segmentation pour la granularité la plus basse (point de vue le plus précis).

A film with the scope of James Cameron's Avatar was always going to be a risk both artistically and financially, especially in today's economic climate. Whether it will pay off monetarily is a question only time can answer, but this viewer can at the very least attest to it being an artistic triumph.
Avatar brings us as close as cinema ever has to actually visiting an alien world. The beautiful environs, the exotic creatures and incredibly lifelike natives of Pandora arrest the senses, visually, aurally and emotionally. The world in Avatar is the true star of the show. The amount of detail and work that has gone into bringing this new world alive is seriously impressive, and it will be a while before we see anything that overtakes it in scope and quality. WETA Workshop and ILM have truly outdone themselves.
Relative newcomer, Aussie Sam Worthington provides a solid human heart amongst all the science-fiction/fantasy beauty and Zoe Saldana gives an impressive performance as the 8 foot tall Na'vi, Neytiri. Even though the characters they both play are blue, giant, catlike aliens, they managed to evoke a chemistry and likability that pierces through the special effects.
That's not to say that everything is perfect. The story is basic and dare I say, banal and predictable. We have seen it plenty of times in all forms of media. **The bad guys are cartoonishly evil, and sadly paper thin. The love story, while charming, is also banal despite being between man and alien.** But in the face of these shortcomings, Avatar is a success because its storytelling lies in the brilliant visuals.
Avatar is a beautiful piece of film and a true event. It does exactly what cinema was always intended to - it takes us away from our problems and worries for a few hours and gives us memorable images which will undoubtedly and deservedly enter into the cultural lexicon to stay for the foreseeable future.

FIGURE 3.11 – Exemple de résultat de segmentation pour le concept "scénario" suivant le premier point de vue.

3.3 Expérimentations et résultats

Cette section a pour objectif de montrer la pertinence de l'approche sur des expérimentations. Les expériences ont été menées sur deux corpus de test d'environ 20000 mots chacun : l'un dans le domaine du cinéma, l'autre dans celui de la restauration. Pour le domaine du cinéma, deux concepts sont considérés : *acteur* ("actor" en anglais) et *scénario* ("scenario" en anglais). Pour le domaine de la restauration, deux concepts sont considérés : *propreté* ("cleanliness" en anglais) et *service* ("service" en anglais). Les indicateurs de Précision, Rappel et F1Score sont utilisés pour évaluer la classification. L'apprentissage des lexiques est effectué en utilisant le moteur de recherche *Google*[3] pour construire le corpus d'apprentissage. Nous utilisons *TreeTagger* comme lemmatiseur et analyseur morphosyntaxique. TreeTagger est un

3. http://www.google.fr

A film with the scope of James Cameron's Avatar was always going to be a risk both artistically and financially, especially in today's economic climate. Whether it will pay off monetarily is a question only time can answer, but this viewer can at the very least attest to it being an artistic triumph.
Avatar brings us as close as cinema ever has to actually visiting an alien world. The beautiful environs, the exotic creatures and incredibly lifelike natives of Pandora arrest the senses, visually, aurally and emotionally. The world in Avatar is the true star of the show. The amount of detail and work that has gone into bringing this new world alive is seriously impressive, and it will be a while before we see anything that overtakes it in scope and quality. WETA Workshop and ILM have truly outdone themselves.
Relative newcomer, Aussie Sam Worthington provides a solid human heart amongst all the science-fiction/fantasy beauty and Zoe Saldana gives an impressive performance as the 8 foot tall Na'vi, Neytiri. Even though the characters they both play are blue, giant, catlike aliens, they managed to evoke a chemistry and likability that pierces through the special effects.
That's not to say that everything is perfect. The story is basic and dare I say, banal and predictable. We have seen it plenty of times in all forms of media. **The bad guys are cartoonishly evil, and sadly paper thin. The love story, while charming, is also banal despite being between man and alien.** But in the face of these shortcomings, Avatar is a success because its storytelling lies in the brilliant visuals.
Avatar is a beautiful piece of film and a true event. It does exactly what cinema was always intended to - it takes us away from our problems and worries for a few hours and gives us memorable images which will undoubtedly and deservedly enter into the cultural lexicon to stay for the foreseeable future.

FIGURE 3.12 – Exemple de résultat de segmentation pour le concept "scénario" suivant le second point de vue.

outil qui permet d'annoter un texte avec des informations comme le genre des mots (noms, verbes, infinitifs et particules) ainsi que des informations de lemmatisation. Il a été développé par *Helmut Schmid* dans le cadre du projet "TC" dans l'ICLUS (Institute for Computational Linguistics of the University of Stuttgart). TreeTagger permet l'étiquetage de plusieurs langues : allemand, anglais, français, italien, espagnol, bulgare, russe, le grec, le portugais, le chinois et les textes français anciens. Il est adaptable à d'autres langages dès lors que des lexiques et des corpus étiquetés manuellement sont disponibles.

Nous identifions cinq points clés pour valider notre approche et déterminer les paramètres algorithmiques optimaux. Le premier point concerne l'apprentissage, et plus particulièrement la taille de la fenêtre utilisée dans la classe, celle utilisée dans l'anti-classe, ainsi que le gain apporté par l'utilisation d'une gaussienne pour la fonction de lissage h de la formule de l'influence $I(m, g_j^q, sz, doc)$. Le second point

concerne la taille de la fenêtre lors de la phase d'extraction. Le troisième point fait référence au nombre de documents web nécessaires à un apprentissage de qualité. Le quatrième point détermine de nombre de mots germes nécessaires pour réaliser un apprentissage pertinent. Le dernier point montre l'efficacité du système sur les deux domaines considérés (cinéma et restauration) ainsi que l'intérêt d'enrichir les lexiques en réapprenant les mots candidats.

Le paramètre th correspondant au niveau de granularité est choisi automatiquement par l'algorithme. Celui-ci est identifié en observant pour chaque th potentiel celui pour lequel la segmentation proposée correspond au mieux à celle de l'expert.

Remarque : Les expériences pour déterminer les paramètres algorithmiques ont été réalisées sur les deux domaines d'application et sur leurs concepts respectifs, pour des raisons de clarté évidentes, nous présentons uniquement les résultats pour le domaine du cinéma en considérant les valeurs moyennes (sur les indicateurs) pour les deux concepts "actor" et "scenario". Les résultats obtenus sur le domaine de la restauration sont similaires à ceux obtenus pour le domaine du cinéma.

3.3.1 Détermination de la taille des fenêtres dans la classe et dans l'anti-classe

L'objectif est d'étudier le "gain" (en informations) apporté, en fonction de la taille de la fenêtre utilisée dans l'apprentissage de la classe et de l'anti-classe. L'étude est bornée en considérant une taille de fenêtre variant de 1 à 7. La taille de la fenêtre dans la classe varie alors de $i = 1$ à $i = 7$, et la taille de fenêtre dans l'anti-classe de $j = 1$ à $j = 7$. Naturellement, plus la taille de la fenêtre sera grande, plus la taille du lexique sera conséquente ; mais ce qui nous intéresse ici c'est la rapidité à laquelle le lexique croît en fonction de la taille de fenêtre. Le gain est calculé en fonction de la taille moyenne des lexiques $T_{ws_{(i,j)}}$ obtenus pour chacun des critères pour une configuration $ws(i, j)$ (taille de la fenêtre dans la classe et dans l'anti-classe), tel que :

$$Gain(i, j) = T_{ws_{(i,j)}}/T_{ws_{(i+1,j)}} \tag{3.15}$$

La figure 3.13 montre comment évolue la taille du lexique (nombre de descripteurs qu'il contient) en fonction de la taille de la fenêtre considérée dans la classe et dans l'anti-classe lors de la phase d'apprentissage. Nous pouvons remarquer que plus la taille des fenêtres est importante, plus le lexique construit est conséquent.

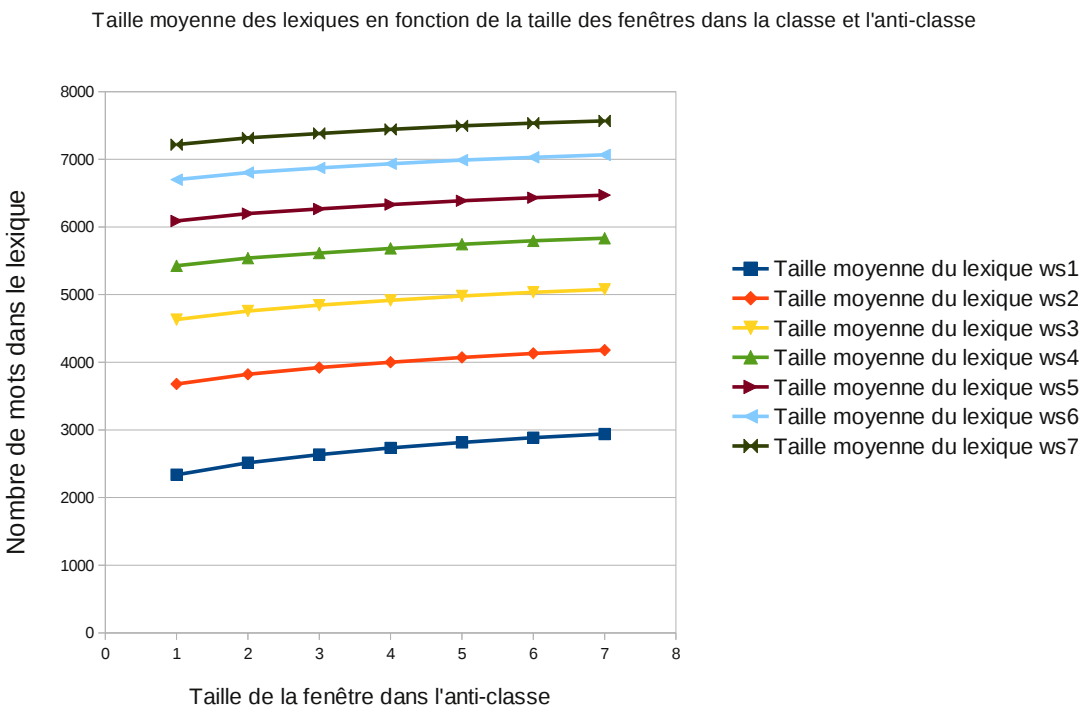

FIGURE 3.13 – Évolution de la taille du lexique en fonction de la taille de la fenêtre dans la classe et dans l'anti-classe

La figure 3.14 montre le résultat de l'expérience. Sur cette figure, chaque série de points (associée à un symbole géométrique sur la figure) représente le gain pour une valeur donnée de la taille de fenêtre de la classe. La taille de fenêtre de l'anti-classe est en abscisse. Nous constatons que le gain maximum, c'est-à-dire la variation de taille de lexique la plus grande, est obtenu lorsque l'on passe d'une fenêtre de taille 1 à 2 pour la classe (lecture en colonne des séries de la figure 3.14 : pour i=2 (colonne de gauche) par exemple, les gains sont 0.35, 0.21, 0.15, puis 0.12, 0.09, 0.07, les variations de gain sont donc de plus en plus faibles). Cette remarque reste valide quelle que soit la taille de la fenêtre de l'anti-classe (les 6 autres colonnes de la figure) : l'accroissement de la taille du lexique est toujours le plus conséquent pour

une taille de fenêtre passant de 1 à 2 pour la classe. Le gain varie finalement peu pour des tailles de fenêtre supérieures à deux. Ce résultat signifie que considérer des voisinages (longueurs de fenêtre) trop importants lors de l'apprentissage n'est pas forcément utile compte-tenu du peu de vocabulaire additionnel appris. Le gain apporté entre une fenêtre de taille 1 et une fenêtre de taille 2 dans la classe est le plus conséquent indépendamment de la taille de fenêtre de l'anti-classe : la variation "d'entropie" (qui est mesurée par la taille du lexique) est la plus significative pour ce changement de taille de fenêtre sur la classe.

Cette expérience montre qu'il n'est pas nécessaire d'évaluer l'ensemble des mots du langage par rapport au concept (taille de fenêtre infinie) pour obtenir suffisamment d'information pour le caractériser. **Une fenêtre de taille 2** dans la classe semble être la meilleure configuration pour notre méthode d'apprentissage (gain maximal en informations).

La taille de la fenêtre dans l'anti-classe reste cependant à déterminer. Pour cela, nous proposons d'étudier l'influence de la taille de la fenêtre dans l'anti-classe sur la qualité de l'apprentissage, en évaluant par les indicateurs de *Précision*, de *Rappel* et de *F1Score*, la qualité de l'extraction sur le corpus de test. L'expérience consiste donc à calculer ces indicateurs de qualité de classification en fonction de la taille de fenêtre de l'anti-classe sur le corpus test, c'est-à-dire sur un ensemble de documents indexés par les concepts. La figure 3.15 montre le résultat de l'expérience et met en évidence qu'**une fenêtre de taille 5** pour l'anti-classe donne les meilleurs résultats (FScore maximal et dispersion entre le rappel et la précision acceptable).

3.3.2 Influence de la taille de la fenêtre sur l'extraction

La taille de la fenêtre Fs est l'unique paramètre de l'algorithme d'extraction. Nous proposons de l'identifier en observant la dispersion des résultats de segmentation par rapport à la taille de la fenêtre choisie. Pour avoir des résultats significatifs, nous évaluons l'algorithme sur 98 lexiques différents. Ces lexiques ont été construits à partir de notre méthode d'apprentissage pour le domaine du cinéma : 49 lexiques sont relatifs au concept "actor", et 49 autres sont relatifs au concept "scenario". L'objectif est d'avoir des lexiques les plus différents possibles, pour être sûr que l'influence de la taille de la fenêtre d'extraction Fs dépende le moins possible des lexiques considérés. Pour assurer cette hétérogénéité sur les lexiques, nous proposons de faire varier la taille des fenêtres dans la classe et dans l'anti-classe de 1 à 7 ce qui explique les 49 (7×7) lexiques générés par mot germe

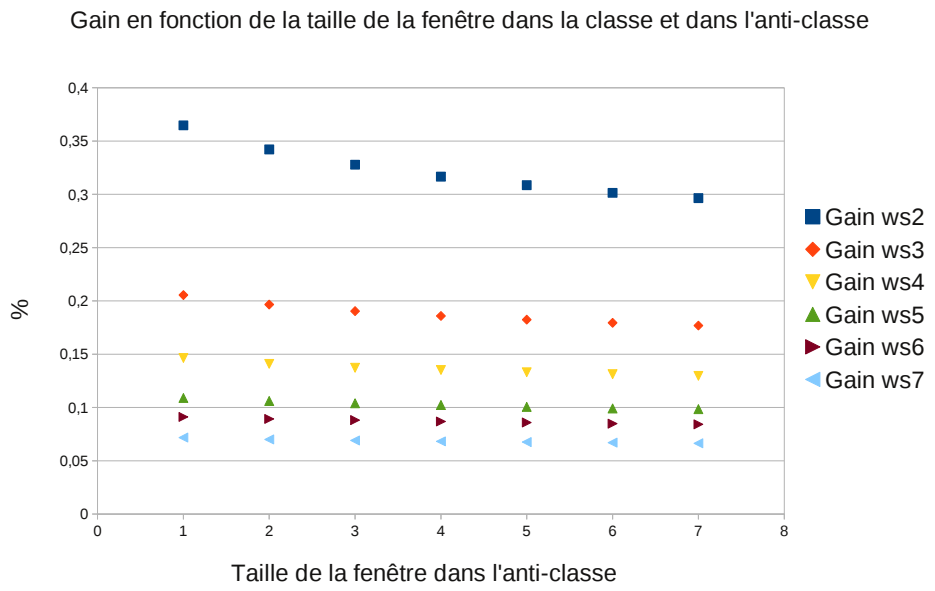

FIGURE 3.14 – Gain apporté en fonction de la taille des fenêtres dans la classe et dans l'anti-classe.

La figure 3.16 montre la dispersion des résultats obtenus (*F1Score* moyen pour les deux concepts "actor" et "scenario" sur les 98 lexiques considérés) pour chacune des tailles de fenêtre Fs (la taille variant de 1 à 7). Nous avons choisi de considérer le *F1Score* comme indicateur car il permet d'obtenir une valeur synthétique de la *Précision* et du *Rappel*. Nous pouvons remarquer qu'une fenêtre Fs de taille 2 minimise la dispersion (valeur max - valeur min) et maximise le *F1Score*. Par ailleurs, plus la taille de Fs augmente, plus la dispersion est importante, ce qui signifie que le système devient de plus en plus instable en donnant des résultats très aléatoires. Il devient alors difficile de quantifier la croyance en l'évaluation . C'est pourquoi, nous conserverons par la suite une **fenêtre Fs de taille 2 comme meilleure configuration**. Ce résultat confirme le choix intuitif de conserver la même fenêtre pour l'apprentissage de la classe et l'extraction de segments.

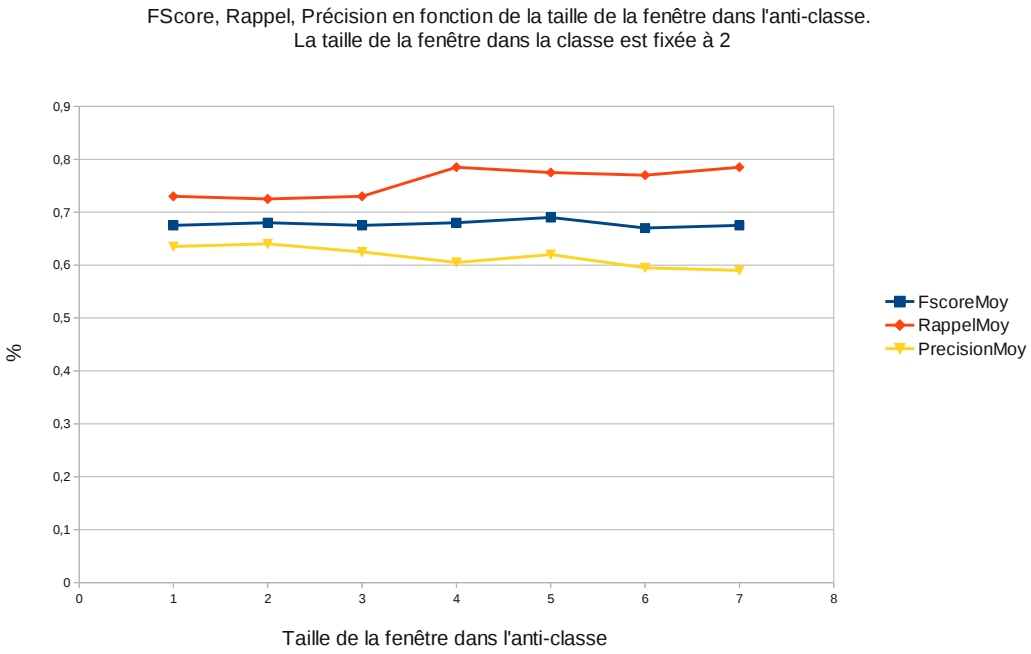

FIGURE 3.15 – FScore, Rappel et Précision en fonction de la taille de la fenêtre dans l'anti-classe pour une fenêtre de taille 2 dans la classe.

3.3.3 Influence du nombre de documents sur la qualité de l'apprentissage

Nous cherchons dans cette expérience à déterminer le nombre de documents nécessaires lors de la phase d'**apprentissage** des descripteurs. L'idée est de trouver le nombre minimal de documents qui suffisent pour construire un lexique pertinent avec une fenêtre de taille 2 pour l'apprentissage de la classe et de taille 5 pour l'anti-classe comme nous l'avons montré en section 3.3.1. Nous faisons varier le nombre de documents de 10 à 300 par pas de 10 documents doc_n^q dans chacun des ensembles de documents Doc^q. À partir des lexiques construits, nous étudions grâce aux indicateurs de *F1Score*, de *Précision* et de *Rappel* la qualité de l'extraction. La figure 3.17 montre qu'à partir de 100 documents, le système continu d'apprendre

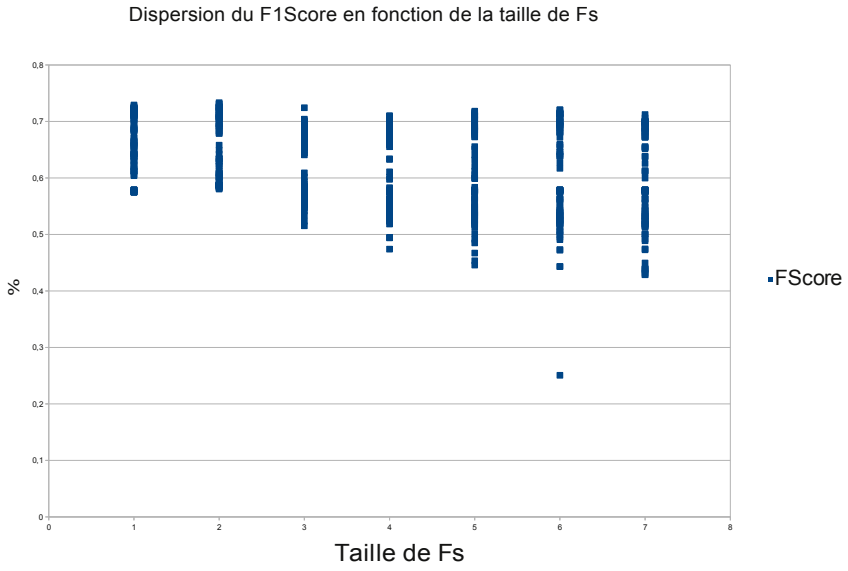

FIGURE 3.16 – Dispersion des résultats de segmentation (F1Score) en fonction de la taille de fenêtre FS

des descripteurs, mais que ceux-ci n'influent plus sur la qualité de l'extraction. Nous pouvons remarquer que c'est lorsque l'on considère environ 60 documents que le système affiche les meilleurs résultats, cependant, la différence entre un apprentissage avec 60 documents et un apprentissage avec 100 documents n'est pas significative pour considérer qu'il s'agisse d'un point caractéristique. De plus, le système n'est pas stabilisé sur la plage de 10 à 100 documents, c'est pourquoi, nous considérons qu'il faut un minimum de 100 documents pour obtenir suffisamment d'information pour construire un lexique significatif.

3.3.4 Nombre de mots germes nécessaires à l'apprentissage d'un concept

Pour amorcer l'apprentissage et construire le corpus d'apprentissage, il est nécessaire de définir des "mots germes". Ces mots germes permettent de caractériser le concept recherché. Nous proposons de déterminer le nombre minimal de mots germes nécessaires pour être en mesure de caractériser un concept. Nous utilisons pour l'apprentissage une fenêtre de taille 2 pour la classe, et une fenêtre de taille 5 pour l'anti-classe (c.f. section 3.3.1). L'idée est d'étudier l'influence du nombre de

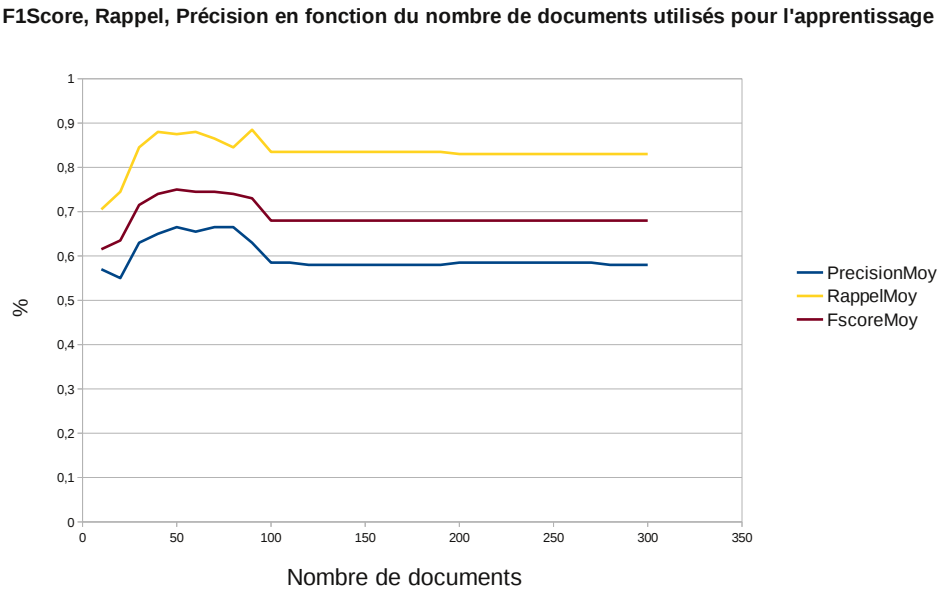

FIGURE 3.17 – F1Score moyen, Rappel moyen et Précision moyenne en fonction du nombre de documents considérés pour l'apprentissage

mots germes (le nombre de mots germes variant de 1 à 7) par rapport à la qualité de l'extraction. Nous construisons alors 7 lexiques pour chacun des concepts "actor" et "scenario", en considérant un mot germe pour le premier lexique, deux mots germes pour le second lexique, et ainsi de suite. À partir des lexiques construits, nous évaluons la qualité de l'extraction en fonction du nombre de mots germes considérés. L'évaluation utilise les indicateurs moyens (moyenne sur les deux concepts considérés) de *Rappel*, de *Précision* et de *F1Score*. L'expérience met en évidence une phase transitoire qui se situe entre une fenêtre de taille 1 et une fenêtre de taille 4 (c.f. figure 3.18). Ce "régime transitoire" du classifieur montre une instabilité du système due à l'imprécision de l'information, le lexique est trop pauvre en vocabulaire, ou alors beaucoup trop général et non spécifique. Nous pouvons donc conclure qu'il est nécessaire de définir au **minimum 4 mots germes** pour être en mesure de caractériser un concept.

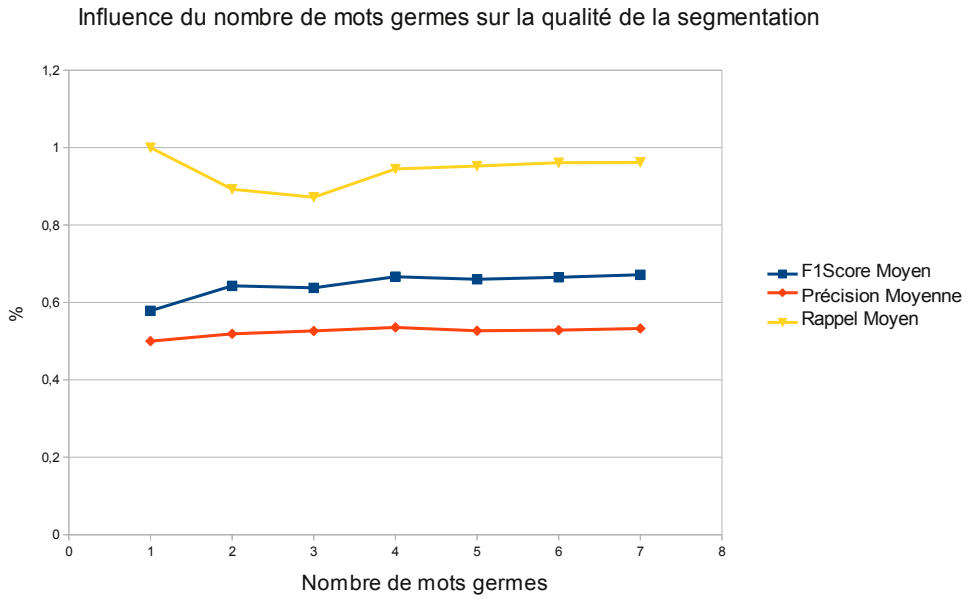

FIGURE 3.18 – Influence du nombre de mots germes par rapport à la qualité de l'extraction

3.3.5 Intérêt d'intégrer l'influence lors de l'apprentissage

Nous montrons dans cette section l'intérêt de considérer la gaussienne de lissage h dans l'influence $I(m, g_j^q, sz, doc)$ lors de la phase d'apprentissage. (c.f. section 3.4). Nous considérons ici une fenêtre de taille 2 dans la classe et une fenêtre de taille 5 dans l'anti-classe pour réaliser l'apprentissage. L'idée est de comparer la qualité de l'extraction en fonction du lexique considéré : un lexique construit avec un lissage de l'influence (h est un filtre gaussien), et un autre lexique construit sans lissage de l'influence (h est l'identité).

Le tableau 3.4 met en évidence l'intérêt de considérer l'influence filtrée, entre les mots lors de la phase d'apprentissage. Nous pouvons remarquer que le système est, en moyenne pour tous les concepts, plus précis : augmentation d'environ 10% de la précision avec le lissage.

	Influence=**Gauss**	Influence=**1**
Précision	**0.64**	0.55
Rappel	**0.78**	0.74
F1Score	**0.70**	0.63

TABLE 3.4 –

3.3.6 Évaluation de l'extraction

Cette section a pour objectifs de montrer la pertinence de notre approche. L'apprentissage des lexiques est effectué en utilisant le moteur de recherche *Google*[4] pour construire les corpus d'apprentissage.
Le paramètre *th* correspondant au niveau de granularité est choisi automatiquement par l'algorithme. Celui-ci est identifié en observant pour chaque *th* potentiel celui pour lequel la segmentation proposée est la plus en adéquation avec celle proposée par l'expert ayant évalué le corpus de test. Nous proposons d'évaluer le système sur chacun des critères des deux domaines d'étude proposés. De plus, nous évaluons le système avant, et après l'enrichissement des lexiques par l'apprentissage des mots candidats pour ainsi montrer l'intérêt de ce traitement particulier. Le tableau 3.5 montre que le traitement des mots candidats augmente de plus de 20% le *F1Score* (22% pour le concept acteur), et de plus de 30% la précision (31% pour le concept acteur). Le traitement particulier appliqué aux mots candidats est donc indispensable pour obtenir une précision acceptable. L'apport sur le critère *scénario* est moins important, mais reste cependant significatif.

	Sans mots candidats		**Avec mots candidats**	
	acteur	***scénario***	***acteur***	***scénario***
Précision	0.56	0.64	**0.87**	**0.67**
Rappel	0.84	0.86	**0.90**	**0.88**
F1Score	0.67	0.74	**0.89**	**0.77**

TABLE 3.5 – Résultats obtenus montrant l'intérêt du traitement particulier effectué sur les mots candidats

Nous avons testé notre algorithme dans le domaine "restauration", en considérant deux critères : "propreté" et "service". Les résultats sont présentés dans le tableau 3.6. **Ces résultats n'ont pas fait l'objet d'un traitement sur les mots**

4. http://www.google.fr

candidats.

	propreté	*service*
Précision	0,52	0,61
Rappel	0,93	0,94
F1Score	0,67	0,74

TABLE 3.6 – Performance de l'algorithme dans le domaine "restauration" pour les critères "propreté" et "service".

.

3.4 Discussion

Dans ce chapitre, nous avons proposé une méthode appelée *Synopsis* qui permet d'identifier les extraits d'un texte qui traitent d'un concept. Cette méthode non supervisée permet, à partir d'un nombre restreint de mots germes définis par l'utilisateur, de caractériser le concept recherché. Ces mots germes permettent d'amorcer l'apprentissage de descripteurs du concept. Nous avons montré que l'apprentissage automatique de ces descripteurs était un réel problème de par leur répartition (polysémie, etc.), et qu'il était nécessaire de considérer l'anti-classe du concept. De plus, nous nous sommes attaqués au problème du pragmatisme du langage, c'est-à-dire à l'interprétation que chaque individu peut avoir à la lecture d'un même texte. Nous pensons que cette dimension est capitale pour prétendre à une caractérisation réelle d'un concept par le langage.

Les principes d'apprentissage mis en avant sont issus des sciences cognitives et sont adaptés ici à l'apprentissage automatique de la langue par la machine. L'objectif dans ce chapitre était d'établir une méthode statistique qui permette d'apprendre et d'identifier des concepts dans un texte, tout en se rapprochant le plus possible des principes d'apprentissage du langage chez l'enfant. L'objectif étant de pouvoir extraire et interpréter la langue comme le ferait l'esprit humain, nous avons proposé de formaliser l'interprétation sous la forme : $interpretation(texte) = f(Se, Sy, Pr, texte)$. La dimension syntaxique est gérée par *TreeTagger* comme lemmatiseur et analyseur morphosyntaxique, les fenêtres d'apprentissage et d'extraction portent la notion de voisinage ; la dimension sémantique est liée à la définition du contexte par les mots germes associés au concept ; la dimension pragmatique est appréhendée par le biais d'une précision d'apprentissage paramétrable. Les

mots germes font référence à l'idée qu'il existe un niveau objectif de description du concept, la précision paramétrable du lexique appris est liée à la perception subjective de l'individu (son expertise). L'interprétation des données lors d'extraction de connaissances est directement liée au pragmatisme de l'information, et il est nécessaire de considérer cette dimension souvent évoquée mais rarement étudiée dans la littérature alors qu'elle est pourtant riche de sens et rend compte souvent des nuances du discours.
La définition des mots germes est l'unique expertise à fournir au système pour réussir à caractériser un concept, la qualité de l'apprentissage en dépend. Une telle approche nous permet de définir assez facilement un concept, et donc d'imaginer des applications pour une plus grande classe d'utilisateurs.

Pour améliorer la qualité de l'extraction, il serait envisageable d'enrichir le vocabulaire (le lexique de descripteurs) par la mise en place d'un apprentissage continu. L'idée serait d'apprendre tous les mots inconnus du système au cours des processus de segmentation. Ainsi, le système pourrait enrichir son lexique et ainsi raffiner sa segmentation. Le nombre de mots d'une langue étant fini, le système devrait converger vers un lexique comportant tous les mots d'une langue, évalués par rapport au concept considéré. Dans l'absolu, chaque concept potentiel serait caractérisé par un lexique qui contiendrait tous les mots de la langue mais avec des scores propres qui le caractériseraient pour établir simplement un modèle vectoriel.

L'apprentissage d'un concept aboutit à la construction d'un lexique de descripteurs. Ce lexique permet d'avoir une liste de mots dont le score tient compte des trois dimensions énoncées par Morris [Morris 1938]. Cependant, ce score n'est pas interprétable directement, bien qu'il donne une idée de la proximité avec le concept considéré. En effet, il est nécessaire de replacer les descripteurs dans leur contexte, c'est-à-dire que nous partons de l'hypothèse qu'un descripteur n'a pas de sens sans les descripteurs qui l'entourent (notion de contexte). Cependant, nous pensons qu'il existe un modèle permettant d'établir des relations logiques entre chacun des descripteurs qui serait en mesure de "modéliser le langage", et ainsi obtenir une structure hiérarchisée où la dimension pragmatique dépendrait du système cognitif de l'individu qui utiliserait cette ressource. À partir de cette ressource, il serait peut-être envisageable de trouver des liens avec les ontologies ou les thésaurus, et ainsi pouvoir prétendre à la mise en place de systèmes permettant l'évolution, l'enrichissement, ou pourquoi pas l'alignement d'ontologies ?

Les approches conceptuelles (ressources ontologiques) permettent de s'affranchir

de ce type de problématique et ainsi d'avoir une vision "universelle, ou synthétique" d'une idée commune. Le langage est imprécis, à l'image de l'esprit humain qui le manipule, ce qui complexifie grandement la tâche d'apprentissage et d'identification. À l'aube du web sémantique, nous pensons que cette approche est une solution convaincante pour, notamment, simplifier les tâches d'annotation et d'extraction de connaissances. Donner un sens aux données pour qu'elles soient interprétables par la machine est un vrai problème, et ces tâches sont, dans la majorité des cas, effectuées par l'Homme. Cela serait une première étape pour aider l'Homme dans ces tâches cognitives complexes. De plus la vision conceptuelle n'est pas directement interprétable par l'homme, qu'il doit plutôt jouer sur les nuances de la langue pour exprimer précisément ce qu'il a à dire, ce qui est un réel problème lors des phases d'interaction entre l'Homme et un système de recherche d'information par exemple, qui traite uniquement des données conceptuelles. De même, il n'est pas intuitif pour un être humain de pouvoir émettre une requête avec des concepts. Les méthodes terminologiques ont dû mal à gérer l'incertitude du langage (manque de précision), ce qui n'est pas le cas des méthodes conceptuelles. En revanche, les systèmes terminologiques sont intuitifs par rapport aux approches conceptuelles, mais difficiles à modéliser, c'est ici une de nos contribution.

Chapitre 4

Extraction d'opinion

"***Nous souhaitons la vérité, et ne trouvons en nous qu'incertitude.***"
Blaise Pascal

"***Le mensonge et la crédulité s'accouplent et engendrent l'Opinion.***"
Paul Valéry

Sommaire

4.1 Extraction d'opinion

Dans le chapitre 2 nous avons montré que la majorité des approches d'extraction d'opinion se basent sur la détection de termes explicitant directement une appréciation comme, par exemple : bon, agréable, excellent, mauvais, méchant, brutal, etc. Malheureusement, on se rend compte qu'une extraction d'opinion avec ces seuls termes explicites n'est pas suffisante pour assurer un résultat satisfaisant : l'expression d'opinion est au moins en partie propre au contexte ("Ce garçon est tout de même un bon gars" et "*Avatar* est un bon film", mais aussi au domaine "l'autofocus de cet appareil photo est parfaitement silencieux" et "ce député est resté parfaitement silencieux tout au long de l'assemblée". Ceci rend les approches purement syntaxiques inefficaces. Il est donc essentiel d'identifier les expressions et concepts propres au contexte ou à un domaine d'utilisation [Harb *et al.* 2008]. De plus, la définition des termes d'opinion est, dans la plupart des approches, généralement faite directement par un expert ou par des connaissances extérieures (ontologie, etc) ou par un apprentissage supervisé (construction manuelle du corpus d'apprentissage). Nous pensons que ces approches nécessitent une expertise trop importante pour être mises en place dans un contexte web où les domaines peuvent être multiples, ce qui supposerait la construction d'un corpus d'apprentissage par domaine. Il est donc nécessaire de développer des méthodes qui ne nécessitent que peu ou pas d'expertise. L'objectif de ce chapitre est de proposer une approche d'extraction d'opinion peu-supervisée qui s'adapte à différents domaines d'utilisation (cinéma, restaturant, politique, etc) tel que cela est décrit dans les travaux de [Harb *et al.* 2008]. Nous proposons d'adapter les techniques et méthodes utilisées dans le chapitre 3 à l'extraction d'opinion, et d'améliorer la méthode proposée par [Harb *et al.* 2008], notamment en supprimant les interventions humaines nécessaires lors de la phase d'extraction (définitions des quantificateurs, etc) en apprenant, non seulement les adjectifs, mais aussi les adverbes et les expressions (adverbe+adjectifs) ce qui nous permettra de nous affranchir d'une expertise coûteuse.

Dans le chapitre précédent, nous avons pu voir que *Synopsis* était adaptée à l'apprentissage des critères [Duthil *et al.* 2011a]. Nous proposons d'utiliser la même philosophie pour extraire de l'opinion. Cependant, dans l'approche *Sysnopsis*, deux classes étaient considérées : la classe et l'anti-classe. Cette distinction permettait d'identifier le concept et son "anti-concept". Dans ce chapitre, le problème n'est plus bipolaire puisque nous ne cherchons pas à caractériser un concept mais plutôt quadripolaire : l'opinion positive (respectivement l'opinion non positive) et l'opinion

négative (respectivement l'opinion non négative). Il est donc nécessaire d'adapter la méthode à ce contexte en définissant quatre classes : positif, non-positif, négatif, non-négatif. Ce chapitre est structuré de la manière suivante. D'abord nous précisions les hypothèses de notre approche éclairée par les notions de concepts opposés de la sémiotique. Puis, nous détaillons notre approche de la constitution du corpus d'apprentissage à l'attribution d'une polarité à un segment de texte. Nous montrons que notre solution au problème de la détection d'opinion n'est autre qu'une déclinaison de la méthode de segmentation thématique du chapitre précédent. Enfin, des expérimentations concluent ce chapitre.

4.1.0.1 Construction d'une opinion

La construction d'une opinion est un processus cognitif complexe qui peut être décomposé en deux étapes principales : la première est une phase d'apprentissage qui consiste à comparer les situations du vécu pour en établir un référentiel (c.f. figure 4.1) permettant de rattacher à chacune des situations des émotions (notamment différencier "le bien" du "mal"). Ce processus est comparable à une approche de clustering qui aurait pour objectif d'identifier des groupes de situations distincts relatifs à des sentiments. Sur la base de son référentiel de situations vécues, l'individu construit son ressenti, ses émotions, ses sentiments. C'est à partir de ce moment qu'il lui est possible d'exprimer une opinion en utilisant ses émotions, de les expliquer par son vécu. Cette opinion est donc directement liée aux sentiments éprouvés par l'individu. Par ailleurs, la construction d'une opinion ne repose pas seulement sur les sentiments directs, les sentiments sont aussi le résultat d'influences des réseaux sociaux dans lequel évolue l'individu. Les débats entre individus sont souvent au cœur de la prise de position et d'émission d'opinions. C'est lors de cette phase que l'opinion peut être renforcée, affaiblie, confirmée ou infirmée. Les argumentaires vont chercher à provoquer des émotions, et ainsi inciter l'individu à rattacher des émotions à la situation, au concept mis en cause dans l'objectif de le convaincre.

4.1.0.2 L'expression d'opinions

Pour résumer, plusieurs éléments sont nécessaires à la construction d'opinions. Nous pouvons identifier l'environnement dans lequel est plongé le sujet, son vécu, son mode d'expression (langage), sa culture. Ce sont ces quatre facteurs qui permettent à l'individu de construire et d'exprimer une opinion. La vision présentée en section 4.1.0.1 illustrée par la figure, 4.1 montre que l'esprit humain, lors d'une prise

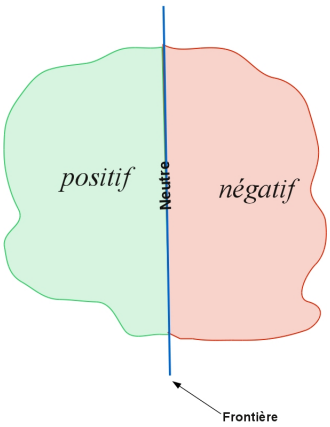

FIGURE 4.1 – Frontière référentiel

de position pourrait être comparé à une machine à états avec trois états potentiels : positif, négatif ou neutre. Cette vision est suffisante si l'on se place au niveau décisionnel (il faut choisir une action), mais elle est insuffisante lorsque l'on s'intéresse à l'expression d'une opinion par le langage qui peut être plus ou moins nuancée. Par exemple, un terme polysémique peut être positif dans un cas, négatif dans un autre et peut parfois être neutre. Les figures 4.2 et 4.3 illustrent cette problématique et montrent que la langue utilise des mécanismes plus complexes que dans la vision précédemment considérée et qu'un même mot peut être présent dans quatre ensembles de mots à la fois selon le contexte : les mots positifs, les mots négatifs, les mots non-positifs et les mots non-négatifs. Les mots peuvent donc être vus comme des objets sémiotiques [Piaget & Inhelder 1967] et, d'un point de vue cognitif, il est possible de se référer à un modèle logique représenté par le *carré sémiotique* (c.f. figure 4.3) de [Ermine 1989]. Cette vision permet d'introduire des règles entre les différents groupes de mots : les axes horizontaux expriment l'*opposition* (exclusion), les axes diagonaux expriment la *contradiction* (proposition tautologique), les axes verticaux expriment la *généralisation* (proposition). Par exemple, on peut *opposer* le concept du "bien" au concept du "mal ", ce qui précise le contexte du discours (la morale par exemple), mais le *contraire* du bien n'est pas pour autant le mal, le "non bien" est une *généralisation* du mal qui inclut le "ni mal", "ni bien ", nuance le manichéisme naïf. Si l'on raisonne sur un ensemble fini de concepts, par exemple $\{bien, mal, neutre\}$, l'opposé de $\{bien\}$, est $\{mal\}$, mais le contraire de *bien* appartient à la paire $\{mal, neutre\}$ qui est une généralisation de $\{mal\}$. Les figures 4.2 et 4.3 illustrent ces relations sémiotiques. Piaget a utilisé le support cognitif de la

sémiotique et ses représentations formelles pour expliquer le processus de l'apprentissage chez l'enfant. La première figure illustre ces règles par rapport à un nuage de termes, la seconde figure met en évidence la partie logique entre les différents ensembles de mots précédemment introduits. Ceci met en évidence que tout ce qui n'est pas positif n'est pas nécessairement négatif. Réciproquement, tout ce qui n'est pas négatif n'est pas nécessairement positif. Cette constatation complexifie l'apprentissage, car le problème ne se limite plus à considérer deux catégories de termes, mais plutôt à en apprendre quatre : les termes positifs, les négatifs, les non-positifs, les non-négatifs.

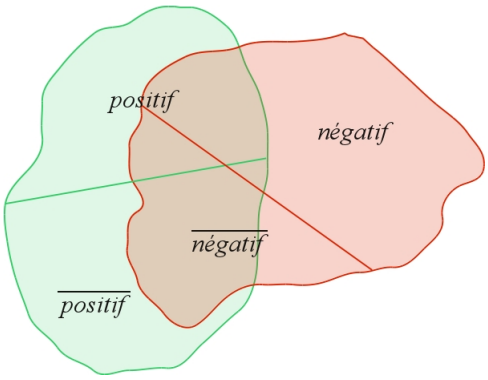

FIGURE 4.2 – Frontière référentiel

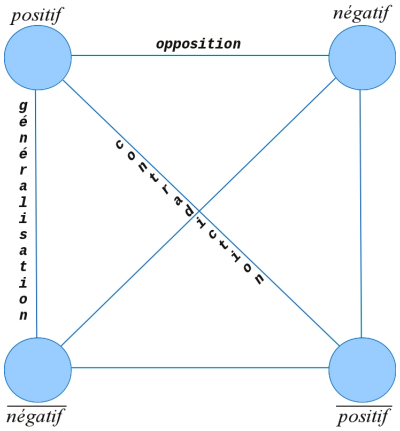

FIGURE 4.3 – Carré sémiotique de Piaget

4.1.0.3 Proposition et hypothèse

L'apprentissage des mots d'opinion est une tâche complexe qui fait appel à de nombreux mécanismes cognitifs d'un niveau conceptuel élevé (pragmatisme), et qui utilise des stimulus complexes comme l'émotion, les sentiments, etc. Ce manuscrit ne s'intéresse pas directement au processus psychologique, mais tente de donner une dimension sémantique évoluée à la machine lors de l'analyse d'opinions en appliquant les mécanismes d'apprentissage et d'analyse observés chez l'Homme (cf section précédente), en opposant ce qui est jugé bien à ce qui est jugé mal, en distinguant que tout ce qui n'est pas jugé bien, n'est pas forcément jugé mal. Notre étude se limite à l'apprentissage du langage par la machine par des méthodes statistiques, c'est pourquoi seront considérés par la suite uniquement les mots porteurs d'opinions, appelés *descripteurs*. Les dimensions relatives au vécu, à l'environnement ne seront pas abor-

dés directement. En effet, notre objectif ici n'est pas de reproduire fidèlement les mécanismes d'apprentissage de l'Homme (construction du référentiel et expression d'opinion), mais plutôt de reconstruire un référentiel à partir d'opinions déjà établies en étudiant, statistiquement, leur probabilité d'être positive (respectivement négative). Nous proposons un apprentissage qui nous permet, dans un premier temps, d'étudier la répartition des termes d'opinion (c.f. Figure 4.2) pour pouvoir, dans un second temps, se rapporter au processus décisionnel de la figure 4.1.
Nous avons montré dans de précédents travaux [Harb *et al.* 2008] que les mots d'opinions sont spécifiques à la thématique du texte (domaine), et peuvent représenter des opinions radicalement différentes selon ce contexte : considérons par exemple, les deux phrases "*The picture quality of this camera is hight*" et "*The ceilings of the building are high*". La première phrase (e.g. une opinion exprimée sur un film), l'adjectif *high* est positif, en revanche, dans la seconde (e.g. un document sur l'architecture), l'adjectif est neutre. Il est donc primordial d'apprendre les descripteurs sur une thématique précise. D'autre part, nous nous plaçons dans un contexte d'apprentissage différent de celui adopté par l'enfant dans la construction de son ressenti puisque nous reconstruisons un référentiel à partir d'opinions fondées, à l'inverse de l'enfant qui s'éduque à partir de facteurs environnementaux (vécu, culture, etc.). Pour pouvoir apprendre un maximum de points de vue (manière d'exprimer une opinion), il est nécessaire d'étudier une variété importante de documents, c'est pourquoi nous considérerons des documents web pour assurer cette diversité.
Le référentiel établi par l'enfant lui permet de pouvoir déterminer l'orientation sémantique d'un mot, d'une expression, dans un contexte particulier. Lors d'un tel processus décisionnel, l'individu est capable d'interpréter les mots ou expressions d'opinion avec une certitude qui lui est propre : c'est ici la dimension pragmatique. En effet, selon le nombre de confrontations à un descripteur, l'individu a une certaine croyance sur l'orientation sémantique du descripteur considéré. C'est pourquoi, nous proposons une méthode statistique pour réaliser notre apprentissage et ainsi étudier l'orientation sémantique (positif/négatif) d'un descripteur. Cette approche nous permet de nous prononcer sur l'orientation sémantique d'un descripteur avec une certaine croyance (statistique).

4.1.1 Présentation de l'approche

L'objectif de l'approche est de pouvoir classer des textes en fonction de leur orientation sémantique (positif/négatif) (c.f. figure 4.4) :. Nous nous intéressons à

deux types de classification. D'une part, la classification binaire qui consiste à attribuer une opinion globale à un document (critère de synthèse unique), et d'autre part la classification multicritère, qui, à partir d'un ensemble de critères prédéfinis attribue une note d'opinion relative à chacun d'entre eux. L'approche est basée sur une technique d'apprentissage permettant, à partir d'un nombre restreint de mots génériques d'opinion, de construire un lexique de descripteurs d'opinion positifs/négatifs pour ensuite calculer un score d'opinion relatif à un document, ou à segment de ce document. Les mots génériques expriment une opinion unique quelle que soit la thématique dans laquelle ils sont employés. Ces mots sont propres à chaque langue et sont utilisés pour amorcer un apprentissage des descripteurs. Ils permettent d'assurer l'orientation sémantique des mots *environnants* i.e. les mots qui sont souvent trouvés dans leur voisinage. Nous partons de l'hypothèse que plus un mot est proche d'un descripteur générique dans une phrase, plus il est probable qu'il ait la même orientation sémantique. Nous proposons d'utiliser les mots génériques positifs P et négatifs N établis par [Turney & Littman 2002] :
$P = \{good, nice, excellent, positive, fortunate, correct, superior\}$;
$N = \{bad, nasty, poor, negative, unfortunate, wrong, inferior\}$;
Dans cette section, nous présentons l'approche dans sa globalité, en expliquant les objectifs de chacune des trois étapes du processus :

- La première étape consiste à construire automatiquement un corpus d'apprentissage le plus riche possible (contenant des documents variés) et adapté à l'apprentissage.
- La seconde étape consiste à construire un lexique de descripteurs d'opinion en utilisant une approche statistique.
- La troisième étape consiste à calculer un score d'opinion pour un document en utilisant le lexique construit à l'étape précédente.

4.1.2 Constitution du corpus d'apprentissage

Comme nous l'avons vu dans le chapitre précédent, la constitution d'un corpus d'apprentissage est une étape délicate dans un processus d'apprentissage. Ce corpus doit être relatif à un domaine, c'est pourquoi nous exigeons que le document contienne au moins une occurrence du domaine D : ici *"movie"*. D'autre part, le corpus doit posséder des opinions positives et négatives. Pour assurer la présence d'opinions dans ces textes, nous utilisons des critiques du domaine (textes où l'on exprime nécessairement une opinion) et nous assurons l'orientation sémantique de

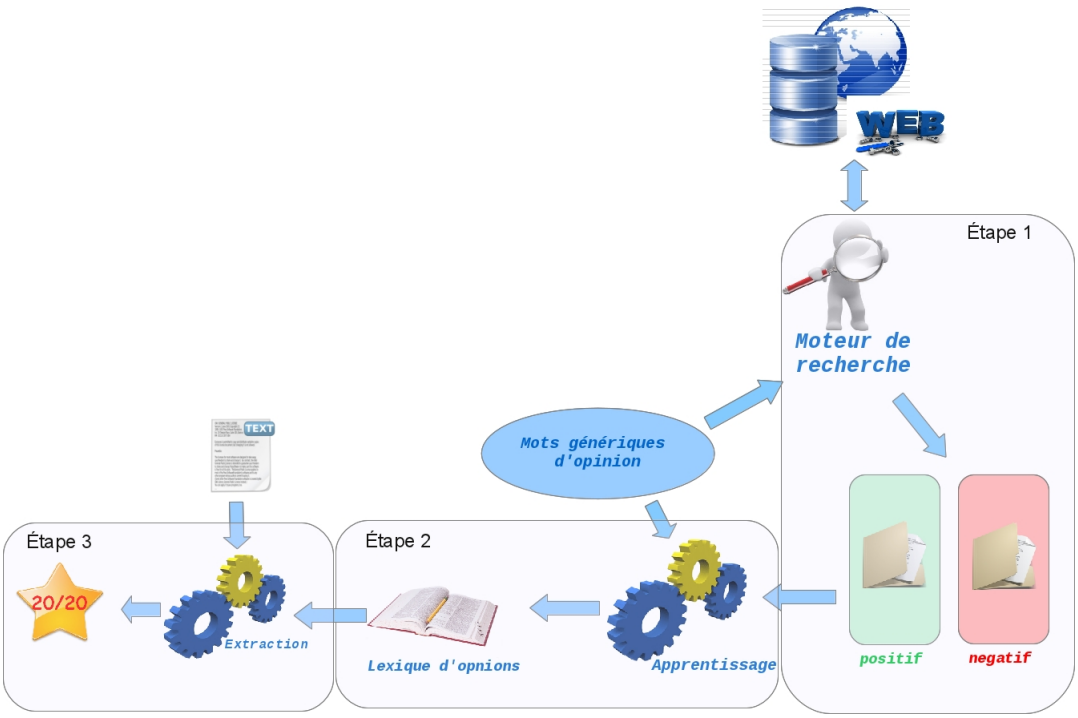

FIGURE 4.4 – Présentation de l'approche.

chacune par la présence des mots génériques : pour être positif, le document doit contenir au moins une occurrence d'un mot générique positif de P et aucun mot générique négatif de N (respectivement pour être négatif, le document doit contenir au moins une occurrence d'un mot générique négatif de N et aucun mot générique positif de P).
Pour obtenir des documents web vérifiant de telles propriétés, nous utilisons un moteur de recherche web en utilisant le domaine D et les mots génériques de N et P.

Exemple 4 *La requête suivante exprimée via le moteur de recherche Google nous permet d'obtenir des documents positifs relatifs au mot générique good :* $+opinion+review+movie+good-bad-nasty-negative-poor-unfortunate-wrong-inferior$.

Ainsi, pour chaque mot générique g de l'ensemble P (respectivement N), et pour un domaine D donné, nous collectons automatiquement K documents d'opinion relatifs à D (*movie*) et qui contiennent l'opinion exprimée par le mot germe $g \in P \cup N$, soit 14 corpus S_g de K documents (c.f. Figure 4.5). Nous définissons alors $Doc^P = \bigcup_{g \in P} S_g$ et $Doc^N = \bigcup_{g \in N} S_g$. Doc^P (respectivement Doc^N) est le corpus

de textes obtenu à partir des mots génériques positifs (respectivement négatifs).

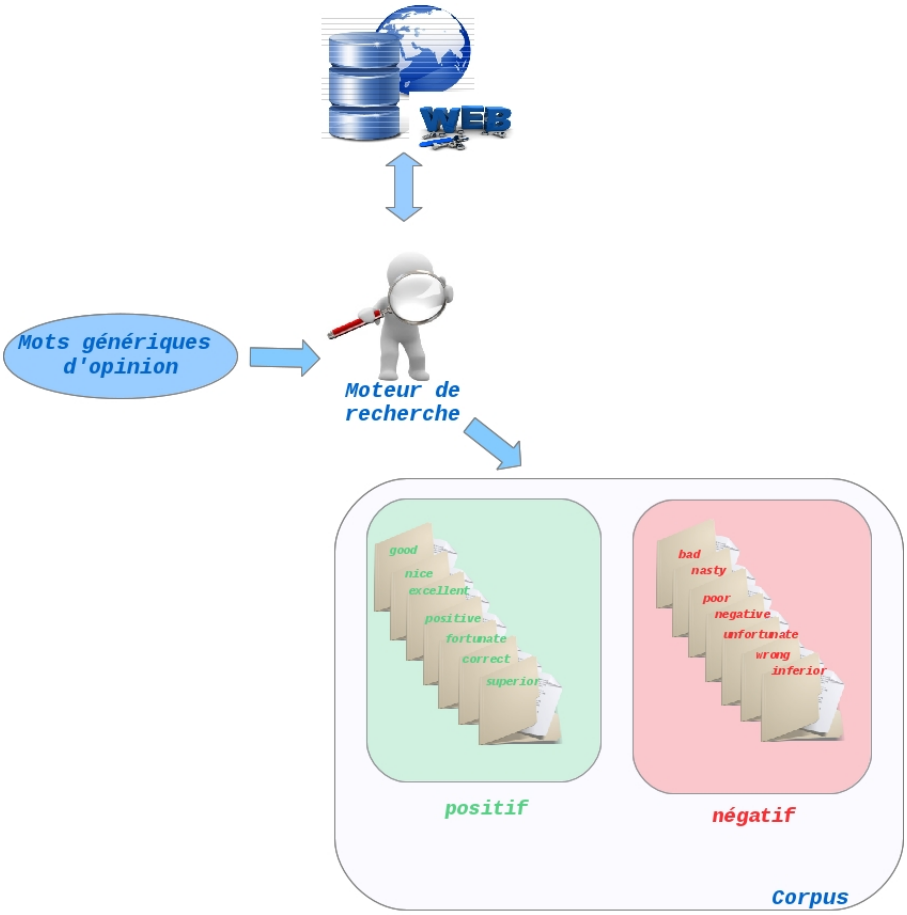

FIGURE 4.5 – Constitution du corpus d'apprentissage.

4.1.3 Apprentissage des descripteurs d'opinion

Il s'agit dans cette section de présenter le processus d'apprentissage des descripteurs d'opinion relatifs au domaine D. L'objectif est de rechercher dans les corpus construits les descripteurs d'opinions : adjectifs et expressions spécifiques au domaine porteurs d'opinion. Nous appelons "expression" la concaténation des adverbes précédant un adjectif.

Exemple 5 *Exemple d'expressions : "the ridiculously uneducated", "all bad", "very very good", "very nice", "simply not good", "so very good".*

L'intérêt de considérer les "expressions" est de ne pas avoir à effectuer un traitement particulier de la langue pour notamment gérer la négation qui est susceptible

de pondérer, voire d'inverser l'orientation sémantique de l'adjectif auquel est rattaché l'adverbe (ou autre modulation). Par exemple, l'expression "not good" n'est pas une expression très négative, et peut aussi être vue comme une opinion plutôt positive. D'une part, les traitements automatiques de la langue peuvent être lourds à mettre en place car ils sont basés sur des règles complexes, d'autre part, l'efficacité de tels traitements comportent des lacunes, notamment sur la gestion de la polysémie, c'est pourquoi nous avons choisi une méthode basée sur une analyse statistique de descripteurs potentiellement complexes.

4.1.3.1 Fenêtrage

Le but de l'apprentissage est d'enrichir les ensembles de mots génériques avec des descripteurs qui ont une orientation sémantique proche de celle des mots génériques. Nous considérons que l'opinion exprimée par des descripteurs proches d'un mot générique est similaire à celle exprimée par ce dernier (ou du moins de même orientation sémantique (positif/négatif)). Ainsi plus un descripteur est corrélé, *i.e*, il apparaît souvent dans le voisinage d'un mot générique, plus il est vraisemblable qu'il ait la même orientation sémantique que le mot germe auquel il est associé statistiquement. De la même manière, nous considérons les mots "éloignés" (loin de tout mot générique) comme non pertinents pour le descripteur générique considéré. Pour assurer un apprentissage tenant compte de la notion de proximité précédemment énoncée pour le calcul de corrélation, nous proposons comme dans *Synopsis* d'utiliser une fenêtre F_{op} de taille sz centrée sur chacun des mots génériques g d'un document t appartenant à S_g (les documents porteurs d'opinion rattachés au germe g) par :

$$Fop(g, sz, t) = \{m \in t / d^t_{JJ}(g, m) \leq sz\} \tag{4.1}$$

où $d^t_{JJ}(g, m)$ est la distance correspondant au nombre d'adjectifs (JJ) séparant le mot m de g.

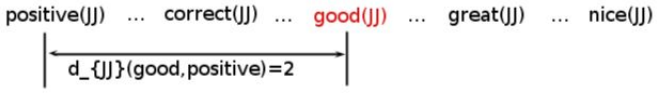

FIGURE 4.6 – Exemple de fenêtre de taille 2

Exemple 6 *La Figure 4.6 illustre un exemple de fenêtre de taille 2, i.e., il y a 2 adjectifs à gauche (**positive**, **correct**) et 2 adjectifs à droite (**great**, **nice**) du germe*

good. *Les pointillés entre les adjectifs symbolisent tous les autres mots présents d'un type grammatical autre qu'adjectif (i.e., noms, déterminants, adverbes, verbes, expressions, etc).*

4.1.3.2 Représentativité

De par la complexité de leur répartition, les descripteurs doivent être appris en deux temps pour pouvoir se rapporter à l'apprentissage mis en place dans *Synopsis*. L'objectif est d'arriver dans un premier temps à étudier la répartition des descripteurs dans les quatre ensembles positifs/non-positifs et négatifs/non-négatifs (c.f. Figure 4.2), puis, dans un second temps, se rapporter à un processus décisionnel proche de celui de l'Homme [Schärlig 1985] en construisant un lexique de descripteurs positifs/négatifs (c.f. Figure 4.1).
À chaque mot générique g est associé un ensemble de mots qui lui sont fortement corrélés, appelé *classe* X_g de g et un ensemble de mots qui ne sont pas corrélés, appelé *anti-classe* $\overline{X}$ de g. Cette technique de discrimination permet comme dans *Synopsis* de supprimer les descripteurs non pertinents, en étudiant la corrélation de chacun d'eux dans la classe et dans l'anti-classe (fréquence). Si un descripteur est plus fréquent dans la classe que de l'anti-classe $\overline{X}$, il est considéré comme pertinent et de même orientation sémantique que les mots de la classe X. À l'inverse un descripteur plus fréquent dans l'anti-classe que dans la classe est considéré comme non pertinent avec une orientation sémantique opposée. Si le descripteur est corrélé de la même manière à la classe qu'à l'anti-classe, il est alors considéré comme non pertinent. Nous définissons alors quatre ensembles de mots :

- L'ensemble des termes positifs : $X_P = \bigcup_{g \in P} X_g$.
- L'ensemble des termes négatifs : $X_N = \bigcup_{g \in N} X_g$.
- L'ensemble des termes non-positifs : $\overline{X}_P = \bigcup_{g \in P} \overline{X}_g$.
- L'ensemble des termes non-négatifs : $\overline{X}_N = \bigcup_{g \in N} \overline{X}_g$.

Nous proposons, pour étudier la corrélation d'un mot par rapport à un mot générique, d'étudier la fréquence d'apparition $\rho(M)$ dans chacune des fenêtres centrées sur g de taille sz, et qui appartiennent à Doc_g^P (respectivement Doc_g^N). Pour tous les descripteurs M une fréquence $\rho(M)$ dans la classe X_P (respectivement X_N) est calculée. Nous définissons le nombre d'occurrences $\mathcal{O}(w,t)$ d'un descripteur w (qui peut être un adjectif, une expression ou un mot générique) dans un texte t ou une

partie du texte t. $\rho(M)$ est définie par :

$$\rho(M) = \sum_{g \in \mathcal{P} \cup \mathcal{N}} \sum_{t \in S_g} \sum_{\gamma \in \mathcal{O}(g,t)} |\mathcal{O}(M, Fop(\gamma, sz, t))| \tag{4.2}$$

$\rho(M)$ correspond à l'action de cumuler les fréquences d'un descripteur présent dans toutes les fenêtres de tous les textes appartenant à Doc_g^P (resp. Doc_g^N). De la même manière, nous calculons la corrélation des descripteurs appartenant à $\overline{X}_P$ (resp. $\overline{X}_N$) par rapport à un mot germe g en utilisant une fenêtre de taille sz, ce sont les descripteurs situés en dehors des fenêtres qui sont comptabilisés. Nous calculons leur fréquence d'apparition $\overline{\rho}(M)$ à l'extérieur des fenêtres des textes appartenant à Doc_g^P (resp. Doc_g^N). Nous définissons ces (anti)fenêtres par : $\overline{F}op(g, sz, t) = \{m \in t / d_{JJ}^t(g, m) > sz\}$. $\overline{\rho}(M)$ est définie par :

$$\overline{\rho}(M) = \sum_{g \in \mathcal{P} \cup \mathcal{N}} \sum_{t \in \mathcal{T}(g)} |\mathcal{O}(M, \bigcap_{\gamma \in \mathcal{O}(g,t)} \overline{F}op(\gamma, sz, t))| \tag{4.3}$$

	Positif ($g \in Doc_g^P$)			**Négatif** ($g \in Doc_g^N$)	
M	$\rho(M)$	$\overline{\rho}(M)$		$\rho(M)$	$\overline{\rho}(M)$
good	1467	53	wrong	509	20
very nice	832	61	incorrect	190	35
dramatic	0	157	not important	12	87

TABLE 4.1 – Exemples de résultats pour Doc_g^P et Doc_g^N

Le tableau 4.1 est un exemple des fréquences obtenues pour six descripteurs dans chacun des quatre ensembles. Ces résultats, nous donnent déjà certaines indications sur l'orientation sémantique des descripteurs. Il est désormais possible de pouvoir déterminer l'orientation sémantique d'un descripteur à l'un ou à l'autre des deux pôles.

4.1.3.3 Calcul du score d'opinion

Chacun des descripteurs est classé en utilisant une fonction de discrimination f permettant d'établir, pour chacun des descripteurs, un score d'opinion tenant compte de sa représentativité dans chacune des classes. Cette fonction est identique à celle utilisée dans le chapitre 3. Le score que nous calculons nous permet, d'une part d'obtenir la probabilité qu'un descripteur appartienne à l'une des quatre classes, et

d'autre part de connaître son orientation sémantique. Ce score d'opinion noté $Sc(M)$ est alors calculé comme suit :

$$Sc(M) = f^2(\rho(M), \overline{\rho}(M)) \tag{4.4}$$

où f est définie telle que :

$$f^2(x, y) = \frac{(x-y)^3}{(x+y)^2} \tag{4.5}$$

Positif ($g \in Doc_g^P$)		**Négatif** ($g \in Doc_g^N$)	
	Sc(M)		Sc(M)
good	1414	wrong	489
very nice	771	incorrect	155
dramatic	-157	not important	-75

TABLE 4.2 – Exemples de résultats pour Doc_g^P et Doc_g^N après le calcul de $Sc(M)$

Le tableau 4.2 montre la répartition de six descripteurs. Nous en déduisons l'orientation sémantique de chacun d'eux à partir de leur score : si $Sc(M) > 0$ alors M appartient à la classe, si $Sc(M) < 0$ alors M appartient à l'anti-classe. Nous pouvons donc déduire que les descripteurs "good" et "very nice" ont une orientation sémantique positive car $Sc(good)$ et $Sc(very\ nice)$ sont positifs. En revanche, le descripteur "dramatic" n'a pas une orientation sémantique positive car $Sc(dramatic)$ est négatif, il appartient donc à l'anti-classe des positifs. De la même manière, les descripteurs "wrong" et "incorrect" ont une orientation sémantique négative, car $Sc(wrong)$ et $Sc(incorrect)$ sont positifs. En revanche, le descripteur "not important" n'a pas une orientation sémantique négative car $Sc(not\ important) < 0$. Ces résultats correspondent bien à l'intuition que l'adjectif *dramatic* et l'expression *not important* n'expriment pas nécessairement d'opinions et illustrent le principe du carré sémiotique.

Le fait qu'un descripteur ait un score $Sc(M) < 0$ ne donne donc pas l'orientation sémantique du descripteur. Considérons par exemple le descripteur *not important* qui a un score $Sc(not\ important)$ négatif. La seule information que nous donne le signe de Sc est que *not important* n'a pas une orientation sémantique négative, mais en aucun cas nous ne pouvons affirmer que ce descripteur a plutôt une orientation sémantique positive, et qu'il a potentiellement une orientation sémantique positive g. La figure 4.7 illustre cette classification des descripteurs dans leurs

classes (classe/anti-classe) respectives.

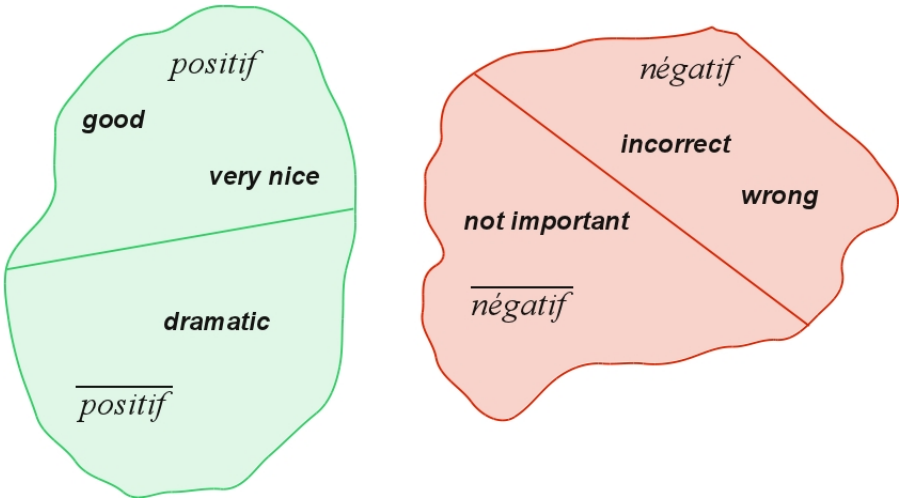

FIGURE 4.7 – Classification des descripteurs.

Nous proposons de ne conserver que les descripteurs avec une orientation sémantique ("good" et "very nice" pour les positifs, "wrong" et "incorrect" pour les négatifs), c'est-à-dire les descripteurs appartenant aux deux classes "polarisées" (les descripteurs ayant un score $Sc > 0$) comme l'illustre la figure 4.8 et ainsi pouvoir rassembler ces deux classes pour construire un lexique L_{Op} d'opinions.

Pour construire le lexique L_{Op} nous proposons de considérer deux classes de descripteurs : les descripteurs ayant un score $Sc(M)$ positifs dans la classe des positifs noté $Sc_P(M)$ et les descripteurs ayant un score $Sc(M)$ positifs dans la classe des négatifs noté $Sc_N(M)$. À partir de ces deux classes nous construisons un lexique L_{Op} tel que L_{Op} est l'ensemble des descripteurs ayant un score positif dans l'une des deux classes. Le score $Sc_{Lop}(M)$ d'un descripteur M dans le lexique est calculé par :

$$Sc_{Lop}(M) = Disc(Sc_P(M), Sc_N(M)) \tag{4.6}$$

avec $Disc(Sc_P(M), Sc_N(M))$ une fonction de discrimination permettant de calculer le score $Sc_{Lop}(M)$ d'un descripteur M dans le lexique L_{Op} tel que :

$$Disc(Sc_P(M), Sc_N(M)) = \frac{(Sc_P(M) - Sc_N(M))^3}{(Sc_P(M) + Sc_N(M))^2} \tag{4.7}$$

Le tableau 4.3 est un extrait du lexique d'opinion construit à partir des fréquences obtenues lors du calcul de $Sc(M)$ (c.f. Tableau 4.2). Nous pouvons remarquer que les descripteurs positifs "good" et "very nice" sont conservés et que le

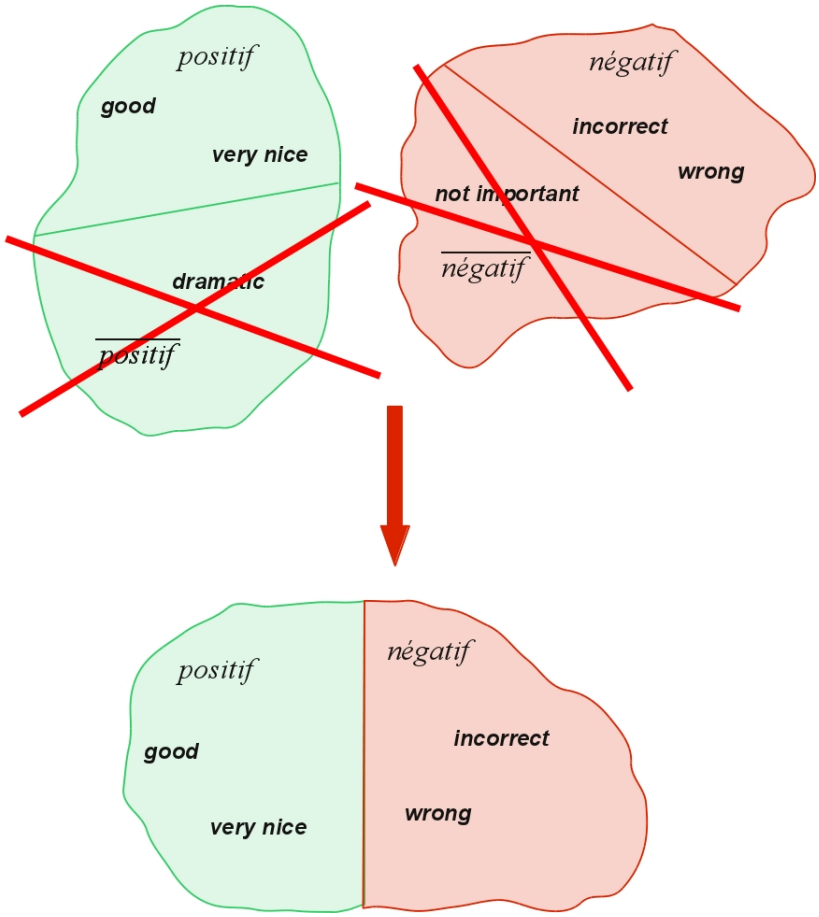

FIGURE 4.8 – Suppression des anti-classes.

Mot	Score(M)
very nice	771
good	93
incorrect	-155
wrong	-489

TABLE 4.3 – Extrait du lexique L_{Op}

descripteur "dramatic" a été supprimé car il n'est pas positif. Ceci est le résultat de la discrimination. De même, les descripteurs négatifs "wrong" et "incorrect" sont conservés, et le descripteur "not important" est supprimé car il n'est pas négatif. De cette manière, nous ne conservons que les mots discriminants dans chacune des classes (c.f. figure 4.8). L'objectif ensuite est d'utilisé le lexique L_{Op} pour déterminer l'opinion globale exprimée dans un document ou une partie de ce document.

4.1.4 Détection d'opinion

L'extraction d'opinion peut avoir deux objectifs : classer des textes selon l'opinion globale qui y est exprimée, ou attribuer des polarités partielles à un document relativement à une base de critères déterminée au préalable. La *classification de textes* d'opinion globale se limite à déterminer la polarité d'un document sans tenir compte de critères particuliers. Ce type de classification est suffisant pour des applications spécifiques où, notamment, les documents considérés traitent d'une seule thématique (un seul critère) où le sentiment global du texte suffit à l'analyse. L'extraction d'opinion multicritère est une forme raffinée de la classification de textes. Elle ne considère plus le texte dans sa globalité, mais plutôt l'opinion relative à chacun des critères qui peuvent y être évoqués. Ce type d'évaluation sur un espace multicritère est largement utilisé dans les systèmes de recommandations où l'évaluation par critère permet d'expliquer et de justifier un avis.

4.1.4.1 Classification de textes d'opinion

À partir du lexique d'opinion précédemment construit (c.f. section 4.1.3.3), nous cherchons à déterminer l'opinion qui se dégage d'un texte, ou d'une portion de texte, en rapport avec la thématique $\mathcal{T}$.
Pour un texte t, nous utilisons une notion de fenêtre glissante de taille sz successivement centrée sur chaque occurrence d'un adjectif dans le texte t. À partir des éléments du lexique, un score est calculé pour chacune des fenêtres Fop de la manière suivante :

$$ScoreOp(Fop) = \frac{1}{mc} \sum_{M \in Fop} Sc_{Lop}(M) \tag{4.8}$$

où mc est le nombre de mots ayant un score > 0, c'est-à-dire les mots ayant contribué (classe(s) grammaticale(s) apprise(s)) au score de $ScoreOp(Fop)$.

La polarité générale de t est alors déterminée par le signe de $Sc(t)$ défini par :

$$Sc(t) = \sum_{Fop \in t} ScoreOp(Fop) \tag{4.9}$$

Si $Sc(t) < 0$, le texte t sera considéré comme négatif, si $Sc(t) > 0$, t est positif et si $Sc(t) \simeq 0$, alors t n'exprime pas d'opinion, ou alors une opinion neutre

4.1.4.2 Extraction d'opinion multicritère

La classification considère que chaque mot porteur d'une opinion contribue au score attribué au document. Cette technique est efficace pour des documents traitant d'un seul sujet où les arguments qui y sont avancés concernent le même concept (critère). En revanche, lorsqu'il s'agit de textes où l'opinion exprimée repose sur plusieurs critères, la méthode précédente s'avère inadaptée. En effet, le fait de considérer l'opinion exprimée dans sa globalité, sans considérer les critères auxquels elle se rattache, peut masquer la divergence potentielle entre les scores relatifs aux critères : l'opinion sur un critère peut être excellente, tandis que sur un autre elle peut être catastrophique malgré une note globale plutôt bonne. Ce constat nous amène alors à considérer l'opinion par rapport à des critères pour pouvoir calculer un score d'opinion plus précis et plus représentatif de la réalité.

Pour réaliser cette opération, nous combinons l'approche *Synopsis* et la méthode d'extraction d'opinion présentée dans ce chapitre. Dans un premier temps, *Synopsis* extrait les segments de textes relatifs à chacun des critères désirés, puis, dans un second temps, nous calculons un score d'opinion sur chacun des extraits identifiés par *Synopsis* (le score est donc un score partiel relatif au critère du segment). Cette approche permet d'obtenir une classification binaire (positif/négatif) de textes par rapport à des critères. Dans cette approche, un texte est vu comme un vecteur de scores partiels (+ ou -) sur l'ensemble des critères. L'évaluation s'arrête là sans complément d'information sur l'importance relative de chacun des critères. Tous les critères sont a priori d'importance équivalente.Considérer que les critères ont la même importance relative suffit pour calculer un score d'opinion global sur des textes : on peut par exemple, de façon purement ordinale, simplement compter le ratio de critères positifs versus négatifs, ou encore ne s'intéresser qu'à un seul critère véto. Nous verrons comment intégrer à notre approche de détection d'opinion des outils pour l'identification de modèles de préférences afin de couvrir dans sa totalité la problématique de la recommandation multicritère (c.f. chapitre 6).

Par ailleurs, il nous semble nécessaire d'intégrer la représentativité ou l'intensité du critère dans le calcul de la polarité du texte. Par exemple, il nous paraît utile de distinguer le cas où un critère est très représenté dans un texte mais avec un score de polarité faible, du cas où le critère est faiblement représenté mais avec un score de polarité très élevé. Par exemple, l'opinion sur le critère "actor" est positive, mais le critère "acteur" n'est que très peu présent dans le texte (une phrase sur 100), de la même manière, l'opinion sur le critère "scénario" est négative, et le critère "scénario"

est présent dans 80 phrases. Cet exemple montre bien qu'il est nécessaire de considérer la représentativité du critère dans le texte pour obtenir une opinion plus juste. L'idée est donc de pondérer le score d'un texte en fonction de la représentativité du critère dans le texte. D'autre part, le fait d'utiliser *Synopsis* pour l'extraction de critères apporte une information sur le poids de chacun des critères considérés pour chacune des fenêtres Fop d'opinion d'un texte puisque le score calculé au chapitre précédent était assimilable à une intensité (plus le concept est présent, plus il est vraisemblable que le texte parle du critère). Par la suite, nous considérons qu'un extrait e de texte identifié par *Synopsis* peut être considéré comme un ensemble de fenêtres Fs du critère C^q ou comme un ensemble de fenêtres d'opinion Fop. Pour simplifier la notation, nous utiliserons e_c lorsque nous considérons la partition de l'extrait e en fenêtres thématiques (critère) et e_o lorsque nous considérons la partition de l'extrait e en fenêtres d'opinion. Nous pouvons remarquer que le nombre de fenêtres appartenant à e_o et e_c est imposé par les constructions respectives de chacun des deux types de fenêtre (opinion et critère). On notera n_e le nombre de fenêtres relatives au critère pour un extrait e, et n_{op} le nombre de fenêtres d'opinion du même extrait e.

Le poids $Imoy$ d'un extrait (e), pour un texte t est calculé en considérant les fenêtres construites lors de l'extraction de critères e_c et qui ont au moins un de leurs mots présent dans une des fenêtres d'opinion (e_o) contenues dans cet extrait. L'idée est de calculer, pour chaque fenêtre d'opinion, un poids $I(e, Fop)$, associé au critère analysé, qui caractérise la représentativité (l'intensité) du critère : le score des fenêtres utilisées lors de l'extraction conceptuelle du critère couvertes par une fenêtre d'opinion (recouvrement des partitions) va être utilisé pour pondérer la polarité par l'intensité d'un critère . $I(e, Fop)$ est la moyenne arithmétique des scores de chacune des fenêtres de critère ayant un mot en commun avec la fenêtre d'opinion Fop considérée (la fenêtre d'opinion peut recouvrir plusieurs fenêtres de critère), soit :

$$I(e, Fop) = \frac{\sum_{Fs \in e_c} Score(Fs).[Fs \wedge Fop]}{\sum_{Fs \in e_c} .[Fs \wedge Fop]} \tag{4.10}$$

où $Score(Fs)$ est le score de représentativité du critère calculé dans le chapitre précédent, et est une formule logique qui caractérise les fenêtres de critère ayant au moins un mot en commun avec la fenêtre d'opinion Fop, soit :

$$[Fs \wedge Fop] = \left[\begin{array}{l} 1 \textbf{ si } Fs \wedge Fop \neq 0 \\ 0 \textbf{ sinon} \end{array} \right. \tag{4.11}$$

Pour un extrait e, l'influence moyenne est ensuite définie par :

$$Imoy(e) = \frac{\sum_{Fop \in e_o} I(e, Fop)}{n_{op}} \tag{4.12}$$

Par ailleurs, un score moyen d'opinion $Scmoy$ est défini pour chaque extrait e par :

$$Scmoy(e) = \frac{\sum_{Fop \in e_o} ScoreOp(Fop)}{n_{op}} \tag{4.13}$$

Lorsqu'il y a plusieurs critères en jeu, il faut encore indicer $Scmoy$ et $Imoy$ par q.

Le calcul du "score d'opinion" $Sc^q(t)$ d'un texte par rapport à un critère C^q est basé sur un calcul d'aire qui considère les deux dimensions précédemment introduites : la représentativité ou intensité du critère $Imoy$ et le score d'opinion $Scmoy$. La figure 4.9 est un exemple de calcul d'aire pour un texte donné contenant quatre extraits relatifs à un critère C^q donné. L'aire de chaque carré est alors le score d'opinion relatif au critère C^q pour un extrait du texte t. Le score d'opinion $Sc^q(t)$ du texte t pour un critère C^q est enfin défini par :

$$Sc^q(t) = \frac{\sum_{e \in t} Sc^q moy(e) \times I^q moy(e)}{\mid e \in t \mid} \tag{4.14}$$

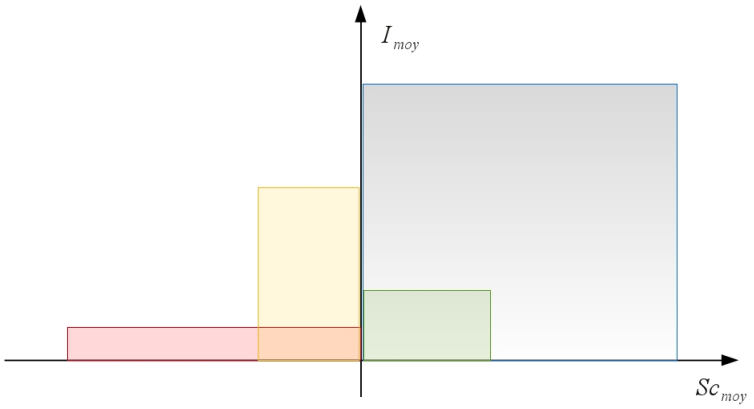

FIGURE 4.9 – Exemple de calcul d'aire pour un texte donné.

Nous avons montré dans le chapitre 3 que l'extraction conceptuelle (ici le concept est relatif à un critère) n'était pas unique, et que l'expression d'un concept dépend du point de vue de la personne. Un texte t peut alors être analysé avec des niveaux de granularité différents au moment de l'extraction conceptuelle. Ainsi, et compte tenu des calculs précédents, l'intensité du critère et donc son poids dans le calcul

de l'opinion va s'en trouver modifier, et par la suite son score d'opinion différera aussi en fonction de cette granularité. Le score d'opinion ne peut plus être considéré comme un précis, mais comme un intervalle de scores caractérisant l'imprécision relative à la polarité du texte. Nous proposons de considérer cette incertitude par un score d'opinion imprécis $[Sc^q](t)$ pour un critère C^q donné, en calculant l'intervalle des valeurs possibles tenant compte des différents niveaux de granularité relatifs au critère :

$$[Sc^q](t) = \left[\overset{min}{Sc^q}(t), \overset{max}{Sc^q}(t)\right] \tag{4.15}$$

$\overset{min}{Sc^q}(t)$ correspond au score minimal $Sc^q(t)$ obtenu en considérant toutes les granularités potentielles. De la même manière, $\overset{max}{Sc^q}(t)$ correspond au score maximal $Sc^q(t)$ obtenu en considérant toutes les granularités potentielles. Notons que $Sc^q(t)$ n'est pas une fonction monotone de la granularité ce qui nécessite de faire les calculs pour toutes les granularités pour trouver les valeurs minimales et maximales du score d'opinion.

4.2 Expérimentations et résultats

Pour valider notre approche de détection d'opinions nous avons choisi d'utiliser le corpus de tests proposé par [Pang *et al.* 2002] et comme thématique $\mathcal{T}$ le cinéma. Ce corpus est composé de critiques cinématographiques issues de `http://reviews.imdb.com/Reviews`. Chacun des textes est étiqueté comme positif ou négatif. Pour valider l'approche dans un contexte utilisant des critères de choix, nous avons repris le même corpus proposé par [Pang *et al.* 2002], nous en avons extrait les parties de texte concernant les critères "acteur" et "scénario", puis nous avons détecté l'opinion relative à chacun d'eux.
Pour comparer nos résultats, nous avons choisi un outil de détection d'opinions de la littérature : *SenticNet* [Cambria *et al.* 2010], qui propose une collection de concepts polarisés (positifs/négatifs) constituant un réseau sémantique. Un score $\in [-1, 1]$ est attribué à chaque concept : plus le score est proche de 1 (respectivement de -1) et plus le score est de polarité positive (respectivement de polarité négative). Un score moyen est établi pour chacune des phrases, et l'agrégation du score de chacune des phrases donne une polarité au texte.

Les paramètres de l'algorithme d'apprentissage sont identiques à ceux utilisés pour l'approche *Synopsis* et ont les mêmes justifications que celles évoquées au

chapitre 3.

4.2.1 Validation de l'approche en classification de textes

Nous proposons ici d'évaluer l'approche dans un contexte de classification de textes. Pour cela, nous utilisons le corpus annoté traitant de la thématique du cinéma proposé par [Pang *et al.* 2002]. Nous avons pour réaliser cette expérience, appris les descripteurs relatifs à la thématique du cinéma et ainsi pu calculer une opinion globale sur chacun des documents.

	Notre approche		**SenticNet**	
	positif	***négatif***	***positif***	***négatif***
FScore	**0,73**	**0,69**	0,68	0,37
précision	**0,68**	**0,75**	0,54	0,74
rappel	0,79	**0,63**	0,91	0,25

TABLE 4.4 – Résultats sur le corpus de [Pang *et al.* 2002] en classification de textes

Le tableau 4.4 met en évidence l'efficacité de l'approche dans un contexte de classification de documents. On remarque que le $FScore$ obtenu par notre approche est meilleur que celui obtenu par *SenticNet*. On observe une petite faiblesse de notre approche dans la détection de textes positifs mais qui reste meilleur que *SenticNet*, mais qui est largement compensée par la détection de textes négatifs (le rappel (0.63) est plus de deux fois plus important que celui de SenticNet(0.25)). Ces résultats mettent en évidence qu'un vocabulaire d'opinion non spécifique à une thématique pourrait suffire pour une évaluation globale d'un document. Nous pouvons constater que notre approche obtient globalement des résultats similaires à *SenticNet* pour la classification de textes (positifs/négatifs) et que le lexique appris est pertinent. Le seul apprentissage des adjectifs et expressions est donc suffisant pour réaliser cette tâche. Par ailleurs, il est important de souligner que notre méthode est très peu supervisée (définition des mots génériques) contrairement à *SenticNet* qui, elle, est entièrement supervisée.

4.2.2 Validation de l'approche sur deux critères de choix, ici les critères "*acteur*" et "*scénario*"

Nous proposons ici d'évaluer l'extraction d'opinions par rapport à des critères de choix. Pour cela, nous utilisons le corpus annoté traitant de la thématique du

cinéma proposé par [Pang *et al.* 2002]. Nous avons, pour réaliser cette expérience, appris les descripteurs relatifs à la thématique du cinéma. Nous avons choisi deux critères de choix en rapport avec cette thématique : *scénario* et *acteur*. Nous avons identifié pour chaque document du corpus annoté les parties de texte relatives à chacun des deux critères. Pour cela, nous utilisons *Synopsis* et nous choisissons le *th* qui maximise la taille du texte sélectionné, en pratique le premier état stable qui considère le point de vue le plus large. Sur chacun des extraits retenus, nous calculons l'opinion qui se dégage, puis nous en déduisons une opinion globale du document en agrégeant les opinions détectées sur chacun des extraits (moyenne) comme proposé dans ce chapitre. Nous obtenons ainsi une opinion globale d'un document par rapport à un critère de choix. Nous procédons de la même manière pour valider l'approche avec *SenticNet*, en identifiant dans un premier temps les parties de textes relatives à chacun des critères, puis nous calculons une opinion globale du document par rapport aux critères de choix.

	Critère *Acteur*			
	Notre approche		**SenticNet**	
	positif	***négatif***	***positif***	***négatif***
FScore	**0,83**	**0,80**	0,69	0,38
Précision	**0,76**	**0,90**	0,55	0,74
Rappel	**0,92**	**0,71**	0,91	0,26

TABLE 4.5 – Résultats sur le corpus de [Pang *et al.* 2002] en classification de textes sur le critère "*acteur*"

	Critère *Scénario*			
	Notre approche		**SenticNet**	
	positif	***négatif***	***positif***	***négatif***
FScore	**0,92**	**0,92**	0,70	0,50
précision	**0,90**	**0,95**	0,60	0,74
rappel	**0,95**	**0,90**	0,87	0,38

TABLE 4.6 – Résultats sur le corpus de [Pang *et al.* 2002] en classification de textes sur le critère "*scénario*"

Le tableau 4.6 montre que l'opinion exprimée par les internautes dans des critiques cinématographiques dépend de critères de choix. Ici, on remarque que le critère "scénario" correspondrait mieux au point de vue des personnes ayant rédigé

les critiques que le critère "acteur" pour lequel les résultats de la classification sont moindre (le classifieur avec le seul critère *scénario* retrouve très bien les polarités de la base indexée). Dans une approche multicritère de l'évaluation d'opinions, on pourrait par exemple en déduire que le critère "scénario" devrait avoir un poids important pour la rédaction d'un avis sur un film.
Les résultats du tableau 4.4 correspondent à un score d'opinion moyen (polarité globale sans considérer de critère particulier) et que, au vu des résultats, nous mettons en évidence que la méthode d'agrégation qu'a l'esprit humain ne se limite pas à une simple moyenne arithmétique. Nous pouvons remarquer que dans un contexte d'extraction d'opinions relatives à des critères, *SenticNet* est beaucoup moins efficace que notre méthode pour plusieurs raisons. Tout d'abord, le fait d'introduire des critères de choix a mis en évidence que le vocabulaire d'opinion est spécifique à la thématique. Nous pouvons en déduire que dans l'expérience précédente, les mots d'opinions spécifiques à la thématique étaient noyés par les autres mots d'opinion du langage courant (non spécifiques à la thématique), ce qui baissait significativement les performances du système. Enfin, la combinaison de notre approche de détection d'opinions à un processus d'extraction thématique non supervisé ne requiert qu'une expertise minimale contrairement à un processus supervisé comme *SenticNet*.

4.3 Discussion

Dans ce chapitre nous avons proposé une approche de fouille d'opinion qui permet d'extraire l'opinion exprimée dans des textes. Cette approche nécessite une phase d'apprentissage d'un vocabulaire d'opinion relatif à un domaine d'utilisation. En effet, nous avons montré que le vocabulaire d'opinion employé était propre au domaine dans lequel il était employé. L'objectif de l'apprentissage est la construction d'un référentiel, sous forme de lexique, qui permet de déterminer l'orientation sémantique d'un texte (ou d'un extrait).

Notre apprentissage se base sur l'utilisation de mots germes d'opinion génériques. Ces mot germes sont spécifiques à chaque langue, et ils sont définis une seule fois pour une langue donnée, tout domaine confondu. L'apprentissage consiste à enrichir le vocabulaire d'opinion à partir des mots germes qui permettent d'amorcer l'apprentissage. L'idée est d'observer, statistiquement, les mots porteurs d'opinions (adjectifs, adverbes et expressions) qui gravitent autour des mots germes. Nous supposons que ces mots ont la même orientation sémantique que le mot germe qu'ils

précisent et complètent.

Nous avons choisi de considérer, comme descripteurs d'opinions, uniquement les adjectifs et les adverbes, ainsi que les expressions "adverbes+adjectifs", car ces mots sont reconnus comme porteurs d'opinions [Turney & Littman 2002]. Cependant, il serait nécessaire de considérer aussi certains noms communs et certains verbes qui sont aussi porteurs d'opinion, mais cette considération, dans un contexte "non-supervisé", est assez difficile à mettre en œuvre, car cela reviendrait à détecter la subjectivité d'un verbe, ou d'un nom, et à déterminer son orientation sémantique. Cette problématique nécessiterait d'établir un certain nombre de règles sur la langue, et ainsi de se placer dans un contexte de traitement automatique du langage. De plus, nous nous sommes intéressés ici à des textes d'opinion (critiques), mais il serait intéressant de pouvoir adapter l'extraction d'opinion aux nouveaux médias comme les "SMS", les "tweets", etc, qui sont des textes courts et pour lesquels les descripteurs d'opinion sont différents de ceux considérés dans ce chapitre, nous pensons notamment aux "smileys".

Nous avons mis en évidence que pour apprendre les descripteurs relatifs à chaque pôle (orientation sémantique positive respectivement négative), il était nécessaire de considérer quatre classes de mots selon le principe de la sémiotique de Piaget [Piaget & Inhelder 1967] : les mots positifs, les mots non-positifs, les mots négatifs et les mots non-négatifs. Cette considération est un point clé de l'apprentissage (tout ce qui n'est pas bon, n'est pas forcément mauvais).
Nous avons montré que l'expression d'une opinion repose sur des critères de choix, et qu'il est nécessaire de les considérer pour avoir une opinion plus précise d'un texte, plutôt que de le considérer dans sa globalité. Les polarités par critère permettent de mieux comprendre les raisons du ressenti exprimé dans un texte.

Nous avons donc proposé de considérer ces deux dimensions : le critère et l'opinion relative. Nous utilisons *Synopsis* pour extraire les parties de textes traitant d'un critère et notre approche de détection d'opinion. En combinant les deux approches, nous sommes en mesure d'attribuer un score imprécis à un document sous la forme d'un intervalle qui tient compte à la fois de la dimension pragmatique du critère (paramétrée par le niveau de granularité de l'extraction conceptuelle), et de l'opinion exprimée. Cette combinaison des deux approches nous permet de mieux caractériser l'opinion exprimée dans le document, et ainsi de mieux comprendre comment le locuteur a construit son opinion. C'est ici la première étape, pour comprendre le système de préférences utilisé par le locuteur. Nous y reviendrons dans le chapitre

6 consacré à la recommandation multicritère.

Une extension possible à notre approche serait la détection de l'ironie. La détection de l'ironie est une problématique liée au traitement automatique du langage, et elle dépend notamment de la structure des phrases (ordre des mots, etc.), pour cela il faudrait intégrer au système proposé certaines règles (pattern) du langage afin de développer de nouvelle fonctionnalités.

Chapitre 5

Applications

"***La tristesse de l'intelligence artificielle est qu'elle est sans artifice, donc sans intelligence.***"

Jean Baudrillard

Sommaire

5.1 Introduction

Au cours des chapitres 3 et 4 nous avons présenté des approches théoriques permettant la segmentation thématique de textes d'une part et l'extraction d'opinions d'autres part. Nous proposons dans ce chapitre de présenter des prototypes qui reposent sur les algorithmes utilisés dans ces approches. Nous avons choisi de diviser ce

chapitre en deux sections où chacun des prototypes est plus particulièrement orienté vers l'une ou l'autre des deux approches théoriques. *Synopsis* et *Colexir* sont dédiés à la segmentation conceptuelle ; *ExOp* et ses dérivées apportent une solution à l'extraction d'opinion. La section 5.2 correspond au chapitre 3 sur la segmentation thématique de texte, la section 5.3 correspond au chapitre 4 sur l'extraction d'opinion. Les applications présentées ici ont toutes été développées en langage JAVA. De par leur conception, elles peuvent s'intégrer rapidement à des architectures existantes, qu'il s'agisse d'applications déportées, web, ou de bureautique.

5.2 Segmentation thématique

5.2.1 "Synopsis" un outil de segmentation de texte

Dans cette section nous proposons un prototype logiciel qui permet de segmenter des textes en fonction de concepts définis par l'utilisateur. Le logiciel est composé de deux modules : le premier concerne l'apprentissage des concepts définis par l'utilisateur : à chaque concept est associé (appris) un lexique de mots caractéristiques. Le second module concerne la segmentation de texte. Ce module utilise les lexiques précédemment construits pour segmenter des textes selon chacun des concepts appris (lexique associé). Aujourd'hui, un livre électronique ne peut que surligner dans le texte tous les passages dans lesquels se trouve le mot qu'un utilisateur recherche. Par exemple, si le mot recherché dans *Tristan et Iseult* est "amour ", alors toutes les phrases contenant explicitement "amour " seront mises en évidence. En revanche, une phrase comme "Lorsque Iseult rejoint Tristan dans la mort, un pied de vigne et un pied de lierre poussèrent enlacés sur les lieux du drame " sera laissée de côté par la fonctionnalité de recherche du livre électronique alors que dans notre approche l "'amour " sera considéré comme un concept et lors de l'apprentissage le lierre sera forcément associé au lexique de l'amour comme étant un symbole de l'étreinte amoureuse. Par conséquent, cette phrase sera proposée à l'utilisateur (ce qui au passage peut constituer une connaissance nouvelle pour lui s'il ne connaît pas cette symbolique). Nous avons vu que notre approche d'extraction conceptuelle permettait de tenir compte de l'expertise de l'utilisateur. La précision avec laquelle l'utilisateur souhaite extraire la thématique dépend évidemment de son expérience qui, en pratique ici, est modélisée par la taille du lexique associé au concept. L'application propose alors plusieurs modes de gestion de la précision (point de vue) de la segmentation : le premier mode une possibilité de segmentation, en conservant

uniquement le point de vue "le plus large" (peu précis puisqu'on filtre peu). Le second mode une interaction avec l'utilisateur qui peut choisir le point de vue (granularité) qu'il souhaite en lui proposant plusieurs points de vue, du plus spécifique (très précis donc avec un fort filtrage des extraits), au plus large (c.f. le curseur de la figure 5.2). Pour l'utilisateur, il est alors possible de faire une segmentation avec une précision donnée. Si le pourcentage de texte restitué est trop important, alors il peut augmenter la précision pour un filtrage plus sévère. Au contraire, s'il ne lui est quasiment rien retourné (silence) alors l'utilisateur diminuera la précision afin que le système renvoie un minimum de segments de texte.

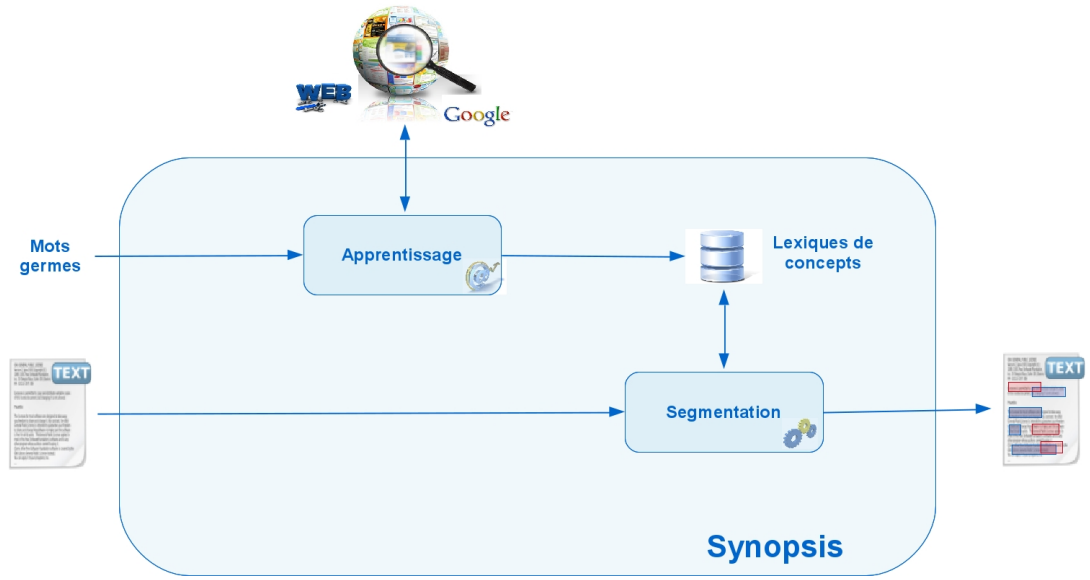

FIGURE 5.1 – Schéma global de l'application "Synopsis" de segmentation de texte.

La figure 5.1 présente l'application Synopsis dans sa globalité avec ses modules d'apprentissage et de segmentation. Le système prend en entrée un texte et les mots germes nécessaires à la définition du concept à identifier. L'apprentissage du concept est réalisé une seule fois. Le lexique construit est ensuite ajouté à la base. Le module de segmentation utilise le lexique relatif au concept recherché.

5.2.2 Utilisation de Synopsis dans un contexte de recherche d'information

Durant la dernière décennie, les ontologies ont pris une place prépondérante dans les modèles d'indexation de corpus et les techniques de requêtage. L'efficacité des

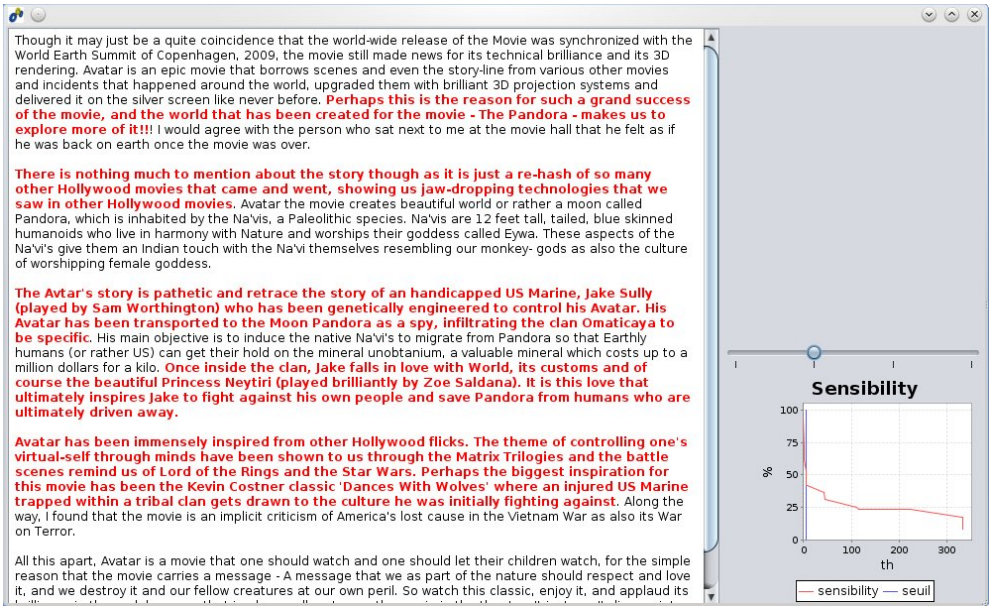

FIGURE 5.2 – Système "Synopsis" : exemple d'interaction avec l'utilisateur.

Systèmes de Recherche d'Information (SRI) à base d'ontologie est aujourd'hui reconnue. Ces SRI utilisent généralement une ontologie de domaine et un corpus de documents annotés par des concepts de l'ontologie. Ils ont pour objectif, à partir d'une requête (ici conceptuelle) fournie par l'utilisateur, de rechercher les documents qui correspondent le mieux à la requête émise. Cette recherche utilise une mesure de similarité (similarité entre un document et la requête) pour déterminer les documents les plus pertinents. Une liste ordonnée de documents est ensuite proposées à l'utilisateur.

Nous pensons qu'il est nécessaire de fournir à l'utilisateur plus qu'une liste de documents jugés pertinents par le système, et qu'il faut, d'une part, justifier la raison pour laquelle un document a été retourné pour que le SRI gagne en crédibilité vis à vis de l'utilisateur et d'autre part, aider l'utilisateur à accéder plus rapidement à l'information pertinente contenu dans le document. C'est pourquoi nous avons proposé de combiner un SRI nommé *OBIRS* avec *Synopsis* pour répondre aux problèmes de crédibilité et d'aide à l'utilisateur précédemment exposés. Cette combinaison permet, à la fois, de bénéficier de la précision du SRI, mais aussi, de l'identification des parties des documents retournés traitant des concepts spécifiés par l'utilisateur. Le système complet, nommé *CoLexIR* [Ranwez *et al.* 2012] permet de gagner la crédibilité de l'utilisateur en identifiant les parties des documents traitant des concepts voulus tout en l'aidant à accéder rapidement à l'information qu'il souhaite. Enfin,

cette fonctionnalité peut aussi être utilisée en reformulation de requête : en effet, l'accès direct à des passages pertinents dans les documents retournés peut permettre à l'utilisateur de repérer de nouveaux concepts qui lui paraissent au final pertinents à intégrer dans une nouvelle requête susceptible d'améliorer la restitution du SRI.

5.2.2.1 Principes

Pour être en mesure de segmenter des textes selon un concept, il est nécessaire de construire un lexique de descripteurs caractérisant le concept voulu. La construction d'un lexique nécessite qu'un corpus d'apprentissage soit formé et qu'un apprentissage soit réalisé. Ces tâches nécessitent la définition de mots germes. Dans notre contexte, nous disposons d'une ontologie du domaine médical appelé *Mesh* (Medical Subject Headings). Cette ontologie contient 25603 concepts pour la version 2010, qui sont organisés hiérarchiquement dans une structure de graphe acyclique direct (DAG) avec plusieurs héritages (relation "*is a*" entre les concepts). Chacun des concepts dispose d'un label qui est la désignation (un terme) du concept considéré. L'objectif final est d'être en mesure d'identifier dans des documents les parties de textes relatives à chacun des 25603 concepts de l'ontologie potentiellement présents. Il faut donc construire un lexique pour chacun des concepts de l'ontologie. La définition des mots germes nécessaires à la construction du lexique est effectuée automatiquement en considérant les labels des concepts proches du concept à caractériser. Nous avons choisi comme "concepts proches" les concepts fils, qui sont une spécialisation (relation "*is a*") du concept père (le concept à caractériser). Ce choix est certes discutable, mais les résultats obtenus sont pertinents. Bien sûr, il serait envisageable d'utiliser une mesure de similarité spécifique qui permettrait de conserver uniquement les "7 concepts les plus proches", qui ne sont pas nécessairement les fils du concept à caractériser. Le processus de constitution de corpus d'apprentissage et l'apprentissage sont coûteux en temps (environ 20 minutes pour apprendre un concept) aujourd'hui. Cette limite technique est due au nombre limite de requêtes imposées par le moteur de recherche *Google*, mais cette contrainte devient minime lorsque nous utilisons un corpus et un moteur de recherche local (1 minute). C'est pourquoi les lexiques relatifs à chacun des concepts sont construits au préalable. Il en est de même pour l'ensemble des documents dont dispose le système *CoLexIR*, ils sont segmentés au préalable pour chacun des concepts avec lesquels ils sont annotés, ceci par souci de rapidité du système.

5.2.2.2 Description

Le système *CoLexIR* est un système de Recherche d'Information qui combine une recherche conceptuelle et une ressource terminologique qui permet de répondre plus efficacement au besoin de l'utilisateur.

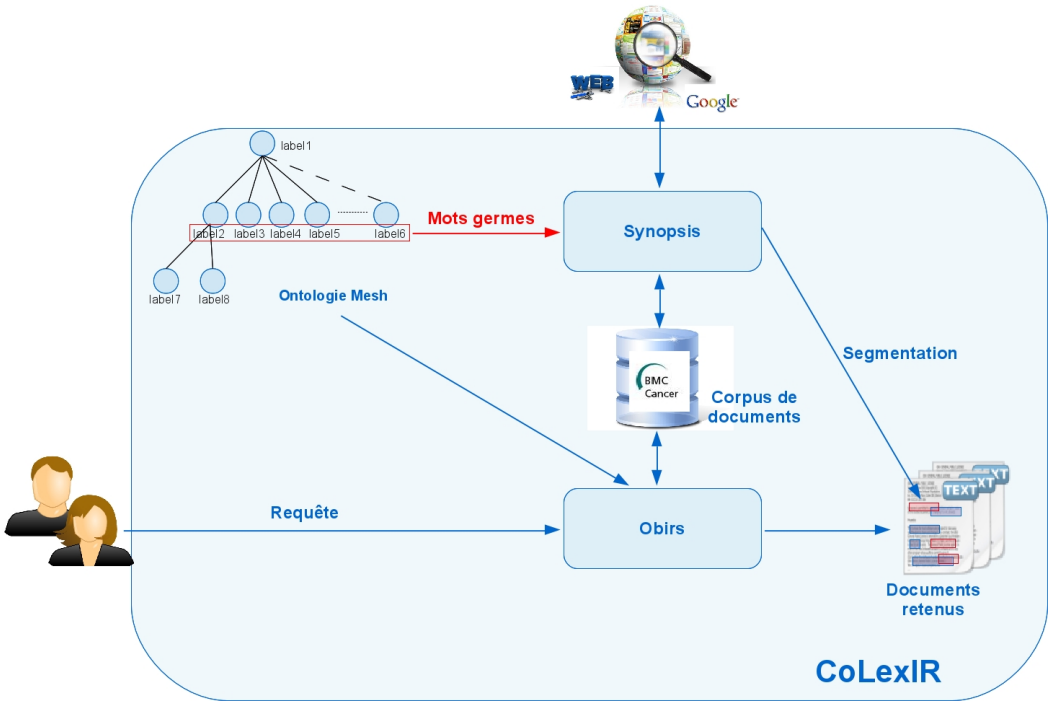

FIGURE 5.3 – Le système CoLexIR

La figure 5.3 illustre le système dans sa globalité. *CoLexIR* est un système en ligne accessible par tous les navigateurs disposant d'un "plugin flash player". Le prototype utilise aujourd'hui une seule base de documents qui correspond aux articles publiés dans la revue *BMC Cancer*.
La figure propose une copie d'écran du système *CoLexIR*. Les histogrammes de couleur indiquent la pertinence des documents (concept recherché présent dans le document, concept représenté par un fils, un père, etc.) ce qui permet à l'utilisateur de comprendre la sélection de *CoLexIR*,puis en fonction de ce premier niveau d'explication, de regarder la segmentation du document jugé opportun proposée par le logiciel afin de se rendre compte rapidement de la pertinence du document vis-à-vis de sa requête en mode "full-text " filtré.

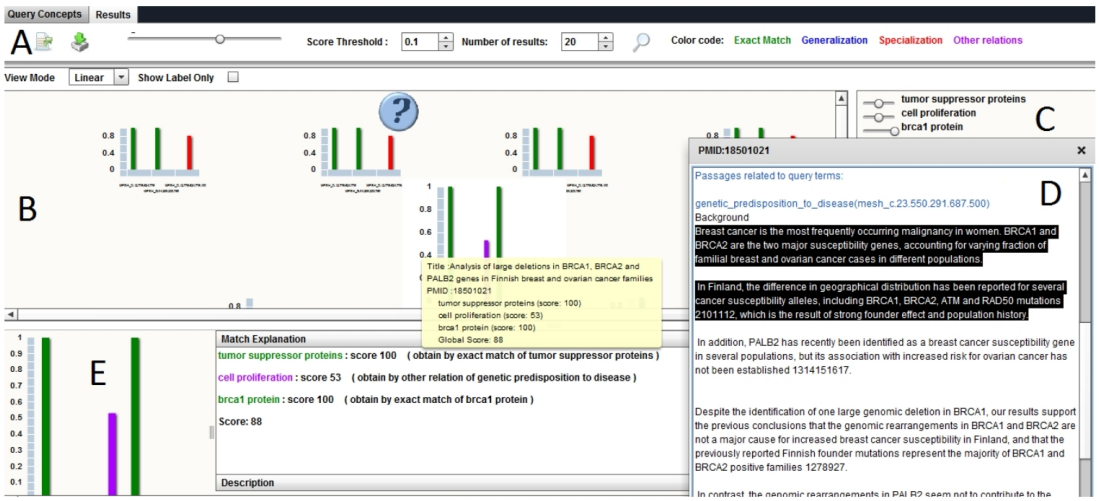

FIGURE 5.4 – Interface du système *CoLexIR*

5.3 Extraction d'opinions

Dans cette section nous proposons un prototype logiciel qui permet d'extraire l'opinion dans un texte. Les principes et algorithmes mis en place ont été décrits en détail dans le chapitre 4.

5.3.1 Classification de texte

Le système *ExOP* est un prototype logiciel qui permet d'annoter des textes selon la polarité de l'opinion qu'ils contiennent (positif/négatif).

ExOp a besoin, en entrée, de mots germes d'opinion (ici la langue choisie est l'anglais) pour créer un lexique d'opinion si celui-ci n'a pas déjà été construit. Le système est constitué de deux modules, le premier est un module qui permet la constitution d'un corpus d'apprentissage à partir de documents web et l'apprentissage des descripteurs d'opinion, le second module permet l'extraction d'opinion et l'annotation du texte par sa polarité. En sortie, le système fourni la polarité du texte fourni en entrée.

5.3.2 Classification de texte multicritère

Nous proposons dans cette section d'étendre le système *ExOp* à la problématique du multicritère. Nous avons montré dans le chapitre 4 que l'opinion exprimée dans un document repose sur des critères, c'est pourquoi nous proposons, plutôt que de

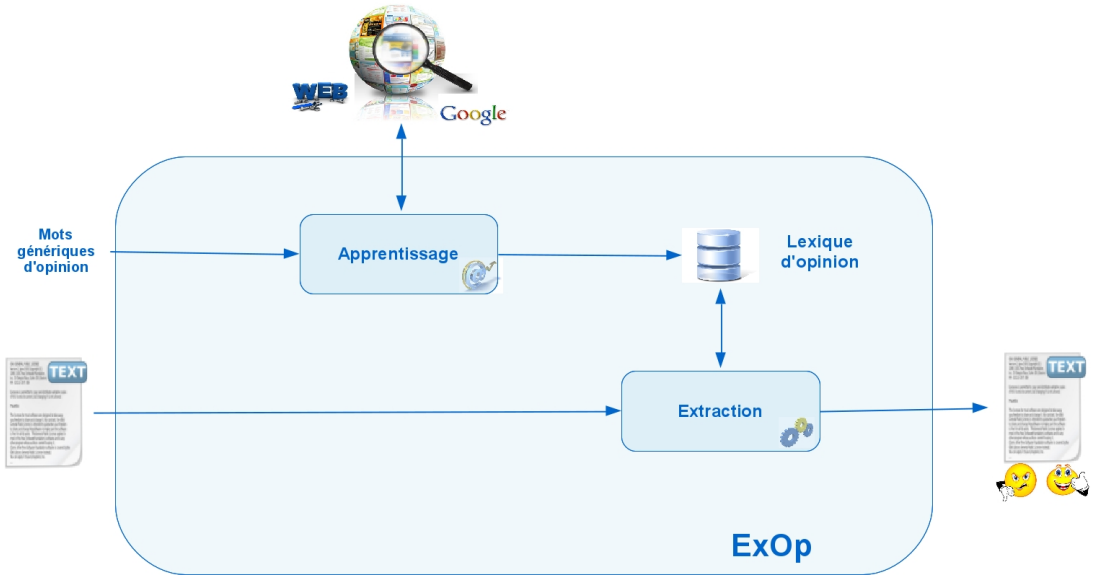

FIGURE 5.5 – Architecture du système *SchemExOp*

considérer uniquement la polarité d'un texte (opinion globale), de considérer les opinions relatives à des critères abordés dans ce texte. Le prototype *ExOPMulticritère* présenté dans cette section combine l'approche *Synopsis*, qui permet d'extraire les parties de texte qui traitent des différents critères (concepts), et l'approche d'extraction d'opinion *ExOp* (c.f. figure 5.7). Le système permet, à partir d'un texte fourni en entrée d'extraire l'opinion contenue dans le texte relativement à chacun des critères prédéfinis. L'évaluation donnée en sortie peut être binaire ou sous forme d'un intervalle de valeur compris entre 0 et 20 (c.f. chapitre 4 section 4.1.4.2).

$$[Sc^q](t) = \left[\overset{min}{Sc^q}(t), \overset{max}{Sc^q}(t)\right]$$

5.3.3 Utilisation du système *ExOpMulticritère* dans un contexte temporel d'analyse d'opinion : Le système "*StraussOp*"

L'analyse de données temporelles apporte souvent des informations pertinentes notamment sur la supervision et l'analyse de processus dynamiques. Étudier l'évolution d'un phénomène au cours du temps s'avère être une information très prisée dans les domaines du marketing, de la politique, de la bourse, etc. Les informations temporelles apportent des indices essentiels pour la prévision d'événements futurs, comme un crash boursier par exemple, ou pour déterminer des stratégies marketing

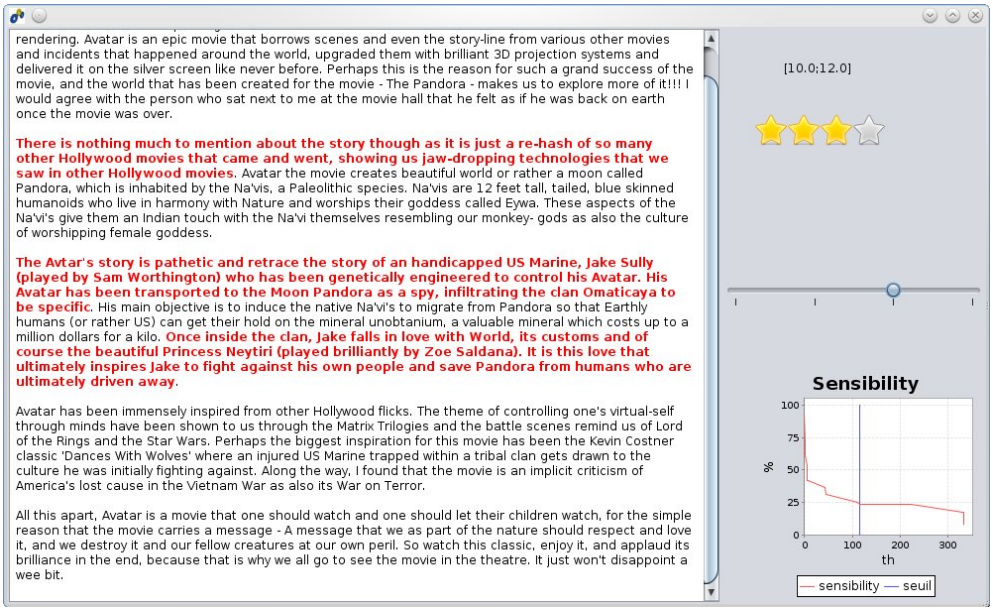

FIGURE 5.6 – Copie d'écran du système *ExOPMulticritère* sur un exemple.

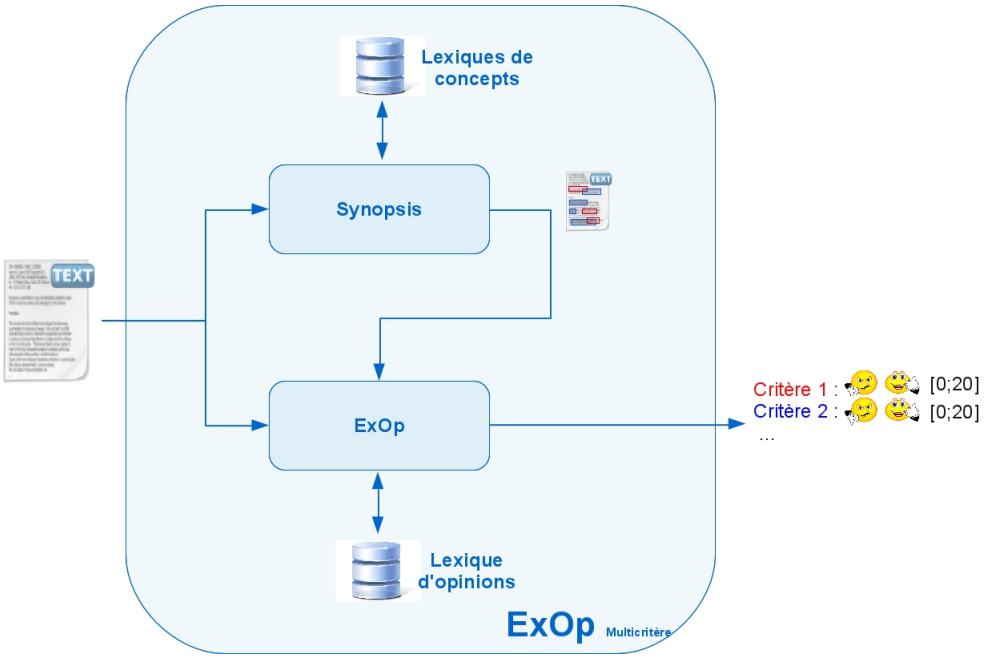

FIGURE 5.7 – Architecture du système *ExOpMulticritère*

à long et court terme, où encore pour adapter la production d'un produit dont la consommation augmente. Grâce à ce type d'approches, il est possible d'avoir des modèles statistiques robustes qui tiennent compte des événements précédents mar-

quants . C'est le cas des nouveaux systèmes boursiers basés sur le "Trading hautes fréquences" (HFT : High Frequency Trading) qui développent des algorithmes intelligents réalisant automatiquement des transactions financières en suivant une multitude de règles acquises au cours du temps.
Une des nombreuses autres applications d'analyse temporelle de données est de mettre en relation des phénomènes à première vue indépendants. L'idée est de rattacher, ou d'expliquer un phénomène par un autre en trouvant un lien de cause à effet : une corrélation entre les phénomènes considérés. Par exemple, trouver une corrélation entre le fait qu'un pays touristique soit subitement moins visité et qu'une insurrection civile y ait eu lieu, suffit à expliquer, diagnostiquer la situation. L'exemple précédent est trivial, mais des problèmes plus complexes peuvent apparaître, notamment lorsqu'il faut apporter des éléments de preuves qui expliquent l'apparition d'un phénomène. C'est le cas par exemple du dérèglement climatique pour lequel nous cherchons des corrélations entre des phénomènes climatiques et/ou géologiques et les activités humaines sur le globe. L'idée est de trouver, à partir de phénomènes simultanés, des corrélations entre les phénomènes observés à partir de données fournies par des systèmes physiques (capteurs :température, hygrométrie, gaz CO_2, précipitations, etc) (les causes possibles) qui seraient en mesure d'apporter une explication complète ou partielle au phénomène principal (l'effet).

Nous proposons ici un prototype logiciel qui analyse des données temporelles dans un but explicatif sur des phénomènes observés. L'objectif ici, est d'analyser l'opinion d'un grand nombre de personnes à l'échelle mondiale sur une affaire, ou un fait marquant, au cours du temps, et d'expliquer cette opinion perçue par des faits concrets. À titre d'exemple, nous proposons, d'analyser "*l'affaire Dominique Strauss-Kahn (DSK)*" qui a éclaté le 14 mai 2011, à partir d'articles de journaux publiés dans le "New York Times" entre le 1er Mai 2011 et le 31 Juin 2012 . Cette analyse n'a en aucun cas pour but de prendre une position politique quelconque, et le suivi d'une telle affaire a un objectif purement illustratif. Nous avons choisi d'étudier les opinions exprimées sur DSK selon deux critères : "Fond Monétaire Internationale" (FMI ou IMF en anglais) qui fait référence à la carrière politique de DSK, "sexe" (ou "sex" en anglais).
Le "New York Times" propose un service web qui donne accès aux articles publiés chaque jour. Chaque article est géolocalisé, c'est-à-dire que le service fournit le pays où l'article a été rédigé (écrit par un envoyé spécial qui résume la situation dans un pays par exemple). C'est à partir de cette information géolocalisée que nous

sommes en mesure, en analysant l'article avec *ExOpMulticritère*, d'extraire l'opinion exprimée dans un pays pour chacun des critères "FMI" et "sexe".
L'idée, est d'analyser les articles sur la période du 1er Mai 2011 au 31 Juin 2012, d'en extraire l'opinion exprimée sur chacun des critères, puis, pour chaque mois, expliquer l'opinion des personnes dans le monde (par pays), par les événements qui ont jalonné l'affaire durant cette période. Nous pouvons ainsi déterminer dans quelle mesure l'affaire a eu un impact sur la carrière politique de DSK par exemple, ou si les gens ont, au final, une opinion plus favorable sur l'un ou l'autre des critères par exemple.

L'application utilise les modules "Synopsis" et "ExOpMulticritère" pour extraire l'opinion sur chacun des critères *FMI* et *sexe*. L'architecture globale est décrite par la figure 5.8.

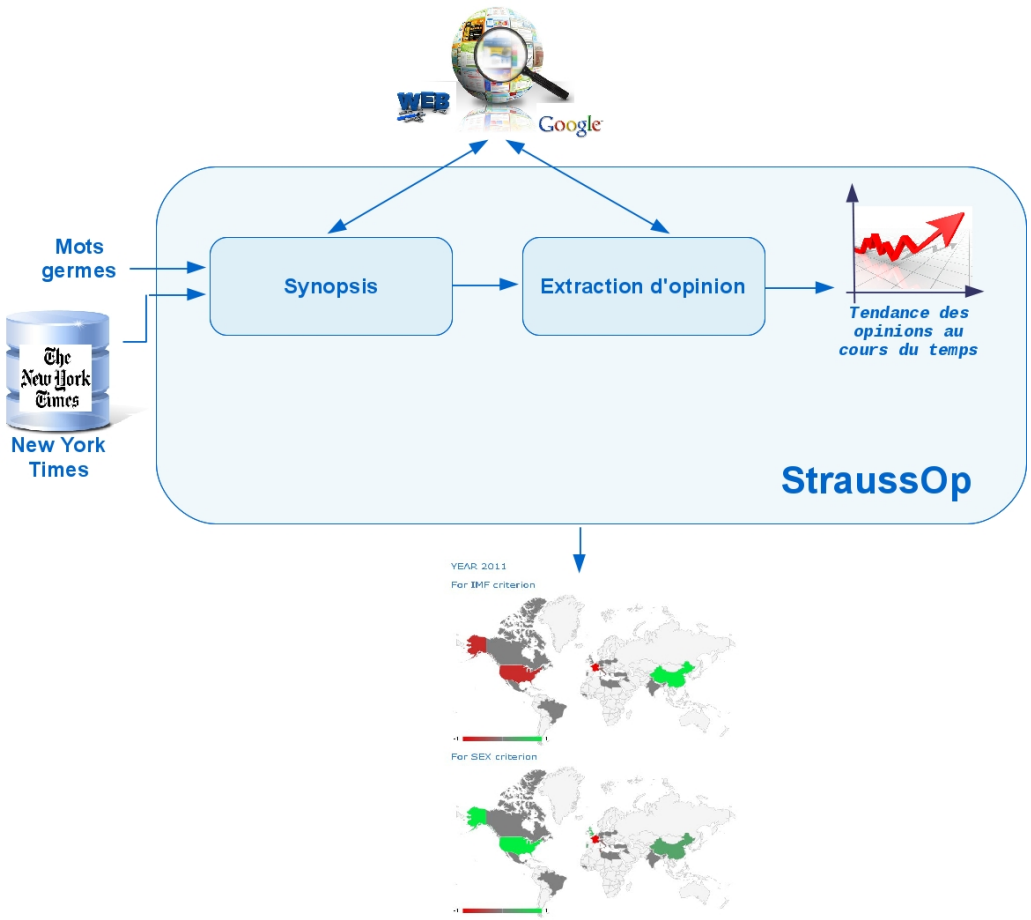

FIGURE 5.8 – Architecture du système *StraussOp*

Interprétation de l'analyse La figure 5.9 regroupe les différentes cartes du monde pour chaque mois de "l'affaire DSK" sur la période du 1er Mai 2011 au 31 Juin 2012. On y retrouve, sur chacune d'elles, l'opinion exprimée dans chacun des pays sur cette affaire et en particulier pour les deux critères retenus. La couleur verte fait référence à une opinion positive, la couleur rouge à une opinion négative, et la couleur grise à une opinion neutre. Nous proposons dans cette section d'identifier les faits les plus marquants de l'affaire qui expliqueraient l'opinion exprimée à un instant précis.

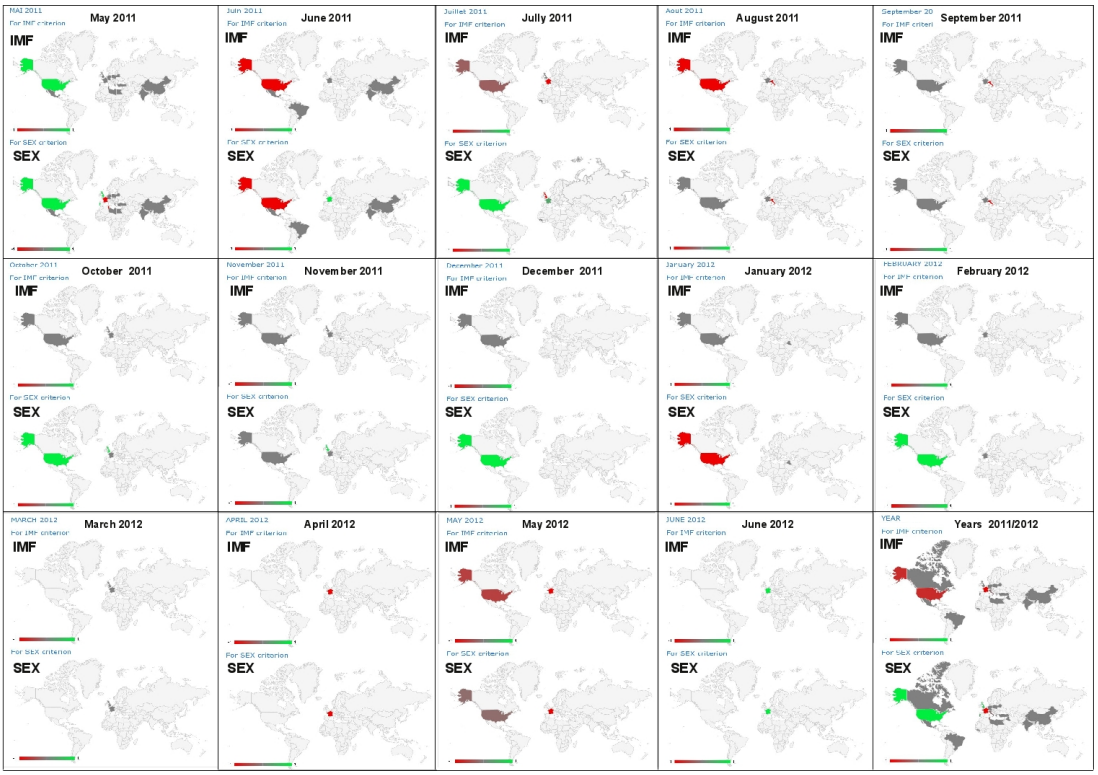

FIGURE 5.9 – Distributions des opinions selon les critères "FMI" et "sexe"

- **Mai 2011 :** L'affaire DSK éclate, l'opinion exprimée en France est plutôt négative sur le critère "sexe" (il existe déjà quelques accusations au niveau national) alors qu'elle est neutre pour le critère "FMI". Les États-unis expriment une opinions positive pour les deux critères.
- **Juin 2011 :** Les USA expriment une opinion négative sur DSK, ils le condamnent aussi bien sur sa carrière politique que sur l'affaire Nafissatou Diallo. En revanche, la France exprime une opinion neutre sur sa carrière politique, et a une

opinion positive sur le critère "sexe". Les français mettent de côté les anciennes rumeurs nationales et restent optimistes quant à l'innocence de DSK.

- **Août 2011 :** DSK est innocenté le 23 août 2011 et l'opinion dans le monde change (opinions neutres), sauf pour l'Italie qui parle de l'affaire DSK en la comparant à "l'affaire Berlusconi" sur la prostitution et qui exprime une opinion négative sur les deux critères considérés.
- **Avril 2012 :** L'opinion en France sur DSK est négative sur les deux critères. La France est en pleine période électorale (présidentielle) et la France s'exprime sur le fait que DSK était candidat à l'élection présidentielle.
- **Mai 2012 :** DSK porte plainte contre Nafissatou Diallo. L'opinion aux USA devient négative sur les deux critères, de même que pour la France.
- **Sur la totalité de la période**, nous pouvons retenir que l'opinion aux USA et en France est négative pour le critère "FMI", c'est-à-dire que les gens ont globalement une opinion plutôt négative sur sa carrière politique. En revanche, les USA ont une opinion plutôt positive sur le critère "sexe". Cela peut s'expliquer par le fait qu'il a été innocenté. La France quant à elle, garde une opinion plutôt négative pour le critère "sexe". Cela peut s'expliquer par les différents affaires et chefs d'accusation à l'égard du comportement de DSK (Tristane Banon et Nafissatou Diallo notamment).

Ces prototypes ont eu pour principal objet d'instrumenter les approches théoriques des deux chapitres précédents. Ils ont chacun mis en évidence soit l'extraction conceptuelle et la segmentation, soit l'extraction d'opinions. Il nous reste maintenant à revenir sur la problématique de départ : la recommandation multicritère dans les systèmes de recommandation. Nous avons construit les pièces maîtresses du traitement de l'information nécessaires à la recommandation multicritère, il nous reste maintenant à en expliquer l'agencement pour répondre à notre problématique initiale afin de conclure ce manuscrit.

Chapitre 6

Systèmes de recommandation

"*Trop de connaissance ne facilite pas les plus simples décisions.*"

Frank Herbert

"*Les perceptions des sens et les jugements de l'esprit sont des sources d'illusion et des causes d'incertitude.*"

Anatole France

Sommaire

6.1 Introduction

Nous avons montré dans l'état de l'art (chapitre 2) que les SIAD multicritères sont des éléments clés du web actuel. Nous avons mis en évidence que le SIAD qui utilise un Modèle Multicritère basé sur la découverte des Préférences à partir d'Évaluations (MMPE) permet d'obtenir les recommandations les plus pertinentes vis-à-vis des préférences de l'utilisateur. Cependant, la mise en place d'un tel modèle nécessite une base de connaissances où les items sont évalués par rapport à un ensemble de critères. Cette contrainte est imposée par le modèle. Elle est très lourde pour l'utilisateur qui doit évaluer chacun des items. Des approches comme [Plantie *et al.* 2008] ont tenté de réduire une partie de cette tâche décourageante, mais la plupart sont spécifiques au domaine d'utilisation, et le fait de changer de domaine nécessite un nouvel apprentissage supervisé (annotation manuelle d'un nouveau corpus d'apprentissage). Nous pensons que cette charge de travail nuit largement à la diffusion des MMPE. C'est pourquoi, nous avons proposé une méthode non-supervisée d'extraction d'opinion multicritère (c.f. chapitre 4) qui permet aux utilisateurs de s'affranchir de ces évaluations partielles sur les items à recommander : les utilisateurs peuvent déposer leurs critiques comme à l'accoutumé, sous la forme d'un simple texte exprimant leur opinion, et le système se charge de l'analyser automatiquement pour la décomposer selon les critères d'évaluation avant de calculer les polarités pour chacun de ces derniers. En combinant cette méthode à un SIAD multicritère, il devient possible d'avoir un système automatisé pour la recommandation. Ainsi, la mise en place de ce type de SIAD ne présente plus d'obstacles majeurs. Il nous reste donc à exposer l'organisation du MMPE doté de nos outils d'extraction et d'un système d'évaluation multicritère adéquat.

6.2 Intégration à un SIAD multicritère (MMPE)

La figure 6.1 illustre le fonctionnement global d'un système de recommandation multicritère de type MMPE. Le système utilise les techniques d'extraction des connaissances présentées dans les chapitres 3 et 4. Ces dernières sont employées ici pour évaluer les opinions exprimées dans des critiques cinématographiques sur un ensemble de critères définis. Les évaluations sur chacun des critères fournies pour chaque critique se présentent sous forme d'intervalles de valeurs (obtenus en faisant varier la précision de l'extraction) comme cela a été décrit au chapitre 4. Cette imprécision nous permet de modéliser la subjectivité relative à l'opinion exprimée pour

mieux restituer la richesse sémantique de l'évaluation. Cette considération implique l'utilisation d'outils mathématiques adaptés qui permettent de gérer cette imprécision. De plus, la multitude d'avis imprécis exprimés par les internautes sur un sujet donné fait apparaître une incertitude sur l'avis finalement attribué à ce sujet. Pour représenter ces données, nous avons choisi d'utiliser la théorie des possibilités proposée par Lofti Zadeh [Zadeh 1978] en considérant chacune des évaluations relatives à une critique comme un intervalle. Cette théorie nous permet de synthétiser un ensemble d'intervalles sous forme de distributions de possibilités tout en contrôlant l'imprécision et l'incertitude lors de la fusion d'opinions. Cette évaluation de synthèse s'avère utile dans un contexte de fouille de données massives pour avoir une représentation simplifiée de la pluralité d'opinions imprécises riches de sens. Grâce à cette représentation, nous sommes en mesure d'avoir un système adapté aux données considérées sans perte d'informations (imprécision liée à l'évaluation sous forme d'intervalles et incertitude liée à la multiplicité de ces intervalles). De plus, le fait d'évaluer chacun des éléments par un processus automatique d'opinion-mining nous permet d'assurer que toutes les évaluations extraites des critiques sont faites par le même système, et ainsi d'assurer la comparabilité des évaluations issues d'une "même subjectivité" d'interprétation. Le fait d'introduire ce processus automatique de notation nous permet de rendre crédible l'idée d'un SIAD MMPE non-supervisé, puisque la seule expertise nécessaire est de fournir une simple critique sous forme de texte. Le problème d'expertise qui rendait quasi impossible la mise en place d'un SIAD multicritère permettant l'individualisation des critères devient alors envisageable. Il est important de gérer l'incertitude (dispersion des avis exprimés) et l'imprécision liées aux données (évaluation imprécise de la polarité) pour obtenir une recommandation fiable. Le MMPE doit donc mettre en place les outils nécessaires à la manipulation de telles données .

6.3 Gestion des données imprécises

L'idée d'un SIAD multicritère est d'obtenir une évaluation globale de toutes les évaluations disponibles sur un item afin de pouvoir le comparer à d'autres items. Par exemple, attribuer une note moyenne de tous les avis exprimés sur le film "Avatar" permet d'obtenir une idée globale sur la qualité du film. Les utilisateurs sont très friands de cette évaluation moyenne qui leur permet d'émettre un jugement rapide sur le film considéré. Cependant, obtenir cette évaluation globale masque générale-

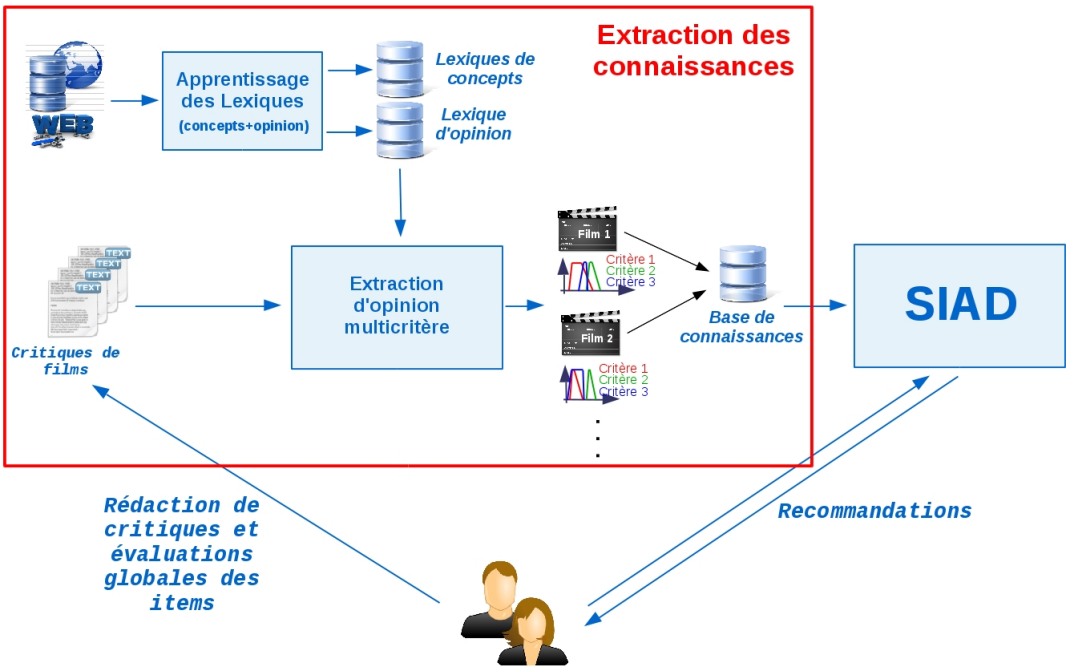

FIGURE 6.1 – Système de recommandation intégrant l'extraction automatique des connaissances

ment beaucoup d'informations et notamment la répartition des évaluations. Nous pouvons isoler trois cas de figures possibles de répartition des évaluations avec une même évaluation moyenne et laissant pourtant envisager des décisions bien différentes :

1. Les avis sont proches : nous sommes en situation de consensus. La distribution des avis est centrée autour de la valeur moyenne des évaluations des internautes ;
2. Les avis sont disparates : la distribution d'avis est uniforme, il ne se dégage aucun avis plus vraisemblable que les autres, nous sommes dans une situation de parfaite ignorance ;
3. Les témoignages sont controversés : plusieurs décisions sont crédibles, la distribution d'opinions est multimodale. Il y a certes des opinions dominantes correspondant aux modes, mais la valeur moyenne ne donne aucune idée des modes, il se peut même qu'aucune évaluation ne corresponde à la valeur moyenne.

La simple considération de valeurs de synthèse de type "moyenne" n'est donc pas suffisante pour émettre une recommandation représentative de la réalité. En

effet, une même valeur moyenne pourrait correspondre aux trois cas précédents et pourtant la décision afférente ne serait sans doute pas la même selon le cas de figure. Il est donc nécessaire de considérer l'information sous forme de distributions d'avis sans les réduire a priori à quelques moments statistiques (en pratique, l'avis de tous les internautes dans les MMPE aujourd'hui est réduit à un seul moment d'ordre un, la moyenne !). Si les évaluations étaient précises, il serait possible d'utiliser la théorie des probabilités pour synthétiser ces données : à partir des évaluations des internautes serait construit un histogramme, puis une distribution de probabilités qui synthétiserait plus fidèlement l'information brute. Dans notre cas où les évaluations sont imprécises, il est nécessaire d'utiliser la théorie des possibilités qui permet de gérer l'imprécision des évaluations. Cependant, si la représentation par des distributions de possibilités permet de distinguer les situations précédentes dans le processus d'évaluation, leur interprétation reste difficile pour le profane. Il est donc nécessaire de leur adjoindre des indicateurs de lecture permettant leur interprétation pour une recommandation. C'est le résultat que nous exposons dans ce chapitre.

6.3.1 La théorie des possibilités

À ce stade de notre travail, le problème qui se pose celui de construire une distribution à partir des scores imprécis $\left[Sc_k^q\right]$ attribués par les internautes. La théorie des possibilités a été introduite en 1978 par Lofti Zadeh [Zadeh 1978] pour donner une sémantique d'incertitude à la notion de sous-ensembles flous qu'il avait proposées dans le années 1960 [Zadeh 1965, Dubois & Prade 1988]. La théorie des possibilités [Dubois & Prade 1988] permet de tenir compte de l'imprécision des données ainsi que de l'incertitude à partir de deux mesures : la mesure de possibilité et la mesure de nécessité.

Les valeurs x appartenant au noyau de la distribution sont totalement possibles $\pi(x) = 1$, et celles appartenant au support de la distribution sont possibles avec un certain degré $\pi(x) \in]0, 1[$. Les valeurs qui se trouvent à l'extérieur du support de la distribution sont impossibles. Les valeurs possibles au degré α appartiennent à un ensemble appelé α-coupe. L'annexe B présente plus précisément la théorie des possibilités, c'est pourquoi nous ne présentons ici que les principes généraux nécessaires à notre représentation.

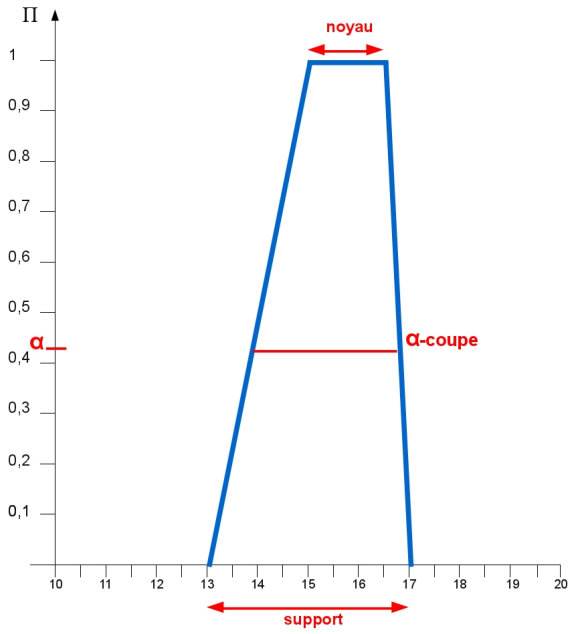

FIGURE 6.2 – Distribution de possibilités

6.3.1.1 Mesure et distribution de possibilité

Une mesure de possibilité, notée Π est une application qui attribue à chaque événement défini sur X , c'est-à-dire tout élément de $\Pi(X)$, un coefficient entre 0 et 1 évaluant à quel point cet événement est possible. Plus précisément, $\Pi : P(X) \longrightarrow [0, 1]$ est telle que :

- $\Pi(\oslash) = 0, \Pi(X) = 1$;
- $\forall A_1 \in P(X), A_2 \in P(X), \ldots$

$$\Pi(\bigcup_{i=1,2,\ldots} A_i) = sup_{i=1,2,\ldots}\Pi(A_i) \tag{6.1}$$

6.3.1.2 Mesure et distribution de nécessité

La mesure de nécessité est une mesure duale associée à la mesure de possibilité. Elle a été introduite dans la théorie des possibilités pour donner plus d'information sur un événement A relatif à X. En effet, l'inconvénient de la mesure de possibilité est qu'elle peut produire pour un événement A les deux résultats : $\Pi(A) = 1$ et $\Pi(A^c) = 1$, où A^c est le complémentaire de A dans X. Ce qui signifie qu'il peut y avoir une ignorance totale sur la réalisation de A avec cette seule information, d'où le besoin d'une information complémentaire sur la réalisation de A qui est apportée par la mesure de nécessité. Une mesure de nécessité N est définie sur $N : P(X) \longrightarrow [0, 1]$

de la manière suivante :

- $N(\oslash) = 0, N(X) = 1$;
- $\forall A_1 \in P(X), A_2 \in P(X), \ldots$

$$\Pi(\bigcap_{i=1,2,\ldots} A_i) = inf_{i=1,2,\ldots} N(A_i) \tag{6.2}$$

6.3.1.3 Relations entre mesures de possibilité et de nécessité

Le couple $(N(A), \Pi(A))$ mesure la croyance en la réalisation d'un événement A relatif à X . Il joue le rôle de la probabilité de la réalisation de l'événement A lorsque celle-ci ne peut être calculée. La quantité $\Pi(A)$ mesure le degré avec lequel l'événement A est susceptible de se réaliser, alors que la quantité $N(A)$ indique le degré de certitude que l'on peut attribuer à cette réalisation. Ainsi, la réalisation de l'événement A est certaine $(N(A) = 1)$ si et seulement si celle de son complémentaire est impossible $(\Pi(A^c) = 0)$.
Si la mesure de possibilité Π est définie à partir d'une distribution de possibilité π, la mesure de nécessité N duale de Π peut être définie à partir de π comme suit :

$$\forall A \in P(X), N(A) = inf_{x \notin A}(1 - \pi(x)) \tag{6.3}$$

Une mesure de possibilité Π est liée à la mesure duale de nécessité N par les propriétés de dualité suivantes :

- $\forall A \in P(X), N(A) = 1 - \Pi(A^c)$;
- $\Pi(A) \geqslant N(A)$;
- $max(\Pi(A), 1 - N(A)) = 1$;
- $N(A) \neq 0 \Rightarrow \Pi(A) = 1$;
- $\Pi(A) \neq 1 \Rightarrow N(A) = 0$;

6.3.2 Construction des distributions de possibilités

À partir des méthodes d'extraction des connaissances présentées aux chapitres 3 et 4 nous sommes en mesure d'extraire l'opinion relative à un critère prédéfini C^q dans un document doc_k sous forme d'un intervalle $\left[Sc_k^q\right](doc_k)$. Construire une distribution de possibilité à partir de l'ensemble des intervalles $\left[Sc_k^q\right](doc_k)$ obtenus à partir de critiques k de la base de connaissances du système nous permettra d'obtenir une image synthétique des données, qui prendra en compte toute l'information des évaluations fournies par les utilisateurs sur un critère C^q. La polarité

du document doc_k pour l'ensemble des internautes est alors vue comme une distribution de possibilité $\pi^q(doc_k)$ qui fusionne tous les $\left[Sc_k^q\right](doc_k)$ des internautes (un nombre flou). $\pi^q(doc_k)$ est donc une représentation floue des évaluations individuelles disponibles sur l'item relativement au critère C^q qui permet de gérer l'imprécision des évaluations $\left[Sc_k^q\right](doc_k)$ et l'incertitude liée à la dispersion des $\left[Sc_k^q\right](doc_k)$ (par construction un avis est imprécis dans notre système puisqu'il est le résultat des évaluations selon toutes les granularités d'analyse, ce qui nous semble dans la pratique assez légitime puisqu'on travaille sur avis "d'experts" et non pas sur des capteurs numériques). L'arsenal mathématique nécessaire à la construction de telles distributions ne fait pas l'objet de ce manuscrit, c'est pourquoi nous avons opté pour l'utilisation d'un algorithme proposé par [Imoussaten 2011] qui permet la fusion d'intervalles. Ainsi, les évaluations fournies par notre méthode d'extraction sont fusionnées, critère par critère, pour obtenir une représentation synthétique des avis sur chacun des items de la base. Plus formellement, r^q est le nombre de textes exprimant des opinions relatives au critère C^q et $\left[Sc_k^q\right]$ modélise l'intervalle d'opinion du texte k, $1 \leq k \leq r^q$ (calculé avec $ExOp$). L'idée générale est de construire une distribution de scores imprécis représentative de la diversité d'opinions exprimées en fusionnant les intervalles. Le principe consiste à étudier à travers une fonction de masse de croyance, la répartition des intervalles sur Ω : "plus un intervalle est fréquent, plus il aura de poids". Pour plus d'explications se référer à l'annexe B qui présente l'algorithme de fusion mis en place ainsi qu'une illustration sur la fusion d'avis.

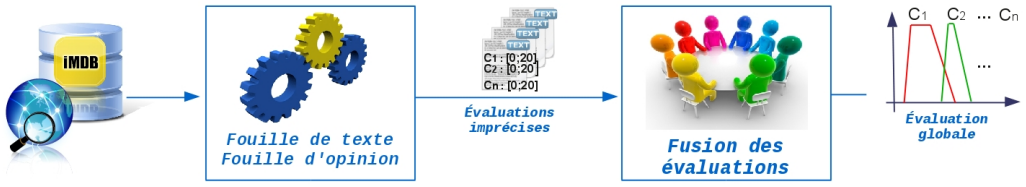

FIGURE 6.3 – Exemple de fusion d'avis pour un film donné sur plusieurs critères.

La figure 6.3 illustre le principe de construction des distributions de possibilité. Pour un film donné, le système construit, à partir des avis relatifs à chacun des critères considérés (notes sous forme d'intervalles) une distribution de possibilité pour chacun des critères. Ces distributions de possibilité synthétisent l'information contenue dans les critiques sans perte d'information pour le processus d'évaluation, c'est-à-dire que la granularité d'interprétation a été modélisée par l'intervalle de

scores, puis l'imprécision et l'incertitude ont été modélisées et intégrées dans la fusion. Le fait de traiter des données sous forme de distributions de possibilité simplifie le traitement de l'information en limitant le nombre de tuples à considérer et nous donne ainsi des informations sur la variabilité des données.

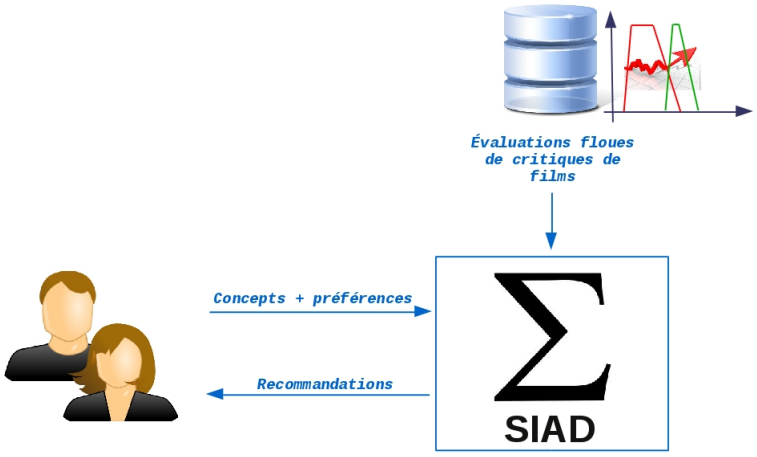

FIGURE 6.4 – Exemple de SIAD MMPE pour la recommandation de films ou de journaux spécialisés dans le cinéma.

6.3.3 Évaluation multicritère

Dans un SIAD multicritère basé sur un modèle MMPE, chaque item est évalué selon des critères définis. Nous obtenons ainsi pour chaque film une distribution de possibilité relative à chacun des critères considérés (c.f. figure 6.3.3).

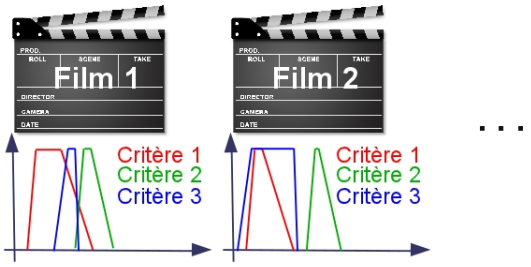

FIGURE 6.5 – Illustration dans le cas multicritère.

6.3.3.1 De la relation de préférence à la théorie de l'utilité multi-attributs (MAUT).

L'objectif d'un SIAD multicritère MMPE est d'établir un modèle de préférences dans un espace multicritère, cependant, en considérant indépendamment chacune des distributions, il n'est pas possible d'obtenir un tel modèle : un item (un film par exemple) est à ce stade représenté par un vecteur de scores flous résultats de la fusion des avis rédigés par les internautes, il n'y a pas d'information qui permette de construire un ordre partiel sur les préférences sans plus d'information sur l'importance relative des critères : il ne peut leur être accordé a priori qu'un rôle symétrique dans la fonction de choix. Il est donc intéressant de connaître l'influence de chacun des critères pour pouvoir élaborer une recommandation, un choix. C'est pourquoi, il est nécessaire de considérer la contribution (poids) de chacun des critères pour évaluer un item.

Plaçons nous quelques instants dans l'hypothèse d'évaluations numériques précises pour présenter la formalisation du problème. Considérons par exemple deux films F_1 et F_2 qui sont évalués sur quatre critères : *scénario*, *acteur*, *mise en scène*, *bande son* avec une note x_i^q comprise entre 0 et 20 relative au film i pour le critère q, et un poids w^q relatif à chaque critère q.

- Pour $F_1 = 16.1/20$: *scénario* $:w^1 = 0.7\,;x_1^1 = 18$, *acteur*$=:w^2 = 0.1\,;x_1^2 = 14$, *mise en scène*$=:w^3 = 0.1\,;x_1^3 = 12$, *bande son*$=:w^3 = 0.1\,;x_1^3 = 9$.
- Pour $F_2 = 16.1/20$: *scénario*$=:w^1 = 0\,;x_2^1 = 11$, *acteur*$=:w^2 = 0.6\,;x_2^2 = 17$, *mise en scène*$=:w^3 = 0\,;x_2^3 = 18$, *bande son*$=:w^3 = 0.4\,;x_2^3 = 14.75$.

Si nous considérons un SIAD multicritère qui base sa recommandation uniquement sur le poids de chacun des critères, la tendance de la recommandation portera sur les critères ayant les poids les plus importants, il n'y aura pas de compensation entre les critères. De la même manière, en considérant uniquement les évaluations de chacun des critères (sans poids), la priorité de tel ou tel critère disparaît. Il est donc nécessaire de considérer ces deux notions dans la contribution d'un critère à la décision : le poids du critère et l'évaluation associée. La contribution d'un critère à l'évaluation globale doit être fonction du couple poids-score. Nous proposons alors d'utiliser un modèle de type $\sum_{q=1}^{p} \omega^q.x_i^q$ où le terme $\omega^q.x_i^q$ tient compte des deux notions précédemment identifiées. $\omega^q.x_i^q$ représentera la contribution du critère q à l'évaluation (c.f. section 6.3.3.2).

Le tableau 6.1 montre les contributions de chacun des critères à l'évaluation glo-

Film	*scénario*	*acteur*	*mise en scène*	*bande son*	**note globale**
F1	12.6	1.4	1.2	0.9	16.1
F2	0	10.2	0	5.9	16.1

TABLE 6.1 – Contribution sur chacun des critères.

bale. Nous pouvons remarquer que malgré une même note globale identique pour les deux films (16.1/20), la contribution de chacun des critères est bien différente. Tandis que pour $F1$, le critère *scénario* contribue majoritairement à la note ($\simeq 78\%$ de la note globale), pour le film $F2$ c'est le critère *acteur* qui contribue majoritairement à la note globale ($\simeq 63\%$ de la note globale). Si nous avions considéré uniquement les scores sans les poids sur chacun des critères, $F1$ aurait été évalué comme un bon film grâce à son *scénario* (18/20) mais aussi de par le jeu des *acteurs* (14/20) ; $F2$ aurait été lui évalué comme un bon film grâce au jeu des *acteurs* (17/20), sa *mise en scène* (18/20) et *sa bande son* (14.75/20). Il faut nécessairement prendre en compte scores et poids pour avoir un modèle minimal du système de préférences des utilisateurs et ne pas retomber dans les biais des systèmes de recommandation cités dans l'introduction de ce chapitre. Nous pouvons donc conclure qu'il est essentiel de connaître, le poids et l'évaluation de chacun des critères pour avoir une recommandation pertinente et fiable dans le cadre d'un véritable système de recommandation multicritère MMPE, et ainsi être en mesure d'expliquer une évaluation globale par ses critères d'intérêt (la recommandation). Pour cela, il est donc nécessaire de connaître le système de préférences de l'utilisateur et ainsi d'identifier les poids, c'est-à-dire l'importance qu'il attache à chacun des critères.

La représentation des préférences est un sujet central dans la prise de décision [Modave & Grabisch]. Habituellement, cela consiste à trouver une fonction d'utilité réelle U telle que pour toute paire d'alternatives s, $s' \in S$ où S est un ensemble d'alternatives, $s \succeq s'$ (s est préférée à s') si et seulement si $U(s) \geq U(s')$. Lorsque les alternatives sont n-dimensionelles $s^k = (s_1^k, \ldots, s_n^k)$, on introduit u_q^k l'évaluation partielle de la k*ième* alternative s^k selon le critère C^q. Un modèle largement étudié est le modèle décomposable de Krantz [Krantz *et al.* 1971], où U est de la forme $U(s_1, \ldots, s_p) = g(u_1(s_1), \ldots, u_n(s_n))$ où les u_q sont des fonctions d'utilité élémentaire. Si $\succeq$ est un ordre partiel sur S, il existe une représentation avec g strictement croissante si et seulement si $\succeq$ vérifie les propriétés d'indépendance et de séparabilité [Modave & Grabisch]. La théorie de l'utilité multi-attributs (MAUT) [Fishburn 1970, Fishburn 1982] est basée sur la théorie de l'utilité qui permet de

quantifier des préférences individuelles. La théorie de l'utilité consiste à interpréter toute mesure comme un degré de satisfaction dans l'intervalle $[0, 1]$ où 0 correspond à la pire alternative et 1 à la meilleure. Une utilité u_q^k est attachée à chaque mesure s_q^k, ainsi toute évaluation partielle a une interprétation unique : un degré de satisfaction. Ensuite, la MAUT fournit une utilité U de synthèse qui apporte une réponse au problème de comparaison de deux situations décrites au moyen de leurs expressions d'utilité élémentaires. La relation d'ordre sur $\mathbb{R}n$ nécessaire à la comparaison et au classement des alternatives est ramenée à un ordre sur R grâce à l'utilité de synthèse.

En considérant deux alternatives X et Y, nous pouvons écrire que X est préférée à Y si :

$$X \succeq Y \Leftrightarrow g(X) \geq g(Y) \tag{6.4}$$

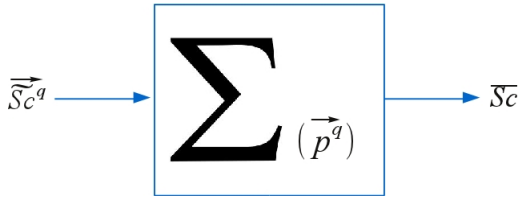

FIGURE 6.6 – Modélisation du système de préférences.

6.3.3.2 Identification du système de préférences.

À partir de l'équation 6.4, il est possible d'expliquer la préférence d'un film sur un autre dès lors que la fonction g est identifiée. La forme analytique de g qui modélise le système de préférences (c.f. figure 6.6) facilite l'identification, la comparaison des alternatives ainsi que la recommandation basée sur la notion de contribution à g. Dans un premier temps, il s'agit donc de déterminer les paramètres du système de préférences. Lors d'une évaluation, l'esprit humain agrège selon différents critères pour émettre une note globale. L'objectif est alors de déterminer la fonction d'agrégation qu'il utilise, et qui synthétise l'expression de ses préférences lors de l'évaluation. Pour un utilisateur donné, dans un système d'évaluation avec des notes précises et avec un modèle linéaire pour g , l'évaluation globale pour un film i est donnée par :

$$\sum_{q=1}^{p} \omega^q . x_i^q \tag{6.5}$$

où x_i^q est le score attribué au film i sur le q *ième* critère, ω^q est le poids donné au critère. L'identification des poids fournira le système de préférences qui rend le mieux compte des évaluations de l'utilisateur. L'idée ici est de trouver les poids ω^q relatifs à chacun des critères qui permettraient de retrouver la note globale attribuée à un film par l'utilisateur dont on cherche la forme analytique du système de préférences. La méthode des moindres carrés est utilisée pour identifier le système de préférences qui reflète le mieux le processus d'évaluation chez l'utilisateur sur un ensemble de films. La méthode recherche la combinaison linéaire qui minimise l'écart entre la valeur globale x_i fournie par l'évaluateur pour le film i, et la valeur calculée par le système de préférences avec le modèle linéaire pour g. Plus formellement, l'identification du modèle d'évaluation de l'utilisateur est définie comme suit :

$$\min_{\omega^q, q=1...p} \sum_{films} \left(\sum_{q=1}^{p} \omega^q . x_i^q - x_i \right)^2 \tag{6.6}$$

La solution de ce problème d'optimisation fournit le jeu de poids qui permet de savoir quel score l'utilisateur attribuera à un film i' qui aurait obtenu les scores $x_{i'}^q$ sur l'ensemble des critères s'il est cohérent avec son système de préférences. C'est la meilleure estimation possible du système de préférences de l'utilisateur avec une forme analytique linéaire pour g . Plus le nombre de films jugés par l'utilisateur est grand, plus l'identification des poids est fiable (plus il donne d'exemples mieux on connaît son système de préférences). L'utilisation de g pour prédire la note globale que devrait attribuer l'utilisateur au vu des scores partiels recueillis sur un film i' est la base de la recommandation. Par exemple, imaginons que l'on recueille les évaluations disponibles dans une base de données du web sur le film i'. Ces évaluations nous fournissent des scores obtenus par i' sur chaque critère. Si on en fait la moyenne statistique sur chaque critère et que l'on applique le jeu de poids propre à l'utilisateur, on peut alors estimer le score global qu'attribuerait l'utilisateur et donc savoir si oui ou non il faut lui recommander i'. De plus, en calculant les contributions $\omega^q . x_i^q$, nous sommes en mesure de justifier la recommandation par les critères ayant le plus contribué à l'évaluation du film i'.

La méthode des moindres carrés est bien adaptée pour des évaluations précises. Dans notre problématique les évaluations unitaires sont imprécises (modélisées sous

la forme d'intervalles), c'est pourquoi nous proposons d'adapter la méthode des moindres carrés à notre type de données. Les valeurs x_i^q précises deviennent imprécises, et elles correspondent aux intervalles $[Sc_i^q]$ de l'évaluation du film i. Supposons donc que l'on dispose d'évaluations partielles selon une base de critères : nos outils d'extraction de critères et d'opinion ont permis de générer la base de $[Sc_i^q]$ pour un ensemble de films depuis les critiques textuelles écrites par les internautes. L'évaluation par tous les évaluateurs i (les critiques recueillies) du critère C^q correspond à la fusion des $[Sc_i^q]$: $\widetilde{Sc}^q$. Enfin, l'évaluation globale calculée sur l'ensemble des critères est elle-même imprécise et prend la forme d'une distribution de possibilités, résultat de l'agrégation des $\widetilde{Sc}^q$, telle que :

$$\widetilde{Sc} = \sum_{q=1}^{p} p^q . \widetilde{Sc}^q \tag{6.7}$$

Le lecteur intéressé pourra se reporter aux thèses de Afef Denguir-Rekik [Rekik 2007] et de Abdelhak Imoussaten [Imoussaten 2011] pour de plus amples détails sur le processus d'agrégation de nombres flous.

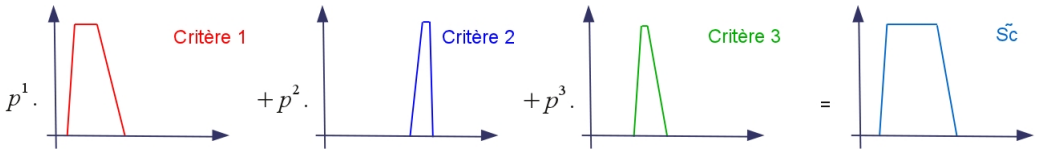

FIGURE 6.7 – Exemple d'agrégation sur trois critères.

Dans la problématique des systèmes de recommandation, l'évaluation globale $\overline{Sc}$ d'une alternative est donnée par ailleurs de manière imprécise (3 étoiles, 4 barres, 5 smileys, etc.) quel que soit le type de modèle sous-jacent au SIAD (c.f. le chapitre 2). $\overline{Sc}$ est fournie par un utilisateur donné (qui peut être un journal spécialisé par exemple). Dans notre approche, $\overline{Sc}$ est modélisée par distribution de possibilité (c.f. figure 6.8). Comme dans la plupart des systèmes de recommandation, seule l'évaluation globale $\overline{Sc}$ est disponible, l'idée est d'extraire des critiques des notes partielles $\widetilde{Sc}^q$ pour chaque critère, de les agréger en $\widetilde{Sc}$ et de vérifier que $\overline{Sc}$ et $\widetilde{Sc}$ coïncident. Ce processus permet d'expliquer les évaluations globales exprimées dans les critiques par les scores partiels des critères. Ce diagnostic permet ainsi d'enrichir la recommandation puisqu'au lieu de recommander *Avatar* parce que c'est un film qui a plu au public, il peut être précisé à l'utilisateur que si ce sont des effets spéciaux qu'il cherche alors *Avatar* est peut-être le film qu'il cherche, par contre si c'est un scénario qu'il veut, il risque d'être déçu.

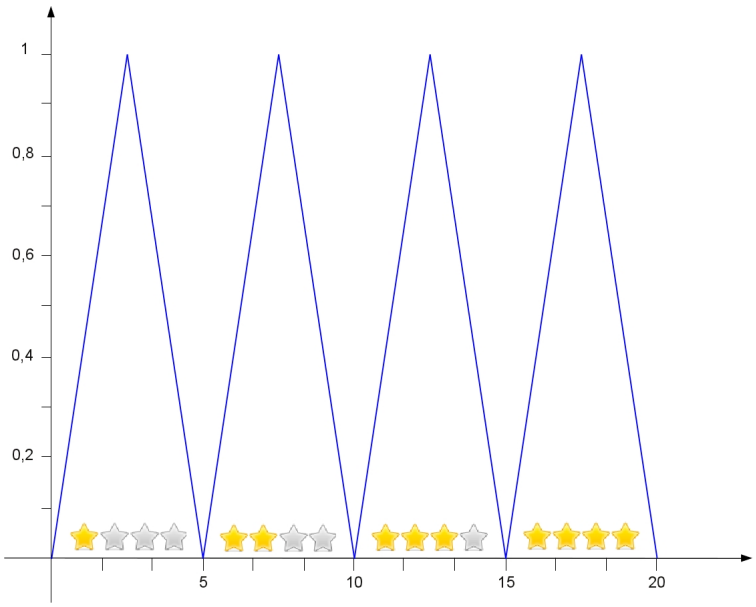

FIGURE 6.8 – Exemple de distribution de possibilités construites à partir des étoiles d'évaluation.

Comme il serait difficile de définir une erreur quadratique (comme dans le cas précis précédent) qui permettrait d'établir un modèle de préférence dans ce contexte imprécis, nous avons choisi de raisonner en terme d'implication. Autrement dit, la question que l'on se pose, est : connaissant les évaluations partielles floues $\widetilde{Sc}^q$ est-il possible d'en déduire $\overline{Sc}$? Mathématiquement, cela se traduit par :

$$\widetilde{Sc} \Rightarrow \overline{Sc} \tag{6.8}$$

Nous cherchons donc le jeu de poids relatifs qui garantit au mieux cette implication.
En logique classique, l'implication $v \Rightarrow w$ équivaut à $\neg v \vee w$. On obtient alors la table de vérité suivante :

v	w	$v \Longrightarrow w$
1	1	1
1	0	0
0	1	1
0	0	1

TABLE 6.2 – Table de vérité en logique classique.

En logique floue, il n'existe pas de définition unique et il existe plusieurs propositions pour l'implication $Imp(\pi, \pi')$. Une extension possible de $\neg v \vee w$ est appelée l'implication de *Kleene-Dienes*. Elle est définie telle que $\widetilde{Sc} \Rightarrow \overline{Sc}$ équivaut à :

$$Imp(\widetilde{Sc}, \overline{Sc}) = inf\ max(1 - \widetilde{Sc}, \overline{Sc}) \tag{6.9}$$

L'implication de *Mandani* est également utilisée. Elle est définie par :

$$Imp(\widetilde{Sc}, \overline{Sc}) = sup\ min(\widetilde{Sc}, \overline{Sc}) \tag{6.10}$$

L'implication de *Kleene-Dienes* est très restrictive et celle *Mandani* souvent trop permissive. Un compromis entre ces deux implications serait de calculer la surface commune entre $\widetilde{Sc}$ et $\overline{Sc}$ telle que cette mesure représente leur pourcentage de recouvrement :

$$Imp(\widetilde{Sc}, \overline{Sc}) = \frac{100 * \int \bigwedge(\widetilde{Sc}, \overline{Sc})}{\int \widetilde{Sc}} \tag{6.11}$$

Les implications précédemment définies ne sont pas exhaustives, et d'autres implications peuvent être établies pour être mieux adaptées aux conditions d'utilisation.

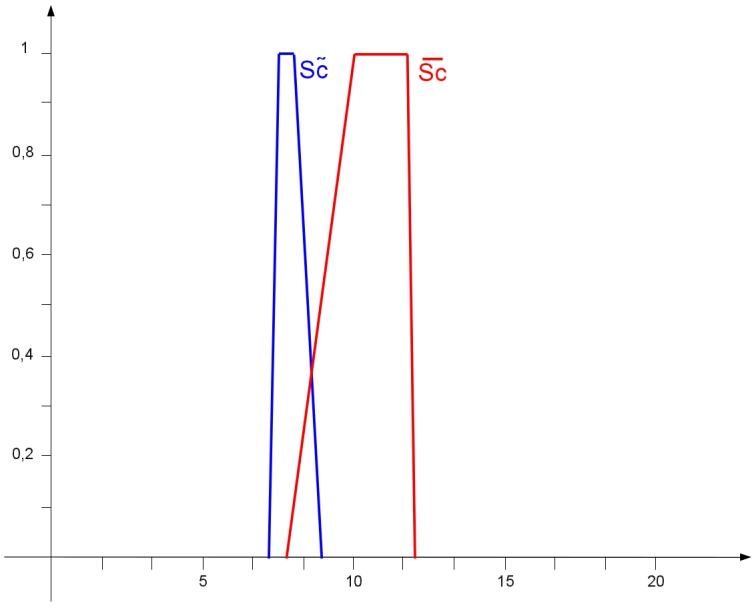

FIGURE 6.9 – Exemple de distributions.

En considérant les deux distributions $\widetilde{Sc}$ et $\overline{Sc}$ telles qu'elles sont représentées sur la figure 6.9 et la fonction d'implication définie par l'équation 6.9, l'implication

$\widetilde{Sc} \Rightarrow \overline{Sc}$ est de 0. En considérant la fonction d'implication définie par l'équation 6.10, l'implication vaut 1. Cet exemple montre que le choix de l'implication de deux distributions est important car les résultats peuvent être radicalement différents. Par ailleurs, nous pouvons remarquer que l'implication définie par l'équation 6.9 est plus restrictive que l'implication définie par l'équation 6.10. Il ne s'agit pas ici de discuter la sémantique de ces implications, pour notre application, nous avons simplement retenu l'implication définie par la surface de recouvrement.

Nous proposons maintenant de déterminer les poids de chacun des critères par l'intermédiaire d'une analyse de sensibilité. Cette dernière consiste, selon le même principe que les moindres carrés dans le cas précis, à faire varier les poids de chacun des critères et d'évaluer, pour chaque variation des w^q, l'implication de $\widetilde{Sc}$ avec $\overline{Sc}$. La distribution de poids $\overrightarrow{w} = \left[w^1 \dots w^q \dots w^p\right]$ retenue sera celle qui maximise l'implication. L'idée est de faire varier la valeur des poids w^q telle que $\forall q \in N$, $w^q + \delta w^q$ avec $\delta w^q \in [0, 1]$ de façon à déterminer les poids du modèle de préférences.

Si l'on dispose d'une base de films évalués avec les critiques associées, on cherche $\overrightarrow{w} = \left[w^1 \dots w^q \dots w^p\right]$ telle que :

$$max_{\overrightarrow{w}}[\sum_{films évalués} (\sum_{q=1}^{p} w^q . \widetilde{Sc^q} \Rightarrow \overline{Sc})] \tag{6.12}$$

Si comme dans le cas précis, on dispose d'une base de films évalués par l'utilisateur, on peut utiliser ce principe pour calculer le modèle utilisateur. En pratique s'il s'agit de calculer le modèle d'un utilisateur en particulier dans le cas imprécis, nous pensons qu'il est plus efficace d'avoir recours à une méthode d'identification indirecte du type Macbeth ([Bana E Costa & Chagas 2004], [Bana E Costa & Vansnick 1997]). Mais ce point est hors du cadre de nos travaux.

L'application de 6.12 est à notre avis plus pertinente dans le cas suivant. Sur le site d'un journal spécialisé, nous recueillons les scores attribués par le journal sur une base de N films. Pour ces films, nous cherchons ensuite les critiques disponibles sur le web. Nous en extrayons les critères et les opinions associées. Nous essayons finalement avec 6.12 d'estimer la distribution de poids du journal $\overrightarrow{w} = \left[w^1 \dots w^q \dots w^p\right]$ qui caractérisera sa stratégie (son système de préférences), c'est-à-dire les critères qui sont systématiquement mis en avant par les experts du journal. Les applications sont ensuite multiples : ce processus permet d'identifier le journal spécialisé qui colle le mieux à un utilisateur donné, de faire du clustering de journaux, de faire du monitoring de l'évolution de la stratégie des journaux, etc.

L'objet de ce chapitre n'est pas d'imposer tel ou tel processus d'évaluation multicritère pour les systèmes de recommandation. En effet, suivant la richesse de l'information disponible sur un site de recommandation, tous les calculs proposés dans ce chapitre ne sont pas forcément nécessaires. Notre objectif est davantage de montrer que quel que soit le type du système de recommandation, du moment qu'il offre une base d'items associés à des critiques et des évaluations globales, notre approche (extraction conceptuelle des critères et extraction des opinions, associée aux outils d'analyse multicritère), permet de le transformer en un site de recommandation multicritère de type MMPE.

6.3.3.3 Indicateurs pour l'aide à la décision.

La propagation des scores flous partiels permet de conserver toute l'information produite par le processus d'évaluation. Néanmoins, le décideur dispose au final du score flou agrégé de chacune des alternatives. L'interprétation de ce résultat n'est pas forcément évidente pour l'utilisateur. Nous pensons qu'il est nécessaire dans ce cas de concevoir des indicateurs $Indic(\widetilde{Sc})$ sur les scores flous générés afin que le décideur dispose de grandeurs qu'il soit en mesure d'interpréter facilement tels que :

$$Indic(\widetilde{Sc}) = \sum_{q=1}^{p} w^q.Indic(\widetilde{Sc^q}) \tag{6.13}$$

Cette écriture permet en effet de trouver la contribution de chaque critère à l'évaluation globale $\widetilde{Sc}$ pour chaque indicateur $Indic$ vérifiant l'équation 6.11. Une fois les poids identifiés, on pourra donc comme dans le cas précis, non seulement estimer la note globale que l'utilisateur devrait attribuer à un film sur la base des notes partielles collectées sur le web, mais encore justifier la recommandation avec les $w^q.Indic(\widetilde{Sc^q})$ les plus conséquents, *i.e.* ceux qui contribuent le plus à la valeur de l'indicateur global $Indic(\widetilde{Sc})$.

Nous présentons ici deux indicateurs : un indicateur de position qui correspond à une note de synthèse précise (en probabilités ce serait la moyenne), et un indicateur d'imprécision moyenne qui correspond à une estimation de l'imprécision moyenne sur cette valeur de synthèse (en probabilités ce serait la variance). Ces deux indicateurs ne sont pas exhaustifs, et il est possible de définir d'autres indicateurs. Notons que ces indicateurs sont définis a posteriori : l'information contenue dans la distribution ne se résume pas à deux indicateurs, et comme toute l'information a été conservée dans ce type de processus de fusion-agrégation, il est parfaitement possible de définir a posteriori tous les indicateurs nécessaires à l'utilisateur pour faire son choix. Le

principe est très différent si l'on décide a priori de ne propager que les indicateurs retenus (et non pas les distributions) : les calculs sont beaucoup plus simples, mais la perte d'information est irrémédiable.

Pour définir nos indicateurs, il est nécessaire de revenir sur des éléments sémantiques essentiels des distributions de possibilités. Une façon d'aborder les possibilités est de considérer qu'une distribution de possibilité définit une famille de probabilités (puisqu'on ne peut définir une unique distribution de probabilité sur la base d'un histogramme d'évaluations imprécises) : cela permet de représenter une connaissance incomplète des probabilités et de définir des bornes inférieures et supérieures de ces probabilités imprécises. On peut en effet définir une fonction de répartition supérieure (F^*) et une fonction de répartition inférieure (F_*) telles que $\forall u$, $F_*(u) \leq F(u) \leq F^*(u)$, avec $\forall u \in \mathbb{R}$, $F^*(u) = \Pi(]-\infty, u])$ et $F_*(u) = N(]-\infty, u[)$. La différence entre les bornes $F^*(u)$ et $F_*(u)$ rend compte de l'imprécision de l'information. Nous pouvons ensuite définir la valeur moyenne inférieure $E_*(\widetilde{Sc}) = \int_{-\infty}^{+\infty} x dF^*(x)$ et la valeur moyenne supérieure $E^*(\widetilde{Sc}) = \int_{-\infty}^{+\infty} x dF_*(x)$. L'intervalle $\left[E_*(\widetilde{Sc}), E^*(\widetilde{Sc})\right]$ est la valeur moyenne de l'imprécision de $\widetilde{Sc}$ (*i.e.*, la valeur moyenne d'un nombre flou est un intervalle).

L'indicateur de position $MD(\widetilde{Sc})$ permet d'obtenir une valeur réelle de synthèse du sous-ensemble flou $\widetilde{Sc}$ telle que :

$$MD(\widetilde{Sc}) = (E_*(\widetilde{Sc}) + E_*(\widetilde{Sc})/2 \tag{6.14}$$

L'indicateur $\triangle(\widetilde{Sc})$ permet d'obtenir l'imprécision moyenne de $\widetilde{Sc}$ telle que :

$$\triangle(\widetilde{Sc}) = E_*(\widetilde{Sc}) - E_*(\widetilde{Sc}) \tag{6.15}$$

Grâce à l'indicateur MD, il sera possible de fournir une information sur les critères qui justifient le score global $\widetilde{Sc}$. L'indicateur $\triangle$ permet de déterminer les critères qui ont le plus contribué à la dispersion sur le score global.

Ainsi, la justification d'une recommandation est plus riche que dans le cas précis. En raisonnant sur l'indicateur MD, on se ramène à l'explication de la recommandation comme dans le cas précis puisque $MD = \sum_q \omega_q.MD_q$. La contribution à l'indicateur de position nous permet d'expliquer le score global attribué à un film par les critères les plus influents sur le calcul de MD. En revanche, $\Delta = \sum_q \omega_q.\Delta_q$ nous permet d'avoir un raisonnement analogue sur l'imprécision de l'évaluation (et

donc la fiabilité de la recommandation) : plus la contribution $\omega_q.\Delta_q$ est conséquente, plus le critère q explique la valeur de l'imprécision moyenne Δ. Plus l'imprécision est grande, moins la recommandation est fiable et le système peut donc fournir les critères qui nuisent le plus à la fiabilité de la recommandation (ceux où il y a le plus d'avis disparates, controversés, etc.).

6.3.3.4 Système de recommandation Multicritère

La pénurie d'information est un vrai problème lorsqu'il faut prendre des décisions, la "sur-information", en est un autre, qui rend les plus simples décisions complexes. Aider l'utilisateur à explorer cette masse de données pour l'aider dans ses choix est l'objectif principal des systèmes de recommandation.
Nous proposons un prototype logiciel qui permet de recommander à un utilisateur des films et des journaux spécialisés dans le cinéma, en fonction de ses préférences. Ce prototype utilise le module d'extraction d'opinion multicritères (c.f. section 5.3.2), ainsi qu'une base de critiques de films, en langage naturel, issue du célèbre site de critiques cinématographiques IMDB. Chaque critique de film a été évaluée par notre système d'extraction d'opinions multicritères selon deux critères : *acteur* et *scénario*, une note sous forme d'intervalle étant attribuée à chaque critique sur les deux critères. Nous disposons par ailleurs, pour chaque film présent dans la base, d'évaluations globales sous forme d'un nombre d'étoiles, attribuées par la presse spécialisée (nous avons retenu 20 journaux dans cette application). Le nombre d'étoiles attribué à un film représente la qualité du film considéré (1 étoile, le film est médiocre, 4 étoiles, le film est excellent).
Le système permet deux choses. D'une part, de recommander au lecteur, des journaux qui en identifiant les politiques d'évaluation des journaux (poids sur chacun des critères considérés). D'autre part, le système est capable de recommander des films à l'utilisateur par rapport à ses préférences et aux notes attribuées par une communauté d'internautes réputés avoir des goûts proches des siens (systèmes de préférences proche du sien). L'architecture du logiciel est décrite par les figures 6.10 et 6.11.

Détection de stratégie pour la recommandation de journaux L'objectif est de trouver le jeu de poids utilisé par un journal (l'agrégation étant supposée être une moyenne pondérée) permettant d'expliquer au mieux la note qu'il a attribuée à un film. Les critiques fournies par IMDB, de par le nombre (environ 3000

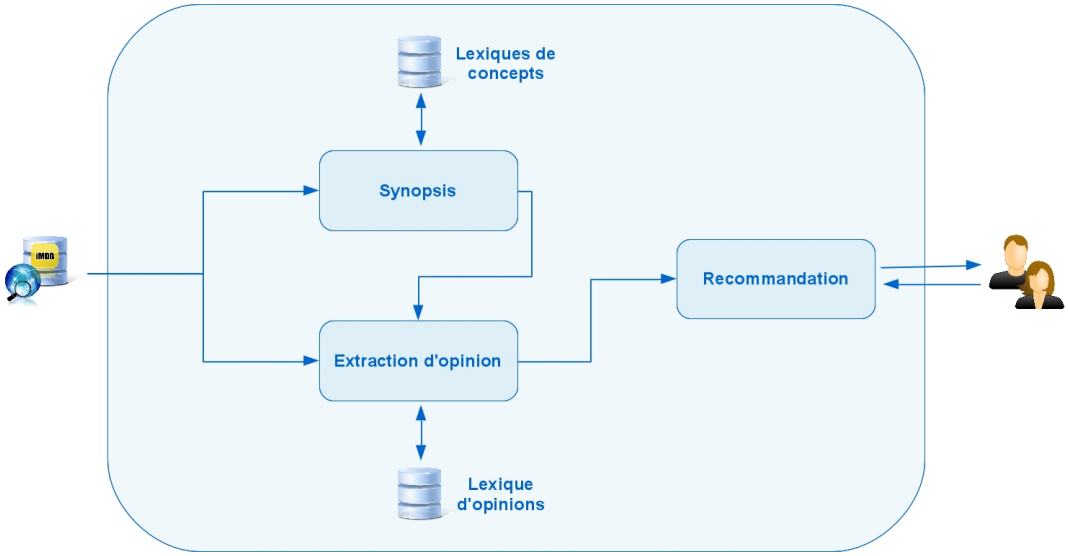

FIGURE 6.10 – Architecture du systèmes de recommandation

critiques par film), fournissent assez d'information pour avoir une idée suffisante de l'opinion globale des gens sur un film. L'idée, est de fusionner les intervalles fournis par notre système d'extraction multicritère, pour obtenir des distributions de possibilités relatives à chacun des critères considérés (c.f. chapitre 4 section 6.3.2). Les journaux fournissent seulement des notes sous forme "d'étoiles", nous transformons cette notation par le biais des distributions de possibilités, puis nous utilisons la méthode décrite en section 6.3.3.2 du chapitre 6 pour extraire les poids relatifs à chacun des critères. On en déduit la politique d'évaluation du journal, c'est-à-dire dans cet exemple à deux critères, l'importance relative des deux critères dans la perception du cinéma par le journal. Après, selon les préférences d'un utilisateur, il lui sera facile de repérer le journal qui partage le mieux ses valeurs (c'est-à-dire qui a les mêmes priorités sur les critères d'évaluation que lui) et de majoritairement se fier aux critiques issues de ce journal.

La figure 6.12 montre un exemple de résultats sur le problème considéré. Pour chaque journal, les deux barres (la bleue et la rouge) représentent les valeurs des poids qui ont été identifiées (leur somme fait 1) à l'aide de l'équation 6.12. Il devient alors facile d'identifier les journaux qui se rapprochent le plus du système de préférences de l'utilisateur et de les utiliser en priorité pour une recommandation personnalisée. .

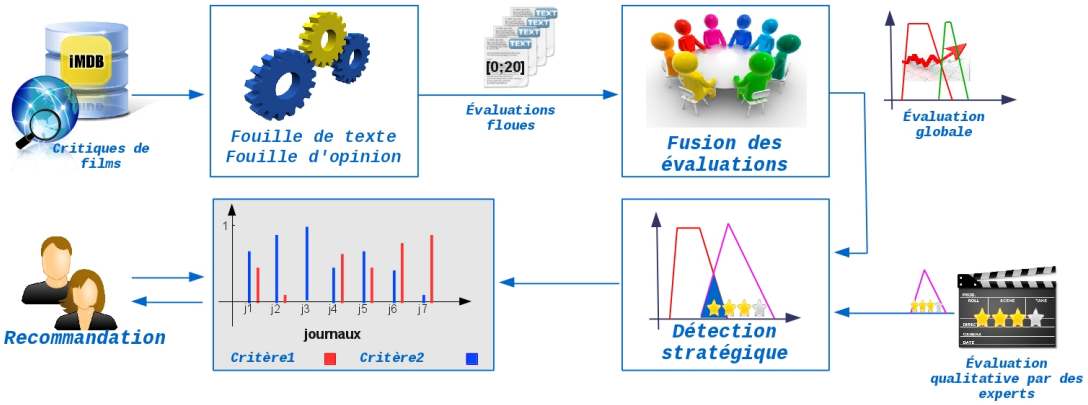

FIGURE 6.11 – Architecture de détection de stratégique

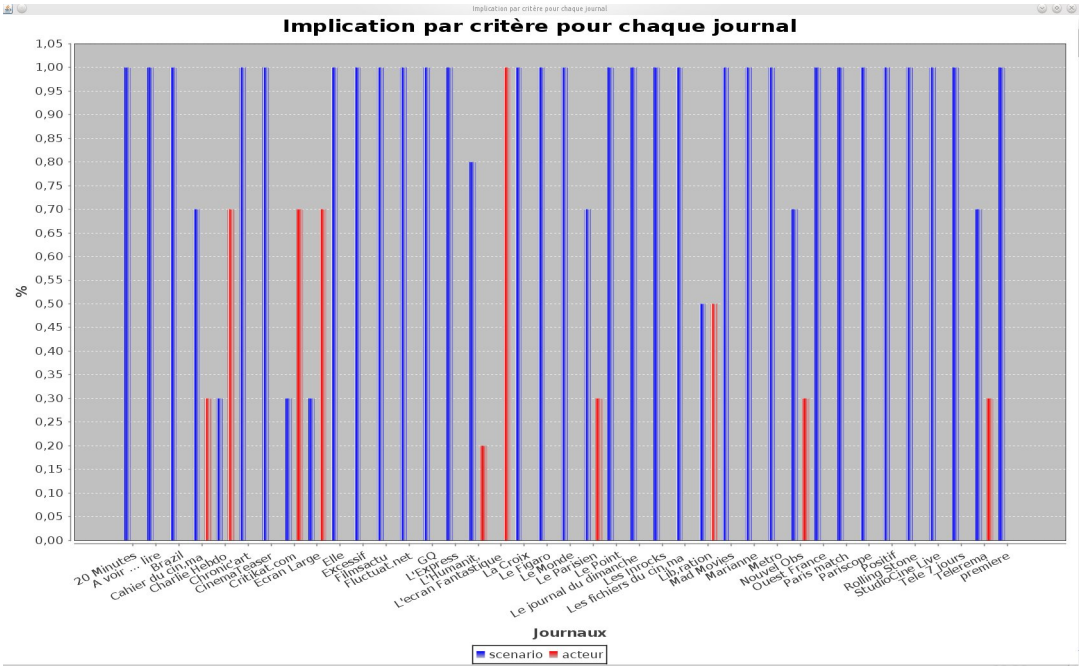

FIGURE 6.12 – Exemple de détection de stratégie.

6.4 Discussion

Dans ce chapitre, nous avons proposé un système de recommandation qui permet de traiter des évaluations imprécises (floues). Nous avons également proposé une méthode qui permet l'identification des poids considérés sur chacun des critères par l'évaluateur en adaptant la méthode des moindres carrées au cas imprécis

des évaluations expertes. Nous avons, grâce à ces calculs, pu offrir un système de recommandation personnalisé, en recommandant à l'utilisateur les entités les plus en adéquation avec son système de préférences. Nous avons montré que la mise en place d'un système de recommandation évolué est possible en utilisant des techniques automatiques de fouille de textes et de fouille d'opinions avec une intervention humaine minimale. Ceci constitue une avancée importante car jusqu'à présent la nécessité d'évaluer manuellement un grand nombre de documents et ce selon plusieurs critères, représentait un frein majeur à la mise en place de tels systèmes.

Le couplage de notre approche d'extraction conceptuelle pour les critères et d'extraction d'opinions pour les évaluations avec les modèles et outils de l'analyse multicritère proposés dans ce chapitre a permis de montrer que quelle que soit la nature du système de recommandation multicritère dans la classification d'Amovicius (c.f. chapitre 2), notre approche permet d'en faire un système de recommandation de type MMPE pour une recommandation plus fiable, pertinente et personnalisée. La chaîne de traitement que nous proposons dans nos travaux établit un processus d'automatisation cognitive qui laisse raisonnablement envisager un large déploiement et la généralisation des SIAD de type MMPE dans les années à venir pour traiter des données issues du web.

Chapitre 7

Conclusion générale

"Toute connaissance est une réponse à une question."

Gaston Bachelard

Sommaire

Le web 2.0 a considérablement changé le comportement des internautes en leur permettant d'échanger de l'information et d'interagir de façon simple, à la fois avec le contenu et la structure des pages web, mais aussi entre eux, créant ainsi notamment le Web social. L'internaute devient, grâce aux outils mis à sa disposition, une personne active sur la toile. Cette facilité d'expression a permis aux internautes de pouvoir s'exprimer en donnant leurs opinions sur de multiples sujets dans différents domaines (e-commerce, politique, etc.). Ces avis disponibles sur le web ont très vite intéressé les internautes lorsqu'il s'agissait de se faire une opinion sur un nouveau produit avant de l'acheter par exemple. En profitant de la connaissance des uns, il devient alors plus facile de décider. Aujourd'hui, la réalité est tout autre car les avis sont si nombreux que l'internaute est plongé dans l'indécision la plus totale et il n'arrive plus à accorder de crédibilité à des avis subjectifs et souvent contradictoires. Les systèmes de recommandation apportent un soutien à l'utilisateur en filtrant l'information pertinente sur la base d'un profil utilisateur rudimentaire. Les statistiques sur un grand nombre de critiques et des évaluations seulement qualitatives sont les outils employés par les systèmes de recommandation pour traiter les problèmes d'imprécision et d'incertitude dans le processus d'évaluation par monsieur tout le monde. Pour être en mesure de répondre au mieux au exigences de l'utilisateur, les systèmes de recommandation sont désormais multicritères, cependant, leur mise en place reste lourde car elle nécessite une implication des internautes beaucoup trop lourde. En effet, quel internaute a envie de prendre du temps pour évaluer un produit ou un service sur un ensemble de critères parmi lesquels peut-être deux ou trois seulement l'intéressent ? Au-delà de l'investissement chronophage que cela représente, les internautes qui contribuent au bon fonctionnement de ces plate-formes collaboratives ne sont pas des experts dont on peut attendre des remarques pertinentes sur une multitude de critères. La plupart du temps, leurs critiques ne sont que la mise en forme d'un retour d'expérience, une "petite rédaction pour exprimer leur ressenti". Dans ces travaux, nous avons expliqué pourquoi les systèmes de recommandation étaient les outils les plus précis pour l'internaute car ils permettent de comprendre quel critère d'évaluation a fait la bonne ou la mauvaise réputation d'un produit ou d'un service. Cependant, si l'on veut profiter du comportement collaboratif des internautes qui alimentent les bases web de critiques, il ne faut pas les contraindre à changer leur mode d'expression : la critique en langage naturel. Il faut alors au contraire faire évoluer les outils d'analyse de l'informatique : être capables d'identifier systématiquement les critères évoqués dans une critique et de leur asso-

cier automatiquement une évaluation détectée par la machine, autrement dit tenter d'interpréter, en termes d'évaluation par critère, la donnée brute que constitue la critique en langage naturel.

Nous avons proposé dans cette thèse des méthodes qui permettent d'extraire automatiquement les informations utiles contenues dans des critiques écrites en langage naturel pour procéder à une évaluation multicritère systématique. La première concerne l'extraction des critères (ou extraction conceptuelle) et la seconde concerne l'extraction d'opinion (ou analyse de sentiments). Nous avons montré que ces deux méthodes dédiées à l'extraction automatique de connaissances pouvaient, si elles étaient associées à des outils d'analyse multicritère adéquats, permettre de transformer automatiquement un système de recommandation classique en un système de recommandation multicritère avec une intervention humaine minimale. Toutefois, si le travail réalisé permet de répondre à cette problématique d'extraction des données pour les systèmes de recommandation et offre des outils efficaces, certains problèmes restent non résolus. Nous proposons donc de synthétiser nos contributions à cette problématique vaste de la recommandation multicritère dans la section 7.1 , puis nous donnons un certain nombre de perspectives dans la section 7.2 pour palier leurs limites.

7.1 Travail réalisé

Dans le cadre de cette thèse, nous nous sommes intéressés à l'extraction de critères et à l'extraction d'opinions multicritères dans des données textuelles écrites en langage naturel. Nous dressons dans cette section un bilan de ce travail.

7.1.1 Extraction de critères

Tout d'abord, nous avons montré que l'extraction des critères revient à segmenter un texte par rapport à un ensemble de concepts, où chaque critère correspond à un concept précis. Nous avons ensuite mis en évidence qu'il est nécessaire de tenir compte de la dimension pragmatique du langage pour être en mesure d'identifier les différentes interprétations possibles d'un texte, le vocabulaire employé dépendant directement du niveau d'expertise du lecteur sur le concept considéré.

Nous avons montré dans le chapitre de l'état de l'art que les méthodes actuelles de segmentation qui pourraient permettre l'extraction de critère ne sont pas adaptées à notre contexte web. En effet, soit elles sont complètement non supervisées, mais

elles ne sont pas en mesure d'indexer les segments par leur thématique (elles ne détectent que les changements thématiques), soit elles sont supervisées, mais alors elles nécessitent une intervention humaine trop conséquente lors de la phase de construction du corpus d'apprentissage ou lors de la construction d'une connaissance extérieure comme un thésaurus ou une ontologie par exemple. Notons encore que l'objectif est une recommandation toujours plus personnelle ; à terme, l'utilisateur pourrait rentrer ses propres critères, ce qui rend définitivement inaptes les techniques supervisées.

Nous avons proposé une méthode statistique appelée *Synopsis* pour extraire les parties d'un document qui traitent de concepts définis au préalable. *Synopsis* tient compte des différentes interprétations possibles du texte considéré grâce à un contrôle du lexique associé à un concept (précision dans l'expression propre au niveau d'expertise). Le fait d'introduire cette dimension pragmatique dans un processus de segmentation de texte est une contribution significative compte tenu de l'hétérogénéité des documents disponibles sur le web. De plus, *Synopsis* est une méthode que l'on peut considérer comme "non-supervisée", car elle ne nécessite qu'une expertise infime qui réside en la définition, par quelques mots germes, du concept à désigner. Ces mots germes sont nécessaires pour construire automatiquement un corpus d'apprentissage et effectuer un apprentissage automatique du vocabulaire propre au concept considéré (la sémantique associée à un concept est déterminée par la contextualistion des mots germes). Nous avons proposé une technique d'apprentissage en raisonnant par analogie avec les principes d'apprentissage du langage chez l'enfant. Nous ne prétendons pas avoir étudié précisément les mécanismes adoptés par l'enfant au cours de l'apprentissage de la langue, mais nous nous sommes plutôt inspirés des principes en se référant en particulier à la sémiotique de Piaget. Nous nous sommes également intéressés aux travaux de [Morris 1938] qui considère que le mot a trois dimensions (syntaxique, sémantique et pragmatique), et nous avons mis en évidence qu'il était nécessaire de considérer chacune de ces dimensions tout au long du processus pour être en mesure d'identifier les différentes interprétations possible d'un texte. L'interprétation d'un texte par *Synopsis* peut donc être vue comme une fonction de la syntaxe, mais aussi de la sémantique (constitution du lexique sur la base de la contextualisation par les mots germes) et du pragmatisme du texte (gestion du niveau d'expertise par la taille du lexique appris).

L'approche *Synopsis* a été utilisée dans plusieurs domaines comme le cinéma, la restauration, la politique (application *StraussOp*) ainsi que le domaine médical

(application de Recherche d'Informations *Ontolex*). Nous sommes conscients qu'il serait nécessaire de valider l'approche dans d'autre domaines d'application, mais la difficulté à trouver des corpus de tests admis par la communauté est un vrai problème et il n'existe pas, à notre connaissance, de corpus de test qui offriraient un ensemble de documents segmentés où chaque extrait traitant de concepts spécifiques soient annotés par ceux-ci. Nous avons cherché dans la communauté de la Recherche d'Information notamment, mais les corpus de références comme le TREC n'offrent pas, à notre connaissance, une annotation de ce type. L'annotation proposée est relative à une requête vue comme un ensemble de concepts en interaction et non pas à un seul concept, il n'est donc pas possible de valider notre approche sur ce type de corpus de références.

Nous avons proposé d'utiliser *Synopsis* dans un système de recherche d'informations conceptuelle car nous avons constaté que les systèmes de recherche d'informations basées sur des ontologies (les requêtes d'une recherche conceptuelle sont les concepts de l'ontologie) ont selon nous une faiblesse majeure lorsqu'il s'agit de restituer à l'utilisateur les documents réputés pertinents pour la requête conceptuelle (c.f. chapitre 6). En effet, le système de RI (Recherche d'Information) manipule des concepts en fonction de l'annotation conceptuelle des documents, mais le plus souvent ne dispose pas des extraits, des passages du document (pas de recherche full text) qui traitent du ou des concepts recherchés. C'est à l'utilisateur de faire le lien entre les concepts de sa requête et les documents qui lui sont restitués. Nous avons pensé qu'il était judicieux de montrer à l'utilisateur les passages du document traitant du concept (c'est-à-dire les passages où l'on trouve des mots caractéristiques du lexique propre au concept) et ainsi de faciliter sa lecture et un accès rapide à l'information pertinente. Par ailleurs, cette fonctionnalité confère au système de RI conceptuelle une capacité à justifier ses résultats qui rend plus transparente la notion de pertinence. Nous pensons que cette dimension explicative est primordiale pour l'utilisateur, et le fait de lui proposer les extraits relatifs à ce qu'il cherche est un plus non-négligeable.

7.1.2 Extraction d'opinion

Nous avons montré dans le chapitre de l'état de l'art que les méthodes actuelles d'extraction d'opinion nécessitent une intervention humaine conséquente, soit pour construire manuellement une base de connaissances recensant les descripteurs d'opinion sous forme d'ontologie, de thésaurus, etc, soit pour construire une base d'ap-

prentissage. Cette expertise nuit selon nous à la diffusion des techniques d'opinion mining. De plus, nous avons montré dans de précédent travaux [Harb *et al.* 2008] que le vocabulaire d'opinion est propre au domaine dans lequel il est employé, ce qui multiplie la charge de travail car il faut construire une base d'apprentissage pour chacun des domaines considérés pour être en mesure d'apprendre le vocabulaire d'opinion spécifique aux domaines considérés. Il s'agit là d'une contrainte quasi rédhibitoire. C'est pourquoi nous avons adapté l'approche *Synopsis* à l'apprentissage des descripteurs d'opinion et ainsi pouvoir construire automatiquement une base d'apprentissage et un lexique de descripteurs d'opinion relatif à un domaine.

Nous avons mis en évidence qu'il était nécessaire de considérer quatre classes (positif, non-positif, négatif, non-négatif) pour assurer un apprentissage pertinent des descripteurs. Cette considération est nécessaire car les descripteurs d'opinion peuvent être présents dans plusieurs classes à la fois (c.f. section sur les relations sémiotiques du chapitre 4).
La quantification de la polarité des descripteurs est à ce jour un réel problème que nous ne savons pas encore résoudre sans intervention humaine, c'est pourquoi nous avons limité notre validation à la classification binaire (positif/négatif). Les scores que nous obtenons sur chacun des descripteurs ne doivent pas être théoriquement assimilés à cette quantification, la polarité est en fait pondérée par la représentativité (ou l'intensité) du critère dans le texte (combien le texte parle en mal ou en bien de tel ou tel critère).
Nous avons proposé une méthode qui permet de déterminer l'orientation sémantique d'un texte (positif/négatif) en utilisant le lexique d'opinion construit lors de la phase d'apprentissage. Nous avons par la suite étendu cette méthode au multicritère, où l'idée est d'extraire les opinions relatives à un ensemble de critères. En cherchant à extraire l'opinion par rapport à des critères nous avons été amenés à combiner les approches d'extraction de critères (*Synopsis*) et d'extraction d'opinions. Nous avons montré que combiner l'approche *Synopsis* à l'extraction d'opinion apportait une dimension pragmatique à l'évaluation et qu'il était nécessaire de la considérer pour obtenir une extraction plus fiable : en effet, l'extraction conceptuelle fournit une segmentation thématique, fonction du lexique associé au concept, or ce lexique dépend du niveau d'expertise avec lequel le texte a été analysé, par suite la détection d'opinions portera sur des segments différents selon la granularité d'analyse. Les différentes possibilités d'interprétation d'un texte peuvent nuancer voire inverser la polarité d'un texte selon le point de vue ou le niveau d'expertise considéré. Nous

avons donc choisi de représenter l'évaluation par un quidam sous la forme d'un intervalle de valeurs possibles afin de modéliser l'imprécision due à la méconnaissance du niveau d'expertise de ce quidam.

7.1.3 Système de recommandation

La contribution principale de cette thèse est l'extraction des connaissances (extractions conceptuelle et d'opinion) pour les systèmes de recommandation. Nous avons focalisé notre étude sur les systèmes de recommandation multicritère de type MMPE. Nous avons montré qu'il était nécessaire de développer des outils d'extraction des connaissances pour pouvoir mettre en place de tels systèmes qui sont à ce jour peu utilisés car ils nécessitent une intervention humaine trop importante qui rebute utilisateurs et administrateurs de site.

Nous avons proposé d'intégrer nos méthodes d'extraction automatique des connaissances à des outils d'analyse multicritère adéquats pour concevoir un système de recommandation de type MMPE. Notre méthode d'extraction d'opinion multicritère permet d'évaluer un ensemble de textes selon une liste de critères. A chaque critère est donc associé, par notre extraction d'opinion, un ensemble d'évaluations imprécises modélisées sous la forme d'intervalles. Puisque les évaluations sont imprécises, il n'est pas possible d'avoir recours aux probabilités pour construire la distribution représentative d'un critère. Nous avons donc choisi de construire des distributions de possibilité pour modéliser l'évaluation incertaine associée à chaque critère. Nous avons ensuite utilisé des opérations arithmétiques élémentaires sur les nombres flous pour calculer une évaluation globale sur la base des évaluations partielles des critères. Nous avons expliqué (lorsque c'est possible) comment synthétiser le modèle de préférence d'un utilisateur sous une forme analytique. Enfin, une analyse de sensibilité de cette forme analytique nous fournit l'explication en termes de critères des propositions du système de recommandation.

Nous avons choisi de valider notre approche d'extraction d'opinion sur le corpus proposé par [Pang & Lee 2002] sur le domaine du cinéma ainsi que sur le domaine politique à travers l'application *StraussOp*. Nous sommes conscients qu'il serait nécessaire de valider cette approche dans d'autres domaines d'application pour vérifier si l'extraction reste efficace.

7.2 Perspectives

Ces travaux offrent de nombreuses perspectives, nous proposons dans cette section d'en exposer les principales.

7.2.1 Extraction de critères

Synopsis pourrait être étendue au contexte d'indexation automatique de documents selon des concepts définis, en utilisant une ontologie par exemple. L'idée n'est pas de se limiter à la classification de documents (le texte parle ou ne parle pas du concept), mais de calculer la représentativité du concept dans le document afin de construire un modèle vectoriel classique en recherche d'informations, en clustering de textes, etc. Ces composantes de représentativité pourraient être modélisées par un intervalle si l'on tient compte du niveau de pragmatisme (on ne peut calculer qu'un intervalle de valeurs pour caractériser la représentativité d'un concept dans un document car son appréciation dépend du niveau d'expertise d'un quidam).

Une application intéressante concerne le tri automatique de dépêches. À ce jour les journaux sont très friands de ce type d'application qui leur permettent de gagner un temps précieux en triant automatiquement un grand nombre de dépêches reçues en temps réel par leur système d'information. La plupart de ces applications sont aujourd'hui basées sur la recherche, dans les dépêches, de quelques mots clés définis par la rédaction. Malheureusement ces approches basées sur les mots clés n'identifient pas de nombreuses dépêches pourtant pertinentes simplement parce que le vocabulaire employé par le journaliste n'est pas identique au lexique constitué par les mots clés de la rédaction. Une approche conceptuelle comme Synopsis devrait typiquement régler ce problème de classification en permettant de classer toutes les dépêches reçues, y compris celles qui ne comporteraient pas les mots clés choisis.

Dans un contexte d'aide à l'écriture, il est possible d'utiliser *Synopsis* pour aider le locuteur à préciser son discours. En effet, en définissant le concept que le locuteur veut exprimer, le système est en mesure d'identifier les interprétations possibles du concept dans le texte proposé. En fonction du résultat, le locuteur peut rectifier son propos en précisant les points qui ont été jugés flous par *Synopsis* par exemple, c'est-à-dire les passages où le système n'a pas identifié correctement le concept exprimé par le locuteur. De plus, le recours à un tel processus met en avant la dimension pragmatique du concept, et ainsi, aide le locuteur à avoir des propos interprétables par un maximum de lecteurs.

7.2.2 Extraction d'opinion

La quantification d'opinion est un problème sur lequel nous concentrons actuellement nos efforts. L'objectif est d'obtenir une échelle ordinale discrète qui permettrait d'évaluer automatiquement l'opinion selon différents degrés comme "très bien", "bien", "neutre","mauvais", "très mauvais", mais l'objectif final est de réussir à établir une échelle cardinale des descripteurs. Nos recherches actuelles ont montré que l'utilisation de mots germes quantifiés n'était pas suffisante. Par exemple, pour obtenir un lexique de descripteurs très positifs, il ne suffit pas de considérer comme mots germes "very good", "perfect", "amazing", "mind-blowing", etc, pour générer un lexique significatif. Nous pensons qu'il faut attacher plus d'importance au contexte et à la syntaxe, en particulier en intégrant la notion de N-grammes pour résoudre une partie du problème.

Les applications à l'extraction d'opinions sont nombreuses. Citons encore le domaine politique très demandeur de ce type d'applications qui permettent notamment l'analyse automatique de textes de loi par exemple. Les objectifs peuvent être multiples, comme la détection de contradiction dans le débat, ou alors l'analyse de l'opinion des électeurs sur un candidat (c.f. chapitre 5), l'analyse de tendances, etc. L'analyse de rumeurs pourrait se baser sur l'écart (ou l'évolution de l'écart) qui peut exister entre l'opinion officielle diffusée par les médias et l'opinion extraite des blogs et autres supports moins formels.

7.2.3 Système de recommandation

En améliorant notre processus d'extraction des connaissances et plus particulièrement si nous arrivons à quantifier l'opinion extraite, nous serons en mesure d'obtenir pour chaque item considéré à la fois l'intensité de chacun des critères sur chaque item et l'intensité d'opinion qui est exprimée sur chacun de ceux-ci. Le système serait plus précis et les recommandations plus fiables.

Greffer nos techniques d'extraction automatique de connaissances (extractions conceptuelle et d'opinion) à des outils d'analyse multicritère dans un contexte d'évaluations imprécises (typiquement les opinions) nous laisse envisager la transformation automatisée de n'importe quel site de recommandation en un site de recommandation multicritère de type MMPE. Il nous suffit de récupérer les critiques du site, de les décomposer par rapport à un référentiel de critères, puis de procéder aux évaluations.

Notons que cette chaîne de traitement automatisée permet également d'envisager

la recommandation personnalisée. Au lieu de se contenter des critères imposés par la FNAC par exemple, l'utilisateur pourra spécifier lui-même ses critères d'intérêt et obtenir des tableaux d'évaluation personnalisés qui confrontent uniquement les items de son choix à partir de ses propres critères. Ce sera la fin du bruit organisé par le site commerçant dans la recommandation.

Annexe A

Algorithmes de fouille de données existants

Sommaire

A.1 Le mot comme descripteur

Un texte écrit en langage naturel est constitué de mots qui permettent l'expression d'idées, de sentiments, etc. Deux types de mots peuvent être identifiés : les mots variables, et les invariables. Les mots invariables sont les interjections, les propositions, les conjonctions et les prépositions. Les mots variables comme les noms, les verbes, les articles, les pronoms ou les adjectifs, ont la propriété de pouvoir être conjugués ou déclinés. La forme brute d'un mot, c'est à dire, non conjugué et non décliné, est appelé **lemme**. Le lemme est l'unité autonome constituante du lexique d'une langue. C'est une suite de caractères formant une unité sémantique et pouvant constituer une entrée de lexique.

Cependant, le fait de considérer le mot comme élément unitaire est souvent insuffisant notamment pour les mots successifs qui dépendent les uns des autres pour être compréhensibles ("corde à sauter" par exemple). La notion de **n-grammes** est apparue la première fois, selon [Shannon 1948]. [Shannon 1948] introduisait la notion n-gramme dans un système prédictif qui permettait de deviner les caractères suivant une suite de caractères précédemment entrées. Un n-gramme peut être vu comme une suite de mots(ou lemme) consécutifs, qui dépendent les uns des autres, c'est à dire que les mots successifs sont liés. Par exemple, le tri-gramme "corde à sauter" peut être considéré comme un seul mot. Cette représentation, permet de désambiguïser les mots composés, et il est ensuite possible de les considérer comme descripteur.

A.2 Sélection des descripteurs

La sélection des descripteurs est le plus souvent réalisée par des approches statistiques qui ont pour objectifs de déterminer l'appartenance d'un descripteur dans un corpus d'apprentissage par exemple. Nous appelons approches statistiques, les approches qui se basent sur la fréquence d'apparitions des termes, et notamment sur leur répétition. Cette sélection est uniquement basée sur la présence des mots dans un contexte précisément établi (textes choisis).

A.2.1 Le *tf-idf*

L'approche la plus utilisée dans la littérature reste le *tf-idf* [Salton & Yang 1973]. Le *tf* signifie "term frequency", et le *idf* "inverse document frequency". L'objectif est de déterminer la représentativité (ou importance) d'un descripteur pour un corpus

donné. Le principe est de pondérer la méthode fréquentielle *tf* par le nombre de documents dans lesquels le terme considéré apparait *df*. Cette pondération est issue de la recherche d'information. Elle est basée sur la loi de Zipf [Zipf 1941] qui considère qu'un terme très fréquent n'est pas nécessairement représentatif du corpus, et qu'il n'est pas, non plus, le plus informatif. Plus formellement, le *tf-idf* peut s'écrire comme suit pour un descripteur i dans un document j :

$$w_{ij} = tf_{ij} \times idf_i \tag{A.1}$$

où

$$idf_i = log\frac{N}{n_i} \tag{A.2}$$

où N est le nombre de documents du corpus, n_i le nombre de documents dans lequel apparait le ième descripteur.

A.2.2 L'entropie

L'entropie s'avère une mesure efficace pour déterminer la dispersion d'un descripteur dans un corpus. Nous nous intéressons ici à la $log-entropy$ telle que décrite dans [Dumais 1991] qui révèle par ailleurs que cette approche obtient les meilleurs résultats pour des tâches de RI que le $tf - idf$. Plus formellement l'entropie est définie par :

$$E(i) = \sum_j \frac{p_{ij} log_2(p_{ij}}{log_2 N} \tag{A.3}$$

avec

$$p_{ij} = \frac{tf_{ij}}{gf_i} \tag{A.4}$$

où gf_i représente le nombre total de fois où le descripteur i apparaît dans le corpus de N documents. Une représentation avec l'approche fréquentielle (tf) peut alors être la suivante avec pour un terme i et un document j :

$$w_{ij} = (1 + E(i))log(tf_{ij} + 1) \tag{A.5}$$

Cette approche permet une meilleure répartition des probabilités et est moins restrictive que le $tf - idf$ qui attribue une probabilité de 0 à un descripteur présent dans tous les documents du corpus par exemple.

A.3 La similarité

La plupart des approches de sélection de descripteurs adoptent une représentation vectorielle des documents. Outre une représentation vectorielle de qualité, il est également nécessaire de définir une mesure permettant de comparer la proximité des différents vecteurs afin de pouvoir, par exemple, regrouper des termes ou des documents. Ainsi, une telle proximité correspondra à des vecteurs proches. L'objectif final est une classification de termes ou de documents.

A.3.1 Le coefficient de Jaccard

Cette mesure introduite par Paul Jaccard, permet de traduire le nombre de descripteurs communs à deux documents par le nombre de descripteurs non communs aux deux documents. Plus formellement ce coefficient est défini par :

$$sim_{Jaccard} = \frac{D_{commun}}{D_{total} - D_{commun}} \tag{A.6}$$

avec D_{commun} le nombre total de descripteurs communs et D_{total} le nombre de descripteurs des deux documents considérés. Cette mesure s'intéresse uniquement à des valeurs binaires présence/absence.

A.3.2 Le cosinus

La mesure de similarité *cosinus* est l'une des plus utilisées, surtout dans le domaine de la Recherche d'Information [Salton & Yang 1973]. Elle mesure le degré de similarité entre deux vecteurs par l'intermédiaire d'un calcul d'angle compris entre 0 et 1. Une valeur de 1 indique une forte proximité des vecteurs, et elle correspond à un angle faible. À l'inverse, une valeur de 0 indique un angle important entre les vecteurs, ce qui marque une forte dissimilarité entre les vecteurs. Le cosinus entre deux vecteurs est obtenu en calculant le produit scalaire entre ces deux vecteurs ($\overrightarrow{u}$ et $\overrightarrow{v}$) , que nous divisons par le produit de la norme des deux vecteurs. Soit :

$$\theta = arcos \frac{\overrightarrow{u} . \overrightarrow{v}}{\| \overrightarrow{u} \| . \| \overrightarrow{v} \|} \tag{A.7}$$

A.4 L'apprentissage

L'apprentissage consiste à acquérir des connaissances à partir de phénomènes observés. L'objectif étant par la suite, d'être en mesure d'identifier des connais-

sances similaires à celles apprises, c'est à dire que contrairement aux approches par sélections, le système apprend dans un **contexte**, ce qui lui permet de considérer (d'apprendre) comme descripteurs tous les mots présents dans les documents sélectionnés. Ce principe permet, plutôt que de considérer les "n premiers mots" (les plus fréquents) comme le font les approches par sélection (c.f. section A.2), de prendre en compte l'ensemble des mots caractérisant le corpus d'apprentissage. Ceci permet aussi de gagner en précision lors d'un classement de documents (ou de mots) par exemple, car le calcul de similarité sera plus précis, dû au fait que l'apprentissage fournit plus d'informations (nombre de descripteurs) que les approches par sélection. Nous distinguons alors deux types d'apprentissage : l'**apprentissage supervisé** et l'**apprentissage non supervisé**. Ces deux types d'apprentissage ont des objectifs généralement différents. L'apprentissage supervisé est utilisé lors de tâches d'extraction ou de classification, l'apprentissage non supervisé est plutôt utilisé à des fins exploratoires. L'apprentissage cherche à caractériser le mieux possible le corpus sur lequel il apprend, pour, dans un second temps, retrouver des documents similaires à ceux constituant les corpus d'apprentissage. Ce type d'apprentissage est très dépendant de la qualité du corpus d'apprentissage, et c'est généralement sur cette partie qu'il est nécessaire de consacrer un maximum d'attention. L'apprentissage non supervisé a pour but de découvrir des groupes homogènes (classes) de documents (ou termes) en n'ayant aucune idée du nombre de groupes.

A.4.1 Apprentissage supervisé

Nous présentons ci-dessous un nombre non exhaustif d'algorithmes d'apprentissage supervisés couramment utilisés lors de processus de classification. Nous en distinguons cinq types :

- Les probabilistes
- Les approches fondées sur les réseaux de neurones
- Les "plus proches voisins"
- Les approches minimisant l'erreur de classification

A.4.1.1 Les probabilistes

Les approches probabilistes consistent à classer un nouvel élément dans une classe en fonction de sa probabilité d'appartenance à cette classe. L'algorithme le plus répandu est le *Naïves Bayes*. Le classifieur de type Naïve Bayes est fondé sur le théorème de Bayes [Bayes 1763]. Considérons $v_j = (v_{j1}, \ldots, v_{jk}, \ldots, v_{jd})$ un vecteur

de variables aléatoires représentant un document d_j et C un ensemble de classes. La probabilité que d_j appartienne à la classe c_i est définie par :

$$P(c_i \mid v_j) = \frac{P(c_i)P(v_j \mid c_j)}{P(v_j)} \tag{A.8}$$

La variable aléatoire v_{jk} du vecteur v_j représente l'occurrence de l'unité linguistique retenue pour la classification du document d_j. La classe d'appartenance c_k du document est déterminée par la probabilité maximale d'appartenir à c_k. Plus formellement :

$$c_k = arg\ max\ P(c_i \in C) \prod_k P(v_{jk} \mid c_j) \tag{A.9}$$

$P(c_i)$ est alors définie par :

$$P(c_i) = \frac{nombre\ de\ documents \in c_i}{nombre\ total\ de\ documents} \tag{A.10}$$

De plus, en faisant l'hypothèse d'indépendance sur les v_j, la probabilité conditionnelle $P(v_j \mid c_j)$ est définie par :

$$P(v_j \mid c_j) = P(v_{jk} \mid c_j) \tag{A.11}$$

A.4.1.2 Les approches fondées sur les réseaux de neurones

Les approches neuronales se basent sur le fonctionnement du système nerveux humain. Elles utilisent des neurones artificiels qui permettent d'effectuer les tâches d'apprentissage. Les automates à seuil permettant de modéliser l'activité neuronale [Mcculloch & Pitts 1943] ainsi que les règles d'apprentissage locales introduites par [Hebb 1961] et [Widrow & Stearns 1985] qui ont fortement inspirés les méthodes d'apprentissages neuronal. Les premières approches étaient "monocouches", c'est le cas du *Perceptron* proposé par [Rosenblatt 1988], mais aussi de *Adaline* [Widrow & Hoff 1988]. Puis sont apparus les méthodes multicouches, dont le *Perceptron multicouche* [Cybenko 1989].

Un réseau de neurones artificiels est composé d'une ou plusieurs couches qui se succèdent, où chaque entrée est la sortie de la couche précédente. Ainsi, la nième couche du réseau aura pour entrées, les sorties de la couche n-1ème. Les neurones humains sont reliés par l'intermédiaire de *synapses*. Dans un réseaux de neurones artificiels, les synapses sont modélisées par des liaisons pondérées unidirectionnelles. Un réseau neuronal peut finalement être vu comme un graphe orienté dont les nœuds

sont des neurones artificiels. L'objectif ensuite, est de déterminer les poids à mettre sur chacun des arcs pour obtenir le modèle attendu, c'est à dire, pour une entrée donnée, la sortie souhaitée.

A.4.1.3 Les "plus proches voisins"

Les approches de type "plus proches voisins" cherchent à définir des zones propres à un concept. Une approche classique est l'algorithme *k-ppv* [Cover & Hart 1967].
Le principe est de mesurer la similarité entre un nouveau document et les documents déjà classés. Cela revient à constituer un espace vectoriel où chaque document est modélisé par un vecteur de mots. L'idée ensuite est de déterminer par un classement, les documents les plus similaires au document à classer. Le classement dépendant du score de la fonction de similarité, généralement le *cosinus* est utilisé. Selon la valeur de k.
Pour classer le nouveau document à partir des k plus proches voisins, k correspondant au nombre de document retenus, nous distinguons deux types d'approches :

- Classer le document dans la même classe que le document qui a eu la plus grande similarité au sein du jeu d'apprentissage [Yang & Liu 1999].
- si k>1 de considérer les k meilleurs documents. Deux méthodes sont envisageables :
  - Calculer parmi les k documents les plus proches pour chacune des classes potentielles et considérer la classe à laquelle est rattaché un maximum des k documents.
  - Considérer le rang de chacun des k documents. Il s'agit pour toutes les classes, d'effectuer la somme des occurrences d'une classe multipliée par l'inverse de son rang.

A.4.1.4 Les approches minimisant l'erreur de classification

Ces approches cherchent à minimiser l'erreur de classification. L'algorithme le plus populaire est celui des machines à support vectoriel connu sous le sigle "SVM". SVM est un classifieur binaire. La première étape de l'algorithme est de projeter les données dans un espace vectoriel. Ce type de classifieur part du principe que les données sont linéairement séparables, et qu'il existe une fonction de transformation appelée "noyau" qui permet de rendre le problème linéaire. Les noyaux les plus fréquemment utilisés sont linéaires, polynomiaux ou gaussiens. L'idée ensuite est de

chercher un hyperplan afin de séparer, au mieux, les exemples positifs des exemples négatifs afin d'obtenir la marge (frontière) la plus importante entre les positifs et les négatifs. Cette approche est facilement adaptable aux problèmes multi-classes en utilisant le principe "One-against-the-rest". Cela revient à comparer chaque classe à l'ensemble des autres pour trouver un hyperplan séparateur.

A.4.2 Apprentissage non-supervisé

Les approches non supervisées se différencient des approches supervisées par le fait qu'elles ne disposent pas de connaissances sur les classes à identifier. De plus, aucune des données disponibles ne sont étiquetées, nous ne pouvons donc pas construire de modèle d'apprentissage. L'objectif ne se limite donc plus à seulement être capable de classer des documents, mais aussi à être en mesure de détecter les classes présentent dans le corpus. Nous parlerons de "cluster" plutôt que de classes dans ce cas. La notion de "plus proches voisins" est très souvent utilisée dans ce type d'apprentissage non supervisé, notamment avec l'approche *k-moyennes* ou "k-means" [MacQueen 1967].

A.4.2.1 K-means

La première étape de l'approche est de sélectionner arbitrairement k centres autour desquels sont regroupés les éléments les plus proches de ces centres. Il est alors calculé le centre de gravité de chaque classe ainsi définie au fur et à mesure du regroupement des éléments. Ces derniers vont définir les nouveaux centres des classes. Cette opération est répétée jusqu'à ce que la dispersion des membres de chaque classe soit minimale. Les classes deviennent en effet à chaque itération plus compactes, permettant une convergence de l'algorithme. Cet algorithme a évolué, et de nouveaux algorithmes comme *x-means* [Pelleg & Moore 2000] et *c-means* [Bezdek 1981] ont vu le jour. Ce dernier permet une classification floue, où chaque document peut appartenir à plusieurs classes avec un certain degré d'appartenance. C'est ici une contribution non négligeable par rapport à l'algorithme original qui conduit souvent à un manque de précision lorsque des classes se chevauchent.

A.4.2.2 Apprentissage par renforcement

Le principe de l'apprentissage par renforcement est de trouver le meilleur choix possible pour une action avec un processus d'essais et d'erreurs. Ainsi, à chaque

action le système d'apprentissage par renforcement effectue un certain nombre d'essais dont il évalue la pertinence par le biais d'une fonction de récompense. Ce type d'apprentissage est non supervisé car la récompense donne juste un indice de qualité et non pas le résultat optimal. Parmi les premiers algorithmes, nous pouvons citer le *Q-learning* [Watkins & Dayan 1992] et le *TD-learning* [Sutton 1988].

A.5 Évaluation des algorithmes

Pour pouvoir évaluer les algorithmes, il est nécessaire d'utiliser des **indicateurs** communs pour permettre une comparaison et une évaluation de qualité. Plusieurs indicateurs sont proposés dans cette section. Ils mesurent chacun un point précis des algorithmes.

A.5.1 *Précision* et *Rappel*

La *Précision* et le *Rappel* sont les indicateurs les plus couramment utilisés en Recherche d'Information (RI) et en classification. Les éléments à évaluer peuvent être de différentes natures, comme des documents, des mots, etc. Dans un processus de classification, l'objectif est de valider si les éléments ont été correctement attribués à une classe. Nous identifions quatre groupes d'éléments :

- Les Vrais Positifs (VP), qui correspondent aux éléments correctement identifiés par le système comme appartenant à la classe.
- Les Faux Positifs (FP), qui correspondent aux éléments qui ont été identifiés par le système comme appartenant à la classe, mais qui, en réalité n'appartiennent pas à la classe (erreur).
- Les Vrais Négatifs (VN), qui correspondent aux éléments correctement identifiés par le système comme n'appartenant pas à la classe.
- Les Faux Négatifs (FN), qui correspondent aux éléments identifiés par le système comme n'appartenant pas à la classe, mais qui, en réalité appartiennent à la classe (erreur)..

Le *Rappel* mesure le pourcentage d'éléments correctement identifiés par rapport au nombre total d'éléments à identifier, il est défini par :

$$Rappel = \frac{VP}{VP + FN} \tag{A.12}$$

La *Précision* mesure le pourcentage d'éléments correctement identifiés par rapport à ceux identifiés par le système. Elle est définie par :

$$Précision = \frac{VP}{VP + FP} \tag{A.13}$$

A.5.2 Justesse

Il peut être utile de connaitre la "Justesse" du système, c'est à dire son erreur globale. La justesse est définie par :

$$Justesse = \frac{VP + VN}{VP + FP + VN + FN} \tag{A.14}$$

A.5.3 $F\beta Score$

Le $F\beta Score$ est un indicateur qui représente la moyenne harmonique de la *précision* et du *rappel*. Elle permet ainsi d'avoir une mesure de synthèse sur ces deux indicateurs. Plus formellement, le $F\beta Score$ s'écrit comme :

$$F\beta Score = \frac{(1 + \beta^2).(Précision.Rappel)}{\beta^2.Précision + Rappel} \tag{A.15}$$

La mesure la plus couramment utilisée est le $F1Score$ ($\beta = 1$), puisqu'elle donne le même poids à la précision et au rappel. La mesure $F2$ est aussi utilisée, et elle donne plus de poids au *Rappel* qu'à la *Précision*, à l'inverse, la mesure $F0.5$ donne plus de poids à la *Précision* qu'au *Rappel*. Cette mesure est paramétrable selon le contexte d'utilisation.

A.5.4 *WindowDiff*

L'évaluation de l'efficacité d'algorithmes de segmentation thématique est généralement effectuée en quantifiant le degré d'accord entre une segmentation hypothétique (hyp) et une segmentation de référence (ref).Les indices classiques en extraction d'information que sont le rappel et la précision ont été critiqués parce qu'ils ne font aucune différence entre des erreurs légères, comme le fait de placer une frontière juste à côté de la position attendue, par comparaison aux erreurs plus graves, comme placer une frontière à une grande distance de cette position attendue, manquer une frontière (faux négatif) ou en ajouter une (faux positif). C'est pourquoi, *WindowDiff* [Pevzner & Hearst 2002] s'est imposé comme l'indice de référence.

$$WD(ref, hyp) = \frac{1}{N - k} \sum_{1}^{N-k} (\mid b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k}) \mid > 0) \tag{A.16}$$

où $b(i, j)$ représente le nombre de frontières entre les positions i et j, N le nombre de positions, k correspond à la moitié de la longueur moyenne d'un segment dans l'annotation de référence (longueur de phrases). On peut décrire le fonctionnement de WD de la manière suivante. Une fenêtre de taille k est déplacée tout au long des unités minimales de segmentation d'un texte (habituellement les phrases). Pour chaque position de la fenêtre, on compare le nombre de frontières de segments que celle-ci englobe selon la norme de référence au nombre de frontières détectées par l'algorithme. Celui-ci est pénalisé d'un point chaque fois que ces nombres sont différents. Le dénominateur permet d'obtenir un score compris entre 0 et 1. S'agissant d'une mesure d'erreur, plus sa valeur est proche de 0, meilleure est la performance.

A.5.5 *Distance de Hamming généralisée*

La distance de Hamming généralisée [Bookstein *et al.* 2002] est une extension de la mesure de Hamming proposée par Richard Hamming, utilisée en informatique, en traitement du signal et dans les télécommunications. Elle joue un rôle important en théorie algébrique des codes correcteurs. Elle permet de quantifier la différence entre deux séquences de symboles.
[Bookstein *et al.* 2002] s'est intéressé à la mesure de la distance entre des vecteurs binaires de mêmes longueurs. Ce genre de données s'obtient en traitement du signal, mais aussi en segmentation thématique où la présence d'une frontière entre deux unités minimales peut-être codée par un "1" et l'absence de frontière par un "0". Pour ce type de données, une mesure possible est la distance de Hamming généralisée (ici appliquée à des données binaires) qui est basée sur le nombre de bits qu'il est nécessaire de modifier ou de déplacer pour transformer une séquence en une autre. Considérée comme une forme particulière de distance d'édition, elle correspond au **coût minimal** des opérations nécessaires pour effectuer cette transformation lorsque les seules opérations sont :

- l'opération d'insertion qui change un 0 en 1 pour un coût $Ci = 1$
- l'opération de suppression qui change un 1 en 0 pour un coût $Cs = 1$
- l'opération de déplacement qui fait glisser un "1" vers la gauche ou vers la droite de la séquence afin de le mettre en correspondance avec un "1" dans l'autre séquence. Le coût de cette opération (Cd) est une fonction strictement positive et monotone croissante de la longueur du déplacement nécessaire tel que $Cd = a. \mid i - j \mid$ où a est un paramètre de l'algorithme et i est la position de la frontière de référence et j, la position de la frontière de l'expérience.

Annexe B

Éléments mathématiques sur la fusion d'avis

B.1 Fusion d'opinion sous forme d'intervalles

Les principaux problèmes mathématiques considérés dans cette section concernent la représentation et la fusion de sources d'information sur un critère donné utilisées pour le choix d'une alternative dans un processus décisionnel. Dans le cas présent, les sources d'information proviennent d'opinions de textes. Dans ce processus les informations fournies sont considérées incertaines, dû à la multiplicité des points de vue et qu'elles sont imprécises, dû à la subjectivité de l'opinion. L'imprécision est représenter par un intervalle de valeurs qui permet de modéliser la subjectivité de l'évaluation fournie : pour un texte, selon le point de vue choisi, l'opinion peut être différente, l'intervalle symbolisant les valeurs possibles. Cette idée de représenter les opinions comme des intervalles a été introduite par [Kaufmann 1988] pour la modélisation de l'opinion d'experts dans des environnements incertains.

Exemple 7 *L'exemple proposé dans le tableau B.1 considère 26 critiques de films noté X (textes d'opinion) évalués sur le critère "scénario". Chacune des évaluations est un intervalle dans l'espace [0,20] qui représente le degré de satisfaction. Une opinion [12,15] signifie que le point de vue (opinion) exprimé dans le texte considère que le degré de satisfaction (la note), pour le critère scénario, peut prendre les valeurs {12,13,14,15}.*

Plusieurs raisons justifient la nécessité de la synthèse de ces opinions : une représentation synthétique des données, l'agrégation avec d'autres critères, l'étude de la dispersion entre les points de vue, etc. Si les opinions étaient précises, la théorie des probabilités proposeraient une solution pour fusionner les opinions. Dans le cas où les évaluations sont incertaines, la théorie des possibilités propose un moyen de

Index	Valeurs pour le critère scénario	Nombre de textes ayant émis l'évaluation
1	*[7,12]*	*4*
2	*[9,14]*	*3*
3	*[10,12]*	*2*
4	*[10,13]*	*4*
5	*[10,15]*	*3*
6	*[3,6]*	*2*
7	*[4,7]*	*4*
8	*[5,7]*	*4*
	Total	*26*

TABLE B.1 – Exemple d'évaluations sous forme d'intervalles

fusionner ces évaluations imprécises. En effet, un cadre possibiliste permet de faire face à la fois à l'imprécision et à l'incertitude des évaluations [Dubois & Prade 1988].

Cette section résume une méthode basée sur la théorie des possibilités qui permet de fusionner des intervalles d'opinion. Ces intervalles d'opinion sont synthétisés dans des distributions de possibilité qui fournissent un encadrement des valeurs de probabilités que l'on aurait obtenues si les informations avaient été précises [Dubois & Prade 1986].

Avant de présenter un bref rappel sur la théorie des possibilités, nous allons introduire les notations suivantes : n le nombre de critères considérés, r_q le nombre de textes contenant le critère $q, 1 \leq q \leq n$ et $\left[Sc_k^q\right]$ est l'intervalle qui modélise l'évaluation d'un texte $k, 1 \leq k \leq r_q$. La théorie des possibilités est utilisée pour prendre en compte les différents aspects de l'imperfection de l'information inhérente aux évaluations données par l'ensemble des textes par rapport à un critère donné. L'information synthétique qui résulte de la fusion de toutes les évaluations imprécises (intervalles) relatives au critère q est considérée comme valeur incertaine, c'est pourquoi elle est représenté par une distribution de possibilité notée π_q

B.1.1 La théorie des possibilités

La théorie des possibilités a été introduite en 1978 par Lofti Zadeh [Zadeh 1978] pour donner une sémantique d'incertitude à la notion de sous-ensembles flous qu'il avait proposée dans le années 1960 [Zadeh 1965] [Dubois & Prade 1988].

Soit un ensemble de référence Ω, A et B deux sous-ensembles de Ω, une mesure de

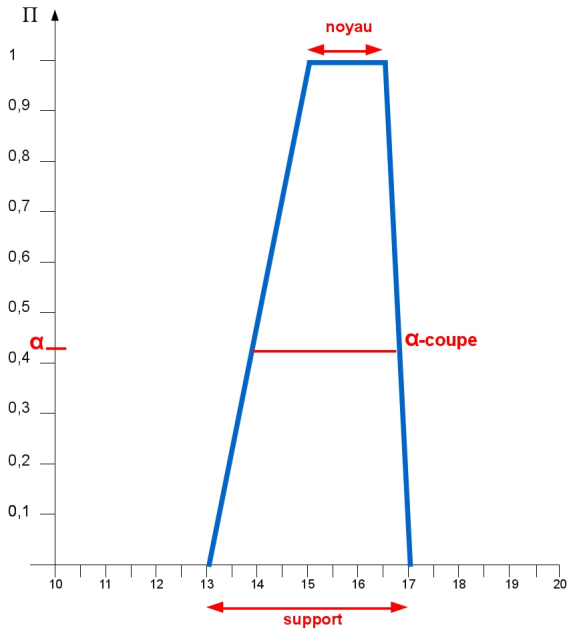

FIGURE B.1 – Distribution de possibilités

possibilité Π, est une fonction définie sur $\mathcal{P}(\Omega)$, qui prend ses valeurs dans l'intervalle $[0, 1]$, tel que : $\Pi(\emptyset) = 0$ and $\Pi(A \cup B) = max(\Pi(A), \Pi(B))$. $\Pi(A)$ quantifie à quel point un événement $A \subseteq \Omega$ est possible : $\Pi(A) = 1$ signifie que l'événement A est totalement possible, et $\Pi(A) = 0$ signifie que l'événement A est impossible.

À la mesure de possibilité, une mesure dual appelée *nécessité* notée N est associée telle que $\Pi(A) = 1 - N(\overline{A})$. Une mesure de *possibilité* Π peut être caractérisée par une distribution de possibilité π qui est une fonction $\pi : \Omega \longrightarrow [0, 1]$ tel que $\Pi(A) = sup_{\omega \in A}\pi(\omega)$, $\forall A \subseteq \Omega$. Cette fonction est normalisée si $\exists \omega \in \Omega, \pi(\omega) = 1$. N peut alors être définie à partir de π tel que : $N(A) = inf_{\omega \notin A}(1 - \pi(\omega))$, $\forall A \subseteq \Omega$. La distribution de possibilité π évalue le degré auquel une valeur peut être possible et est caractérisée par deux ensembles : son noyau et son support (c.f. figure B.1).

Les valeurs appartenant au noyau de la distribution sont totalement possibles, et celles appartenant au support de la distribution sont possibles avec un certain degré. Les valeurs qui se trouvent à l'extérieur du support de la distribution sont impossibles. Les valeurs possibles au degré α appartiennent à un ensemble appelé α-coupe.

Un cadre général pour la théorie des probabilités et de la possibilité est la théorie de l'évidence (ou théorie des croyances) [Dempster 1967, Shafer 1976]. Ce cadre général permet d'introduire des fonctions de croyances et de plausibilités qui permettent

de déterminer les bornes supérieur et inférieur d'une distribution de probabilités.

Définition 1 *Une masse de croyance m est une fonction de $\mathcal{P}(\Omega)$ dans $[0,1]$ tel que : $\sum_{\{A\in\mathcal{P}(\Omega)\}} m(A)$ et $m(\emptyset)=0$.*

Un élément focal $E_j, 1 \leq j \leq p$ est une partie de $\mathcal{P}(\Omega)$ tel que $m(E_j) > 0$. L'ensemble des éléments focaux est noté $\mathbb{F}$. Lorsque les éléments E_j sont des observations imprécises, la probabilité de tous les événements $A \subseteq \Omega$ notés $Pr(A)$ est imprécise et appartient à $[Bel(A), Pl(A)]$ avec les définitions suivantes :

Définition 2 *La fonction de croyance notée Bel d'un ensemble A de Ω est définie par : $Bel(A) = \sum_{\{E\in\mathbb{F}/A\supseteq E\}} m(E)$ où $\{E \in \mathbb{F}/A \supseteq E\}$ est l'ensemble des témoignages en faveur de A.*

Définition 3 *La fonction de plausibilité notée Pl d'un ensemble A de Ω est définie par : $Pl(A) = \sum_{\{E\in\mathbb{F}/A\cap E\neq\emptyset\}} m(E)$ où $\{E \in \mathbb{F}/A \cap E \neq \emptyset\}$ est l'ensemble des éléments focaux qui ont une relation avec A et le rendent possible.*

B.1.2 Fusion des intervalles

L'intervalle $\left[Sc_k^q\right]$, $1 \leq k \leq r_q$ représente un élément focal relatif au critère C^q, $\{[Sc_1^q], \left[Sc_k^2\right], \ldots, \left[Sc_{r_q}^q\right]\}$ l'ensemble des opinions distinctes ($\forall k \neq k', \left[Sc_k^q\right] \neq \left[Sc_{k'}^q\right]$) et $p(\left[Sc_k^q\right])$ la probabilité d'occurrence des intervalles $\left[Sc_k^q\right]$ parmi $\{\left[Sc_k^q\right], 1 \leq k \leq r_q\}$. Dans [Imoussaten 2011] on identifie quatre type d'intervalles : des singletons, emboîtés , cohérents, incohérents.

B.1.2.1 Les intervalles sont des singletons

Dans ce cas la construction d'un histogramme est possible : $Bel(A) = Pl(A) = Pr(A)$, $\forall A \subseteq \Omega$.

B.1.2.2 Les intervalles sont emboîtés

Lorsque les éléments sont emboîtés,$[Sc_1^q] \subseteq [Sc_2^q] \subseteq \ldots \subseteq [Sc_{r^q}^q]$, la fonction de croyance Bel est une mesure de nécessité et la fonction de plausibilité Pl est une mesure de possibilité [Shafer 1976]. Une distribution de possibilité π approchant les données peut être construite de la manière suivante [Dubois & Prade 1986] :

$$\forall \omega \in \Omega, \pi(\omega) = \begin{cases} 0 & \text{if} \quad \omega \notin [Sc^q_{r^q}] \\ \sum_{\{l \leq k \leq r^q\}} p([Sc^q_k]) & \text{if} \quad \omega \in [Sc^q_l] \setminus [Sc^q_{l-1}] \\ 1 & \text{if} \quad \omega \in [Sc^q_1] \end{cases} \tag{B.1}$$

B.1.2.3 Les intervalles sont cohérents

Les éléments focaux sont cohérents i.e., $\bigcap_{1 \leq k \leq r^q} [Sc^q_k] = [Sc^q] \neq \emptyset$, mais non emboîtés, il est alors possible de construire une distribution de possibilités à partir des intervalles $[Sc^q_k]$, $1 \leq k \leq r_q$ [Dubois & Prade 1986]. Il s'agit de construire de nouveaux intervalles emboîtés E^k_q, $1 \leq k \leq s'$ et leurs masses associées $m(E^k_q)$ pour se rapporter au cas précédent, de telle sorte que : $Sc \subseteq E^1_q \subseteq E^2_q \subseteq ... \subseteq E^s_q = \bigcup_{1 \leq j \leq s} [Sc^q_k]$ et $m(E^k_q) = \sum_{\{[Sc^q_{k'}] \ associ \ \acute{a} \ E^k_q\}} p([Sc^q_{k'}])$. Pour plus d'explications sur la construction des éléments E^k_q, $1 \leq k \leq s'$ voir [Dubois & Prade 1986]. Dans ce cas où les données sont cohérentes, deux distributions sont construites pour donner un encadrement de la probabilité de tous les sous-ensembles de Ω : (1) une approximation possibiliste inférieure des intervalles $[Sc^q_k]$, $1 \leq k \leq r_q$ est définie par :

$$\forall \omega \in \Omega, \pi_1(\omega) = \sum_{k=1,s} p([Sc^q_k]) \chi_{[Sc^q_k]}(\omega) \tag{B.2}$$

où $\chi_{[Sc^q_k]}$ est la fonction caractéristique classique de $[Sc^q_k]$.

(2) une approximation possibiliste supérieure des intervalles E^k_q, $1 \leq k \leq s'$ (B.1) tel que :

$$\forall \omega \in \Omega, \pi_2(\omega) = \begin{cases} 0 & \text{if} \quad \omega \notin E^{s'}_q \\ \sum_{\{l \leq k \leq s'\}} p(E^k_q) & \text{if} \quad \omega \in E^l_q \setminus E^{l-1}_q \\ 1 & \text{if} \quad \omega \in E^1_q \end{cases} \tag{B.3}$$

B.1.2.4 (d) Les intervalles sont incohérents

Les données sont généralement ni précises, ni cohérentes. Les probabilités d'une part et la possibilité-nécessité, d'autre part, correspondent à des situations extrêmes et idéales [Dubois & Prade 1986]. Malheureusement, les opinions exprimées peuvent correspondre à des évaluations contradictoires. Cela signifie que la contrainte de cohérence n'est pas toujours vérifiée dans la pratique, i.e., $\bigcap_{\{1 \leq k \leq s\}} [Sc^q_k] = \emptyset$. Pour faire face à cette situation, [Imoussaten 2011] propose une méthode qui consiste à considérer la cohérence des groupes d'intervalles avec une intersection non vide :

$K_\beta \subseteq \{1, ..., s\}, 1 \leq \beta \leq v$ tel que $\bigcap_{\{k \in K_\beta\}} \left[Sc_k^q\right] \neq \emptyset$ (où v est le nombre de groupes K_β). À chaque groupe sont associées deux distributions π_1^β et π_2^β (equation B.2 et equation B.3) Alors, v distributions mono-modales sont construites pour synthétiser les données $\left[Sc_k^q\right]$, $1 \leq j \leq r_i$ en utilisant une approximation possibiliste inférieure $\pi_1 = \vee_{\{1 \leq \beta \leq v\}} \pi_1^\beta$ et une approximation possibiliste supérieure $\pi_2 = \vee_{\{1 \leq \beta \leq v\}} \pi_2^\beta$.

B.1.3 Illustration de l'exemple 7

Comme nous pouvons facilement le remarquer, les données utilisée dans l'exemple 7 sont incohérentes ($[5, 7] \cap [10, 12] = \emptyset$).

Les données représentent, en réalité, deux groupes d'intervalles cohérents : $K_{\beta_1} = \{1, 2, 3, 4, 5\}$ et $K_{\beta_2} = \{6, 7, 8\}$. La figure B.2 illustre la construction des intervalles E_q^k, et des approximations inférieure et supérieure pour les deux groupes d'intervalles.

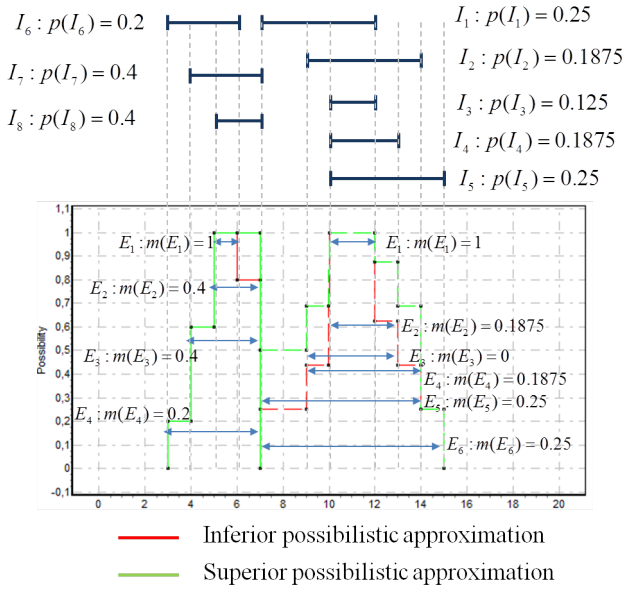

FIGURE B.2 – Approximations inférieure et supérieure

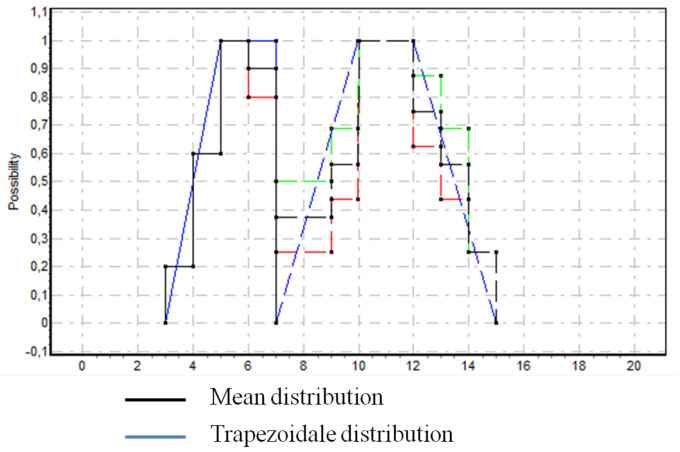

FIGURE B.3 – Approximation des distributions

Bibliographie

[Adomavicius & Kwon 2007] Gediminas Adomavicius et YoungOk Kwon. *New Recommendation Techniques for Multicriteria Rating Systems*. IEEE Intelligent Systems, pages 48–55, 2007. (Cité en pages 10, 11 et 22.)

[Adomavicius *et al.* 2011] Gediminas Adomavicius, Nikos Manouselis et YoungOk Kwon. Multi-criteria recommender systems. Springer, 2011. (Cité en pages 10, 11 et 12.)

[Akharraz 2004] A. Akharraz. Acceptabilité de la décision et risque décisionnel : un système explicatif de fusion d'informations par l'intégrale de Choquet. 2004. (Cité en page 23.)

[Amini *et al.* 2000] Massih-reza Amini, Hugo Zaragoza et Patrick Gallinari. *Learning for Sequence Extraction Tasks*, 2000. (Cité en page 34.)

[Andreevskaia & Bergler 2006] Alina Andreevskaia et Sabine Bergler. *Mining WordNet for Fuzzy Sentiment : Sentiment Tag Extraction from WordNet Glosses*. In Proceedings EACL-06, Trento, Italy., 2006. (Cité en page 38.)

[Argamon *et al.* 1998] Shlomo Argamon, Moshe Koppel et Galit Avneri. *Routing Documents According to Style*. In Proceedings of First International Workshop on Innovative Information Systems, 1998. (Cité en page 36.)

[Bai 2011] Xue Bai. *Predicting consumer sentiments from online text*. Decis. Support Syst., vol. 50, pages 732–742, March 2011. (Cité en pages 35 et 37.)

[Baizet 2004] Y. Baizet. La gestion des connaissances en conception : application à la simulation numérique chez renault-diec. 2004. (Cité en page 4.)

[Bana E Costa & Chagas 2004] Carlos Bana E Costa et Manuel P. Chagas. *A career choice problem : An example of how to use MACBETH to build a quantitative value model based on qualitative value judgments*. European Journal of Operational Research, pages 323–331, 2004. (Cité en page 137.)

[Bana E Costa & Vansnick 1997] Carlos Bana E Costa et J. Vansnick. *Applications of the Macbeth approach in the framework of an additive aggregation model*. Journal of Multi-Criteria Decision Analysis, vol. 6, pages 107–114, 1997. (Cité en page 137.)

[Bayes 1763] T. Bayes. *An essay towards solving a problem in the doctrine of chances*. Phil. Trans. of the Royal Soc. of London, vol. 53, pages 370–418, 1763. (Cité en page 159.)

[Beineke *et al.* 2004] Philip Beineke, Trevor Hastie, Christopher Manning et Shivakumar Vaithyanathan. *Exploring Sentiment Summarization*. In Yan Qu, James Shanahan et Janyce Wiebe, editeurs, Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text Theories and Applications, volume 07, pages 1–4. AAAI Press, 2004. (Cité en page 37.)

[Bezdek 1981] James C. Bezdek. Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers, Norwell, MA, USA, 1981. (Cité en page 162.)

[Bigi *et al.* 2000] Brigitte Bigi, Renato de Mori, Marc El-Bèze et Thierry Spriet. *A fuzzy decision strategy for topic identification and dynamic selection of language models*. Signal Processing, vol. 80, no. 6, pages 1085–1097, Juin 2000. (Cité en page 34.)

[Bookstein *et al.* 2002] Abraham Bookstein, Vladimir A. Kulyukin et Timo Raita. *Generalized Hamming Distance*. Inf. Retr., vol. 5, no. 4, pages 353–375, Octobre 2002. (Cité en page 165.)

[Boullier & Lohard 2012] Dominique Boullier et Audrey Lohard. Opinion mining et sentiment analysis : méthodes et outils. OpenEdition Press, 2012. (Cité en page 35.)

[Brooking 1998] Annie Brooking. Corporate memory : Strategies for knowledge management. International Thomson Publishing, 1998. (Cité en page 4.)

[Caillet *et al.* 2004] Marc Caillet, Jean francois Pessiot, Massih reza Amini et Patrick Gallinari. *Unsupervised Learning with Term Clustering for Thematic Segmentation of Texts*. In Proceedings of RIAO, pages 648–656, 2004. (Cité en page 34.)

[Cambria *et al.* 2010] Erik Cambria, Robert Speer, Catherine Havasi et Amir Hussain. *SenticNet : A Publicly Available Semantic Resource for Opinion Mining*. Artificial Intelligence, pages 14–18, 2010. (Cité en page 100.)

[Cardie *et al.* 2003] Claire Cardie, Janyce Wiebe, Theresa Wilson et Diane Litman. *Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering*. In Working Notes - New Directions in

Question Answering (AAAI Spring Symposium Series, pages 20–27, 2003. (Cité en page 37.)

[Chaovalit & Zhou 2005] Pimwadee Chaovalit et Lina Zhou. *Movie Review Mining : a Comparison between Supervised and Unsupervised Classification Approaches*. Hawaii International Conference on System Sciences, vol. 4, page 112c, 2005. (Cité en page 36.)

[Cheung *et al.* 2003] Kwok-Wai Cheung, James T. Kwok, Martin H. Law et Kwok-Ching Tsui. *Mining customer product ratings for personalized marketing*. Decis. Support Syst., vol. 35, no. 2, pages 231–243, Mai 2003. (Cité en page 10.)

[Choi 2000] Freddy Y. Y. Choi. *Advances in domain independent linear text segmentation*. proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, vol. 23, pages 26–33, 2000. (Cité en pages 30, 31 et 32.)

[Chuang & Yang 2000] Wesley T. Chuang et Jihoon Yang. *Extracting Sentence Segments for Text Summarization : A Machine Learning Approach*. Proceedings of the 23 th ACM SIGIR, pages 152–159, 2000. (Cité en page 35.)

[Clarke & Terra 2003] Charles L. A. Clarke et Egidio L. Terra. *Passage retrieval vs. document retrieval for factoid question answering*. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, SIGIR '03, pages 427–428, New York, NY, USA, 2003. ACM. (Cité en page 37.)

[Cover & Hart 1967] T. Cover et P. Hart. *Nearest neighbor pattern classification*. vol. 13, no. 1, pages 21–27, Janvier 1967. (Cité en page 161.)

[Cybenko 1989] G. Cybenko. *Approximation by superpositions of a sigmoidal function*. Mathematics of Control, Signals, and Systems (MCSS), vol. 2, no. 4, pages 303–314, Décembre 1989. (Cité en page 160.)

[Dempster 1967] A.P Dempster. *Upper and Lower probabilities induced by multivalued mapping*. Annals of Mathematical Statistics, vol. 38, 1967. (Cité en page 169.)

[Denguir-Rekik *et al.* 2006] Afef Denguir-Rekik, Gilles Mauris et Jacky Montmain. *Propagation of uncertainty by the possibility theory in Choquet integral-based decision making : application to an E-commerce website choice support*. IEEE

T. Instrumentation and Measurement, vol. 55, no. 3, pages 721–728, 2006. (Cité en page 4.)

[Denguir-Rekik *et al.* 2009] Afef Denguir-Rekik, Jacky Montmain et Gilles Mauris. *A possibilistic-valued multi-criteria decision-making support for marketing activities in e-commerce : Feedback Based Diagnosis System.* European Journal of Operational Research, vol. 195, no. 3, pages 876–888, 2009. (Cité en page 4.)

[D'hondt *et al.* 2011] Joris D'hondt, Paul-Armand Verhaegen, Joris Vertommen, Dirk Cattrysse et Joost R. Duflou. *Topic identification based on document coherence and spectral analysis.* Information Sciences, 2011. (Cité en pages 30 et 33.)

[Dimitrova *et al.* 2002] Maya Dimitrova, Aidan Finn, Nicholas Kushmerick et Barry Smyth. *Web Genre Visualization.* In Proceedings conference on human factors in, 2002. (Cité en page 36.)

[Dubois & Prade 1986] Didier Dubois et Henri Prade. *Fuzzy sets and statistical data.* European Journal of Operational Research, vol. 25, 1986. (Cité en pages 168, 170 et 171.)

[Dubois & Prade 1988] Didier Dubois et Henri Prade. Possibility theory : An approach to computerized processing of uncertainty. Plenum Press, New York, 1988. (Cité en pages 125 et 168.)

[Dumais 1991] Susan T. Dumais. *Improving the retrieval of information from external sources.* Behavior Research Methods, Instruments, & Computers, vol. 23, no. 2, pages 229–236, 1991. (Cité en page 157.)

[Durbin *et al.* 2003] Stephen D. Durbin, J. Neal Richter et Doug Warner. *A system for affective rating of texts.* In Proceedings of OTC-03, 3rd workshop on operational text classification, Washington, USA, 2003. (Cité en page 36.)

[Duthil *et al.* 2011a] Benjamin Duthil, François Trousset, Mathieu Roche, Gérard Dray, Michel Plantié, Jacky Montmain et Pascal Poncelet. *Towards an Automatic Characterization of Criteria.* In DEXA (1), pages 457–465, 2011. (Cité en page 82.)

[Duthil *et al.* 2011b] Benjamin Duthil, François Trousset, Gérard Dray, Jacky Montmain et Pascal Poncelet. *Vers une caractérisation automatique de critères pour l'opinion-mining.* In Les Cahiers du numérique : special issue

Sciences et technologies de l'information et de la communication, volume 7, pages 41–62, 2011. (Cité en pages 39 et 45.)

[Duthil *et al.* 2012a] Benjamin Duthil, François Trousset, Gérard Dray, Jacky Montmain et Pascal Poncelet. *Opinion Extraction Applied to Criteria.* In Stephen W. Liddle, Klaus-Dieter Schewe, A Min Tjoa et Xiaofang Zhou, editeurs, DEXA (2), volume 7447 of *Lecture Notes in Computer Science*, pages 489–496. Springer, 2012. (Cité en pages 22 et 38.)

[Duthil *et al.* 2012b] Benjamin Duthil, François Trousset, Gérard Dray, Pascal Poncelet et Jacky Montmain. *Extraction d'opinions appliquée à des critères.* In Yves Lechevallier, Guy Melançon et Bruno Pinaud, editeurs, EGC, volume RNTI-E-23 of *Revue des Nouvelles Technologies de l'Information*, pages 483–488. Hermann-Éditions, 2012. (Cité en page 38.)

[Ermine *et al.* 1996] J.-L. Ermine, M. Chaillot, P. Bigeon, B. Charreton et D. Malavielle. *MKSM, méthode pour la gestion des connaissances.* Ingénierie des systèmes d'information (1993), vol. 4, no. 4, pages 541–575, 1996. fre. (Cité en page 4.)

[Ermine 1989] J.L. Ermine. Systèmes experts : Théorie et pratique. Technique et Documentation-Lavoisier, 1989. (Cité en page 84.)

[Esuli & Sebastiani 2005] Andrea Esuli et Fabrizio Sebastiani. *Determining the semantic orientation of terms through gloss classification.* In Proceedings of the 14th ACM international conference on Information and knowledge management, CIKM '05, pages 617–624, New York, NY, USA, 2005. ACM. (Cité en page 38.)

[Fishburn 1970] Peter C Fishburn. Utility theory for decision making, volume 6. Wiley, 1970. (Cité en page 131.)

[Fishburn 1982] P.C. Fishburn. The foundations of expected utility. D. Reidel Publishing, Dordrecht, 1982. (Cité en page 131.)

[Galley *et al.* 2003] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier et Hongyan Jing. *Discourse segmentation of multi-party conversation.* In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03, pages 562–569, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. (Cité en page 30.)

[Ghani & Fano 2002] Rayid Ghani et Andrew Fano. *Building Recommender Systems using a Knowledge Base of Product Semantics.* In 2nd International

Conference on Adaptive Hypermedia and Adaptive Web Based Systems, Malaga, 2002. (Cité en page 10.)

[Gopal *et al.* 2011] Ram Gopal, James R. Marsden et Jan Vanthienen. *Information mining ? Reflections on recent advancements and the road ahead in data, text, and media mining.* Decision Support Systems, vol. 51, no. 4, pages 727–731, Novembre 2011. (Cité en page 37.)

[Harb *et al.* 2008] Ali Harb, Michel Plantié, Gérard Dray, Mathieu Roche, François Trousset et Pascal Poncelet. *Web opinion mining : how to extract opinions from blogs ?* International Conference on Soft Computing as Transdisciplinary Science and Technology, 2008. (Cité en pages 38, 82, 86 et 150.)

[Hatzivassiloglou & McKeown 1997] Vasileios Hatzivassiloglou et Kathleen R. McKeown. *Predicting the semantic orientation of adjectives.* In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, ACL '98, pages 174–181, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics. (Cité en page 37.)

[Hatzivassiloglou & Wiebe 2000] Vasileios Hatzivassiloglou et Janyce M. Wiebe. *Effects of Adjective Orientation and Gradability on Sentence Subjectivity.* In Proceedings of the 18th international conference on computational linguistics., pages 299–305, 2000. (Cité en page 38.)

[He & Zhou 2010] Yulan He et Deyu Zhou. *Self-training from labeled features for sentiment analysis.* Information Processing and Management, vol. In Press, Corrected Proof, pages –, 2010. (Cité en page 35.)

[Hearst 1997] Marti A. Hearst. *TextTiling : segmenting text into multi-paragraph subtopic passages.* ACM Computational Linguistics, vol. 23(1), pages 33–64, 1997. (Cité en pages 30, 31, 32 et 49.)

[Hebb 1961] D.O. Hebb. The organization of behavior : a neuropsychological theory. Science editions. Science Editions, 1961. (Cité en page 160.)

[Hong & Hatzivassiloglou 2003] Yu Hong et Vasileios Hatzivassiloglou. *Towards answering opinion questions : Separating facts from opinions and identifying the polarity of opinion sentences.* In Proceedings of EMNLP-03, pages 129–136, 2003. (Cité en page 38.)

[Hu & Liu 2004] Minqing Hu et Bing Liu. *Mining and summarizing customer reviews.* In Proceedings of the tenth ACM SIGKDD international conference

on Knowledge discovery and data mining, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM. (Cité en page 37.)

[Imoussaten 2011] Abdelhak Imoussaten. *Modélisation et pilotage de la phase de délibération dans une décision collective - vers le management d'activités à risques*. PhD thesis, MINES ParisTech, 2011. (Cité en pages 23, 128, 134, 170 et 171.)

[Jain *et al.* 1999] A. K. Jain, M. N. Murty et P. J. Flynn. *Data clustering : a review*, Septembre 1999. (Cité en page 34.)

[Jin *et al.* 2009] Wei Jin, Hung Hay Ho et Rohini K. Srihari. *OpinionMiner : A Novel Machine Learning System for Web Opinion Mining and Extraction*. IEEE Symposium on Visual Analytics Science and Technology, 2009. (Cité en page 37.)

[Kamps *et al.* 2004] Jaap Kamps, Maarten Marx, Robert J. Mokken et Maarten de Rijke. *Using WordNet to Measure Semantic Orientations of Adjectives*. In Proceedings of LREC-04, 4th international conference on language resources and evaluation, Lisbon, PT, volume 4, pages 1115–1118, 2004. (Cité en page 38.)

[Kaufmann 1988] A. Kaufmann. *Theory of expertons and fuzzy logic*. Fuzzy Sets and Systems, vol. 28, no. 3, 1988. (Cité en page 167.)

[Kessler *et al.* 1997] Brett Kessler, Geoffrey Numberg et Hinrich Schütze. *Automatic detection of text genre*. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, ACL '98, pages 32–38, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics. (Cité en page 36.)

[Kim & Hovy 2004] Soo-Min Kim et Eduard Hovy. *Determining the sentiment of opinions*. In Proceedings of the 20th international conference on Computational Linguistics, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. (Cité en page 38.)

[Kleiber 1996] Georges Kleiber. *Noms propres et noms communs : un problème de dénomination*. Meta, pages 567–589, 1996. (Cité en page 53.)

[Krantz *et al.* 1971] D.H. Krantz, R.D. Luce, P. Suppes et A. Tversky. Foundations of measurement. Academic Press, New York, 1971. (Cité en page 131.)

[Ku *et al.* 2005] Lun-Wei Ku, Li-Ying Lee, Tung-Ho Wu et Hsin-Hsi Chen. *Major topic detection and its application to opinion summarization.* In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '05, pages 627–628, New York, NY, USA, 2005. ACM. (Cité en page 37.)

[Kupiec *et al.* 1995] Julian Kupiec, Jan Pedersen et Francine Chen. *A trainable document summarizer.* Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pages 68–73, 1995. (Cité en page 35.)

[Lamprier *et al.* 2007] Sylvain Lamprier, Tassadit Amghar, Bernard Levrat et Frédéric Saubion. *SegGen : A Genetic Algorithm for Linear Text Segmentation.* In IJCAI, pages 1647–1652, 2007. (Cité en page 32.)

[Lin 1998] Dekang Lin. *Automatic retrieval and clustering of similar words.* In Proceedings of the 17th international conference on Computational linguistics - Volume 2, COLING '98, pages 768–774, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics. (Cité en pages 37 et 38.)

[Liu *et al.* 2005] Bing Liu, Minqing Hu et Junsheng Cheng. *Opinion observer : analyzing and comparing opinions on the Web.* In Proceedings of the 14th international conference on World Wide Web, WWW '05, pages 342–351, New York, NY, USA, 2005. ACM. (Cité en page 36.)

[Loh *et al.* 2004] Stanley Loh, Fabiana Lorenzi, Ramiro Saldaña et Daniel Licthnow. *A tourism recommender system based on collaboration and text analysis.* Information Technology and Tourism, vol. 6, 2004. (Cité en page 10.)

[MacQueen 1967] J. B. MacQueen. *Some Methods for Classification and Analysis of MultiVariate Observations.* In L. M. Le Cam et J. Neyman, editeurs, Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297. University of California Press, 1967. (Cité en page 162.)

[Malioutov & Barzilay 2006] Igor Malioutov et Regina Barzilay. *Minimum cut model for spoken lecture segmentation.* In Proceedings of the Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006, pages 25–32, 2006. (Cité en page 33.)

[Manouselis & Costopoulou 2007] Nikos Manouselis et Constantina Costopoulou. *Analysis and Classification of Multi-Criteria Recommender Systems.* World

Wide Web, vol. 10, no. 4, pages 415–441, Décembre 2007. (Cité en pages 10 et 11.)

[Mcculloch & Pitts 1943] Warren Mcculloch et Walter Pitts. *A Logical Calculus of Ideas Immanent in Nervous Activity*. Bulletin of Mathematical Biophysics, vol. 5, pages 127–147, 1943. (Cité en page 160.)

[McNee *et al.* 2003] S. M. McNee, S. K. Lam, C. Guetzlaff, J. A. Konstan et J. Riedl. *Confidence Displays and Training in Recommender Systems*. In INTERACT '03 IFIP TC13 International Conference on Human-Computer Interaction, pages 176–183, 2003. (Cité en page 3.)

[Mélès 1971] J. Mélès. La gestion par les systèmes. Editions Hommes et Technique, 1971. (Cité en page 65.)

[Modave & Grabisch] F. Modave et M. Grabisch. *Preference Representation by the Choquet Integral : The Commensurability Hypothesis*. (Cité en page 131.)

[Moens & De Busser 2001] Marie-Francine Moens et Rik De Busser. *Generic topic segmentation of document texts*. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '01, pages 418–419, New York, NY, USA, 2001. ACM. (Cité en page 34.)

[Montmain *et al.* 2005] J. Montmain, A. Denguir-Rekik et G. Mauris. *How deriving benefits from expert advices to make the right choice in multi-criteria decisions based on the Choquet integral ?* In European workshop on the Use of Expert Judgement in Decision-Making, 2005. (Cité en page 3.)

[Morinaga *et al.* 2002] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi et Toshikazu Fukushima. *Mining Product Reputations on the Web*. In ACM SIGKDD 2002, pages 341–349. ACM Press, 2002. (Cité en page 36.)

[Morris & Hirst 1991] Jane Morris et Graeme Hirst. *Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text*. Computational Linguistics, pages 21–48, 1991. (Cité en page 30.)

[Morris 1938] C.W. Morris. Foundations of the theory of signs. International encyclopedia of unified science. University of Chicago Press, 1938. (Cité en pages 28, 29, 43, 44, 53, 78 et 148.)

[Mullen & Malouf 2006] Tony Mullen et Robert Malouf. *A preliminary investigation into sentiment analysis of informal political discourse*. In AAAI Symposium

on Computational Approaches to Analysing Weblogs (AAAI-CAAW), pages 159–162, 2006. (Cité en page 37.)

[Oelke *et al.* 2009] Daniela Oelke, Ming Hao, Christian Rohrdantz, Daniel A. Keim, Umeshwar Dayal, Lars-Erik Haug et Halldór Janetzko. *Visual Opinion Analysis of Customer Feedback Data*. KDD'09, 2009. (Cité en pages 38 et 39.)

[Pang & Lee 2002] Bo Pang et Lillian Lee. *Thumbs up ? Sentiment Classification using Machine Learning Techniques*. In Proceedings of EMNLP, pages 79–86, 2002. (Cité en pages 35 et 151.)

[Pang *et al.* 2002] Bo Pang, Lillian Lee et Shivakumar Vaithyanathan. *Thumbs up ? Sentiment Classification using Machine Learning Techniques*. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002. (Cité en pages 100, 101 et 102.)

[Pelleg & Moore 2000] Dan Pelleg et Andrew W. Moore. *X-means : Extending K-means with Efficient Estimation of the Number of Clusters*. In Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00, pages 727–734, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. (Cité en page 162.)

[Penalva & Montmain 2002] J.-M. Penalva et J. Montmain. *Travail collectif et intelligence collective : les référentiels de connaissances*. In Proceedings of IPMU'2002, 2002. (Cité en page 4.)

[Pereira *et al.* 1993] Fernando Pereira, Naftali Tishby et Lillian Lee. *Distributional clustering of English words*. In Proceedings of the 31st annual meeting on Association for Computational Linguistics, ACL '93, pages 183–190, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. (Cité en pages 37 et 38.)

[Pevzner & Hearst 2002] Lev Pevzner et Marti A. Hearst. *A critique and improvement of an evaluation metric for text segmentation*. Comput. Linguist., vol. 28, no. 1, pages 19–36, Mars 2002. (Cité en page 164.)

[Piaget & Inhelder 1967] J. Piaget et B. Inhelder. La psychologie de l'enfant. Que sais-je ? Presses universitaires de France, 1967. (Cité en pages 84 et 104.)

[Plantie *et al.* 2008] Michel Plantie, Mathieu Roche, Gérard Dray et Pascal Poncelet. *Is a voting approach accurate for opinion mining ?* Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery (DaWaK'08), 2008. (Cité en pages 22, 23, 24, 25 et 122.)

[Ponte & Croft 1998] Jay M. Ponte et W. Bruce Croft. *A Language Modeling Approach to Information Retrieval.* pages 275–281, 1998. (Cité en page 34.)

[Ranwez *et al.* 2012] Sylvie Ranwez, Benjamin Duthil, Mohameth-François Sy, Jacky Montmain et Vincent Ranwez. How ontology based information retrieval systems may benefit from lexical text analysis. Springer, 2012. (Cité en page 110.)

[Rekik 2007] A.D. Rekik. Un cadre possibiliste pour l ?aide à la décision multicritère et multi-acteurs : Application au marketing et benchmarking de sites e-commerce. phd. Université de Savoie, 2007. (Cité en pages 22, 23 et 134.)

[Reynar 2000] Jeffrey C. Reynar. *Topic segmentation : Algorithms and applications.* PhD thesis, 2000. (Cité en page 31.)

[Rosenblatt 1988] F. Rosenblatt. *Neurocomputing : foundations of research.* chapitre The perception : a probabilistic model for information storage and organization in the brain, pages 89–114. MIT Press, Cambridge, MA, USA, 1988. (Cité en page 160.)

[Salton & Yang 1973] G. Salton et C. S. Yang. *On the specification of term values in automatic indexing.* Journal of Documentation., vol. 29, no. 4, pages 351–372, 1973. (Cité en pages 156 et 158.)

[Schärlig 1985] A. Schärlig. Décider sur plusieurs critères : Panorama de l'aide à la décision multicritère. Diriger L'Entreprise. Presses Polytechniques Romandes, 1985. (Cité en page 91.)

[Shafer 1976] G Shafer. A mathematical theory of evidence. Princeton University Press, Princeton, 1976. (Cité en pages 169 et 170.)

[Shannon 1948] C. E. Shannon. *A mathematical theory of communication.* Bell system technical journal, vol. 27, 1948. (Cité en page 156.)

[Shi & Malik 2000] Jianbo Shi et Jitendra Malik. *Normalized Cuts and Image Segmentation.* IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 8, pages 888–905, 2000. (Cité en page 33.)

[Sitbon & Bellot 2007] Laurianne Sitbon et Patrice Bellot. *Topic segmentation using weighted lexical links (WLL).* In Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr et Noriko Kando, editeurs, SIGIR, pages 737–738. ACM, 2007. (Cité en page 30.)

[Spertus 1997] Ellen Spertus. *Smokey : Automatic Recognition of Hostile Messages.* In Proceedings IAAI, pages 1058–1065, 1997. (Cité en page 36.)

[Sutton 1988] Richard S. Sutton. *Learning to Predict by the Methods of Temporal Differences*. Mach. Learn., vol. 3, no. 1, pages 9–44, Août 1988. (Cité en page 163.)

[Taboada *et al.* 2006] Maite Taboada, Mary Ann Gillies et Paul McFetridge. *Sentiment classification techniques for tracking literary reputation*. In Proceedings of LREC 2006 workshop "Towards Computational Models of Literary Analysis ?, pages 36–43, 2006. (Cité en page 36.)

[Terveen & Hill 2001] Loren Terveen et Will Hill. *Beyond Recommender Systems : Helping People Help Each Other*. In HCI in the New Millennium, pages 487–509. Addison-Wesley, 2001. (Cité en page 3.)

[Thomas *et al.* 2006] Matt Thomas, Bo Pang et Lillian Lee. *Get out the vote : Determining support or opposition from Congressional floor-debate transcripts*. In Proceedings of EMNLP, pages 327–335, 2006. (Cité en page 37.)

[Tsuchiya 1995] S. Tsuchiya. *Commensurability, a key concept of business reengineering*. In Proceedings of 3rd international symposium of the management of information and corporate knowledge, institut national pour l'intelligence artificielle, 1995. (Cité en page 4.)

[Turney & Littman 2002] Peter Turney et Michael Littman. *Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus*, 2002. (Cité en pages 38, 47, 87 et 104.)

[Utiyama & Isahara 2001] Masao Utiyama et Hitoshi Isahara. *A Statistical Model for Domain-Independent Text Segmentation*. In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, pages 491–498, 2001. (Cité en page 34.)

[Watkins & Dayan 1992] Christopher J. C. H. Watkins et Peter Dayan. *Q-Learning*. Machine Learning, vol. 8, no. 3-4, pages 279–292, 1992. (Cité en page 163.)

[Widrow & Hoff 1988] Bernard Widrow et Marcian E. Hoff. *Neurocomputing : foundations of research*. chapitre Adaptive switching circuits, pages 123–134. MIT Press, Cambridge, MA, USA, 1988. (Cité en page 160.)

[Widrow & Stearns 1985] Bernard Widrow et Samuel D. Stearns. Adaptive signal processing. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1985. (Cité en page 160.)

[Wiebe & Riloff 2005] Janyce Wiebe et Ellen Riloff. *Creating Subjective and Objective Sentence Classifiers from Unannotated Texts*. In Sixth international

conference on intelligent text processing and computational linguistics, 2005. (Cité en page 38.)

[Xu *et al.* 2011] Kaiquan Xu, Stephen Shaoyi Liao, Jiexun Li et Yuxia Song. *Mining comparative opinions from customer reviews for Competitive Intelligence.* Decision Support Systems, vol. 50, no. 4, pages 743–754, 2011. (Cité en pages 37 et 38.)

[Yang & Liu 1999] Yiming Yang et Xin Liu. *A Re-Examination of Text Categorization Methods.* pages 42–49. ACM Press, 1999. (Cité en page 161.)

[yen Kan *et al.* 1998] Min yen Kan, Judith L. Klavans et Kathleen R. McKeown. *Linear Segmentation and Segment Significance.* In Proceedings of the 6th International Workshop of Very Large Corpora, pages 197–205, 1998. (Cité en page 31.)

[Yi *et al.* 2003] Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu et Wayne Niblack. *Sentiment Analyser : Extraction Sentiments about a Given Topic using Natural Language Processing Techniques.* In IEEE Intl. Conf. on Data Mining (ICDM), 2003. (Cité en page 39.)

[Zadeh 1965] L. Zadeh. *Fuzzy sets.* Information and Control, vol. 8, 1965. (Cité en pages 125 et 168.)

[Zadeh 1978] L. Zadeh. *Fuzzy sets as a basis for a theory of possibility.* Fuzzy Sets and Systems, vol. 1, 1978. (Cité en pages 123, 125 et 168.)

[Zipf 1941] G.K. Zipf. National unity and disunity : the nation as a bio-social organism. The Principia Press, inc., 1941. (Cité en page 157.)