



Sur les traces du futur : entre comprendre et predire

Armelle Brun

► To cite this version:

Armelle Brun. Sur les traces du futur : entre comprendre et predire. Intelligence artificielle [cs.AI]. Université de Lorraine, 2018. tel-01832540

HAL Id: tel-01832540

<https://hal.science/tel-01832540>

Submitted on 8 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sur les traces du futur : entre comprendre et prédire

Mémoire déposé et présenté publiquement le 10 avril 2018

pour l'obtention d'une

Habilitation de l'Université de Lorraine
(mention informatique)

par

Armelle Brun

Composition du jury

- Président :* Bernard Girau, Professeur, Université de Lorraine, LORIA
- Rapporteurs :* Pascale Kuntz, Professeur, Université de Nantes, LS2N
Philippe Lenca, Professeur, IMT Atlantique, Lab-STICC
Agathe Merceron, Professeur, Beuth University of Applied Sciences
- Examineurs :* Anne Boyer, Professeur, Université de Lorraine, LORIA
Patrick Gallinari, Professeur, Sorbonne Université, LIP6
Shengrui Wang, Professeur, Université de Sherbrooke
- Invité :* Kamel Smaïli, Professeur, Université de Lorraine, LORIA

Mis en page avec la classe thesul.

Sommaire

1	Introduction	1
1.1	Choix effectués	1
1.2	Caractéristiques des données	3
1.3	Approche adoptée	4
1.4	Contraintes	5
1.5	Problématique et thématiques de recherche	6
2	Trajectoire de recherche	9
2.1	Premiers pas, premières années : la modélisation statistique du langage	9
2.1.1	Problématiques scientifiques	10
2.1.2	La détection de thèmes dans des énoncés	11
2.1.3	Les modèles de langage	13
2.1.4	En résumé	14
2.2	Amorce d'un virage thématique : vers la modélisation utilisateur et les systèmes de recommandation	14
2.2.1	Premiers travaux : changement de cadre, mêmes modèles	15
2.2.2	Modélisation de comportement : le problème du manque de données	15
2.2.3	Modélisation de comportement sur le Web : le problème de la résistance au bruit	16
2.2.4	Les utilisateurs : modélisation et préférences	17
2.2.5	En résumé	18
2.3	Les systèmes de recommandation : une nouvelle dynamique, de nouvelles problématiques	18
2.3.1	La gestion des préférences utilisateurs	19
2.3.2	La modélisation de relations au sein de données de comportement (PS4)	21
2.3.3	Le problème du manque de données (PS3)	23
2.3.4	La conception de modèles légers (PS2)	24
2.3.5	Des algorithmes transparents	25
2.3.6	Vers un nouveau cadre : l'éducation	26
2.3.7	En résumé	27
2.4	Synthèse	27
3	Quelques contributions majeures	31
3.1	Des utilisateurs singuliers	31
3.1.1	Notations	32

3.1.2	Les utilisateurs dits “représentatifs”	32
3.1.3	Les utilisateurs “moutons gris”	37
3.1.4	En résumé	38
3.2	Le problème de manque de données et un problème spécifique : le démarrage à froid (PS3)	39
3.2.1	Conception d’approches dédiées	39
3.2.2	Exploitation ou intégration de nouvelles sources de données	42
3.2.3	En résumé	44
3.3	Les données séquentielles : bruit, distance et influence pour la prédiction et la recommandation	44
3.3.1	Le bruit dans les données (PS1)	45
3.3.2	Les relations de distance au sein des données (PS4)	47
3.3.3	Les relations d’influence au sein des données	50
3.3.4	En résumé	50
4	Projet de recherche	53
4.1	L’éducation : domaine d’application principal	54
4.2	De la prédiction à la prescription : comment mener un utilisateur vers un objectif donné ?	58
4.2.1	Synthèse des travaux effectués en prédiction	59
4.2.2	La prescription ou comment mener un utilisateur à un but fixé	60
4.2.3	Caractéristiques et défis apportés par le cadre applicatif	63
4.2.4	La prescription vue comme un problème de fouille de données	65
4.2.5	La prescription vue comme un problème de prise de décision	68
4.3	Un système de qualité pour tous et en permanence	71
4.3.1	De la particularité des utilisateurs à des modèles particuliers	71
4.3.2	Des données véloce	72
4.3.3	De nouveaux phénomènes	73
4.4	Thématiques à lancer	75
5	Animation, administration et responsabilités	77
5.1	Vie d’une équipe de recherche et transfert	77
5.2	Rayonnement	78
5.3	Activités de valorisation	79
5.3.1	Projets en e-commerce	79
5.3.2	Projets en e-learning	80
5.3.3	Projets soumis ou en cours d’élaboration	81
5.4	Vie de la recherche	81
5.5	Mandats et activités collectives	82
5.5.1	Mandats	82
5.5.2	Expertise	82
5.5.3	Activités locales	83
5.6	Animation et administration de l’enseignement	83
5.6.1	Direction de formations	83

	3
5.6.2 Diffusion	83
Bibliographie	85

Chapitre 1

Introduction

Les phénomènes complexes, qu'ils soient physiques, météorologiques, géologiques, sonores, etc. constituent un sujet d'intérêt pour de nombreux chercheurs issus de disciplines variées telles que la biologie, les mathématiques, la physique, la chimie, etc., mais également pour le monde industriel. Tous sont très intéressés par la compréhension et la modélisation de ces phénomènes.

L'objectif de mes travaux de recherche est la modélisation numérique de tels phénomènes, et plus précisément leur modélisation prédictive. Je cherche ainsi à construire automatiquement des modèles, dans le but de comprendre les phénomènes, de raisonner sur ces derniers, d'en inférer des informations ou d'en prédire des réalisations.

Je ne m'intéresse pas à un unique phénomène, mais à un ensemble de phénomènes, partageant la caractéristique d'être réalisés par l'humain. Cette dimension humaine les rend variables, parfois incertains, mais aussi cohérents, modélisables et souvent prédictibles. Dans mes travaux, je me suis intéressée à trois phénomènes en particulier : le comportement utilisateur dans le contexte d'un système informatisé, les préférences utilisateur dans ce même contexte, et le langage naturel.

La modélisation du comportement utilisateur permet par exemple de comprendre le comportement des utilisateurs, d'identifier des comportements-type, de rapprocher différents utilisateurs, de prédire des comportements futurs, incluant des évolutions de comportement, etc.

En modélisation de préférences, il est possible d'identifier des groupements d'utilisateurs avec des préférences similaires, ou à l'opposé d'identifier des utilisateurs avec des préférences spécifiques, de rapprocher des ressources sur les préférences, de tracer l'évolution des préférences, d'inférer des préférences inconnues, etc. La modélisation du langage permet, quant à elle, de comprendre le contenu de documents, d'identifier des structures fréquentes, ou au contraire peu fréquentes, de la langue ou entre langues, de prédire la suite de documents ou de phrases, d'identifier des documents traitant de thèmes similaires, etc.

L'intelligence artificielle, et plus précisément l'apprentissage automatique et la fouille de données, constituent l'approche générale de mes activités. Les travaux que je mène et les modèles que je conçois peuvent par conséquent être aisément adaptés à d'autres phénomènes.

Dans ce document je présente les différents travaux que j'ai menés depuis une quinzaine d'années, qui visent dans la très grande majorité à la modélisation de ces trois phénomènes. Pour mener à bien ces travaux, j'ai effectué plusieurs choix, que j'expose ici.

1.1 Choix effectués

Exploiter des traces de réalisation

Pour modéliser un phénomène, je choisis d'adopter une approche qui ne s'appuie pas sur des connaissances *a priori* de ce dernier, mais sur des données empiriques : des **traces de réalisation** / d'observation du phénomène. Ces traces concernent des **observés** (l'objet d'étude), en lien avec des **objets** (éléments) relatifs au phénomène. Je fais l'hypothèse que les traces exploitées sont représentatives du phénomène à modéliser, je ne les mets pas en doute *a priori*.

Je considère que le fait de ne pas exploiter de connaissances *a priori* sur le phénomène à modéliser constitue un avantage sur deux plans :

1. Cela permet d'éviter la phase, potentiellement complexe et coûteuse, de collecte des connaissances. Plus encore, de telles connaissances peuvent ne pas être disponibles ; c'est notamment le cas en modélisation du comportement et en modélisation des préférences. Leur collecte est donc inenvisageable.
2. Un modèle appris sur les traces pourra contenir une information différente de celle d'un modèle représentant les connaissances *a priori*, je suis même convaincue qu'elle peut être plus précise. En effet, les lois générales/règles régissant un phénomène constituent une première information relative à celui-ci, mais je pense que les traces de réalisation constituent une source d'informations qui représente mieux la réalité du phénomène. Elles peuvent par exemple contenir un sous-ensemble des règles ou être en contradiction partielle avec ces dernières. Le modèle appris devrait par conséquent être plus performant.

Considérons les trois phénomènes mentionnés ci-dessus. En modélisation du langage naturel, l'observé est par exemple le document textuel ; l'objet est l'entrée lexicale (le mot) et les traces de réalisation représentent le contenu de textes : les différents objets et les relations entre eux. Le contenu de ces textes peut ne pas respecter certaines règles de la langue. Par exemple, il n'est pas rare qu'en langue française *ne* soit omis avant *pas* : "*ils ont pas de chance*". Cette spécificité se retrouvera ainsi dans le modèle appris sur ces traces. Un tel modèle permettra également de rendre compte que certaines structures sont plus communes que d'autres ou encore que certaines, bien que respectant les règles générales, ne sont jamais ou très rarement utilisées. Le modèle de langue appris sur des traces de réalisation permettra donc de représenter les réalisations effectives du phénomène.

En modélisation du comportement dans le contexte d'un système informatisé, l'observé est l'utilisateur, l'objet représente la ressource du système et les traces de réalisation représentent les interactions (consultation, impression, etc.) d'utilisateurs avec des ressources. Ces traces sont des données dites implicites, elles ne sont en règle générale pas laissées consciemment par les utilisateurs. Ici, aucune règle/connaissance *a priori* n'est disponible, les traces constituent donc la principale source d'informations pour la modélisation du phénomène.

En modélisation des préférences, l'observé et l'objet sont également respectivement l'utilisateur et la ressource du système. Les traces de réalisation seront, quant à elles, les préférences des utilisateurs, qui peuvent se présenter sous la forme de notes, de *like*, de commentaires, etc. que les utilisateurs auront laissés sur des ressources. Les traces sont dites explicites. Le système peut par ailleurs inférer des données explicites (notamment des notes) à partir du comportement/activité des utilisateurs (les traces implicites). Ici également, aucune règle/connaissance *a priori* n'est disponible.

Mon objectif est donc d'apprendre un modèle de ces traces, qui sera exploité, dans un second temps, pour inférer des informations ou prédire d'autres réalisations du phénomène.

N'exploiter aucune connaissance sur les observés ni sur les objets

Outre le fait de ne pas utiliser de connaissances *a priori* sur les phénomènes à modéliser, je fais également le choix, dans la majorité des travaux que je mène, de ne pas exploiter d'information ou de connaissance supplémentaire sur les objets ou sur les observés. Ainsi, lorsqu'un objet est un mot, je n'exploite ni sa nature, ni sa fonction dans la phrase. Lorsqu'il est une ressource d'un service en ligne, par exemple un film dans le cadre d'un service de vidéo à la demande, je n'exploite ni son titre, ni les mot-clés le représentant, ni son année de sortie, etc. Lorsqu'un observé est un utilisateur, je ne cherche pas à connaître son nom, son âge, sa CSP, etc. Les observés et les objets sont par conséquent uniquement connus au travers d'un identifiant non interprétable. Je fais ce choix pour trois raisons :

1. Des informations supplémentaires, qu'elles soient sur les objets ou sur les observés, peuvent, comme précédemment, ne pas être disponibles. Par conséquent, concevoir un modèle reposant sur ces informations présente le risque de ne pouvoir être envisageable dans le cas où ces informations ne pourraient être collectées. Par ailleurs, l'acquisition de ces informations peut nécessiter une intervention humaine coûteuse, non seulement en temps mais également d'un point de vue financier (tout comme cela pouvait être coûteux de collecter des informations *a priori* sur les phénomènes).

2. La législation impose de concevoir des modèles qui garantissent la préservation de la vie privée. Cette contrainte est donc importante dans le cas où les traces de réalisation concernent des utilisateurs. Ne pas exploiter d'informations supplémentaires sur les observés, ici des données à caractère personnel sur les utilisateurs, est un moyen de favoriser la préservation de la vie privée de ces derniers. Toutefois, je n'exclus pas la possibilité d'exploiter des informations supplémentaires lorsqu'elles sont indispensables à la compréhension du phénomène, et qu'elles sont disponibles. C'est notamment le cas en e-éducation, où les traces d'apprentissage des étudiants, utilisées seules, sont insuffisantes pour modéliser de façon précise la tâche d'apprentissage. Des données supplémentaires, telles que des informations issues du système d'information des institutions d'enseignement, sont indispensables. Ce cas sera traité dans le chapitre 4 de ce manuscrit.
3. N'utilisant pas d'informations supplémentaires, les modèles que je conçois ne reposent que sur les traces. L'algorithme d'apprentissage des modèles n'est donc dépendant ni des observés ou objets sur lesquels il porte, ni de leur nature ; il pourra donc être réutilisé sur des traces portant sur d'autres phénomènes, sans nécessiter d'adaptation conséquente.

Collecter automatiquement les traces

Je fais le choix de collecter automatiquement les traces, sans aucune sollicitation ou intervention humaine, ce qui permet de disposer de données en très grande quantité, en permanence, et à un coût limité.

Le Web s'étant largement démocratisé, il est désormais une source de traces très riche, permettant la collecte d'un gros volume de données et, d'un point de vue technique, il est adapté à la collecte automatique. Le Web est donc une des sources que j'ai choisi d'exploiter. Bien évidemment, ces traces ont des caractéristiques dont il faudra que je tienne compte dans mes travaux.

Par ailleurs, toute source numérique de données est une source que je pourrai considérer.

1.2 Caractéristiques des données

En raison des choix que j'ai effectués, les traces de réalisation, que j'exploite, ont plusieurs caractéristiques, que j'expose ci-dessous.

Les traces sont incertaines, variables et peuvent être bruitées

L'incertitude des données est principalement due à la modalité de leur collecte, certaines pouvant même tout simplement être fausses.

Le bruit dans les données est quant à lui lié à la nature de la source de données. En effet, grâce à la démocratisation d'Internet, les utilisateurs ne sont désormais plus de simples consommateurs, ils produisent une énorme quantité de données au travers de leurs nombreuses activités en ligne. Ces différentes informations sont soit laissées intentionnellement : écriture dans des blogs, partage de ressources (vidéos, photos, etc.), dépôt d'avis, etc., soit involontairement : navigation, requêtes, etc.

La variabilité intrinsèque à l'humain se retrouve également dans les données collectées. Cependant, ces données, issues d'activités humaines, sont majoritairement cohérentes.

Les traces sont collectées en très grand nombre et sont hétérogènes

Comme mentionné précédemment, la collecte automatique des traces a pour effet de pouvoir disposer de données en très grand nombre et en temps réel. Ces données peuvent même être très véloces.

Par ailleurs, les multiples sources de données disponibles et collectables ont pour conséquence que les données exploitables, et exploitées, sont ou peuvent être hétérogènes.

Les traces sont parcimonieuses

Bien qu'elles soient collectées en très grand nombre, les traces sont parcimonieuses à l'échelle des observés, mais aussi à l'échelle des objets. Cette parcimonie est due au fait que pour chaque observé

ou chaque objet, les traces existent uniquement en très petite quantité. Si l'on considère le cas de la modélisation de préférences, bien que disposant de très grandes quantités de traces, nous ne disposons que de peu de données de préférences sur chaque utilisateur, car chacun exprime ses préférences sur peu d'objets, relativement au nombre d'objets possibles (notons que le problème reste le même lorsque les préférences sont déduites des données de comportement). Dans le cas de la modélisation du langage, bien que de nombreux textes soient automatiquement collectables, la taille des documents est finie et le nombre de réalisations de chaque entrée lexicale est relativement réduit, limitant ainsi la possibilité de modélisation du phénomène.

Les traces collectées sont relatives aux observés

Un élément de trace n'a pas la même signification ou la même importance selon l'observé auquel il est rattaché. En modélisation du langage naturel par exemple, un même mot peut ne pas être interprété de la même manière dans deux textes ; en modélisation du comportement, une même action peut ne pas avoir le même sens ou le même rôle pour deux utilisateurs.

Les traces peuvent être périssables et évolutives

Les données disponibles, ou collectées à un moment donné, ne sont potentiellement plus pertinentes après un certain temps, voire deviennent fausses. Par ailleurs, de nouvelles données apparaissant régulièrement, elles évoluent, de même que les phénomènes qu'elles représentent.

Les modèles que je conçois doivent par conséquent prendre en compte l'ensemble des caractéristiques énoncées ci-dessus.

1.3 Approche adoptée

Pour modéliser un phénomène à partir de traces de réalisation, et sachant les caractéristiques de ces traces, je choisis d'adopter une approche par **apprentissage statistique**, avec un intérêt particulier pour la **fouille de données**. Ces approches sont connues pour être adaptées au traitement de données volumineuses et bruitées. Par ailleurs, ces approches étant automatiques, elles peuvent tenir compte des nouvelles données apparaissant régulièrement.

Dans la littérature, la majorité des travaux portant sur la modélisation de phénomènes, suppose que les réalisations sont identiquement distribuées : un processus unique les a générées, et ces réalisations sont indépendantes. Par conséquent, l'ensemble des traces est utilisé pour apprendre un unique modèle. Cependant, en fonction des phénomènes auxquels on s'intéresse, cette supposition peut ne pas être réaliste. La majorité des travaux que je mène ne repose pas sur l'hypothèse qu'un seul modèle est à la source des réalisations.

En modélisation du langage, on peut aisément supposer qu'un unique modèle ne peut représenter l'ensemble de la production en langage naturel. En effet, un discours, un article journalistique, un roman, etc. sont très différents. En modélisation des préférences, il n'est, là non plus, pas envisageable de considérer qu'il existe un unique modèle de préférences. Chaque utilisateur ayant des préférences qui lui sont propres, il serait judicieux de développer un modèle par utilisateur. De la même façon, en modélisation du comportement, chaque utilisateur a sa propre façon de se comporter, et un unique modèle trouvera rapidement ses limites si l'objectif est de représenter de façon précise les spécificités de chacun.

Cependant, étant donné la **parcimonie** des traces à l'échelle des observés, et le choix que j'ai fait de ne pas exploiter de connaissances *a priori* sur le phénomène, ni, dans la mesure du possible, de connaissances externes sur l'observé ou sur les objets, il n'est pas envisageable d'apprendre un modèle pour chaque observé.

Pour apprendre des modèles, je choisis de tirer profit de la **cohérence entre les observés**. De cette façon, pour modéliser un observé (un document, un utilisateur), j'exploite les informations relatives à d'autres observés. Dans le cadre de la modélisation statistique du langage, cela signifie que certains documents ont des structures ou des vocabulaires similaires voire identiques dans certains cas, et que la connaissance de cette proximité entre documents peut être exploitée pour apprendre un modèle de la

langue. En modélisation des préférences, cela signifie que les préférences entre certains utilisateurs sont proches, et que des préférences d'autres utilisateurs peuvent être exploitées pour modéliser celles d'un utilisateur donné.

1.4 Contraintes

Je choisis d'imposer deux principales contraintes aux modèles que je conçois.

Modèles interprétables

Les choix que j'ai faits relativement à la collecte des traces, font qu'elles ne sont pas interprétables *a priori*, notamment pour un utilisateur grand public. Par ailleurs, le volume de ces traces fait que même les experts du phénomène ne peuvent plus les interpréter.

Une de mes préoccupations récentes est de former des modèles dits transparents, c'est-à-dire des modèles qui puissent être interprétés par des humains. La transparence peut concerner deux aspects dans les modèles. Tout d'abord, ce peut être le modèle en lui-même qui est transparent : un humain (expert) peut comprendre ce dernier. L'avantage d'une telle interprétabilité est la possibilité, pour l'expert, d'évaluer la validité du modèle en amont de son exploitation. Ensuite, ce peut être la sortie du modèle, lorsqu'il est mis en œuvre, qui est transparente : le modèle est capable de justifier la/les raison(s) de ses résultats.

De nombreux travaux de la littérature ont pour objectif principal la qualité de la modélisation. Cependant, ni les modèles résultants, ni les résultats ne sont en général directement exploitables par un humain. Cela constitue, pour moi, une limite de ces modèles car ils sont composés de données numériques uniquement exploitables par un autre processus.

Plusieurs approches permettent cependant de former des modèles interprétables. Je pense notamment aux approches qui visent à découvrir des indicateurs (ou indices) et/ou des structures (régularités) dans les traces. Ces régularités et indicateurs peuvent porter sur les objets, sur les observés ou encore sur les objets et les observés à la fois. Je me suis naturellement tournée vers ces approches, et notamment vers la fouille de données, que j'adapte aux données et aux problématiques sur lesquelles je travaille. La fouille de données vise à identifier des motifs ou des règles utiles (dans le sens où ils représentent une connaissance nouvelle), statistiquement significatifs, dans une grande base de données. La fouille de données est utilisée dans de nombreux domaines : en commerce dans le but d'identifier les habitudes d'achat des clients, en santé pour découvrir les liens entre gènes et maladies ou pour prédire de futures épidémies, en météorologie, politique, psychologie, etc. Dans le cadre de mes travaux, les motifs et règles fouillés sont exploités pour expliquer ou comprendre des phénomènes, résumer les données ou prédire des éléments relatifs aux traces. Par exemple, en langage naturel les motifs pourront être des groupes (ordonnés ou non) de mots partageant des caractéristiques, en modélisation de comportement elles pourront être des suites (contiguës ou non) de ressources représentant des séquences-type de comportements, etc.

Modèles performants, robustes, peu complexes et dynamiques

Pour que les régularités et motifs obtenus par un algorithme de fouille de données soient pertinents, la quantité de données sur laquelle le modèle est appris doit être significative. Notons qu'en l'espace de dix-quinze ans, la relation au volume de données a complètement changé. Lorsque j'ai débuté mon activité de recherche, concevoir un modèle à partir de traces nécessitait de s'assurer de pouvoir collecter suffisamment de données pour que les modèles appris soient fiables et représentatifs du phénomène modélisé. Le problème actuel est tout autre : il n'est plus de s'assurer que la quantité de données collectées est suffisante, mais de s'assurer de concevoir un algorithme capable de traiter l'énorme volume de données disponible. En 2016 par exemple, 2,5 exabytes de données étaient produits chaque jour sur le Web¹, parmi lesquels 100 teraoctets de données générées quotidiennement par Twitter [Lin and Ryaboy, 2013]. La chaîne américaine Walmart enregistre 1 milliard de transactions toutes les 28 heures [Reynolds, 2016], et les traces d'achats sont aisément collectables.

1. <http://www.northeastern.edu/levelblog/2016/05/13/how-much-data-produced-every-day/>

Bien que très (trop) nombreuses, ces données sont cependant potentiellement toutes utiles, car elles peuvent toutes contribuer à la couverture du modèle formé. Les modèles doivent donc être **peu complexes en temps d'apprentissage**.

Par ailleurs, le volume de données disponible est désormais tel qu'il n'est même plus envisageable d'étudier manuellement les données pour identifier les sources de données, ou les données, qui ne sont pas intéressantes, non fiables ou trop bruitées, etc. Les modèles conçus doivent donc pouvoir automatiquement écarter, ou tout du moins gérer les données qui peuvent être sources de problème (données bruitées par exemple), les modèles doivent donc être **robustes**.

Enfin, de nouvelles données arrivant fréquemment, le modèle doit être capable de se mettre à jour aisément et rapidement, il doit être **dynamique**.

Je rappelle ici que dans les phénomènes auxquels je m'intéresse, malgré le gros volume de données, on est face à un paradoxe : la quantité totale de données pouvant être recueillie est énorme, alors que pour chaque observé (utilisateur par exemple) la quantité est relativement (voire très) faible, les données sont parcimonieuses. Malgré cela, le modèle doit être **performant**.

1.5 Problématique et thématiques de recherche

Le contexte, les choix effectués, les caractéristiques des données, l'approche et les contraintes étant posés, je peux maintenant introduire la problématique scientifique générale à laquelle je m'intéresse dans mes recherches :

Comment construire, à partir de traces de réalisation volumineuses, parcimonieuses, bruitées, relatives, périssables et non interprétables, des modèles performants, robustes, peu complexes et dynamiques d'indices/indicateurs ou de régularités qui permettent de prédire des réalisations ou d'inférer des données relatives aux observés ?

Ces modèles ont pour but de mieux connaître et comprendre les réalisations et les observés. Par ailleurs, dans le cas où les observés sont des utilisateurs, ces modèles peuvent être exploités pour les aider, prédire des réalisations voire les anticiper.

Je précise ici que l'apprentissage (la construction) de ces modèles peut se faire *hors ligne*, en exploitant un corpus de traces : un jeu de données de la communauté scientifique, un jeu de données réel fourni par des partenaires de recherche notamment des partenaires industriels ou encore un corpus de traces collecté par nos soins. Il peut également se faire *en ligne* si l'on souhaite construire un modèle qui intègre les traces au fur et à mesure de leur disponibilité, ce qui permet également une boucle de retour pour adapter le modèle.

De la même façon, l'évaluation de ces modèles peut se faire soit *hors ligne*, soit *en ligne*. Lorsque l'évaluation est faite hors ligne, elle permet d'évaluer la qualité de l'inférence ou de la prédiction, mais également la robustesse. Lorsqu'elle est faite en ligne, la principale différence concerne l'évaluation de la qualité de l'inférence/prédiction, qui est faite en interaction directe avec les utilisateurs des systèmes sur lesquels les modèles sont implémentés.

Le processus de fouille de données est traditionnellement vu comme un processus en cinq étapes : sélection de la source de données, prétraitement des données, transformation des données, fouille de données et enfin interprétation/évaluation du résultat de la fouille [Fayyad et al., 1996]. L'étape de prétraitement des données a pour but de supprimer le bruit, de gérer les données manquantes, les erreurs de mesure, les données aberrantes (outliers), de réduire la dimension des données en trouvant les caractéristiques utiles pour les représenter. L'étape de fouille consiste à choisir ou à proposer des algorithmes de fouille adaptés aux caractéristiques des données, dans le but d'identifier des motifs au sein de ces dernières.

Mon intérêt porte majoritairement sur cette dernière étape de fouille, et à la marge sur l'étape de prétraitement. Dans les travaux que j'ai menés, les données fouillées peuvent se présenter sous deux formes. Elles peuvent être sous forme matricielle, où chaque élément de la matrice représente une interaction (quantifiée ou non) entre un observé et un objet, ou sous forme de graphe où chaque nœud est un observé ou un objet et les arcs représentent les interactions ou les liens entre ces derniers. Les possibilités de traitement des données sous forme matricielle et sous forme de graphes sont, dans la majorité des

cas, similaires. Elles peuvent également se présenter sous forme séquentielle, ce qui reflète également une interaction entre un observé et un objet, mais dans un contexte séquentiel d'interactions.

Dans le cadre de ces formats de données, je me suis intéressée à plusieurs défis, j'en présente un sous-ensemble ici :

- **La modélisation prédictive séquentielle sur des données bruitées.** Du point de vue des phénomènes que j'étudie, cette problématique concerne le langage naturel et le comportement utilisateur. Concernant le langage naturel, l'objectif est la conception de modèles de prédiction de mots (j'utilise le terme "mot" au sens très large) qui peuvent constituer la suite d'une phrase ou d'un texte (écrit ou oral), sachant la réalisation du début de cette phrase/texte. Les domaines d'application sur lesquels j'ai travaillé regroupent la reconnaissance de la parole, la détection de thèmes dans des textes et le routage d'e-mails. Concernant le comportement utilisateur, l'objectif est la conception de modèles de prédiction de comportement : prédire les objets vers lesquels un utilisateur va se diriger connaissant son historique de comportement. Les cadres d'application sur lesquels j'ai travaillé regroupent le e-(m)-commerce, le Web (navigation, prévisions), les intranets documentaires, le domaine bancaire et plus récemment la e-éducation et la détection d'émergence. Ces problématiques ont été abordées durant mon doctorat, les premières années de mon activité en tant que maître de conférences et dans les thèses d'Ilham Esslimani, Geoffray Bonnin, Lina Fahed et Yacine Abboud que j'ai co-encadrées. Dans ces travaux, je me suis penchée sur plusieurs aspects : la classification, la maîtrise de qualité de la prédiction, la gestion de la distance de prédiction, la robustesse, l'identification de motifs, en incluant des contraintes sur ces motifs, l'émergence de motifs, la prédiction de comportements futurs, à plus ou moins brève échéance, etc.
- **L'inférence de données dans des données matricielles.** Du point de vue des phénomènes que je modélise, cette problématique concerne la modélisation de préférences et l'objectif est d'inférer des préférences (futures) d'utilisateurs sachant leurs préférences passées. Le cadre applicatif concerne majoritairement les systèmes de recommandation. Cette problématique a été abordée dans les thèses d'Ilham Esslimani, Marharyta Aleksandrova et Oleksandr Palchenko, que j'ai co-encadrées. Je me suis penchée sur les aspects du manque de données, la qualité de l'inférence (des recommandations), la complexité du modèle, le problème du démarrage à froid, etc.
- **La représentation synthétique et interprétable des données.** Du point de vue des phénomènes à modéliser, cette problématique concerne les trois phénomènes. L'objectif est à la fois d'identifier des éléments dans les données qui permettent de représenter de façon simple (et interprétable) celles-ci, et de concevoir des modèles qui permettent d'inférer/prédire des données de façon interprétable et compréhensible. Cette problématique a été abordée dans le cadre d'une collaboration avec Amine Boumaza, chercheur spécialiste des algorithmes évolutionnaires, mais également avec Oleg Chertov Professeur à KPI (Ukraine) dans le cadre du co-encadrement de la thèse de Marharyta Aleksandrova et enfin dans la thèse de Julie Budaher que je co-encadre.
- **L'identification et la modélisation d'éléments impossibles, rares ou différents.** Du point de vue des phénomènes à modéliser, cette problématique concerne également les trois phénomènes. L'objectif est l'identification d'utilisateurs avec des préférences qui ne sont pas en cohérence avec celles des autres, l'identification de comportements rares qui peuvent être signe d'une émergence, ou encore l'identification de séquences impossibles de la langue. Cette problématique a été abordée durant mon doctorat et lors des thèses de Lina Fahed, Benjamin Gras et Yacine Abboud que j'ai co-encadrées. Le défi ici est de concevoir des modèles qui, dans le cadre de données bruitées et parcimonieuses, identifient des motifs impossibles, peu rencontrés ou atypiques (par définition peu rencontrés et différents des autres), et qui, pour les deux derniers cas, les modélisent.

Le chapitre 2 constitue une synthèse de mes activités de recherche et d'encadrement, présentées de façon chronologique. Il présente également de façon succincte les projets académiques et industriels dans le cadre desquels ces activités ont été menées, le cas échéant. Le chapitre 3 s'intéresse à trois thématiques spécifiques de mes recherches, et présente mes contributions scientifiques en lien avec ces thématiques. Le chapitre 4 introduit les activités que je souhaite mener dans les années à venir. Le chapitre 5 résume, quant à lui, mes activités d'animation de la recherche, mais également de l'enseignement. Enfin, le chapitre 6 comprend un sous-ensemble de publications représentatives de mon activité.

Chapitre 2

Trajectoire de recherche

Ce chapitre présente un panorama de mon activité de recherche, de ses débuts lors de mon DEA jusqu'à mes travaux actuels. Pour décrire et expliciter mes travaux, je choisis ici une présentation chronologique, pour comprendre leur évolution sur les différentes années et périodes. J'y présenterai notamment les thèmes abordés, les problématiques identifiées, les approches choisies, les résultats saillants, les points d'originalité, mais aussi les encadrements scientifiques et les projets académiques et industriels dans lesquels je me suis investie. Je choisis de diviser mon activité en 3 périodes principales.

Durant la première période, qui correspond aux années 1999 à 2007, mon activité a porté sur la modélisation statistique du langage avec pour cadre applicatif principal la reconnaissance automatique de la parole. C'est durant cette période que j'ai effectué mes premiers encadrements scientifiques (Master recherche).

Le début de la seconde période est marqué par ma volonté d'évoluer thématiquement. J'ai choisi de m'intéresser à la modélisation utilisateurs : modélisation du comportement et des préférences, avec pour cadre applicatif principal les systèmes de recommandation. Bien que cette seconde période soit marquée par une évolution de la thématique de mes recherches, j'ai choisi de conserver la même approche scientifique, à savoir une approche statistique, qui est d'ailleurs toujours celle que j'adopte dans mes activités actuelles. Durant cette seconde période, qui se situe sur les années 2006 à 2010, outre le fait de contribuer activement à la recherche en systèmes de recommandation, j'ai effectué mes premiers co-encadrements de thèse et j'ai participé à plusieurs projets académiques et industriels.

Durant la troisième période, débutée en 2010, j'ai confirmé mon évolution thématique. J'ai proposé de nouvelles orientations scientifiques, j'ai également co-dirigé plusieurs thèses et j'ai aussi assumé la responsabilité de lots et de projets.

Une quatrième et dernière période, débutée très récemment, ouvre une nouvelle voie à mes recherches. Cette période sera présentée dans le chapitre projet de ce manuscrit.

2.1 Premiers pas, premières années : la modélisation statistique du langage

J'ai effectué mes premiers pas en recherche lors de mon stage de DEA en 1999. Mon travail a porté sur le domaine de la modélisation statistique du langage, avec pour cadre applicatif la reconnaissance automatique de la parole (RAP). Il s'est effectué sous la direction de Kamel Smaïli de l'équipe Parole du laboratoire LORIA (Laboratoire lOrrain de Recherche en Informatique et ses Applications). J'ai choisi de poursuivre en doctorat, toujours dans le domaine de la modélisation statistique du langage. Les recherches que j'ai menées durant mes premières années d'activité en tant que maître de conférences (j'ai été nommée en 2003) ont également porté sur ce domaine, avant d'amorcer une évolution thématique en 2006.

Durant cette première période, j'ai identifié des grandes problématiques de recherche sur lesquelles j'ai souhaité travailler, problématiques introduites dans la section ci-dessous. Plusieurs d'entre elles sont d'ailleurs présentes tout au long de mon activité de recherche, elles en constituent un fil directeur. Au fur

et à mesure des périodes, je les ai cependant abordées sous un angle différent, dans des cadres applicatifs autres et j’y ai introduit de nouvelles contraintes. J’ai par ailleurs identifié de nouvelles problématiques.

Un système de RAP est composé de deux modules : un module de traitement de signal (le décodeur acoustico-phonétique), dont le but est de déterminer la suite de phonèmes correspondant au signal acoustique d’un énoncé, et un module de modélisation du langage, dont le but est de valider la suite de phonèmes déterminée par le premier module. Le modèle de langage sous-jacent capture et caractérise les règles ou les régularités du langage naturel. Notons ici qu’un modèle de langage peut être utilisé pour d’autres applications que la RAP, comme la traduction automatique, le résumé automatique, etc. C’est la conception de modèles de langage qui constitue le cœur de mes recherches.

Pour être considéré comme étant de qualité, un modèle de langage doit représenter de façon précise les caractéristiques des énoncés sur lesquels il est utilisé. Dans un système de RAP, les énoncés sont ceux qui doivent être reconnus par le système. Ainsi, meilleure est la qualité du modèle de langage, meilleure devrait être la qualité de la reconnaissance.

Plusieurs approches sont possibles pour la modélisation du langage. J’ai choisi une approche statistique, pour laquelle le modèle de langage est appris automatiquement sur un ensemble de données, données dites de réalisation ou d’observation.

J’introduis ci-dessous les problématiques scientifiques que j’ai identifiées et traitées durant cette période. Je présente ensuite des travaux que j’ai menés, en lien avec ces problématiques, en soulignant les contributions principales et les encadrements associés. Ces travaux se divisent en deux grands ensembles : ceux dédiés à la détection de thème et ceux plus généraux sur les modèles de langage.

2.1.1 Problématiques scientifiques

Dans un système de RAP, les énoncés sur lesquels le modèle de langage est utilisé sont bruités et contiennent des erreurs, puisqu’ils sont issus d’un module de reconnaissance de la parole. Un modèle de langage de qualité dans le cadre d’un système de RAP doit donc pouvoir être robuste aux erreurs, ainsi qu’au bruit, dans les données. La première problématique scientifique à laquelle j’ai choisi de m’intéresser a donc été **PS1 - la conception de modèles robustes au bruit et aux erreurs dans les données**.

Lorsque des modèles sont appris sur des volumes de données conséquents et sur des données en grande dimension, la complexité des modèles conçus est une question qui se pose naturellement. Le problème de la complexité se pose également lorsque l’exploitation des modèles se fait en situations réelles. En modélisation du langage, plus le vocabulaire exploité est large et les structures évoluées, plus le modèle est de qualité, mais plus il est complexe. La seconde problématique sur laquelle j’ai souhaité me pencher a donc été **PS2 - la conception de modèles de faible complexité à large couverture**.

La qualité d’un modèle statistique de langage dépend, entre autres, de l’adéquation entre les données sur lesquelles il a été appris et celles sur lesquelles il est exploité. Il est donc judicieux de choisir des données d’apprentissage (de réalisation) les plus proches possibles du futur contexte d’application. Cela a évidemment pour conséquence que le volume de données utilisé pour l’apprentissage du modèle est réduit. Il est classique de lire que le monde fait actuellement face à une surabondance de données disponibles, et de très nombreux travaux de recherche ont pour objectif de concevoir des algorithmes permettant de gérer la masse d’informations. Cependant, nombreux sont les phénomènes qui restent totalement ou en partie sous-dotés en données d’apprentissage. La troisième problématique sur laquelle je me suis penchée est donc **PS3 - l’apprentissage de modèles sur des données disponibles en faible quantité**.

La validation des modèles sur des données réelles et en conditions réelles est une constante dans mes activités de recherche. Cela me permet d’évaluer de façon plus poussée les modèles. Les nombreux projets industriels auxquels j’ai activement participé et dans lesquels je suis investie, m’ont permis de réaliser cette validation.

Dans les problématiques identifiées, j’ai volontairement utilisé le terme “modèle” et non pas “modèle de langage”. En effet, ces problématiques sont génériques, elles seront et pourront être abordées dans d’autres cadres.

Les problématiques scientifiques étant introduites, je présente dans les sections ci-dessous les travaux que j’ai menés en modélisation statistique du langage, et j’explique en quoi ils permettent de contribuer

à répondre à ces problématiques.

2.1.2 La détection de thèmes dans des énoncés

Les travaux que je présente ici reposent sur l'hypothèse que les caractéristiques d'un énoncé sont impactées par le thème traité dans celui-ci, notamment au niveau lexical. Ainsi, si un système de RAP exploite un modèle de langage en adéquation avec le thème de l'énoncé qu'il est en train de reconnaître, alors la qualité de la reconnaissance devrait être accrue. Cependant, le thème d'un énoncé n'est généralement pas connu en amont de la tâche de reconnaissance. Par conséquent, celui-ci devra être identifié au cours de la reconnaissance pour pouvoir utiliser un modèle de langage adéquat.

Je me suis donc intéressée au problème de la détection automatique de thèmes dans des énoncés.

J'ai choisi de construire automatiquement des modèles de détection, par apprentissage automatique, à partir d'énoncés. J'ai abordé le problème de la détection de thèmes comme un problème de classification. En disposant d'un ensemble d'énoncés dont le/les thème(s) (la classe) était(ent) connu(s), mon but a été d'apprendre le modèle, ici le classifieur, permettant de déterminer le/les thèmes (la/les classe(s)) d'un document donné. J'ai cherché à proposer des algorithmes de classification performants, relativement aux caractéristiques du cadre applicatif et apportant des réponses aux trois premières problématiques scientifiques posées précédemment.

Données d'évaluation

J'ai choisi de travailler sur deux types de d'énoncés. Le premier type est constitué d'énoncés de qualité (sans bruit, sans erreur), longs et disponibles en grande quantité. Le second type est constitué d'énoncés relativement courts, bruités, contenant des erreurs et disponibles en quantité réduite. Le choix de ces deux types a été fait pour évaluer non seulement la qualité des modèles, leur robustesse au bruit et aux erreurs dans les énoncés (**PS1**), mais aussi leur dépendance à la quantité de données d'apprentissage (**PS3**). Concrètement, les énoncés sur lesquels j'ai travaillé sont des articles journalistiques (énoncés sans bruit, sans erreur, longs et disponibles en grande quantité) et des e-mails (énoncés avec fautes, courts, et disponibles en plus petite quantité).

Contributions

Pour traiter du problème de la détection automatique de thèmes, je me suis naturellement posée trois questions : Comment représenter un document ? Quel vocabulaire exploiter ? Quel type de classifieur utiliser ? Les réponses à ces questions sont bien évidemment liées puisque la construction d'un classifieur dépend de la représentation des documents et du vocabulaire utilisé, et inversement.

Après un premier état de l'art des différentes approches et techniques de sélection de vocabulaire, de représentation de documents et de classifieurs [Brun, 1999], j'ai mené une large étude expérimentale pour l'évaluation des différentes approches de sélection de vocabulaire et des classifieurs de la littérature [Brun et al., 2000b]. J'ai ainsi pu identifier les limites de l'état de l'art, notamment en ce qui concerne la sélection de vocabulaire et la représentation des documents. J'ai donc proposé de nouvelles approches pour ces deux aspects.

Sélection de vocabulaire

- Plus un vocabulaire utilisé est riche, plus le modèle qui l'exploite devrait être performant, mais plus il est complexe. J'ai souhaité ici contribuer à **PS2**, relativement à la complexité de l'algorithme. Partant du constat que chaque classe (thème) n'a pas la même représentativité dans le corpus d'apprentissage, ce qui limite la qualité du modèle, j'ai proposé de constituer un vocabulaire par classe, indépendamment les uns des autres (les approches classiques forment un vocabulaire général à partir de l'ensemble des données d'apprentissage). Un vocabulaire global est ensuite formé par l'union des vocabulaires de thème, il a donc la caractéristique de représenter équitablement chaque thème, ne biaisant *a priori* pas les résultats. Les expérimentations menées ont montré non seulement une amélioration des performances des algorithmes de l'état de l'art, mais également une diminution significative de la taille du vocabulaire, diminuant par là même le temps d'exécution des modèles

[Bigi et al., 2001c]. Une étude comparative des performances sur les deux types d'énoncés a également permis de montrer une bonne résistance à l'erreur du vocabulaire et des modèles associés [Bigi et al., 2001a] (**PS1**).

Ces travaux ont été poursuivis quelques années plus tard lors du **stage de M2R Informatique de Julien Lourdin, que j'ai encadré** en 2006. Il avait pour objectif d'établir un état de l'art des différentes méthodes d'étiquetage automatique d'e-mails (données bruitées et courtes) dans un but de routage automatique. Ces méthodes reposaient majoritairement sur le vocabulaire utilisé.

- Toujours dans le cadre de la sélection de vocabulaire, j'ai souhaité m'intéresser à une mesure autre que la traditionnelle fréquence des mots. J'ai proposé d'exploiter la similarité des mots avec les thèmes, calculée à partir de l'information mutuelle entre mots. **Cette approche originale a été la première en modélisation du langage à ne pas reposer sur la fréquence des mots.** Les performances en classification obtenues ont non seulement été à nouveau accrues, mais l'avantage principal a, là encore, résidé dans une diminution très significative de la taille du vocabulaire requis, amenant ainsi à un temps d'exécution et à un espace requis significativement plus faibles [Brun et al., 2002a, Brun et al., 2002b]. Cette proposition a ainsi permis de contribuer à la problématique **PS2**.

La conclusion forte de ces travaux est que la présence, l'absence et la fréquence d'éléments dans des données d'apprentissage ne constituent pas la seule information à exploiter pour permettre une modélisation de qualité.

Représentation des documents J'ai également travaillé sur la représentation des documents en proposant une représentation plus riche que ne le fait l'état de l'art, mais tout en limitant la complexité de la représentation, et par conséquent celle des classifieurs. J'ai proposé de représenter les documents non plus sous la forme de vecteurs de mots mais de vecteurs de suites de mots, des vecteurs de n -grammes. Cette représentation est inspirée de la modélisation statistique du langage. J'ai pu montrer une amélioration significative de la qualité de la classification, tout en limitant la complexité du modèle [Bigi et al., 2001d]. Cette représentation a ainsi permis de contribuer aux problématiques **PS1** et **PS2**.

J'ai par ailleurs évalué la capacité des modèles conçus à identifier un thème avec un nombre réduit de mots, c'est-à-dire très tôt dans le processus de reconnaissance de la parole, de façon à adapter le modèle de langage le plus rapidement possible et ainsi améliorer les performances de reconnaissance. J'ai montré que la représentation de documents proposée permettait de détecter le thème d'un texte significativement plus tôt que l'approche traditionnelle [Brun et al., 2000c], ce qui fait de cette représentation une représentation adéquate pour la reconnaissance automatique de la parole.

Evaluations en conditions réelles J'ai conduit des évaluations de ces contributions en conditions réelles, *i.e.* sur un système de reconnaissance automatique de la parole. Ces évaluations ont porté à la fois sur la qualité de la classification (détection de thème) ainsi que sur la qualité de la reconnaissance après adaptation du modèle de langage au thème détecté (combinaison d'un modèle de thème et d'un modèle général). Une amélioration significative de la qualité de la reconnaissance avait ainsi été constatée. J'ai pu en conclure que **la détection de thème et l'adaptation du modèle de langage en conséquence permettent d'améliorer les performances en reconnaissance, tout en ayant une complexité supplémentaire très limitée.** L'hypothèse sur laquelle reposaient mes travaux a donc été validée.

L'incertitude liée à l'humain En marge de ces travaux, je me suis intéressée à la pertinence des étiquettes de classes fournies dans les données d'apprentissage. Une étude sur un grand nombre d'utilisateurs a pu montrer que même pour un petit nombre de thèmes, l'avis d'étiqueteurs humains pouvait différer, voire grandement différer non seulement de l'étiquette fournie par un expert, mais également entre eux. En parallèle, j'ai pu mettre en corrélation la performance des classifieurs et l'homogénéité des thèmes proposés par les évaluateurs [Brun and Smaïli, 2004]. **Ces travaux, en lien direct avec l'humain, m'ont permis de me rendre compte non seulement de la variabilité, mais aussi de l'incertitude, des erreurs, etc. lorsque l'humain est associé à un système.**

2.1.3 Les modèles de langage

En parallèle de ces travaux, je me suis intéressée plus directement à la modélisation statistique du langage, dans le but d’apporter des réponses aux trois problématiques identifiées ci-dessus.

Modèles résistants au bruit (PS1) et peu complexes (PS2)

D’une façon générale, un modèle de langage exploite les mots précédents (les mots déjà reconnus dans le cas d’un système de reconnaissance de la parole), également appelé l’historique, pour évaluer la probabilité d’apparition d’un mot suite à cet historique. Les modèles n -grammes, modèles les plus populaires, exploitent un historique composé des exactement $n - 1$ derniers mots pour déterminer le mot suivant le plus probable ; ils ne sont donc pas résistants au bruit, et leur complexité intrinsèque limite la taille de l’historique considéré. J’ai travaillé sur la proposition d’un modèle qui repousse ces deux limites (problématiques **PS1** et **PS2**). Ce modèle peut ne pas considérer certains mots de l’historique, ce qui permet, à taille d’historique équivalente, soit de diminuer sa complexité, soit d’augmenter la taille de l’historique, à complexité équivalente. Les expérimentations menées ont permis de montrer sa résistance au bruit [Brun et al., 2006, Boyer and Brun, 2007b].

Le bruit et les erreurs dans les données sont une réalité que beaucoup de travaux de la littérature ne considèrent pas. Je continue à m’intéresser, à l’heure actuelle, à cette problématique. Elle est en effet toujours d’actualité car avec le Web 2.0, les très nombreuses données produites par le “grand public” contiennent énormément de bruit ; les erreurs dans les données sont désormais légion.

Modèles peu complexes (PS2), appris sur peu de données (PS3)

Les modèles de langage à base de classes, n -classes (n -grammes de classes), ont l’avantage de réduire la complexité des modèles de type n -grammes (**PS2**) et sont un moyen de pallier le manque de données d’apprentissage (**PS3**). Dans le but de contribuer à la réduction de la complexité des modèles, j’ai travaillé sur une méthode de classification automatique de mots. Le problème de la recherche de la classification optimale étant complexe, j’ai proposé d’exploiter une approche par recuit simulé, pour limiter le temps d’exécution. Elle a montré une amélioration de la qualité de la modélisation [Smaïli et al., 1999]. **Ce travail était le premier à exploiter cette approche pour former automatiquement une classification.** J’ai souhaité poursuivre ce travail sur la conception de nouvelles approches pour la classification automatique de mots. Cela a fait l’objet de l’encadrement du stage de M2 Informatique de François Campana, en 2005.

Le manque de données (PS3)

Dans le but d’améliorer la performance des modèles statistiques du langage, je me suis intéressée au problème des événements non rencontrés dans le corpus d’apprentissage. La taille du corpus d’apprentissage étant finie, par définition celui-ci ne peut contenir l’ensemble des événements (ici les n -grammes) possibles de la langue. Dans les modèles de la littérature, même si un événement n’est pas rencontré dans les données d’apprentissage, une probabilité non nulle lui est assignée. Dans mes travaux, je suis partie de l’hypothèse que certains événements ne seraient jamais rencontrés, même si la taille des données d’apprentissage était infinie, pour la simple raison que ces événements n’appartiennent pas à la langue, ce sont des événements que j’ai dénommés *événements impossibles*. J’ai proposé plusieurs méthodes d’identification de ces événements impossibles, à base d’heuristiques exploitant la fréquence des événements et à base de classes de mots. Les évaluations ont montré une amélioration significative de la qualité du modèle de langage [Brun et al., 2000a, Langlois et al., 2003]. **Ce travail est le premier de la littérature à traiter d’événements impossibles, notamment en modélisation du langage.**

2.1.4 En résumé

Durant cette période, qui s'est terminée en 2007, mes travaux ont porté sur la modélisation statistique du langage avec pour cadre applicatif la reconnaissance automatique de la parole. Mes contributions ont porté sur la proposition de modèles performants, peu complexes et robustes au bruit dans les données. Ces modèles sont à la fois des modèles de langage, des modèles de classification pour la détection de thèmes et des modèles de sélection de vocabulaire. Plusieurs de ces travaux sont originaux, soit en raison du problème traité (notion d'événements impossibles de la langue), soit par la façon dont ils ont été abordés (sélection de vocabulaire par une mesure d'information mutuelle).

Durant cette période, j'ai identifié trois problématiques scientifiques liées à la conception de modèles reposant sur des données réelles :

- **PS1 - la conception de modèles robustes au bruit et aux erreurs dans les données,**
- **PS2 - la conception de modèles de faible complexité à large couverture,**
- **PS3 - l'apprentissage de modèles sur des données disponibles en faible quantité.**

Ces problématiques sont volontairement générales et indépendantes de toute application. Cela me permettra de les aborder dans d'autres cadres, ce que nous verrons dans les sections suivantes.

C'est durant cette période que j'ai effectué mes premiers encadrements scientifiques : 2 stages de Master 2 recherche.

De cette période, j'ai par ailleurs retiré à la fois le souhait de valider les modèles sur des données réelles et en conditions réelles, de façon à mieux maîtriser les apports des différents modèles, dans des cadres variés. Pour continuer à réaliser ce souhait, la suite de mes travaux ne pourra être dissociée de collaborations avec des industriels.

2.2 Amorce d'un virage thématique : vers la modélisation utilisateur et les systèmes de recommandation

A partir de 2006, j'ai souhaité m'intéresser à d'autres cadres que la reconnaissance automatique de la parole, tout en conservant l'approche générale que j'avais adoptée durant la période précédente, à savoir l'apprentissage de modèles à partir de données de réalisation/d'observation. Bien que ces nouveaux cadres applicatifs aient leurs propres spécificités et soulèvent de nouvelles difficultés, j'ai choisi de m'intéresser, dans un premier temps, aux mêmes problématiques : la résistance des modèles au bruit (**PS1**), la complexité des modèles (**PS2**) et le manque de données d'apprentissage (**PS3**), et ainsi exploiter les connaissances acquises et les conclusions tirées de mes travaux passés. Dans un second temps, j'ai introduit de nouvelles problématiques.

Au début de cette période, j'ai initié une collaboration avec Anne Boyer, alors maître de conférences dans l'équipe MAIA du LORIA. Anne Boyer travaillait dans le domaine des systèmes de recommandation. L'objectif d'un système de recommandation est de pouvoir proposer à un utilisateur des ressources qui correspondent à ses attentes, son comportement, ses préférences, etc. Pour atteindre cet objectif, le système exploite, entre autres, le profil de l'utilisateur, qu'il a constitué grâce aux données qu'il possède ou qu'il a collectées sur ce dernier.

Lors de cette collaboration, notre problématique commune a porté sur **la modélisation et la prédiction de comportement utilisateur**, en interaction avec un système numérique, en vue de proposer des recommandations. Un exemple d'un tel système peut être un navigateur Web. L'objectif est alors de modéliser le comportement de navigation des utilisateurs, de prédire leur comportement futur, et ainsi de leur proposer des recommandations adéquates.

En 2006, les systèmes de recommandation n'en étaient qu'à leurs débuts. Peu de chercheurs, et particulièrement en France, s'intéressaient à ce domaine. Les systèmes de recommandation sont depuis lors devenus un domaine étudié par un très grand nombre de chercheurs, avec des approches très variées.

En terme de données, les traces de comportement utilisateur peuvent être vues comme une suite d'ac-

tions (par exemple la suite des ressources consultées dans le cadre de la navigation Web), tout comme j'avais considéré un document comme une suite de mots dans mes travaux précédents.

En terme d'objectif applicatif, prédire le comportement d'un utilisateur sachant son historique de comportement peut être vu comme un problème similaire à celui de la prédiction d'un mot sachant son historique de mots (objectif d'un modèle de langage).

Le problème de la modélisation et de la prédiction de comportement d'utilisateur dans un objectif de recommandation est donc proche de celui de la modélisation statistique du langage.

2.2.1 Premiers travaux : changement de cadre, mêmes modèles

Ma première proposition lors de cette collaboration a été à la fois d'étudier la capacité des modèles de type n -grammes (modèles sur lesquels j'avais travaillé durant la période précédente) à fournir une prédiction de qualité, mais aussi de proposer des modèles dérivés. Les problématiques du manque de données (**PS3**), du bruit dans les données (**PS1**) et de la complexité des modèles (**PS2**) restent présentes sur les données de comportement ; certaines sont même encore plus présentes. C'est notamment le cas du bruit qui, dans le cas de données de navigation peut correspondre aux erreurs faites par les utilisateurs : des ressources consultées qui ne les intéressent pas, des ressources cliquées par erreur, etc., erreurs que l'on retrouve très fréquemment. Les travaux que nous avons menés nous ont permis de montrer que l'utilisation de modèles n -grammes, sans adaptation à ce nouveau cadre, permettait une prédiction, et donc une recommandation, de bonne qualité [Boyer and Brun, 2007b, Boyer and Brun, 2007a]. Ces modèles ont ainsi constitué une première base intéressante pour la recommandation.

Ces premiers travaux m'ont confortée dans le choix de m'intéresser aux systèmes de recommandation et aux nombreux autres défis qu'ils posent et aux questions qu'ils soulèvent. Cette même année, j'ai travaillé sur la conception d'un système de recommandation basé sur un modèle triggers. Les modèles triggers, également utilisés en modélisation statistique du langage, permettent de modéliser le lien ordonné entre paires d'éléments. C'est sur la définition d'un tel système qu'a porté **le stage de M2R informatique d'Arnaud Villenave, que j'ai encadré** en 2007. L'avantage d'un modèle trigger est qu'il est particulièrement robuste au bruit des données (**PS1**) car il modélise des liens entre éléments distants. De plus il est peu complexe (**PS2**), car il ne modélise que des paires d'éléments. Les expériences ont montré que les performances associées à ce modèle étaient plus faibles que celles d'un modèle n -grammes, notamment en raison de la distance non contrôlée entre les éléments. Des expériences supplémentaires relatives à un modèle hybride n -grammes et triggers ont permis d'améliorer les performances.

2.2.2 Modélisation de comportement : le problème du manque de données

Anne Boyer et moi-même avons débuté le **co-encadrement de la thèse d'Ilham Esslimani** fin 2006, dans le cadre du projet PERCAL. L'objectif applicatif de cette thèse était de concevoir un système de recommandation de ressources provenant d'un Intranet documentaire.

Le projet PERCAL (2006-2009) est une convention de recherche entre l'Université Nancy 2 (laboratoire LORIA) et le Crédit Agricole S.A. Il avait pour objectif le développement de modèles de recommandations personnalisées de documents. Le Crédit Agricole S.A. avait fait le constat que les documents déposés quotidiennement sur son Intranet n'étaient pas ou peu consultés. La raison provenait du fait que trop de documents étaient déposés et que les employés ne savaient pas quels documents pourraient les intéresser. La question qui se posait alors était : quels documents devaient être recommandés, et à qui, en fonction des intérêts estimés des employés.

Le problème ici a été de traiter des séquences de comportement (navigation dans l'Intranet), **dans un cadre de manque de données (PS3)**, et de proposer un modèle de recommandation sans passer par une étape d'élicitation de préférences. Nous avons proposé plusieurs solutions originales pour pallier le manque de données, parmi lesquelles le modèle BNCF "Behavioral Network Collaborative Filtering" qui représente les utilisateurs sous la forme d'un graphe et applique une technique de filtrage collaboratif sur ce dernier [Esslimani et al., 2009b]. Le modèle D-BNCF "Densified-Behavioral Network based Collaborative Filtering" applique, quant à lui, des techniques de densification de graphes pour enrichir les données

et donc limiter le problème du manque de données [Esslimani et al., 2011], mais aussi des techniques de classification automatique dans le but d'identifier un voisinage de qualité [Esslimani et al., 2008a, Esslimani et al., 2009a].

Ilham a également initié un travail sur l'identification d'utilisateurs "leaders" [Esslimani et al., 2010a, Esslimani et al., 2013] exploités pour former un modèle léger (**PS2**). **Ce modèle est le premier modèle de la littérature cherchant à extraire des utilisateurs "leaders" avec pour but la conception d'un modèle peu complexe.** L'article [Esslimani et al., 2010a] a reçu le **3ème Best Paper Award à la conférence Advances in Social Networks Analysis (ASONAM) 2010.**

Le contexte de ce travail a été l'occasion d'aborder pour la première fois le **problème du démarrage à froid** [Esslimani et al., 2010b]. Ce problème est un cas particulier du problème du manque de données (**PS3**). Je m'y suis par ailleurs intéressée à plusieurs reprises par la suite.

Ici également, **le contexte applicatif réel a été au cœur de ces recherches.** Les algorithmes conçus par Ilham ont été pensés pour être utilisables et utilisés sur des données réelles (en l'occurrence celles du Crédit Agricole S.A.) et en conditions réelles. Ils devaient donc permettre une réponse en temps-réel. Ils ont d'ailleurs été mis en œuvre sur l'Intranet du Crédit Agricole S.A. Ces travaux ont confirmé mon souhait de chercher à valider mes propositions, lorsque cela était possible, sur des utilisateurs réels et dans des conditions réelles.

Ilham a soutenu sa thèse à l'automne 2010, elle poursuit actuellement sa carrière dans la recherche et est aujourd'hui chercheur au laboratoire LIST au Luxembourg.

2.2.3 Modélisation de comportement sur le Web : le problème de la résistance au bruit

Toujours dans le cadre de la modélisation de comportement, j'ai par la suite souhaité étudier plus en détails les modèles n -grammes et d'une façon plus générale les modèles permettant la modélisation de phénomènes satisfaisant la propriété de Markov. Mon but était de proposer des modèles de recommandation innovants, tout en garantissant une performance élevée, une résistance au bruit (**PS1**), mais également une faible complexité (**PS2**).

C'est dans ce cadre que j'ai co-encadré, avec Anne Boyer, **la thèse de Geoffray Bonnin** débutée en 2007, et dont l'objectif était de concevoir des modèles de recommandation pour la navigation Web. **Le premier défi de cette thèse a concerné la gestion du bruit dans les données (PS1).** Partant du constat que les modèles n -grammes n'étaient pas résistants au bruit, que les modèles triggers l'étaient, mais ne pouvaient pas modéliser précisément les relations entre éléments, Geoffray a proposé un modèle permettant de gérer des éléments non contigus au sein de l'historique des ressources. Ce modèle ne prend pas en compte certaines ressources de l'historique, notamment celles correspondant à du bruit. Il a montré que le modèle permettait de gérer efficacement le bruit dans les données et ainsi d'atteindre de bonnes performances en recommandation, tout en ayant la caractéristique d'être léger [Bonnin et al., 2008b]. Dans ces travaux, Geoffray s'est également intéressé à la couverture des modèles conçus : la capacité à générer des recommandations, quelle que soit la séquence de comportements considérée (**PS2**).

Dans le cadre de la navigation Web, le choix de consulter une ressource peut ne pas dépendre exclusivement des toutes dernières ressources consultées (propriété de Markov). Il peut également dépendre de certaines ressources plus lointaines, ce que les modèles classiques, pour des raisons de complexité, ne permettent pas de prendre en compte. Nous avons souhaité traiter ce problème en introduisant une nouvelle problématique **PS4 - la conception de modèles d'identification et de gestion de relations distantes entre éléments.** Geoffray a conçu un modèle permettant de gérer un historique long tout en étant peu complexe. Le modèle proposé, le modèle SBR "Skipping-Based Recommender", permet de considérer à la fois des sauts, mais en nombre restreint (pour garantir la légèreté du modèle), tout en pondérant les éléments de l'historique par rapport à leur distance à l'élément à prédire (pour garantir une recommandation de qualité) [Bonnin et al., 2009a, Brun et al., 2009a, Bonnin et al., 2010c, Bonnin et al., 2012]. **Ce travail constitue le premier travail de la littérature sur la modélisation de relations distantes entre éléments dans un objectif de recommandation.** Dans mes travaux actuels, je continue à m'intéresser à cette problématique.

Cet algorithme a pu être exploité dans le cadre du projet TechnoReco.

Le projet TechnoReco (2008) est une convention de recherche entre l'Université Nancy 2 (laboratoire LORIA, équipe KIWI) et la société Technoscope. Il avait pour objectif le développement de modèles de recommandations personnalisées d'articles journalistiques. Le défi était de permettre l'identification des ressources d'intérêt pour un journaliste qui navigue sur le site Web de Technoscope, sans qu'il exprime son opinion et sachant qu'il navigue un peu "au hasard" en recherche d'articles.

Enfin, Geoffray s'est intéressé au traitement automatique de données de comportement issues de fichiers de logs de navigateurs, données qui peuvent correspondre à plusieurs navigations imbriquées (utilisation de plusieurs onglets du navigateur, où chaque onglet est une navigation. Dans les données collectées, aucune indication de l'onglet utilisé n'est disponible). Une première approche possible est de considérer les données correspondant à une session de navigation comme du bruit pour une autre session, et inversement. Geoffray a adopté une autre approche, il a proposé un algorithme original TABAKO "Tabbing-Based All-kth-Order" permettant d'identifier automatiquement les différentes sources (onglets) composant les données, dans le but de séparer les sources, puis il a conçu un algorithme de recommandation gérant les différentes séquences identifiées. Il a ainsi pu montrer la pertinence de l'approche qu'il a proposée, à la fois en terme de qualité des recommandations produites, mais aussi en terme de résistance au bruit et de complexité du modèle résultant (**PS1** et **PS2**) [Bonnin et al., 2010a, Bonnin et al., 2010b, Bonnin et al., 2011b, Bonnin et al., 2011a]. **Ce travail a été le premier à s'intéresser à la tâche complexe de recommandation dans un environnement de navigations parallèles.**

Geoffray a soutenu sa thèse à l'automne 2010. Après avoir effectué un post-doctorat à l'Université de Dortmund (Allemagne) dans le domaine des systèmes de recommandation, il a obtenu un poste de maître de conférences à l'Université de Lorraine en 2014.

2.2.4 Les utilisateurs : modélisation et préférences

Durant cette période, j'ai commencé à m'intéresser au processus d'expression de la satisfaction (des préférences) par les utilisateurs et à leur exploitation dans les systèmes de recommandation. Ce travail a fait l'objet du **stage de M2R Sciences Cognitives de Mounir Katet, que j'ai encadré** en 2007. Mounir a étudié les différents moyens d'exprimer sa satisfaction ou ses préférences, et une étude qu'il a menée a permis d'avoir un premier aperçu des caractéristiques de chacun, à la fois du point de vue de l'utilisateur, mais aussi en vue d'une exploitation dans les systèmes de recommandation. Ce travail préliminaire s'est poursuivi en 2010 par Nicolas Jones.

Ici encore, ce sont les utilisateurs qui sont au cœur de l'étude.

Le problème de la recommandation repose en grande partie sur une étape préliminaire de modélisation des utilisateurs. Bien que de nombreux travaux aient été menés par la communauté, il est difficile d'avoir une vue claire des différents modèles existants, des points communs et des différences de chacun, de ce que chaque modèle apporte, etc. Plus particulièrement, lorsqu'un professionnel souhaite modéliser ses utilisateurs, un large panel de modèles s'offre à lui et il ne sait pas lequel lui permet d'avoir la modélisation adéquate, en fonction des caractéristiques des données dont il dispose.

J'ai souhaité travailler sur la conception d'une carte à destination à la fois des chercheurs et des professionnels, permettant d'aider au choix d'un modèle utilisateur. Ce travail a fait l'objet d'une **collaboration avec Liana Razmerita, Associate Professor à la Copenhagen Business School**, où j'ai passé un peu plus de 3 mois au printemps 2010, et qui a abouti à une publication [Brun et al., 2010a]. Grâce à l'**obtention d'un financement de l'Université Nancy 2** pour un séjour de courte durée, que j'ai à nouveau réalisé à la Copenhagen Business School, il s'est ensuite poursuivi sur la conception de modèles de groupes d'utilisateurs [Razmerita and Brun, 2010, Razmerita and Brun, 2011].

2.2.5 En résumé

Cette deuxième période de mon activité de recherche a débuté en 2006 et s'est terminée en 2010. Elle correspond à mon souhait d'évoluer thématiquement et de travailler en particulier sur les systèmes de recommandation et sur les défis qu'ils apportent.

L'approche (apprentissage à partir de données), ainsi que les problématiques que j'avais abordées jusque là : la complexité, la résistance au bruit et le manque données, ont continué à être au centre de mes préoccupations.

Le changement de cadre applicatif m'a permis de me confronter à de nouvelles spécificités : présence encore plus forte de bruit, données multi-sources imbriquées, nécessité d'enrichir les données, etc. et de proposer d'autres approches pour traiter ces problématiques : exploitation de graphes et des algorithmes associés, proposition de modèles de gestion d'historiques longs avec une faible complexité, etc. En particulier, le modèle SBR que nous avons proposé est un modèle très original non seulement dans la problématique traitée mais également dans la façon dont celle-ci est abordée. La gestion automatique de navigations imbriquées est également un problème original, de même que le modèle TABAKO qui permet de le résoudre.

J'ai par ailleurs introduit une nouvelle problématique liée à la gestion de données séquentielles :

- **PS4 - la conception de modèles d'identification et de gestion de relations distantes entre éléments.**

J'ai continué à appliquer mes modèles sur des données réelles, dans des conditions réelles, et sur des utilisateurs réels, ce qui a été rendu possible grâce aux partenariats montés et concrétisés par des contrats de recherche, notamment industriels.

Enfin, cette période a été l'occasion de renforcer mon activité d'encadrement : 2 encadrements de M2R, et 2 premiers co-encadrements de thèse dans le domaine des systèmes de recommandation, en co-direction avec Anne Boyer. J'ai également pu m'investir dans des contrats de recherche dans lesquels j'étais en première ligne : conventions de recherche avec le Crédit Agricole S.A. et avec la société Technoscope.

Enfin, au cours de cette période, j'ai choisi de participer à la **création d'une nouvelle équipe de recherche** au laboratoire LORIA, l'équipe KIWI (*Knowledge, Information and Web Intelligence*), dont j'ai été un des 3 membres fondateurs et dont la responsable est Anne Boyer. Le thème principal de l'équipe est la conception de méthodes pour la proposition de services personnalisés aux utilisateurs. Les 3 membres fondateurs provenaient tous d'équipes de recherche différentes, avec des approches différentes mais complémentaires. La création de cette équipe a donc été un défi, et a d'ailleurs été un succès, puisqu'elle a maintenant 10 ans, comporte 9 membres permanents, 2 ingénieurs et en moyenne 8 doctorants.

2.3 Les systèmes de recommandation : une nouvelle dynamique, de nouvelles problématiques

La troisième et plus récente période de mon activité a débuté en 2010. Elle fait suite à l'évolution thématique que j'ai initiée dans la période précédente, et confirme ma volonté de faire des systèmes de recommandation le cadre applicatif principal de mes recherches, tout en conservant l'approche que j'ai adoptée dans les périodes précédentes.

Dans cette nouvelle période, j'ai souhaité me confronter à de nouveaux aspects des systèmes de recommandation. Au début de celle-ci, la popularité des systèmes de recommandation avait significativement augmenté et un nombre croissant de travaux étaient menés. J'ai souhaité travailler sur des aspects originaux relatifs à ces systèmes : les modalités d'expression des préférences, les préférences atypiques, les relations et l'influence au sein des données de comportements, récemment l'explication des recommandations faites aux utilisateurs, etc. Je présente ci-dessous les travaux que j'ai menés, les contributions associées, dans quelle mesure ils ont contribué aux problématiques définies précédemment, ainsi que les nouvelles problématiques qu'ils m'ont permis d'introduire.

Plusieurs de ces travaux ont été menés dans le cadre de projets recherche ou industriels, et ont fait ou font l'objet de plusieurs co-encadrements de thèses, dont certains réalisés dans le cadre de ces projets.

Pour des raisons de cohérence et de compréhension des travaux menés, ceux-ci ne sont pas systématiquement présentés par ordre chronologique.

2.3.1 La gestion des préférences utilisateurs

Jusqu’au début de cette période, j’avais presque exclusivement travaillé sur des données de comportement utilisateur. J’ai souhaité m’intéresser ici aux données de préférences, c’est-à-dire des données qui représentent l’intérêt d’utilisateurs pour des ressources. Ce type de données était, d’ailleurs, celui qui était le plus souvent exploité dans la littérature des systèmes de recommandation. Sur ces données, l’objectif est d’inférer des préférences inconnues à partir des préférences connues, pour ensuite recommander les ressources associées aux plus hautes préférences estimées, à des utilisateurs. Contrairement aux travaux portant sur le comportement utilisateur, l’ordre dans les données de préférences n’a généralement pas d’importance, c’est la valeur de la préférence qui est au cœur du problème. J’ai choisi d’aborder la tâche d’inférence des préférences (puis de recommandation), avec une approche par filtrage collaboratif.

La très grande majorité des travaux de la littérature portait sur la proposition d’algorithmes et d’approches pour l’inférence de préférences, avec pour unique objectif l’amélioration de la qualité globale de cette inférence (et de la recommandation), c’est d’ailleurs presque toujours le cas actuellement. J’ai souhaité travailler sur des aspects autres que la pure “performance”. Je me suis en particulier intéressée aux modalités d’expression de préférences, aux utilisateurs avec des préférences différentes de celles des autres et à la recommandation pour des groupes d’utilisateurs.

Une nouvelle modalité d’expression de préférences

Les préférences exploitées par les systèmes de recommandation peuvent se présenter sous la forme de préférences absolues, souvent des valeurs numériques, sur une échelle définie *a priori* (une note/étoile entre 1 et n , un “j’aime”/“je n’aime pas”, etc.). Ces préférences sont en général fournies directement et explicitement par les utilisateurs. En 2010, les préférences “absolues” étaient quasiment la seule forme de préférences exploitée par les systèmes de recommandation. J’ai souhaité étudier plus en détails ce type de préférences et proposer de nouvelles modalités de préférences dans le but de les comparer, à la fois du point de vue de la qualité des recommandations fournies mais aussi du point de vue de l’utilisateur : facilité d’expression des préférences, consistance, etc. Ce travail fait suite au travail amorcé par Mounir Katet lors de son stage de M2R. J’ai introduit une nouvelle problématique **PS5 - la proposition de nouvelles modalités d’expression de préférences**.

Je souhaite préciser ici que les données de préférences utilisateurs peuvent être exploitées par des systèmes autres que les systèmes de recommandation : les systèmes de personnalisation, les services souhaitant avoir un retour sur la popularité de leurs ressources/services/pages, etc. De plus, les travaux que j’ai menés peuvent également être exploités dans d’autres cadres.

J’ai proposé une modalité d’expression des préférences qui permet à l’utilisateur d’exprimer ses préférences sous la forme de préférences relatives. Il exprime sa préférence entre deux ressources : “je préfère la ressource A à la ressource B ”, “je préfère la ressource B à la ressource A ” ou “j’ai apprécié les deux ressources de la même façon”. Je me suis dans un premier temps intéressée à la conception des algorithmes de traitement de ces données : formation du profil utilisateur, recherche de voisinage, puis génération de recommandations. Ce premier objectif a fait l’objet d’une **collaboration avec Olivier Buffet, chargé de recherche Inria** et de l’encadrement du Master 2 de **Ahmad Hamad** en 2010, qui a travaillé sur le **premier modèle de recommandation de la littérature exploitant des préférences relatives**. Nous avons proposé de représenter les profils utilisateurs sous la forme d’une relation de préférences (ordre partiel) et nous avons montré qu’il était possible de raisonner sur de telles représentations et de générer des recommandations dont la qualité était légèrement supérieure à celle obtenue avec des préférences absolues [Brun et al., 2010b, Brun et al., 2010d].

J’ai, dans un second temps, étudié cette nouvelle modalité du point de vue des utilisateurs, car l’expression des préférences sous la forme de préférences relatives implique un changement radical pour ces derniers. J’ai souhaité m’intéresser non seulement à la capacité des utilisateurs à exprimer de telles préférences, à la facilité et à leur envie de les exprimer sous cette forme, mais aussi à l’évaluation *online*

des recommandations générées. Ce travail a fait l'objet **du post-doctorat de Nicolas Jones que j'ai encadré** en 2011, qui, pour répondre à ces questions, a mené une large étude sur un ensemble d'utilisateurs. Les conclusions ont confirmé le double intérêt de manipuler de telles préférences : la facilité d'expression des préférences et l'envie qu'ont les utilisateurs d'adopter cette nouvelle modalité mais aussi la plus grande stabilité des préférences et l'augmentation de la qualité des recommandations fournies [Jones et al., 2011c, Jones et al., 2011a, Jones et al., 2011d, Jones et al., 2011b].

Les travaux menés ici constituent une étude complète et originale allant de l'étude de la capacité et du souhait des utilisateurs à exprimer leurs préférences sous une autre forme, à l'évaluation de la qualité des recommandations, en situation réelle, en passant par la conception d'un algorithme de recommandation gérant ce nouveau type de données. Ils font par ailleurs partie des **premiers travaux de la littérature s'intéressant l'expression de préférences relatives**. Une fois encore, mener des études sur des utilisateurs réels m'a permis de tirer des conclusions plus fortes que ne l'auraient permises des évaluations sur des données de l'état de l'art.

Ces travaux n'ont pour le moment pas été poursuivis. Cependant, la problématique de l'expression des préférences est une problématique qui continue à m'intéresser, en particulier dans les travaux que j'ai initiés récemment dans le cadre de l'éducation.

Les utilisateurs avec des préférences atypiques

L'approche de recommandation par filtrage collaboratif repose sur l'hypothèse que les préférences des utilisateurs sont cohérentes entre utilisateurs ou tout du moins au sein d'un groupe d'utilisateurs. C'est donc en exploitant les préférences d'un groupe d'utilisateurs que le système peut proposer des recommandations de qualité à un utilisateur donné. Cependant, certains utilisateurs reçoivent peu, voire jamais, de recommandations de qualité. Une des raisons bien connue de cet échec provient du manque d'informations dont le système dispose sur certains de ces utilisateurs (problème généralement appelé problème de démarrage à froid) (**PS3**). Cependant, j'ai pu constater que le système échoue également sur des utilisateurs qui ne sont pas en situation de démarrage à froid.

J'ai fait l'hypothèse que pour certains utilisateurs, cette mauvaise qualité pouvait provenir du fait que leurs préférences ne sont pas cohérentes avec celles des autres utilisateurs, elles sont "différentes". L'hypothèse sur laquelle repose le filtrage collaboratif n'est donc pas respectée dans leur cas. J'ai souhaité travailler sur ces utilisateurs et la problématique scientifique que j'ai introduite est : **PS6 - la conception de modèles d'identification et de modélisation de données "différentes"**. Dans le cas de données de préférences, les "données différentes" sont des "utilisateurs singuliers". Les questions sous-jacentes à cette problématique sont : qu'est-ce qu'être différent, comment définir la différence, comment tenir compte de cette différence pour inférer des préférences, etc. Rares sont les travaux de la littérature qui s'intéressent à l'identification des utilisateurs singuliers et aucun ne s'est explicitement intéressé à la proposition d'approches permettant de leur proposer des recommandations de qualité. Pourtant, la présence de tels utilisateurs est évidente et, pour moi, il est très important de pouvoir leur fournir des recommandations de qualité.

Ce travail a été mené par **Benjamin Gras, durant son stage de M2 R Sciences Cognitives, que j'ai encadré** en 2014. Il est actuellement poursuivi, toujours par **Benjamin Gras, dans sa thèse débutée en 2015, que je co-encadre**. Avec pour objectif la recommandation, Benjamin s'intéresse à la fois à l'identification automatique des utilisateurs différents, qu'il appelle "moutons gris" et à la conception de méthodes permettant de leur fournir des recommandations de qualité.

Benjamin fait le choix de considérer ces utilisateurs comme des données aberrantes (*outliers*), et de s'inspirer des travaux du domaine de l'*outlier detection*. La problématique de la détection d'outliers a été largement traitée dans l'état de l'art. Cependant, la caractéristique des données de préférences (manque de données, incertitude des données) ne permet pas d'exploiter directement ces travaux. Par ailleurs, dans la majorité des travaux de la littérature, une fois que de telles données sont identifiées, elles sont simplement écartées du corpus de données. L'objectif ici est de fournir aux utilisateurs identifiés des recommandations de qualité, il n'est donc pas envisageable de les écarter des données. Le défi est donc double. **L'identification et la modélisation des préférences des utilisateurs "moutons gris" est donc une problématique nouvelle dans le domaine.**

Benjamin a abordé la tâche d'identification des "moutons gris" de façon originale, en la réalisant en amont de toute recommandation. Un premier travail a exploité une approche statistique [Gras et al., 2015b], cet **article a été présélectionné pour une nomination pour les Best Paper Awards à la conférence Webist 2015**. Un second travail a exploité une approche probabiliste [Gras et al., 2016], cet article a été **nominé "outstanding paper" à la conférence UMAP 2016**, une des plus grandes conférences du domaine de la modélisation utilisateur. Les résultats obtenus ont montré qu'il est possible d'identifier de façon fiable des utilisateurs qui recevront des recommandations de mauvaise qualité, avant même de leur proposer des recommandations. Benjamin Gras travaille actuellement à la conception d'algorithmes de modélisation de ces utilisateurs et de recommandation [Gras et al., 2017]. La thèse de Benjamin Gras est financée sur le projet PremierSuiveur. Benjamin soutiendra sa thèse en janvier 2018.

Le projet PremierSuiveur (2015-2018), dont je suis responsable, est un projet financé par le Grand Nancy, mené en collaboration avec l'entreprise Nancéenne Sailendra. L'entreprise Sailendra est une entreprise de e-commerce et l'objectif visé de ce projet est de fournir un service de qualité (recommandations satisfaisantes) à chaque client, quelque soit son profil, quelles que soient ses particularités. Le défi ici provient de l'aspect "chaque client".

La recommandation pour des groupes

Effectuer une ou des recommandations à un utilisateur revient à identifier les ressources qui correspondent au mieux à son profil. Parfois, les recommandations à fournir ne sont pas destinées à un utilisateur, mais à un groupe d'utilisateurs. C'est par exemple le cas lorsqu'un groupe de personnes (amis, famille) souhaite regarder un film (cinéma, télévision), ou partir en vacances.

Le défi dans ce cadre ne porte plus uniquement sur l'algorithme de recommandation, mais également sur l'identification et la prise en compte des caractéristiques du groupe, qui doivent impacter les algorithmes de recommandation. En effet, la composition du groupe (présence d'une personne influente, d'enfants, nombre de personnes, etc.) doit être prise en compte par l'algorithme. Une étude des différents types de groupes, des différentes stratégies de formation du profil de groupe et des stratégies de recommandation a été réalisée. Des premières expériences de recommandation ont été menées [Bernier et al., 2010], permettant de montrer l'adéquation de certaines stratégies dans certaines configurations de groupes. Ce **papier a d'ailleurs reçu un Best Paper Award à la conférence SIIE 2010**.

Cet aspect de la recommandation est une thématique qui est à nouveau présente dans les travaux que je mène actuellement dans le cadre de l'éducation.

2.3.2 La modélisation de relations au sein de données de comportement (PS4)

Les relations distantes

Les approches classiques de la modélisation et de la prédiction de comportement ont pour but de modéliser des liens entre éléments/actions de comportements, et/ou de prédire/recommander des actions (comportements). C'est d'ailleurs cet aspect qui a fait l'objet des premiers travaux que j'ai menés sur les systèmes de recommandation. Ces liens et prédictions ont la caractéristique de concerner une fenêtre de temps de taille réduite : liens peu distants et prédiction à un horizon proche. Dans de nombreux cadres applicatifs cependant, une recommandation ou une prédiction à un horizon proche est inutile. Prenons par exemple le cas d'une usine de production, pour laquelle les données disponibles sont une séquence d'éléments relatifs à la production. Prédire l'échec de la production d'un produit ou d'une série de produits peu de temps avant que celle-ci soit terminée, est inutile. Au contraire, si l'échec avait pu être prédit en avance, l'usine aurait pu soit stopper la production pour la relancer par la suite, soit prévoir une intervention de façon à empêcher cet échec. C'est le défi sur lequel j'ai souhaité travailler et la problématique scientifique qui m'intéresse ici est à nouveau **PS4 - la conception de modèles d'identification et de gestion de relations distantes entre éléments**. Cependant, le défi principal provient ici du fait que la distance (l'horizon) est beaucoup plus grande que celle considérée dans mes travaux passés, et dans la littérature en général.

Pour répondre à ce défi, j’ai souhaité m’orienter plus fortement vers une approche fouille de données. La modélisation de relations distantes avec une approche par fouille de données est un problème qui a rarement été abordé dans la littérature. J’ai choisi de travailler sur une séquence de données unique pour permettre la manipulation de grandes fenêtres de données : une séquence dite d’événements. C’est la fouille de règles d’épisodes qui répond au mieux à la problématique de la prédiction et de la recommandation dans une séquence d’événements. La conséquence de la règle représente l’élément prédit ou recommandé alors que l’antécédent représente l’historique. Ce défi a été abordé par **Lina Fahed, lors de sa thèse débutée en décembre 2012, soutenue à l’automne 2016 et que j’ai co-encadrée**. Depuis décembre 2016, Lina Fahed est post-doctorante à l’Institut Mines Telecom Atlantique (Brest). La thèse de Lina a été financée par le projet ARMURES.

Le projet ARMURES (2002-2015) (*Applications de Recherche et de Modélisation d’Utilisateurs dans les Réseaux Sociaux*), **projet dont j’ai été responsable**, est une convention de recherche entre l’Université de Lorraine (équipe KIWI, laboratoire LORIA) et le Crédit Agricole S.A. Ce projet a également été mené avec la filiale Crédit Agricole Consumer Finance. Il avait pour objectif l’analyse automatique des réseaux sociaux et de blogs dans un but de détection automatique de critères discriminants entre prospects, et d’événements déclencheurs ou influenceurs de caractéristiques clients.

Pour permettre la modélisation de relations distantes, Lina a proposé de fouiller des motifs et des règles en imposant une contrainte de distance (ou de temps) entre les éléments. Plus précisément, elle a proposé une contrainte de distance minimale entre l’antécédent et la conséquence. Cette contrainte permet, lorsque les règles sont utilisées, de garantir un temps d’au minimum une valeur fixée entre le moment où la conséquence est prédite et la réalisation effective de celle-ci. **Cette contrainte de distance minimale entre l’antécédent et la conséquence n’a jamais été proposée dans la littérature**, elle est pourtant particulièrement utile au niveau applicatif, pour disposer de temps pour réagir : annuler l’occurrence de la conséquence, l’avancer, la retarder, etc. Lina a proposé un algorithme novateur qui a la caractéristique de former les règles en déterminant la conséquence tôt dans le processus de fouille, à l’opposé des approches de l’état de l’art qui déterminent la conséquence en dernier lieu. **Déterminer la conséquence tôt dans le processus de formation des règles est une approche originale** qui a été expérimentée sur des données réelles issues de blogs (données bruitées), et qui a montré un gain conséquent en temps de fouille, en raison d’un élagage naturel dû à la connaissance de la conséquence [Fahed et al., 2014b, Fahed et al., 2014a, Fahed et al., 2015b, Fahed et al., 2018].

Les relations d’influence

Toujours dans le cadre de la modélisation de relations dans les séquences d’événements, Lina Fahed s’est intéressée à l’identification automatique d’événements dits “perturbateurs”, c’est-à-dire des événements dont la présence semble avoir un effet sur d’autres événements, des événements futurs. Dans un souci de découvrir ces événements avec un algorithme de complexité minimale, Lina a choisi de faire reposer ce travail sur l’algorithme qu’elle avait conçu auparavant. Ainsi, connaissant la conséquence d’une règle, elle a pu évaluer l’impact, sur cette dernière, de la présence d’un événement supplémentaire dans l’antécédent, lors de la formation de la règle. Cet événement peut ainsi être considéré comme un événement perturbateur. Les expérimentations menées ont montré à la fois l’existence de tels événements, la pertinence de l’algorithme en temps d’exécution, et les différents impacts de ces événements sur la conséquence : l’augmentation de sa probabilité ou au contraire sa diminution, l’éloignement de cette dernière ou encore son rapprochement [Fahed et al., 2015a].

D’un point de vue applicatif, ces événements perturbateurs correspondent à des événements à recommander (ou à introduire dans les données) pour impacter l’occurrence d’autres événements (la conséquence des règles). **L’identification d’événements perturbateurs est un défi innovant à la fois en fouille de données mais aussi dans les systèmes de recommandation**.

Les algorithmes que Lina a proposés ont été validés par le Crédit Agricole S.A. et sont mis en place dans leur système. Par ailleurs, les données d’expérimentations étaient, une fois encore, des données réelles.

Avec un objectif proche, j'ai à nouveau souhaité m'intéresser à l'identification automatique d'événements perturbateurs, non plus dans un cadre séquentiel, mais dans un cadre transactionnel. Ce défi a été traité par **Marharyta Aleksandrova, lors de sa thèse démarrée en décembre 2013, soutenue en juillet 2017, effectuée en co-tutelle avec l'Ukraine et que j'ai co-encadrée**. L'approche que Marharyta a adoptée est une approche par *contrast mining*. Marharyta a proposé de former des motifs appelés "ensemble de règles de contraste". Un changement d'état de certains attributs (les attributs source) de cet ensemble impacte la valeur d'autres attributs (les attributs cible). L'algorithme proposé [Aleksandrova et al., 2016a, Aleksandrova et al., 2016b] est original à double titre : non seulement l'identification automatique d'attributs source est un concept nouveau dans la littérature, mais le motif proposé "ensemble de règles de contraste" est également totalement novateur. Dans le cadre de la recommandation, la recommandation porte sur les attributs sources.

Dans les deux travaux présentés ci-dessus, j'ai parlé d'événements perturbateurs et d'impact/influence entre attributs. Cependant, c'est un abus de langage car aucune évaluation réelle n'a été faite, seules des évaluations *offline* ont été effectuées, la perturbation ou l'influence effectives n'ont donc pas pu être prouvées.

2.3.3 Le problème du manque de données (PS3)

La très populaire approche de recommandation par factorisation de matrices, qui forme automatiquement un espace latent, de dimension réduite, dans lequel les utilisateurs et les ressources sont représentés, a l'avantage de fournir des recommandations de qualité tout en proposant une représentation des données en faible dimension. Elle est une approche qui répond naturellement à **PS2**. Cependant, elle souffre du problème du traditionnel démarrage à froid (problème particulier du manque de données (**PS3**)) rencontré par toutes les approches collaboratives.

Je fais l'hypothèse que si l'espace latent peut être interprété, alors cette interprétation peut être exploitée pour répondre au problème du démarrage à froid. Le défi que je me suis fixé est donc l'interprétation automatique cet espace latent. Ce problème a été étudié par Marharyta Aleksandrova lors de sa thèse. Elle a proposé d'interpréter les facteurs comme des utilisateurs du système, qu'elle a nommés "utilisateurs représentatifs", et d'exploiter ces utilisateurs à l'arrivée de nouveaux items dans le système. Ainsi, pour chaque nouvel item, il est demandé aux utilisateurs représentatifs de donner leurs préférences. Partant de ces préférences, celles de l'ensemble des utilisateurs du système sur ces nouveaux items peuvent être estimées. Pour permettre une telle interprétation, Marharyta a fait l'hypothèse qu'un facteur latent pouvait être interprété comme un utilisateur donné, si la représentation de ce dernier, dans l'espace latent, était proche d'un vecteur canonique.

Cette approche a l'avantage d'être complètement automatique, elle ne requiert ni intervention humaine, ni connaissance extérieure, ce qui est totalement nouveau dans la littérature. De plus, cette approche a montré sa capacité à répondre effectivement au problème du démarrage à froid [Brun et al., 2014, Aleksandrova et al., 2017a]. Par ailleurs, c'est à notre connaissance le seul travail de la littérature interprétant les facteurs comme des utilisateurs. Marharyta Aleksandrova débute un séjour post-doctoral à l'Université du Luxembourg à l'automne 2017, pour une durée de 2 ans.

Toujours dans le cadre du problème du démarrage à froid, j'ai souhaité adopter une autre approche, celle qui vise à injecter des préférences manquantes dans les données, pour ainsi limiter le problème du démarrage à froid. C'est une approche que j'ai adoptée par deux fois.

Tout d'abord, j'ai choisi de m'intéresser à l'exploitation d'informations supplémentaires, provenant, par exemple, d'une autre source. Un premier travail a été réalisé par **Emilien Perrin, lors de son stage de M2 Sciences Cognitives, que j'ai encadré** en 2011. L'hypothèse sous-jacente à ce travail est que les préférences d'un utilisateur restent les mêmes entre domaines (ou applications/services), ou sont en tous cas corrélées. Ainsi, si l'on dispose de préférences d'un utilisateur dans un domaine (ou dans une application), elles peuvent être exploitées pour inférer (et injecter) des préférences dans le domaine (application) d'intérêt et pour lequel on dispose de peu de préférences pour cet utilisateur. Emilien a proposé une première technique de transfert d'informations entre domaines. Cette technique a montré son intérêt en permettant d'améliorer la qualité des recommandations pour les utilisateurs en situation

de démarrage à froid [Perrin et al., 2012].

Très récemment, je me suis intéressée à l’injection de données, non pas en exploitant des données externes, mais en exploitant des caractéristiques identifiées dans les données disponibles (par exemple la “densité” des données). C’est un des objectifs de la thèse d’**Oleksandr Palchenko, débutée en juin 2016, en co-tutelle avec l’Ukraine et que je co-encadre**. Oleksandr a conçu différentes stratégies de complétion de matrices. Les expériences menées ont montré une amélioration très significative des performances sur les utilisateurs en situation de démarrage à froid [Palchenko et al., 2017]. Pour des raisons personnelles, Oleksandr a souhaité interrompre sa thèse au bout d’une année.

Dans les travaux que j’ai menés durant ma première période d’activité en modélisation statistique du langage, j’ai travaillé sur des modèles de classe car ils permettaient de limiter le problème du manque de données. J’ai souhaité travailler à nouveau sur ce type de modèles dans le cadre de la recommandation, mais ici ce sont des classes de ressources que j’ai cherché à former, et le but est d’améliorer la qualité de la recommandation. Ce travail a été réalisé par **Brahim Batouche, lors de son post-doctorat, que j’ai encadré** en 2014. Brahim a proposé un algorithme de classification de ressources pédagogiques et l’algorithme de recommandation associé. Cet algorithme permet d’identifier et de gérer les relations entre classes, utilisées pour pallier le manque de données. Ces travaux ont été menés dans le cadre du projet PERICLES.

Le projet PERICLES (2012-2014) (*Projet pour l’Evaluation et la Recherche Informatisée autour des Compétences dans L’Enseignement Supérieur*) est un projet de l’appel PIA2. C’est le premier projet traitant des Learning Analytics en France. Parmi les objectifs de ce projet, la tâche 3 vise à la formulation de recommandations de formations ou de parcours de formation à partir des traces d’apprentissage et des compétences acquises. C’est dans cette tâche que je me suis investie, notamment sur la recommandation de ressources.

Les expérimentations conduites sur des données réelles d’UNT (Université Numérique Thématique) ont montré une amélioration significative de la qualité des recommandations dans un cadre de manque de données [Batouche et al., 2014b, Batouche et al., 2014a].

Récemment, j’ai abordé un autre aspect du manque de données : l’émergence.

Les travaux que je mène ont pour objectif de déterminer l’émergence d’un phénomène avant son émergence effective. Le défi auquel j’ai souhaité faire face est donc : parmi l’ensemble des événements ou motifs apparus peu fréquemment, lesquels vont émerger ? J’ai traité cet aspect à deux reprises. Avec Lina Fahed, nous avons choisi d’exploiter une heuristique, reposant sur la similarité entre des motifs connus et fréquents, et les motifs candidats à l’émergence. L’évaluation expérimentale a permis de montrer que l’utilisation d’une telle heuristique permettait d’identifier, avec une bonne précision, des motifs émergents avant leur émergence effective.

La détection d’émergence est également un des objectifs de **la thèse de Yacine Abboud, débutée en octobre 2015, que je co-encadre**. Le cadre applicatif est cependant différent. Yacine Abboud travaille sur la détection de compétences ou d’activités émergentes dans des corpus d’offres d’emploi. Les premiers travaux menés par Yacine se sont focalisés sur la conception d’un algorithme d’identification d’activités [Abboud et al., 2015]. Plus récemment, il a conçu un algorithme d’identification de motifs contraints, ces contraintes permettant de réduire le temps requis pour la fouille [Abboud et al., 2017]. Il travaille actuellement sur l’aspect émergence de motifs.

Yacine Abboud est mis à disposition pour 3 ans par son entreprise, spécialisée dans les actions de formation et les bilans de compétences dans les entreprises. L’interaction avec l’entreprise de Yacine est très forte, et l’accès à des experts et à des données réelles permettent de mener à bien le projet.

2.3.4 La conception de modèles légers (PS2)

Les techniques de recommandation à base de voisinage sont réputées pour fournir des recommandations de qualité. Cependant, elles ont l’inconvénient de ne pas passer à l’échelle car elles sont particulièrement complexes, notamment en raison de l’étape de sélection de voisins. Pour réduire cette complexité,

j’ai proposé une approche qui identifie les voisins parmi un sous-ensemble réduit d’utilisateurs, que j’ai appelés des “voisins globaux”. Evidemment, les utilisateurs de ce sous-ensemble doivent être choisis de façon à ne pas diminuer la qualité du modèle, ou en tous cas le moins possible. Le défi ici est donc d’identifier le plus petit sous-ensemble de voisins globaux qui diminue le moins possible la qualité de recommandation. J’ai choisi d’aborder ce problème comme un problème d’optimisation. Pour réaliser ce défi, j’ai **collaboré avec Amine Boumaza, alors chercheur au laboratoire LISIC² de l’Université du Littoral Côte d’Opale** et spécialiste des approches évolutionnaires. **J’ai passé un semestre à l’Université du littoral, au laboratoire LISIC**, grâce à une convention d’échange d’un demi-service d’enseignement entre l’Université de Lorraine et l’Université du Littoral. Nous avons proposé une approche de sélection des voisins globaux par évolution artificielle. Nous avons ainsi montré que les performances en recommandation demeurent compétitives, tout en réduisant de façon très significative la taille du modèle de recommandation [Boumaza and Brun, 2012a, Boumaza and Brun, 2012b].

Avec le même objectif de concevoir des modèles légers, j’ai travaillé sur la conception d’un modèle de recommandation pour le m-commerce (commerce sur mobile). En raison des spécificités des systèmes sur lesquels les modèles sont implémentés, la problématique abordée ici est relative à la conception d’un modèle peu complexe, en mémoire et en temps requis pour fournir une recommandation. L’algorithme que j’ai ainsi conçu a pour caractéristique de fournir une recommandation *anytime*, c’est-à-dire que, quel que soit le temps imparti, l’algorithme est capable de fournir une recommandation, et plus le temps alloué est grand, meilleure est la qualité de la recommandation. **Cet algorithme est le premier algorithme de recommandation anytime** [Brun and Boyer, 2010a].

2.3.5 Des algorithmes transparents

Les travaux de la littérature ont mis en avant le fait qu’une recommandation est mieux perçue par un utilisateur si celui-ci comprend la raison de cette recommandation. Une façon pour lui de la comprendre est de recevoir des explications relatives cette recommandation. En lien avec mon intérêt constant pour les utilisateurs, j’ai souhaité travailler sur cet aspect. La problématique reliée et sur laquelle je me suis penchée récemment est donc **PS7 - la conception de modèles de recommandation transparents, permettant de justifier les recommandations**.

Les techniques de recommandation à base de voisinage permettent naturellement de justifier les recommandations fournies aux utilisateurs. En effet, elles permettent d’expliquer quels autres utilisateurs (les voisins) et quelles préférences ont été utilisés pour fournir des recommandations. Cependant cette approche ne passe pas à l’échelle. L’approche par factorisation de matrices est plus performante que l’approche par voisinage, tout en passant à l’échelle. Cependant, l’espace latent formé n’est pas interprétable et les recommandations issues de cette représentation ne le sont pas non plus. Dans l’approche que Marharyta Aleksandrova a proposée et qui permet d’interpréter les facteurs comme des utilisateurs du système, Marharyta y a vu la possibilité, pour ce nouveau modèle, de fournir des recommandations interprétables, sur le même principe que le fait l’approche par voisinage [Aleksandrova et al., 2017a]. **Ce travail est le premier travail qui permet de fournir des recommandations interprétables avec une approche par factorisation de matrices**.

Je travaille actuellement sur le problème de l’explication des recommandations dans deux cadres distincts.

Le premier cadre est l’e-éducation. Le but ici est de recommander des ressources à des apprenants, et plus particulièrement à des enfants, avec pour objectif l’amélioration de leur processus d’apprentissage et l’augmentation de leur chance de réussite. En plus de l’objectif de proposer des recommandations de qualité, un aspect primordial dans ce cadre est la capacité à expliquer des recommandations. En effet, les explications sont un moyen de motiver les apprenants et de les convaincre de choisir les recommandations faites. Ce travail, **débuté en décembre 2016 par Julie Budaher, dans le cadre de sa thèse que je co-encadre**, se déroule dans le cadre du projet METAL.

2. Laboratoire d’Informatique Signal et Image de la Côte d’Opale

Le projet METAL (2016-2020) (*Modèles Et Traces au service de l'Apprentissage des Langues*) est un projet de l'appel eFran du PIA2 dont l'objectif est de concevoir des outils de suivi individualisé destinés aux apprenants ou aux enseignants, et des technologies innovantes pour un apprentissage des langues à l'écrit et à l'oral. Ce projet est particulièrement pluridisciplinaire, rassemblant des élèves, des parents, mais également des chercheurs en éducation, en droit et en intelligence artificielle. Dans ce projet, je suis responsable de la tâche 1.2 qui vise à la conception, au développement et à l'évaluation d'un outil de suivi individualisé de et pour les apprenants. Cet outil sera vu comme un tableau de bord qui permettra aux apprenants de se situer, de se motiver et de recevoir des recommandations et ainsi d'améliorer leur processus d'apprentissage

Le second cadre est le *e-commerce*, et plus précisément la vente en ligne de vin. L'objectif est de concevoir un sommelier virtuel permettant de recommander des paniers d'achats à des clients. Dans ce cadre, non seulement la qualité des recommandations est importante, mais l'explication de ces recommandations est également primordiale, pour que le client ait confiance en les recommandations qui lui sont faites. C'est un algorithme de recommandation hybride (exploitant des données de réalisation et des données de contenu) et transparent (pour permettre d'expliquer les recommandations fournies) qui me semble le plus adéquat. Ce travail est mené par **Jeffrey Honion dans le cadre de son M2R Informatique, que j'encadre** en 2017. Ce travail rentre dans le cadre du projet "Sommelier Virtuel".

Le projet Sommelier Virtuel (2017), projet dont je suis responsable, est une convention de recherche avec l'entreprise Sommelier Particulier, située en Alsace. L'objectif de cette convention est d'étudier la pertinence et la faisabilité d'un système de recommandation dans le domaine du vin, qui permet de simuler un sommelier, c'est-à-dire de prendre en compte les souhaits, envies, habitudes, niveau de connaissance, etc. des clients, tout en exploitant l'expertise et la "patte" d'un sommelier, et qui permet d'expliquer les recommandations proposées. Ce travail fait actuellement l'objet de la soumission d'un projet à la région Grand Est dans le but d'être poursuivi.

Le travail réalisé par Jeffrey a permis la conception d'un modèle de recommandation qui permet d'exploiter des sources d'information multiples, tout en permettant de fournir des premières explications. Le modèle, en cours de validation sur des données d'achat, de description et de recommandation fournies par des sommeliers, a donné de premiers résultats prometteurs.

2.3.6 Vers un nouveau cadre : l'éducation

J'ai commencé récemment à m'intéresser à l'éducation, en tant que cadre applicatif des travaux que je mène. La toute première fois où j'ai travaillé sur l'éducation a été lors de ma collaboration avec Liana Razmerita de la Copenhagen Business School en 2010, au cours de laquelle nous nous sommes penchées sur la conception de modèles de formation de groupes d'apprenants [Razmerita and Brun, 2010, Razmerita and Brun, 2011]. En 2012, je m'y suis à nouveau intéressée au cours du projet PERICLES, où j'ai travaillé sur la recommandation de ressources pédagogiques dans un contexte de manque de données [Batouche et al., 2014b, Batouche et al., 2014a]. En 2014, c'est le projet Interlingua qui m'a permis de poursuivre des travaux dans ce domaine. Les travaux que j'ai menés m'ont permis à la fois de travailler sur les problématiques des utilisateurs (apprenants) et sur des algorithmes de recommandation hybrides [Brun et al., 2015b, Brun et al., 2015a].

Le projet Interlingua (2014-2015) est un projet financé par l'INTERREG IVA. L'objectif de ce projet international est de concevoir un service évolutif pour le problème du soutien pratique pour les élèves qui étudient dans une langue étrangère. J'ai été responsable de la tâche 1, qui visait à recueillir les retours d'expériences d'étudiants ayant effectué une partie de leurs études à l'étranger et de proposer un guide de bonnes pratiques pour la conception du service visé par le projet. J'ai également participé à la tâche 2 sur la proposition d'algorithmes de recommandation.

Depuis 2016, le projet METAL me permet à nouveau de travailler dans ce domaine.

L'éducation constitue le domaine d'application principal de mon projet de recherche. Je le détaillerai dans le chapitre 4.

2.3.7 En résumé

Cette troisième période, débutée en 2010, est celle qui a confirmé mon activité dans le domaine des systèmes de recommandation. J'ai introduit de nouvelles problématiques :

- **PS5 - la proposition de nouvelles modalités d'expression de préférences**
- **PS6 - la conception de modèles d'identification et de modélisation de données "différentes"**
- **PS7 - la conception de modèles de recommandation transparents, permettant de justifier les recommandations**

J'ai par ailleurs poursuivi des travaux sur les problématiques définies dans les périodes précédentes.

J'ai travaillé sur des défis originaux, et j'ai proposé des solutions innovantes qui ont été publiées à de nombreuses reprises. Parmi les défis originaux, j'ai travaillé sur l'identification et la modélisation des utilisateurs "différents" dans la recommandation collaborative, sur la gestion de préférences relatives, sur la modélisation de relations de grande distance et sur l'influence dans les données.

Ces problématiques et travaux ont été menés en collaboration avec d'autres chercheurs. J'ai notamment effectué deux séjours longs à la Copenhagen Business School au Danemark et au laboratoire LISIC de l'Université du Littoral, de respectivement 3,5 et 6 mois. Ils ont également été menés dans le cadre de l'encadrement de 3 M2 R, du co-encadrement de 6 thèses, dont 2 soutenues et 1 qui sera soutenue en janvier 2018, et de l'encadrement de 2 post-doctorats.

Les projets de recherche auxquels j'ai contribué activement,

- qu'ils soient industriels ou académiques,
 - internationaux (projet européen INTERREG), nationaux (projets PIA), locaux (projets d'Université ou régionaux),
 - que j'ai portés, dont j'ai été responsable de lots ou auxquels j'ai participé
- ont permis d'apporter un cadre applicatif à mes travaux, cadre réel la plupart du temps et avec des données réelles. Ces projets ont par ailleurs pour beaucoup été pluridisciplinaires.

J'ai récemment initié des travaux dans le domaine de l'e-éducation, domaine qui constitue le cadre principal des mes futurs travaux de recherche. Ces derniers feront l'objet du chapitre 4 de ce manuscrit.

2.4 Synthèse

Mon activité de recherche se situe en intelligence artificielle numérique. Les travaux que j'ai menés ont pour objectif la modélisation de phénomènes réalisés par l'humain, dans le but d'inférer ou de prédire des réalisations de ces derniers. Mes recherches visent à concevoir des modèles de ces phénomènes. Les phénomènes auxquels je me suis intéressée sont le langage naturel, les préférences utilisateur et le comportement utilisateur. J'ai choisi d'adopter une approche reposant sur des données de réalisation de ces phénomènes. Mes travaux se sont concentrés sur trois aspects :

- **Les données** sur lesquelles les modèles sont appris. Les données ont la caractéristique d'être incertaines, bruitées, volumineuses et par contraste également parcimonieuses. En effet, les données ne contiennent pas et ne peuvent pas contenir toutes les réalisations possibles du phénomène à modéliser. Cependant, l'absence d'une réalisation dans les données ne signifie pas qu'elle est impossible ou rare, simplement que les données disponibles à ce moment là ne contiennent pas cette réalisation. A l'opposé, une réalisation rare peut correspondre à du bruit ou à un phénomène en émergence. Je souhaite que ces données soient des données réelles, de façon à ce que les modèles que je conçois puissent être confrontés à leurs spécificités.

- **L'apprentissage du modèle.** Les modèles que je conçois doivent permettre de modéliser toutes les données disponibles, avec leur variabilité et leurs spécificités, tout en étant robuste au bruit, à l'incertitude et au volume de données. Ils doivent également permettre la modélisation de données non observées (complétude des modèles). Les modèles doivent aussi permettre d'exploiter le temps ou tout du moins l'ordre et la distance au sein des réalisations des phénomènes. La qualité, la complétude, la complexité et la dynamique des modèles sont des caractéristiques importantes à prendre en compte dans la conception de ces derniers.
- **L'exploitation du modèle** sur différents cas d'application. Dans la mesure du possible, les cas d'application sur lesquels j'évalue mes modèles sont des cas réels, avec des utilisateurs réels et sur des données réelles. De cette façon, je ne me contente pas uniquement de mesures théoriques d'évaluation telles que des mesures d'erreur ou de complexité, qui sont bien évidemment une première information très riche et très utile. Un retour d'utilisation est un complément précieux.

Au travers de ces travaux, je me suis intéressée à de nombreuses problématiques que j'ai introduites :

- **PS1 - la conception de modèles robustes au bruit et aux erreurs dans les données,**
- **PS2 - la conception de modèles de faible complexité à large couverture,**
- **PS3 - l'apprentissage de modèles sur des données disponibles en faible quantité,**
- **PS4 - la conception de modèles d'identification et de gestion de relations distantes entre éléments,**
- **PS5 - la proposition de nouvelles modalités d'expression de préférences,**
- **PS6 - la conception de modèles d'identification et de modélisation de données "différentes",**
- **PS7 - la conception de modèles de recommandation transparents, permettant de justifier les recommandations.**

Je me suis intéressée à plusieurs domaines d'application : à l'origine la reconnaissance de la parole, puis une évolution thématique m'a menée vers les systèmes de recommandation sur lesquels je travaille depuis maintenant dix ans et plus récemment la e-éducation.

Les travaux que j'ai menés ont permis de contribuer à ces différentes problématiques scientifiques, et de travailler sur des défis originaux et des approches novatrices.

Au fur et à mesure des années, mon activité s'est donc diversifiée :

- **sur les problématiques scientifiques et les défis.** J'ai commencé par des problématiques classiques : le manque de données, la complexité, la résistance au bruit, pour ensuite proposer des problématiques nouvelles : nouveaux types de données de préférences, relations dans les données, données différentes, données rares. J'ai cependant toujours conservé la même approche pour aborder ces problématiques,
- **sur le cadre applicatif.** J'ai commencé par traiter de la reconnaissance automatique de la parole, pour ensuite m'ouvrir au e-(m-)commerce, puis à l'éducation,
- **sur l'encadrement scientifique.** J'ai tout d'abord débuté par des encadrements M2, puis des encadrements de doctorants, dont le nombre s'est accru ces dernières années, et de post-doctorants. Ces encadrements ont porté sur des sujets qui, sur mes premiers encadrements de thèse, étaient choisis par le directeur de thèse, et qui, sur les encadrements suivants ont été co-réfléchis. J'ai maintenant mes propres problématiques de recherche. Sur l'ensemble de mon activité, j'ai encadré 8 M2R, co-encadré 8 thèses (dont 4 soutenues, 1 interrompue et 1 qui sera soutenue dans les mois qui viennent), et encadré 2 post-doctorants,
- **sur les projets.** J'ai travaillé à la fois sur des projets locaux, nationaux et internationaux. Projets dans lesquels j'ai pu avoir le rôle de participante, la responsabilité de lots ou la responsabilité du projet,
- **sur les collaborations.** Les nombreux projets dans lesquels je me suis investie m'ont permis de collaborer à la fois avec des industriels mais également avec des partenaires académiques. J'ai par ailleurs collaboré avec plusieurs chercheurs, non seulement en local, avec Olivier Buffet chercheur Inria, mais aussi en national avec Amine Boumaza alors enseignant chercheur à l'Université du Littoral et enfin à l'international avec Liana Razmerita de la Copenhagen Business School. Ces deux dernières collaborations se sont réalisées dans le cadre de séjours longs dans les laboratoires

de ces chercheurs (3,5 et 6 mois respectivement).

Dans le chapitre suivant je présente un sous-ensemble des travaux que j'ai menés. J'ai choisi de présenter quelques travaux majeurs de mon activité de recherche.

Chapitre 3

Quelques contributions majeures

Le chapitre 2 a présenté un panorama relativement complet de mon activité, en se focalisant sur l'évolution temporelle des différentes thématiques, défis et problématiques sur lesquels j'ai travaillé. J'y ai également fait mention de mes encadrements scientifiques et des investissements dans les différents projets associés à mon activité.

Dans ce chapitre, je choisis de me focaliser sur un sous-ensemble de mes travaux, que je juge importants, de par l'originalité du problème traité, l'approche adoptée, ou la prépondérance du thème dans mon activité. Les travaux présentés ici ne font référence qu'à mon activité menée sur les systèmes de recommandation.

Je présente 3 thèmes, évoqués dans les grandes lignes dans le chapitre précédent :

- les utilisateurs singuliers, qui représentent une dimension originale de mes recherches. Ils ouvrent de nouvelles perspectives à la recommandation collaborative, notamment sur les aspects qualité/complexité/couverture des modèles, et sur leur transparence. La communauté porte un intérêt croissant à ce dernier aspect.
- le problème du manque de données et le cas particulier du démarrage à froid, qui est une des limites bien connues des systèmes collaboratifs, et qu'il est impératif de considérer lorsque le modèle est exploité en situation réelle. J'ai proposé plusieurs approches originales pour répondre à ce problème.
- la prédiction dans des données séquentielles, car elle représente une partie conséquente de mon activité de recherche, que j'y aborde des problématiques originales, mais aussi parce qu'elle est fortement liée à mon projet de recherche.

Les deux premiers thèmes concernent la modélisation de préférences et le troisième porte sur la modélisation du comportement. L'ordre dans lequel les travaux sont présentés ne reflète aucune évolution temporelle, mais une cohérence thématique.

3.1 Des utilisateurs singuliers

Les systèmes de recommandation à base de filtrage collaboratif (FC), souvent appelés systèmes de recommandation collaboratifs, reposent sur l'hypothèse que des utilisateurs ayant eu des préférences ou des comportements semblables dans le passé, auront des préférences ou des comportements semblables dans le futur. Dans cette section, je m'intéresse aux systèmes reposant sur des données de préférences. Pour inférer les préférences d'un utilisateur sur des items (et ensuite lui recommander ceux pour lesquels les préférences estimées sont élevées), le système identifie :

- soit des utilisateurs qui ont eu des préférences proches : des utilisateurs voisins. C'est l'approche mémoire à base de voisinage utilisateur,
- soit des items similaires à ceux que l'utilisateur a appréciés : des items voisins. C'est l'approche mémoire à base de voisinage item,
- soit des récurrences/cohérences au sein des données, exploitées pour la formation d'un modèle. C'est l'approche modèle. Ces cohérences peuvent, par exemple, être des facteurs latents qui permettent d'explicitement les préférences des utilisateurs sur les items : c'est l'approche modèle à base de

factorisation de matrices.

Ainsi, une fois l’approche choisie (et le modèle formé dans le cas d’une approche modèle), c’est celle-ci qui est exploitée pour déterminer les items à recommander à un utilisateur, quel qu’il soit et quels qu’ils soient. De la même façon, tous les utilisateurs (et leurs préférences) sont de potentiels candidats pour être exploités dans l’inférence des préférences d’autres utilisateurs (que l’approche soit mémoire ou modèle). Nous pouvons dire que dans une approche collaborative, tous les utilisateurs ont le même “rôle” et sont considérés de la même façon. Il n’y a pas de distinction *a priori* entre les utilisateurs.

De mon point de vue cependant, certains utilisateurs ont des caractéristiques spécifiques, qui font qu’ils pourraient être considérés différemment des autres, relativement au rôle qu’ils jouent dans l’inférence des préférences des autres utilisateurs, ou dans la façon dont leurs préférences devraient être inférées. J’appelle ces utilisateurs des utilisateurs singuliers.

Dans mes travaux, je me suis intéressée à deux types d’utilisateurs singuliers : les utilisateurs dits “représentatifs” et les utilisateurs dits “moutons gris”. L’identification, la modélisation et la prise en compte de ces utilisateurs singuliers peuvent avoir un impact sur les modèles de recommandation, leur complexité, leur performance, leur transparence, etc.

Avant de détailler les travaux que j’ai menés, j’introduis ici quelques notations.

3.1.1 Notations

Soit U l’ensemble des utilisateurs du système, avec u un utilisateur particulier et a l’utilisateur actif, c’est-à-dire l’utilisateur à qui le système doit fournir des recommandations. Soit I l’ensemble des items (qui pourront être recommandés aux utilisateurs), et i un item particulier. Les utilisateurs expriment leurs préférences sur les items (les préférences peuvent également être inférées du comportement des utilisateurs [Kelly and Teevan, 2003]), notées $r_{u,i}$ (préférence de l’utilisateur u pour l’item i), elles sont également souvent appelées votes ou notes. L’ensemble des préférences exprimées par les utilisateurs est stocké dans une matrice de préférences R de dimension $|U| \times |I|$. L’observé ici est l’utilisateur.

Les préférences sont en règle générale des valeurs numériques sur une échelle de valeurs allant de 1 à V , où 1 signifie que la personne n’a pas aimé l’item et V signifie qu’elle l’a aimé. La littérature s’est penchée sur les différentes échelles possibles : granularité, moyen de visualiser (notes, étoiles, etc.), nombre de valeurs dans l’échelle, avec présence ou non d’une valeur neutre, etc. Bien que de nombreux travaux sur l’impact de l’échelle de préférences aient été menés dans le cadre des études utilisateur, on trouve relativement peu de travaux en systèmes de recommandation [Kuflik et al., 2012, Cena et al., 2017].

L’objectif d’un système de recommandation est d’identifier, pour l’utilisateur actif a , les items à lui recommander. Pour cela, le système infère les préférences $r_{a,i}^*$ pour tous les items i sur lesquels a n’a pas exprimé ses préférences.

Dans ce chapitre, j’utiliserai de façon indifférenciée les termes préférence, avis, opinion, note, vote, que je considère comme équivalents dans le cadre des systèmes de recommandation.

La qualité des recommandations fournies par un algorithme de recommandation peut être estimée en fonction des erreurs faites dans l’estimation des votes : l’écart entre $r_{a,i}^*$ et $r_{a,i}$. Les mesures associées sont les traditionnelles MAE et RMSE [Shani and Gunawardana, 2011]. Plus récemment, les mesures liées au rang des items (NDPM) dans les recommandations ont gagné en popularité, de même que les mesures de précision [Konstan and Riedl, 2012, Maksai et al., 2015].

3.1.2 Les utilisateurs dits “représentatifs”

Comme mentionné précédemment, dans une approche par filtrage collaboratif classique, les utilisateurs sont *a priori* tous considérés de la même façon : leurs préférences sont/peuvent être utilisées pour inférer celles d’autres utilisateurs. Dans l’approche mémoire, les préférences d’un utilisateur u sont exploitées si u est un des utilisateurs dont les préférences sont les plus proches de celles de l’utilisateur actif a . Dans l’approche modèle, l’importance des préférences de u dans la construction du modèle est fonction de la “cohérence” et/ou de la complémentarité de ses préférences avec celles d’autres utilisateurs, où la notion de cohérence dépend du modèle choisi.

On peut ainsi dire que dans une approche par filtrage collaboratif, les préférences de tous les utilisateurs sont potentiellement exploitées.

Dans les travaux que j’ai menés, j’ai constaté que dans l’ensemble des utilisateurs du système, certains ont des préférences telles qu’ils sont plus utiles (ou plus souvent utilisés) que d’autres dans l’inférence des préférences utilisateurs, que ce soit pour former le modèle ou les voisinages. Ces utilisateurs étant prépondérants pour l’inférence de préférences, je choisis de les appeler des “utilisateurs représentatifs” de la population. La notion de représentativité sera expliquée dans les paragraphes suivants. A mes yeux, il est important d’identifier ces utilisateurs et intéressant d’étudier de quelle façon tirer profit de la connaissance de tels utilisateurs. Mon objectif est donc double : 1) identifier les utilisateurs représentatifs, 2) les exploiter explicitement dans un cadre de recommandations collaboratives.

Je commence par discuter de la seconde partie de mon objectif en répondant à la question : quel objectif / quelle utilisation de ces utilisateurs représentatifs dans un cadre de recommandations collaboratives ?

- Un premier objectif dans l’utilisation des utilisateurs représentatifs peut être la diminution de la complexité des systèmes de recommandation. En effet, ces utilisateurs étant représentatifs de la population, exploiter uniquement leurs préférences pour inférer celles des autres utilisateurs peut être un moyen de diminuer la complexité du modèle, soit à l’apprentissage pour une approche modèle, soit à l’exploitation pour une approche mémoire.
- Un second objectif peut être leur exploitation dans le cadre de la résolution du problème de démarrage à froid côté item. Dans ce cadre, les préférences de ces utilisateurs étant représentatives de la population, ces utilisateurs peuvent être vus comme ceux dont l’opinion sur les nouveaux items doit être sollicitée, pour ensuite inférer les préférences des autres utilisateurs sur ces items. Le problème du démarrage à froid fait l’objet de la section suivante (3.2). Une utilisation similaire, mais en dehors du cadre des systèmes de recommandation, peut être le sondage d’opinions. Interroger les utilisateurs représentatifs peut permettre d’inférer l’opinion d’une population entière.
- Enfin, ces utilisateurs peuvent être ceux dont l’opinion est à suivre, de façon à connaître leur évolution et ainsi inférer l’évolution des préférences de la population entière.

Bien évidemment, chacune de ces utilisations est dépendante de la façon dont les utilisateurs représentatifs sont choisis, et de ce que signifie leur représentativité.

Dans les travaux que j’ai menés, les objectifs visés ont été principalement les deux premiers mentionnés ci-dessus, et j’ai proposé plusieurs approches de sélection des utilisateurs représentatifs, pour des approches à base de voisinage, ou des approches à base de factorisation de matrices.

L’approche de recommandation à base de voisinage

Les travaux de sélection d’utilisateurs représentatifs, que je présente ici, voient leur utilisation dans les approches de recommandation à base de voisinage.

Dans l’approche mémoire à base de voisinage utilisateur, pour chaque item i sur lequel l’utilisateur a n’a pas exprimé sa préférence, les K plus proches voisins de a (K à déterminer) ayant exprimé leur préférence sur i sont identifiés. Ensuite, la préférence $r_{a,i}^*$ est estimée en exploitant les préférences de ces K voisins. Plusieurs mesures de proximité (distance, similarité, etc.) peuvent être exploitées, de même que plusieurs techniques d’inférence des préférences [Adomavicius and Tuzhilin, 2005]. Bien que cette approche soit performante car elle permet de modéliser précisément l’utilisateur a en exploitant les préférences d’utilisateurs lui étant très proches, elle a un inconvénient majeur qui est son non passage à l’échelle. En effet, dans cette approche, chaque utilisateur a son propre voisinage, et même pire : chaque utilisateur a autant de voisinages que d’items pour lesquels son opinion doit être inférée.

Dans l’objectif de réduire la complexité de cette étape (**PS2**), j’ai proposé de limiter le nombre de voisins possibles, en les restreignant à un sous-ensemble d’utilisateurs : les utilisateurs représentatifs. L’approche correspondante est donc à base de voisinage, mais est devenue une approche modèle. **Ce travail est le premier de la littérature à exploiter la notion d’utilisateurs représentatifs pour la recommandation.** Cette approche peut être considérée comme proche d’un modèle à base de clustering [Sarwar et al., 2002], puisque l’ensemble des voisins possibles est réduit. Dans un modèle à base de clustering les voisins potentiels sont en effet les utilisateurs appartenant au même cluster que a . Dans notre cas ce seront les utilisateurs représentatifs. Dans le cas limite, si l’ensemble des utilisateurs

représentatifs correspond à l'ensemble des utilisateurs, cela revient à une approche mémoire. Avec une telle approche, il est évident que les performances en recommandation ne seront pas améliorées, elles seront même, de façon certaine, diminuées. En effet, les voisins utilisés ici ne seront pas ceux qui sont les plus proches de l'utilisateur actif, mais un ensemble d'utilisateurs avec des propriétés plus globales à l'ensemble des utilisateurs. Cependant, cette diminution peut être limitée, en fonction de la façon dont les utilisateurs représentatifs sont sélectionnés. En particulier, les utilisateurs représentatifs peuvent être complémentaires à l'utilisateur actif, complémentaires entre eux, etc.

Dans le premier travail que j'ai mené, en collaboration avec Ilham Esslimani, un utilisateur représentatif est défini comme un utilisateur dont l'opinion peut permettre de représenter celle d'un nombre conséquent d'autres utilisateurs. Ici, la représentativité est donc sa capacité à permettre de représenter les préférences d'autres utilisateurs. Pour permettre l'identification de tels utilisateurs, nous avons proposé de représenter l'ensemble des utilisateurs sous la forme d'un graphe (d'utilisateurs). La valeur associée aux arcs est la similarité entre utilisateurs. Pour répondre à la caractéristique du nombre conséquent d'utilisateurs qu'ils représentent, nous avons choisi d'exploiter la connectivité des nœuds dans le graphe. En particulier, ce sont les nœuds (utilisateurs) à forte connectivité qui nous intéressent. Nous avons choisi de redéfinir la connectivité d'un nœud comme étant le nombre de nœuds auxquels celui-ci est connecté par un arc de valeur élevée. Concrètement, cela représente le nombre d'utilisateurs dont il est fortement similaire. Pour répondre à la caractéristique de l'inférence de l'opinion d'autres utilisateurs, nous nous sommes inspirés des travaux de la littérature menés en analyse de la propagation d'influence et en identification de leaders [Valente, 1995, Domingos and Richardson, 2001, Goyal et al., 2008]. Nous avons défini la capacité de propagation d'un utilisateur représentatif (un leader) comme étant la capacité du nœud du graphe le représentant à inférer les valeurs des nœuds auxquels il est connecté dans le graphe. La fonction d'inférence a dû être définie, et nous avons proposé d'exploiter la similarité entre les nœuds, qui agit comme un facteur d'atténuation sur les préférences. En résumé, les utilisateurs représentatifs sélectionnés sont les utilisateurs qui ont le meilleur couple forte connectivité/qualité de propagation élevée.

Dans le modèle à base de voisinage correspondant, $r_{a,i}^*$ est simplement estimé en exploitant l'opinion moyenne pondérée des utilisateurs représentatifs. Le modèle obtenu est donc très simple, non seulement en raison du nombre réduit de voisins possibles, mais également dans la façon d'inférer les préférences de l'utilisateur actif, puisque les utilisateurs représentatifs ont été choisis pour leur capacité à propager leurs préférences. Il est à noter qu'ici, la préférence d'un utilisateur sur un item $r_{a,i}^*$ peut être déduite d'un seul utilisateur représentatif ou de plusieurs. Les expérimentations menées ont permis d'identifier qu'avec un nombre faible d'utilisateurs représentatifs (3% de l'ensemble des utilisateurs) 1) des recommandations de qualité peuvent être fournies, bien que de qualité plus faible qu'un modèle classique, 2) la taille du modèle associé est significativement plus faible.

Ce travail a été réalisé durant la thèse d'Ilham Esslimani, que j'ai co-encadrée, et a été publié à la fois en conférence internationale [Esslimani et al., 2010a, Esslimani et al., 2010b] et dans un chapitre de livre [Esslimani et al., 2013].

Un inconvénient lié à cette proposition est que les utilisateurs représentatifs sont sélectionnés indépendamment les uns des autres. Cela a une double conséquence : 1) certains utilisateurs représentatifs peuvent être redondants avec d'autres, et donc être inutiles. Cet inconvénient est cependant également présent dans l'approche mémoire à base de voisinage puisque ce sont les meilleurs voisins de a qui sont sélectionnés, peu importe la proximité entre ces derniers, 2) il n'y a aucune garantie que chaque utilisateur du système ait au moins un utilisateur représentatif auquel il est lié, donc aucune garantie que des recommandations puissent être faites à tous les utilisateurs. On est donc face à un problème de couverture du modèle.

Le second travail présenté ici, et mené en collaboration avec Amine Boumaza, a visé à pallier ce double inconvénient, tout en répondant à **SP2**. Les utilisateurs représentatifs sont choisis comme un ensemble, et non pas utilisateur par utilisateur. Contrairement à l'approche précédente qui exploitait une heuristique, l'identification de cet ensemble est vue ici comme un problème d'optimisation : quel est le sous-ensemble d'utilisateurs, de taille minimale (pour éviter de la redondance inutile entre utilisateurs représentatifs), qui permet de fournir des recommandations, de qualité et de couverture maximales ?

Nous avons choisi d'utiliser un algorithme de recherche stochastique, en particulier un algorithme évolutionnaire. La fonction objectif (*fitness*) que nous avons conçue prend en compte non seulement la qualité des recommandations (MAE), mais également la couverture et la taille de l'ensemble d'utilisateurs représentatifs. Un avantage supplémentaire de cette approche est qu'elle détermine automatiquement la taille adéquate de l'ensemble des utilisateurs représentatifs. Les expérimentations menées ont montré que la taille du modèle pouvait être diminuée de plus de 80%, comparé à une approche par voisinage classique, avec une perte de performance et de couverture de moins de 1%.

Ce travail a été réalisé en collaboration avec Amine Boumaza, alors enseignant-chercheur à l'Université du Littoral Côte d'Opale, spécialisé dans l'évolution artificielle. Il a été publié dans deux conférences internationales très sélectives [Boumaza and Brun, 2012b, Boumaza and Brun, 2012a].

En résumé, nous avons proposé deux approches de sélection d'utilisateurs représentatifs, dans un but d'exploitation dans une approche à base de voisinage. Non seulement nous avons montré qu'un ensemble très restreint d'utilisateurs représentatifs est suffisant pour obtenir des recommandations de qualité, s'il est choisi de façon adéquate. De plus, les performances obtenues sont comparables à celles d'une approche classique, tout en diminuant de façon très significative la complexité du modèle. Notons que, comme dans tout système à base de voisinage, cette approche permet de fournir des explications aux recommandations [Bobadilla et al., 2013].

Par ailleurs, bien que l'approche résultante est qualifiée d'approche modèle, elle permet, sous certaines conditions, de facilement tenir compte des nouvelles préférences d'utilisateurs au fur et à mesure qu'elles apparaissent. En effet, si l'on considère que les utilisateurs représentatifs sont stables, c'est-à-dire qu'ils sont représentatifs sur une certaine durée, la mise à jour du modèle est très rapide, puisque seules les valeurs de similarité avec les utilisateurs représentatifs doivent être mises à jour. Par contre, la mise à jour de l'ensemble des utilisateurs représentatifs est plus complexe, puisqu'elle nécessite de réexécuter l'algorithme d'identification.

L'approche de recommandation à base de factorisation de matrices

Je me suis plus récemment intéressée à l'identification des utilisateurs représentatifs dans la technique à base de factorisation de matrices. Cette technique est actuellement la plus performante dans l'approche par filtrage collaboratif, et la plus étudiée. En quelques mots, la technique à base de factorisation de matrices vise à identifier un espace latent, de petite taille f , dans lequel sont représentés à la fois les utilisateurs et les items, formant respectivement deux matrices V et W de taille $|U| \times f$ et $|I| \times f$. Ces matrices et cet espace sont formés de façon à ce que le produit de ces deux matrices approche le plus possible la matrice de votes d'origine $R \approx V^T W$ [Takács et al., 2008, Koren et al., 2009]. Des coefficients de régularisation sont utilisés pour éviter le sur-apprentissage. Les valeurs dans V et W représentent dans quelle mesure les utilisateurs (les items) sont liés aux facteurs latents sous-jacents. Cet espace latent, et notamment chaque dimension de cet espace, représente ce qui permet d'expliquer le lien entre les utilisateurs et les items : les préférences. Bien que très performante, cette technique a le même inconvénient que de nombreuses approches modèle, c'est-à-dire la non interprétabilité des recommandations. En effet, l'espace latent formé est l'espace optimal qui permet de retrouver au plus près la matrice R , il permet donc de fournir les meilleures recommandations. Cependant, cet espace n'est pas directement interprétable, les recommandations associées ne le sont donc pas non plus. Des travaux dans la littérature ont cherché à interpréter ces facteurs latents : [Koren et al., 2009] propose de les interpréter comme des caractéristiques des items (le genre dans le cas de films, par exemple) ; [Zhang et al., 2006] les voit comme des groupes d'intérêt, ils peuvent également être considérés comme des utilisateurs-prototype [Pessiot et al., 2006] représentant un comportement-type. Cependant, chacune de ces interprétations nécessite non seulement une intervention humaine, mais aussi potentiellement d'avoir accès à des ressources extérieures.

J'ai proposé d'interpréter les facteurs comme des utilisateurs réels, des utilisateurs du système. **Le travail associé est le premier travail de la littérature à proposer d'interpréter les facteurs comme des utilisateurs.**

Les valeurs dans la matrice V représentant le lien entre les utilisateurs et les facteurs, si les facteurs sont des utilisateurs, alors ils représentent le lien entre les utilisateurs et les utilisateurs-facteurs.

L'estimation des notes exploitant les valeurs de V , elle se fera donc en fonction des utilisateurs-facteurs. Ceux-ci peuvent donc être considérés comme des utilisateurs représentatifs. La représentativité est ici vue comme la capacité à inférer les préférences d'autres utilisateurs. **Interpréter les facteurs comme des utilisateurs du système permet, par conséquent, d'interpréter automatiquement les recommandations (PS7) issues d'une factorisation de matrices, ce qui est rare dans la littérature.**

Dans la technique sur laquelle nous avons travaillé en collaboration avec Marharyta Aleksandrova, dans le but d'interpréter les facteurs comme des utilisateurs, nous nous sommes imposées la contrainte qu'elle ne devait **nécessiter aucune intervention humaine, ni ressource extérieure**, ce qui est, à nouveau, une originalité du travail car **aucune autre approche de la littérature n'a cette caractéristique**. L'avantage que nous y voyons, en plus d'être indépendant de la disponibilité d'un expert, est la mise à jour automatique de ces interprétations.

L'hypothèse sur laquelle repose notre approche est la suivante : si un utilisateur a un vecteur dans la matrice V de la forme d'un vecteur canonique : une unique valeur non nulle et toutes les autres valeurs à 0, cela signifie qu'il est uniquement lié au facteur où la valeur est non nulle. Cela signifie donc qu'il est ce facteur. L'idée derrière cette interprétation, bien que simple, n'a jamais été proposée dans la littérature. Les travaux, menés sur plusieurs corpus de données, ont montré que non seulement il existait des utilisateurs avec une représentation dans V semblable à celle souhaitée, et que, par ailleurs, certains facteurs avaient même plusieurs utilisateurs avec une telle représentation. Cette dernière caractéristique permet donc de pouvoir varier la recommandation si plusieurs utilisateurs sont candidats. Bien évidemment, rares sont les utilisateurs dont la représentation comporte exactement une seule valeur non nulle, bien souvent c'est un vecteur proche du vecteur canonique que l'on retrouve dans V . La validation de cette proposition a été faite dans le cadre du démarrage à froid, qui fait l'objet de la section 3.2 et sera donc détaillée dans cette section.

Considérer les facteurs latents comme des utilisateurs est par ailleurs un moyen de voir les recommandations comme étant issues d'autres utilisateurs, comme le fait l'approche par voisinage. Ce travail a ainsi permis de montrer que la technique à base de factorisation de matrices et l'approche par voisinage peuvent être vues comme similaires, alors que dans la littérature elles sont souvent présentées comme deux approches concurrentes.

Ces travaux, menés dans le cadre de la thèse de Marharyta Aleksandrova que j'ai co-encadrée, ont été résumés dans un journal international [Aleksandrova et al., 2017a], un journal national ukrainien ^a [Chertov et al., 2015], et une conférence internationale [Brun et al., 2014].

^a. La thèse a été effectuée en co-tutelle avec l'Ukraine

En résumé

Dans ces travaux, je me suis fixé pour objectif d'**identifier un ensemble d'utilisateurs dits représentatifs d'une population et d'exploiter ces utilisateurs dans un système de recommandation. Cet objectif est nouveau dans la littérature des systèmes de recommandation.** Dans un système à base de voisinage, nous avons montré que l'exploitation des préférences de ces utilisateurs permettait de diminuer la complexité du modèle (**PS2**) (notamment en test pour les approches mémoire), tout en conservant une qualité de recommandation élevée, même dans le cas d'un ensemble relativement petit d'utilisateurs représentatifs.

Dans une approche à base de factorisation de matrices, qui est une approche performante et qui passe à l'échelle, de tels utilisateurs permettent d'interpréter les recommandations (**PS7**).

Une suite à ce travail sera l'intégration de la dimension temporelle à la notion d'utilisateur représentatif. Cela permettra d'identifier non seulement des leaders (d'opinion) dans le sens où lorsqu'un de ces leaders a une opinion/préférence/fait une action, alors un nombre significatif d'utilisateurs aura la même opinion/préférence/ fera la même action. Cela permettra également de tracer l'évolution de l'ensemble des utilisateurs représentatifs.

On peut par ailleurs imaginer d'autres méthodes de sélection d'utilisateurs représentatifs, mais également d'autres façons de les exploiter. Par exemple, une fois les utilisateurs représentatifs identifiés, ils peuvent être considérés comme représentant la population entière et un modèle de ces utilisateurs (donc

de la population) peut être appris. Les questions soulevées sont bien évidemment : quels utilisateurs représentatifs, quels modèles, etc.

3.1.3 Les utilisateurs “moutons gris”

Un des objectifs principaux des recherches habituellement menées dans les systèmes de recommandation est l’amélioration de la qualité des recommandations. Les travaux publiés présentent systématiquement les performances des modèles en terme de performance moyenne, que ce soit en MAE, RMSE, TopN, NDPM, etc. Cependant, à performances moyennes équivalentes, deux modèles peuvent différer grandement dans la qualité effective des recommandations qu’ils produisent. Par ailleurs, il est impossible de savoir comment se répartit la qualité de la recommandation au sein de l’ensemble des utilisateurs. De même, il est impossible de savoir si une amélioration des performances est une amélioration effective pour l’ensemble des utilisateurs ou une amélioration significative pour un sous-ensemble d’utilisateurs.

Les travaux et études que j’ai menés sur les systèmes de recommandation collaboratifs m’ont permis de constater que, même lorsque les performances moyennes d’un système sont élevées, certains utilisateurs ne reçoivent pas toujours des recommandations de qualité, voire reçoivent systématiquement des recommandations de mauvaise qualité, et ce bien souvent quelle que soit la technique de recommandation utilisée.

L’approche collaborative de la recommandation repose sur l’hypothèse que les préférences des utilisateurs sont cohérentes et que, si les préférences connues d’un utilisateur a sont cohérentes avec celles d’un groupe d’autres utilisateurs (ses voisins dans l’approche par voisinage), alors les préférences inconnues de a peuvent être approchées/estimées par celles de ce groupe d’utilisateurs ; les préférences des “autres” constituent ainsi le “modèle”/la “norme” des préférences de a .

Je pense qu’une des raisons pour lesquelles certains utilisateurs ne reçoivent pas des recommandations de qualité provient du fait que l’hypothèse sur laquelle l’approche collaborative repose n’est pas vraie pour ces utilisateurs. En effet, certains utilisateurs ont des préférences différentes des autres : ni identiques, ni proches, ni corrélées. Je fais référence à ces utilisateurs comme étant des utilisateurs “moutons gris”. A mon sens, un utilisateur mouton gris n’est pas un utilisateur incohérent, ses préférences ne suivent pas la “norme” représentée par les autres utilisateurs, il est simplement différent des autres mais est tout de même modélisable.

J’ai souhaité m’intéresser non seulement à l’identification de ces utilisateurs moutons gris : ceux qui ont des préférences différentes des autres et qui reçoivent des recommandations de mauvaise qualité, mais également à la proposition d’une approche de recommandation qui leur permet de recevoir des recommandations de qualité (**PS6**).

D’un point de vue applicatif, l’objectif final est de garantir une recommandation de qualité à chaque utilisateur, quel que soit son profil. Dans un cadre de e-commerce par exemple, des recommandations de qualité pour tous est un objectif majeur. En effet, un système qui propose des mauvaises recommandations à certains de ses clients prend le risque de faire baisser ses ventes et sa réputation, car les clients déçus se tourneront probablement vers la concurrence et ne reviendront jamais.

Des travaux de la littérature se sont intéressés à l’identification des caractéristiques utilisateurs permettant d’expliquer la fluctuation de performance des systèmes de recommandation [Bellogín, 2011, Adomavicius and Zhang, 2012]. Dans les travaux que j’ai souhaité mener, mon objectif est un peu différent, puisque je cherche à identifier uniquement les utilisateurs qui recevront des recommandations de mauvaise qualité. Cette identification doit pouvoir se faire avant toute recommandation, en exploitant uniquement le profil des utilisateurs, pour ainsi permettre l’exploitation d’un modèle de recommandation adéquat.

La majeure partie du travail que j’ai mené, en collaboration avec Benjamin Gras, a visé à l’étape d’identification des moutons gris.

Nous avons proposé plusieurs définitions de la notion de préférences différentes :

- Avoir des préférences différentes, c’est être en fort désaccord avec la préférence moyenne des autres,
- Avoir des préférences différentes, c’est être en fort désaccord avec la préférence moyenne des autres sur les items consensuels (pour lesquels la majorité des utilisateurs a une opinion similaire),
- Avoir des préférences différentes, c’est ne pas être d’accord avec la majorité des autres.

Ces trois définitions ont mené à la proposition de plusieurs mesures de différence, dont certaines sont inspirées des travaux menés en *outlier detection* [Aggarwal and Yu, 2005, Han and Kamber, 2006]. Le défi principal ici a été le manque de données, leur incertitude, ainsi que le biais utilisateur. Ces mesures reposent soit sur une approche statistique exploitant le vote moyen des items et les écarts-types des votes sur les items, soit sur une approche probabiliste, reposant sur la distribution de probabilités des notes sur les items.

Les nombreuses expériences réalisées sur plusieurs corpus de données et plusieurs techniques de recommandation, ont montré qu’il est effectivement possible d’identifier, en amont de toute recommandation, des utilisateurs mouton gris : des utilisateurs dont les préférences différentes font qu’ils recevront des recommandations de mauvaise qualité dans une approche collaborative. La précision associée à certaines des mesures proposées dépasse même les 95%.

Des premiers travaux sur la modélisation de ces utilisateurs ont été menés, en proposant de nouvelles mesures de similarité. Ces travaux ont permis d’améliorer la qualité des recommandations de ces utilisateurs, même si pour le moment elle n’est pas encore comparable à celle des utilisateurs “standards”.

Ce travail a été réalisé en collaboration avec Benjamin Gras, durant son M2 que j’ai encadré puis durant sa thèse que je co-encadre et qui sera soutenue en janvier 2018. Il a mené à plusieurs publications [Gras et al., 2015b, Gras et al., 2016, Gras et al., 2017]. Notons que [Gras et al., 2016] est publié à la conférence UMAP, une des meilleures conférences dans le domaine de la modélisation utilisateur, et a reçu la distinction “outstanding paper”.

En conclusion, nous pouvons dire qu’il existe effectivement des utilisateurs qui ne sont pas en situation de démarrage à froid et qui reçoivent tout de même des recommandations de mauvaise qualité, en raison de leurs préférences différentes des autres. Par ailleurs, il est possible d’identifier de façon fiable une partie de ces utilisateurs.

Ce travail est le premier travail de la littérature s’intéressant à l’identification des moutons gris. Le terme mouton gris est souvent mentionné comme étant une des limites des systèmes de recommandation collaboratifs, mais les définitions diffèrent, incluant même les utilisateurs en situation de démarrage à froid. Par ailleurs, la majorité de ces travaux se contente d’écarter les utilisateurs identifiés, et ils tendent en général à identifier uniquement les utilisateurs extrêmement différents : des “moutons noirs”. La question qui se pose concernant ces utilisateurs est : sont-ils des utilisateurs réels ? ou peuvent-ils représenter des attaques ?

L’aspect original de notre travail réside dans le fait qu’il vise à la fois à identifier et à modéliser les utilisateurs partiellement différents, les moutons gris, qu’un système classique n’arrive pas à modéliser.

3.1.4 En résumé

Les travaux que j’ai menés sur les utilisateurs singuliers, partent du constat que dans les approches classiques de recommandations collaboratives, tous les utilisateurs sont *a priori* considérés de la même façon. Dans mes travaux, je fais l’hypothèse que certains utilisateurs ont des caractéristiques qui font qu’ils peuvent ou doivent être considérés différemment des autres dans un système de recommandation collaborative.

Je me suis intéressée à deux types d’utilisateurs en particulier. Le premier type, les utilisateurs dits représentatifs, ont des préférences qui représentent ou qui permettent de déduire les préférences des autres. Le second type, les utilisateurs dits moutons gris ont, à l’opposé, des préférences différentes des autres et sont mal modélisés (et mal recommandés) par les approches classiques de recommandation collaboratives. J’ai montré que non seulement de tels utilisateurs existent mais également que leur identification peut permettre, soit d’améliorer la qualité des recommandations (**PS6**), soit de réduire la complexité des modèles (**PS2**), ou encore d’expliquer certaines recommandations (**PS7**).

Ces travaux sont originaux dans l’état de l’art et ont fait l’objet de nombreuses publications et pour la plupart se sont faits en collaboration, dans le cadre d’encadrements de thèses.

3.2 Le problème de manque de données et un problème spécifique : le démarrage à froid (PS3)

Parmi les problèmes et les limites rencontrés par les systèmes de recommandation collaboratifs, le manque de données est un des plus couramment mentionnés. Dans les systèmes reposant sur des données de préférences, le manque de données représente le fait que, d’une manière générale, les utilisateurs votent peu d’items et que les items sont votés par peu d’utilisateurs. Cela a pour conséquence qu’il est difficile de trouver suffisamment, et de façon fiable, d’utilisateurs/items similaires, ou de régularités dans les données [Guo et al., 2014].

Un problème spécifique du manque de données est le démarrage à froid. La littérature fait état de trois catégories de démarrage à froid : “nouveau système” [Middleton et al., 2004], “nouvel item” [Saveski and Mantrach, 2014] et “nouvel utilisateur” [Lika et al., 2014]. Le démarrage à froid “nouveau système” correspond à la situation où un nouveau système est mis en place, par conséquent un tout petit nombre de données de préférences est disponible, à la fois sur les items, et sur les utilisateurs. Le démarrage à froid “nouvel utilisateur” et “nouvel item” représentent respectivement le cas où un nouvel utilisateur ou un nouvel item intègre un système existant. Dans ce cas, le système ne dispose pas de suffisamment de préférences sur cet utilisateur ou sur cet item pour garantir des recommandations de qualité (pas de calcul de voisinage possible ou pas de possibilité d’exploiter le modèle de façon adéquate). Deux configurations peuvent être identifiées dans le démarrage à froid. Le démarrage à froid complet (appelé aussi démarrage à froid pur) représente le cas où le système ne dispose d’aucune préférence sur l’utilisateur ou sur l’item. Le démarrage à froid partiel est le cas où le système dispose de quelques préférences sur l’utilisateur ou l’item, mais pas suffisamment pour permettre des recommandations de qualité [Pereira and Hruschka, 2015].

Du point de vue applicatif, la très grande majorité des systèmes de recommandation, qu’ils soient utilisés en e-commerce, e-learning, e-santé, etc. sont concernés par le problème du démarrage à froid. En effet, tout système voit régulièrement arriver de nouveaux utilisateurs ou items, cela peut même être tous les jours voire toutes les secondes. Cependant, comme mentionné dans la section 3.1.3, il est primordial qu’un système de recommandation puisse fournir des recommandations de qualité à tous ses utilisateurs, quelles que soient leurs caractéristiques, y compris les nouveaux utilisateurs. De même, il est primordial qu’un nouvel item puisse être recommandé à des utilisateurs qui l’apprécieront effectivement. Le problème du démarrage à froid est donc un des problèmes majeurs des systèmes de recommandation collaboratifs. Je souhaiterais préciser ici que même en cas d’un manque complet de données, il est toujours possible de fournir des recommandations, typiquement des recommandations tirées au hasard. Mais dans ce cas les recommandations ne sont pas personnalisées et leur qualité est loin d’être assurée. Dans mes travaux, je considère qu’une recommandation est faite (ou une préférence estimée) uniquement dans le cas où le système dispose de suffisamment de données pour effectuer une estimation. Le problème de couverture mentionné précédemment est également lié au manque de données.

La littérature a proposé deux grands types d’approches pour traiter le manque de données et le démarrage à froid : la conception d’approches spécifiques pour le manque de données, et l’utilisation/intégration de nouvelles sources de données [Son, 2016].

Dans les travaux que j’ai menés, je me suis intéressée à la fois au problème de manque de données et au démarrage à froid (**PS3**), et j’ai contribué aux deux types d’approches. Je présente ces travaux ci-dessous.

3.2.1 Conception d’approches dédiées

Les mesures de similarité

Le filtrage collaboratif repose en grande partie sur la définition de mesures de similarité. Cependant, dans le cas général du manque de données (et du démarrage à froid) les mesures classiques (coefficient de corrélation de Pearson, similarité cosin) ne sont évidemment pas exploitables telles quelles. Des métriques tenant compte du peu de données ont donc été proposées. De façon générale, ces métriques reposent sur des informations simples, des indicateurs. Par exemple, [Ahn, 2008] a proposé une mesure exploitant des notions de proximité, d’impact et de popularité ; [Bobadilla et al., 2012] a proposé d’utiliser la similarité

de Jaccard, la moyenne des différences absolues entre votes (comme dans [Ahn, 2008]), et la différence des moyennes des votes. Evidemment, ces métriques ne peuvent être utilisées que si un minimum de données est disponible. Elles peuvent par ailleurs être utilisées à la fois pour le manque de données et pour le démarrage à froid.

Le clustering

Les matrices de votes étant particulièrement creuses, des approches visant à réduire leur dimension ont été étudiées. Le clustering (et le co-clustering) est une de ces approches [Cuong et al., 2011, Wang et al., 2012], et j’ai souhaité m’y intéresser.

Je suis partie du constat que de nombreuses approches de recommandation à base de clustering n’exploitent pas le lien existant entre les clusters. Or, dans de nombreux cadres d’applications, et sur la majorité des données réelles, les clusters ne sont pas indépendants les uns des autres, les frontières étant souvent floues. En collaboration avec Brahim Batouche, nous avons fait l’hypothèse qu’exploiter ces liens pourrait être un moyen de pallier le problème du manque de données et donc d’améliorer la qualité des recommandations. La question posée a donc été de déterminer comment exploiter ce lien. Nous avons proposé une approche qui tient compte non seulement de la probabilité d’appartenance d’un item à un cluster (comme de nombreuses approches), mais également du manque d’items dans le cluster et de la proximité entre les clusters. Nous avons montré une amélioration de la qualité et de la couverture des recommandations, en exploitant des données de navigation dans des UNT et de description de ressources pédagogiques.

Ce travail a été réalisé en collaboration avec Brahim Batouche, lors de son post-doctorat dans le projet PERICLES, que j’ai encadré et a été publié dans deux conférences internationales [Batouche et al., 2014b, Batouche et al., 2014a].

La densification de données

Une autre approche possible dans le cas du manque de données est l’injection de (quelques) données (de vote par exemple) inférées à partir des données disponibles. On parle de densification de données. Une fois les données densifiées, on n’est plus dans une situation de manque de données, les techniques classiques peuvent donc être exploitées. Si les données injectées sont de qualité, alors les recommandations proposées devraient également l’être. La technique la plus simple proposée par la littérature vise à injecter des données correspondant aux votes moyens : de l’item, de l’utilisateur, ou encore le vote majoritaire sur l’item ou l’utilisateur [Candillier et al., 2007], mais les performances restent limitées. J’ai souhaité travailler sur cette approche de densification de données.

J’ai tout d’abord fait à nouveau le choix de considérer le cas où les utilisateurs sont représentés sous la forme d’un graphe, dont les nœuds sont les utilisateurs, reliés par leur similarité (de préférence ou de comportement). Ce graphe a la caractéristique d’être sous-connecté en raison du manque de données. En effet, de nombreuses similarités ne sont pas calculables car les utilisateurs ont co-voté très peu voire aucun item. Dans ce graphe, un arc inexistant ne signifie donc pas que les utilisateurs ne sont pas similaires, mais qu’il a été impossible d’évaluer cette similarité. Un arc de valeur nulle signifie, à l’opposé, qu’ils ne sont pas similaires. Par conséquent, identifier de façon précise des voisins, dans ce graphe, ne peut être fait simplement. J’ai fait l’hypothèse que si de nouveaux liens pouvaient être identifiés, et injectés dans le graphe, celui-ci serait moins sous-connecté, la recherche de voisins serait possible, et la qualité des recommandations serait améliorée. Pour découvrir ces liens, nous nous sommes inspirées, avec Ilham Esslimani, des travaux menés dans le cadre de l’analyse de graphes et notamment des réseaux sociaux. En particulier, nous avons étudié les techniques utilisées pour la prédiction de liens futurs. Dans les réseaux sociaux, la prédiction de liens est utilisée pour la recommandation d’utilisateurs, ceux liés par les liens prédits. Ces recommandations peuvent par exemple être des recommandations de co-auteurs [Barabasi et al., 2002]. Notre objectif est un peu différent, les liens découverts sont injectés dans le graphe et exploités dans une étape ultérieure : la recommandation d’items.

Pour prédire ces liens, nous avons choisi une approche topologique [Liben-Nowell and Kleinberg, 2003],

et exploité à la fois les voisinages directs et les chemins dans le graphe. Nous avons étudié plusieurs stratégies : l'attachement préférentiel, la distance entre nœuds dans le graphe, les voisins communs, le coefficient de Jaccard et la mesure Adamic/Adar. Ce sont ces deux dernières qui ont montré les meilleures performances, amenant à une amélioration significative de la qualité des recommandations.

Au moment où nous nous sommes intéressés à la densification de graphes pour la recommandation, cette approche avait été relativement peu étudiée. Depuis, la communauté s'y est intéressée plus activement [Li et al., 2014, Xie et al., 2015].

Ce travail a été réalisé en collaboration avec Ilham Esslimani, durant sa thèse que j'ai co-encadrée, et a été publié en journal international [Esslimani et al., 2011].

Récemment, je me suis à nouveau intéressée à l'approche par densification de données. Les travaux menés reposent sur un double constat : 1) même si la matrice de votes est très creuse, certaines zones de la matrice le sont moins que d'autres, 2) les utilisateurs ayant renseigné le plus de préférences sont en général ceux qui reçoivent les meilleures recommandations.

J'ai souhaité proposer une approche qui densifie la matrice de votes directement. De cette façon, les approches classiques (voisinage, modèle) pourront être exploitées. Le travail mené en collaboration avec Oleksandr Palchenko s'est inspiré de [Wibowo, 2016] qui pré-remplit un certain pourcentage de données de la matrice avec une approche traditionnelle de recommandation (en l'occurrence la factorisation de matrices). Les données à pré-remplir sont choisies aléatoirement, l'objectif de ce travail étant uniquement d'obtenir une matrice moins creuse. Nous avons proposé de ne pas choisir aléatoirement les données pré-remplies, mais celles dont la bonne qualité serait "garantie". Selon le second constat fait précédemment, ces données sont celles relatives aux utilisateurs ayant le plus grand nombre de préférences connues, ce sont donc les préférences de ces utilisateurs qui sont remplies en priorité. Par ailleurs, nous avons proposé de pré-remplir la matrice en plusieurs étapes, sans impacter la complexité d'une approche classique, chaque étape exploitant les données pré-remplies dans les étapes précédentes.

Les expérimentations effectuées ont montré que non seulement une amélioration globale significative était obtenue, mais que, par ailleurs, les utilisateurs étant le plus en situation de démarrage à froid obtenaient des recommandations de qualité presque équivalente à celle des utilisateurs ayant fourni de nombreuses préférences. **Cette approche est très originale, non seulement car elle ne nécessite aucune donnée extérieure et ne requiert aucune nouvelle métrique/technique, de plus elle n'impacte pas la complexité du modèle (PS2)** (hormis le calcul du nombre de votes d'un utilisateur). Par ailleurs, nous avons montré que la complexité de l'approche du modèle d'origine n'était pas impactée.

Ce travail a été réalisé en collaboration avec Oleksandr Palchenko, dans le cadre de sa thèse débutée en juin 2016, en co-tutelle avec l'Ukraine et que j'ai co-encadrée. Il a été publié en conférence internationale [Palchenko et al., 2017]. Pour des raisons personnelles, Oleksandr a souhaité interrompre sa thèse au bout d'une année.

L'utilisation d'utilisateurs représentatifs

Dans la section précédente, j'ai présenté plusieurs approches permettant d'identifier des utilisateurs représentatifs. Ces utilisateurs étaient utilisés soit pour réduire la complexité du modèle, soit pour permettre l'explication de recommandations.

J'ai souhaité à nouveau exploiter ces utilisateurs, mais pour résoudre le problème du démarrage à froid nouvel utilisateur. Les utilisateurs représentatifs sont ceux qui "représentent" l'avis de la population dans son ensemble, ou en tous cas ils sont ceux dont les préférences permettent d'inférer celles de chaque autre utilisateur. Par conséquent, dans le cas d'un nouvel utilisateur, nous pouvons utiliser les préférences des utilisateurs représentatifs pour inférer celles de ce nouvel utilisateur. Dans le cas où aucune donnée n'est disponible sur ce nouvel utilisateur, il est possible de lui recommander simplement les items aimés par les utilisateurs représentatifs. Dans le cas où quelques données de préférence sont connues sur cet utilisateur, une proximité/similarité entre le nouvel utilisateur et les utilisateurs représentatifs peut être estimée, et le système peut recommander les items aimés par les utilisateurs représentatifs les plus proches. Dans le cas où ce sont des données de comportement qui sont disponibles, le principe reste le même. Les

expérimentations menées ont d'ailleurs été faites sur ce type de données et ont montré la pertinence de l'approche.

Ce travail a fait l'objet d'une publication en conférence internationale [Esslimani et al., 2010b].

L'originalité de ce travail repose sur le fait que ce ne sont ni des leaders d'opinions connu/reconnus, ni des experts déclarés *a priori* qui sont utilisés pour inférer les préférences des nouveaux utilisateurs. Ce sont des utilisateurs identifiés automatiquement et qui ont la caractéristique d'être représentatifs des autres. L'identification d'utilisateurs représentatifs pour la recommandation a depuis intéressé la communauté pour résoudre le problème du démarrage à froid. Récemment, des travaux similaires ont été menés [Mazumdar et al., 2017]. Dans ce travail, les utilisateurs représentatifs sont nommés experts et sont identifiés selon leur taux d'activité. Dans le même ordre d'idées, [Chen et al., 2013] a proposé d'exploiter la confiance entre utilisateurs pour identifier les utilisateurs en qui la population a le plus confiance. Les votes de ces utilisateurs sont alors exploités pour inférer les préférences d'utilisateurs en situation de démarrage à froid.

D'autres travaux liés au problème du manque de données

Le manque de données a également été présent dans d'autres travaux que j'ai menés, même si l'objectif principal n'était pas explicitement le manque de données.

Le premier aspect que je souhaite évoquer concerne mes travaux sur les utilisateurs moutons gris (présentés section 3.1.3). Ils reposent sur le fait que les utilisateurs mouton gris, bien qu'ayant des préférences différentes des autres, ne sont pas pour autant incohérents. Ils ne suivent simplement pas la "norme" représentée par les autres, ils ont leur propre logique. Par conséquent, dans une approche collaborative, il est impossible de leur fournir des recommandations de qualité car ils n'ont pas/peu de voisins suffisamment proches, ou qu'il est impossible de trouver des cohérences entre eux et les autres. On peut donc dire que, dans ce cas, on est face à un problème de manque de données. Ce point de vue est confirmé par [Claypool et al., 1999, Terveen and Hill, 2001] qui mentionne que la notion de mouton gris n'existe que pour des communautés de petite ou de moyenne taille.

Au début des années 2000, je me suis intéressée à un problème spécifique du manque de données, dans le cadre de la modélisation statistique du langage. Dans les données de réalisation de la langue, sur lesquelles les phénomènes de la langue sont appris, certaines réalisations ne sont pas présentes. Une des raisons principales de cette absence est la taille finie des données d'apprentissage, donc le manque de données. Les approches de l'état de l'art leur assignent une probabilité non nulle en répartissant une partie de la masse de probabilités issue d'autres réalisations. Cependant, certaines réalisations ne seront jamais présentes, quelle que soit la taille des données d'apprentissage, car ce sont des réalisations impossibles dans la langue. La difficulté ici, et qui constitue son originalité, a été d'identifier, parmi toutes les réalisations manquantes dans les données, celles qui étaient des réalisations possibles mais non présentes en raison du manque de données, et celles qui étaient des réalisations tout simplement impossibles. Un autre aspect a porté sur la gestion d'erreurs dans les données. En effet, des erreurs font que des réalisations sont présentes dans les données, alors qu'elles ne correspondent pas à des événements effectivement possibles. Le défi a été de les identifier et d'en tenir compte dans la modélisation. En collaboration avec David Langlois, nous avons travaillé sur la proposition d'heuristiques, reposant sur la fréquence des réalisations et sur des classes de mots, permettant l'identification des deux types de réalisations. Ce travail, réalisé durant mon doctorat, a été publié en journal national [Langlois et al., 2003].

3.2.2 Exploitation ou intégration de nouvelles sources de données

Une approche concurrente de celle présentée ci-dessus, vise à exploiter des données externes, utilisées en complément des données existantes. La très grande majorité des travaux exploite des données externes de contenu, des données démographiques ou encore des données de préférences déclarées (obtenues au travers de questionnaires). Dans les approches à base de voisinage, les similarités calculées entre utilisateurs et/ou items sont donc des similarités de contenu, de profil démographique ou de profil déclaré [Schein et al., 2002, Vozalis and Margaritis, 2004]. Plus récemment, c'est l'utilisation de données issues de réseaux sociaux qui a été étudiée [Lin et al., 2014]. Des approches hybrides collaboratives/données externes sont également souvent proposées, et ont montré leur capacité à pallier effectivement le manque de

données [Li and Kim, 003, de Campos et al., 2010, Sahebi and Cohen, 2011]. Dans le cas du démarrage à froid complet (aucune connaissance sur l'utilisateur), l'exploitation de données externes est la seule approche proposée dans la littérature [Lenhart and Herzog, 2016].

La médiation inter-domaines

Des données de préférences peuvent également être exploitées en tant que données externes. En règle générale, ces données proviennent d'un autre domaine (ou d'un autre système). L'hypothèse sous-jacente de ces travaux est que certaines préférences restent inchangées entre domaines. Par conséquent, si des préférences d'un domaine (appelé domaine source) sont connues, alors il peut être possible de les transférer dans le domaine pour lequel il y a un manque de données (domaine cible) [Berkovsky et al., 2007].

Des travaux ont été menés à la fois dans le cas où les utilisateurs étaient communs aux deux domaines (et identifiables) et, dans ce cas, ce sont les préférences de ce dernier qui sont transférées [Pan et al., 2010], mais aussi dans le cas où les utilisateurs ne sont pas communs entre domaines, ou qu'il est impossible de faire le lien entre les utilisateurs des deux domaines. Dans ce cas, les travaux identifient un espace latent qui permet un transfert entre les deux domaines [Zhang et al., 2010]. Ces travaux appartiennent au domaine du *transfer learning* et de la médiation de modèles. J'ai souhaité m'intéresser à cette approche.

En collaboration avec Emilien Perrin, nous avons proposé d'exploiter des informations de contenu et des meta-données sur les items (des attributs) sur les deux domaines. Nous sommes partis de l'hypothèse qu'entre domaines, les préférences sur certains attributs étaient stables. La question a donc été : quels attributs ? Nous avons étudié la corrélation entre attributs communs à deux domaines et avons proposé de transférer des préférences entre attributs fortement corrélés. Ainsi, dans le domaine cible, sachant les préférences inférées sur les attributs, et ayant des informations de contenu et des meta-données sur les items, il est possible d'inférer les préférences sur les items. De la même façon, nous avons étudié les corrélations entre attributs différents entre domaines et avons proposé un transfert comme dans le cas précédent. Les expérimentations menées ont montré que les préférences ainsi inférées étaient de bonne qualité.

Ce travail a été mené en collaboration avec Emilien Perrin, lors de son stage de M2 R que j'ai encadré. Il a donné lieu à une publication en conférence nationale [Perrin et al., 2012].

De mon point de vue, cette approche peut être considérée comme une approche par densification. En effet, prenons l'exemple d'un unique domaine (une unique matrice de votes) que l'on divise en "sous-domaines", obtenant ainsi plusieurs matrices de votes. Il est alors possible d'exploiter des préférences d'un de ces sous-domaines pour inférer celles d'un autre sous-domaine. La définition de sous-domaine est évidemment à préciser, elle peut être thématique, liée aux préférences, à la densité des votes, etc. Ce dernier point de vue se rapproche des travaux que j'ai menés avec Oleksandr Palchenko.

L'utilisation d'utilisateurs représentatifs

J'ai une nouvelle fois exploité les utilisateurs représentatifs dans le cadre du démarrage à froid, mais cette fois-ci pour le démarrage à froid complet, nouvel item et en exploitant des données supplémentaires. Les préférences des utilisateurs représentatifs étant "représentatives" de celles de la population, j'ai proposé de collecter leurs opinions sur les nouveaux items (au travers de sollicitations explicites). De cette façon, leurs avis peuvent être exploités pour inférer les préférences des autres utilisateurs. La façon dont les avis sont inférés dépend de la signification de la représentativité de ces utilisateurs. Ce sont donc bien de nouvelles connaissances, externes, qui sont utilisées pour pallier le manque de données. Exploiter les avis d'utilisateurs spécifiques pour obtenir des préférences sur de nouveaux items n'est pas nouveau dans la littérature [Rashid, 2007, Amatriain et al., 2009, Liu et al., 2011]. L'originalité du travail mené ici est la façon dont les utilisateurs sont choisis (voir section 3.1.2, travaux de Marharyta Aleksandrova). Par ailleurs, il est évident que certains utilisateurs représentatifs ne pourront pas ou ne souhaiteront pas répondre aux sollicitations, ou uniquement partiellement. L'approche d'identification d'utilisateurs représentatifs que nous avons proposée a l'avantage et l'originalité d'identifier automatiquement des utilisateurs représentatifs de "remplacement", de façon à s'assurer de disposer de préférences sur les nouveaux items en nombre, et de qualité. **Elle est la seule approche de la littérature à avoir cette propriété**

et permet une utilisation dans des conditions réelles, sans sur-solliciter les utilisateurs représentatifs identifiés. Ce travail a été conduit dans le cadre de la thèse de Marharyta Aleksandrova [Aleksandrova et al., 2017a].

D'autres travaux liés au problème du manque de données

L'identification de l'émergence est une tâche à laquelle je me suis intéressée récemment. De mon point de vue, elle peut être vue comme un problème de manque de données. Pour identifier qu'un phénomène émerge, la majorité des travaux exploite simplement la fréquence d'apparition du phénomène ainsi que son évolution ; la fréquence devant être élevée, et l'évolution croissante. Je considère que l'intérêt applicatif est perdu si l'émergence d'un phénomène est identifiée alors que celui-ci a déjà émergé. J'ai donc souhaité m'intéresser à la détection d'émergence de phénomènes n'ayant pas encore émergé. Dans ce cas, la fréquence d'apparition du phénomène est encore faible et l'évolution de sa fréquence est complexe à évaluer, il est donc difficile de déterminer si le phénomène va émerger ou non. On est face à un problème de manque de données, ce qui n'est plus le cas une fois que le phénomène a émergé. L'approche que j'ai adoptée vise à exploiter des informations supplémentaires, en plus de la fréquence du phénomène. Partant de l'hypothèse qu'un phénomène similaire à un phénomène ayant déjà émergé, va probablement également émerger, j'ai choisi de tenir compte de la similarité entre phénomènes pour déterminer l'émergence. La difficulté, ici, a résidé dans la définition de la similarité entre phénomènes. La connaissance de phénomènes ayant déjà émergé peut être vue comme une source de données externes. Ce travail, réalisé durant la thèse de Lina Fahed, est un des premiers travaux de la littérature visant à identifier les futurs phénomènes émergents [Fahed, 2016].

3.2.3 En résumé

Dans les travaux que j'ai menés sur le manque de données, je me suis intéressée à plusieurs facettes du problème : j'ai travaillé sur le manque général de données, sur le démarrage à froid nouvel item et nouvel utilisateur, qu'il soit complet ou partiel.

J'ai proposé des solutions relativement à chacune des deux grandes approches de la littérature : celle exploitant des données externes dans le but d'enrichir les données, et celle qui définit des techniques adaptées au manque de données.

Dans la première grande approche, j'ai travaillé à la fois sur l'intégration des données issues d'autres domaines, mais également sur l'exploitation d'utilisateurs représentatifs qui jouent le rôle d'experts. Dans la seconde approche, j'ai proposé des techniques visant à exploiter des données fiables, soit en identifiant des utilisateurs spécifiques dont les préférences sont exploitées, soit en injectant ces données fiables pour densifier les données.

Les techniques de densification de données et d'identification d'utilisateurs représentatifs (et notamment ceux de remplacement) que j'ai proposées, constituent les deux points les plus originaux de mes travaux.

Ces travaux ont fait l'objet de nombreuses publications et, pour la plupart, se sont faits en collaboration dans le cadre d'encadrements de M2 et de thèses.

3.3 Les données séquentielles : bruit, distance et influence pour la prédiction et la recommandation

Une partie significative de mon activité a porté sur le traitement de données séquentielles. En terme de phénomène, c'est en général le comportement utilisateur que ces données représentent. Les travaux que j'ai menés sur ces données, visent à modéliser, prédire et recommander des comportements ou des événements.

Le Web constitue le contexte général de ces travaux. Je me suis intéressée à deux types de données issues du Web. Tout d'abord des données de navigation Web, puis des données de publication de messages sur le Web (blogs).

Dans le cadre de données de navigation, les traces correspondent à des séquences d'items consultés par un utilisateur. L'utilisateur, auteur d'une séquence, n'est pas connu, mais peut être identifié. Par choix (voir section 1.1), un item est connu uniquement par un identifiant. Les données de navigation auxquelles je me suis intéressée ont des caractéristiques qui les différencient des données de préférence traitées dans les deux sections précédentes :

- elles ne contiennent pas de préférences explicites ou valuées. La seule information disponible est la présence de l'item dans la séquence (l'utilisateur a consulté l'item), on ne sait pas si il a été apprécié ou non,
- l'ordre d'apparition des items au sein de la séquence est important, il permet de connaître le contexte dans lequel l'item est consulté.

Même si aucune information d'appréciation n'est connue sur les items, l'état de l'art suppose que la répétition d'un item dans les traces permet de savoir si l'item est pertinent ou non. En effet, si un item est régulièrement consulté dans un contexte donné, alors cela signifie qu'il est intéressant (pertinent) dans ce contexte [Zimdars et al., 2001, Shani et al., 2005].

Les données de navigation se présentent sous la forme d'un ensemble de séquences de données, dans lequel une séquence de données est la suite d'items consultés par un utilisateur.

La tâche de recommandation reposant sur des données séquentielles revient à prédire le/les items suivant(s) dans une séquence donnée. Ce sont les items prédits qui sont ensuite recommandés. Ainsi, dans la suite de cette section, je parlerai de façon indifférenciée de prédiction et de recommandation.

De nombreux travaux se sont intéressés et s'intéressent à la modélisation de données séquentielles. Des défis ont été identifiés, dont certains sont toujours au cœur des recherches actuelles. J'ai travaillé sur trois de ces défis, avec pour objectif principal la recommandation. Ces défis sont la prise en compte du bruit dans les données, la modélisation de relations distantes au sein des données et l'influence dans les données. Ils font l'objet des trois sections suivantes.

3.3.1 Le bruit dans les données (PS1)

Les approches de l'état de l'art

En 2007, lorsque je me suis intéressée à la prédiction de consultations de ressources Web, deux approches principales étaient proposées dans la littérature.

La première approche repose sur l'hypothèse que la réalisation des phénomènes dans les données séquentielles satisfait la propriété de Markov à l'ordre k , c'est-à-dire qu'un état dépend uniquement des k états précédents, les états antérieurs à ces k états n'ont aucune influence. Dans ces travaux, un état correspond à un item consulté [Chimphlee et al., 2006, Mobasher, 2007]. Dans le cas de données de navigation Web, il est logique de considérer que la consultation d'un item dépend uniquement des derniers items consultés (l'historique de l'item, son contexte), et non pas des items consultés auparavant.

Notons qu'il est également possible de définir un état comme étant une suite d'items, en particulier les derniers items de la séquence. Dans ce cas, le phénomène de navigation satisfait la propriété de Markov (à l'ordre 1).

Toujours avec cette même approche, la littérature a également proposé d'aborder la prédiction pour la navigation Web en exploitant des modèles n -grammes (modèles utilisés en modélisation statistique du langage [Rosenfeld, 2000], voir également section 2.1.3). Un modèle n -grammes représente le langage sous la forme d'un ensemble de probabilités conditionnelles d'éléments du langage, à la suite de $n - 1$ éléments. L'apprentissage des modèles de langage repose sur la fréquence de suites de n éléments. Un modèle n -grammes est donc équivalent à un modèle de Markov d'ordre $k - 1$.

Ces modèles (n -grammes ou Markov) ont l'inconvénient de ne pas être résistants au bruit. En effet, lorsqu'une prédiction doit être faite, ils considèrent systématiquement la suite exacte des k (ou $n - 1$) derniers items consultés. Cependant, si un de ces items représente en fait du bruit, alors le modèle utilisé ne contiendra probablement pas cette suite d'items et ne pourra donc pas fournir de recommandations. Dans le cadre de la navigation, le bruit peut correspondre à une erreur de navigation, au souhait de l'utilisateur de consulter un item qui n'est pas en lien direct avec la navigation courante, etc. Ces modèles ont, par conséquent, une couverture limitée : si la suite des items consultés par l'utilisateur n'est pas

présente dans le modèle, alors il ne pourra pas fournir de recommandations. Ces modèles ont également une couverture limitée en raison de la taille de l'historique pris en compte (n ou $k - 1$). Ceci, combiné à des données de réalisation de taille limitée (sur lesquelles les modèles sont appris), la couverture est à nouveau impactée.

Pour pallier le problème de la résistance au bruit et de la couverture de ces modèles, la littérature a introduit le all- k^{th} -order Markov model [Pitkow and Piroli, 1999] qui forme k sous-modèles d'ordre allant de 0 à k . Pour prédire des items, le modèle exploite en priorité celui d'ordre k et se rabat pas à pas sur celui d'ordre inférieur, tant que le modèle n'est pas capable de fournir une recommandation. Bien qu'ayant une couverture plus élevée, de mon point de vue, ce modèle ne peut pas garantir des prédictions de qualité dans le cas de données bruitées, notamment si le bruit se situe dans l'historique très récent.

La seconde approche de la modélisation de données séquentielles est la fouille de données (séquentielles), qui vise, entre autres, à identifier des motifs séquentiels fréquents dans les données [Han and Kamber, 2006].

De nombreux algorithmes de fouille de données séquentielles ont été proposés dans la littérature. Pour n'en citer que quelques uns : les algorithmes GSP [Srikant and Agrawal, 1996], SPADE [Zaki, 1998], SPAM [Ayres et al., 2002], PrefixSpan [Han et al., 2001]. Ils sont généralement répartis en deux catégories [Mabroukeh and Ezeife, 2010] : les algorithmes basés sur Apriori [Agrawal et al., 1993] et ceux basés sur une structure en arbre (comme FP-Growth [Han et al., 2000]).

Le cadre de la navigation Web a une spécificité : les motifs fouillés sont des motifs d'items (à chaque pas de temps un seul item est consulté), alors que dans le cas général de la fouille de données, ils sont des motifs d'ensembles d'items (dans des traces d'achats, à chaque pas de temps, plusieurs produits sont achetés). La fouille de motifs pour la navigation Web est également connue sous le terme *Web Usage Mining* [Cooley et al., 1997], et les motifs fouillés représentent des séquences-types de navigation.

Une fois les motifs séquentiels formés, les algorithmes construisent des règles (d'association) qui identifient les items probables (la conséquence) à la suite d'un motif séquentiel (l'antécédent). C'est la conséquence qui constitue le/les items à recommander [Mobasher, 2007]. L'antécédent des règles a un rôle similaire à celui de la suite de k états précédents dans les modèles de Markov d'ordre k , et l'item composant la conséquence est l'état suivant prédit.

Un avantage de cette approche réside dans le fait que les motifs séquentiels sont composés d'items qui peuvent être non contigus dans les séquences [Nakagawa and Mobasher, 2003]. Par conséquent, elle est naturellement plus robuste au bruit. De plus, la taille des motifs séquentiels n'est pas imposée *a priori* (au contraire des modèles de Markov d'ordre k). Une grande couverture est donc possible. Cependant, un inconvénient est sa complexité en temps et en espace, due à la non contiguïté des items. Des contraintes temporelles ont donc été introduites pour limiter cette complexité : la contrainte de *gap* maximal, qui impose une durée maximale entre deux éléments du motif, et la contrainte de *span* maximal, qui impose une durée maximale entre le premier et le dernier élément du motif [Srikant and Agrawal, 1996, Zaki, 2000]. Notons qu'il est possible de forcer l'algorithme à fouiller des motifs contigus, et dans ce cas le modèle se rapproche d'un all- k^{th} -order Markov model. Les modèles peuvent même être équivalents, selon la façon dont les motifs sont exploités pour la prédiction.

Dans ces deux grandes approches, un compromis doit cependant être fait entre pouvoir de représentation, fonction de la longueur des séquences-types fouillées (n -grammes ou motifs séquentiels) et du volume de données de réalisation, et la complexité des modèles.

Le modèle SBR

Partant des deux approches exploitées dans la littérature et de leurs limites respectives, j'ai souhaité m'intéresser à un modèle résistant au bruit, avec une grande couverture et une complexité plus faible que celle des modèles à base de motifs séquentiels discontigus.

Avec Geoffroy Bonnin, nous avons proposé de conserver la caractéristique des modèles à base de motifs séquentiels qui, pour être résistants au bruit, fouillent des motifs discontigus. Cependant, nous avons supposé que, même si les données de navigation étaient bruitées, la quantité de bruit était limitée, et par conséquent, fouiller des motifs totalement discontigus ne nous semblait pas pertinent. Il est en effet fort probable que la majorité des motifs découverts dans ce cas ne représentent aucune séquence-type

réelle. Par ailleurs, c'est cette possibilité de non-contiguïté entre tous les éléments du motif qui rend le modèle complexe. Aussi, nous avons proposé un compromis : autoriser des motifs discontigus, mais avec une discontiguïté limitée. Par ailleurs, toujours pour des raisons de complexité, nous avons proposé de fixer la taille des motifs fouillés (contrainte de *span* maximal), comme c'est aussi le cas dans les modèles de Markov. Le modèle résultant est le modèle SBR (Skipping-Based Recommender), un modèle de fouille de motifs séquentiels exploitant le concept de *skipping*³. En modélisation du langage, les modèles *skip*-grammes [Huang et al., 1992, Shani et al., 2005] sont des modèles n -grammes dans lesquels les éléments ne sont pas obligatoirement consécutifs, mais des sauts peuvent être autorisés entre les items des données. Cette notion est similaire à celle de *gap* en fouille de motifs séquentiels [Pei et al., 2007].

Le modèle SBR est un modèle de n -grammes non contigus (des *skip*-grammes). Pour gérer le bruit, tout en limitant la complexité du modèle, nous avons introduit deux contraintes :

- les n -grammes non contigus sont identifiés dans une fenêtre de taille fixe k , à l'apprentissage et au test. Il permet donc de gérer le bruit, mais en quantité limitée. Par ailleurs, le modèle ne gérant que des n -grammes ($n < k$), il est relativement peu complexe,
- un nouveau schéma de *skipping* est introduit, de façon à pouvoir, à nouveau, extraire des séquences réalistes. Ce nouveau skipping est propre à la navigation Web. Nous considérons qu'une personne qui navigue peut à un moment faire une ou plusieurs erreurs (ou aller consulter un item autre), mais pas à plusieurs moments au sein d'une séquence de taille n (la valeur de n étant limitée). Par conséquent, le skipping que nous avons introduit autorise un unique saut dans les séquences, mais ce saut n'a pas de taille maximale (dans la limite de la taille de la fenêtre k).

Ce modèle emprunte donc des caractéristiques aux modèles n -grammes, aux modèles de Markov et à la fouille de données.

Lorsqu'une prédiction/recommandation est requise, ce modèle exploite l'historique de l'utilisateur, restreint aux k derniers items, au sein duquel toutes les séquences de taille $n - 1$ sont recherchées dans le modèle. En cas de multiples séquences identifiées dans le modèle, il faut combiner les différentes prédictions. Nous avons proposé plusieurs stratégies de combinaison qui rendent l'algorithme de prédiction *anytime* [Bonnin et al., 2008b, Bonnin et al., 2009a]. De cette façon, quel que soit le temps disponible, l'algorithme sera capable de fournir une/des recommandations, et plus ce temps est grand, plus la qualité de ces recommandations sera grande également.

Les expérimentations menées ont montré que le modèle SBR fournit des recommandations dont la qualité est soit meilleure que celle des modèles de l'état de l'art, soit comparable mais avec une complexité en temps et en mémoire très inférieure, et montre soit une meilleure résistance au bruit que les modèles de l'état de l'art, soit une résistance comparable mais avec une complexité en temps et en mémoire significativement réduites, caractéristiques qu'ont peu de modèles de la littérature. Cette caractéristique est due à la fois à la notion de contiguïté limitée et au fait d'approcher un historique de taille k avec des n -grammes ($n \ll k$), que SBR est le seul à proposer.

Ce travail a été réalisé en collaboration avec Geoffray Bonnin, durant sa thèse que j'ai co-encadrée, et a été publié en conférences internationales [Bonnin et al., 2008b, Bonnin et al., 2010c] et en journal national [Bonnin et al., 2012].

Rappelons qu'en plus d'être bruitées, les données de navigation, tout comme les données de traces d'activité humaine, sont incertaines et parcimonieuses (voir Chapitre 1). L'approche de gestion du bruit dans les données que nous avons proposée ici peut également être exploitée pour gérer à la fois l'incertitude et le manque de données.

3.3.2 Les relations de distance au sein des données (PS4)

Très rares sont les modèles permettant de modéliser explicitement les relations distantes au sein des données. La raison principale est la complexité des modèles permettant une telle modélisation. C'est d'ailleurs pour cette raison que ce sont des modèles n -grammes ou de Markov qui sont exploités dans la littérature, de même que des contraintes temporelles sont ajoutées dans les approches de fouille de données.

3. "saut" en français

J'ai souhaité m'intéresser à la modélisation de relations distantes, à la fois dans des séquences de données, mais également dans une unique séquence de données.

Dans des séquences de données

Rappelons que dans le modèle SBR introduit ci-dessus, un historique de taille k d'un utilisateur est approché par un ensemble de n -grammes (*skip*-grammes) (avec $n < k$), et la taille de cet ensemble grandit lorsque $n \ll k$. Par conséquent, pour évaluer la probabilité d'items futurs (puis recommander ces derniers), les n -grammes composant l'historique et présents dans le modèle doivent être combinés.

Dans le cas où $n \ll k$, SBR permet de modéliser des relations distantes (dans la limite de k) entre items d'une séquence-type, tout en ayant une complexité limitée. Cependant, certains n -grammes identifiés dans l'historique peuvent être plus significatifs que d'autres. En particulier, nous pensons que le modèle n -grammes exploitant des items éloignés (en terme de distance dans l'historique) de l'item à prédire, ces derniers doivent avoir une importance moindre par rapport à celle de n -grammes plus proches. Nous avons pensé que tenir compte de cet éloignement, notamment lors de la combinaison des n -grammes, pourrait permettre d'améliorer la qualité de la prédiction.

Nous avons proposé plusieurs schémas de pondération, dont un schéma de décroissance linéaire et un schéma de décroissance exponentielle, fonction de la distance entre chaque élément qui compose les n -grammes et l'item à prédire [Bonnin et al., 2008a]. Nous avons par ailleurs proposé d'apprendre automatiquement les paramètres de ces deux schémas [Bonnin et al., 2009a]. Le schéma de décroissance exponentielle est celui qui améliore le plus, et significativement, les performances de l'état de l'art.

Le modèle SBR permet donc, en plus d'être résistant au bruit, de prendre en compte les relations distantes entre éléments de données séquentielles. Son avantage le plus prépondérant est sa faible complexité.

Dans une unique séquence de données

Plus récemment, je me suis de nouveau intéressée aux relations distantes au sein de données, mais dans un tout autre cadre. Ici, les données sont toujours des données séquentielles, mais se présentent sous la forme d'une unique séquence de données.

Dans le cas d'une séquence unique, on parle de fouille d'épisodes et de fouille de règles d'épisodes [Mannila et al., 1997], et les éléments composant la séquence sont en général appelés des événements. La fouille d'épisodes a des difficultés qui lui sont propres, telle que la définition d'une mesure de fréquence. En effet, dans le cas de séquences de données, la fréquence est définie comme le nombre de séquences contenant le motif. Dans le cas d'une séquence unique, la fréquence correspond au nombre d'occurrences du motif dans la séquence. Des problèmes liés à l'entrelacement et à la superposition d'occurrences complexifient la tâche de fouille [Achar et al., 2012].

Par ailleurs, dans les séquences de données, la distance entre les éléments des données est naturellement limitée par la taille de la séquence. Dans le cas de séquences trop longues, une contrainte de *span* peut être exploitée. Dans le cas d'une unique séquence de données, par définition très longue, voire infinie (dans le cas de flux), il est impossible de fouiller des épisodes (ou des règles d'épisodes) dans la séquence entière. Non seulement le temps de fouille serait infini, de même que le nombre d'épisodes/règles fouillés. De plus, la significativité des épisodes/règles fouillés serait mise en doute. Par conséquent, les algorithmes de fouille d'épisodes/règles reposent systématiquement sur l'exploitation d'une contrainte de *span*. Plus la valeur du *span* est petite, plus l'algorithme sera rapide et moins il identifiera d'épisodes, mais plus la représentativité des épisodes et du modèle associé seront limitées. Par conséquent, la valeur du *span* exploité est généralement réduite, mais dans une certaine limite.

Dans certains phénomènes représentés par une unique séquence de données, certaines relations peuvent être distantes, même largement au delà de la valeur traditionnelle du *span*. Ces relations ne peuvent donc pas être identifiées par les algorithmes traditionnels.

J'ai souhaité travailler sur un algorithme de fouille de règles d'épisodes permettant de modéliser les relations distantes entre événements. En particulier, c'est une grande distance entre l'antécédent et la conséquence de la règle qui m'intéresse. Cette caractéristique permet, en terme de prédiction, de prédire des événements dont l'occurrence est lointaine.

Les algorithmes classiques de fouille de règles d'épisodes procèdent tous en deux étapes : 1) la fouille des épisodes, puis 2) la construction des règles. Les mesures de support et de confiance sont également utilisées pour la construction des épisodes et des règles d'épisodes, tout comme pour la fouille de motifs et de règles d'association. Le fait que ces algorithmes reposent sur une première étape de fouille d'épisodes a pour conséquence qu'ils ne sont pas adaptés à la fouille de règles avec une conséquence distante. En effet, lors de la construction des épisodes, qui se fait en post-fixant itérativement un événement à l'épisode en cours de construction, l'algorithme ne peut pas savoir si cet événement fera partie de la conséquence ou de l'antécédent de la règle formée dans une seconde étape. Par conséquent, l'algorithme ne peut pas décider s'il doit contraindre la distance entre l'événement ajouté à l'épisode et le reste de l'épisode. Le seul moyen pour ces algorithmes de fouiller des règles avec une conséquence distante nécessite de (1) fouiller tous les épisodes en utilisant un large *span*, (2) former toutes les règles confiantes, (3) filtrer les occurrences de ces règles en éliminant celles qui ne respectent pas la distance minimale. En raison du large *span* et du post-traitement, ce processus est très gourmand en temps d'exécution.

L'algorithme que j'ai conçu, en collaboration avec Lina Fahed, DEER (Distant and Essential Episode Rules) permet de fouiller de telles règles. Cela est rendu possible grâce à l'originalité de cet algorithme qui fouille les règles d'épisodes sans passer par une première étape de fouille d'épisodes, caractéristique qui permet également de limiter la complexité de la fouille de telles règles. Pour ne pas reposer sur la fouille d'épisodes, DEER fixe la conséquence de la règle très tôt dans le processus, et par conséquent, maîtrise sa distance à l'antécédent. C'est la connaissance de la conséquence pendant le processus de fouille qui permet de diminuer la complexité de l'algorithme. En effet, un très grand nombre d'élagages de règles candidates se fait lors du processus, diminuant ainsi de façon significative le temps d'exécution.

L'approche consistant à fixer la conséquence tôt dans le processus permet également de fouiller des règles dites essentielles. Une règle essentielle est une règle dont l'antécédent est minimal, c'est-à-dire, non seulement le plus petit possible en nombre d'événements le constituant, mais aussi en durée.

D'un point de vue applicatif, la fouille de règles avec une conséquence distante permet de prédire des éléments dans une séquence de données, très tôt avant leur occurrence, ce qui permet à des tiers de pouvoir réagir en cas de prédictions d'événements non souhaités. Les règles essentielles permettent de prédire la conséquence lorsqu'un nombre minimal d'éléments est apparu, impactant d'autant plus la distance entre l'antécédent et la conséquence.

Les données sur lesquelles nous avons travaillé ne sont plus des données de navigation Web, mais des données de publications sur le Web, et notamment de blogs. Les données, dans ce cadre, sont plus proches des données traditionnelles utilisées dans le cadre de la fouille de données séquentielles : un ensemble d'items à chaque pas de temps (le message posté). Les expérimentations menées ont confirmé que le temps d'exécution de DEER est très significativement inférieur à celui de l'état de l'art.

Des expérimentations supplémentaires ont été conduites, dans lesquelles DEER a été exploité mais pour former des règles dont la conséquence n'est pas distante de l'antécédent. DEER a également montré une amélioration significative du temps d'exécution, en raison d'un très grand nombre d'élagages au sein du processus. Nous en avons donc conclu qu'identifier la conséquence tôt dans le processus de fouille est un moyen de réduire significativement le temps de fouille requis.

Ce travail a été mené en collaboration avec Lina Fahed, durant sa thèse que j'ai co-encadrée, et a été publié en conférence internationale [Fahed et al., 2014a] et en journal international [Fahed et al., 2018].

L'algorithme DEER a une double originalité : **DEER est le premier algorithme de l'état de l'art permettant de fouiller des règles sans passer par l'étape de fouille d'épisodes, ce qui a pour conséquence une diminution de la complexité en temps grâce au grand nombre d'élagages que cette caractéristique permet. Il est, par ailleurs, le premier algorithme de fouille de règles avec une conséquence distante.**

Récemment, des travaux sur la fouille de règles d'épisodes composées d'éléments distants ont été proposés. Notamment [Ao et al., 2017], qui propose une distance fixe entre chaque élément des épisodes composant la règle, et notamment entre l'antécédent et la conséquence.

3.3.3 Les relations d'influence au sein des données

L'algorithme DEER, présenté ci-dessus, permet de fouiller des règles dont la conséquence est éloignée. Par conséquent, ces règles peuvent être utilisées pour prédire des événements lointains. Dans le cas où l'événement prédit n'est pas souhaité par la/les personnes concernée(s) par cet événement, elle/elles peut/peuvent agir de façon à empêcher l'occurrence de cet événement.

J'ai souhaité travailler à l'identification automatique des événements qui impactent l'occurrence d'autres événements.

Avec Lina Fahed, nous avons proposé l'algorithme IE (Influencer Events), un algorithme pour la détection automatique d'événements que nous avons appelés "influenceurs" dans une séquence d'événements. IE modélise les associations entre événements au travers de l'"impact" porté par certains événements sur d'autres événements. Cet algorithme a l'originalité de détecter les événements qui, une fois injectés dans un contexte spécifique, permettront d'influencer des événements futurs. IE a été conçu de façon à être une brique supplémentaire de DEER, ce qui a pour conséquence que la complexité de ce dernier n'est que très légèrement impactée.

Ce sont les événements influenceurs qui sont recommandés. Trois types d'influences ont été étudiées : annuler l'occurrence de l'événement prédit, rapprocher temporellement l'occurrence de cet événement ou simplement augmenter sa probabilité d'occurrence.

D'un point de vue applicatif, soit les événements influenceurs sont automatiquement injectés dans la séquence de données, ce qui permet de moins reposer sur une intervention humaine, soit ils sont recommandés à la personne concernée, qui agit elle-même. Dans ce dernier cas, les événements influenceurs identifiés doivent également respecter une contrainte de distance entre le moment où l'événement est prédit, et celui où l'événement déclencheur peut être injecté dans la séquence, de façon à ce que la personne concernée puisse disposer du temps nécessaire pour injecter cet événement.

L'impact d'un événement sur un événement prédit est évalué au travers de l'évolution de la probabilité d'occurrence de l'événement prédit. Le terme "impact" est cependant un abus de langage puisque les expérimentations menées n'ont été faites que sur des données de séquences *offline*. Des expérimentations futures devront être conduites en situations réelles, et notamment effectuer des tests A/B permettrait de valider notre approche.

L'originalité de l'algorithme IE réside dans le fait qu'il détecte automatiquement les événements influenceurs adéquats pour chaque contexte (comme influencer les achats d'un client particulier, le comportement d'un utilisateur particulier dans un blog, etc.) et non pas un ensemble d'événements influenceurs en général, exploitables quel que soit le contexte. Les expérimentations menées montrent la pertinence du modèle, notamment en détectant des événements qui influencent en même temps plusieurs caractéristiques des événements prédits.

L'identification automatique d'événements influenceurs est novateur dans la littérature.

Ce travail a également été mené en collaboration avec Lina Fahed, durant sa thèse que j'ai co-encadrée, et a été publié en conférence internationale [Fahed et al., 2015a].

3.3.4 En résumé

Dans cette section, j'ai présenté quelques travaux que j'ai menés en fouille de données séquentielles. Ces travaux avaient pour objectif principal la prédiction/recommandation.

Je me suis intéressée à trois défis relatifs à la gestion de données séquentielles : la gestion du bruit, la modélisation de relations distantes et la détection d'influence dans les données.

Pour gérer le bruit dans les données, j'ai proposé l'algorithme SBR, dont l'objectif est de fouiller des séquences-types non contiguës, mais avec une discontiguïté limitée, ce qui permet de limiter la complexité. L'algorithme SBR est le seul de l'état de l'art à proposer une telle discontiguïté.

Pour modéliser les relations distantes, j'ai à nouveau travaillé sur l'algorithme précédent et montré qu'il pouvait être utilisé pour de telles relations. Par ailleurs, j'ai conçu une version *anytime* de cet algorithme. Je me suis également intéressée aux données se présentant sous la forme d'une séquence de données unique (qui peut également être un flux). La modélisation de relations distantes dans ce type de données est naturellement complexe. J'ai proposé un algorithme de fouille de règles représentant des

relations distantes, notamment entre l'antécédent et la conséquence. Cet algorithme procède d'une façon très originale pour fouiller ces règles, en fixant la conséquence tôt dans le processus. Cette approche est nouvelle dans la littérature de la fouille de données, elle permet de diminuer de façon très significative la complexité de la tâche de fouille.

Enfin, je me suis intéressée à la proposition d'un algorithme de détection automatique d'événements influenceurs dans une séquence de données. L'algorithme que nous avons conçu est le premier de la littérature à permettre la fouille de tels événements. Par ailleurs, ce dernier reposant sur l'algorithme précédent, sa complexité additionnelle est très limitée.

Ces travaux ont fait l'objet de nombreuses publications et, pour la plupart, se sont faits en collaboration dans le cadre d'encadrements de thèses.

Le chapitre suivant s'intéresse à mon projet de recherche, qui est en lien direct avec la dernière section de ce chapitre.

Chapitre 4

Projet de recherche

L'ensemble des travaux que j'ai présentés dans les chapitres précédents m'a mené à soulever de nombreuses questions. Les activités de recherche que je souhaite entreprendre dans les prochaines années visent, entre autres, à y répondre. Elles se situent donc dans la continuité de ces travaux passés, tout en s'intéressant à de nouveaux défis et problématiques de recherche.

Mes travaux passés reposaient sur l'exploitation de données, qu'elles soient des données explicites (données de préférence utilisateur, tels que des votes), ou des données implicites (données de comportement utilisateur). Cependant, bien que les données explicites soient très riches (porteuses de beaucoup d'informations), et qu'elles permettent en général une modélisation de qualité, je fais le constat qu'elles ne sont finalement disponibles qu'en faible nombre. En effet, il s'est avéré que dans de très nombreuses applications, les utilisateurs ne fournissent que rarement un retour explicite aux systèmes, car ils manquent de temps, car ils n'y voient aucun bénéfice direct, car ils ne souhaitent pas dévoiler leurs appréciations, car ils sont submergés de demandes de retours, etc. Les données implicites (les traces de comportement par exemple), bien que porteuses de moins d'informations, peuvent, quant à elles, être disponibles en très grand nombre, car elles peuvent être automatiquement collectées sans que les utilisateurs soient explicitement sollicités. Par ailleurs, il est possible que cette masse de données puisse compenser la faible quantité d'information présente dans cette source. Rappelons qu'à partir des données implicites, on peut déduire des données se rapprochant de données explicites, en exploitant, par exemple, la durée de consultation, la fréquence, des actions spécifiques : achat, sauvegarde, impression, partage, etc. Cependant, les données résultantes sont relativement peu fiables, car elles ont bien souvent des degrés d'expressivité plus faibles et sont souvent bruitées [Hu et al., 2008, Jawaheer et al., 2010].

Pour mes travaux futurs, je fais donc le choix d'**exploiter principalement des données implicites, en particulier des traces d'activité**, de comportement, ce qui me permettra de me confronter à nouveau aux problèmes du volume et du bruit dans les données (caractéristique inhérente à la nature de la source), mais également à celui de l'hétérogénéité des données (dans le cas où plusieurs sources de données sont disponibles).

Par ailleurs, je me suis jusqu'à présent intéressée à la modélisation de phénomènes avec pour double objectif, de les comprendre (modèles descriptifs) et d'en prédire des réalisations (modèles prédictifs). J'ai récemment travaillé sur les liens existant entre événements et sur la possibilité de les exploiter pour "agir" sur la réalisation future d'événements (co-encadrement de la thèse de Lina Fahed). Dans mes travaux futurs, je souhaite travailler cet aspect plus en profondeur. Plus précisément, mon objectif ne sera plus simplement de prédire des réalisations futures, mais, sachant un objectif futur que l'on souhaite atteindre, de déterminer les actions à réaliser pour atteindre cet objectif. Ce travail rentrera dans le cadre de **l'analyse prescriptive**.

La performance de systèmes, et tout particulièrement des systèmes de recommandation, est généralement évaluée comme une performance moyenne. Elle ne permet donc de mettre en avant ni des performances fluctuantes, ni des performances moindres sur certains utilisateurs ou items. Dans mon activité passée, je me suis intéressée à un sous-ensemble d'utilisateurs qui avaient la caractéristique d'être différents des autres et recevaient des recommandations de faible qualité, les utilisateurs dits "moutons

gris” (co-encadrement de la thèse de Benjamin Gras). J’ai travaillé non seulement sur leur identification en amont de toute recommandation, mais également sur la conception de modèles de recommandation dédiés, de façon à leur garantir des recommandations de qualité. Je souhaite poursuivre ces travaux en les adaptant à des données implicites, mais également en les élargissant à toutes spécificités d’utilisateurs, et d’items. Cela peut, par exemple, concerner des comportements cycliques, des façons spécifiques de se comporter, comme le fait de systématiquement se rendre sur les items nouvellement arrivés, ou, au contraire, de ne jamais s’y rendre, etc. Cela peut également ne pas concerner des utilisateurs, mais des phénomènes présents dans les données et relatifs aux items. Ici, l’objectif applicatif est à nouveau de garantir des recommandations de qualité à l’ensemble des utilisateurs, quelles que soient leurs caractéristiques. Mon objectif scientifique est donc d’**identifier, comprendre et prendre en compte les particularités des données (utilisateurs, items)**.

Je rappelle ici mon intérêt fort pour l’exploitation des modèles que je conçois, sur des données et applications réelles. Par conséquent, les aspects relatifs à la complexité, et notamment à la légèreté des modèles demeurent des aspects à considérer dans mes travaux futurs, tout comme ils l’ont été dans mes travaux passés [Esslimani et al., 2010a, Bonnin et al., 2010c, Boumaza and Brun, 2012b, Fahed et al., 2014a, Aleksandrova et al., 2017a]. De la même façon, **la conception de modèles dynamiques**, c’est-à-dire de modèles qui garantissent la prise en compte de nouvelles données en temps réel, mais qui, également, identifient et prennent en compte leur évolution, est une dimension que je souhaite considérer de façon plus approfondie dans mes travaux futurs. Ici, l’objectif est encore de garantir un service de qualité à tous et à tout moment.

Partant de ces quatre éléments, je peux formuler de la façon suivante la problématique générale à laquelle je souhaite désormais m’intéresser :

Concevoir des modèles utilisateurs et de recommandation, dynamiques et légers, appris sur des données implicites très volumineuses (mais parcimonieuses au niveau des utilisateurs), véloce et hétérogènes, permettant à chaque utilisateur, quelles que soient ses caractéristiques, d’atteindre un but (une cible) donné.

Au sein de cette problématique générale, mon défi principal est “comment définir les items à recommander à un utilisateur, de façon à ce qu’il puisse atteindre un but donné?”. Il est en effet le défi qui, pour moi, soulève le plus de questions, de tout ordre, et il sera traité sur le long terme. Pour cette raison, il constituera la première et plus grande partie de mon projet de recherche.

Les aspects relatifs à la gestion des particularités des utilisateurs, quelles qu’elles soient, mais aussi aux modèles dynamiques et légers, seront vus comme des aspects complémentaires au défi principal et feront l’objet de la seconde partie, moins conséquente, de mon projet.

L’ensemble des travaux liés à cette problématique sera mis en œuvre sur un domaine applicatif principal qu’est l’éducation, dont les spécificités impacteront certains aspects des travaux que je mènerai. Bien évidemment, d’autres domaines seront considérés mais dans une moindre mesure, tels que le e-commerce sur lequel je travaille depuis de nombreuses années.

Dans ce chapitre, je présente tout d’abord le domaine applicatif principal. Ensuite, je me focaliserai sur le défi qui est au cœur de ma problématique générale, en détaillant les sous-défis associés, puis je m’attarderai sur les aspects complémentaires au défi principal, le tout en lien avec le cadre applicatif.

Bien évidemment, ce projet de recherche ne pourra pas être mené à bien sans collaboration avec des psychologues et spécialistes de l’éducation et du numérique dans l’éducation. La partie du projet de recherche présentée dans ce document est cependant principalement axée sur les propositions et défis liés à la dimension informatique de ce projet.

4.1 L’éducation : domaine d’application principal

Au sein du domaine de l’éducation, la dimension qui m’intéresse est l’éducation appuyée sur le numérique (connue dans la littérature sous le terme *Technology-Enhanced Learning* (TEL) ou e-éducation), qui

regroupe toutes les solutions ou dispositifs pédagogiques, d'appui aux apprenants et/ou aux enseignants, reposant sur le numérique. Cela inclut les dispositifs permettant l'accès à des supports, en ligne ou non ; les dispositifs d'apprentissage individuel ou collectif, en présence ou non d'un enseignant, en présentiel ou à distance ; d'accès à une aide (tuteur) ; d'accès à des corrections, etc. Ces dispositifs peuvent donc être exploités à la fois dans un contexte d'apprentissage traditionnel individuel en présentiel, avec accès à des ressources pédagogiques numériques, jusqu'à un apprentissage en ligne, sans contrainte de lieu, ni de temps, et sans enseignant. Ces dispositifs peuvent, par exemple, reposer sur des plateformes numériques d'apprentissage.

Dans ce document, j'utiliserai volontairement uniquement le terme "étudiant" pour faire référence de façon indifférenciée à tout type d'apprenant : élèves du primaire, du secondaire, étudiants de l'enseignement supérieur, mais également tout apprenant en formation initiale ou continue, en formation tout au long de la vie, inscrit ou non dans une formation.

L'éducation appuyée sur le numérique est un domaine très vaste, qui va de la détermination de la nature de l'appui à fournir, du public concerné (étudiant/enseignant), du contexte possible d'apprentissage et/ou du moment où le dispositif peut/doit être utilisé, au choix de la partie prenante initiant l'utilisation de l'outil (étudiant ou enseignant ?).

D'un point de vue plus technologique, cela concerne également la conception de ressources éducatives numériques, la gestion des aspects techniques, la scénarisation de l'apprentissage, la conception d'outils collaboratifs, etc.

De leur côté, les chercheurs, qu'ils soient en sciences de l'éducation, mais aussi en EIAH ou en intelligence artificielle, s'intéressent également au numérique pour l'éducation.

L'objectif partagé par tous est de permettre un enseignement adapté, voire personnalisé, et plus flexible, en vue de l'amélioration du processus d'apprentissage des étudiants.

L'éducation est un domaine d'intérêt relativement récent pour moi, j'y ai initié quelques travaux, notamment dans le cadre des projets PIA2 PERICLES (2012-2015), INTERREG IVA Interlingua (2014-2015), PIA2 e-FRAN METAL (2016-2020). Ces travaux m'ont permis non seulement de me familiariser avec ce domaine, de comprendre les raisons de l'intérêt que les acteurs académiques et les chercheurs y portent, mais aussi d'en identifier les nombreux enjeux, à la fois applicatifs et scientifiques. C'est ce dernier point qui a motivé mon choix de poursuivre mes travaux dans ce domaine en particulier.

Les learning analytics

Mon intérêt au sein de l'e-éducation porte sur les *learning analytics*, qui visent à collecter, mesurer, analyser et synthétiser les données sur les étudiants, les enseignants ainsi que sur leurs contextes respectifs [JISC, 2011]. Leur objectif est non seulement de comprendre et d'évaluer le processus d'apprentissage et l'environnement dans lequel il se fait [Ferguson, 2012], mais également de l'améliorer, en fournissant aux acteurs de l'enseignement (étudiants, enseignants ou institutions) des recommandations ou des modèles du processus d'apprentissage [Long and Siemens, 2011]. Les *learning analytics* traitent ainsi principalement de la réussite des étudiants, de la qualité de la formation, mais peuvent aussi concerner la dimension éthique et financière de l'enseignement.

Les *learning analytics* constituent un domaine très récent, le terme a été mentionné pour la première fois en 2011 [Duval, 2011]. Ils suscitent un intérêt croissant de la part de l'ensemble des acteurs de l'enseignement, notamment en raison de leur grand potentiel et des nombreuses retombées que les enseignants, étudiants et institutions entrevoient. Malgré les nombreux travaux qui sont actuellement menés, le domaine des *learning analytics* en est encore à ses débuts : "*Learning analytics is in its infancy*"⁴ et les questions ouvertes restent très nombreuses.

La mise en œuvre des *learning analytics* est classiquement divisée en cinq étapes : 1) la génération des données, qui est en général faite par les environnements d'apprentissage, mais également par les systèmes d'information des institutions, 2) la collecte, le pré-traitement et le stockage des données, 3) l'analyse des données, qui vise à y découvrir des informations cachées et à en extraire de la connaissance (également appelée analyse descriptive), 4) la prédiction, qui vise à prédire les réalisations et tendances futures (également appelée analyse prédictive) et l'action, qui vise à personnaliser l'environnement d'apprentissage,

4. Barbara Wasson, ICDE Summit, 2017

à prescrire des stratégies d'apprentissage, de formation, etc. (également appelée analyse prescriptive), 5) leur mise en place dans un dispositif et l'analyse de leur usage. Cette dernière étape comprend également une boucle de retour sur la première étape, permettant ainsi d'adapter les environnements d'apprentissage.

La ressource au cœur des *learning analytics*, et sans laquelle ils ne peuvent être envisagés, sont les données relatives aux étudiants, enseignants et institutions. Je vais donc m'attarder sur cet aspect.

Des données

Depuis plusieurs années, l'utilisation du numérique dans l'enseignement explose littéralement, qu'il concerne le primaire, le secondaire, le supérieur ou la formation tout au long de la vie, grâce, notamment, à la généralisation des environnements numériques d'apprentissage.

Les enseignants et les étudiants voient dans ces environnements la possibilité de diffuser et d'accéder à des données et à de la connaissance : cours, exercices, etc. ; d'évaluer et de se faire évaluer : quizz, exercices, etc. ; d'échanger avec d'autres étudiants ou avec des enseignants : forums, etc. ; d'étudier aux moments qui leur conviennent et non pas à des créneaux prédéfinis ; d'étudier de la façon qui leur convient, etc.

Dès qu'un enseignant ou un étudiant effectue une action dans un environnement numérique d'apprentissage : consulter un cours, un exercice, réaliser une auto-évaluation, discuter avec un pair, déposer un message sur un forum à destination de son enseignant ou d'un pair, etc., il laisse une trace "numérique", une trace de son activité, c'est une donnée implicite. La génération de ces données correspond à la première étape de la mise en œuvre des *learning analytics*. Ces données de traces représentent donc des activités d'apprentissage et d'enseignement.

Les étudiants et les enseignants utilisant de plus en plus fréquemment ces environnements (ne serait-ce que parce qu'ils rassemblent toute l'information/matériel dont ils ont besoin), **la quantité de données** qu'ils laissent, et qui peut être exploitée, **est donc énorme** et **des données peuvent être générées en continu**. On peut même considérer ces données comme un flux. Comme toutes les traces de réalisation, et notamment celles sur lesquelles j'ai travaillé par le passé, elles **sont bruitées et incertaines**. Ces données peuvent, par ailleurs, être complétées par d'autres sources, notamment les systèmes d'information des institutions (qui contiennent également des données collectables). Ces données contiennent des informations relatives aux étudiants (démographiques, évaluations, etc.), aux enseignants, aux formations et aux cours (description, composition, objectifs, prérequis, etc.), etc. **Ces données sont donc hétérogènes**, entre et au sein de chaque source. Par ailleurs, elles sont riches, probablement complémentaires, et concernent les étudiants, les enseignants, les institutions, les formations, les ressources d'apprentissage, mais également l'interaction entre tous ces éléments.

La collecte et le pré-traitement de ces données constituent la deuxième étape des *learning analytics*. La pré-traitement peut avoir pour but de filtrer le bruit, de fusionner les différences sources de données, etc. Notons que la dimension éthique et de conservation de la vie privée est au centre des *learning analytics*, et de cette deuxième étape en particulier.

L'analyse automatique de ces données permet de déduire de nombreuses informations, comme les habitudes de travail des étudiants, les liens explicites et implicites entre étudiants, les centres d'intérêt, les ressources pédagogiques peu consultées ou au contraire très populaires, les ressources qui semblent complexes aux étudiants, les comportements menant souvent à un échec, etc., qui peuvent ensuite être exploitées pour adapter les dispositifs d'enseignement.

Cette analyse constitue la troisième étape des *learning analytics* : la modélisation descriptive. Notons que la génération, la collecte, le pré-traitement et l'analyse peuvent être réalisées sans jamais solliciter explicitement ni les étudiants, ni les enseignants.

Des questions

Les institutions, les enseignants, les étudiants et les chercheurs se posent de très nombreuses questions relatives aux possibilités, aux difficultés, aux limites et aux risques des *learning analytics*. Ces questions portent à la fois sur l'aspect opérationnel, technique et recherche. Je cite ici quelques questions que les acteurs se posent, en lien avec l'aspect recherche :

Comment définir le profil d'un étudiant ? Cette première question est souvent abordée dans la littérature, le profil d'un étudiant étant au cœur de beaucoup de systèmes. Le profil d'un étudiant doit bien évidemment rendre compte de l'état de ses connaissances, qu'il les ait acquises au travers des activités qu'il a réalisées dans un cadre formel d'enseignement, *via* l'environnement d'apprentissage ou en dehors de celui-ci, mais également en dehors de ce cadre. Le profil d'un étudiant doit également représenter les éléments qui caractérisent l'étudiant, au-delà de ses connaissances pures. On peut, par exemple, citer son rythme de travail, le type de ressources d'apprentissage qu'il préfère étudier (média, nature, longueur, etc.), son style d'apprentissage, etc. [Henze and Nejd, 2004]. C'est ce qui est classiquement appelé les caractéristiques et préférences de l'étudiant [Karampiperis and Sampson, 2005]

L'identification de l'état des connaissances d'un étudiant constitue un premier défi. En effet, à activités équivalentes, les connaissances acquises par deux étudiants peuvent ne pas être les mêmes. L'identification des caractéristiques et préférences des étudiants constitue un second défi. Notamment, il faudra distinguer les activités (nature, moment, vitesse, etc.) qui ont été réalisées sous l'impulsion et par choix de l'étudiant, de celles qui lui ont été imposées (instruction de l'enseignant, temps court avant une évaluation, etc.).

Ainsi, deux questions sont classiquement soulevées dans la littérature : quelles informations doivent être présentes dans le profil, et comment peuvent-elles être évaluées ? Par ailleurs, elles ont fait et font l'objet de nombreux travaux [Wei and Yan, 2009, Dwivedi and Bharadwaj, 2015, Wu et al., 2015]. C'est une question traditionnellement posée, notamment en EIAH, à laquelle les *learning analytics* peuvent contribuer à répondre.

Comment gérer la mobilité des étudiants ? les technologies mobiles se généralisant, elles ont contribué au développement de l'apprentissage ubiquitaire. Des questions relatives à la prise en compte des spécificités liées au support mobile, au débit de la connexion, à l'environnement dans lequel l'apprenant est immergé, se posent pour les tâches de modélisation, de prédiction et de prescription. Les *learning analytics* peuvent aider à fournir une réponse à ces questions, en exploitant les traces laissées par les étudiants, traces contenant des informations sur les terminaux utilisés, la date/heure, etc. De la même façon, cette mobilité a également pour conséquence de générer de nombreuses sources de données, sources par essence hétérogènes, associées à des contraintes spécifiques, que les *learning analytics* doivent prendre en compte.

Comment s'adapter à l'apprentissage tout au long de la vie ? Dans le cas de la formation tout au long de la vie, un étudiant doit combiner vie professionnelle, vie privée et activités d'apprentissage. Les questions qui se posent alors sont : comment faciliter le passage d'un contexte de vie à un autre et comment exploiter ce qui a été appris dans un contexte pour un autre contexte ? comment construire un modèle de l'apprenant valable pour les différents contextes dans lesquels il évolue et qui intègre des informations long terme ?, etc. Les *learning analytics* peuvent également contribuer à répondre à ces questions en permettant, notamment, de déterminer automatiquement le contexte d'une activité donnée, de comprendre/comparer les activités issues de différents contextes, etc.

Comment gérer l'évolution du profil des étudiants et identifier les nouvelles pratiques ? Le profil des étudiants évolue constamment, notamment en raison de l'acquisition progressive de nouvelles connaissances. Comment prendre en compte en permanence les dernières données, pour que le système agisse au plus près du profil de l'étudiant ? Comment identifier les nouvelles pratiques, comment estimer si elles seront pérennes ou non ? Comment comprendre et tenir compte d'un changement de contexte (lieu, environnement, et même objectif) ?

Comment s'assurer d'un service de qualité pour chaque étudiant ? Dans le domaine de l'éducation, les dispositifs visant à accompagner les étudiants dans leur processus d'apprentissage doivent garantir un service de qualité à chacun de leurs étudiants. Sous le terme "service", je regroupe toute adaptation, personnalisation, aide, recommandation, etc. proposée à un étudiant. En éducation, des conséquences très négatives peuvent résulter d'un manque de qualité : non acquisition ou mauvaise acquisition des connaissances, démotivation, abandon, etc.

Le numérique permet également d'offrir un accès à l'enseignement à un très grand nombre d'étudiants, non présents sur un lieu physique d'enseignement. Ces étudiants peuvent, par conséquent, avoir un profil, une culture, et des pratiques très hétérogènes, dont le dispositif doit tenir compte. Les *learning analytics* peuvent-ils permettre et garantir une telle prise en compte ?

Il est par ailleurs important que les services proposés par les dispositifs ne supplantent pas, même en partie, le travail de l'étudiant. En effet, les étudiants doivent être aidés, au plus près, de la meilleure façon

possible, mais comment s'assurer qu'ils ne perdent pas leur autonomie ?

Mes questions

Mon intérêt dans les *learning analytics* porte principalement sur l'analyse des données recueillies, mais également sur les étapes suivantes de prédiction et de prescription (troisième et quatrième étapes). Ma cible privilégiée est la population étudiante, et je souhaite notamment concevoir des modèles et les dispositifs associés permettant de fournir à chaque étudiant un service adapté, voire personnalisé, dans le but d'améliorer son processus d'apprentissage et, par conséquent, ses connaissances. Un tel service peut, par exemple, proposer à un étudiant, au moment opportun, des activités correspondant à son profil, dans le but de lui faire acquérir des connaissances.

Dans la problématique générale de mon projet, j'ai mentionné **deux questions particulières**, qui constituent mes défis principaux. Ces questions trouvent leur instantiation en *learning analytics*.

- **Comment mener un utilisateur vers un objectif donné ?**, ou comment lui prescrire les actions qu'il doit effectuer pour atteindre un but ?
Cette question fera l'objet de la section 4.2 - De la prédiction à la prescription : comment mener un utilisateur vers un objectif donné ?
- **Comment garantir un modèle de qualité pour tous ?** permettant de modéliser précisément tous les utilisateurs, quelles que soient leurs caractéristiques et leurs évolutions ?
Cette question fera l'objet de la section 4.3 - Un système de qualité pour tous et en permanence.

Les solutions à ces défis constituent le cœur de mon projet de recherche. Dans les sections ci-dessous, je discuterai chacun de ces deux défis, tout d'abord en dehors de tout cadre applicatif, puis je les instancierai sur les *learning analytics*.

4.2 De la prédiction à la prescription : comment mener un utilisateur vers un objectif donné ?

Dans cette section, je traite le défi principal de ma problématique de recherche : sachant un objectif ou un but donné, comment arriver à ce dernier ?

Plusieurs éléments caractérisent ce défi :

- L'objectif/le but visé doit être atteint par un utilisateur, et non pas par un système. Je détaillerai ci-dessous l'impact, sur les modèles à concevoir, du fait que c'est un utilisateur qui doit atteindre le but.
- Pour mener l'utilisateur au but, je choisis de passer par des recommandations, qui seront faites à ce dernier. Ces recommandations sont des actions ou des items, qu'il doit suivre, consulter, consommer, étudier, réaliser, etc., et qui lui permettront d'arriver au but fixé.

La recommandation est une problématique sur laquelle a porté la majorité des travaux que j'ai menés ces dernières années. Dans ces travaux, ainsi que dans la littérature, les recommandations destinées aux utilisateurs ont trois caractéristiques :

- elle sont utilisées en tant que conseil ou suggestion donnés aux utilisateurs,
- l'objectif de ces suggestions est de satisfaire, soit le recommandeur, soit la personne recevant les recommandations, et en général c'est ce dernier cas qui est considéré,
- la satisfaction visée est bien souvent à court terme, c'est-à-dire que la satisfaction de l'utilisateur est immédiate, dès qu'il adopte la recommandation.

En ce sens, la notion de recommandation se rapproche de son sens premier, où une recommandation est l' "action d'exhorter quelqu'un à faire quelque chose"⁵, elle est "un conseil".

5. Définition du dictionnaire Larousse

Abordée ainsi, elle ne permet pas de répondre exactement au défi sur lequel je souhaite travailler. En effet, dans la question qui m'intéresse :

- l'objectif visé, bien qu'étant également de satisfaire l'utilisateur, n'est pas un objectif à court terme, mais un objectif à plus long terme (le but à atteindre),
- la recommandation ne vise pas simplement à être un conseil donné à l'utilisateur, mais une recommandation que celui-ci est très fortement incité à suivre, de façon à maximiser ses chances d'atteindre le but fixé.

De mon point de vue, c'est plutôt la tâche de prescription qui permet de répondre à la question que je pose. Une prescription peut être définie comme : "un ordre formel et détaillé énumérant ce qu'il faut faire"⁶.

La prescription et la recommandation sont donc deux notions et deux tâches différentes, avec des implications différentes des acteurs concernés, notamment des bénéficiaires de la recommandation ou de la prescription (clients, utilisateurs, etc.). Cependant, les systèmes de recommandation qui se sont largement démocratisés ces dernières années, ont contribué à rapprocher ces deux notions⁷, les considérant en général comme très proches, voire équivalentes, les deux termes étant souvent exploités de façon indifférenciée. L'existence d'une réelle différence entre ces deux tâches est même parfois remise en cause⁸.

De mon point de vue, ces deux tâches diffèrent réellement de par l'importance de suivre la recommandation/prescription, et de par l'horizon du but fixé. Dans la suite de ce document, je vais donc différencier ces deux termes. Le terme recommandation sera utilisé pour faire référence à une suggestion qui a un objectif visé à court terme : une satisfaction directe liée à la recommandation. A l'opposé, le terme prescription sera utilisé dans le cadre d'une incitation forte avec un objectif de satisfaction à plus long terme. L'objectif visé d'une prescription n'étant pas immédiat, il est, par conséquent, très important que cette dernière soit suivie. Le défi que je traite se rapporte bien à des **prescriptions**.

Avant d'exposer mon projet relativement à ce défi, je commence par résumer les travaux que j'ai menés dans le cadre de la recommandation (reposant sur la prédiction), ainsi que mes premiers travaux en prescription. Ce sont sur ces travaux que repose mon projet.

4.2.1 Synthèse des travaux effectués en prédiction

Je rappelle que l'analyse prédictive est la tâche qui vise à prédire la réalisation d'événements futurs, ou l'occurrence d'items, à la suite de la réalisation d'une séquence (observée jusqu'au moment où la prédiction doit être réalisée : l'historique). Les cadres applicatifs de l'analyse prédictive sont très variés : météorologie, chaînes de production, e-commerce, BI, etc.

Dans les systèmes de recommandation, l'analyse prédictive a pour but de prédire les items les plus adéquats pour un utilisateur, à la suite d'une séquence relative à ce dernier (séquence d'achats, de consultations, d'impressions, etc. ou toute combinaison de ces différentes actions). Reposant sur l'hypothèse qu'un item probable est un item qui correspond aux besoins ou aux souhaits des utilisateurs, c'est cet item que le système recommande alors. Dans le cas où c'est un ensemble d'items qui doit être proposé à l'utilisateur, des notions de diversité [Niemann and Wolpers, 2013, Castagnos et al., 2014] et de nouveauté [Castells et al., 2015] au sein de cet ensemble, sont à considérer. Il est à noter que la très grande majorité des travaux menés en recommandation a pour objectif d'identifier les items les plus probables suivant immédiatement la séquence de l'utilisateur (l'historique). Dans un cadre de e-commerce ou d'un intranet documentaire (cadres applicatifs sur lesquels j'ai travaillé), cette caractéristique se justifie pleinement. Elle permet ainsi de recommander des items qu'il serait intéressant que les utilisateurs achètent ou consultent dans l'immédiat.

Mes travaux récents ont porté sur un aspect original de cette prédiction, à savoir qu'ils n'ont pas visé à prédire des items suivant immédiatement une séquence, mais à prédire ceux qui apparaîtront ultérieurement dans la séquence [Fahed et al., 2014a, Fahed et al., 2015b]. D'un point de vue exploitation du modèle, il permet de prédire un item dont l'occurrence est éloignée dans le temps, ou, d'un autre point de vue, prédire en avance l'occurrence d'un item. Ce type de prédiction permet donc à la personne ou à l'organisme exploitant ce modèle d'avoir connaissance, par avance, de l'occurrence future d'un item.

6. Définition du dictionnaire Larousse

7. <https://edc.revues.org/6582> - accédé le 06/06/2017

8. www.predictiveanalyticsworld.com/patimes/prescriptive-versus-predictive-analytics-distinction-without-difference/

Cela lui permet également de pouvoir anticiper cette occurrence. Le travail réalisé repose sur la fouille de motifs, il est un des rares travaux de la littérature s'intéressant à la fouille de motifs avec une contrainte de distance ou de temps entre l'antécédent et la conséquence. Il est même le premier à permettre la fouille de tels motifs en maintenant un temps d'exécution faible.

Contrairement aux cadres classiques qui prédisent puis recommandent des items apparaissant dans l'imédiat, ici, les items distants prédits ne sont pas obligatoirement recommandés. Par exemple, lorsque l'occurrence d'un item prédit n'est pas souhaitée par la personne exploitant le modèle, celle-ci peut souhaiter agir de façon à empêcher son occurrence. Etant donné que l'item n'apparaîtra pas dans un horizon proche (à l'opposé d'une prédiction classique), elle aura le temps pour cela.

A la suite de ce travail, j'ai souhaité aller une étape plus loin, je me suis intéressée à l'identification automatique d'items, dits perturbateurs, c'est-à-dire les items qui semblent avoir un impact sur l'occurrence d'items futurs [Fahed et al., 2015a]. Ainsi, sachant un item prédit à un horizon lointain, et un ensemble d'items perturbateurs identifiés, si la perturbation de ces derniers sur les items prédits est celle souhaitée, alors les items perturbateurs sont ceux qui doivent être recommandés. Le système reposant sur cet algorithme peut être vu comme un outil d'aide à la décision. Ce premier travail est lié à la tâche de prescription, il vise à prescrire des items en vue d'un but qui n'est pas directement lié à la prescription effectuée, mais qui est ultérieur. L'objectif visé par ces premiers travaux est de mieux "maîtriser" ou d'"agir" sur le futur.

Mes futurs travaux constituent la suite de ces derniers travaux.

4.2.2 La prescription ou comment mener un utilisateur à un but fixé

Dans les sections précédentes, j'ai introduit la notion de prescription. L'analyse prescriptive (*prescriptive analytics*) [Ransbotham et al., 2015] permet de déterminer les prescriptions adéquates pour un problème ou une situation donnés. L'analyse prescriptive va un pas plus loin que l'analyse prédictive (qui vise à prédire ce qui va se passer), dans le sens où elle cherche à déterminer ce qui devrait être fait (ce qui sera prescrit), sachant ce qui devrait se produire (ce qui est prédit). Pour pouvoir effectuer une prescription, l'analyse prescriptive a donc besoin non seulement de prédire ce qui va se passer, mais également d'en comprendre les raisons. Elle peut ainsi être vue comme un moyen de planifier le futur ; son objectif est plutôt à moyen et long terme.

Elle peut également être utilisée pour identifier l'impact d'éventuelles décisions sur le futur et pour informer les acteurs concernés en amont de leurs décisions. Dans le même ordre d'idées, elle peut être utilisée pour optimiser une prise de décision, notamment pour indiquer (prescrire) quelles actions mener afin d'atteindre un objectif donné ou afin de maximiser un résultat. Elle est liée aux problèmes d'optimisation et de planification. Elle est donc un moyen de passer de la situation "ce qu'il va se passer" (prédiction) à la situation "ce qu'il devrait se passer" (prescription).

L'analyse prescriptive est donc une réponse possible au défi posé ici. Elle permettra d'identifier la suite d'actions qu'un utilisateur doit effectuer pour atteindre l'objectif souhaité. Cet objectif peut soit être visé par le système, soit être visé par un acteur du dispositif (par exemple l'enseignant dans le cas de l'éducation). La suite d'actions identifiée est donc celle qui sera prescrite, c'est une **séquence de prescriptions** qui sera faite.

Chercher à atteindre un objectif distant donné *a priori*, change le paradigme traditionnel des algorithmes de recommandation. En effet, le problème passe d'une recommandation "en un coup" à une recommandation (prescription) "en plusieurs coups". Dans le premier cas (cadre classique du e-commerce, par exemple), la recommandation doit impérativement être la meilleure recommandation possible au moment où elle est faite, puisque l'objectif est immédiat (court terme). L'objectif doit être atteint au travers de cette seule recommandation : l'utilisateur doit consulter/adopter/acheter/consommer la ressource recommandée, et, bien évidemment, il est attendu que l'utilisateur revienne sur le système plus tard. Dans le second cas, qui nous intéresse ici, l'objectif est à plus long terme. Ainsi, chaque prescription prise isolément, peut être moins "pertinente", puisqu'à elle seule elle ne permet pas d'atteindre l'objectif. C'est la suite de prescriptions qui doit être la plus pertinente, ou du moins maximiser la probabilité d'atteindre l'objectif. Bien évidemment, il est attendu que l'utilisateur ne quitte pas le système avant la fin de la

suite de prescriptions.

Je présente ci-dessous un exemple relatif au e-commerce, l'objectif à atteindre et les prescriptions associées. Supposons que l'objectif général du vendeur (représenté par le système) est de maximiser le nombre d'exemplaires du produit P vendus. Soit un client qui navigue sur le site de vente en ligne. L'analyse prédictive peut prédire 3 situations : 1) l'utilisateur n'achètera probablement pas le produit P , 2) l'utilisateur va, dans quelques temps, fort probablement acheter le produit P , 3) l'utilisateur va, sous peu, fort probablement acheter le produit P . En fonction de ces situations, l'analyse prescriptive pourra être utilisée, afin que le vendeur atteigne son objectif : maximiser ses ventes. Dans la situation 1), l'analyse prescriptive permettra d'identifier la suite de produits ou informations à prescrire au client, produits ou informations qu'il va consulter/acheter, qu'il va aimer ou non, et qui impacteront ses envies, son profil, augmentant, par là même, la probabilité que celui-ci achète P . Chaque élément de la suite de prescriptions, pris isolément, ne permet pas d'atteindre l'objectif, et peut même ne pas satisfaire l'utilisateur. Dans la situation 2), même s'il est prédit que le produit sera acheté par le client, l'analyse prescriptive peut être utilisée pour identifier les produits ou informations à prescrire au client pour rapprocher l'échéance à laquelle P sera acheté. Dans la dernière situation, étant donné que la prédiction à court terme correspond à l'objectif du vendeur, l'analyse prescriptive n'est pas utile.

Définition du problème

D'une façon plus formelle, le problème sur lequel je souhaite travailler peut être défini de la façon suivante : soit S l'ensemble des états possibles, S_i un état initial (l'état courant), S_g un état but à atteindre. $A = \{a_1, \dots, a_n\}$ est un ensemble prédéfini d'actions possibles, applicables à un état et menant à un nouvel état $f : S \times A \rightarrow S$. La question est : quelle(s) suite(s) d'actions (quel(s) chemin(s)) permet(tent) d'atteindre S_g en partant de S_i ?

Dans mes travaux, l'état initial S_i correspond au profil de l'utilisateur, S_g est le profil visé pour cet utilisateur et A est l'ensemble des items accessibles à un utilisateur (au travers d'achats, consommation, etc. en fonction du domaine d'application). Le problème est donc le suivant : **quelle suite d'items prescrire à l'utilisateur** (items qu'il consultera, achètera, consommera, etc.) **afin que celui-ci atteigne l'objectif fixé ?**

L'analyse prescriptive est généralement abordée comme un problème d'optimisation dans la littérature [Gröger et al., 2014, Souza, 2014]. Sachant les données du problème : connaissances sur l'état courant, connaissances externes, connaissances du passé, but visé, contraintes imposées par le problème, critère de qualité, etc. l'objectif est d'identifier la solution optimale (chemin optimal) à ce problème.

L'analyse prescriptive à destination d'utilisateurs

La littérature considère l'analyse prescriptive comme un des défis principaux des années à venir [Evans and Lindner, 2012, Soltanpoor and Sellis, 2016], particulièrement dans le domaine des affaires où des premiers travaux ont déjà été menés récemment, mais à destination des décideurs majoritairement [Appelbaum et al., 2017].

De mon point de vue, la conception d'un modèle de prescription à destination de décideurs est une tâche différente de la conception d'un modèle de prescription à destination d'utilisateurs (cadre d'application auquel je m'intéresse). En effet, les décideurs sont en général conscients de l'objectif à atteindre, et leur intérêt est même de maximiser la probabilité de l'atteindre effectivement. Ils suivent donc généralement scrupuleusement les prescriptions qui leur sont faites. Ce n'est pas le cas des utilisateurs, qui :

- ne sont pas toujours conscients de l'objectif visé,
- ne sont même pas forcément motivés pour atteindre cet objectif.

Par conséquent, les utilisateurs peuvent décider de ne pas suivre les prescriptions, voire de quitter le système si les prescriptions ne leur conviennent pas.

Notons que concevoir un modèle de prescription à destination d'un système est plus simple. Un système, un robot par exemple, ne remet pas (jamais ?) en cause une prescription qui lui est faite, il

l'exécutera toujours. Un décideur, même motivé et conscient de l'objectif, peut cependant décider de ne pas suivre une prescription si celle-ci est trop contraignante pour lui.

Par conséquent, la conception de modèles de prescription pour des utilisateurs, même si ceux-ci sont motivés et connaissent l'objectif, est soumise à des contraintes, que je présente ci-après.

Les contraintes sur les modèles et les algorithmes

Etant donné que la séquence de prescriptions composant la solution est destinée à un utilisateur, libre de quitter le système lorsqu'il le souhaite, l'algorithme doit chercher à minimiser ce risque de départ. Cela a un impact sur **la nature de la séquence de prescriptions** :

- *Cohérence de la séquence.* Plus la séquence de prescriptions est cohérente aux yeux de l'utilisateur, plus il la comprend, et plus il devrait accepter et adopter chacune des prescriptions. A l'opposé, si l'utilisateur ne comprend pas la séquence ou sa logique, il pensera que le système fait "n'importe quoi", il est donc fort probable que celui-ci quitte le système, ne comprenant pas ce vers quoi le système le mène.

Cette cohérence peut, par ailleurs, avoir un impact sur la confiance et l'acceptation du système par l'utilisateur.

La cohérence de la séquence prescrite est donc une contrainte que devra respecter l'algorithme dans sa stratégie de formation d'une séquence de prescriptions. La question qui se pose donc est qu'est-ce qu'une séquence cohérente pour un utilisateur ?

- *Optimalité de la séquence.* Le critère utilisé pour déterminer la séquence à prescrire (la séquence optimale) doit prendre en compte le fait que celle-ci est dédiée à un utilisateur. Le nombre d'éléments composant le chemin, qui est un critère classique des algorithmes d'optimisation, n'est pas forcément le critère à utiliser lorsque la séquence est destinée à un utilisateur. Un tel chemin peut, en effet, soit paraître incohérent au sens de l'utilisateur, soit être complètement inadapté à ce dernier. Par ailleurs, chaque utilisateur peut avoir son propre critère d'optimalité, fonction de son profil ou de ses spécificités. Ce critère devrait donc être automatiquement identifié par le système, pour chaque utilisateur. D'une façon générale, ce critère doit être celui qui maximise la probabilité que l'utilisateur atteigne le but.
- *Une séquence d'ensembles.* Même si le système vise à guider l'utilisateur vers un but, dans certains cas comme par exemple en e-commerce, l'utilisateur peut vouloir se sentir libre de ses choix, ne pas se sentir trop restreint, au risque qu'il quitte le système. Dans ce cas, les prescriptions faites ne devraient pas être composées d'une unique séquence d'items, mais d'une séquence d'ensembles d'items (ou d'un ensemble de séquences d'items).

Une prescription étant plus associée à un ordre qu'à un simple conseil, s'assurer que chaque prescription (ou chaque ensemble de prescriptions) de la séquence de prescriptions soit suivie par l'utilisateur, a un impact sur **la façon dont la séquence est présentée**.

- *Justification des prescriptions.* Si chaque prescription peut être justifiée, argumentée, expliquée, ou tout du moins lorsque l'utilisateur en ressent le besoin, elle devrait être comprise par celui-ci. De cette façon, celui-ci sera plus enclin à suivre chacune d'elles.

La justification d'une prescription peut également être vue comme un moyen d'augmenter la confiance de l'utilisateur en le système.

Notons que, dans les systèmes de recommandation, il est maintenant largement reconnu que pouvoir justifier la raison d'une recommandation est une réelle plus-value [Tintarev and Masthoff, 2007, Cleger et al., 2014] car l'utilisateur accueille celle-ci avec bienveillance et a plus confiance en cette dernière. Dans le cadre de la prescription d'une séquence d'items, l'explication est d'autant plus primordiale qu'il est important que l'utilisateur accepte non pas une seule prescription, mais chacune des prescriptions de la séquence. Par ailleurs, comme mentionné précédemment, les utilisateurs n'ont pas toujours conscience de l'objectif à atteindre, celui visé par le système. La justification sera d'autant plus utile.

Une autre contrainte liée aux algorithmes est donc qu'ils doivent être capables de justifier les prescriptions faites, soit en lien avec la séquence complète de prescriptions, soit en lien avec le but, si ce dernier est connu de l'utilisateur. Cet aspect rentre dans le cadre des systèmes persuasifs

[Yoo et al., 2012].

D'une façon plus générale, l'explication/la justification des prescriptions a pour but de persuader l'utilisateur. Cependant, la question de savoir comment le système peut stimuler ou persuader les utilisateurs de réaliser certaines actions, est un domaine encore sous-étudié à l'heure actuelle [Jannach et al., 2016].

- *Moment de la prescription.* Une surcharge de prescriptions, ou une prescription faite à un mauvais moment, peut avoir un impact négatif sur l'utilisateur et sur sa relation avec le système. L'algorithme doit donc pouvoir identifier automatiquement le/les moment(s) où l'utilisateur a besoin de prescriptions, ou les moments où l'utilisateur est enclin à recevoir des prescriptions. Evidemment, chaque utilisateur étant unique, le moment opportun devra être automatiquement appris.

Et si, malgré tout, l'utilisateur ne choisit pas la/les prescriptions ? Malgré les incitations fortes à adopter les items qui lui sont prescrits, un utilisateur peut tout de même choisir de se tourner vers d'autres items, sans pour autant quitter le système. Ceci a deux conséquences sur l'algorithme :

- il doit pouvoir être ré-exécuté et la séquence recalculée à chaque item choisi et non prescrit, donc potentiellement à chaque action faite par l'utilisateur. L'algorithme doit donc pouvoir s'exécuter en temps-réel.
- Comme mentionné précédemment, la cohérence de la suite des prescriptions doit continuer à être respectée. La nouvelle séquence formée par l'algorithme doit donc être non seulement cohérente mais également cohérente avec la séquence précédemment proposée à l'utilisateur. Elle doit également être en cohérence avec les items que l'utilisateur a choisis, plutôt qu'avec ceux prescrits auparavant.

Une question qui peut également être traitée en parallèle est de pouvoir identifier la raison pour laquelle la prescription qui a été faite n'a pas été choisie. Une boucle de retour sur l'algorithme pourra ainsi être mise en place.

Et si l'état de l'utilisateur n'est pas celui escompté ? Dans le cas où l'utilisateur choisit de ne pas consulter un item prescrit, ou que l'état de l'utilisateur après une consultation n'est pas celui attendu, il est important de pouvoir non seulement évaluer la "distance" entre cet état et celui prévu par l'algorithme, mais également la distance entre la séquence prévue et celle effectivement suivie par l'utilisateur. En exploitant ces informations, le système doit pouvoir déterminer dans quelle mesure cette distance impacte ou non le chemin qui avait été prévu pour l'utilisateur. En d'autres mots, l'utilisateur est-il en train de dériver significativement de ce qui avait été prévu, auquel cas le système doit-il agir ou non (proposer d'autres prescriptions, modifier des paramètres) ?

En résumé, si un utilisateur se voit prescrire une séquence qui lui semble incohérente, qu'il ne comprend pas, ou, qui plus est, ne lui est pas expliquée, il est probable que, soit il ne la suive pas, soit il quitte le service.

D'autres considérations sont à prendre compte du fait qu'un utilisateur soit le destinataire des prescriptions. Ces considérations n'ont pas d'impact sur l'algorithme en lui-même, mais sur son utilisation. Elles sont plus du ressort de l'interaction entre le système et l'utilisateur :

- Le système doit-il afficher, ou non, la distance de l'utilisateur au but ? Cela peut être un moyen d'augmenter les chances pour l'utilisateur d'atteindre le but, en le motivant. Au contraire, le système peut avoir intérêt à ne pas montrer à l'utilisateur sa distance au but, de peur de le décourager.
- Le système doit-il recommander/présenter à l'utilisateur la séquence de prescriptions dans son intégralité ? Doit-il la lui faire découvrir progressivement ?

4.2.3 Caractéristiques et défis apportés par le cadre applicatif

Dans le cadre de l'éducation, un utilisateur est un étudiant et l'objectif visé est généralement un niveau de connaissances donné.

Dans ce cadre, la question que je me pose, en lien avec le défi traité dans cette section, est la suivante : soit un étudiant dont on connaît le profil et un objectif à atteindre, quelle est la séquence de ressources/activités pédagogiques (cours, leçons, exercices, quizz, etc.) qui doit être prescrite à l'étudiant

pour garantir, ou au moins augmenter ses chances, qu'il atteigne l'objectif visé ? L'approche que j'adopte repose sur les traces collectées sur les étudiants et sur leurs activités, complétées par toute information collectable en lien avec l'objectif (données des systèmes d'information des institutions par exemple). Elle entre dans le cadre des *learning analytics*.

L'analyse prédictive a beaucoup été étudiée en *learning analytics*, notamment dans le but de prédire l'état futur d'un étudiant. Le plus souvent, c'est la prédiction de l'échec d'un étudiant qui est recherchée [McQuiggan and Sapp, 2014]. A l'opposé, l'analyse prescriptive n'a, pour le moment, été que très peu étudiée, bien qu'elle soit de plus en plus souvent mentionnée dans la littérature [Van Barneveld et al., 2012, Rajni and Malaya, 2015, Daniel, 2015, Wilson et al., 2017, Daniel, 2017].

Parmi les caractéristiques de l'éducation dans un cadre numérique, on peut mentionner que :

- le profil de l'étudiant (utilisateur) est composé de nombreuses caractéristiques, qui le rendent encore plus unique que dans un cadre de e-commerce : son parcours d'apprentissage, l'état actuel de ses connaissances, ses caractéristiques, sa façon de travailler, sa dynamique, s'il préfère travailler sur des concepts abstraits ou détaillés, sur des exemples, sur des exercices, sur des études de cas, etc. Ainsi, même deux étudiants d'une même classe peuvent avoir des profils radicalement différents car, bien que partageant *a priori* le même objectif, ils peuvent avoir des caractéristiques, des acquis et des expériences différents, de même qu'une motivation et une attente différentes.
- L'objectif à atteindre peut ne pas simplement correspondre à un niveau de connaissances, il peut également être associé à une date, correspondant par exemple au jour de l'examen. Dans ce cas, il est primordial que l'objectif soit atteint pour le jour de l'examen ou juste avant. En effet, en aucun cas on ne doit prendre le risque qu'il soit atteint après l'examen. Par ailleurs, si un étudiant atteint le but trop longtemps à l'avance, ses connaissances peuvent diminuer jusqu'au moment de l'examen.
- Il peut exister un chemin d'apprentissage-type, préconisé par l'enseignant. Celui-ci peut être exploité par l'algorithme de prescription, en tant que connaissance supplémentaire. Chaque étudiant étant unique, le meilleur chemin pour un étudiant n'est probablement pas celui préconisé par l'enseignant, mais dérivé de celui-ci, en fonction de son profil.
- En éducation, le bénéfice d'une prescription n'est mesurable ni directement, ni immédiatement. A l'opposé, en e-commerce, le fait qu'un utilisateur clique sur un item recommandé peut signifier que celui-ci l'intéresse, mais si l'item est acheté, alors l'intérêt de l'utilisateur pour cet item est confirmé, de même que la qualité de la recommandation (et donc du modèle). En éducation, ce n'est pas forcément l'intérêt de l'étudiant pour la ressource qui doit être mesuré. C'est plutôt l'impact de la ressource sur le niveau de connaissances de l'étudiant par rapport à l'objectif fixé, voire l'enrichissement de la connaissance du profil étudiant par le modèle. On peut imaginer que si l'item recommandé est un exercice, alors le niveau de connaissances peut être estimé grâce au résultat de l'étudiant sur l'exercice. Mais qu'en est-il si la recommandation est un cours ? La question est donc : comment évaluer le bénéfice d'une ressource, à partir des traces et/ou grâce aux futures prescriptions, dont certaines peuvent être faites dans le but d'obtenir une information, et non pas simplement d'améliorer la connaissance de l'étudiant ?
- L'objectif est-il unique ? En éducation, il peut y avoir plusieurs objectifs à atteindre, objectifs qui peuvent être liés ou non, et dont les prescriptions peuvent également être liées ou non. Par exemple, pour réussir une année universitaire complète, chaque examen (matière) doit être réussi (ou la moyenne des examens), chacun peut ainsi être vu comme un sous-objectif. Les liens entre ces derniers (redondance par exemple) et entre les prescriptions faites à chacun d'eux, doivent être gérés par le système. Et si les objectifs sont contradictoires ?
- Les étudiants peuvent travailler collaborativement, dans ce cas les prescriptions peuvent être des prescriptions de camarades, et non pas d'items. Par ailleurs, les algorithmes doivent également tenir compte des interactions entre eux, voire faire des prescriptions destination du groupe.
- Le système n'est pas le seul prescripteur, l'enseignant peut l'être également. Dans ce cas, le système doit tenir compte, dans la construction de la séquence de prescription, qu'elle est à destination de l'enseignant, mais concerne l'apprenant. La question qui se pose est donc : comment en tenir compte ?

A l'heure actuelle, peu de travaux en *learning analytics* cherchent à modéliser le chemin d'apprentissage personnalisé à faire suivre aux étudiants pour les amener vers un objectif donné. La majorité des travaux exploitent uniquement le chemin passé (l'historique, le profil de l'étudiant) dans le but de pré-

dire/inférer l'intérêt des utilisateurs pour les ressources, à un horizon "instantané" [Huptych et al., 2017]. C'est pourtant une problématique dont l'importance est incontestable. D'un point de vue applicatif, Ian Dolphin d'Apereo⁹ a mentionné cette problématique comme étant le futur des *learning analytics*, lors de sa présentation au sommet ICDE de mai 2017¹⁰.

Plusieurs approches me semblent intéressantes pour traiter le problème de prescription présenté dans les sections ci-dessus, en particulier la fouille de données et les modèles de prise de décision. Je présente, dans les sections ci-dessous, en quoi ces approches peuvent être un cadre de solution au problème qui nous intéresse.

4.2.4 La prescription vue comme un problème de fouille de données

La fouille de données, et notamment la fouille de données temporelles, a été au cœur de mon activité de recherche ces dernières années. Elle est une approche possible pour l'analyse descriptive et prédictive [Han and Kamber, 2006]. En effet, elle vise à extraire l'information présente dans les données (données de traces, dans notre cas), sous la forme de motifs temporels fréquents, de règles d'association temporelles, de traits implicites (features), de clusters, etc., dans le but de comprendre les données ou de prédire des réalisations futures dans ces dernières.

Je pense que la fouille de données temporelles peut également être une approche adéquate pour l'analyse prescriptive, ce qui fait l'objet de cette section.

Considérer la fouille de données comme une solution à la tâche de prescription revient à considérer cette dernière comme répétitive, c'est-à-dire que :

- la situation courante (l'état courant), qui nécessite une prescription, est une situation connue, elle est présente dans les données. Notons que la même hypothèse est faite lorsque la fouille de données est utilisée pour la prédiction,
- la/les solution(s) (la/les prescription(s)) est/sont présentes dans les données,
- l'association situation et solution est également présente dans les données.

Comme évoqué dans la section 3.3.1, la situation courante/l'état courant peut être représenté par la suite des dernières actions menées.

A ma connaissance, la tâche de prescription n'a été que très rarement abordée avec une approche par fouille de données. C'est par exemple le cas dans le domaine médical, pour prescrire des solutions médicamenteuses [Wright et al., 2015].

Pour construire un modèle de prescription à base de motifs ou de règles temporels, deux questions principales se posent : quel type de motifs ou de règles est adéquat ? quel ensemble de motifs ou de règles doit être conservé ? Je vais donner quelques premiers éléments de réponse à ces deux questions.

Dans la tâche de prédiction, ce sont des règles d'association séquentielles qui sont adéquates. Pour former les règles, les algorithmes traditionnels reposent sur une première étape de fouille de motifs séquentiels fréquents, puis identifient l'item/une séquence d'items/la plus grande sous-séquence d'items postfixant le motif, et qui a une probabilité sachant les autres éléments du motif, supérieure à un seuil donné. Pour identifier les règles applicables dans une situation donnée, c'est l'antécédent des règles qui est considéré, en identifiant ceux qui correspondent à la situation. C'est ensuite l'item (ou les items) composant la conséquence des règles applicables qui sont recommandés à l'utilisateur (voir section 3.3.1 pour plus de détails).

Dans une tâche de prescription, ce sont également des règles séquentielles qu'il me semble judicieux d'exploiter. Cependant, l'objectif n'étant plus de prédire, mais de prescrire, cela a un impact sur les règles. Je considère deux points de vue pour les caractéristiques des règles permettant de répondre à la tâche de prescription.

9. <https://www.apereo.org>

10. <https://psummit2017.wordpress.com>

Premier point de vue : les items à prescrire sont présents dans la conséquence

Je considère ici que les items à prescrire sont une sous-séquence de la conséquence des règles. Dans ce cas, l'étape d'identification des règles applicables est soumise à une contrainte supplémentaire. En effet, les règles applicables sont celles dont, non seulement l'antécédent correspond à la situation courante (comme dans la tâche classique de prédiction), mais également dont un élément de la conséquence correspond au but à atteindre. Il est naturel de penser que ce sont les items de la conséquence, précédant l'item but, qui doivent être prescrits.

Dans le cas où le premier élément de la conséquence est l'item but, aucune prescription n'est nécessaire, puisque l'on est déjà au but.

Cependant, dans les règles séquentielles que fouillent les algorithmes traditionnels, il n'y a *a priori* aucun lien de "déclenchement" entre les items qui composent la conséquence. Ce ne sont pas ceux-ci qui "déclenchent" l'item but de la conséquence. En effet, c'est l'antécédent qui déclenche la conséquence dans sa globalité.

Dans une tâche de prescription, si l'on considère que ce sont les items de la conséquence, précédant l'item but, qui sont ceux à prescrire, alors chaque item de la conséquence devrait avoir une "influence" ou devrait "déclencher" l'item but de la conséquence. Les algorithmes devront donc être adaptés en ce sens, de façon à former des règles avec cette caractéristique supplémentaire.

Second point de vue : les items à prescrire sont présents dans l'antécédent

Je considère ici que les items à prescrire sont une sous-séquence de l'antécédent des règles. Plus précisément, ce sont les derniers items de l'antécédent qui sont à prescrire. Les premiers items composant l'antécédent représentent, quant à eux, la situation courante. Cependant, les algorithmes traditionnels ne permettent pas de former des règles dans lesquelles ces derniers items contribuent effectivement à déclencher la conséquence. De même, ces algorithmes ne sont pas conçus pour évaluer dans quelle mesure ils y contribuent. En effet, les algorithmes classiques de fouille de données ne cherchent pas tous à identifier des règles dont l'antécédent n'est composé que d'items utiles. Par conséquent, il n'y a pas de garantie que prescrire les derniers items d'un antécédent permette d'augmenter la probabilité d'atteindre le but. Les algorithmes devront donc, une fois de plus, être adaptés pour former des règles avec cette caractéristique.

Au vu des deux considérations ci-dessus, nous pouvons dire que les règles fouillées par les algorithmes classiques ne permettent pas de répondre complètement à notre problème. Un nouvel algorithme de fouille de règles devra donc être conçu.

Une nouvelle forme de règles

Je pense notamment à la proposition d'une nouvelle forme de règles. Ces règles pourront être composées, comme les règles classiques, d'*un antécédent* et d'*une conséquence* qui seront exploités pour identifier si une règle est applicable. Elles seront, de plus, composées d'un élément intermédiaire, une *prescription*, qui contiendra la suite d'éléments qui, s'ils sont consultés/achetés/consommés, déclencheront la conséquence. En l'occurrence, ce seront les éléments composant la *prescription* qui seront à prescrire. Une règle aura donc la forme suivante $R : \text{antecedent} \rightarrow \text{prescription} \rightarrow \text{consequence}$.

Ce type de règle permettra de connaître, sachant une situation (l'antécédent) et un but (la conséquence) donnés, la suite d'items à prescrire pour atteindre le but.

Pour concevoir un algorithme permettant de fouiller de telles règles, il faudra redéfinir les notions classiques de support et de confiance et proposer de nouvelles mesures permettant d'évaluer la qualité d'une *prescription*, et notamment de chaque item composant la *prescription*. Je pense, par exemple, à la mesure de gain d'information. Des questions relatives à la formation de ces règles devront être étudiées : l'antécédent doit-il être minimal ? la même question se pose relativement à la prescription : minimale, maximale, toutes les prescriptions ? D'autres critères tels que la longueur, l'homogénéité, l'hétérogénéité, etc. peuvent également être considérés. Ces choix vont potentiellement impacter la couverture associée aux règles (pourcentage de cas où au moins une règle est applicable). La réponse à ces questions dépendra évidemment du cadre applicatif.

Une fois ces règles formées, une question relative à leur utilisation se pose : comment déterminer la suite d'items à prescrire lorsqu'aucune règle ne contient à la fois la situation actuelle et le but, mais uniquement la situation actuelle, ou uniquement le but, voire aucun des deux. Il sera dans ce cas nécessaire de passer par une étape de **composition de règles** et/ou par l'utilisation de mesures de similarité entre items et/ou motifs séquentiels. Il est à noter que le choix des règles à conserver dans l'étape précédente pourra également être impacté par cette possibilité de composition. En effet, si l'on sait qu'*a priori* les règles devront être composées, la contrainte de l'état but dans la conséquence peut-être levée. Par ailleurs, le problème du manque de données et de l'émergence inciteront peut-être à former des règles composées de peu d'éléments, mais avec un support élevé, ou, à l'opposé, de former des règles composées de nombreux éléments mais avec un support faible. Des notions telles que les points de passage obligé (items obligatoires) pour atteindre un but pourront alors être étudiées pour limiter le nombre de règles, mais également augmenter la couverture et la qualité de la prescription. Les problèmes traditionnels d'inconsistance, de confluence, et d'insuffisance [De Bra et al., 2004] rencontrés classiquement dans la gestion de règles devront bien évidemment être traités.

Notons que les approches à base de règles ont, par ailleurs, un autre avantage pour la tâche de prescription. Comme pour la prédiction, elles offrent la possibilité de pouvoir expliquer les sorties des algorithmes : savoir pourquoi tel item (ou telle séquence d'items) est prescrit. Je rappelle que **l'explication des prescriptions** (ou des recommandations) est une caractéristique importante lorsque celles-ci sont destinées à des utilisateurs (ou étudiants).

Spécificités apportées par le cadre de l'éducation

L'exploitation de la fouille de données en éducation constitue un domaine de recherche à part entière, l'*Educational Data Mining*, auquel une société internationale ¹¹, une conférence annuelle ¹² et même une revue ¹³ sont dédiées.

Au début de ce chapitre, j'ai mentionné le fait que le profil d'un étudiant devait à la fois représenter le niveau de connaissances de ce dernier, mais également ses caractéristiques. De fait, plus un profil est complet, meilleure devrait être la qualité de la prescription.

Une première question qui se pose ici est relative à la nature des éléments qui composeront les motifs et les règles fouillés. Ces éléments seront-ils uniquement des identifiants de ressources ? Cette approche semble *a priori* trop simple, puisqu'elle ne permettra pas de modéliser des aspects relatifs au comportement, tels que le temps associé à chaque ressource, ni relatifs à la nature de la ressource : le type, la discipline, la difficulté (ou toute autre meta donnée), etc. ni même aucune information sur le niveau de connaissances des étudiants effectuant la séquence correspondante. Il sera donc vraisemblablement plus approprié de considérer chaque ressource comme un ensemble de descripteurs. Ainsi, les motifs et règles fouillés seront composés de suites d'ensembles de descripteurs, de façon à représenter au mieux, à la fois les connaissances et les caractéristiques. Par ailleurs, ces motifs pourraient également contenir d'autres informations, relatives aux caractéristiques ou au niveau des étudiants. Ces informations peuvent être présentes, soit en tant que meta-données associées aux motifs, soit en tant qu'éléments présents dans les motifs. Ils permettront, ainsi, d'affiner la modélisation. Dans ce deuxième cas, bien évidemment, les éléments supplémentaires ne sont pas des éléments à prescrire, mais aideront à identifier les règles applicables. Pour permettre de former de telles règles, il pourra être intéressant de s'intéresser aux travaux menés dans la fouille de règles granulaires, qui permettent de former des règles provenant de multiples sources [Min and Zhu, 2013].

Une autre façon d'aborder le problème serait de dissocier l'état de connaissances d'un étudiant de ce qui caractérise ce dernier. C'est l'état de connaissances qui serait utilisé pour identifier les règles applicables. Dans ce cas, l'antécédent n'est pas composé de suite de ressources, mais de la description d'un état de connaissances. Ensuite, les caractéristiques de l'étudiant seront utilisées pour adapter la suite de prescriptions. Celle-ci nécessitera également de définir des classes d'équivalence de caractéristiques ou d'ensembles de caractéristiques (profil-type de caractéristiques).

11. <http://www.educationaldatamining.org>

12. <http://educationaldatamining.org/EDM2017/>

13. <http://www.educationaldatamining.org/JEDM/index.php/JEDM>

Par ailleurs, la définition du but devra être repensée pour l'éducation. Le but étant un ensemble de connaissances plutôt qu'un item, comment représenter le but de façon la plus précise ? La réponse à cette question sera liée au paragraphe précédent, pour lequel l'identification de l'antécédent devait également refléter un état de connaissances.

Un dernier aspect me semble primordial dans le cadre éducatif : la réussite de l'étudiant. Les règles fouillées et exploitées pour la prescription sont celles qui devraient amener à la réussite de l'étudiant. Je souhaite étudier l'apport de l'analyse par contraste pour identifier les règles permettant effectivement de favoriser la réussite des étudiants. J'ai déjà étudié cette approche dans le cadre de la thèse de Marharyta Aleksandrova, mais dans un contexte non séquentiel, non temporel. Ici cette dimension est au cœur du problème. Dans cette analyse par contraste, deux jeux de données seront exploités : les données relatives aux étudiants qui réussissent et celles relatives aux étudiants qui échouent. Cette approche permettra notamment d'identifier les éléments (items, caractéristiques, séquences) qui impactent la réussite d'un étudiant.

Dans la section précédente, j'ai évoqué la tâche de composition de règles. Ici, les items sont représentés par un ensemble d'attributs, dont certains peuvent être associés à des concepts. Ainsi, l'approche par règles pourra permettre d'identifier des suites de concepts à connaître pour atteindre un but donné. Notons que ces suites pourront être présentées et validées par des experts. Ensuite, ces règles pourront être affinées pour descendre au niveau des ressources, pour déterminer la/l'ensemble de suites de ressources à suivre pour atteindre un concept donné, comme proposé dans [Karamperis and Sampson, 2005].

4.2.5 La prescription vue comme un problème de prise de décision

Définir la suite d'items à prescrire à un utilisateur pour lui permettre d'atteindre un but donné peut également être vu comme un problème de prise de décision. Plus précisément, c'est un problème de prise de décision séquentielle, un problème de planification. C'est ce second point de vue qui va être considéré dans cette section.

La décision à prendre (la prescription) dépend de la situation dans laquelle l'utilisateur est (l'état de l'utilisateur) et, sous certaines conditions, peut dépendre uniquement de celle-ci. Dans ce dernier cas, le problème de prescription satisfait donc la propriété de Markov. Par ailleurs, dans le cadre des applications qui m'intéressent, par exemple le e-commerce ou l'éducation, l'impact d'une action sur un état n'est pas parfaitement connu *a priori*. En effet, il est impossible de savoir *a priori* si un utilisateur souhaitera découvrir d'autres livres de science-fiction après avoir lu un livre précis, ni si sa compréhension d'une notion sera augmentée après avoir étudié tel cours. Il est cependant possible d'estimer la distribution de probabilités des différentes possibilités (états) à l'issue d'une action. La prescription pourra donc être abordée comme un processus de décision Markovien (MDP).

Pour aborder la prescription comme un MDP, il est indispensable de connaître la liste des états possibles (les situations), la liste des actions, les pré-conditions de chacune et leur effet (leurs effets possibles), c'est-à-dire les différents états résultants (et la distribution de probabilité associée). Par ailleurs, le coût des actions (ou les récompenses), l'éventuelle durabilité de leur effet et le critère à optimiser doivent également être connus.

Dans le cas de la prescription à destination d'utilisateurs, il est courant que l'état de l'utilisateur ne soit pas parfaitement connu, ni de façon sûre : a-t-il vraiment aimé tel livre ? dans l'achat de tel appareil photo, est-ce la résolution ou le zoom qui a fait pencher son choix ? maîtrise-t-il vraiment telle notion ? De même, suite à l'application d'une action, seule une observation partielle de l'état est disponible. Il y a donc une incertitude quant à l'état effectif d'un utilisateur. Par conséquent, **le problème de la prescription à un utilisateur est plutôt apparenté à un processus de décision markovien partiellement observable (POMDP).**

Les différents cadres d'application qui m'intéressent correspondent à des **problèmes dits de grande taille**. En effet, le nombre d'actions possibles (nombre d'items) est grand, voire très grand. Par ailleurs, bien que la définition d'un état ne soit pas complètement spécifiée pour le moment, il semble que l'on s'oriente également vers un très grand nombre d'états.

Pour aborder un problème avec un formalisme POMDP, deux étapes sont nécessaires : 1) la construction du POMDP (le modèle), c'est-à-dire définir ce qu'est un état, établir la liste des états possibles,

des actions, des transitions entre états, des récompenses, etc. 2) la résolution du POMDP, c'est-à-dire déterminer quelle action (prescription) est optimale pour un état donné.

Construction du POMDP

En ce qui concerne l'étape de construction du POMDP, dans les cadres d'applications auxquels je m'intéresse, il est impossible de spécifier *a priori* la fonction de transition (quels états sont possibles suite à l'application d'une action sur un état donné, et avec quelle probabilité). Le POMDP devra donc être appris automatiquement à partir de traces de réalisation [Buffet and Sigaud, 2008]. Si une connaissance experte du domaine, même partielle, est disponible, elle pourra également être exploitée lors de l'apprentissage du POMDP. Apprendre le POMDP à partir des traces de réalisation, par définition finies et partielles, peut avoir la conséquence que certaines combinaisons d'états et actions peuvent ne pas être présentes dans les traces, rendant ainsi certains états partiellement voire non atteignables. Des solutions devront donc être proposées pour pallier ce problème.

Plusieurs définitions d'états sont envisageables, comme c'était le cas dans l'approche fouille de données. Intuitivement, plus un état contiendra d'informations, meilleure devrait être la qualité du POMDP. De même, plusieurs types de récompenses sont envisageables. Des évaluations expérimentales de chacun de ces deux éléments devront donc être conduites, ce qui permettra notamment de déterminer quels attributs ne sont pas importants dans la définition de l'état, ou de déterminer si une définition d'un état est suffisamment précise.

Résolution du POMDP

L'étape de résolution du POMDP est réputée très complexe, notamment pour les problèmes de grande taille. Pour résoudre un modèle, deux approches sont possibles : la résolution *offline* et la résolution *online*. Dans la résolution *offline*, la politique optimale est apprise en amont de toute exécution, et vise à identifier la politique optimale dans son ensemble. Cette dernière pourra ensuite être exploitée dans un grand nombre de situations (quel que soit l'état initial par exemple). Des solutions ont été proposées pour réduire la complexité inhérente à cette approche. Une approche a retenu mon attention, ce sont les POMDP hiérarchiques, qui, gardant la même approche, visent à réduire *a priori* la complexité du problème. Ils reposent sur la connaissance de la structure du domaine, et décomposent le problème soit en sous-tâches, soit en sous-ensembles (cohérents) d'actions [Toussaint et al., 2008]. Un sous-ensemble d'actions peut même correspondre à une unique sous-tâche. Ainsi, chaque "sous-POMDP" est moins complexe. C'est cette décomposition qui est appelée hiérarchie. De mon point de vue, dans le cas de la décomposition en sous-tâches, les différentes sous-tâches peuvent être vues comme des points de "passage obligé" pour atteindre l'objectif, qui pourront d'ailleurs être appris automatiquement, s'ils ne sont pas connus *a priori*. En éducation, ces points peuvent même être associés à des concepts à apprendre. On peut imaginer que les buts intermédiaires correspondent à des leçons définies par les enseignants ou à des notions définies comme importantes. Dans ces cas ils sont connus *a priori*. Notons que, par ailleurs, ces points de passage obligé, ou concepts, pourront être un moyen de faciliter l'explication des prescriptions présentées aux utilisateurs.

Dans la résolution *online* [Ross et al., 2008], la politique optimale est apprise au moment où une action est requise, et apprise pour la situation courante (à l'opposé de la résolution *offline* qui recherche une politique optimale pour l'ensemble des situations). La résolution *online* a été conçue pour résoudre les problèmes de grande dimension. Dans le cas où le temps imparti n'a pas permis de trouver la politique optimale, elle a l'avantage de permettre de fournir une politique sous-optimale. L'avantage que je vois à cette résolution provient de sa capacité à calculer une politique en temps-réel. En effet, comme mentionné à plusieurs reprises, non seulement les données sur lesquelles je travaille sont véloces, mais, par ailleurs, elles évoluent (apparition de nouveaux phénomènes). L'environnement est donc non stationnaire. La politique optimale sera donc identifiée au moment souhaité, en exploitant, entre autres, les dernières données, ce qui devrait permettre d'améliorer la qualité des prescriptions. Par ailleurs, en fonction du profil de l'utilisateur, on peut imaginer que les récompenses que le modèle doit exploiter peuvent être différentes. Elles peuvent être fonction des caractéristiques de l'utilisateur. La résolution pourra en tenir compte.

Pour toutes ces raisons, la résolution *online* me semble donc la plus adéquate pour la tâche de prescription à un utilisateur.

Rappelons qu'un utilisateur peut décider de faire une action qui n'est pas celle qui lui a été prescrite. Dans ce cas, le POMDP doit tout de même être capable d'estimer l'état résultant, en fonction de l'observation suivante. Deux cas se présentent : 1) l'action choisie par l'utilisateur suite à l'état précédent (estimé) a été rencontrée dans les données d'apprentissage. Dans ce cas, même si cette action ne fait pas partie de la politique optimale, elle est connue du POMDP. Par conséquent, la distribution de probabilités sur les états suivants est évaluée naturellement. 2) l'action choisie par l'utilisateur suite à l'état estimé, n'a jamais été rencontrée dans les données d'apprentissage (soit en raison d'un manque de données d'apprentissage, soit parce que l'action est nouvelle dans cet état). Dans ce cas, même si l'observation est disponible, il est impossible d'estimer la distribution de probabilité des états possibles. Une question qui se pose est donc la suivante : comment estimer l'état suite à l'action effectuée par l'utilisateur ? Une première solution, simple, consiste à exploiter la similarité entre ressources (actions), et considérer que l'action effectivement faite est celle la plus similaire pour laquelle la distribution de probabilité sur les états suivants est connue (qui a donc été rencontrée dans les données d'apprentissage). De même, la similarité entre états peut être exploitée.

Pour limiter le nombre de situations comme celles présentées ci-dessus, il est envisageable de fournir des prescriptions aux utilisateurs qui permettent d'apporter une information supplémentaire et importante au modèle, qui pourra ensuite être intégrée dans celui-ci. C'est le paradigme exploitation/exploration [Cai et al., 2009].

Spécificités apportées par le cadre de l'éducation

La prescription en éducation a déjà été abordée comme un problème de prise de décision, notamment dans [Durand et al., 2011] qui aborde cet aspect comme un MDP et non pas comme un POMDP. Ce choix a été justifié par le fait que la prescription ou la planification est à destination d'enseignants et non pas d'étudiants. Je pense, cependant, que même lorsque la prescription est à destination d'un enseignant, l'observation est également partielle, un POMDP se justifie pleinement.

La modélisation de l'apprentissage avec un formalisme POMDP, dans le but de prescrire des ressources, ou des séquences de ressources à des étudiants, a été très peu étudiée dans la littérature. Récemment, [Rafferty et al., 2015] s'est intéressé à ce problème. Dans ce travail, le POMDP est un POMDP de concepts pédagogiques, et seul l'état des connaissances est géré. Aucune information propre aux caractéristiques de l'étudiant n'est considérée, limitant, de mon point de vue, la qualité de personnalisation, et donc de la prescription. Par ailleurs, le POMDP est considéré connu *a priori*, ce qui est impossible dans le cas que je souhaite traiter. Ce travail constitue cependant une première base pour mes recherches futures. Récemment, l'équipe d'Emma Brunskill de Stanford, spécialisée dans les POMDP, a commencé à s'intéresser à l'éducation comme cadre applicatif.

En éducation, la définition d'un état est au cœur du problème, comme mentionné précédemment. C'est l'état qui permet de déterminer la situation courante. Dans notre cas, l'état concerne principalement l'étudiant. Le nombre d'attributs à intégrer dans l'état est potentiellement très grand : niveau de connaissances de l'étudiant pour chaque concept, habitudes de travail, caractéristiques d'apprentissage (apprend vite, lentement), nature des ressources préférées, terminal utilisé, etc. Ce sont ces mêmes attributs et cette même définition d'un état qui peuvent être utilisés pour l'approche par fouille de données. Pour certains de ces attributs, il n'est cependant pas possible de savoir *a priori* s'ils sont pertinents ou non dans la résolution du problème. Notons que l'on peut imaginer déterminer la pertinence d'un attribut *a posteriori*, ce qui permettra potentiellement de diminuer le nombre d'états, mais également de prendre en compte l'évolution des données sur lesquelles le POMDP est appris.

Dans un formalisme POMDP, étant donné que, par définition, l'action à exécuter ne doit dépendre que de l'état courant, il peut être intéressant d'ajouter des attributs à la définition de l'état : la liste des ressources précédemment consultées avec les dates associées, les résultats, l'évolution du niveau, etc. De même, ici, il n'est pas possible de savoir *a priori* quels attributs sont utiles. Ces très nombreux attributs que doit comporter un état vont, par conséquent, faire exploser le nombre d'états possibles.

Pour limiter la complexité associée, et sur le même principe que les POMDP hiérarchiques de l'état de

l'art, qui définissent des sous-ensembles d'actions ou des sous-ensembles de tâches, je pense qu'une première solution pourrait reposer sur la définition de sous-ensembles d'états. Les états qui seront regroupés dans le même sous-ensemble pourraient correspondre à des états proches (similaires). Par exemple, deux états ne se différenciant que par une ressource dans la liste des ressources précédemment étudiées ou par la nature des ressources préférées, peuvent être considérés comme proches. Ainsi, regrouper les états proches en "meta-états" permettrait de diminuer significativement le nombre d'états et permettrait, dans un premier temps, de déterminer une politique "gros grain". Dans un second temps, l'introduction des états "détaillés" permettrait de représenter les transitions correspondant à ces états détaillés, mais en lien avec les meta-états. Les questions relatives à la détermination de mesures de similarité entre états, le nombre de meta-états, les attributs les plus importants, etc. devront donc être traitées.

Dans le cas spécifique de l'éducation, certaines actions (consultation de certaines ressources) n'ont aucun impact sur l'état effectif de l'étudiant. Ce peut être, par exemple, le cas de la résolution d'un exercice, qui n'a pas d'impact sur l'état des connaissances d'un étudiant. Ces actions permettent cependant au système d'obtenir une information supplémentaire ou de préciser/confirmer une information. Par conséquent, suite à ces actions, seule la distribution de probabilités de chaque état du modèle sera impactée (la fonction de croyance).

La définition de la récompense, utilisée pour la recherche de la politique optimale, devra être définie de façon précise. Une approche simple consistera à définir une récompense positive, si l'état atteint est d'un niveau (de connaissances) supérieur de l'état dans lequel le système était. C'est d'ailleurs le choix fait par [Rafferty et al., 2015]. Des récompenses plus évoluées devront cependant être étudiées.

J'ai mentionné précédemment que, dans certains cas, l'objectif à atteindre pouvait être associé à un temps donné (la date de l'examen, par exemple). Aussi, le temps restant jusqu'à cet objectif à atteindre est une dimension dont il faut pouvoir tenir compte. Par exemple, s'il reste 6 mois à un étudiant avant son examen, les ressources qui devraient lui être prescrites ne seront évidemment pas les mêmes que si il lui reste 6 jours. Une approche classique consiste à intégrer le temps dans la définition des états, en tant qu'attribut supplémentaire. Pour des raisons de simplicité, il est préférable de discrétiser cette dimension. Le pas de temps utilisé devra par conséquent être étudié. Par ailleurs, des POMDP à horizon fini peuvent également être étudiés [Dujardin et al., 2015].

En résumé, le formalisme POMDP semble être un formalisme adéquat pour le problème de la prescription à des utilisateurs, et notamment dans le cadre applicatif qu'est l'éducation. Cependant, de nombreux défis ont été identifiés, liés à plusieurs éléments :

- la construction du POMDP repose sur des traces de réalisation finies et partielles,
- la définition d'un état qui peut potentiellement contenir un très grand nombre d'attributs,
- la définition de la récompense peut être dépendante de l'utilisateur,
- le problème est un problème de grande taille,
- les utilisateurs peuvent décider de ne pas suivre les actions qui leur sont prescrites,
- les données de réalisations sont véloces et évolutives, etc.

4.3 Un système de qualité pour tous et en permanence

En plus de chercher à guider des utilisateurs vers un but donné, une autre dimension de mon projet vise à pouvoir proposer des prescriptions (ou tout autre service) de qualité à chaque utilisateur du système, quelles que soient ses caractéristiques. Par ailleurs, les données étant véloces et évolutives, je souhaite pouvoir prendre en compte, à tout moment, non seulement les dernières données, mais également l'évolution des caractéristiques de ces dernières, toujours dans l'objectif de fournir un service de qualité.

4.3.1 De la particularité des utilisateurs à des modèles particuliers

Même si chaque utilisateur est unique, en règle générale, des cohérences entre utilisateurs peuvent être identifiées. C'est l'hypothèse sur laquelle repose l'approche collaborative de la recommandation, et les dif-

férentes approches que j’ai proposées pour la prescription. Elle permet de former des groupes d’utilisateurs de profils cohérents (identification de voisinage, clustering), d’identifier des profils-type d’utilisateurs, d’identifier des utilisateurs représentatifs (voir les travaux menés section 3.1.2 et [Boumaza and Brun, 2012b, Esslimani et al., 2013, Aleksandrova et al., 2017a]), etc. à partir desquels des recommandations ou des prescriptions peuvent être établies pour un utilisateur donné. Toutes ces tâches reposent sur la définition et sur l’utilisation de la similarité entre utilisateurs.

Dans mes travaux passés, j’ai travaillé sur la tâche de recommandation pour deux types d’utilisateurs particuliers : les utilisateurs en situation de démarrage à froid, voir section 3.2 et [Esslimani et al., 2011, Aleksandrova et al., 2017a] et les utilisateurs “moutons gris”, voir section 3.1.3 et [Gras et al., 2016]. Dans ces deux cas, la similarité traditionnelle ne pouvait être exploitée et l’identification des spécificités de ces utilisateurs a été au cœur du problème. Ces spécificités m’ont permis à la fois d’identifier les utilisateurs de façon fiable, mais aussi de les modéliser pour leur proposer des recommandations de “bonne” qualité. Ces travaux ont été réalisés avec pour objectif la modélisation de préférences.

Je souhaite à nouveau travailler sur ces deux types d’utilisateurs, mais également sur d’autres types d’utilisateurs spécifiques. Cependant, ces futurs travaux porteront sur des données de comportement. Les **données de comportement** (données implicites) ont la caractéristique d’être à la fois bruitées et hétérogènes (c’est notamment le cas dans le cadre des *learning analytics*). Ces caractéristiques introduisent une première difficulté relative à la définition de la similarité de comportement. Par ailleurs, la variabilité naturelle entre utilisateurs accroît plus encore la difficulté de cette définition. Enfin, la synonymie des items et les échelles de temps potentiellement différentes entre utilisateurs et entre sources de données sont une difficulté supplémentaire. L’objectif sera donc non seulement d’identifier la/les particularité(s) de chaque utilisateur, de déterminer l’importance de cette/ces particularité(s) dans le modèle de l’utilisateur, dans quelle mesure cela impacte sa proximité aux autres utilisateurs, mais aussi dans quelle mesure cela doit impacter les prescriptions qui lui sont faites.

Dans les approches statistiques, reposant sur des traces, les modèles de recommandation cherchent toujours à trouver des cohérences dans les données, et ce sont ces cohérences qui sont présentes dans les modèles et exploitées pour déterminer les recommandations. L’approche que je souhaite étudier reposera, à l’opposé, sur la modélisation explicite des spécificités des utilisateurs, de leur différence aux autres. Une telle approche pourra reposer sur un modèle général (ou de profil-type), qui permettra de définir, dans un premier temps, les prescriptions “standards” à présenter à un utilisateur (en fonction de son comportement). Le modèle de spécificités sera exploité, dans un second temps, pour adapter les prescriptions à ces dernières. Je suis convaincue que la modélisation des spécificités, des différences entre utilisateurs, est un moyen de permettre une modélisation plus précise et, bien évidemment, de permettre de proposer des prescriptions de meilleure qualité aux utilisateurs avec un nombre significatif de spécificités, en particulier dans le cas où le modèle des différences est couplé à un modèle classique, reposant sur les similarités. La modélisation de cette différence pourra, par exemple, passer par une approche par *contrast mining*, dont l’objectif premier est d’identifier les attributs permettant de justifier l’appartenance à un groupe. Les différences modélisées peuvent concerner les comportements différents des autres sur un sous-ensemble d’items, sur une période de temps donnée, sur des cycles temporels, etc.

En éducation

En éducation, des profils-type d’étudiants sont classiquement identifiés [Majumdar, 2017], et pourront être exploités en tant que modèle de base. Cependant, même si chaque étudiant peut être associé à un de ces profils-type, il a ses propres caractéristiques. Il est important de pouvoir les identifier, pour impacter la prescription, afin d’être au plus près du profil des étudiants.

Cependant, la grande dimension de la définition du profil d’un étudiants, de même que la variabilité de comportement entre étudiants, en raison du potentiel grand nombre de ressources pédagogiques accessibles, constituent une difficulté supplémentaire.

4.3.2 Des données véloces

Dans de nombreux systèmes, de nouvelles données sont constamment disponibles. Dans notre cas, ce sont de nouvelles traces d’activités utilisateur (données de comportement). Il est primordial que ces

données soient prises en compte au fur et à mesure de leurs apparitions. Elles peuvent simplement être intégrées en mettant à jour le modèle, mais avec le risque qu'elles soient "noyées" dans le modèle, car elles représentent une quantité infime de données, en comparaison du volume total de données.

Ces données ont cependant une importance spécifique, puisqu'elles représentent le profil à court terme de l'utilisateur. Dans le modèle de prescription, notamment, elles devront impérativement être prises en compte pour modéliser au mieux l'utilisateur. Par exemple, deux modèles peuvent être exploités en parallèle : un modèle court-terme et un modèle moyen/long-terme [Li et al., 2007]. La difficulté proviendra de la façon de tenir compte, de façon appropriée, du modèle court-terme pour la tâche de prescription.

De la même façon, des nouvelles données contenant des informations sur les items sont régulièrement disponibles, il est également important de les prendre en compte dès que celles-ci apparaissent.

Ainsi, le modèle conçu doit être capable d'intégrer, en temps réel, toute nouvelle donnée. Cette caractéristique de **dynamisme du modèle**, bien que classique, est très importante, le modèle doit donc être mis à jour *online*. Le défi, ici, en plus de proposer des modèles qui prennent en compte les dernières données, est non seulement lié à la **vélocité des données** (problématiques liées au big data), mais également aux sources variées de données, et donc à l'**hétérogénéité des données**. C'est notamment à nouveau la différence d'échelle temporelle entre les sources de données qui sera complexe à gérer. En effet, certaines sources peuvent être plus véloces que d'autres. C'est également le cas de l'information que chacune de ces sources représente.

4.3.3 De nouveaux phénomènes

En plus d'avoir de nouvelles données disponibles en continu, et dont le modèle doit tenir compte, ces données évoluent, dans le sens où de nouveaux phénomènes peuvent apparaître. Plusieurs questions se posent relativement à ces phénomènes :

- A quoi les phénomènes sont-ils relatifs ? un/des utilisateur(s), un/des item(s) ou des utilisateurs sur des items ?
- Qu'est ce qu'un phénomène ? Un phénomène peut représenter tout motif au sein des données de trace, il peut simplement être un item, représenter des relations entre items (sous la forme de séquences d'items, d'ensembles d'items), ou encore une séquence de comportement-type (composée d'items mais aussi d'autres éléments présents dans les données tels que les caractéristiques des items, etc.), ou encore un lien (corrélation) entre comportements.
- Quel type de nouveauté ? une apparition, une émergence, une évolution, une substitution, une disparition ?

Il est primordial d'identifier ces phénomènes, mais également de les modéliser et d'en tenir compte dans les différents modèles de prescription.

La littérature s'intéresse au démarrage à froid nouveau système, nouvel utilisateur et nouvel item. Dans mes travaux passés, je me suis intéressée au démarrage à froid nouvel utilisateur et nouvel item [Esslimani et al., 2011, Aleksandrova et al., 2017a]. Dans le problème qui m'intéresse ici, bien sûr le problème des nouveaux utilisateurs et des nouveaux items devra être considéré, mais ce sont les **nouveaux phénomènes** qui m'intéressent. Un nouveau phénomène concerne des items et des utilisateurs qui ne sont pas dans un état de démarrage à froid. Cependant, c'est, comme mentionné précédemment, la combinaison de plusieurs de ces éléments qui est nouvelle : de nouvelles associations d'items, pour l'ensemble des utilisateurs, un groupe d'utilisateurs, un seul utilisateur, etc.

Une ressource ou une séquence de ressources qui n'est plus réalisée, ne doit plus être recommandée. En effet, elles peuvent concerner des ressources qui ne sont plus accessibles, plus d'actualité, etc. A l'opposé, si de nouvelles pratiques apparaissent, il peut être adéquat de prescrire ces dernières, car elles peuvent être plus en adéquation avec les utilisateurs. Ici encore, le bruit dans les données, la variabilité des sources, l'hétérogénéité des données et les différentes échelles de temps dans les données seront les difficultés principales de l'identification de ces deux types de nouveautés.

Un nouveau phénomène ? Je fais la différence entre l'apparition et l'émergence d'un phénomène. Pour moi, un phénomène apparaît lorsque celui-ci devient subitement commun (fréquent) sur une très courte période. L'apparition d'un tel phénomène peut être due à un événement extérieur qui force, qui incite ou qui permet ces changements.

A l’opposé, un phénomène émerge lorsque celui-ci apparaît sur une période plus longue. Une fois encore, c’est le problème de l’échelle de temps qui fait la différence entre ces deux phénomènes, et qui est à prendre en compte. Ces deux types de nouveaux phénomènes doivent-ils être considérés différemment dans les modèles ?

Rappelons que les données étant bruitées, il faudra donc également différencier une donnée qui est du bruit (ou une anomalie) d’une donnée qui apparaît effectivement.

A l’inverse, des phénomènes peuvent par ailleurs complètement disparaître. Ici également, l’identification au plus tôt de cette disparition, mais également des raisons de cette disparition sont très importantes, de façon à pouvoir anticiper la disparition éventuelle d’autres phénomènes, et les intégrer dans les modèles.

Un phénomène qui évolue ? Certains phénomènes qui apparaissent peuvent correspondre à l’évolution de phénomènes existants. Concrètement, cette évolution peut représenter l’apparition d’un phénomène, identifié comme légèrement différent d’un phénomène existant. La caractéristique ici est que chacun des deux phénomènes perdure. On peut parler de phénomène qui se dédouble. Dans ce cas, la connaissance du fait que le phénomène existe par ailleurs et qu’il perdure, peut et doit être exploité, constituant des informations supplémentaires sur le nouveau phénomène.

Dans d’autres cas, l’apparition peut correspondre à un phénomène dont la forme change : le même phénomène, dont les deux formes sont relativement proches, mais la première forme disparaît au profit de la seconde. Dans ce cas, le défi sera d’identifier automatiquement la raison de cette mutation, de façon à en tenir compte dans la modélisation du second phénomène, mais également dans les prescriptions.

Deux phénomènes qui se substituent ? La substitution de phénomènes représente un phénomène qui apparaît au détriment d’un autre, qui disparaît. Au contraire de l’évolution précédemment décrite, ici, ce sont bien deux phénomènes différents, dans le sens où ils ne sont pas proches dans leur signification, ni dans les éléments qui les composent. La difficulté, ici, sera d’identifier le lien entre ces deux phénomènes qui ne sont pas similaires *a priori*.

Dans le même esprit de modélisation de plusieurs phénomènes, je souhaite également m’intéresser aux liens/corrélations entre évolutions de phénomènes. Il est en effet intéressant d’identifier qu’un ou des phénomènes apparaissent/évoluent, mais il est d’autant plus intéressant de savoir quels phénomènes évoluent en même temps/en séquence, de la même façon/en opposition. Cela permet d’avoir une connaissance plus précise des données, une meilleure modélisation, qui pourront être exploitées toujours dans la tâche de prescription.

Les phénomènes évoluent ... pour tous ou pour certains utilisateurs ? Au delà de la modélisation des phénomènes, il pourra être intéressant de les relier aux utilisateurs. Je pense notamment à nouveau à l’identification des utilisateurs représentatifs [Boumaza and Brun, 2012b, Esslimani et al., 2013, Aleksandrova et al., 2017a], mais surtout aux utilisateurs “précurseurs”, c’est-à-dire ceux qui sont les premiers à réaliser les nouveaux phénomènes. Ce seront les utilisateurs dont il faudra impérativement suivre les traces, de façon à pouvoir prédire l’évolution des phénomènes. De la même façon, il sera intéressant d’identifier les utilisateurs “retardataires”, ceux qui adoptent certains comportements/phénomènes une fois que ceux-ci sont avérés, et après la majorité des utilisateurs. L’utilité ici n’est pas de suivre leurs activités, mais, au contraire, de pouvoir permettre au modèle de leur proposer des prescriptions adaptées à leur spécificité.

La prise en compte de ces nouveaux phénomènes Quel que soit le type de nouveauté que l’on cherche à identifier, l’objectif principal est d’améliorer la qualité du système (modélisation et/ou prescription). Il est donc primordial que les phénomènes concernés soient identifiés, et pris en compte également le plus tôt possible, donc avec peu de données.

Les approches par fouille de données ou de simples statistiques ne me semblent pas les plus adaptées.

Une première approche qui me semble *a priori* la plus pertinente est la **dérive des concepts** [Tsymbal, 2004, Gama et al., 2014], qui est devenue très populaire ces dernières années, notamment sur des données de flux [Wang et al., 2003, Yang and Fong, 2015, Webb et al., 2016]. Les algorithmes proposés permettent d’identifier de nombreuses évolutions dans les données, en temps-réel, et sur des flux. Ce-

pendant, comme mentionné précédemment, les multiples sources de données, le bruit qu'elles contiennent et leur volume, qui caractérisent notre problème, constitueront une difficulté pour l'exploitation de cette approche.

Une seconde approche est l'identification d'anomalies dans des séquences de données, où une anomalie est une sous-trajectoire dans la séquence [Banerjee et al., 2016]. Cette approche peut constituer une base pour l'identification de la nouveauté. Elle permettra non seulement de différencier le bruit de données effectives, mais pourrait également être exploitée pour identifier des phénomènes “candidats à l'évolution”, et exploités directement par les algorithmes dédiés à l'identification de concepts qui dérivent pour diminuer la complexité.

Nouveaux phénomènes en éducation En éducation, il est évident que les toutes dernières traces d'activités doivent être prises en compte. Soit au niveau de l'étudiant pour connaître et tenir compte de ses toutes dernières activités, interactions, résultats, intérêts, etc. dans les prescriptions. Soit au niveau global pour tous / un groupe d'étudiants, afin d'identifier de nouveaux phénomènes.

Ces phénomènes peuvent représenter de nouvelles pratiques pédagogiques pour l'ensemble des étudiants ou pour des groupes d'étudiants : nouveaux parcours d'apprentissage, nouvelles ressources d'intérêt, nouvelles association de ressources, etc. Ce sont, dans tous les cas, les évolutions de pratiques/ et d'intérêt qu'il est primordial d'identifier, et le plus rapidement possible, pour en tenir compte dès les premiers indices d'un changement (signaux faibles).

4.4 Thématiques à lancer

Si je devais mentionner des thématiques que je souhaiterais lancer, au travers, par exemple, d'encadrements d'étudiants de Master ou de doctorat, elles seraient les suivantes. L'ensemble de ces thématiques a pour cadre applicatif les *learning analytics*.

- La fouille de données hétérogènes pour la prescription. Application à l'éducation pour la personnalisation de parcours pédagogiques
- La prise de décision dans un domaine partiellement observable, de grande taille et non stationnaire, à partir de données bruitées et hétérogènes.
- Stratégies de recommandation : comment déterminer automatiquement le moment adéquat pour fournir une recommandation.
- Identification des dérives des utilisateurs et modélisation : réceptivité aux recommandations, compréhension des activités, adaptation, etc.
- Modélisation de l'évolution de phénomènes dans des données bruitées et volumineuses, pour une modélisation temps-réel de parcours d'apprenants.

Chapitre 5

Animation, administration et responsabilités

En parallèle de mon activité de recherche, d’encadrement et de publication, j’assume régulièrement des responsabilités et assure de tâches d’animation et d’administration, de la recherche et de l’enseignement. Cet autre pan de mon activité est très varié, allant d’activités de relectures à la responsabilité de projets, en passant par des activités éditoriales, la responsabilité de formations, des mandats électifs, etc. Pour des raisons de clarté, je me focalise sur les éléments les plus récents de mon activité.

5.1 Vie d’une équipe de recherche et transfert

Co-création d’une équipe de recherche

En 2007, en collaboration avec Anne Boyer et Azim Roussanaly, nous avons décidé de créer une nouvelle équipe de recherche, au LORIA. Ce souhait provenait de travaux que nous avions menés, soit ensemble, soit indépendamment mais pour lesquels la cohérence et la complémentarité nous semblaient fortes.

La thématique principale de cette équipe, nommée KIWI - *Knowledge, Information and Web Intelligence* est la modélisation utilisateurs, la personnalisation de services et la recommandation. La responsable scientifique de cette équipe est Anne Boyer.

La création de cette équipe a été un réel défi. En effet, les 3 permanents la composant, vite rejoints par un enseignant chercheur de Besançon, provenaient de 3 équipes de recherche, et donc de 3 thématiques de recherche différentes. La création de l’équipe n’était donc pas la conséquence d’une scission d’une équipe existante, mais bien de la création d’une nouvelle thématique commune, partiellement existante auparavant. C’est en ceci que cette création a été un défi.

Le travail de fond, les très nombreuses réunions, interactions et travaux communs que nous avons menés, nous ont permis de mener à bien et ensemble plusieurs recherches et de définir des objectifs communs. L’équipe a officiellement été créée en 2008, et aura bientôt 10 ans. Elle est maintenant composée de 9 permanents, et d’en moyenne 8 doctorants, post-doctorants et ingénieurs.

Animation d’un axe de l’équipe

Dans cette équipe je suis depuis quelques années responsable de l’axe « modélisation prédictive et caractérisation d’utilisateurs ». Cet axe comprend 5 permanents et 6 doctorants. A ce titre, j’anime des réunions où doctorants et permanents présentent et discutent leurs travaux respectifs, initient des travaux communs, réfléchissent à des recherches futures et au dépôt de projets de recherche communs (européens et nationaux).

Transfert

Outre la participation active à des projets de recherche et industriels, qui seront détaillés dans une section ultérieure, les travaux que j'ai menés dans le domaine du *e-learning*, ainsi que ceux menés par 2 des permanents de l'équipe KIWI, ont permis de dégager un savoir-faire unique et original. Ce savoir-faire nous a permis d'envisager la création d'une entreprise dans ce domaine. Cette entreprise, dont l'objectif est de proposer des outils de *learning analytics* aux institutions d'enseignement, sera créée au début de l'année 2018.

5.2 Rayonnement

Les travaux et activités de recherche que je mène m'ont permis d'obtenir un certain rayonnement, au travers de la récompense de certains travaux, de présentations invitées et de plusieurs séjours dans d'autres universités.

Prix

A 4 occasions, des travaux de recherche dont j'étais co-auteur ont été soit nominés, soit primés dans des conférences. Sur chacun de ces articles, c'est un des doctorants que je co-encadrais ou pour lequel j'ai participé à l'encadrement, qui était l'auteur principal.

- **ASONAM 2009**, 3ème Best Paper Award pour l'article *From Social Networks to Behavioral Networks in Recommender Systems*, à la conférence Advances in Social Networks Analysis and Mining (ASONAM).
- **SIIE 2010**, Best Paper Award pour l'article *Topology of Communities for Collaborative Recommendations to Groups*, à la conférence Information Systems and Economic Intelligence (SIIE).
- **Webist 2015**, pré-sélection pour les best paper awards pour l'article *Identifying users with atypical preferences to anticipate inaccurate recommendations*, à la conférence 11th International Conference on Web Information Systems and Technologies (Webist).
- **UMAP 2016**, nomination « outstanding paper » pour l'article *Identifying Grey Sheep Users in Collaborative Filtering : a Distribution-Based Technique*, à la conférence User Modeling, Adaptation and Personalization (UMAP).

Séminaires invités

J'ai, à plusieurs reprises, été invitée à présenter mon activité de recherche, sur le thème des systèmes de recommandation.

- **2009**, conférencière invitée à la journée « Analyser le Web : méthodes et savoirs interdisciplinaires ». Bordeaux. *D'une analyse des traces de navigation à une recommandation personnalisée de pages Web*.
- **2010**, conférencière invitée à la ACM International Conference on Intelligent Interactive Technologies and Multimedia (IITM 2010). Thème : *From cooperative to collaborative filtering*.
- **2013**, présentation invitée, dans le cadre d'une journée de travail de l'ANR blanche B4RCP (Business Recommendation for Configurable Products), LIRMM Montpellier. *Systèmes de recommandation : approches séquentielles et diversité*.
- **2013**, présentation invitée au laboratoire CRESTIC, Reims. *La recommandation sociale : de l'individu à la communauté*.
- **2017**, présentation invitée au *Applied Mathematics Department*, KPI, Kiev, Ukraine.

Mobilité

J'ai par deux fois eu l'occasion d'effectuer un séjour long dans une autre université, ce qui m'a permis de mener des recherches communes avec un chercheur de ces universités.

En 2009, j'ai initié une collaboration avec le Dr Liana Razmerita de la Copenhagen Business School (CBS), Center of Applied Information and Communication Technologies (CAICT), Copenhagen, Danemark, et j'ai effectué un séjour d'une semaine au printemps 2009. Cette première collaboration nous a

permis d'initier des travaux et d'identifier de nombreux défis. Cette collaboration s'est donc poursuivie, notamment lors d'un **séjour de février à juin 2010**, que j'ai effectué à CBS, durant lequel nous nous sommes penchées sur le problème de l'exploitation des systèmes de recommandation dans le domaine du e-learning. Ce séjour a pu être réalisé grâce à l'obtention d'un semestre de CRCT. Cette collaboration a mené à la publication de 2 articles en conférences internationales et 1 article en conférence nationale.

Durant l'année 2011-2012, j'ai effectué un **séjour de 6 mois à l'Université du Littoral Côte d'Opale, au laboratoire d'Informatique Signal et Image de la Côte d'Opale (LISIC)** - EA 4491, dans le but de collaborer avec Amine Boumaza, alors enseignant chercheur en informatique, dont la thématique de recherche est l'évolution artificielle. Ce séjour a pu avoir lieu grâce à un échange de 1/2 service entre nos deux Universités : Amine Boumaza est venu durant le premier semestre à l'Université de Lorraine et y a assuré 1/2 service d'enseignement, puis je suis allée à l'Université du Littoral pour le second semestre où j'ai également assuré 1/2 service d'enseignement.

Notre collaboration a porté sur la conception d'algorithmes d'identification automatique d'utilisateurs représentatifs d'une population (Global Neighbors) par approche évolutionnaire. Celle-ci a abouti à la publication de deux articles dans des conférences internationales de rang A.

5.3 Activités de valorisation

Comme j'ai pu le souligner dans le chapitre 2, mon activité de recherche est caractérisée par une participation forte dans des projets de recherche ou industriels, de nature et d'objectifs variés. Ce lien fort avec des partenaires industriels est un choix que je fais car il me permet à la fois de me confronter à des contraintes réelles, et d'avoir accès à des données de terrain, des données réelles, qui me permettent d'évaluer la pertinence des modèles et algorithmes que je propose.

Par ailleurs, les projets sur lesquels je travaille sont un moyen de financer les doctorants et post-doctorants.

Je présente ici les projets les plus récents, dans lesquels j'ai eu un investissement important. Une liste exhaustive de ces projets peut être trouvée dans mon CV, notamment ceux liés à mon activité en modélisation du langage.

Je divise les projets en deux parties : les projets ayant pour domaine d'application le e-commerce, qui est le domaine auquel je m'intéresse depuis de nombreuses années et sur lequel mon expertise est confirmée puis les projets ayant pour domaine d'application le e-learning, auquel je m'intéresse depuis plus récemment.

5.3.1 Projets en e-commerce

Mon investissement dans les premiers projets relatifs au domaine du e-commerce a consisté en une participation active à ces projets. Depuis quelques années ma participation à ces projets s'est accrue, je suis responsable d'une grande partie des projets dans lesquels je suis investie, projets que j'ai montés et dont j'assure le suivi.

La convention de recherche PERCAL, 2006–2010, avec le Crédit Agricole S.A. avait pour objectif de proposer des modèles de recommandation pour un intranet documentaire, à partir de traces. Cette convention a permis de financer la thèse d'Ilham Esslimani. Dans ce projet j'ai contribué à la proposition de modèles de recommandation : recommandation par une technique de fouille de données séquentielles et recommandation collaborative à base de voisinage identifié sur des traces d'activité. J'ai par ailleurs co-encadré la thèse d'Ilham Esslimani sur la conception de modèles de recommandation à partir de traces. Les algorithmes proposés par Ilham sont actuellement implantés sur l'intranet du Crédit Agricole S.A..

La convention de recherche ARMURES (Applications de Recherche et de Modélisation d'Utilisateurs dans les Réseaux Sociaux), 2012-2015, également avec le Crédit Agricole S.A., et en partenariat avec sa filiale Crédit Agricole Consumer Finance, avait pour objectif l'analyse automatique des réseaux sociaux et blogs dans un but de détection automatique de critères discriminants entre prospects, et d'événements déclencheurs ou influenceurs de caractéristiques clients. C'est cette convention de recherche qui a financé

la thèse de Lina Fahed. Mon rôle dans ce projet, en plus d'en avoir été la **responsable**, a consisté à proposer des approches innovantes de fouille de motifs et d'identification de critères discriminants. J'ai par ailleurs co-dirigé la thèse de Lina Fahed.

Le projet PremierSuiveur, 2015-2018, est une convention de recherche avec l'entreprise Nancéenne Sailendra et est financée par le Grand Nancy, dans l'objectif d'apporter un avantage concurrentiel à l'entreprise. L'objectif de ce projet est de pouvoir identifier et modéliser les utilisateurs précurseurs d'un système de vente en ligne, ainsi que les premiers suiveurs de ces utilisateurs, de même que les utilisateurs ayant un comportement distinct des autres. L'objectif visé de ce projet est de fournir un service de qualité (recommandations satisfaisantes) à chaque client, quelque soit son profil, quelles que soient ses particularités (précurseur ou atypique). Mon rôle dans ce projet en a été la **responsabilité**, de même que l'encadrement de la thèse de Benjamin Gras, que ce projet a permis de financer. J'ai par ailleurs contribué à la proposition de modèles permettant de répondre aux objectifs du projet.

Le projet Sommelier Virtuel, 2017-2018, est un projet avec l'entreprise Alsacienne Sommelier Particulier, financé par la Région Grand Est. L'objectif du projet, dont je suis **responsable**, est de concevoir un système de recommandation permettant de simuler un sommelier, dans sa tâche de conseil d'achat de vins. Mon rôle dans ce projet est non seulement la caractérisation des données, la définition des objectifs, et d'identifier les solutions de l'état de l'art envisageables. Ce projet a par ailleurs permis de financer le stage de M2 R de Jeffer Honion, qui a proposé un premier algorithme hybride de recommandation, et se poursuit actuellement dans le cadre de l'encadrement d'un ingénieur de recherche pour 18 mois.

5.3.2 Projets en e-learning

J'ai été et je suis actuellement investie dans plusieurs projets en e-learning. Ce domaine d'application est relativement nouveau dans mon activité. Par conséquent, mon rôle dans ces projets est limité, pour le moment, à une participation active dans ces derniers et à la responsabilité de lots. Le e-learning étant au cœur de mes perspectives de recherche, je souhaite monter de nouveaux projets dans ce domaine, projets que je porterai.

Le projet PERICLES (Projet pour l'Evaluation et la Recherche Informatisée autour des Compétences dans L'Enseignement Supérieur), 2012-2014, est un projet de l'appel PIA2. L'objectif général de ce projet est de la réalisation d'un démonstrateur d'outils permettant aux institutions d'enseignement supérieur de mettre en œuvre une démarche d'assurance qualité interne fondée sur les critères de leur choix vise. Les partenaires de ce projet sont à la fois des acteurs éducatifs publics et privés et des équipes de recherche. La tâche 3 de ce projet vise à la détection automatique de squelettes d'apprentissage et à la proposition de recommandations personnalisées. Mon rôle dans ce projet a consisté en une participation active à tâche, dans laquelle j'ai contribué à la conception d'algorithmes de recommandation de ressources pédagogiques. J'ai par ailleurs encadré le post-doctorat de Brahim Batouche sur cette thématique.

Le projet Interlingua 2014-2015, du programme Interreg IVA a pour objectif de concevoir une solution évolutive pour le problème du soutien pratique pour les élèves qui étudient dans une langue étrangère. Les partenaires de ce projet étaient à la fois des partenaires éducatifs et des partenaires recherches, belges, luxembourgeois, français et allemands. Dans ce projet, j'ai été **responsable de la tâche 1** qui visait à recueillir des expériences d'étudiants ayant effectué une partie de leurs études en langue étrangère et de proposer un guide de bonnes pratiques pour la conception du service visé par le projet. J'ai donc participé à l'élaboration de cette tâche, en partenariat avec les partenaires belges et allemands. je me suis également investie dans la conception d'un système de recommandation (tâche 2).

Le projet METAL (Modèles Et Traces au service de l'Apprentissage des Langues), 2016-2020 est un projet de l'appel e-Fran du PIA2. L'objectif du projet METAL est concevoir, développer et évaluer des outils de suivi individualisé destinés aux élèves ou aux enseignants (*Learning Analytics*), et des technologies innovantes pour un apprentissage personnalisé des langues à l'écrit et à l'oral. Mon rôle dans ce projet, outre la **responsabilité de la tâche 1.2**, est de travailler sur la conception d'un tableau de bord à destination des élèves, de façon à améliorer leur expérience d'apprentissage et leur motivation.

C'est dans le cadre de ce projet que je co-dirige la thèse de Julie Budaher, débutée en décembre 2016, et qui travaille, dans un premier temps, sur la proposition d'algorithmes de recommandation permettant d'expliquer ces dernières.

Le projet METAL fait partie des 9 projets retenus lors du premier appel e-Fran.

5.3.3 Projets soumis ou en cours d'élaboration

- Je suis partenaire du projet européen BigWebData - Projet Marie Curie Innovative Training Network, soumis en janvier 2017. Mon rôle consiste en la proposition de modèles de recherche de motifs fréquents dans des traces d'apprentissage dans le but d'améliorer l'expérience d'apprentissage.
- Je suis partenaire et co-rédactrice du projet européen *Target System*, Projet Horizon 2020, actuellement en soumission, portant sur l'amélioration du circuit de production et des habitudes de consommation de produits alimentaires, au travers de systèmes d'abonnement. Mon rôle porte sur l'élaboration de systèmes de recommandation à destination des consommateurs et en la conception d'algorithmes d'identification de consommateurs décrocheurs.
- Depuis l'automne 2016, je suis l'acteur académique principal du projet « Sommelier Virtuel », avec l'entreprise « Sommelier Particulier ». Objectif : conception d'algorithmes de recommandation de vins à partir de données de traces d'interaction (d'achats) et de comportements d'experts (conversations avec des sommeliers). Ce projet fait l'occasion de l'encadrement d'un M2 R au printemps 2017 et devrait mener à une convention de recherche de longue durée à partir de l'automne 2017.
- Depuis l'automne 2016, je travaille à la mise en place d'une collaboration entre KIWI et l'entreprise TRACKING (Luxembourg) et plusieurs autres partenaires Belges et Néerlandais pour la modélisation de traces d'exploitation d'utilisateurs et la modélisation prédictive de phénomènes.

5.4 Vie de la recherche

J'ai une activité régulière d'animation de la recherche, en participant à des jurys ou comités, ou en organisant des événements.

Membre de jurys de thèse

Outre la participation aux jurys de thèse des doctorants que j'encadre, j'ai participé à des jurys de thèse, en tant que rapporteur (thèse à l'étranger) ou examinateur au titre d'expert du domaine de la recommandation.

En 2012, j'ai été rapporteur de la thèse de Yanir Seroussi, Monash University, Australie. La thèse portait sur *Text Mining and Rating Prediction with Topical User Models*.

En 2016, j'ai été examinateur de la thèse de Rajani Chulyadyo, Université de Nantes (LINA, section 27). Le titre de la thèse était *A new horizon for the recommendation Integration of spatial dimensions to aid decision making*.

Par ailleurs en 2016, j'ai été membre du jury de thèse à mi-parcours de Matias Callara, Université de Haute Alsace (Laboratoire MIPS, section 27). Le titre de la thèse était *Suivi et prédiction des comportements utilisateurs par apprentissage pour l'optimisation des ressources et des services informatiques*.

Participation à des comités de sélection

En 2013, j'ai été membre du comité de sélection du poste MCF0542 section 27, Université de Lorraine, UFR MI.

En 2016, j'ai été membre du comité de sélection du poste MCF4108 section 27/61, Université de Haute Alsace, Mulhouse - IUT de Mulhouse.

Activités éditoriales, comités de programme et évaluations

Je suis membre du **comité éditorial de plusieurs revues** :

- 2010–présent : Revue Technique et Sciences Informatiques (TSI).
- 2012 et 2016 : Les Cahiers du Numérique, numéros spéciaux.
- 2017–présent : Revue d'Intelligence Artificielle (RIA)

Je suis par ailleurs **membre du comité scientifique ou du comité de programme** de nombreuses conférences en intelligence artificielle, apprentissage automatique, systèmes de recommandation, etc.

Je cite ici un sous-ensemble de ces conférences : CORIA (chaque année), CAP (2008 à 2014), Workshop Intelligent Personalisation (IP) à IJCAI 2015, Workshop Interdisciplinaire sur les systèmes de recommandation (AISR) 2017 (comité scientifique), consortium doctoral de la conférence User Modeling, Adaptation and Personalization (UMAP 2010), ACM HT 2010, ACM IITM 2010, etc.

Je suis bien évidemment relecteur régulier pour des revues (TSI, Advances in Social Networks Analysis and Mining, Information Sciences (Elsevier), Journal of Systems and Software (Elsevier)), TCS (Theoretical Computer Science), TIST (Transactions on Intelligent Systems and Technology), JIIS (Journal of Intelligent Information Systems), EAIT (Education and Information Technologies), et pour de nombreuses conférences.

Présidence et participation à des comités d'organisation de conférences

En 2018, je serai **présidente du comité d'organisation de PFIA 2018** - Plate-Forme Intelligence Artificielle, qui aura lieu à Nancy du 2 au 6 juillet 2018.

A plusieurs reprises, j'ai choisi de m'investir à plusieurs reprises dans l'organisation de conférences, en intégrant leur comité d'organisation :

- TALN - Traitement Automatique des Langues Naturelles, 2002, Nancy.
- Ecole d'été ESSLLI - European Summer School in Logic, Language, and Information, 2004, Nancy.
- CAp - Conférence sur l'Apprentissage automatique, 2012, Nancy.
- CORIA - COnférence en Recherche d'Information et Applications, 2014, Nancy.

5.5 Mandats et activités collectives

En parallèle de ces activités, j'interviens également de façon régulière au service de la communauté scientifique au travers de ma présence dans différents conseils et d'expertises.

5.5.1 Mandats

- **2011–présent** membre élue du CNU 27 (en 2011, réélue en 2015)
- **2013–présent** membre élue du conseil de laboratoire LORIA. - **2014–présent** membre élue du conseil d'UFR MI.
- **2017–présent** membre élue du conseil de collegium LMI (Lorraine Management et Innovation).
- **2017–présent** membre nommé du CLHSCT (comité local d'hygiène, de sécurité et des conditions de travail) du laboratoire LORIA.

5.5.2 Expertise

- **2010–présent** : experte auprès du Ministère de la Recherche pour les accréditations Crédit Impôt Recherche. Une dizaine d'expertises et contrôles par an.
- **2011–présent** : experte régulière auprès de l'ANR pour l'évaluation de projets (différents appels).
- **Reviewer occasionnel** de projets COST (European Cooperation in Science and Technology), Suisse.

5.5.3 Activités locales

- **2008–2010** membre de la commission ingénieur de INRIA Nancy Grand Est.
- **2009–présent** membre du comité d'organisation du séminaire IPAC « Image, Perception, Action, Cognition » au LORIA, séminaire commun à plusieurs équipes de recherche du laboratoire.
- **2012** membre de la commission bureau du laboratoire LORIA. Objectif : repenser l'attribution des bureaux aux différentes équipes de recherche.
- **2012–présent** membre de la commission de recrutement ATER et DCCE de l'UFR MI, en lien avec le laboratoire LORIA en vue d'une harmonisation des recrutements entre composantes.
- **2014** membre nommé du groupe de travail du LORIA sur les Masters en rapport avec l'informatique en Lorraine. Groupe formé à l'initiative du Laboratoire LORIA.

5.6 Animation et administration de l'enseignement

5.6.1 Direction de formations

2012–présent : Responsabilité d'un diplôme de Licence Depuis 2012, je suis responsable de la Licence MIASHS - Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales. Je suis par ailleurs **présidente de jury** et **responsable du processus de recrutement en Licence** (recrutement sur dossier et entretien oral pour les candidatures en L2 et L3). Je suis également en charge de l'étude des dossiers Campus France (>300 dossiers/an).

Je suis **porteur de la maquette 2018-2022** (maquette déposée à l'Université en Octobre 2016) : j'ai animé les réunions de réflexion autour des évolutions pédagogiques, définition des nouveaux contenus, redéfinition du socle commun aux parcours. J'ai également mis en place des accords privilégiés avec les porteurs d'autres licences, notamment dans le but de faciliter les réorientations.

Je suis par ailleurs **directrice des études** du parcours Sciences Cognitives (2ème et 3ème année de Licence). A ce titre, j'assure le recrutement des enseignants, définis et gère les emplois du temps. Par ailleurs j'accompagne et conseille les étudiants au jour le jour (administratif, pédagogique, sur d'éventuelles réorientations si nécessaire, semestres ERASMUS, etc.).

2004–2012 : Responsabilité du cycle de formation continue ADSIO

J'ai eu la responsabilité du cycle de formation continue ADSIO *Analyste Développeur en Systèmes d'Information des Organisations*, géré par l'UFR et financé par la région Lorraine. Cette formation qualifiante est destinée à un public en recherche d'emploi. La formation a lieu sur sept mois, chaque année, et se déroule à temps plein. Mes responsabilités ont concerné la sélection des candidats (dossiers + entretiens oraux), le recrutement des enseignants, la gestion de l'emploi du temps, le suivi du déroulement de la formation, du stage et le lien avec le service de formation continue de l'Université.

5.6.2 Diffusion

Je suis très active dans la **promotion de l'offre de formation** des diplômes de l'UFR MI, pour lesquels j'effectue chaque année un grand nombre d'actions de communication auprès des enseignants, conseillers d'orientation, lycéens et des étudiants de BTS/DUT.

Je co-organise chaque année le **forum des Sciences Cognitives de Nancy**. Ce forum, d'une durée d'une journée, a pour objectif de faire découvrir aux étudiants de l'UFR MI, aux étudiants de toutes filières de l'Université, aux lycéens mais aussi au grand public, les Sciences Cognitives sous toutes leurs formes. Tout au long de la journée, des exposés scientifiques et professionnels, ainsi que des démonstrations sont faites par des industriels et des chercheurs.

En 2013, j'ai co-organisé (nous étions 2 organisateurs) les **Journées Nationales MIAGE** à Nancy (16-18 mai 2013), qui ont rassemblé plus de 500 personnes : des représentants des équipes pédagogiques et des étudiants des 20 MIAGE de France, ainsi que des professionnels.

Bibliographie

- [Abboud et al., 2017] Abboud, Y., Boyer, A., and Brun, A. (2017). CCPM : A Scalable and Noise-Resistant Closed Contiguous Sequential Patterns Mining Algorithm. In *13th International Conference on Machine Learning and Data Mining MLDM 2017*, volume 89, page 15, New York, United States.
- [Abboud et al., 2015] Abboud, Y., Brun, A., and Boyer, A. (2015). Predict the emergence – application to competencies in job offers. In *International Conference on Tools with Artificial Intelligence (ICTAI)*.
- [Achar et al., 2012] Achar, A., Laxman, S., Viswanathan, R., and Sastry, P. (2012). Discovering injective episodes with general partial orders. *Data Mining and Knowledge Discovery*, 25(1) :67–108.
- [Adomavicius and Tuzhilin, 2005] Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6) :734–749.
- [Adomavicius and Zhang, 2012] Adomavicius, G. and Zhang, J. (2012). Impact of data characteristics on recommender systems performance. *ACM Transactions on Management Information Systems (TMIS)*, 3(1) :3 :1–3 :17.
- [Aggarwal and Yu, 2005] Aggarwal, C. and Yu, S. (2005). An effective and efficient algorithm for high-dimensional outlier detection. *The VLDB Journal*, 14 :211–221.
- [Agrawal et al., 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD*, pages 207–216.
- [Agrawal and Srikant, 1995] Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the International Conference on Data Engineering (ICDE’95)*, pages 3–14.
- [Ahn, 2008] Ahn, H. J. (2008). A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Information Sciences*, 178(1) :37 – 51.
- [Aleksandrova et al., 2014] Aleksandrova, M., Brun, A., Boyer, A., and Chertov, O. (2014). What about interpreting features in matrix factorization-based recommender systems as users? In *Workshop on Social Personalization at ACM HT conference*.
- [Aleksandrova et al., 2017a] Aleksandrova, M., Brun, A., Boyer, A., and Chertov, O. (2017a). Identifying Representative Users in Matrix Factorization-based Recommender Systems : Application to Solving the Content-less New Item Cold-start Problem. *Journal of Intelligent Information Systems*, 48(2) :365–397.
- [Aleksandrova et al., 2016a] Aleksandrova, M., Brun, A., Chertov, O., and Boyer, A. (2016a). Automatic formation of sets of contrasting rules to identify trigger factors. In *ECAI – European Conference on Artificial Intelligence*.
- [Aleksandrova et al., 2016b] Aleksandrova, M., Brun, A., Chertov, O., and Boyer, A. (2016b). Sets of contrasting rules : a supervised descriptive rule induction pattern for identification of trigger factors. In *Proceedings of the annual IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*.
- [Aleksandrova et al., 2017b] Aleksandrova, M., Chertov, O., Brun, A., and Boyer, A. (2017b). Contrast classification rules for mining local differences in medical data. In *9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems : Technology and Applications (IDAACS’17)*.

- [Amatriain et al., 2009] Amatriain, X., Lathia, N., Pujol, J. M., Kwak, H., and Oliver, N. (2009). The wisdom of the few : a collaborative filtering approach based on expert opinions from the web. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 532–539. ACM.
- [Anastasiu et al., 2014] Anastasiu, D. C., Iverson, J., Smith, S., and Karypis, G. (2014). *Big Data Frequent Pattern Mining*, pages 225–259. Springer International Publishing.
- [Anaya et al., 2016] Anaya, A., Luque, M., and Peinado, M. (2016). A visual recommender tool in a collaborative learning experience. *Expert Systems With Applications*, 45 :248–259.
- [Ao et al., 2017] Ao, X., Luo, P., Wang, J., Zhuang, F., and He, Q. (2017). Mining precise-positioning episode rules from event sequences. In *Proceedings of the 33rd IEEE International Conference on Data Engineering (ICDE)*, pages 83–86.
- [Appelbaum et al., 2017] Appelbaum, D., Kogan, A., Vasarhelyi, M., and Yan, Z. (2017). Impact of business analytics and enterprise systems on managerial accounting. *International Journal of Accounting Information Systems*, 25 :29 – 44.
- [Araujo, 2007] Araujo, L. (2007). How evolutionary algorithms are applied to statistical natural language processing. *Artificial Intelligence Review*, 28 :275–303.
- [Ayres et al., 2002] Ayres, J., Flannick, J., Gehrke, J., and Yiu, T. (2002). Sequential pattern mining using a bitmap representation. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD’02, pages 429–435.
- [Banerjee et al., 2016] Banerjee, P., Yawalkar, P., and Ranu, S. (2016). Mantra : A scalable approach to mining temporally anomalous sub-trajectories. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD’16, pages 1415–1424.
- [Barabasi et al., 2002] Barabasi, A., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaboration. *Physica A*, 311(3-4).
- [Batouche et al., 2014a] Batouche, B., Brun, A., and Boyer, A. (2014a). Clustering based recommendation of pedagogical resources. In *EDEN - Research Workshop*.
- [Batouche et al., 2014b] Batouche, B., Brun, A., and Boyer, A. (2014b). Unsupervised Machine Learning based recommendation of pedagogical resources. In *EC-TEL*, Graz, Austria.
- [Bellogín, 2011] Bellogín, A. (2011). *Performance Prediction in Recommender Systems*, pages 401–404. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3 :1137—1155.
- [Berkovsky et al., 2007] Berkovsky, S., Kuflik, T., and Ricci, F. (2007). Cross-domain mediation in collaborative filtering. In *Proceedings of the 11th international conference on User Modeling*, UM ’07, pages 355–359.
- [Bernier et al., 2010] Bernier, C., Brun, A., Aghasaryan, A., Bouzid, M., Picault, J., and Senot, C. (2010). Topology of Communities for the Collaborative Recommendations to Groups. In *3rd International Conference on Information Systems and Economic Intelligence - SIIE’2010*, Sousse, Tunisia.
- [Bigi et al., 2001a] Bigi, B., Brun, A., Haton, J., Smaïli, K., and Zitouni, I. (2001a). A Comparative Study of Topic Identification on Newspaper and E-Mail. In *8th Proc. of String Processing and Information Retrieval (SPIRE2001)*, pages 238–241, Laguna de San Rafael, Chile.
- [Bigi et al., 2001b] Bigi, B., Brun, A., Smaïli, K., and Haton, J. (2001b). A Hierarchical Approach for Topic Identification. In *International Workshop "Speech and Computer" (SPECOM2001)*, pages 85–88.
- [Bigi et al., 2001c] Bigi, B., Brun, A., Smaïli, K., and Haton, J. (2001c). Dynamic Topic Identification : Towards Combination of Methods. In *European Conference on Recent Advances in Natural Language Processing (RANLP2001)*, pages 255–257, Tzigov Chark, Bulgaria.
- [Bigi et al., 2001d] Bigi, B., Brun, A., Smaïli, K., and Haton, J.-P. (2001d). A Hierarchical Approach for Topic Identification. In *Proceedings of the international workshop Speech and Computer - SPECOM’01*, page 4 p, Moscow, Russia, France. Colloque avec actes et comité de lecture. internationale.

- [Bobadilla et al., 2012] Bobadilla, J., Ortega, F., Hernando, A., and Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems*, 26 :225 – 238.
- [Bobadilla et al., 2013] Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, 46 :109 – 132.
- [Bonnin et al., 2008a] Bonnin, G., Brun, A., and Boyer, A. (2008a). Collaborative Filtering Inspired from Language Modeling. In *Proc. of the First International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2008), Workshop on Recommender Systems and Personalized Retrieval (RSPR)*, Ostrava, Czech Republic.
- [Bonnin et al., 2008b] Bonnin, G., Brun, A., and Boyer, A. (2008b). Using Skipping for Sequence-Based Collaborative Filtering. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT'08)*, pages 775–779, Sydney, Australia. University of Technology.
- [Bonnin et al., 2009a] Bonnin, G., Brun, A., and Boyer, A. (2009a). A Low-Order Markov Model Integrating Long-Distance Histories for Collaborative Recommender Systems. In *International Conference on Intelligent User Interfaces (IUI)*, pages 57–66, Sanibel Island, United States.
- [Bonnin et al., 2009b] Bonnin, G., Brun, A., and Boyer, A. (2009b). Renforcement des modèles probabilistes en utilisant l'Information Mutuelle pour des Recommandations contextualisées. In Hassoun, M. and Hachani, M. E., editors, *7e colloque du chapitre français de l'ISKO - Intelligence collective et organisation des connaissances*, Lyon, France. Université Jean Moulin Lyon3 / ENSSIB.
- [Bonnin et al., 2010a] Bonnin, G., Brun, A., and Boyer, A. (2010a). Detecting Parallel Browsing to Improve Web Predictive Modeling. In *International Conference on Knowledge Discovery and Information Retrieval - KDIR 2010*, pages 504–509, Valencia, Spain.
- [Bonnin et al., 2010b] Bonnin, G., Brun, A., and Boyer, A. (2010b). Towards Tabbing Aware Recommendations. In *International Conference on Intelligent Interactive Technologies and Multimedia - ACM IITM 2010*, Allahabad, India.
- [Bonnin et al., 2010c] Bonnin, G., Brun, A., and Boyer, A. (2010c). *Web Intelligence and Intelligent Agents*, chapter Skipping-Based Collaborative Recommendations Inspired from Statistical Language Modeling, pages 263–288. IN-TECH.
- [Bonnin et al., 2011a] Bonnin, G., Brun, A., and Boyer, A. (2011a). Handling Tabbing and Backward References for Predictive Web Usage Mining. In *Proc. of the International Joint Conference on Knowledge Discovery and Information Retrieval (KDIR 2011)*.
- [Bonnin et al., 2011b] Bonnin, G., Brun, A., and Boyer, A. (2011b). Taking into account tabbed browsing in predictive web usage mining. In *Proc. of the First International Conference on Social Eco-Informatics (SOTICS 2011)*.
- [Bonnin et al., 2012] Bonnin, G., Brun, A., and Boyer, A. (2012). Exploitation du skipping pour la modélisation prédictive des usages du web. Vers une meilleure prise en compte du bruit. *Revue des Sciences et Technologies de l'Information - Série RIA : Revue d'Intelligence Artificielle*, 26(6) :609–642.
- [Boumaza and Brun, 2012a] Boumaza, A. and Brun, A. (2012a). From Neighbors to Global Neighbors in Collaborative Filtering : an Evolutionary Optimization Approach. In *GECCO - Genetic and Evolutionary Computation Conference - 2012*, pages 345–352, Philadelphia, United States. ACM.
- [Boumaza and Brun, 2012b] Boumaza, A. and Brun, A. (2012b). Stochastic search for global neighbors selection in collaborative filtering. In *27th ACM Symposium on Applied Computing (ACM SAC 2012)*,.
- [Boyer and Brun, 2007a] Boyer, A. and Brun, A. (2007a). Natural Language Processing for Usage Based Indexing of Web Resources. In *29th European Conference on Information Retrieval (ECIR 2007)*, pages 517–524, Roma.
- [Boyer and Brun, 2007b] Boyer, A. and Brun, A. (2007b). Towards a statistical grammar of usage for document retrieval in digital libraries. In *International Symposium on Signal Processing and its Applications*, United Arab Emirates.
- [Boyer et al., 2010] Boyer, A., Brun, A., and Skaf-Molli, H. (2010). Human Computer Collaboration to Improve Annotations in Semantic Wikis. In *6th International Conference on Web Information Systems and Technologies - WEBIST 2010*, pages 89–94, Valencia, Spain. INSTICC Press.

- [Brown et al., 1990] Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2) :79–85.
- [Brun, 1999] Brun, A. (1999). Identification de thèmes pour la modélisation statistique du langage. Master’s thesis, Université Nancy 2.
- [Brun, 2003] Brun, A. (2003). *Détection de thèmes et adaptation des modèles de langage pour la reconnaissance automatique de la parole*. PhD thesis, Université Henri Poincaré Nancy1.
- [Brun et al., 2014] Brun, A., Aleksandrova, M., and Boyer, A. (2014). Can latent features be interpreted as users in matrix factorization-based recommender systems? In *Proceedings of the IEEE/ACM WI-IAT conference*.
- [Brun et al., 2009a] Brun, A., Bonnin, G., and Boyer, A. (2009a). History Dependent Recommender Systems Based on Partial Matching. In *First and Seventeenth International Conference on User Modeling, Adaptation and Personalization - UMAP 2009*, volume 5535 of *LNCS*, pages 343–348.
- [Brun and Boyer, 2007] Brun, A. and Boyer, A. (2007). Usage based Indexing of Web Resources with Natural Language Processing. In *3rd International Conference on Web Information Systems and Technologies*, Barcelona.
- [Brun and Boyer, 2009] Brun, A. and Boyer, A. (2009). Towards Privacy Compliant and Anytime Recommender Systems. In *10th International Conference on Electronic Commerce and Web Technologies - EC-Web 09*, volume 5692 of *LNCS*, pages 276–287. Johannes Kepler University of Linz.
- [Brun and Boyer, 2010a] Brun, A. and Boyer, A. (2010a). Are Recommender Systems Real-Time in Mobile Environment? Towards Instantaneous Recommenders. In *6th International Conference on Web Information Systems and Technologies (Webist 2010)*, pages 101–106, Valencia, Spain.
- [Brun and Boyer, 2010b] Brun, A. and Boyer, A. (2010b). Du e-commerce au m-commerce : vers une Recommandation Incrémentale. In *Proc. of the 7th COnférence en Recherche d’Information et Applications (CORIA)*.
- [Brun and Boyer, 2010c] Brun, A. and Boyer, A. (2010c). Linking Collaborative Filtering and Social Networks : Who are my Mentors? In *The 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2010)*, Odense, Denmark.
- [Brun and Boyer, 2011] Brun, A. and Boyer, A. (2011). Inspiration des sondages d’opinion pour réduire la latence en filtrage collaboratif. In *COnférence en Recherche d’Information et Applications - CORIA 2011*, pages 49–56, Avignon, France.
- [Brun and Boyer, 2012] Brun, A. and Boyer, A. (2012 (à paraître)). Détection de communautés d’intérêt et recommandation sociale par leaders. *Ingénierie des Systèmes d’Information (ISI) - Numéro Spécial "Impact des réseaux sociaux et du Web 2.0 dans/sur/pour les systèmes d’information"*.
- [Brun et al., 2010a] Brun, A., Boyer, A., and Razmerita, L. (2010a). Compass to Locate the User Model I need : Building the Bridge Between Researchers and Practitioners in User Modeling. In *User Modeling, Adaptation and Personalization - UMAP 2010*, volume 6075 of *LNCS*, pages 303–314, Kona, United States.
- [Brun et al., 2009b] Brun, A., Castagnos, S., and Boyer, A. (2009b). A Positively Directed Mutual Information Measure for Collaborative Filtering. In *2nd Conférence Internationale Systèmes d’Information et Intelligence Economique (SIIE 2009)*, pages 943–958, Hammamet, Tunisia.
- [Brun et al., 2011a] Brun, A., Castagnos, S., and Boyer, A. (2011a). From Community Detection to Mentor Selection in Rating-Free Collaborative Filtering. *Advances in Multimedia Journal*, 2011 :1–19.
- [Brun et al., 2011b] Brun, A., Castagnos, S., and Boyer, A. (2011b). *Social Network Mining, Analysis and Research Trends : Techniques and Applications*, chapter Social Recommendations : Mentor and Leader Detection to Alleviate the Cold-Start Problem in Collaborative Filtering. IGI Global.
- [Brun et al., 2004] Brun, A., Cerisara, C., Fohr, D., Illina, I., Langlois, D., Mella, O., and K., S. (2004). ANTS : Le système de Transcription Automatique du Loria. In *Journées d’Etudes sur la Parole*, Fès, Maroc.

- [Brun et al., 2015a] Brun, A., Grandbastien, M., Henry, J., and Vandeput, E. (2015a). Analyse de besoins pour un service en ligne. In *Environnements Informatiques pour l'Apprentissage Humain (EIAH)*, Agadir, Morocco.
- [Brun et al., 2015b] Brun, A., Grandbastien, M., Henry, J., and Vandeput, E. (2015b). Needs analysis for an online learning service. In *IFIP TC3 Working Conference*, Vilnius, Lithuania.
- [Brun et al., 2010b] Brun, A., Hamad, A., Buffet, O., and Boyer, A. (2010b). From "I like" to "I prefer" in Collaborative Filtering. In *International Conference on Tools with Artificial Intelligence - ICTAI 2010*, pages 365–367, Arras, France. IEEE.
- [Brun et al., 2010c] Brun, A., Hamad, A., Buffet, O., and Boyer, A. (2010c). Towards Preference Relations in Recommender Systems. In *Workshop on Preference Learning (PL2010) in ECML-PKDD*, Barcelona, Spain. Eyke Hüllermeier and Johannes Fürnkranz.
- [Brun et al., 2010d] Brun, A., Hamad, A., Buffet, O., and Boyer, A. (2010d). Vers l'utilisation de relations de préférence pour le filtrage collaboratif. In *17eme congrès francophone Reconnaissance des Formes et Intelligence Artificielle - RFIA 2010*, Caen, France.
- [Brun et al., 2006] Brun, A., Langlois, D., and Smaïli, K. (2006). Exploration et utilisation d'informations distantes dans les modèles statistiques de langage. In *Traitement Automatique des Langues Naturelles (TALN2006)*, pages 425–434, Leuven, Belgium.
- [Brun et al., 2007] Brun, A., Langlois, D., and Smaïli, K. (2007). Improving language models by using distant information. In *International Symposium on Signal Processing and its Applications*, United Arab Emirates.
- [Brun et al., 2000a] Brun, A., Langlois, D., Smaïli, K., and Haton, J. (2000a). Discarding Impossible Events from Statistical Language Models. In *International Conference on Spoken Language Processing (ICSLP2000)*, pages 981–984.
- [Brun et al., 2001a] Brun, A., Langlois, D., Smaïli, K., and Haton, J.-P. (2001a). Improving Statistical Language Models by Removing Impossible Events. In *Proceedings of the International Workshop "Speech and Computer" - SPECOM 2001*, page 4 p, Moscow, Russia. Colloque avec actes et comité de lecture. internationale.
- [Brun et al., 2001b] Brun, A., Langlois, D., Smaïli, K., and Haton, J. P. (2001b). Improving Statistical Language Models by Removing Impossible Events. In *International Workshop "Speech and Computer" (SPECOM2001)*.
- [Brun et al., 2010e] Brun, A., Skaf-Molli, H., and Boyer, A. (2010e). Raising up Annotations In Pedagogical Resources by Human-Computer Collaboration. In *European Distance and E-learning Network (EDEN 2010)*, Budapest, Hungary.
- [Brun et al., 2010f] Brun, A., Skaf-Molli, H., and Boyer, A. (2010f). Raising up Annotations In Pedagogical Resources by Human-Computer Collaboration. In *Research Workshop European Distance and E-learning Network (EDEN 2010)*, Budapest, Hungary.
- [Brun and Smaïli, 2004] Brun, A. and Smaïli, K. (2004). Fiabilité de la référence humaine dans la détection de thème. In *Traitement Automatique des Langues Naturelles (TALN2004)*, Fès, Maroc.
- [Brun et al., 2000c] Brun, A., Smaïli, K., and Haton, J. (2000c). Topic Identification Challenge Based on Short Word History. In *Traitement automatique des Langues Naturelles (TALN2000)*, pages 383–392, Lausanne, Suisse.
- [Brun et al., 2002a] Brun, A., Smaïli, K., and Haton, J. (2002a). Contribution to Topic Identification by Using Word Similarity. In *International Conference on Spoken Language Processing (ICSLP2002)*, pages 1965–1968, Denver, Colorado.
- [Brun et al., 2002b] Brun, A., Smaïli, K., and Haton, J. (2002b). WSIM : une méthode de détection de thème fondée sur la similarité entre mots. In *Traitement Automatique des Langues Naturelles (TALN2002)*, pages 145–154, Nancy, France.
- [Brun et al., 2003] Brun, A., Smaïli, K., and Haton, J. (2003). Nouvelle approche de la sélection de vocabulaire pour la détection de thème. In *Traitement Automatique des Langues Naturelles (TALN2003)*, pages 45–54, Nantes, France.

- [Brun et al., 2000b] Brun, A., Smaïli, K., and Haton, J. (Sep 2000b). Experiment Analysis in Newspaper Topic Detection. In *String Processing and Information Retrieval, IEEE Computer Society (SPIRE2000)*, pages 55–64, A Coruña, Spain.
- [Buffet and Sigaud, 2008] Buffet, O. and Sigaud, O. (2008). *Processus décisionnels de Markov en intelligence artificielle*. IC2 - informatique et systèmes d'information. Lavoisier - Hermes Science Publications.
- [Bull and Kay, 2010] Bull, S. and Kay, J. (2010). Open learner models. *Advances in intelligent tutoring systems*, pages 301–322.
- [Cai et al., 2009] Cai, C., Liao, X., and Carin, L. (2009). Learning to explore and exploit in pomdps. In *Advances in Neural Information Processing Systems*, pages 198–206.
- [Campos et al., 2014] Campos, P. G., Díez, F., and Cantador, I. (2014). Time-aware recommender systems : a comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction*, 24(1) :67–119.
- [Candillier et al., 2007] Candillier, L., Meyer, F., and Boullé, M. (2007). Comparing state-of-the-art collaborative filtering systems. In *Proc. of 5th International Conference on Machine Learning and Data Mining in Pattern Recognition, MLDM'07*, pages 548–562.
- [Castagnos et al., 2008a] Castagnos, S., Brun, A., and Boyer, A. (2008a). Probabilistic Association Rules for Item-Based Recommender Systems. In *4th European Starting AI Researcher Symposium (STAIRS 2008), in conjunction with the 18th European Conference on Artificial Intelligence (ECAI 2008)*, pages 36–46, Patras, Greece. University of Patras.
- [Castagnos et al., 2008b] Castagnos, S., Brun, A., and Boyer, A. (2008b). Probabilistic Reinforcement Rules for Item-Based Recommender Systems. In ECCAI, editor, *18th European Conference on Artificial Intelligence (ECAI 2008)*, pages 823–824, Patras, Greece. University of Patras.
- [Castagnos et al., 2013a] Castagnos, S., Brun, A., and Boyer, A. (2013a). Utilité et perception de la diversité dans les systèmes de recommandation. In *Proceedings of the 10th Conférence en Recherche d'Information et Applications*, pages 237–252.
- [Castagnos et al., 2013b] Castagnos, S., Brun, A., and Boyer, A. (2013b). When diversity is needed... but not expected! In *Proceedings of the Third International Conference on Advances in Information Mining and Management IMMM 2013*, pages 44–50.
- [Castagnos et al., 2014] Castagnos, S., Brun, A., and Boyer, A. (2014). La diversité : entre besoin et méfiance dans les systèmes de recommandation. *Journal Interaction Intelligence Information*.
- [Castells et al., 2015] Castells, P., Hurley, N. J., and Vargas, S. (2015). Novelty and diversity in recommender systems. In *Recommender Systems Handbook*, pages 881–918. Springer.
- [Cena et al., 2017] Cena, F., Gena, C., Grillo, P., Kuflik, T., Vernerio, F., and Wecker, A. J. (2017). How scales influence user rating behaviour in recommender systems. *Behaviour & Information Technology*, 0(0) :1–20.
- [Chen et al., 2013] Chen, C. C., Wan, Y.-H., Chung, M.-C., and Sun, Y.-C. (2013). An effective recommendation method for cold start new users using trust and distrust networks. *Information Sciences*, 224 :19–36.
- [Chen and Goodman, 1998] Chen, S. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University.
- [Chen and Goodman, 1996] Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310—318.
- [Chertov et al., 2015] Chertov, O., Brun, A., Boyer, A., and Aleksandrova, M. (2015). Comparative analysis of neighborhood-based approach and matrix factorization in recommender systems. *Eastern-European Journal of Enterprise Technologies*, 3.
- [Chimphlee et al., 2006] Chimphlee, S., Salim, N., Bin Ngadiman, M. S., and Chimphlee, W. (2006). *Using Association Rules and Markov Model for Predit Next Access on Web Usage Mining*, pages 371–376. Springer Netherlands.

- [Chomsky, 1956] Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2 :113–124.
- [Clarkson and Robinson, 1997] Clarkson, P. R. and Robinson, A. J. (1997). Language model adaptation using mixtures and an exponentially decaying cache. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'97)*.
- [Claypool et al., 1999] Claypool, M., Gokhale, A., and Miranda, T. (1999). Combining content-based and collaborative filters in an online newspaper. In *Proceedings of the SIGIR Workshop on Recommender Systems : Algorithms and Evaluation*.
- [Cleger et al., 2014] Cleger, S., Fernández-Luna, J., and Huete, J. (2014). Learning from explanations in recommender systems. *Information Sciences*, 287 :90–108.
- [Cooley et al., 1997] Cooley, R., Mobasher, B., and Srivastava, J. (1997). Web mining : Information and pattern discovery on the world wide web. In *Proceedings of the International Conferences on Tools with Artificial Intelligence*.
- [Cuong et al., 2011] Cuong, P., Cao, Y., Klamma, R., and Jarke, M. (2011). A clustering approach for collaborative filtering recommendation using social network analysis. *Journal of Universal Computer Science (j-jucs)*, 17 :583–604.
- [Daniel, 2015] Daniel, B. (2015). Big data and analytics in higher education : Opportunities and challenges. *British journal of educational technology*, 46(5) :904–920.
- [Daniel, 2017] Daniel, B. K. (2017). Big data in higher education : The big picture. In *Big Data and Learning Analytics in Higher Education*, pages 19–28. Springer.
- [De Bra et al., 2004] De Bra, P., Aroyo, L., and Cristea, A. (2004). *Web Dynamics, Adaptive to Change in Content, Size, Topology and Use*, chapter Adaptive Web-based Educational Hypermedia. Heidelberg.
- [de Campos et al., 2010] de Campos, L. M., Fernández-Luna, J. M., Huete, J. F., and Rueda-Morales, M. A. (2010). Combining content-based and collaborative recommendations : A hybrid approach based on bayesian networks. *International Journal of Approximate Reasoning*, 51(7) :785 – 799.
- [Domingos and Richardson, 2001] Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'01, pages 57–66, New York, NY, USA. ACM.
- [Dujardin et al., 2015] Dujardin, Y., Dietterich, T., and Chades, I. (2015). α -min : A compact approximate solver for finite-horizon pomdps. In *IJCAI*, pages 2582–2588.
- [Durand et al., 2011] Durand, G., LaPlante, F., and Kop, R. (2011). A learning design recommendation system based on markov decision processes. In *17th ACM SIGKDD conference on knowledge discovery and data mining (KDD'11)*.
- [Duval, 2011] Duval, E. (2011). Attention please ! : Learning analytics for visualization and recommendation. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, LAK '11, pages 9–17. ACM.
- [Dwivedi and Bharadwaj, 2015] Dwivedi, P. and Bharadwaj, K. K. (2015). e-learning recommender system for a group of learners based on the unified learner profile approach. *Expert Systems*, 32(2) :264–276.
- [Emami et al., 2003] Emami, A., Xu, P., and Jelinek, F. (2003). Using a connectionist model in a syntactical based language model. In *Proceedings of ICASSP*, pages 372–375.
- [Esslimani et al., 2008a] Esslimani, I., Brun, A., and Boyer, A. (2008a). Behavioral similarities for collaborative recommendations. *Journal of Digital Information Management*, 6(6) :442–448.
- [Esslimani et al., 2008b] Esslimani, I., Brun, A., and Boyer, A. (2008b). Enhancing Collaborative Filtering by frequent usage patterns. In *First International Conference on the Applications of Digital Information and Web Technologies - ICADIWT 2008, Workshop on Recommender Systems and Personalized Retrieval (RSPR)*, pages 180–185, Ostrava, Czech Republic.
- [Esslimani et al., 2009a] Esslimani, I., Brun, A., and Boyer, A. (2009a). A Collaborative Filtering Approach Combining Clustering and Navigational Based Correlations. In *5th International Conference on Web Information Systems and Technologies - WEBIST 2009*, pages 364–369, Lisbonne, Portugal.

- [Esslimani et al., 2009b] Esslimani, I., Brun, A., and Boyer, A. (2009b). From Social Networks to Behavioral Networks in Recommender Systems. In *The 2009 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2009)*, pages 143–148, Athènes, Greece.
- [Esslimani et al., 2010a] Esslimani, I., Brun, A., and Boyer, A. (2010a). Detecting Leaders in Behavioral Networks. In *Proc. of the 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2010)*, pages 281–285, Odense, Denmark. IEEE.
- [Esslimani et al., 2010b] Esslimani, I., Brun, A., and Boyer, A. (2010b). Detecting Leaders to Alleviate Latency in Recommender Systems. In *International Conference on Electronic Commerce and Web Technologies (EC-Web 2010)*, pages 229–240, Bilbao, Spain.
- [Esslimani et al., 2011] Esslimani, I., Brun, A., and Boyer, A. (2011). Densifying a Behavioral Recommender System by Social Networks Link Prediction Methods. *Social Network Analysis and Mining*, 1(3) :159–172.
- [Esslimani et al., 2013] Esslimani, I., Brun, A., and Boyer, A. (2013). *The Influence of Technology on Social Network Analysis and Mining*, volume 6, chapter Towards Leader-Based Recommendations, pages 455–470. Springer.
- [Evans and Lindner, 2012] Evans, J. and Lindner, C. (2012). Business analytics : The next frontier for decision sciences. retrieved on 15 apr. 2017 http://www.cbpp.uaa.alaska.edu/afef/business_analytics.htm.
- [Facca and Lanzi, 2005] Facca, F. M. and Lanzi, P. L. (2005). Mining interesting knowledge from weblogs : a survey. *Data & Knowledge Engineering*, 53(3) :225 – 241.
- [Fahed, 2016] Fahed, L. (2016). *Prédire et influencer l'apparition des événements dans une séquence complexe*. PhD thesis, Université de Lorraine.
- [Fahed et al., 2014a] Fahed, L., Brun, A., and Boyer, A. (2014a). Episode Rules Mining Algorithm for Distant Event Prediction. In *KDIR - 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management.*, rome, Italy.
- [Fahed et al., 2014b] Fahed, L., Brun, A., and Boyer, A. (2014b). Extraction de règles d'épisodes minimales dans des séquences complexes. In *In 14e Conférence Internationale Francophone sur l'Extraction et la Gestion de Connaissance (EGC 2014)*.
- [Fahed et al., 2015a] Fahed, L., Brun, A., and Boyer, A. (2015a). Influencer events in episode rules : a way to impact the occurrence of events. In *19th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*.
- [Fahed et al., 2015b] Fahed, L., Brun, A., and Boyer, A. (2015b). *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, chapter Efficient Discovery of Episode Rules With a Minimal Antecedent and a Distant Consequent, pages 3–18. Springer-Verlag.
- [Fahed et al., 2018] Fahed, L., Brun, A., and Boyer, A. (2018). Deer : Distant and essential episode rules for early prediction. *Expert Systems with Applications*, 93 :283–298.
- [Farabet et al., 2013] Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8).
- [Fayyad et al., 1996] Fayyad, U., Piatetsky-shapiro, G., and Smith, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3).
- [Ferguson, 2012] Ferguson, R. (2012). Learning analytics : drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5-6) :304–317.
- [Gama et al., 2014] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)*, 46(4) :44.
- [Garrido et al., 2016] Garrido, A., Morales, L., and Serina, I. (2016). On the use of case-based planning for e-learning personalization. *Expert Systems With Applications*, 60 :1–15.
- [Goyal et al., 2008] Goyal, A., Bonchi, F., and Lakshmanan, L. V. (2008). Discovering leaders from community actions. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, pages 499–508.

- [Gras et al., 2015a] Gras, B., Brun, A., and Boyer, A. (2015a). Identification des utilisateurs atypiques dans les systèmes de recommandation sociale. In *EGC - Extraction et Gestion de Connaissances*, Esch sur Alzette, Luxembourg.
- [Gras et al., 2015b] Gras, B., Brun, A., and Boyer, A. (2015b). Identifying users with atypical preferences to anticipate inaccurate recommendations. In *Webist 2015 – 11th International Conference on Web Information Systems and Technologies*.
- [Gras et al., 2015c] Gras, B., Brun, A., and Boyer, A. (2015c). *WEBIST (Revised Selected Papers)*, chapter When Users with Preferences Different from Others Get Inaccurate Recommendations. Springer.
- [Gras et al., 2016] Gras, B., Brun, A., and Boyer, A. (2016). Identifying grey sheep users in collaborative filtering : a distribution-based technique. In *ACM UMAP User Modeling and Adaptive Personalisation*.
- [Gras et al., 2017] Gras, B., Brun, A., and Boyer, A. (2017). Can Matrix Factorization Improve the Accuracy of Recommendations Provided to Grey Sheep Users? In *13th International Conference on Web Information Systems and Technologies (WEBIST)*, pages 88 – 96, Porto, Portugal.
- [Gröger et al., 2014] Gröger, C., Schwarz, H., and Mitschang, B. (2014). Prescriptive analytics for recommendation-based business process optimization. In Systems, B. I., editor, *International Conference on Business Information Systems*.
- [Guo et al., 2014] Guo, G., Zhang, J., and Thalmann, D. (2014). Merging trust in collaborative filtering to alleviate data sparsity and cold start. *Knowledge-Based Systems*, 57 :57–68.
- [Han and Kamber, 2006] Han, J. and Kamber, M. (2006). *Data Mining : Concepts and Techniques*. Morgan Kaufmann.
- [Han et al., 2001] Han, J., Pei, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., and Hsu, M. (2001). Prefixspan : Mining sequential patterns efficiently by prefix-projected pattern growth. In *proceedings of the 17th international conference on data engineering*, pages 215–224.
- [Han et al., 2000] Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. *SIGMOD Rec.*, 29(2) :1–12.
- [Henze and Nejdl, 2004] Henze, N. and Nejdl, W. (2004). A logical characterization of adaptive educational hypermedia. *Hypermedia - Special issue : Adaptive hypermedia in the age of the adaptive web*, 10(1) :77–113.
- [Hu et al., 2008] Hu, Y., Koren, Y., and Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 263–272. Ieee.
- [Huang et al., 1992] Huang, X., Allewa, F., wuen Hon, H., yuh Hwang, M., and Rosenfeld, R. (1992). The sphinx-ii speech recognition system : An overview. *Computer, Speech and Language*, 7 :137–148.
- [Huptych et al., 2017] Huptych, M., Bohuslavek, M., Hlosta, M., and Zdrahal, Z. (2017). Measures for recommendations based on past students' activity. In *Proceedings of the Learning Analytics and Knowledge (LAK 17)*.
- [Jannach et al., 2016] Jannach, D., Resnick, P., Tuzhilin, A., and Zanker, M. (2016). Recommender systems — beyond matrix completion. *Communications of the ACM*, 59(11) :94–102.
- [Jawaheer et al., 2010] Jawaheer, G., Szomszor, M., and Kostkova, P. (2010). Comparison of implicit and explicit feedback from an online music recommendation service. In *proceedings of the 1st international workshop on information heterogeneity and fusion in recommender systems*, pages 47–51. ACM.
- [Jelinek, 1985] Jelinek, F. (1985). Self-organized language modeling for speech recognition. Technical report, IBM T.J. Watson Research Center, Continuous Speech Recognition Group.
- [Jelinek, 1997] Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. MIT Press.
- [JISC, 2011] JISC (2011). http://repository.jisc.ac.uk/6560/1/learning-analytics_and_student_success.pdf, retrieved on 18 May 2017.
- [Johnson et al., 2012] Johnson, L., Adams, S., and Cummins, M. (2012). Nmc horizon report : 2012 higher education edition.

- [Jones et al., 2011a] Jones, N., Brun, A., and Boyer, A. (2011a). An Exploratory Work in Using Comparisons Instead of Ratings. In *Proc. of the 12th International Conference on Electronic Commerce and Web Technologies (EC-Web 11)*, pages 184–195.
- [Jones et al., 2011b] Jones, N., Brun, A., and Boyer, A. (2011b). Comparisons Instead of Ratings : Towards More Stable Preferences. In *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT'11)*, pages 451–456.
- [Jones et al., 2011c] Jones, N., Brun, A., and Boyer, A. (2011c). Improving Reliability of User Preferences : Comparing Instead of Rating. In *Proc. of the Sixth International Conference on Digital Information Management (ICDIM 2011)*, pages 316–321, Melbourne, Australia.
- [Jones et al., 2011d] Jones, N., Brun, A., and Boyer, A. (2011d). Initial Perspectives From Preferences Expressed Through Comparisons. In *The 14th International Conference on Human-Computer Interaction - HCI International 2011*, pages 33–37, Orlando, United States.
- [Jurafsky and Martin, 2008] Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing, 2nd Edition*. Prentice Hall.
- [Karampiperis and Sampson, 2005] Karampiperis, P. and Sampson, D. (2005). Adaptive learning resources sequencing in educational hypermedia systems. *Educational Technology and society*, 8(4).
- [Katz and Lazarsfeld, 1955] Katz, E. and Lazarsfeld, P. (1955). *Personal influence : The part played by people in the flow of mass communications*. The Free Press.
- [Katz, 1987] Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3).
- [Kelly and Teevan, 2003] Kelly, D. and Teevan, J. (2003). Implicit feedback for inferring user preference : A bibliography. *SIGIR Forum*, 37(2) :18–28.
- [Konstan and Riedl, 2012] Konstan, J. A. and Riedl, J. (2012). Recommender systems : from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1) :101–123.
- [Koren et al., 2009] Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8) :30–37.
- [Kuflik et al., 2012] Kuflik, T., Wecker, A. J., Cena, F., and Gena, C. (2012). *Evaluating Rating Scales Personality*, pages 310–315. Springer Berlin Heidelberg.
- [Kuhn and De Mori, 1990] Kuhn, R. and De Mori, R. (1990). A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6) :570–583.
- [Lafferty et al., 2001] Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *18th International Conf on Machine Learning (ICML)*, pages 282–289.
- [Langlois et al., 2003] Langlois, D., Brun, A., Smaïli, K., and Haton, J.-P. (2003). Événements impossibles en modélisation stochastique du langage. *Traitement Automatique des Langues*, 44(1) :33–61. Article dans revue scientifique avec comité de lecture. nationale.
- [Laporte, 2005] Laporte, E. (2005). Symbolic Natural Language Processing. In Lothaire, editor, *Applied Combinatorics on Words*, pages 164–209. Cambridge University Press.
- [Lenhart and Herzog, 2016] Lenhart, P. and Herzog, D. (2016). Combining content-based and collaborative filtering for personalized sports news recommendations. In *Proceedings of the 3rd Workshop on New Trends in Content-Based Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2016)*, pages 3–10.
- [Li et al., 2014] Li, J., Zhang, L., Meng, F., and Li, F. (2014). Recommendation algorithm based on link prediction and domain knowledge in retail transactions. *Procedia Computer Science*, 31 :875 – 881. 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014.
- [Li et al., 2007] Li, L., Yang, Z., Wang, B., and Kitsuregawa, M. (2007). *Dynamic Adaptation Strategies for Long-Term and Short-Term User Profile to Personalize Search*, pages 228–240. Springer Berlin Heidelberg.

- [Li and Kim, 003] Li, Q. and Kim, B. (003). Clustering approach to hybrid recommendation. In *Proceedings of IEEE/WIC International Conference on Web Intelligence (WI'03)*.
- [Liben-Nowell and Kleinberg, 2003] Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM'03*, pages 556–559.
- [Lika et al., 2014] Lika, B., Kolomvatsos, K., and Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41(4) :2065 – 2073.
- [Lin et al., 2014] Lin, C., Xie, R., Guan, X., Li, L., and Li, T. (2014). Personalized news recommendation via implicit social experts. *Information Sciences*, 254 :1 – 18.
- [Lin and Ryaboy, 2013] Lin, J. and Ryaboy, D. (2013). Scaling big data mining infrastructure : The twitter experience. *SIGKDD Explor. Newsl.*, 14(2) :6–19.
- [Liu et al., 2011] Liu, N. N., Meng, X., Liu, C., and Yang, Q. (2011). Wisdom of the better few : cold start recommendation via representative based rating elicitation. In *Proceedings of the 5th ACM Conference on Recommender Systems*, pages 37–44. ACM.
- [Long and Siemens, 2011] Long, P. and Siemens, G. (2011). Penetrating the fog : analytics in learning and education. *EDUCAUSE Review*.
- [Mabroukeh and Ezeife, 2010] Mabroukeh, N. R. and Ezeife, C. I. (2010). A taxonomy of sequential pattern mining algorithms. *ACM Computing Survey*, 43(1) :3 :1–3 :41.
- [Majumdar, 2017] Majumdar, A. (2017). <https://elearningindustry.com/creating-elearning-varying-learner-profiles-5-learners-know>, retrieved on 29th August 2017.
- [Maksai et al., 2015] Maksai, A., Garcin, F., and Faltings, B. (2015). Predicting online performance of news recommender systems through richer evaluation metrics. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 179–186. ACM.
- [Mannila et al., 1997] Mannila, H., Toivonen, H., and Verkamo, A. I. (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3) :259–289.
- [Mazumdar et al., 2017] Mazumdar, P., Patra, B., and Babu, K. (2017). An effective poi recommendation in various cold-start scenarios. In *The 22nd International Conference on Management of Data (COMAD)*.
- [McQuiggan and Sapp, 2014] McQuiggan, J. and Sapp, A. W. (2014). *Implement, Improve and Expand Your Statewide Longitudinal Data System : Creating a Culture of Data in Education*. Wiley.
- [Middleton et al., 2004] Middleton, S., Shadbolt, N., and De Roure, D. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, 22(1) :54–88.
- [Min and Zhu, 2013] Min, F. and Zhu, W. (2013). Granular association rules for multi-valued data. In *26th Annual IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–5. IEEE.
- [Mobasher, 2007] Mobasher, B. (2007). *Data Mining for Web Personalization*, chapter 3, pages 90–135. LNCS 4321 - Brusilovsky, P. and Kobsa, A. and Nejdl, W.
- [Mobasher et al., 2002] Mobasher, B., Dai, H., Luo, T., and Nakagawa, M. (2002). Mining sequential and non-sequential patterns in predictive web usage mining tasks. In *ICDM*, pages 669–672.
- [Nakagawa and Mobasher, 2003] Nakagawa, M. and Mobasher, B. (2003). Impact of Site Characteristics on Recommendation Models Based On Association Rules and Sequential Patterns. In *Intelligent Techniques for Web Personalization*.
- [Narvekar and Banu, 2015] Narvekar, M. and Banu, S. S. (2015). Predicting user’s web navigation behavior using hybrid approach. *Procedia Computer Science*, 45 :3 – 12. International Conference on Advanced Computing Technologies and Applications (ICACTA).
- [Newell et al., 1998] Newell, A., Langer, S., and Hickey, M. (1998). The role of natural language processing in alternative and augmentative communication. *Natural Language Engineering*, 4(1) :1–16.

- [Niemann and Wolpers, 2013] Niemann, K. and Wolpers, M. (2013). A new collaborative filtering approach for increasing the aggregate diversity of recommender systems. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 955–963. ACM.
- [Nowakowski et al., 2012] Nowakowski, S., Bernard-Issenmann, N., Cherqui-Houot, I., and Brun, A. (2012). Technical and pedagogical feedback on the deployment of an ePortfolio. Models of the uses, analysis and perspectives. In *10th ePortfolio and Identity conference - ePIC 2012*, pages 177–185, Londres, United Kingdom. ISBN : 9782954014418.
- [Nowakowski et al., 2009] Nowakowski, S., Boyer, A., Brun, A., Skaf-Molli, H., Dinet, J., Gicquel, R., and Antoine, A. (2009). P2CeL - Collaborative Knowledge Construction and eLearning : an Approach Based on Semantic Wikis. In *Online Educa - 15th International Conference on Technology Supported Learning and Training*, Berlin, Germany.
- [Nowakowski et al., 2010a] Nowakowski, S., Brun, A., and Boyer, A. (2010a). Towards Recommender Systems Based on Kalman Filters - A new Approach by State Space Modeling. In *6th International Conference on Web Information Systems and Technologies - WEBIST 2010*, pages 345–349, Valencia, Spain.
- [Nowakowski et al., 2010b] Nowakowski, S., Brun, A., Boyer, A., Skaf-Molli, H., Gicquel, R., Dinet, J., and Antoine, A. (2010b). Production collaborative de connaissances et eLearning : une approche par wikis sociaux sémantiques. In *7ème Colloque Technologies de l'Information et de la Communication pour l'Enseignement - TICE 2010*, Nancy, France.
- [Palchenko et al., 2017] Palchenko, O., Brun, A., Boyer, A., and Chertov, O. (2017). Using n-step matrix factorization for solving new user cold-start problem. In *9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems : Technology and Applications (IDAACS'17)*.
- [Pan et al., 2010] Pan, W., Xiang, E. W., Liu, N. N., and Yang, Q. (2010). Transfer learning in collaborative filtering for sparsity reduction. In *Association for the Advancement of Artificial Intelligence 2010*.
- [Pei et al., 2007] Pei, J., Han, J., and Wang, W. (2007). Constraint-based sequential pattern mining : the pattern-growth methods. *Journal of Intelligent Information Systems*, 28(2) :133–160.
- [Pereira and Hruschka, 2015] Pereira, A. L. V. and Hruschka, E. R. (2015). Simultaneous co-clustering and learning to address the cold start problem in recommender systems. *Knowledge-Based Systems*, 82 :11 – 19.
- [Perrin et al., 2012] Perrin, E., Brun, A., and Boyer, A. (2012). Utilisation d'invariants pour une médiation inter-domaines de modèles utilisateurs : ressources invariantes et invariants sémantiques. In *12e Conférence Internationale Francophone sur l'Extraction et la Gestion de Connaissance (EGC 2012)*.
- [Pessiot et al., 2006] Pessiot, J.-F., Truong, T.-V., Usunier, N., Amini, M.-R., and Gallinari, P. (2006). Factorisation en matrices non-négatives pour le filtrage collaboratif. In *Proceedings of 3rd Conference en Recherche d'Information et Applications*, pages 315–326.
- [Pitkow and Pirolli, 1999] Pitkow, J. and Pirolli, P. (1999). Mining Longest Repeating Subsequences to Predict World Wide Web Surfing. In *USITS'99 : Proceedings of the 2nd conference on USENIX Symposium on Internet Technologies and Systems*, pages 139–150.
- [Ponte and Croft, 1998] Ponte, J. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the ACM SIGIR'98*, pages 275–281.
- [Potamianos and Jelinek, 1998] Potamianos, G. and Jelinek, F. (1998). A study of n-gram and decision tree letter language modeling methods. *Speech Communication*, 24(3) :171–192.
- [Quillian, 1963] Quillian, R. (1963). A notation for representing conceptual information : An application to semantics and mechanical english paraphrasing. sp-1395,. System Development Corporation.
- [Rabiner, 1989] Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *IEEE*, pages 257–286.
- [Rafferty et al., 2015] Rafferty, A., Brunskill, E., Griffiths, T., and Shafto, P. (2015). Faster teaching via pomdp planning. *Cognitive Science*, pages 1–43.

- [Rajni and Malaya, 2015] Rajni, J. and Malaya, D. B. (2015). Predictive analytics in a higher education context. *IT Professional*, 17(4) :24–33.
- [Ransbotham et al., 2015] Ransbotham, S., Kiron, D., and Prentice, P. K. (2015). Minding the analytics gap. *MIT Sloan Management Review*.
- [Rashid, 2007] Rashid, A. M. (2007). *Mining influence in recommender systems*. PhD thesis, University of Minnesota.
- [Razmerita and Brun, 2010] Razmerita, L. and Brun, A. (2010). Assigning Students in Groups : Self-formed Groups versus Automatically-formed Groups. In *7ème Colloque Technologies de l’Information et de la Communication pour l’Enseignement - TICE 2010*, Nancy, France.
- [Razmerita and Brun, 2011] Razmerita, L. and Brun, A. (2011). Collaborative Learning in Heterogeneous Classes : Towards a Group Formation Methodology. In *International Conference on Computer Supported Education (CSEDU 2011)*, pages 189–194, Noordwijkerhout, Netherlands.
- [Reilly and Sharkey, 1992] Reilly, R. and Sharkey, N. (1992). Connectionist approaches to natural language. *The American Journal of Psychology*, 107(2) :291–299.
- [Reynolds, 2016] Reynolds, G. (2016). *Information Technology for Managers*. Cengage Learning. 2nd edition.
- [Rosenfeld, 2000] Rosenfeld, R. (2000). Two decades of statistical language modeling : Where do we go from here? *Proceedings of the IEEE*, 88(8).
- [Ross et al., 2008] Ross, S., Pineau, J., Paquet, S., and Chaib-draa, B. (2008). Online planning algorithms for pomdps. *Journal of Artificial Intelligence Research*, 32(2) :663–704.
- [Sahebi and Cohen, 2011] Sahebi, S. and Cohen, W. W. (2011). Community-based recommendations : a solution to the cold start problem. In *Workshop on recommender systems and the social web, RSWEB*, page 60.
- [Sarwar et al., 2002] Sarwar, B.M. and Karypis, G., Konstan, J., and Riedl, J. (2002). Recommender systems for large-scale e-commerce : Scalable neighborhood formation using clustering. In *The 5th Int. Conf. on Computer and Information Technology*.
- [Saveski and Mantrach, 2014] Saveski, M. and Mantrach, A. (2014). Item cold-start recommendations : Learning local collective embeddings. In *Proceedings of the 8th ACM Conference on Recommender Systems, RecSys’14*, pages 89–96. ACM.
- [Schein et al., 2002] Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR’02*, pages 253–260. ACM.
- [Shani and Gunawardana, 2011] Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. *Recommender systems handbook*, pages 257–297.
- [Shani et al., 2005] Shani, G., Heckerman, D., and Brafman, R. (2005). An MDP-Based Recommender System. *JMLR : The Journal of Machine Learning Research*, pages 453–460.
- [Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27 :398–403.
- [Smaïli et al., 1999] Smaïli, K., Brun, A., Zitouni, I., and Haton, J. (1999). Automatic and Manual Clustering for Large Vocabulary Speech Recognition : A Comparative Study. In *European Conference on Speech Communication and Technology (EUROSPEECH’99)*, pages 1795–1798, Budapest, Hungary.
- [Soltanpoor and Sellis, 2016] Soltanpoor, R. and Sellis, T. (2016). Prescriptive analytics for big data. In *Australasian Database Conference*.
- [Son, 2016] Son, L. H. (2016). Dealing with the new user cold-start problem in recommender systems : A comparative review. *Information Systems*, 58 :87 – 104.
- [Souza, 2014] Souza, G. (2014). Supply chain analytics. *Business Horizon*, 57 :595–605.
- [Srikant and Agrawal, 1996] Srikant, R. and Agrawal, R. (1996). Mining sequential patterns : Generalizations and performance improvements. In *EDBT*, pages 3–16.

- [Takács et al., 2008] Takács, G., Pilászy, I., Németh, B., and Tikk, D. (2008). Investigation of various matrix factorization methods for large recommender systems. In *Proceedings of the 2Nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, NETFLIX '08, pages 6 :1–6 :8. ACM.
- [Terveen and Hill, 2001] Terveen, L. and Hill, W. (2001). Beyond recommender systems : Helping people help each other. *HCI in the New Millenium*, pages 487–509.
- [Tintarev and Masthoff, 2007] Tintarev, N. and Masthoff, J. (2007). A survey of explanations in recommender systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*, ICDEW '07, pages 801–810.
- [Toussaint et al., 2008] Toussaint, M., Charlin, L., and Poupart, P. (2008). Hierarchical pomdp controller optimization by likelihood maximization. In *UAI*, volume 24, pages 562–570.
- [Tsymbal, 2004] Tsymbal, A. (2004). The problem of concept drift : definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2).
- [Ullrich and Melis, 2009] Ullrich, C. and Melis, E. (2009). Pedagogically founded courseware generation based on htn-planning. *Expert Systems with Applications*, 36(5).
- [Valente, 1995] Valente, T. (1995). *Network Models of the Diffusion of Innovations*. Hampton Press.
- [Van Barneveld et al., 2012] Van Barneveld, A., Arnold, K. E., and Campbell, J. P. (2012). Analytics in higher education : Establishing a common language. *EDUCAUSE learning initiative*, 1(1) :1–11.
- [Vassileva and Deters, 1998] Vassileva, J. and Deters, R. (1998). Dynamic courseware generation on the www. *Journal of Educational Technology*, 29(1).
- [Vozalis and Margaritis, 2004] Vozalis, M. and Margaritis, K. (2004). Collaborative filtering enhanced by demographic correlation. In *Proceedings of the International Conference on Intelligent Systems Design and Applications*.
- [Wang et al., 2003] Wang, H., Fan, W., Yu, P. S., and Han, J. (2003). Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 226–235. ACM.
- [Wang et al., 2012] Wang, W., Zhang, D., and Zhou, J. (2012). *COBA : A Credible and Co-clustering Filterbot for Cold-Start Recommendations*, pages 467–476. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Webb et al., 2016] Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., and Petitjean, F. (2016). Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4) :964–994.
- [Wei and Yan, 2009] Wei, X. and Yan, J. (2009). Learner profile design for personalized e-learning systems. In *Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on*, pages 1–4. IEEE.
- [Whittaker and Woodland, 1998] Whittaker, E. and Woodland, P. (1998). Comparison of language modelling techniques for russian and english. In *Proceedings of ICSLP'98*.
- [Wibowo, 2016] Wibowo, A. T. (2016). Generating pseudotransactions for improving sparse matrix factorization. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 439–442.
- [Wilson et al., 2017] Wilson, A., Watson, C., Thompson, T. L., Drew, V., and Doyle, S. (2017). Learning analytics : challenges and limitations. *Teaching in Higher Education*, pages 1–17.
- [Wright et al., 2015] Wright, A., Wright, A., McCoy, A., and Sittig, D. (2015). The use of sequential pattern mining to predict next prescribed medications. *Journal of Biomedical Informatics*, 53 :73–80.
- [Wu et al., 2015] Wu, D., Lu, J., and Zhang, G. (2015). A fuzzy tree matching-based personalized e-learning recommender system. *IEEE Transactions on Fuzzy Systems*, 23(6) :2412–2426.
- [Xie et al., 2015] Xie, F., Chen, Z., Shang, J., Feng, X., and Li, J. (2015). A link prediction approach for item recommendation with complex number. *Knowledge-Based Systems*, 81 :148 – 158.
- [Yang and Fong, 2015] Yang, H. and Fong, S. (2015). Countering the concept-drift problems in big data by an incrementally optimized stream mining model. *Journal of Systems and Software*, 102 :158–166.

- [Yoo et al., 2012] Yoo, K., Gretzel, U., and Zanker, M. (2012). *Persuasive Recommender Systems*. Springer.
- [Zaki, 1998] Zaki, M. J. (1998). Efficient enumeration of frequent sequences. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, CIKM'98, pages 68–75.
- [Zaki, 2000] Zaki, M. J. (2000). Sequence mining in categorical domains : Incorporating constraints. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, CIKM'00, pages 422–429. ACM.
- [Zhang et al., 2006] Zhang, S., Wang, W., Ford, J., and Makedon, F. (2006). Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of the 6th SIAM Conference on Data Mining*, volume 6, pages 548–552.
- [Zhang et al., 2010] Zhang, Y., Cao, B., and yan Yeung, D. (2010). Multi-domain collaborative filtering. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 725–732.
- [Zimdars et al., 2001] Zimdars, A., Chickering, D. M., and Meek, C. (2001). Using Temporal Data for Making Recommendations. In Breese, J. S. and Koller, D., editors, *UAI*, pages 580–588. Morgan Kaufmann.

Résumé

La modélisation empirique, reposant sur des données de réalisation ou de traces, est une approche de modélisation de phénomènes, systèmes ou objets, et a la caractéristique de s'intéresser à la "réalité" de ces derniers. Les travaux de recherche que je mène s'intéressent à la modélisation descriptive et prédictive. Je me suis intéressée à des problématiques générales telles que la robustesse, la complexité, et la qualité des modèles, tout en me focalisant sur des défis plus spécifiques tels que le démarrage à froid et le manque général de données, mais aussi l'identification de facteurs influents ou explicatifs au sein des données.

Mes contributions ont été appliquées et validées principalement sur en contexte e-commerce et plus récemment en éducation : traces de comportement, de préférences, etc.

Mes recherches futures iront un pas plus loin dans la modélisation, et auront pour objectif la modélisation prescriptive : que faire pour arriver à un but fixé ? Des aspects relatifs à la transparence et à l'explicabilité des algorithmes, de même qu'à la gestion de sources de données multiples seront au cœur de ces travaux.

Abstract

Empirical modelling, which relies on data, also referred to as traces, is an approach for modelling phenomena, systems or objects. It has the characteristics of modeling the "reality" of these phenomena. The researches I have conducted are dedicated to both descriptive and predictive modelling. They focused on robustness, complexity and quality of the models, but also on the identification of triggering or explanatory factors in data.

My contributions have been applied and validated in the frame of e-commerce and, more recently, on e-education through the use of traces of behavior, of preferences, etc.

My future research goes a step further and will focus on prescriptive modelling : what can be done to reach a given objective ? Some considerations related to algorithms that can explain themselves and that are transparent (explainable AI), as well as the management of multiple sources of data will be studied.