UNE APPROCHE BASÉE SUR L'ANALYSE DES SÉQUENCES POUR LA RECONNAISSANCE DES ACTIVITÉS ET COMPORTEMENTS DANS LES ENVIRONNEMENTS INTELLIGENTS

par

Belkacem Chikhaoui

Thèse présentée au Département d'informatique
en vue de l'obtention du grade de philosophiæ doctor (Ph.D.)

FACULTÉ DES SCIENCES

UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, 17 décembre 2013

Le 09-12-2013

le jury a accepté la thèse de Monsieur Belkacem Chikhaoui
dans sa version finale.

Membres du jury

Professeur Shengrui Wang
Directeur de recherche
Département d'informatique

Professeure Hélène Pigot
Co-directrice de recherche
Département d'informatique

Professeur Abdenour Bouzouane
Évaluateur externe
Université de Chicoutimi

Professeur Jean-Pierre Dussault
Évaluateur interne
Département d'informatique

Professeur Marc Frappier
Président rapporteur
Département d'informatique

Sommaire

Cette thèse vise à étudier deux problématiques différentes : 1) la reconnaissance des activités de la vie quotidienne des personnes dans un habitat intelligent, et 2) la construction du profil comportemental de la personne. Nos contributions sont présentées dans deux chapitres illustrant les solutions proposées. La première contribution de cette thèse est liée à l'introduction d'une nouvelle approche non supervisée de reconnaissance d'activités nommée ADR-SPLDA (Activity Discovery and Recognition using Sequential Patterns and Latent Dirichlet Allocation). Contrairement aux approches existantes, ADR-SPLDA permet la découverte et la reconnaissance des activités de façon non supervisée sans faire nécessairement recours à l'annotation des données. En outre, ADR-SPLDA est basée sur l'analyse de patrons fréquents, ce qui permet de réduire significativement la quantité du bruit dans les données. La fiabilité de ADR-SPLDA est illustrée à travers une série de tests et de comparaisons avec les approches existantes sur une variété de données réelles.

Le deuxième travail vise la construction du profil comportemental de la personne en se basant sur ses activités. Nous avons développé une approche qui permet de découvrir les différents comportements dans les séquences, et d'extraire les relations causales entre les différents comportements. Notre contribution inclut l'introduction de l'analyse causale dans la construction du profil, ce qui nous a permis aussi de découvrir les relations causales entre les différentes activités. Une série de tests a été également effectuée pour illustrer la fiabilité de notre approche sur une variété de données. Le travail de recherche entrepris dans cette thèse constitue l'une des nombreuses étapes importantes dans l'accomplissement d'un système d'assistance efficace dans l'objectif d'assurer le bien-être des personnes.

Remerciements

Je tiens à remercier sincèrement mes directeurs de recherche, le professeur Shengrui Wang, et la professeure Hélène Pigot, pour la pertinence de leur encadrement, leurs valeureux conseils et encouragements, et surtout pour leurs qualités humaines. Leur rigueur scientifique, leurs conseils judicieux et leur disponibilité tout au long de cette thèse ont joué un rôle déterminant et m'ont permis de mener à terme ce travail de recherche avec les qualités escomptées. Avec mes directeurs de recherche, j'ai appris la flexibilité, l'efficacité et la patience. Ils demeurent des professeurs envers qui j'ai beaucoup de respect.

Je remercie également le Conseil de Recherches en Sciences naturelles et en Génie du Canada pour le soutien financier durant deux ans.

Je ne peux oublier de remercier ma famille : ma mère, mon père, mes frères et ma soeur pour leur soutien et leurs encouragements durant toute cette période.

Mes vifs remerciements vont particulièrement à ma femme. Merci de m'avoir soutenu et encouragé dans les moments difficiles. Merci d'être aussi patiente avec moi. Merci pour tout.

Abréviations

LDA Latent Dirichlet Allocation

HMM Hidden Markov Model

BN Réseau Bayésien (Bayesian Network)

NB Bayésien Naif (Naive Bayes)

TE Transfert d'Entropie

Table des matières

Liste des figures

Liste des figures

Liste des tableaux

Liste des tableaux

Introduction

"***Une personne qui n'a jamais commis d'erreurs n'a jamais tenté d'innover***". Albert Einstein

0.1 Problématique

Dans les pays occidentaux, la baisse de la natalité et l'augmentation de l'espérance de vie se traduisent par la modification de la pyramide de l'âge et une augmentation de la proportion des personnes âgées. Selon le rapport de la deuxième assemblée mondiale sur le vieillissement qui s'est déroulée à Madrid en avril 2002, le nombre des personnes âgées devrait d'ici 2050 dépasser celui des jeunes et ce pour la première fois dans l'histoire de l'humanité [30].

L'une des principales préoccupations associées à la vieillesse est le risque de dépendance associée à la maladie. Cependant, en raison du problème grandissant posé par les soins de plus en plus coûteux, et des besoins accrus en terme d'assistance, on craint qu'une population vieillissante n'entraîne une charge supplémentaire pour les services de santé déjà surchargés.

Outre les maladies chroniques, c'est sans doute l'affaiblissement intellectuel que les personnes âgées craignent le plus. Ceci engendre une perte d'autonomie dans l'accomplissement des activités de la vie quotidiennes, qui entraîne une augmentation progressive du niveau de dépendance. Selon les statistiques de l'Institut de Vieillissement du Canada[1], il est estimé que 16 % des personnes âgées de plus de 65 ans souffrent de troubles cognitifs, 8 % d'entre elles sont atteintes de maladies neurodégénératives comme la démence. Cette prévalence augmente de façon significative avec

1. http ://www.cihr-irsc.gc.ca/f/8671.html

l'âge, atteignant 30 % (troubles cognitifs) et 35 % (démence) chez les personnes de plus de 85 ans. De tels problèmes menacent la qualité de vie de ces personnes.

Au Canada, cependant, les services de santé sont essentiellement conçus pour le traitement à court terme des maladies aiguës. Cela pose une question très importante : Qui s'occupe des personnes âgées et plus précisément des personnes présentant des troubles cognitifs ?

La dépendance est la première conséquence de la vieillesse. La dépendance physique et/ou mentale empêche la personne âgée de vivre seule dans un logement ordinaire. Elle se traduit soit par le placement de la personne en institution, soit par le besoin d'un soutien ou aide d'un proche. Durant les années à venir, l'augmentation du nombre de personnes souffrant de troubles cognitifs ou d'autres maladies liées à l'âge va aggraver la surcharge de travail et la pénurie de personnels soignants (infirmière, médecin, spécialiste, assistants sociaux etc.) dans les établissements médico-sociaux. Pour faire face au problème de surcharge des établissements et pour répondre aux souhaits des malades de rester chez eux, le maintien à domicile apparaît comme une solution socio-économique en faveur de ce type de population. Toutefois, pour rester plus longtemps à domicile dans de bonnes conditions, quatre critères doivent être réunis. Ils concernent la santé, l'environnement familial, le niveau de ressources et l'habitat [31].

- L'état de santé : c'est le critère le plus important dans le choix du maintien à domicile. En effet, la dégradation de l'état de santé, entraînant une incapacité totale ou partielle, implique un placement en institution. Les troubles physiques ne sont pas toujours un obstacle au maintien à domicile. Or, les troubles mentaux sont trop difficiles à prendre en charge.
- La famille : elle constitue la clé du maintien à domicile malgré le niveau de dépendance. Pour favoriser le maintien à domicile, les professionnels viennent compléter l'aide familiale sans pour autant s'y substituer.
- Le niveau de ressources (coûts) : le maintien à domicile n'est pas toujours possible en raison du coût d'une aide professionnelle rémunérée. En France par exemple, au moment de l'évaluation de la dépendance (dans le cadre de la Prestation Expérimentale Dépendance), les coûts pour les cas les plus lourds ont été estimés à plus de 12 000 Francs par mois) [31].

– L'habitat : s'il est inadapté, il peut accélérer le placement en institution. L'adaptation du logement permet à la personne âgée de conserver une partie de son autonomie.

Au début, l'idée du maintien à domicile supposait un déplacement régulier du personnel médical pour fournir aux malades des soins médicaux ou paramédicaux de qualité similaires à ceux donnés dans les institutions de soins. Toutefois, cela exige un soutien très fort de la part de la famille. Une solution consiste alors à impliquer l'environnement physique pour qu'il assiste la personne. Cette solution introduit un nouveau système de maintien à domicile dans lequel l'environnement joue un rôle important dans le système de soins. Grâce aux nouvelles technologies de l'information et de la télécommunication, l'environnement est devenu l'interlocuteur entre la personne à son domicile et ses soignants qui se trouvent sur des sites différents, voire un acteur dans le système de soins. Cette tendance peut être garantie par la mise en place d'un système informatique, qui surveille la personne et n'alerte le personnel médical qu'en cas d'urgence. Les soins fournis à domicile constituent une alternative à l'hospitalisation et au recours aux établissements d'hébergement de longue durée. Mais le problème des personnes atteintes de troubles cognitifs n'est pas seulement de leur offrir des soins médicaux, mais aussi d'assurer leur sécurité et de les assister dans les activités de la vie quotidienne. Cette problématique a donné naissance au concept d'habitat intelligent.

Un habitat intelligent est défini par sa capacité à réagir à ce qui se passe dans son environnement. Il doit être capable d'identifier toute situation inhabituelle ou inadéquate et de fournir l'aide appropriée à l'occupant en cas de besoin. En d'autres termes, un habitat intelligent peut s'adapter aux besoins de l'occupant. Plusieurs projets de recherche sur les habitats intelligents sont lancés à travers le monde afin de favoriser le maintien à domicile. Le tableau 1 présente un exemple de ces habitats intelligents.

Dans un habitat intelligent, l'assistance aux personnes provient de l'intérieur en analysant et interprétant les données issues des différents capteurs par un système informatique intelligent. En effet, l'habitat intelligent perçoit les actions de la personne et analyse l'adéquation de ce qui a été fait et ce qui devrait l'être. On distingue deux types d'assistance selon la nature du déficit de la personne : une assistance physique

tableau 1 – Quelques habitats intelligents construits à travers le monde

Nom de l'habitat intelligent	**Nom de l'université**
CASAS	Université de Washington State
DOMUS	Université de Sherbrooke
PlaceLab	Massachusetts Institute of Technology(MIT)
Aware Home	Georgia Tech
ISLab	Université d'Amsterdam
iSpace/iDorm	Université d'Essex
MARC	Université de Virginia
Gator Tech	Université de Florida
Tiger Place	Université de Missouri

et une assistance cognitive.

L'assistance physique consiste à compenser le déficit physique (handicap) de la personne en utilisant des moyens appropriés (commande à distance, chaise roulante, robot mobile, bras manipulateur, etc.). L'assistance cognitive consiste à fournir de l'aide à la personne pour pallier aux déficits cognitifs causés par les maladies telles que l'Alzheimer, la schizophrénie, les traumatismes crâniens et la déficience intellectuelle. Elle permet de rappeler au besoin, les activités à réaliser et les procédures pour les faire. Toutefois, dans des situations délicates, une aide extérieure est nécessaire particulièrement si les risques sont trop élevés [103]. Donc, le système d'assistance cognitive doit être en mesure d'analyser les différentes situations, ce qui peut être réalisé par l'intégration de mécanismes de reconnaissance d'activités, d'apprentissage et de raisonnement très pertinents.

La reconnaissance d'activités constitue un domaine clé de recherche dans les environnements intelligents. Des travaux sur la reconnaissance d'activités dans les environnements intelligents ont été réalisés dans la dernière décennie. Ces travaux peuvent être divisés en : 1) approche invasive dans laquelle les analyses d'images et vidéos sont utilisées ; 2) approche non-invasive dans laquelle le système de reconnaissance est basé principalement sur l'analyse des données issues de capteurs non-invasifs disséminés dans l'environnement. L'approche non-invasive préserve mieux l'intimité des personnes car aucune caméra n'est utilisée pour capter ce qui se passe à l'intérieur de l'habitat intelligent.

0.1. Problématique

Dans le cas des systèmes basés sur les approches non-invasives, la personne ne doit en aucun cas sentir qu'elle est surveillée ou contrôlée. Cela peut se faire par l'utilisation de capteurs non-invasifs tels que les capteurs infrarouges, électromagnétiques, RFID, pression, humidité, température, tapis tactiles, électricité, ..., etc. C'est cette approche qui sera retenue dans ce travail. À noter que la reconnaissance d'activités dans cette approche est une tâche difficile. Les algorithmes existants pour la reconnaissance d'activités peuvent être caractérisés par les méthodes de représentation et de classification comme suit :

1. la reconnaissance d'activités en utilisant les méthodes statistiques comme les modèles Bayesiens ou Markoviens,
2. la reconnaissance d'activités en utilisant les méthodes de classification supervisée ou non supervisée,
3. la reconnaissance d'activités par l'analyse des séquences d'événements.

Toutes ces approches rencontrent quelques difficultés pour faire face aux défis associés aux problèmes de bruit, de données manquantes, l'incertitude de données, l'annotation de données, la progression temporelle des activités ainsi que la reconnaissance d'activités concurrentes et entrelacées. Les activités concurrentes sont des activités qui s'exécutent en parallèle. Par exemple, prendre un café en regardant la télévision. Les activités entrelacées sont des activités pour lesquelles l'utilisateur peut alterner entre les étapes de réalisation de plusieurs activités.

Dans le contexte de reconnaissance d'activités dans les environnements intelligents, la problématique de ce projet de recherche consiste à créer une approche non-invasive de reconnaissance d'activités permettant de remédier aux limites des systèmes existants. Elle devra aussi être capable de reconnaître des activités plus complexes comme les activités les plus longues. Cette problématique s'inscrit dans le contexte général de forage de données. Dans ce projet, le forage de données est placé au coeur de la problématique à cause de la très grande quantité d'événements émis par des capteurs dans un environnement intelligent. L'approche proposée dans ce projet ainsi que la démarche adoptée doivent permettre de dessiner un cadre théorique pour cette problématique, et de proposer une base solide pour un processus complexe de reconnaissance d'activités.

Par ailleurs, le développement, le test et la validation d'un système de reconnaissance d'activités requiert un environnement équipé d'une infrastructure matérielle et logicielle permettant de détecter et de recueillir les informations sur les activités en cours de réalisation. Cette infrastructure permet également de fournir une aide appropriée lorsqu'un comportement inadéquat est détecté ou une aide est sollicitée pour la personne. À cet effet, le laboratoire DOMUS [2] dispose d'un appartement intelligent à la fine pointe de la technologie où l'on retrouve un équipement domiciliaire de base. Il s'agit d'un 4 et 1/2 construit à l'intérieur des murs de l'université, pouvant loger une personne seule. Cet appartement est équipé d'une technologie de pointe dispersée dans l'environnement. Les capteurs, haut-parleurs, écrans d'ordinateurs et de télévisions, les réseaux filaires et non filaires et les serveurs transforment cet appartement en un environnement intelligent capable de s'adapter aux actions de l'occupant. Ce type d'appartement s'adresserait à des personnes présentant des troubles cognitifs qui les empêchent de mener à bien les activités de la vie quotidienne. Par conséquent, les projets du laboratoire Domus s'adressent à une population bien ciblée comme les personnes atteintes de démence de type Alzheimer, les clientèles de type schizophrène, traumatisé crânien, personnes âgées, etc. L'habitat intelligent de Domus servira comme un champ d'expérimentation et de validation pour toutes les applications développées au profit de cette population.

0.2 Objectifs de recherche

Notre projet de recherche s'est déroulé en trois principales étapes comportant chacune des objectifs bien déterminés.

La première étape du projet avait pour but de faire une analyse et investigation approfondie de notre problématique. Cela a pu être effectué par une revue des principales approches de reconnaissance d'activités dans la littérature. Cette phase s'est déroulée en deux volets. D'abord, le premier volet consistait à faire une étude ciblée des travaux proposés dans la littérature pour la reconnaissance d'activités de façon non invasive. Le respect de la vie privée des personnes et leur intimité est très important dans un système d'assistance à domicile, et constitue l'une des conditions primaires pour

2. http ://domus.usherbrooke.ca

l'acceptabilité des systèmes d'assistance. Beaucoup d'approches se sont basées sur l'analyse d'images et vidéos collectées en utilisant des caméras pour développer des systèmes d'assistance. Cependant, ces approches respectent peu l'intimité des gens et risquent de ne pas être utilisables dans la pratique. Il y a lieu de signaler que le fait de ne pas utiliser des caméras compliquera davantage le problème de reconnaissance. Plus particulièrement, lorsque nous n'avons pas suffisamment d'information sur les objets dans l'habitat intelligent. Cela constitue l'un des défis majeurs des approches de reconnaissance d'activités auquel nous devons faire face. Ce premier volet de la recherche nous a permis, d'un coté, de mieux cerner les besoins inhérents à notre contexte applicatif, et, de l'autre, d'effectuer une étude bibliographique orientée.

Par la suite, le second volet consistait à faire un état de l'art sur les différentes approches non invasives de reconnaissance d'activités. Le résultat de cette étude bibliographique, qui sera présenté dans le chapitre suivant, nous a permis de dégager les limites des différentes approches, et d'entrevoir les différentes pistes de solutions que nous pouvions explorer pour faire face aux limites des approches existantes. En conséquence, cette première phase nous a permis de constater que la vaste majorité des approches proposées souffrent du problème d'annotation des données. De plus, nous avons également pu constater qu'il existait très peu de travaux qui utilisent les approches non supervisées pour la reconnaissance d'activités [111, 48]. Pourtant, de telles approches constituent l'une des principales pistes pour résoudre le problème d'annotation des données. Donc, cette phase nous a permis de mieux comprendre notre problématique de recherche, et d'évaluer les différentes pistes de solutions à la lumière de nos conclusions tirées de notre état de l'art.

L'objectif de la deuxième étape de notre projet consistait, suite aux investigations et conclusions précédentes, à proposer une approche permettant la découverte et la reconnaissance d'activités. Cette étape s'est déroulée aussi sur deux volets. Dans le premier volet, et pour répondre à nos besoins, nous avons opté pour le développement d'une approche non supervisée de reconnaissance d'activités afin de surmonter les limites dont souffrent les approches existantes. À partir de là, nous avons choisi de formaliser notre approche en s'appuyant sur des modèles statistiques combinés avec des modèles de forage de données. Plus précisément, cette phase nous a permis de poser les fondements théoriques de notre approche et de créer un modèle formel

de découverte et de reconnaissance d'activités. Ensuite, le second volet consistait à valider l'approche développée dans le premier volet.

Pour mettre en oeuvre une approche systématique de reconnaissance d'activités, plusieurs paramètres doivent être réunis à savoir l'environnement, l'infrastructure matérielle et logicielle, les données et les personnes avec lesquelles se déroulent les expérimentations. Comme nous l'avons mentionné auparavant, le laboratoire Domus dispose des moyens matériels et logiciels ainsi que de l'environnement dans lequel peuvent se dérouler des expérimentations. Beaucoup d'expérimentations ont été faites dans le laboratoire Domus avec des personnes saines et d'autres présentant des déficits cognitifs légers. Nous avons utilisé les données qui correspondent à la réalisation des activités de la vie quotidiennes et qui ont été obtenues dans le cadre d'un projet de recherche au laboratoire Domus [68]. Nous avons également utilisé d'autres données qui sont disponibles sur le Web notamment au laboratoire CASAS à Washington State University [3], PlaceLab au MIT [4], et StarHome [5].

Les résultats de ces expérimentations nous ont permis d'évaluer les performances de l'approche développée selon différents aspects, de comparer les résultats obtenus de notre approche avec d'autres approches existantes, et finalement de dresser les voies de développement futur pour l'amélioration et l'extension de notre approche.

L'objectif de la troisième et ultime étape du notre projet consistait à proposer une approche permettant la construction du profil comportemental des personnes. En effet, nos investigations précédentes nous ont amené à conclure que la mise en oeuvre d'un système d'assistance et de maintien à domicile requiert non seulement un système de reconnaissance d'activités, mais aussi, un système de profilage permettant de dresser un modèle global sur les comportements des personnes lors de la réalisation des activités de la vie quotidiennes, que se soient des comportements normaux ou aberrants. Par conséquent, l'étude du profil comportemental de la personne est une étape importante et centrale dans le développement d'un système sophistiqué d'assistance et de maintien à domicile.

Cette phase s'est déroulée en trois principaux volets. Le premier volet consistait à

3. $http://ailab.eecs.wsu.edu/casas/$
4. $http://architecture.mit.edu/house_n/$
5. $http://www.imada.sdu.dk/\sim gu/$

faire une étude ciblée des travaux proposés dans la littérature pour l'étude du profil comportemental humain. Le profil humain se divise en deux parties importantes : le profil basé sur les connaissances et le profil comportemental. Le profil basé sur les connaissances comporte des informations sur la personne à savoir les informations sociales, démographiques, etc. Le profil comportemental comporte des informations sur le comportement de la personne et ses habitudes de vie. Ces deux profils sont complémentaires mais notre revue de la littérature nous a révélé l'intérêt porté au profil comportemental vu l'importance du comportement humain et son influence sur tout système d'assistance [77, 42]. Cette phase de recherche nous a permis de mieux cerner nos besoins et de comprendre les limites des systèmes existants pour la construction du profil comportemental. Par la suite, le deuxième volet consistait, à la lumière de nos études et investigations dressées dans le premier volet, à proposer une nouvelle approche pour la construction du profil comportemental des personnes. À partir de là, nous avons choisi de formaliser notre approche en nous appuyant sur l'apprentissage non supervisé et la théorie de la causalité, de manière à exploiter le principe des relations causales pour introduire une nouvelle définition, représentation et construction du profil comportemental. Cette phase nous a permis de poser les fondements théoriques de notre approche et de créer un modèle formel pour la représentation du profil comportemental.

Finalement, le troisième volet consistait à valider l'approche développée dans le deuxième volet. Pour ce faire, nous avons adopté deux scénarios différents pour la validation. Une validation subjective dans laquelle nous avons évalué la capacité de notre approche à construire un modèle graphique représentant les relations causales entre les différents comportements, et une validation objective dans laquelle nous avons proposé deux algorithmes simples pour deux applications pratiques du profil : un algorithme pour l'identification des personnes et un autre pour la prédiction des activités. Les résultats de ces expérimentations nous ont permis d'évaluer les performances de l'approche développée, de comparer les résultats obtenus de notre approche avec d'autres approches existantes, et finalement de dresser les voies de développement futur pour l'amélioration et l'extension de notre approche.

0.3 Contributions originales de cette thèse

Afin de s'attaquer à la problématique de recherche définie dans cette thèse, qui s'inscrit dans le contexte général d'assistance des personnes dans un habitat intelligent, et de proposer des solutions permettant de surmonter les limites des approches existantes, nous proposons dans cette thèse deux approches novatrices pour la reconnaissance d'activités et la construction du profil comportemental des personnes respectivement. Nos contributions seront détaillées dans les sous-sections suivantes :

0.3.1 Contribution 1 : Nouvelle approche de reconnaissance d'activités non-supervisée

La raison principale ayant motivé notre choix de l'approche non supervisée concerne le problème d'annotation des données qui est difficile à résoudre en utilisant l'approche supervisée, car elle suppose une connaissance a priori des activités. De plus, notre approche combine un modèle non supervisé avec l'utilisation des patrons fréquents. La raison ayant motivé notre décision d'adopter l'analyse de patrons fréquents réside principalement dans l'objectif de réduire la quantité de bruit dans les données. En effet, le bruit apparaît de façon irrégulière dans les données. À cet effet, l'utilisation de patrons fréquents nous paraît comme une solution prometteuse pour filtrer le bruit dans les données. Par conséquent, les contributions originales de ce travail sont :

1. La proposition d'une approche de reconnaissance d'activités permettant de surmonter l'annotation des données.
2. La réduction significative du bruit dans les données par l'utilisation des patrons fréquents.
3. La découverte automatique des activités par la combinaison d'un modèle statistique non supervisé avec l'analyse de patrons fréquents.

Cette approche est une contribution originale à la fois dans le domaine des habitats intelligents et le domaine de forage de données.

0.3.2 Contribution 2 : Nouvelle approche pour la construction du profil comportemental

La reconnaissance d'activités est une étape centrale dans un système d'assistance. Cependant, la variabilité du comportement humain rend ce processus plus difficile, plus particulièrement lorsque les personnes présentent des déficits cognitifs, physiques ou autres. Notre approche sera basée sur l'analyse des séquences d'événements. Il s'agit d'une direction qui est peu exploitée par les chercheurs de la communauté. L'intérêt majeur de ce choix est la possibilité d'utiliser notre approche avec tout type de données séquentielles. En outre, l'analyse des séquences permettra de caractériser les comportements des personnes en utilisant le concept de patrons fréquents.

Une particularité de notre approche est l'exploration des relations entre les comportements qui pourrait mener au développement de différentes applications de personnalisation, d'adaptation et d'assistance. Notre ultime objectif consiste à créer une approche capable d'extraire les relations causales entre les différents comportements et activités de la vie quotidienne. Au meilleur de notre connaissance, l'extraction des relations causales entre les activités de la vie quotidienne et la découverte des activités causales constituent le premier travail dans la littérature qui traite le problème de la causalité entre les activités de la vie quotidienne. Il est espéré que ce travail aura des impacts aussi bien dans le domaine des habitats intelligents, que dans d'autres domaines tels que le Web et les réseaux sociaux par l'introduction de la théorie de l'analyse causale entre les différentes activités et comportements. Les contributions originales de ce travail sont résumées dans les points suivants.

1. La détection des différents comportements des personnes de façon non supervisée en utilisant les techniques de clustering appliquées sur les séquences.
2. L'introduction de la théorie de causalité dans la construction du profil comportemental et extraction des relations causales entre les différents comportements et activités des personnes.
3. La détection des activités causales qui jouent un rôle central dans l'assistance, la planification et la prédiction des activités des personnes.
4. La validation de notre approche selon différents aspects en utilisant des jeux de données provenant de plusieurs domaines et représentant différents types de

données.

La découverte des différents comportements et l'extraction des relations causales entre ces comportements constituent une contribution originale à la fois dans le domaine des habitats intelligents et dans le domaine de forage de données.

En somme, ce travail de recherche s'attaque à des problématiques épineuses rencontrées dans le domaine des habitats intelligents à savoir la reconnaissance d'activités et la construction du profil comportemental. Les solutions proposées ont permis d'attaquer ces problématiques et de surmonter les limites des approches existantes. Ce travail peut être considéré comme l'une des nombreuses composantes nécessaires au développement d'une plateforme qui s'inscrit dans le contexte de réalisation d'un projet de recherche de grand envergure au laboratoire Domus. Les solutions proposées pourront servir de base à des projets futurs visant des applications différentes dans le contexte des habitats intelligents. Ce travail de recherche résulte en un ensemble de méthodes, d'outils d'analyse, de tutoriels et de solutions à des fins éducatives et de recherche.

0.4 Structure et organisation de la thèse

Le contenu de cette thèse est organisé en trois principaux chapitres, correspondant chacun à l'une des phases de réalisation de notre projet de recherche.

Le premier chapitre décrit de façon détaillée l'état de l'art ciblé des principales approches de reconnaissance d'activités proposées dans la littérature. Ce chapitre présentera les quatres approches importantes de reconnaissance d'activités à savoir les approches statistiques, de classification et d'analyse des séquences. Il se focalisera particulièrement sur les approches d'analyse des séquences, qui sont beaucoup plus proche de nos travaux. Ce chapitre permettra de déceler les limites des approches existantes et de positionner nos travaux par rapport aux approches antérieures. Il permettra également de mettre en relief nos contributions au domaine de la reconnaissance d'activités.

Le deuxième chapitre présentera en profondeur notre approche de reconnaissance d'activités. En premier lieu, nous donnerons les détails du modèle statistique (LDA) [11] que nous avons utilisé et qui est le fondement de notre approche. Par la suite, nous

introduirons notre approche en mettant l'accent sur la combinaison entre les patrons fréquents et le modèle statistique LDA afin de reconnaître les activités. Tout au long de ce chapitre, des définitions ainsi que des notions théoriques seront illustrées par des exemples afin de comprendre leur application. Après avoir présenté les concepts théoriques ainsi que les détails de notre approche, nous présenterons les études expérimentales que nous avons menées pour valider notre approche. Cela comprend une présentation des données utilisées ainsi que la démarche adoptée dans les expérimentations. Cette étude expérimentale permettra aussi de dégager les forces et les faiblesses de notre approche. À la fin du chapitre, nous présenterons les résultats obtenus ainsi qu'une comparaison avec les approches existantes de reconnaissance d'activités. Ce chapitre correspond à notre contribution au domaine de la reconnaissance d'activités.

Le dernier chapitre permettra d'introduire notre nouvelle approche de construction du profil comportemental des personnes. Ce chapitre est constitué de deux parties. La première partie fera une présentation panoramique sur les différentes approches de construction du profil des personnes. Cette partie porte une attention particulière aux approches de construction du profil comportemental vu son importance dans les systèmes d'assistance et dans d'autres systèmes. Cela nous permettra de déceler les limites des approches existantes et de positionner nos travaux par rapport aux approches antérieures. Dans la deuxième partie, nous présenterons également quelques notions théoriques comme la notion de la causalité ainsi que la classification non supervisée des données catégoriques sur lesquelles se base notre nouvelle approche. La validation de notre approche sera également présentée à la fin de ce chapitre avec une comparaison avec d'autres approches existantes. À la lumière des résultats présentés, ce chapitre discutera des avantages et des limites de notre approche.

Finalement, nous conclurons sur un bilan de toutes nos contributions originales et leur apport à la résolution de certains problèmes épineux et à l'avancement du domaine des habitats intelligents. Nous mettrons en évidence les pistes de recherches qui peuvent se découler à partir de ces travaux. Dans cette conclusion, nous discuterons les différentes voies de développements futures pour l'amélioration et l'extension de notre approche de reconnaissance d'activités ainsi que l'approche de construction du profil.

Chapitre 1

État de l'art sur la reconnaissance d'activités

1.1 Introduction

Dans cette section, nous allons présenter les travaux concernant la reconnaissance d'activités dans les environnements intelligents. Dans ce qui suit, nous allons nous concentrer sur les approches non-invasives de reconnaissance d'activités. Les approches invasives e.g. les approches basées sur le traitement d'images ou vidéos ne seront pas abordées.

En premier abord, il est très important de commencer par donner une définition claire d'une activité. Cela permettra de réunir les différents travaux traitant le même objectif et de dégager ainsi une ligne directrice dans la présentation des travaux proposés.

Dans la littérature, les activités sont classées en trois types :

- Activités de la Vie Quotidienne AVQ (*Activities of Daily Living ADLs* en anglais) introduites par le docteur en médecine Sidney Katz en 1960 [64]. Ce sont des activités basiques qu'une personne doit être capable de réaliser pour prétendre à une certaine autonomie. Ces activités comprennent entre autres les activités d'hygiène, d'habillement, de déplacement, de prise de repas,..., etc.
- Activités Instrumentales de la Vie Quotidienne AIVQ (*Instrumental ADLs* en

anglais) introduites par Lawton et Brody en 1969 [74]. Ce sont des activités plus complexes qui nécessitent un certain effort physique, un bon jugement et un sens de l'organisation pour les accomplir. Ces activités comprennent entre autres les activités d'utilisation du téléphone, de prise de médicaments, de préparation de repas, d'entretien ménager,..., etc.

– Activités Étendues de la Vie Quotidienne AEVQ (*Enhanced ADLs* en anglais) introduites par Rogers et al. en 1998 [114]. Ce sont des activités permettant à une personne de s'adapter aux changements de l'environnement. Par exemple, l'utilisation des services d'internet comme un moyen de communication avec le monde extérieur, la famille, les amis, et même avec des communautés locales.

La plupart des approches proposées traitent le problème de reconnaissance d'activités de la vie quotidienne AVQ et des activités instrumentales AIVQ. Dans notre recherche, nous nous sommes intéressés aussi aux AVQ et AIVQ. Selon la façon de la réalisation des activités, nous distinguons plusieurs types d'activités :

1. Activités séquentielles : dans ces activités, un certain ordre est imposé entre les tâches. Si une activité séquentielle est composée de deux tâches T_1 et T_2, alors la tâche T_1 doit être exécutée avant la tâche T_2. Par exemple, pour regarder la télévision, il faut tout d'abord l'allumer.
2. Activités concurrentes : deux activités sont dites concurrentes si elles apparaissent simultanément ou parallèlement. Par exemple, lors de la préparation de repas, le téléphone sonne.
3. Activités entrelacées : deux activités sont dites entrelacées si l'utilisateur peut alterner entre les étapes de réalisation de ces deux activités. Par exemple, l'utilisateur commence l'activité 'travailler sur ordinateur', puis il interrompe cette activités pour faire l'activité 'préparer un café', par la suite il revient pour compléter son travail sur l'ordinateur.

Il n'existe pas de consensus sur la définition d'une activité dans les environnements intelligents. Dans notre travail, nous définissons une activité comme étant un ensemble de patrons. Chaque patron est une sous-séquence d'événements représentant une tâche de l'activité.

Il est important de mentionner que la reconnaissance d'activités se fait généralement en suivant un modèle. La plupart des modèles [27, 111, 112] adoptent la

représentation hiérarchique dans laquelle les activités de haut niveau sont reconnues à partir de celles de bas niveau. Mais cette représentation n'est pas toujours valable pour toutes les activités. Par exemple, les activités ouvrir le robinet, s'asseoir, se mettre debout ne peuvent pas être décomposées en structure hiérarchique et ne peuvent pas être reconnues suivant cette structure.

Dans les environnements intelligents, la reconnaissance d'activités est basée sur le traitement des événements issus des différents capteurs placés dans l'environnement. Ces événements sont caractérisés par : un identifiant relatif au capteur (par exemple : IR1000 est l'identifiant du capteur infrarouge 1000), la date d'apparition de l'événement et l'instant d'activation de capteur (par exemple : 15 :04 :2011 10 :10 :34 est la date et instant d'activation du capteur IR1000), et un état ou valeur du capteur (par exemple : ON est l'état du capteur IR1000). Selon les capteurs, les états seront binaires comme (ON, OFF, OPEN, CLOSE, etc) ou des valeurs numériques comme dans le cas de capteurs de température, d'humidité, de pression, etc. Tableau 1.1 présente un exemple d'événements enregistrés dans un habitat intelligent.

tableau 1.1 – Exemple d'événements enregistrés dans un habitat intelligent

Date	Temps	Nom de capteur	État / Valeur
2009-02-02	12 :18 :44	CapteurInfraRouge16	ON
2009-02-02	12 :18 :46	CapteurInfraRouge17	OFF
2009-02-02	12 :28 :50	CapteurPorte12	OPEN
2009-02-02	12 :29 :55	CapteurRFID03	PRESENT
2009-02-05	08 :05 :52	CapteurEauChaude-B	0.0448835
2009-02-05	12 :21 :51	CapteurPorte09	CLOSE
2009-02-10	17 :03 :57	CapteurTiroir03	ON

Les travaux publiés dans la littérature sur la reconnaissance d'activités dans les environnements intelligents sont classés en deux catégories selon la façon d'utilisation des événements. La première catégorie représente les travaux basés sur les événements. Une activité est définie comme un ensemble d'événements sans aucune contrainte sur l'ordonnancement de ces événements. La deuxième catégorie représente les travaux basés sur les séquences d'événements. Une activité est composée d'un ensemble d'événements ordonnés ou partiellement ordonnés. Dans la deuxième catégorie les événements sont ordonnés dans le temps et cet ordre est pris en compte dans le processus de reconnaissance d'activités. Ces deux catégories sont représentées graphiquement

dans la figure 1.1.

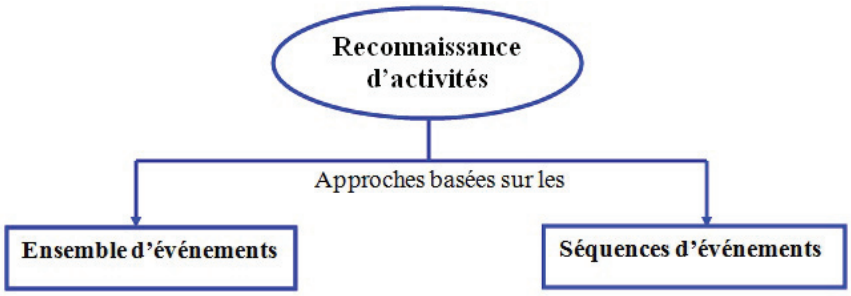


figure 1.1 – Principales catégories des approches de la reconnaissance d'activités

Plusieurs travaux sur la reconnaissance d'activités ont été réalisés dans chacune des catégories. À noter qu'il existe des travaux qui peuvent être classés dans les deux catégories comme les modèles Markoviens ou Bayésiens, mais cela ne constitue pas une contrainte pour le processus de reconnaissance.

Tous les travaux de reconnaissance d'activités partagent l'aspect de représentation de données et de classification, et diffèrent principalement au niveau des modèles utilisés. Ces modèles visent à modéliser la **relation** entre les événements et les activités. Les événements représentent les effets et les activités représentent les causes. Donc, la relation entre les événements et les activités est une relation de cause/effet. Par conséquent, chaque approche proposée traite un aspect particulier en utilisant un modèle particulier. Dans ce qui suit nous présentons les travaux publiés sur la reconnaissance d'activités en spécifiant la catégorie à laquelle appartient chaque approche.

1.2 Approches statistiques

Les approches statistiques ont connu une large utilisation dans le domaine de la reconnaissance d'activités. Elles se basent principalement sur le calcul des probabilités conditionnelles liant les événements aux activités. Cette probabilité s'écrit sous la forme **p(événement|activité)**. Ces approches peuvent également être divisées en plusieurs modèles statistiques.

1.2.1 Modèles Bayésiens

Les modèles Bayésiens permettent de modéliser des processus stochastiques, i.e. de phénomènes aléatoires évoluant avec le temps. Le modèle Bayésien fait partie des modèles basés sur les événements. Le modèle le plus connu des approches Bayésiennes est celui des réseaux Bayésiens.

Définition 1. Réseau Bayésien :

Un réseau Bayésien (appelé aussi réseau probabiliste ou réseau de croyance) [90, 23, 98] est un modèle probabiliste représentant des connaissances incertaines sur un phénomène complexe, et permettant d'élaborer un raisonnement à partir des données. Un réseau Bayésien, souvent noté BN (Bayesian Network), est un graphe acyclique orienté dont les nIJud sont des variables aléatoires qui peuvent prendre des valeurs discrètes ou continues. Les arcs reliant les noeuds sont rattachés à des probabilités conditionnelles. Notons que le graphe est acyclique : il ne contient pas de boucles. Les arcs représentent des relations entre les variables qui sont déterministes ou probabilistes.

Un réseau Bayésien s'appuie sur le théorème de Bayes. C'est un résultat de base en théorie des probabilités, issu des travaux du révérend Thomas Bayes (1702-1761) [97]. Étant donné deux événements A et B, le théorème de Bayes est donné par la formule suivante :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1}$$

Le terme P(A) est la probabilité a priori de A. Elle est "antérieure" au sens qu'elle précède toute information sur B. P(A) est aussi appelée la probabilité marginale de A. Le terme P(A|B) est appelé la probabilité a posteriori de A sachant B (ou encore de A sachant B). Elle est "postérieure", au sens qu'elle dépend directement de B. Le terme P(B|A), pour un B connu, est appelé la fonction de vraisemblance de A. De même, le terme P(B) est appelé la probabilité marginale ou a priori de B.

Le réseau Bayésien possède beaucoup d'avantages qui peuvent être résumés dans les points suivants :

– Les réseaux Bayésiens peuvent gérer le problème de connaissances incertaines grâce au modèle probabiliste.

- Les réseaux Bayésiens possèdent une sémantique très claire et représentent un modèle solide mathématiquement en se basant sur la théorie de probabilité.
- Les réseaux Bayésiens permettent une représentation graphique des connaissances, ce qui augmente la lisibilité et rend intuitive la compréhension du domaine.

Construction d'un réseau Bayésien

La construction d'un réseau Bayésien est liée à la construction du graphe acyclique. Deux éléments importants sont nécessaires pour construire un réseau Bayésien :

1. le graphe acyclique décrivant le modèle. Le graphe acyclique est généralement appelé la structure du réseau Bayésien. Cette structure est souvent construite par des experts du domaine.
2. les tables de probabilités de chaque variable conditionnellement à ses parents. Les tables de probabilités sont appelées les paramètres du réseau Bayésien. Ces tables de probabilités sont généralement définies à partir des données.

Un réseau Bayésien est basé sur l'hypothèse d'indépendance conditionnelle des variables. Cette contrainte permet de réduire considérablement l'espace de recherche qui est de l'ordre exponentiel. Notons que l'espace de recherche des réseaux Bayésiens dépend en général du nombre de variables et du nombre d'arcs.

Les tables de probabilités sont définies par des statistiques relatives au problème étudié. Ces tables peuvent parfois être déterminées par des experts. Chaque variable dispose d'une table de probabilités conditionnelles relatives aux variables parents dont elle dépend.

En d'autres termes, un réseau Bayésien est un modèle graphique permettant de représenter les indépendances conditionnelles entre un ensemble de variables. Une variable A est conditionnellement indépendante de la variable B étant donné C si $P(A, B|C) = P(A|C)P(B|C)$ pour tout A, B et C avec P(C) $\neq 0$. Soient un ensemble de variables aléatoires $X = (X_1, X_2, X_3, ..., X_N)$ et $P(X)$ sa distribution jointe de probabilité. Le graphe acyclique indique les indépendances conditionnelles entre les noeuds (variables aléatoires). La paramétrisation est donnée en terme de probabilités conditionnelles des noeuds sachant leurs parents. Par conséquent, la probabilité jointe

peut être écrite sous la forme suivante :

$$P(X_1, X_2, X_3, ..., X_N) = \prod_{i=1}^{N} P(X_i|parents(X_i)) \tag{1.2}$$

Exemple 1. Réseau Bayésien : Exemple tiré de [93]

Prenons un exemple de réseau Bayésien[1] à cinq variables : C (cambrioleur), S (séisme), A (alarme), R (annonce radio), V (appel voisin). Les variables "cambrioleur" et "séisme" sont indépendantes, et "cambrioleur" et "annonce radio" sont indépendantes étant donné le "séisme". Cela peut être traduit par le fait qu'aucun événement n'affecte les variables "cambrioleur" et "séisme". De même, les propositions "cambrioleur" et "annonce radio" sont indépendantes étant donné le "séisme" signifie que si "annonce radio" est le résultat d'un séisme, il ne peut pas être le résultat d'un cambriolage. Le voisin appelle le propriétaire pour lui annoncer que l'alarme est déclenchée. Le réseau Bayésien associé à cet exemple est présenté dans la figure 1.2.

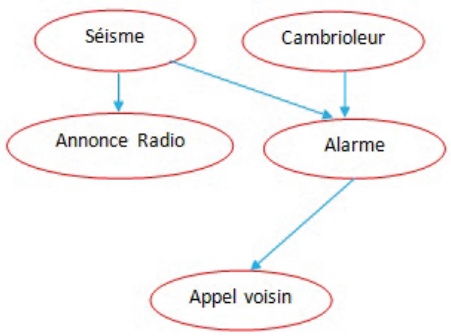


figure 1.2 – Exemple d'un réseau Bayésien

Vu l'indépendance entre ces variables, la probabilité jointe est présentée par la formule suivante :

$P(C, S, A, R, V) = P(S).P(C).P(A|S, C).P(R|S).P(V|A).$

1. Cet exemple est extrait de [93] avec traduction

Apprendre les paramètres d'un réseau Bayésien

Apprendre les paramètres d'un réseau Bayésien revient à chercher la structure du réseau Bayésien qui décrit le mieux les données observées. Plusieurs heuristiques ont été proposées pour apprendre la structure d'un réseau Bayésien. Il existe deux catégories de méthodes pour l'apprentissage des structures d'un réseau Bayésien [21, 20, 49, 44, 57, 69], selon la nature des données (complètes ou non-complètes) [71].

Données complètes :

Il existe deux méthodes principales pour apprendre la structure du réseau Bayésien avec des données complètes. Apprentissage statistique et apprentissage Bayésien.

- Apprentissage statistique : cette méthode est la méthode la plus simple et la plus utilisée lorsque toutes les variables sont observées. L'estimation statistique consiste donc à estimer la probabilité d'un événement par sa fréquence d'apparition dans la base de données. Cette méthode est appelée aussi le maximum de vraisemblance dont l'estimation peut s'écrire sous la forme suivante (θ représente la paramétrisation du réseau Bayésien) :

$\hat{P}(X_i = x_k | parents(X_i) = x_j) = \hat{\theta}_{i,j,k} = \frac{N_{i,j,k}}{\sum_k N_{i,j,k}}$.

où $N_{i,j,k}$ est le nombre d'événements dans la base de données pour lesquels la variable X_i est dans l'état x_k et ses parents sont dans la configuration x_j [71].

- Apprentissage Bayésien : cette méthode utilise des a priori sur les paramètres θ pour trouver les paramètres les plus probables sachant que les données ont été observées. La règle de Bayes est utilisée dans cette méthode et s'écrit sous la forme suivante : $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta)$ où $\mathcal{D}$ représente les données. Cette méthode s'appelle le maximum a posteriori. La formule du maximum a posteriori dépend de la distribution a priori employée sur les paramètres θ.

Données incomplètes :

Les méthodes que nous avons vues pour les données complètes ne sont valables que si toutes les variables sont observées. Cependant, dans la pratique, les bases de données sont très souvent caractérisées par des données manquantes ou incomplètes. Cela est probablement dû au fait que certaines variables ne sont observées que partiellement ou bien jamais dépendamment du contexte ou de certains facteurs qui peuvent influencer les observations. Il existe des méthodes pour le cas des données incomplètes, et la méthode la plus connue est l'algorithme EM (Expectation Maximization).

Algorithme EM : L'algorithme EM, proposé par Dempster et al. [29] en 1977, est un algorithme itératif permettant de trouver le maximum de vraisemblance des paramètres de modèle probabiliste. Cet algorithme a été utilisé pour la première fois par [23, 92] dans le cadre des réseaux Bayésiens. Nous allons donc introduire brièvement l'algorithme EM. Pour cela nous adopterons le formalisme de [71].

Soit :

- $X_v = \{X_v^{(I)}\}, I = 1, .., N$, l'ensemble des données observées (visibles).
- $\theta^{(t)} = \{\theta_{i,j,k}^{(t)}\}$, les paramètres du réseau Bayésien à l'iteration t.

L'algorithme EM s'applique pour trouver les paramètres du réseau Bayésien en répétant les deux étapes suivantes jusqu'à convergence.

- **Espérance** (Expectation) : cette étape permet d'estimer les $N_{i,j,k}$ manquants, en calculant leur moyenne conditionnellement aux données et aux paramètres actuels du réseau Bayésien.
  $N_{i,j,k}^* = E[N_{i,j,k}] = \sum_{l=1}^{N} P(X_i = x_k | parents(X_i), X_v^{(l)}, \theta^{(t)})$.
  Cette étape revient à calculer une série d'inférences en utilisant les paramètres courants du réseau Bayésien, et de remplacer les valeurs manquantes par des valeurs de probabilités obtenues par inférence.
- **Maximisation** (Maximization) : comme son nom l'indique, cette étape permet, en remplaçant les $N_{i,j,k}$ par leur moyenne, de calculer dans l'étape d'espérance, les paramètres $\theta^{(t+1)}$ par maximum de vraisemblance comme suit : $\theta_{i,j,k}^{(t+1)} = \frac{N_{i,j,k}^*}{\sum_N N_{i,j,k}^*}$

Les Réseaux Bayésiens Dynamiques (DBN) [17] sont une extension des réseaux Bayésiens qui représentent l'évolution temporelle des variables aléatoires. En considérant un ensemble $X_t = \{X_t^1, X_t^2, ..., X_t^N\}$ de variables évoluant dans $[0, T]$.

Il existe beaucoup de travaux dans la littérature qui utilisent les réseaux Bayésiens ou les réseaux Bayésiens dynamiques pour la reconnaissance d'activités. Par exemple, le modèle Bayésien naïf basé sur le théorème de Bayes est utilisé par Kasteren et al. [127] pour la reconnaissance d'activités basiques dans les résidences pour personnes âgées. Un modèle similaire basé sur les filtres de particules [106], une variante du modèle Bayésien naïf, a été utilisé pour apprendre l'activités d'utilisation de transport. Dans ce dernier modèle, les données sont des signaux de GPS. Les filtres de particules permettent de réduire le bruit dans les données. Cependant ces modèles ne prennent

pas en compte l'aspect temporel des activités et souffrent du problème d'annotation des données qui est une tâche laborieuse et coûteuse en terme de temps.

Le réseau Bayésien dynamique a été utilisé par Philipose et al [101] pour la reconnaissance d'activités dans un habitat intelligent. Les données dans ce cas proviennent de capteurs RFID (Radio Frequency IDentification). Un modèle similaire a été utilisé par Gu et al. [35] pour résoudre le problème d'incertitude de contextes dans la reconnaissance d'activités. Ces modèles prennent en compte l'aspect temporel des activités. Cependant, ils souffrent du problème d'annotation des données.

1.2.2 Modèles Markoviens

Les modèles Markoviens sont aussi fréquemment utilisés dans le domaine de reconnaissance d'activités. Les modèles Markoviens sont des automates probabilistes à états finis. Ils se basent sur l'hypothèse de Markov qui stipule que le futur ne dépend que de l'état présent. Plus précisément ce sont des modèles Markoviens de premier ordre. Cette hypothèse implique que l'état du modèle doit contenir suffisamment d'informations pour permettre une bonne prédiction du comportement futur du système, ce qui n'est pas nécessairement toujours le cas dans des situations réelles.

Définition 2. Chaîne de Markov :

Un processus discret $X_0, X_1, ..., X_n, ...$ défini sur l'espace de probabilités et à valeurs dans E (espace mesurable) est une chaîne de Markov s'il possède la propriété suivante (dite de Markov) :
$\forall n, P(X_n = x_n | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, ..., X_0 = x_0) = P(X_n = x_n | X_{n-1} = x_{n-1}), \forall x_0, x_1, ..., x_n \in E$.

Les chaînes de Markov telles que définies (chaque état correspond à un événement observable) sont assez limitées pour des problèmes plus complexes comme la reconnaissance d'activités ou la reconnaissance de la parole par exemple. Une chaîne de Markov est incapable de capturer des dépendances à grande distance. Pour surmonter ces limites, une extension de la chaîne de Markov dans laquelle l'observation est modélisée comme une fonction probabiliste de l'état du modèle a été proposée. Dans ce cas, la suite d'observation est dissociée de la suite d'état et devient non observable ou cachée.

Une chaîne de Markov cachée (ou HMM : Hidden Markov Model) est une chaîne de Markov dont les états ne sont pas déterminés mais génèrent une suite de variables aléatoires (observations) indépendantes deux à deux. Les chaînes de Markov cachés (HMM) sont une extension des modèles de Markov. Ces modèles se basent sur deux processus stochastiques dépendants l'un de l'autre. En effet, l'état du système n'est plus directement observable ; il est caché par un processus d'observation.

Dans le domaine des environnements intelligents, les seules variables qui sont observables sont les états de capteurs. L'objectif principal étant de reconnaître les activités qui sont cachées, les HMM apparaissent comme une solution prometteuse et un modèle pratique pour résoudre la problématique de la reconnaissance d'activités. Dans ce qui suit nous allons présenter le principe général d'un modèle de Markov caché, puis nous présenterons les travaux publiés dans la littérature qui utilisent les modèles de Markov cachés pour la reconnaissance d'activités.

Le modèle de Markov caché tire son nom à partir des deux propriétés suivantes :

- Premièrement, le modèle HMM suppose que l'observation au temps t a été générée par un processus dont l'état S_t est cachée de l'observateur.
- Deuxièmement, le modèle HMM suppose que ce processus caché vérifie la propriété de Markov qui stipule que, l'état courant du système S_t ne dépend que de l'état S_{t-1}. En d'autres termes, l'état à un moment donné encapsule toutes les informations nécessaires sur l'historique du processus pour prédire l'état futur du processus.

Notons que les sorties du modèle HMM satisfont la propriété de Markov : étant donné un état S_t, l'observation Y_t est indépendante de tous les états et de toutes les observations à toutes les autres tranches de temps [33]. La probabilité jointe d'une séquence d'états et observations peut être écrite dans un modèle HMM de la façon suivante :

$$P(S_{1:T}, Y_{1:T}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_t|S_t) \tag{1.3}$$

Les notations $S_{1:T}$ et $Y_{1:T}$ sont utilisées pour signifier $S_1, S_2, ..., S_T$ et $Y_1, Y_2, ..., Y_T$ respectivement.

Éléments d'un HMM

Un modèle HMM est composé des éléments suivants (cette présentation est une adaptation de [113]) :

1. N est le nombre des états du modèle : $S = \{S_1, S_2, ..., S_N\}$
2. M est le nombre de symboles d'observations : $Y = \{Y_1, Y_2, ..., Y_M\}$
3. la distribution des probabilités de transitions entre états : $A = [a_{ij}]$ avec $a_{ij} = P(s_{t+1} = S_j | s_t = S_i)$
4. la distribution des probabilités d'observations (probabilités d'émission) : $B = [b_j(m)]$ où $b_j(m) = P(y_t = Y_m | s_t = S_j)$
5. la probabilité d'état initial $\pi = \pi_i$ où $\pi_i = P(s_1 = S_i)$

Soit λ un modèle du HMM. Trois principaux problèmes sont liés au modèle λ :

- **Évaluation** : étant donné le modèle λ, quelle est la probabilité $P(Y|\lambda)$ d'avoir une séquence d'observations $Y = \{y_1, y_2, ..., y_T\}$?
- **Décodage** : étant donné le modèle λ et une séquence d'observations Y, quelle est la séquence d'états $S = \{s_1, s_2, ..., s_T\}$ qui a vraisemblablement généré Y ? $S^* = argmax_S P(S|Y, \lambda)$
- **Apprentissage** : étant donné un jeu de données d'entraînement $\mathcal{D} = \{Y^k\}$ contenant des séquences d'observations, quel est le modèle λ du HMM qui aurait vraisemblablement généré $\mathcal{D}$? $\lambda^* = argmax_\lambda P(\mathcal{D}|\lambda)$

Les techniques proposées dans la littérature pour résoudre ces trois problèmes sont discutées dans l'annexe A. Nous allons discuter dans le reste de cette section les travaux qui utilisent les modèles de Markov cachés.

Panduranga et al. [110] ont utilisé un modèle HMM de base pour la reconnaissance d'activités dans un habitat intelligent. Les données dans ce cas sont issues de capteurs RFID. Les capteurs RFID sont installés sur tout objet utilisé dans la vie quotidienne afin de récupérer l'information sur l'utilisation de cet objet par l'usager. Le même modèle est utilisé pour prédire les événements dans un habitat intelligent. La prédiction de l'événement à apparaître dans un instant t se fait en prenant l'historique des observations $y_1, y_2, ..., y_{t-1}$.

Pour gérer les granularités spatiales et temporelles dans la reconnaissance d'activités, les modèles HMM à couches (les HMM à couches sont des HMM composés)

sont utilisés respectivement par Sanchez et al. [122] et Oliver et al. [95]. Dans les modèles HMM à couches, les résultats d'une couche (un HMM) sont utilisés comme entrée pour la couche suivante. Le résultat final est celui généré par le modèle HMM de la dernière couche. Les modèles HMM à couches ont donné des résultats meilleurs que ceux obtenus avec le modèle HMM. De plus, Les modèles HMM à couches sont beaucoup plus robustes aux changements survenant dans l'environnement.

Kim et al. [65] améliorent le modèle HMM pour prendre en considération la non-indépendance entre les événements en utilisant le modèle CRF (Conditional Random Fields). Le modèle CRF est une variante du modèle HMM avec la particularité de prise en compte des dépendances entre les variables. D'autres variantes des modèles de Markov ont été proposées afin d'améliorer la reconnaissance d'activités. À titre d'exemple, Nguyen et al. [94] introduisent le modèle de Markov hiérarchique, et Osentoski et al. [123] introduisent le modèle de Markov caché abstract. Ces modèles sont composés de plusieurs modèles HMM et permettent de faire face au problème de bruit des données, et d'améliorer les résultats de reconnaissance d'activités. Cependant, tous ces modèles souffrent du problème d'annotation des données.

1.2.3 Résumé sur les approches probabilistes de reconnaissance d'activités

Dans cette section du chapitre, nous avons présenté les fondements des principales approches probabilistes de reconnaissance d'activités à savoir les modèles Bayésiens et les modèles Markoviens. Nous avons également introduit des exemples simples afin d'illustrer les processus inférentiels dans chacun des modèles. Plus précisément, cette partie a permis d'illustrer comment un problème de reconnaissance d'activités peut être modélisé à l'aide d'un modèle probabiliste basé sur un réseau Bayésien ou un modèle Markovien, et comment réaliser des inférences à partir de ces modèles, en spécifiant la séquence d'états cachés susceptible de générer une séquence d'observations dans le cas d'un modèle Markovien par exemple. Après avoir présenté les fondements de chaque modèle probabiliste, nous avons présenté quelques travaux dans la littérature qui utilisent ces modèles probabilistes.

À travers notre présentation des approches probabilistes, nous avons pu constater

que la force de ces approches réside dans leur capacité d'effectuer des raisonnements dans des contextes incertains provenant de la problématique de reconnaissance d'activités dans les environnements intelligents [50, 10]. Donc la notion d'incertitude est prise en compte dans le modèle.

Cependant, les modèles probabilistes présentent certaines limites qui sont principalement liées aux hypothèses sur lesquelles se basent ces approches. La première limite concerne l'hypothèse d'indépendance conditionnelle des variables. Dans la pratique, les variables ne sont généralement pas indépendantes, et le fait d'ignorer cette vérité peut entraîner la perte de certaines informations pertinentes qui rendraient ces modèles plus robustes et performants. Bien que la prise en compte de cette propriété augmente la complexité du système, elle permet de développer des systèmes plus précis et appropriés pour la problématique de la reconnaissance d'activités. C'est l'exemple des CRF(Conditional Random Fields) mentionnés auparavant qui donnent, dans la plupart du temps, des résultats meilleurs que ceux obtenus en utilisant les modèles Markoviens ou Bayésiens [128, 130, 129]. La deuxième limite est liée principalement au problème d'annotation des données. Ces approches nécessitent des connaissances supplémentaires sur les données, souvent représentées sous forme d'association entre les observations et les activités à reconnaître. Chaque observation est liée à une activité bien déterminée. Notons que les données de base telles que collectées dans les environnements intelligents ne sont pas dotées de cette annotation. L'annotation doit alors être effectuée soit par l'expérimentateur qui veille sur le déroulement des expérimentations (le cas le plus adopté), soit par l'utilisateur lui même qui doit annoter les activités une par une au fur et à mesure de l'avancement des expérimentations (un cas très rare, car le déroulement de l'expérimentation introduit un biais lors du processus d'annotation).

Additionnellement aux points sus cités, les modèles Markoviens comportent une limite importante relative au postulat du départ qui stipule que l'environnement intelligent peut être modélisé par un ensemble d'états finis. Cette hypothèse est réductrice dans le sens qu'elle exige une connaissance quasi exhaustive sur les différents états de l'environnement. Cela n'est en pratique pas le cas étant donnés les changements, l'évolution et le dynamisme de ces environnements.

1.3 Approches de classification

Les approches de classification supervisée et non supervisée permettent de reconnaître les activités en utilisant des algorithmes de classification ou de clustering. À noter que ces approches peuvent être jumelées avec des modèles statistiques pour améliorer la reconnaissance d'activités.

L'objectif principal de la classification est d'identifier les classes auxquelles appartiennent des objets à partir des traits descriptifs, appelés aussi attributs, caractéristiques, ou en anglais, "features". Les attributs dans notre cas représentent les événements et les classes représentent les activités. Les approches de classification différent de celles basées sur les statistiques dans la modélisation de la relation entre les événements et les activités. En effet, les approches de classification modélisent cette relation comme une fonction de transformation entre les événements et les activités i.e. (événement $\rightarrow$ activité). La classification est la tâche qui permet d'apprendre une fonction objectif f pour assigner un ensemble d'attributs x à l'une des classes prédéfinies étiquettée y. La fonction f peut aussi être appelée un modèle de classification. Ce modèle de classification peut être utilisé dans différentes situations.

Dans la classification supervisée, les classes sont connues et l'on dispose d'exemples de chaque classe, ce qui n'est pas le cas pour la classification non supervisée où les classes ne sont pas connues. Des mesures de similarité/dissimilarité sont utilisées afin de grouper les objets similaires dans le même groupe. Dans ce qui suit, nous allons introduire la définition du concept de classification ainsi que la description formelle du processus de classification. Ces détails sont inspirés de [125].

Le modèle de classification le plus connu dans ce contexte est le modèle Bayésien naïf [101]. Dans ce qui suit, nous allons présenter les détails de ce modèle et comment la classification peut être effectuée en utilisant ce modèle.

1.3.1 Modèle Bayésien naïf

De manière abstraite, le modèle Bayésien naïf est un modèle probabiliste conditionnel. Il se base principalement sur la règle de Bayes présentée dans la formule 1.1. Le caractère naïf du modèle Bayésien provient du fait que les descripteurs (features, caractéristiques, traits, etc.) sont indépendants conditionnellement à la classe. De manière

plus formelle, soit le système de classification suivant : $x \in X^p \longmapsto y \in \{1, 2, ..., M\}$. Dans ce système, $x = (x_1, x_2, ..., x_p)$ est un vecteur de descripteurs. Notons que $\forall i \in \{1, 2, ..., p\}, x_i \in X$. Le but du système de classification est d'apprendre la probabilité $P(y|x)$.

Par l'application de la règle de Bayes, nous aurons : $P(y|x) = \frac{P(x|y)P(y)}{P(x)}$. Le modèle Bayésien naïf consiste à poser l'hypothèse assez forte que les descripteurs x_i sont indépendants conditionnellement à la classe. Le numérateur devient : $P(x|y) = \prod_{i=1}^{p} P(x_i|y)$.

Exemple 2.

Dans cet exemple, nous illustrons comment le modèle Bayésien naïf peut être utilisé dans un problème de classification. Cet exemple est tiré de [125]. Soit la base de données présentée dans la figure 1.3.

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

figure 1.3 – Exemple d'une base de données

Soit le nouvel enregistrement présenté dans la figure 1.4, et pour lequel nous cherchons la classe à laquelle appartient cet enregistrement.

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

figure 1.4 – Exemple d'une base de données

Soit :

- A : l'ensemble des attributs.
- M : la classe des mammifères (Mammals dans l'exemple).
- N : la classe des non mammifères (Non-Mammals dans l'exemple).
- $P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$.
- $P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$.
- $P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$.
- $P(A|N)P(N) = 0.0042 \times \frac{13}{20} = 0.0027$.

Puisque $P(A|M)P(M) > P(A|N)P(N)$, donc le nouvel enregistrement sera classifié comme 'mammifère' (Mammals).

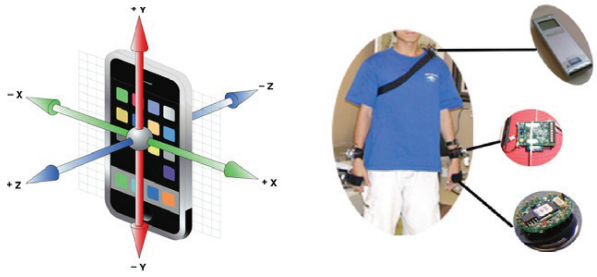
1.3.2 Revue des travaux existants

Dans cette section nous allons présenter les principaux travaux publiés dans la littérature, pour lesquels le problème de reconnaissance d'activités est traité comme un problème de classification.

Bao L. et al. [10] utilisent plusieurs classificateurs tels que les arbres de décision, le modèle Bayésien naïf, les tables de décision, et le modèle d'apprentissage basé sur les instances (IBL : instance-based learning) pour reconnaître des activités. Dans ce cas, les données sont obtenues à partir des accéléromètres placés dans différents endroits du corps humain à savoir les bras, la hanche, les cuisses , les mains, etc. La figure 1.5 montre un exemple d'un accéléromètre et les endroits sur lesquels il peut être placé [2]. Dans ce travail, des descripteurs tels que la moyenne, l'entropie, et la corrélation des données ont été calculés et utilisés pour la classification. Ces descripteurs sont obtenus en utilisant une fenêtre temporelle avec 50% de chevauchement. Les arbres de décision ont obtenu des meilleurs performances que le modèle Bayésien. Par contre ce modèle souffre du problème d'annotation des données qui est faite par les usagers eux mêmes, ce qui augmente la probabilité d'erreur d'annotation.

D'autres travaux utilisent d'autres classificateurs pour la reconnaissance d'activités comme les réseaux de neurones [144], KNN (K nearest neighbor), K-means, HMM et SVM [41], et le classificateur Bayésien [124]. Dans le travail de [41], les accéléro-

2. Les images sont tirées respectivement de [10, 132].



(a) Accéléromètre (b) Un usager porte des accéléromètres

figure 1.5 – Exemple d'un accéléromètre et les endroits sur lesquels il peut être placé

mètres sont utilisés pour collecter les activités réalisées par les usagers. La moyenne et la variance d'un signal ont été choisies comme des attributs (features), et sont calculées pour chaque signal en utilisant une fenêtre temporelle. L'algorithme de K-means est utilisé afin de créer des clusters d'activités. Chaque cluster est annoté par l'activité qui apparaît le plus souvent dans le cluster à partir de l'ensemble d'entraînement. L'algorithme SVM est utilisé pour effectuer la classification en attribuant à un échantillon de test, l'étiquette du centre de cluster le plus proche. L'algorithme SVM est un algorithme dont le but est de résoudre les problèmes de discrimination à deux classes. Le problème de discrimination à deux classes est un problème dans lequel on tente de déterminer la classe à laquelle appartient un individu parmi deux choix possibles. Pour ce faire, ils utilisent les n caractéristiques de cet individu représentées par un vecteur $\mathbf{x} \in \mathbb{R}^n$. La classe à laquelle appartient l'individu est représentée par $y \in -1, 1$, où une classe est représentée par -1, et l'autre par 1. L'objectif de l'algorithme SVM est de trouver un hyperplan séparateur qui permet de séparer les deux classes en maximisant la marges entre les deux classes.

Supposons que les données sont linéairement séparables. Cela suppose qu'il existe un hyperplan $\mathbf{w}.\mathbf{x} + b = 0$, tel que $\mathbf{w}.\mathbf{x} + b > 0$ pout tout $\mathbf{x}$ appartenant à la classe 1, et $\mathbf{w}.\mathbf{x} + b < 0$ pout tout $\mathbf{x}$ appartenant à la classe -1. Avec $\mathbf{w} = (w_1, w_2, ..., w_n) \in \mathbb{R}^n$ le vecteur de coefficients de l'hyperplan et $b \in \mathbb{R}$ est un scalaire appelé le biais.

Une revue exhaustive des méthodes de classification dans le contexte de la reconnaissance d'activités est présentée dans le travail de [8]. Toutes ces méthodes partagent le même principe de classification, mais utilisent des vecteurs de descripteurs diffé-

rents. Cependant, tous ces modèles souffrent du problème d'annotation des activités. De plus, les approches basées sur les accéléromètres présentent certaines limites et risquent de ne pas être utilisables dans la pratique. Par exemple, ces approches ne sont fonctionnelles que si les accéléromètres sont portés par les usagers, et ne peuvent pas fonctionner dans le cas où les usagers oublient de les porter. Les accéléromètres sont des équipements qui requièrent d'être portés par les usagers ce qui crée une certaine invasivité chez les usagers.

La classification non supervisée (clustering) est une approche proposée dans l'objectif de surmonter le problème des classes prédéfinies dans la classification supervisée. Cette technique est utilisée pour faire face au problème d'annotation des données. Rashidi et al [112] proposent une méthode pour reconnaître les activités et leurs occurrences dans les séquences en utilisant le clustering. Leur méthode est composée de plusieurs étapes :

- **Étape 1** : dans cette étape, les patrons significatifs sont extraits à l'aide du principe de longueur descriptive minimale (MDL) [55]. Le MDL permet d'extraire les patrons qui compressent mieux les données.
- **Étape 2** : dans cette étape, les patrons significatifs sont regroupés dans des groupes différents en utilisant la mesure d'édition comme une mesure de similarité.
- **Étape 3** : cette étape constitue le coeur de cette approche. En effet, les activités sont reconnues en utilisant un modèle Markovien caché. La figure 1.6 présente comment les modèles Markoviens sont incorporés dans le processus de reconnaissance d'activités [3]. Les cercles dans la figure 1.6 représente les états cachés, i.e. les activités, et les rectangles représentent les états de capteurs (états observés).
- La reconnaissance d'activités est faite en utilisant un vote sur les différents HMMs utilisés. Chaque HMM donne des résultats de classification pour une activité. La combinaison des résultats de tous les HMMs donne le résultat final de reconnaissance de l'activité en question (Activity Label) dans la figure 1.6.

Bien que ce modèle donne des bons résultats, il souffre du problème de complexité par l'incorporation de plusieurs modèles HMMs. De plus, ce modèle requiert des don-

3. Cette figure est tirée de [112]

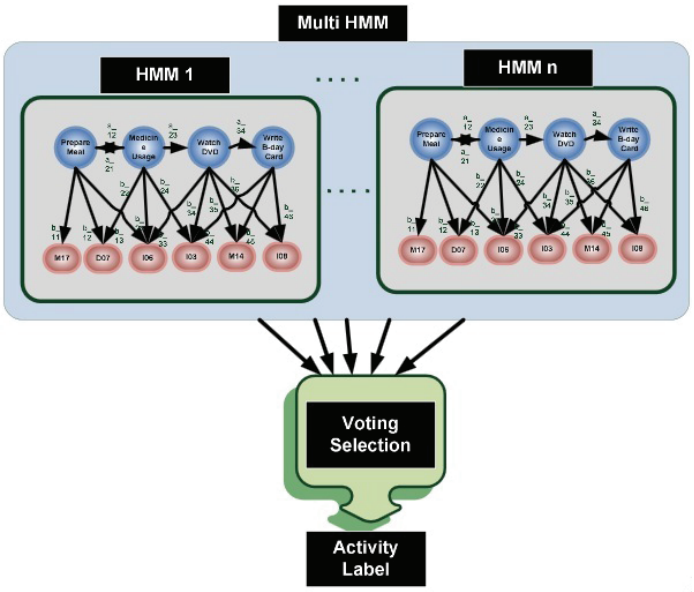


figure 1.6 – Modèle HMM pour la reconnaissance d'activités dans [112]

nées annotées pour pouvoir attribuer les bonnes étiquettes aux activités découvertes par les modèles HMMs.

1.3.3 Résumé sur les approches de classification

Dans cette section du chapitre, nous avons présenté les approches de reconnaissance d'activités faisant appel aux techniques de classification. Nous avons présenté sommairement les différentes techniques de classification adoptées dans le contexte de la reconnaissance d'activités. Nous avons toutefois porté une attention particulière aux modèles probabilistes tel que le modèle Bayésien naïf afin de mettre en relief les approches probabilistes et leurs avantages à résoudre le problème de la reconnaissance d'activités.

Dans la première partie de cette section du chapitre, nous avons présenté le modèle Bayésien naïf et comment celui-ci peut être utilisé dans un problème de reconnaissance d'activités. Nous avons aussi illustré ce principe à travers un exemple concret. Ensuite, nous avons présenté de façon panoramique les différents travaux proposés dans la littérature, et qui se basent sur les méthodes de classification. Nous avons

pu constater que l'avantage majeur des méthodes de classification réside dans leur capacité d'apprendre les modèles d'activités et de les construire à partir des données, contrairement aux approches qui requièrent des modèles d'activités prédéfinis.

Nous avons également discuté les méthodes de classification non supervisée. Nous avons présenté les détails d'un modèle de reconnaissance d'activités combinant le clustering et les modèles HMMs. L'avantage de ces méthodes réside dans leur capacité de surmonter le problème 'annotation des données rencontré par les autres approches de classification. Cependant, les méthodes de clustering souffrent du problème de choix de la mesure de similarité, ce qui influe sur les résultats de clustering.

Par ailleurs, les techniques d'apprentissage présentent certaines faiblesses. La principale faiblesse provient essentiellement des données d'apprentissage. En effet, ces techniques nécessitent beaucoup de données dans la phase d'apprentissage afin de construire les modèles d'activités et ajuster les paramètres de ces modèles. Mais, les données d'apprentissage proviennent en général des capteurs environnementaux. Ces données peuvent être erronées, incomplètes et bruitées. Cela va nécessairement influencer les modèles construits, ce qui faussera sans doute les résultats de la reconnaissance par la suite. De plus, les méthodes de classification présument que les données ne changent pas considérablement au cours du temps. Ceci est efficace dans un contexte qui n'est pas assujetti aux changements. Cependant, dans les environnements intelligents qui sont caractérisés par leur dynamisme et changements, avec des comportements variés des usagers, l'utilisation des méthodes de classification sera plus délicate et peu efficace. De plus, les méthodes de classification nécessitent des connaissances préalables sur l'appartenance des objets aux classes. C'est à dire, les activités sont connues à l'avance, et chaque événement est associé à une classe particulière (problème d'annotation des données). Cela aussi constitue une limite de ces méthodes et présente un fardeau aux utilisateurs qui doivent annoter les données avant de pouvoir utiliser ces méthodes.

1.4 Approches de l'analyse des séquences

Les approches de reconnaissance d'activités basées sur l'analyse des séquences présentent une nouvelle orientation dans le domaine de l'informatique ubiquitaire.

La reconnaissance d'activités basée sur l'analyse des séquences en est à son début et elle n'a été utilisée que récemment. Contrairement aux approches basées sur les événements, l'approche basée sur la séquence prend en considération l'aspect séquentiel des événements. Les événements sont ordonnés dans la séquence selon leur temps d'apparition. La prise en compte des informations temporelles permet d'un côté, la reconnaissance d'activités selon différents niveaux de granularité temporelle, et d'un autre côté de reconnaître des activités plus complexes comme les activités concurrentes et entrelacées.

Malgré les difficultés inhérentes à cette approche, elle possède un potentiel important en terme d'analyse, de détection de patrons, de détection des relations et des associations, et de changement et de progression dans le temps. Un autre avantage de cette approche est qu'elle est générique à toutes les données séquentielles à savoir les données transactionnelles, les séquences biologiques, les données de navigation sur internet (clickstream data), etc.

Une séquence est une suite ordonnée d'événements. Un patron est une sous séquence d'événements qui se répète dans les données. La répétition des patrons dans les données signifie l'importance de ces patrons. Ces patrons représentent des tâches réalisées par l'utilisateur. Plusieurs techniques sont utilisées afin d'extraire les informations à partir des séquences à savoir les algorithmes d'extraction de patrons fréquents, e.g. l'algorithme Apriori [6] et PrefixSpan [104], les algorithmes d'extraction des règles d'association e.g. l'algorithme Apriori [5] et d'autres algorithmes. Des modèles probabilistes peuvent également être utilisés pour enrichir le modèle de reconnaissance et résoudre certaines problématiques telles que le problème de bruit ou les données incomplètes. Le plus important dans l'analyse des séquences c'est comment déterminer si ces patrons sont significatifs et intéressants pour un problème donné. Plusieurs techniques ont été introduites pour déterminer l'importance de patrons selon un domaine particulier, mais il n'existe pas une méthode pertinente pour trouver les patrons significatifs dans une séquence. Cela nous motive à explorer cette approche afin de chercher des patrons significatifs pour chaque activité réalisée, ce qui permettra ainsi de distinguer les activités à la base de patrons significatifs.

Rachidi et al. [111] proposent un modèle pour l'analyse des séquences d'événement afin de reconnaître les patrons d'activités séquentielles dans le temps. Cette approche

permet d'identifier les patrons ainsi que leur changement dans le temps. Détecter le changement de patrons permet entre autres d'identifier les comportements anormaux. Dans cette approche, une fenêtre de temps est utilisée afin d'extraire les patrons d'activités selon certaines contraintes comme la fréquence d'apparition d'items et l'intervalle de temps utilisé. Cette approche souffre du problème d'annotation des données, ainsi que le problème d'identification de la taille de la fenêtre temporelle en fonction des séquences. Un autre modèle de reconnaissance d'activités proposé par les mêmes auteurs [112] incorpore un modèle Markovien après la phase d'extraction des patrons. Ce modèle Markovien avait comme rôle de prédire les activités en se basant sur les patrons extraits.

Pour reconnaître les activités concurrentes et entrelacées, Tao et al. [39] proposent epSICAR (Emerging Patterns based approach to Sequential, Interleaved and Concurrent Activity Recognition). Cette approche est basée sur l'utilisation de patrons émergents pour différencier les différentes activités. L'avantage de cette approche est que les modèles d'activités sont construits par l'extraction des patrons émergents à partir des traces des activités séquentielles. Toutefois, cette approche souffre du problème d'annotation des données et elle ne décrit pas comment faire face au problème de bruit. Un modèle similaire est utilisé par Kim et al. [65] pour la reconnaissance d'activités simples. Cependant, il n'est pas tout à fait claire en quoi ce patron émerge diffère d'un patron fréquent.

L'un des problèmes rencontrés lors de la reconnaissance des activités est l'annotation des données. C'est à dire, le processus qui consiste à faire correspondre les événements de capteurs avec les activités réalisées. Un autre problème également rencontré dans le même contexte est celui du bruit. Le bruit présent dans les données de capteurs est généralement lié au problème du matériel. C'est à dire que le capteur peut se déclencher plusieurs fois successivement pour le même événement. Pour faire face à ces des problèmes, Chikhaoui et al. [27] proposent une approche de reconnaissance d'activités simples basée sur le principe de patrons fréquents. Cependant, cette approche nécessite des connaissances préalables sur les activités pour construire leur modèles.

D'autres méthodes sont utilisées pour reconnaître les activités en prenant en compte l'aspect temporel des activités [54]. Cependant, ces méthodes souffrent du

problème d'annotation des données.

1.4.1 Résumé sur les approches d'analyse des séquences

Dans cette section du chapitre, nous avons présenté les approches basées sur l'analyse des séquences. Nous avons constaté , à travers l'état de l'art des travaux présentés, que ces approches tirent leur force de la prise en compte de l'aspect séquentiel des données. De plus, ces approches cherchent à trouver des patrons qui permettent de décrire et caractériser les activités.

Selon notre revue des travaux existants, les approches basées sur l'analyse des séquences sont généralement jumelées avec d'autres modèles statistiques à savoir les modèles Markoviens tel que présenté dans le modèle de [112]. L'incorporation des modèles statistiques avec les approches d'analyse des séquences permet d'augmenter la performance des systèmes de reconnaissance d'activités [112].

Les approches d'analyse des séquences permettent de prendre en considération les aspects temporels des activités. Cela permet à ces approches de découvrir non seulement les activités, mais aussi les manières selon lesquelles ces activités sont réalisées. En fait, l'ordre d'apparition des événements dans la séquence joue un rôle très important dans le suivi et l'assistance des personnes lors de la réalisation des activités de la vie quotidienne.

Enfin, les approches basées sur l'analyse des séquences peuvent être jumelées avec des approches non supervisées afin de surmonter le problème d'annotation des données.

1.5 Discussion

Dans ce chapitre nous avons présenté des travaux publiés dans la littérature concernant la reconnaissance d'activités dans les environnements intelligents. Nous avons divisé ces travaux en deux grandes classes. La classe des travaux basés sur les événements et celle des travaux basés sur les séquences d'événements. Dans les deux classes, la reconnaissance d'activités est effectuée selon l'approche bottom-up i.e. les événements sont traités pour reconnaître des tâches qui, elles mêmes, sont utilisées

pour reconnaître les activités dans un plus haut niveau d'hiérarchie comme le montre la figure 1.7.

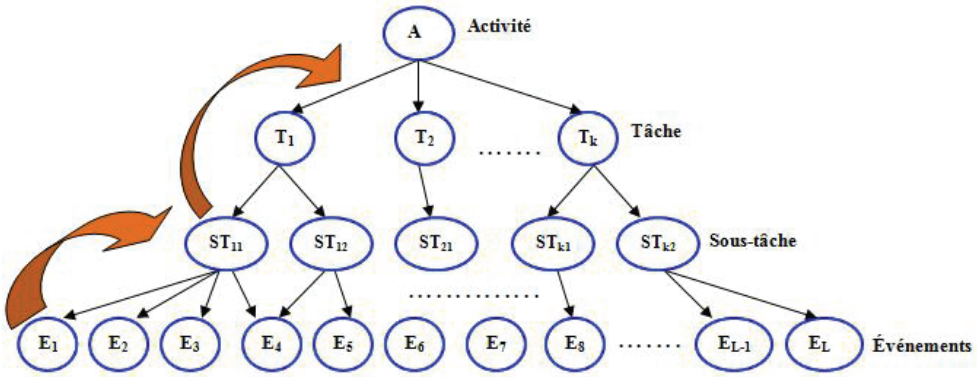


figure 1.7 – Méthode bottom-up de reconnaissance d'activités

Le tableau 1.2 présente un récapitulatif des travaux sus-présentés du point de vue du type d'activités (selon le niveau de complexité) à reconnaître. Quand plusieurs travaux se ressemblent, nous n'en citons qu'un seul.

tableau 1.2 – Tableau récapitulatif des travaux selon le type d'activités

Activités	Approche	Ref	Activités recunues	Attributs	Type de classification	Technique d'apprentissage	Avantages	Limites
Assis/ Debout/ Marcher/ Courir/ Lire/Dormir	Statistique	[106]	Les activités de transport : marcher prendre le bus.	données GPS Map arrêts des bus	Supervisée	Filtre Bayésien	Gère le bruit	1) Ne peut pas prédire les endroits non appris 2) Perte de signal GPS
	Statistique	[101]	Les AVQ : Regarder TV, lire, écouter la musique.	noms des objets données RFID	Supervisée	Réseau Bayésien dynamique	Gère le bruit	1) Annotation manuelle des données 2) activités simples
	Statistique	[14]	Les AVQ : Regarder TV, lire, dormir	noms des objets données RFID	Supervisée	HMM	Gère le bruit	1) Annotation manuelle des données
	Statistique	[95]	Les activités : Conversation distante ou téléphonique, présentation, présence.	énergie, moyenne, variance de la fréquence	Supervisée	HMM à couches	Gère le bruit Réduit la dimension de l'espace	1) Activités simples 2) le nombre de couches rend le modèle moins efficace
	Classification	[10]	Les activités Marcher, assis, courrir, Regarder TV, lire.	Position de l'accéléromètre	Supervisée	Arbre de décision	Reconnaît plusieurs activités	1) Ne prend pas le temps des activités 2) Ne gère pas le bruit
	Classification	[144]	Les activités Assis, debout, marcher, courir.	moyenne, correlation, énergie, RMSE, écart, type, variance	Supervisée	Réseau de neurones	Réduit la dimension de l'espace	1) Ne gère pas le bruit 2) utilise un seul capteur accéléromètre
	Séquence	[54]	AVQ : marcher, courir, descendre et monter les escaliers.	Erreur de reconstruction,	Non supervisée	Espaces propres comme classificateur	1) Prend en charge le temps 2) Pas d'apprentissage	1) Ne gère pas le bruit 2) Activités trés courtes
AVQ : S'habiller/ Préparer un repas/ Manger/ Néttoyer	Statistique	[127]	AVQ : Manger, se baigner,	Étiquette de l'AVQ noms des objets	Supervisée	Réseau Bayésien dynamique	Gère le bruit	1) Annotation d'AVQ
	Statistique	[14]	AVQ : Préparer des repas, Manger, prendre des médicaments.	identifiant d'objets obtenus par les tags RFID	Supervisée	HMM	Gère le bruit	1) Annotation d'AVQ 2) tous les objets doivent être marqués avec les tags RFID
	Séquence	[111]	AVQ : prendre médicament, entrer dans la maison, préparer les repas, dormir.	durée des activités, des patrons	Supervisée	Technique basée sur les patrons	1) Prend en charge le temps	1) Annotation d'AVQ 2) Taille de la fenêtre temporelle
AVQ : Concurrentes et entrelacées	Séquence	[39]	AVQ avec entrelacement : Manger, prendre des médicaments, s'habiller, téléphoner.	noms des objets, patrons émergents	Supervisée	Technique basée sur les patrons	1) Reconnaître les activités complexes	1) Ne gère pas le bruit
	Statistique	[81]	AVQ avec entrelacement : Manger, préparer un café, préparer la table, faire un jus.	noms des objets, étiquette des activités	Supervisée	HMM entrelacé	1) Reconnaître les activités complexes	1) Annotation d'AVQ
	Statistique	[50]	AVQ concurrentes et entrelacées : préparer à manger, préparer un café, manger.	états de capteurs, étiquette des activités	Supervisée	Markov logique	1) Reconnaître les activités complexes	1) Ne gère pas le bruit 2) Difficile d'utiliser la représentation logique des activités

Nous constatons à partir de notre revue de la littérature, que la plupart des approches souffrent de problèmes d'annotation de données et de bruit. En effet, les événements sont généralement caractérisés par la quantité du bruit véhiculé, l'ambiguïté et l'incertitude. Dans les approches basées sur les événements, l'utilisation directe de ces événements nécessite la prise en compte de tous ces problèmes lors du processus de la reconnaissance d'activités. En outre, plusieurs approches ne prennent pas en considération les informations temporelles, ce qui limite le processus de la reconnaissance à des activités simples. À cet effet l'analyse des séquences est introduite comme une nouvelle variante afin de surmonter quelques problèmes comme le bruit et les aspects temporels des activités, et de permettre d'identifier des patrons significatifs de chaque activité. De plus, la détection des patrons significatifs dans les séquences permettra non seulement de reconnaître les activités, mais aussi de caractériser les comportements des personnes, ce qui constitue un grand pas vers l'étude du profil comportemental de la personne. Par conséquent, et vu les avantages de l'analyse des séquences, nous pensons que l'analyse des séquences est une excellente piste qui mérite d'être explorée dans le but de faire face aux problèmes que nous venons de citer.

1.6 Positionnement de notre travail par rapport aux travaux existants

Dans cette thèse, et comparativement aux approches existantes, nous proposons une nouvelle approche à la fois pour la découverte et la reconnaissance des activités. Comme nous avons pu le constater à travers l'état de l'art, il existe peu de travaux qui traitent le problème de la découverte des activités. Dans ce contexte, l'approche que nous proposons définit un cadre théorique innovant qui tire ses bases à partir de deux fameuses approches de forage de données : 1) le forage de patrons fréquents, et 2) l'allocation Dirichlet latente (LDA). La combinaison de ces approches permet de découvrir les activités, de les reconnaître, et aussi d'exploiter les relations entre les patrons fréquents et les activités. Ces relations jouent un rôle très important dans le développement des algorithmes de prédiction en se basant sur l'observation

des patrons et tenant en compte les relations probabilistes qui lient les patrons aux activités.

En outre, nous avons aussi pu constater que les travaux existants ne sont pas dotés de modèles mathématiques d'activités qui vont servir dans le processus de la reconnaissance d'activités. L'apport de notre approche dans ce contexte est de construire les modèles d'activités de façon automatique. Ces modèles sont construits à l'aide du modèle LDA, et sont représentés sous forme de relations probabilistes entre les patrons fréquents et les activités auxquelles ils appartiennent. Contrairement aux approches existantes supervisées dans lesquelles les modèles d'activités sont construits en se basant sur des algorithmes de forage de données comme le modèle de Bayes naïf, SVM, arbre de décision, HMM, réseaux Bayésiens, etc, notre approche permet de construire les modèles d'activités de façon non supervisée et ne requiert pas la tâche d'annotation des données. Par conséquent, notre travail s'attaque à une problématique double dans laquelle deux problèmes importants sont résolus : 1) la découverte et la construction des modèles d'activités, et 2) la reconnaissance des activités.

Un autre apport important de notre approche est d'extraire les patrons significatifs pour chaque activité. En effet, chaque activité est caractérisée par un certain nombre de patrons. Cependant, ces patrons ne sont pas tous significatifs pour cette activité. Par conséquent, l'un des défis majeurs de notre approche est la découverte des patrons significatifs pour chaque activité. En fait, les patrons significatifs dans une activité possèdent des probabilités très élevées par rapport aux autres patrons. Le potentiel de découvrir ces patrons réside dans le fait que ces patrons peuvent être utilisés pour distinguer les activités, et ils peuvent aussi être considérés comme étant des attributs (features) dans un système de classification des activités. Cela constitue un problème nouveau qui n'a jamais été soulevé par les approches existantes basées sur l'analyse des séquences.

Chapitre 2

Découverte et reconnaissance des activités par la combinaison des patrons fréquents séquentiels et l'allocation Dirichlet latente (LDA)

2.1 Introduction

À travers notre revue de la littérature, nous avons constaté que la vaste majorité des approches souffrent du problème d'annotation des données et du problème de bruit dans les données. L'objectif de ce chapitre est de proposer une approche permettant de découvrir et de reconnaître des activités en palliant aux limites rencontrées dans les approches existantes, en s'inspirant des algorithmes de forage de données textuelles.

Notre décision d'opter pour un modèle statistique non supervisé (LDA) combiné avec le principe de patrons fréquents est motivée par deux raisons principales :

1. le modèle LDA est un modèle non supervisé. Par conséquent, la découverte des activités potentielles [1] peut être effectuée sans le besoin des données annotées. Cela constitue une contribution importante de notre travail.

1. Une activité potentielle est une activité découverte mais qui n'est pas encore validée comme activité réelle

2. l'utilisation des patrons fréquents permet de réduire de façon significative le bruit dans les données. Cela permet de faire face au problème de bruit rencontré dans plusieurs approches existantes. En outre, l'utilisation des patrons fréquents permet d'un coté de donner une description riche des activités, et de caractériser les relations de dépendance et d'ordre entre les événements qui composent les patrons de l'autre coté.

Contrairement aux modèles non supervisés existants tels que l'algorithme k-means, FCM, k-representative qui sont principalement basés sur des mesures de similarité, le modèle LDA n'adopte pas une mesure de similarité particulière. Il utilise des distributions de probabilités et la co-occurrence des variables afin de modéliser les relations entre les variables et leurs classes correspondantes. Cela constitue un potentiel important du modèle LDA.

Dans ce qui suit, nous allons introduire tout d'abord quelques notions et définitions sur l'analyse des séquences et les détails du modèle LDA, ensuite nous présenterons notre approche proposée pour la reconnaissance d'activités.

2.2 Analyse des séquences

L'analyse des séquences est un concept clé qui a pris de l'ampleur dans différents domaines tels que le Web, la biologie, la finance et les habitats intelligents récemment. Elle commence par découvrir les sous séquences fréquentes nommées aussi "motifs" ou "patrons fréquents". Les patrons qui apparaissent au moins un certain nombre de fois dans la séquence sont jugés intéressants. Cependant, la tâche de découvrir des patrons fréquents dans les séquences est un problème assez difficile vu que l'espace de recherche est d'ordre exponentiel.

La notion de patron fréquent prend principalement son application dans l'analyse des données de supermarchés, où se retrouvent différents produits avec des consommations différentes par les clients. Connaître les produits les plus demandés permettra une meilleure stratégie de prix, de promotion, de rangement et de publicité. Dans les données relatives aux supermarchés ou les données transactionnelles de façon générale, les items les plus recherchés représentent les itemsets fréquents. Les règles d'association élaborent les produits qui s'achètent ensemble. Si des contraintes tem-

porelles complètent l'achat des produits, par exemple l'achat d'un ordinateur est suivi de l'achat d'un modem plustard, on parle alors d'un patron séquentiel.

L'utilisation des patrons fréquents donne beaucoup d'avantages :

- le patron fréquent contient une description très riche de l'information en prenant un ensemble d'items au lieu d'un seul.
- l'utilisation des patrons fréquents permet de réduire significativement la quantité de bruit présent dans les données. Cela est due au fait que le bruit apparaît de façon irrégulière dans les données.
- l'utilisation des patrons fréquents conduit à identifier les patrons significatifs parmi toutes les données et d'explorer ces patrons pour des fins d'analyse et d'exploitation.
- le patron fréquent est facilement interprétable par la prise en compte des items contenant dans le patron et l'ordre entre ces items.

Dans la dernière décennie, plusieurs méthodes de découverte de patrons fréquents dans les séquences et les données transactionnelles ont eu lieu, depuis le travail fondateur de Agrawal et Srikant [6]. Dans ce qui suit, nous allons introduire quelques définitions formelles ainsi qu'un aperçu sur les différents méthodes proposées dans la littérature pour le forage des patrons fréquents.

Selon Han J. et al. [43], un patron fréquent est un itemset, une sous-séquence ou une sous-structure qui apparaît dans une base de données avec un nombre de fois supérieur ou égal à un certain seuil spécifié par l'utilisateur. Une séquence est une liste ordonnée d'items. Dans ce qui suit, nous adopterons les concepts et définitions de Agrawal et al. [6].

Définition 3. Item : Soit $\mathcal{A} = \{a_1, a_2, ..., a_n\}$ un ensemble d'attributs. Chaque attribut a_i de $\mathcal{A}$ est également appelé item. Un itemset I est un ensemble non vide d'items noté $I = (i_1, i_2, ..., i_t)$ où $i_j \in \mathcal{A}$.

Définition 4. Transaction : Soit $\mathcal{U}$ un ensemble de clients, et $\mathcal{D}$ un ensemble de dates. Une transaction T, pour un client $c \in \mathcal{C}$ dans une date $d \in \mathcal{D}$, est l'ensemble d'items I achetés par le client c à la date d. Une transaction s'écrit sous la forme d'un triplet : $< c, d, I >$.

Définition 5. Patron séquentiel : Un patron séquentiel est dit maximal s'il n'est

pas contenu dans aucun autre patron séquentiel ou séquence. Le nombre d'items dans une séquence s est appelé la longueur de la séquence et est notée $|s|$. Une séquence de longueur k est appelée $k - séquence$. Une séquence $a =< a_1 a_2 ... a_n >$ est contenue dans une autre séquence $b =< b_1 b_2 ... b_m >$ ou a est une $sous - séquence$ de b, s'il existe des entiers $1 < i_1 < i_2 < ... < i_n < m$ tel que $a_1 = b_{i_1}, a_2 = b_{i_2}, ..., a_n = b_{i_n}$.

Exemple 3. Le patron séquentiel $ps_1 =$ <(2)(3 4)(6)> est inclus dans le patron séquentiel $ps_2 =$ <(2 7)(3 4 5)(6)(7 8)> car, (2) $\subseteq$ (2 7), (3 4) $\subseteq$ (3 4 5), et (6) $\subseteq$ (6). Les items dans les parenthèses représentent des items achetés ensemble.

Définition 6. Support : Soit $T = \{t_1, t_2, ..., t_n\}$ une base de données contenant un ensemble de transactions t_i. Soit p_α un patron dans T, et $P_T = \{p_1, p_2, ..., p_l\}$ l'ensemble de ces patrons. On note l'ensemble de transactions dans lesquelles p_α apparaît comme $T_\alpha = \{t_i | p_\alpha \in t_i, t_i \in T\}$. Le support de patron p_α dans la base de données T est défini par la quantité : $\frac{|T_\alpha|}{|T|}$.

Il existe plusieurs définitions pour le support selon le contexte dans la littérature. Par exemple, dans Agrawal et Srikant [6], le support est défini comme le rapport entre les transactions contenant l'itemset I et toutes les transactions dans la base de données. Dans Mannila et al. [89], le support est défini comme le rapport entre le nombre de fenêtres glissantes contenant l'itemset et le nombre total des fenêtres glissantes. Dans notre cas, nous adopterons la définition de Agrawal et Srikant [6] vu que nous n'avons pas utilisé le concept de fenêtre glissante.

Définition 7. Patron fréquent : Soit $T = \{t_1, t_2, ..., t_n\}$ une base de données contenant un ensemble de transactions t_i, qui peuvent être des itemsets, des séquences ou des graphes. Soit p_α un patron dans T, et $P_T = \{p_1, p_2, ..., p_l\}$ l'ensemble de ces patrons. On note l'ensemble de transactions dans lesquelles p_α apparaît comme $T_\alpha = \{t_i | p_\alpha \in t_i, t_i \in T\}$. Un patron p_α est fréquent dans la base de données T, si $\frac{|T_\alpha|}{|T|} \geq \sigma$, où σ est un seuil défini par l'utilisateur, et $\frac{|T_\alpha|}{|T|}$ est le support de p_α.

Le forage de patrons fréquents ou patrons séquentiels fréquents est le processus qui permet d'extraire tous les patrons fréquents tels que définis précédemment dans la base de données. L'annexe B présente en détails les principaux algorithmes d'extraction de patrons fréquents à partir des bases de données.

2.3 Latent Dirichlet Allocation (LDA)

Le LDA [11] est un modèle graphique génératif conçu pour analyser les thèmes latents dans un document. L'idée de base est qu'un document est un mélange probabiliste de thématiques latentes (c'est à dire cachées de nos yeux curieux). Chaque thématique est caractérisée par une distribution de probabilités sur les mots qui lui sont associés. On constate donc que l'élément clé dans le modèle LDA est la notion de thématique, c'est à dire que la sémantique est prioritaire sur la syntaxe (la notion de terme ou mot).

Par ailleurs, le modèle LDA a été conçu pour éviter que la distribution de probabilités qui sert au choix de thématique soit dépendante des documents connus précédemment (cela signifie que le modèle est capable de prendre en compte des documents non connus à l'avance). Tout cela vient de choix techniques judicieux des distributions de probabilités. En particulier on mentionne le nom de Dirichlet. La distribution Dirichlet est une famille de lois de probabilité continues pour des variables aléatoires multinomiales [82]. Elle permet un choix probabiliste efficace (de thématiques) sur les distributions multinomiales. Au lieu de tirer au sort une thématique comme avec le modèle PLSA [51], on tire d'abord au sort une méthode de tirage sur les thématiques [2]. C'est pour cela le modèle LDA est beaucoup plus puissant que les autres modèles latents tel que le PLSA.

Le modèle LDA suppose qu'un document est un mélange de plusieurs thèmes (topics), et qu'un mot dans un document est généré par l'un de ces thèmes (topics). Le modèle LDA peut être vu comme une extension du modèle PLSA [51], qui n'est pas vraiment un modèle génératif vu qu'il n'est pas capable de générer de nouveaux documents, chose qui a été résolue avec le modèle LDA. Les deux critiques essentielles que Blei et al. [11] expriment à l'égard du modèle PLSA sont présentées dans les deux points suivants :

1. le modèle PLSA n'est pas un modèle génératif de documents, car la variable aléatoire qu'il emploie pour les textes suit une loi multinomiale qui prend comme valeur les numéros de textes uniquement dans l'ensemble d'apprentissage. Elle ne peut donc pas estimer la probabilité des documents non vus,

2. http ://www.spoonylife.org/algorithms-and-computation/latent-dirichlet-allocation

2. le nombre de paramètres dans le modèle PLSA augmente en fonction du nombre de documents utilisés pour estimer les paramètres (donc une augmentation linéaire par rapport au nombre de documents). Cela était à l'origine du problème de sur-apprentissage observé dans le modèle PLSA.

Le modèle LDA a été appliqué dans différents domaines, dont le traitement automatique du langage naturel [9], le traitement d'images [78], la modélisation de séquences ADN en biologie [22], la médecine [151], les réseaux sociaux [46].

2.3.1 Formalisme du modèle LDA

Dans cette section, nous allons introduire le formalisme du modèle LDA en se référant à l'article de base de David Blei et al. [11] dans le cadre du forage de textes. Pour ce faire, nous définissons :

- le "mot" est l'unité de base des données. Un mot est défini comme un terme qui fait partie d'un vocabulaire indexé par {1, 2, ..., V}. Le $v^{\text{ème}}$ mot dans le vocabulaire est représenté par un vecteur de $V dimensions$ tel que $w^v = 1$ et $w^u = 0$ pour $u \neq v$.
- V est le nombre de termes t dans le vocabulaire.
- M est le nombre de documents dans le corpus.
- un document est une séquence de N_m mots notée par $m = (w_1, w_2, ..., w_{N_m})$.
- un corpus est un ensemble de M documents noté $D = m_1, m_2, ..., m_M$.
- z désigne un thème. Nous utilisons les mots : topic, thème, classe ou thématique pour dire la même chose.
- K est le nombre de thèmes (topics).
- α est un paramètre de la distribution Dirichlet a priori de la proportion des thèmes dans le document (K-vecteur).
- β est un paramètre de la distribution Dirichlet a priori de la proportion des termes dans un thème (V-vecteur).
- θ_m désigne la distribution des thèmes dans le document m, noté souvent $p(z|d = m)$. Pour tous les documents dans le corpus, on note : $\Theta = \{\theta_m\}_{m=1}^{M}$ une matrice $(M \times K)$.
- ϕ_k désigne la distribution des termes dans le thème k, noté souvent $p(t|z = k)$.

Pour tous les thèmes, on note : $\Phi = \{\phi_k\}_{k=1}^{K}$ une matrice $(K \times V)$.

L'idée de base est que les documents sont représentés comme un mélange aléatoire de topics latents, où chaque topic est caractérisé par une distribution sur les mots comme le montre en détail la figure 2.1.

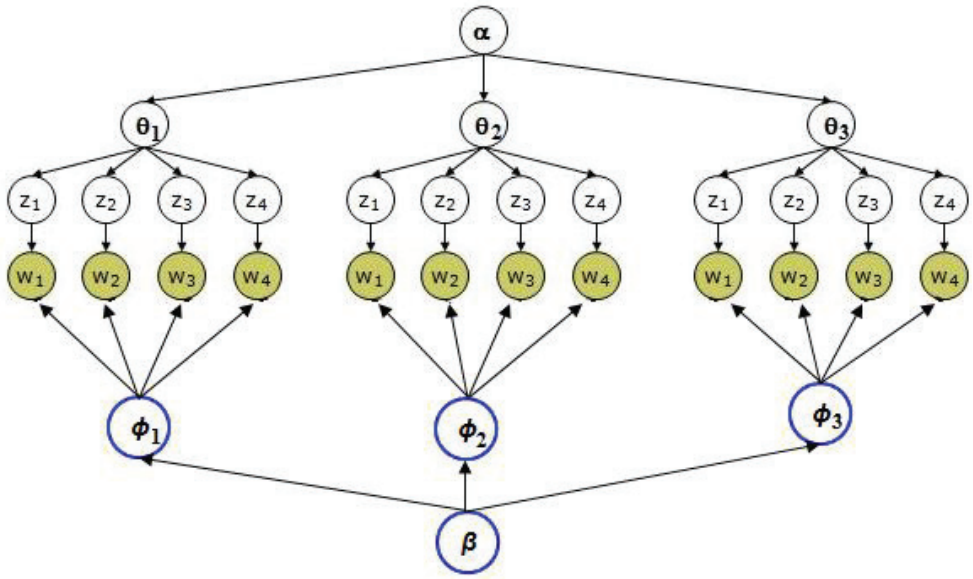


figure 2.1 – Représentation détaillée du LDA

Le modèle de génération proposé par LDA pour chaque document m dans le corpus D est alors le suivant :

1. Pour tous les thèmes $k \in [1..K]$ faire :
   - choisir les composantes $\phi_k \sim Dir(\beta)$
2. Pour tous les documents $m \in [1..M]$ faire :
   - choisir la proportion $\theta_m \sim Dir(\alpha)$
   - choisir la longueur de document $N_m \sim Poisson(\xi)$
   - Pour tous les mots $n \in [1..N_m]$ faire :
     - choisir un index du thème $z_{m,n} \sim Multinomial(\theta_m)$
     - choisir un terme pour le mot $w_{m,n} \sim Multinomial(\phi_{z_{m,n}})$

Ce modèle de génération est présenté graphiquement dans la figure 2.2. Dans le modèle présenté dans la figure 2.2, chaque document est modélisé par un mélange de

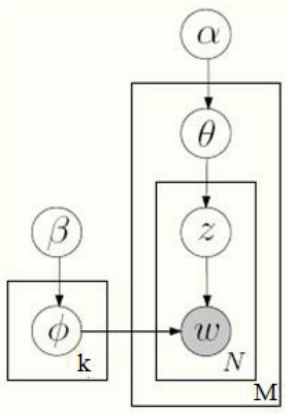


figure 2.2 – Modèle graphique de génération de LDA

thèmes. Chaque thème génère ensuite les mots du document. Quelques hypothèses sont faites sur ce modèle : 1) la dimensionnalité K de la distribution Dirichlet (et donc aussi la dimensionnalité de la variable de topic z) est supposée connue et fixée à l'avance ; 2) le nombre N qui représente la longueur de document est indépendant de toutes les variables (θ et z).

Avant d'introduire les formules d'estimation des paramètres du modèle LDA, nous introduisons tout d'abord la distribution de Dirichlet et comment cette distribution est utilisée dans le modèle LDA. Dans le modèle LDA, le mélange de topics pour chaque document est tiré selon une certaine distribution. Donc, nous voulons mettre une distribution sur le mélange de topics telle que présentée dans la figure 2.1.

La densité de probabilité de la distribution de Dirichlet pour la variable aléatoire $\theta = (\theta_1, \theta_2, ..., \theta_K)$ [3] de paramètres $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K)$ est donnée par la formule suivante :

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{2.1}$$

où le paramètre α est un K-vecteur avec les composantes $\alpha_k > 0$, et $\Gamma(x)$ est la fonction gamma. Comme le montre la figure 2.1, le paramètre de Dirichlet α_k peut être considéré comme un compte apriori (prior count) du k^{eme} topic (classe).

Remarque : Durant les deux dernières décennies, le problème d'estimation des paramètres qui déterminent le mélange de lois a fait l'objet de diverses études. La

3. Cette représentation vise à simplifier le formule 2.1

méthode du maximum de vraisemblance est devenue l'approche la plus utilisée pour résoudre ce problème. Cette approche est utilisée pour traiter le problème d'estimation des paramètres comme un problème d'optimisation.

Selon le modèle LDA, la probabilité qu'un mot $w_{m,n}$ instancie un terme particulier t sachant les paramètres du LDA peut se calculer comme suit :

$$p(w_{m,n} = t|\theta_m, \Phi) = \sum_{k=1}^{K} p(w_{m,n} = t|\phi_k)p(z_{m,n} = k|\theta_m) \tag{2.2}$$

La formule 2.2 signifie que la probabilité d'un mot est calculée en prenant en compte le thème dans la distribution θ qui sera utilisé comme indice dans la distribution Φ.

Nous pouvons donc spécifier la vraisemblance des données d'un document, c-à-d, la probabilité jointe de toutes les variables visibles et cachées étant donnés les hyperparamètres.

$$p(w_m, z_m, \theta_m, \Phi|\alpha, \beta) = \prod_{n=1}^{N_m} p(w_{m,n}|\phi_{z_{m,n}}).p(z_{m,n}|\theta_m).p(\theta_m|\alpha).p(\Phi|\beta) \tag{2.3}$$

Cette formule est utilisée comme une base pour faire d'autres dérivations. En intégrant sur θ_m et Φ et sommant sur tous les $z_{m,n}$, la vraisemblance d'un document w_m peut s'écrire de la façon suivante :

$$\begin{aligned} p(w_m|\alpha, \beta) &= \int\int p(\theta_m|\alpha).p(\Phi|\beta). \prod_{n=1}^{N_m} \sum_{z_{m,n}} p(w_{m,n}|\phi_{z_{m,n}}).p(z_{m,n}|\theta_m) d\Phi d\theta_m \\ &= \int\int p(\theta_m|\alpha).p(\Phi|\beta). \prod_{n=1}^{N_m} p(w_{m,n}|\theta_m, \Phi) d\Phi d\theta_m \end{aligned} \tag{2.4}$$

Finalement, la vraisemblance totale de tout le corpus $D = \{w_m\}_{m=1}^{M}$ s'obtient en multipliant les vraisemblances de tous les documents comme suit :

$$\begin{aligned} p(D|\alpha, \beta) &= \prod_{m=1}^{M} p(w_m|\alpha, \beta) \\ &= \prod_{m=1}^{M} \int\int p(\theta_m|\alpha).p(\Phi|\beta). \prod_{n=1}^{N_m} p(w_{m,n}|\theta_m, \Phi) d\Phi d\theta_m \end{aligned} \tag{2.5}$$

Estimation des paramètres

Dans cette section, nous allons introduire les méthodes les plus utilisées pour l'estimation des paramètres dans LDA.

L'estimation des paramètres consiste à déterminer, pour un corpus de documents, les paramètres α et β qui maximisent la vraisemblance des données. Cela peut se faire en utilisant l'échantillonnage de Gibbs proposé par Griffiths et Steyvers [38], l'algorithm EM variationnel [107], ou d'autres méthodes. Nous allons présenter la méthode de l'échantillonnage de Gibbs car elle est la plus utilisée dans la littérature vu sa simplicité. L'échantillonnage de Gibbs est un cas particulier de l'algorithme de Monte Carlo par chaîne de Markov (Monte Carlo Markov Chain MCMC) nommé l'algorithme de $Metropolis-Hastings$ [87]. Il fournit souvent des algorithmes simples pour des inférences approximatives dans des modèles de haute dimensionnalité [45].

L'idée des méthodes MCMC est d'utiliser une chaîne de Markov pour générer une distribution stationnaire, à partir de laquelle nous pouvons effectuer un échantillonnage [45]. À chaque étape, l'algorithme met à jour chaque dimension des données plusieurs fois durant un processus appelé "burn-in", qui signifie que la distribution stationnaire n'a pas encore été atteinte. Après l'étape de "burn-in", la distribution stationnaire est obtenue et nous pouvons donc obtenir des échantillons. Plutôt que de présenter les détails de la méthode MCMC, nous allons présenter comment la méthode de Gibbs est utilisée dans le cadre du modèle LDA pour estimer les paramètres.

Les lois de Dirichlet a priori α et β sont des conjuguées des lois multinomiales θ et ϕ, ce qui permet de calculer la distribution jointe $P(D, Z)$ en intégrant sur θ et ϕ. En effet, $P(D, Z) = P(D|Z)P(Z)$ et les distributions ϕ et θ apparaissent respectivement dans le premier et deuxième terme, ce qui permet d'intégrer séparément.

En intégrant sur ϕ, nous obtenons d'après [38] :

$$P(D|Z) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K \prod_{j=1}^{K} \frac{\prod_w \Gamma(n_j^{(w)} + \beta)}{\Gamma(n_j^{(.)} + V\beta)} \tag{2.6}$$

où $n_j^{(w)}$ est le nombre de fois où le mot w a été assigné au thème j, et $\Gamma(.)$ est la fonction gamma. Le second terme peut être calculé en faisant l'intégrale sur θ comme

suit :

$$P(Z) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^M \prod_{m=1}^{M} \frac{\prod_j \Gamma(n_j^{(m)} + \alpha)}{\Gamma(n_\cdot^{(m)} + K\alpha)} \tag{2.7}$$

où $n_j^{(m)}$ est le nombre de fois où un mot du document m est assigné au thème j. Notre objectif consiste donc à calculer la distribution a posteriori $P(Z|D)$:

$$P(Z|D) = \frac{P(D,Z)}{\sum_Z P(D,Z)} \tag{2.8}$$

Cependant, la somme du dénominateur ne peut pas être factorisée et prend un temps considérable qui est de l'ordre de K^n, où n est le nombre total d'instances de mots dans le corpus.

La méthode de Gibbs est utilisée afin de réaliser un échantillonnage à partir de la distribution a posteriori $P(Z|D)$. Cependant, pour échantillonner à partir de la distribution $P(Z|D)$, nous aurons besoin de la distribution conditionnelle a posteriori $P(z_i|Z_{-i}, D)$, où Z_{-i} désigne tous les z_j tel que $j \neq i$. Notons qu'en particulier, la méthode de Gibbs ne requiert pas la connaissance exacte de $P(z_i|Z_{-i}, D)$. Il suffit de trouver une fonction d'approximation $f(.)$ qui a la particularité suivante :

$$P(z_i|Z_{-i}, D) \propto P(z_i|Z_{-i}, D) \tag{2.9}$$

L'objectif est donc d'estimer la probabilité $P(z_i = j|Z_{-i}, D)$. Ici les paramètres θ et ϕ sont intégrés. Si nous connaissons les valeurs du vecteur Z pour chaque document, le calcul de θ et ϕ sera facile.

$$\begin{aligned} P(z_i = j|Z_{-i}, D) \propto & P(z_i = j, Z_{-i}, D) \\ = & P(w_i|z_i = j, Z_{-i}, D_{-i}) P(z_i = j|Z_{-i}, D_{-i}) \\ = & P(w_i|z_i = j, Z_{-i}, D_{-i}) P(z_i = j|Z_{-i}) \end{aligned} \tag{2.10}$$

Le premier terme de l'équation 2.10 est la vraisemblance, et le second terme représente un a priori.

Prenons maintenant le premier terme :

$$\begin{aligned} P(w_i|z_i = j, Z_{-i}, D_{-i}) &= \int P(w_i|z_i = j, \phi^{(j)})P(\phi^{(j)}|Z_{-i}, D_{-i})d\phi^{(j)} \\ &= \int \phi^{(j)}_{w_i} P(\phi^{(j)}|Z_{-i}, D_{-i})d\phi^{(j)} \\ P(\phi^{(j)}|Z_{-i}, D_{-i}) &\propto P(D_{-i}|Z_{-i},) \\ &\sim Dirichlet(\beta + n^{(w)}_{-i,j}) \end{aligned} \tag{2.11}$$

Ici, $n^{(w)}_{-i,j}$ représente le nombre d'instances du mot w assignées au thème j.

Par l'utilisation de la propriété de l'espérance de la distribution Dirichlet ($E(X_i) = \alpha_i / \sum \alpha_i$), nous aurons :

$$P(w_i|z_i = j, Z_{-i}, D_{-i}) = \frac{n^{(w_i)}_{-i,j} + \beta}{n^{(.)}_{-i,j} + V\beta} \tag{2.12}$$

Où $n_{-i,j}$ représente le nombre total de mots assignés au thème j.

De la même façon que précédemment, nous pouvons obtenir le deuxième terme de l'équation 2.10.

$$\begin{aligned} P(z_i = j|Z_{-i}) &= \int P(z_i = j|\theta^{(m)})P(\theta^{(m)}|Z_{-i})d\theta^{(m)} \\ P(\theta^{(m)}|Z_{-i}) &\propto P(Z_{-i}|\theta^{(m)})P(\theta^{(m)}) \\ &\sim Dirichlet(\alpha + n^{(m)}_{-i,j}) \end{aligned} \tag{2.13}$$

Où $n^{(m)}_{-i,j}$ représente le nombre de mots assignés au thème j excepté le mot courant. Donc, en utilisant la propriété de l'espérance de la distribution Dirichlet, on obtient :

$$P(z_i = j|Z_{-i}) = \frac{n^{(m)}_{-i,j} + \alpha}{n^{(m)}_{-i,.} + K\alpha} \tag{2.14}$$

Où $n^{(m)}_{-i,.}$ représente le nombre total de thèmes assignés au document m excepté le thème courant.

En conclusion, l'équation 2.10 devient :

$$P(z_i = j|Z_{-i}, D) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(.)} + V\beta} \frac{n_{-i,j}^{(m)} + \alpha}{n_{-i,.}^{(m)} + K\alpha} \tag{2.15}$$

Donc, dans cette formule, nous aurons besoin de quatre variables de comptage :

1. le compteur document-thème $n_{-i,j}^{(m)}$
2. la somme document-thème $n_{-i,.}^{(m)}$
3. le compteur thème-terme $n_{-i,j}^{(w_i)}$
4. la somme thème-terme $n_{-i,j}^{(.)}$

Une fois Z connue, nous pouvons estimer les paramètres θ et ϕ comme suit :

$$\phi_{j,w} = \frac{n_w^{(j)} + \beta}{\sum_{w=1}^{V} n_w^{(j)} + V\beta} \tag{2.16}$$

$$\theta_j^{(m)} = \frac{n_j^{(m)} + \alpha}{\sum_{z=1}^{K} n_z^{(m)} + K\alpha} \tag{2.17}$$

Où $n_w^{(j)}$ est le nombre de fois le mot w est assigné au thème j, et $n_z^{(m)}$ est le nombre de mots assignés au thème z. L'algorithme global de l'échantillonnage de Gibbs pour LDA est présenté dans l'annexe C.

2.4 Notre approche de reconnaissance d'activités

Dans cette section nous allons présenter notre approche de reconnaissance d'activités en se basant sur le modèle LDA présenté précédemment et le principe de forage de patrons fréquents [4].

2.4.1 Modèle d'activités

Les modèles d'activités représentent la composante principale dans un système de reconnaissance d'activités. Comme nous avons pu le constater à travers l'état de

4. Ici les termes patron fréquent et patron séquentiel ont la même signification

l'art, les approches existantes de reconnaissance d'activités supposent des modèles prédéfinis ou construisent ces modèles en utilisant les données d'apprentissage.

Selon les spécialistes d'étude du comportement humain [80], les humains accomplissent leurs activités de façon hiérarchique structurée. Par conséquent, les activités sont composées de tâches et de sous-tâches décomposées jusqu'aux tâches terminales qui ne peuvent plus être décomposées. On distingue différents types d'activités : les activités séquentielles, les activités concurrentes et entrelacées. Dans notre travail, on s'intéresse aux activités séquentielles, i.e. les activités sont supposées être réalisées une après l'autre sans chevauchement ou entrelacement, une activité à la fois.

L'ordre de la réalisation des tâches varie selon le type de l'activité. Par exemple, dans les activités séquentielles, si la tâche T_1 est exécutée avant la tâche T_2, alors toutes les sous-tâches de T_1 doivent être réalisées avant celles de T_2. L'ordre est donc très important.

Nous avons mentionné auparavant que la plupart des approches de reconnaissance d'activités adoptent le principe de reconnaissance de bas vers le haut (bottom-up), i.e. de l'événement à l'activité. Vu que notre modèle est inspiré du modèle LDA, notre méthode de la reconnaissance d'activités sera différente. En effet, dans le modèle LDA, les documents sont représentés sous forme de thèmes et les thèmes sont représentés sous forme de mots. Le modèle LDA commence tout d'abord par choisir les thèmes, puis il passe aux mots associés aux thèmes. Par conséquent, notre modèle choisit tout d'abord les activités puis les patrons qui leur sont associés. Donc notre méthode de la reconnaissance d'activités est plutôt dirigée de haut en bas (top-down) dans un modèle hiérarchique. La figure 2.3 présente la correspondance entre le modèle LDA et notre modèle hiérarchique de reconnaissance d'activités.

Les activités peuvent partager des tâches ou objets en commun. Cela pourrait créer des confusions et ambiguïtés et rendre le processus de la reconnaissance d'activités beaucoup plus difficile. Selon notre revue de la littérature, nous avons constaté que ce problème n'a pas encore été abordé. À cet effet, nous présenterons dans notre approche, une nouvelle piste de solution pour résoudre ce problème.

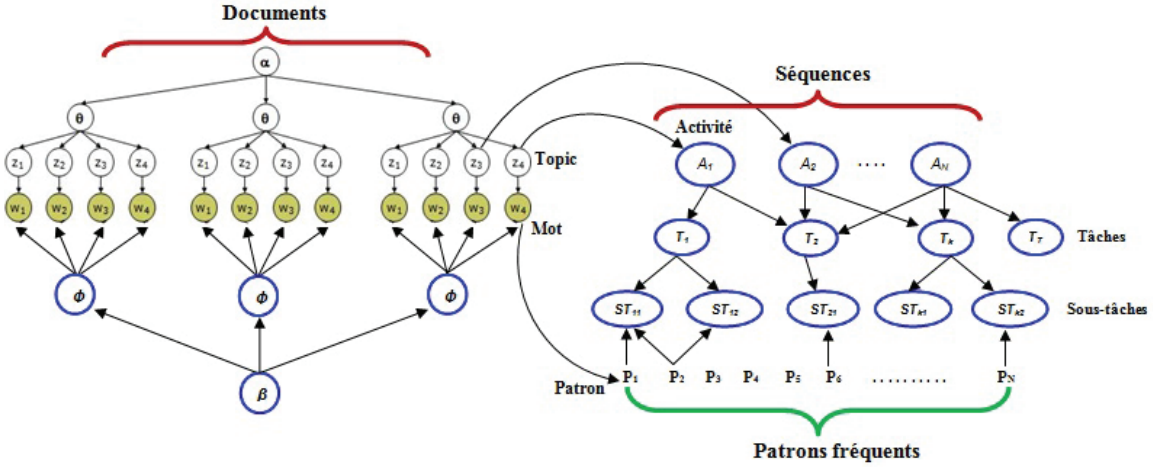


figure 2.3 – Modèle hiérarchique d'activité et correspondance avec le modèle LDA

2.4.2 Motivation pour l'utilisation d'un modèle d'activités basé sur LDA

Dans la section précédente nous avons présenté le modèle LDA et comment ce modèle est utilisé pour la détection des thèmes dans des documents. Malgré sa complexité, il possède un potentiel important dans le traitement et la compréhension des relations entre les mots et les thèmes. Nous avons constaté que la puissance du modèle LDA réside dans la définition des thèmes. Ces thèmes, qui possèdent une certaine signification statistique ou sémantique, permettent d'identifier et de classer les documents. L'intérêt de cette approche est que les thèmes extraits sont des structures sémantiques qui peuvent être facilement interprétables.

Idéalement, s'il existe un modèle mathématique pour décrire et représenter les activités, alors le problème sera plus facilement résolu. Il suffira de représenter les tâches et sous-tâches en utilisant ce modèle et de les chercher ensuite dans les séquences. Malheureusement, ce modèle mathématique n'existe pas. Pour cette raison, nous cherchons une technique permettant d'identifier les activités dans les séquences. Pour relever ce défi, nous avons pensé au modèle LDA qui permet de détecter des thèmes représentant des activités. À l'aide de ce modèle, nous pouvons identifier les activités dans les séquences. Cela nous permet de comparer aussi les séquences à la base des activités identifiées.

En outre, le modèle LDA permet de comprendre la relation entre les activités et les patrons fréquents à l'aide d'un modèle statistique. Cela nous permet par la suite de reconnaître les différents types d'activités comme les activités concurrentes et entrelacées.

2.4.3 Annotation des données

L'annotation des données, appelée aussi annotation d'activités, constitue un autre aspect de la reconnaissance d'activités et un défi majeur que rencontrent les approches de reconnaissance d'activités. L'annotation des activités est une tâche préalable à la reconnaissance et s'inscrit dans le contexte de prétraitement des données requis afin de permettre aux approches de reconnaissance d'activités de s'exécuter.

La plupart des chercheurs annotent leurs activités manuellement. Cela veut dire que l'annotation s'effectue au fur et à mesure de la réalisation des expérimentations en inspectant les états de capteurs comme le travail présenté par Wren et al.[137]. Une autre méthode d'annotation utilisée consiste à demander à l'usager d'annoter ses activités. Donc, c'est l'usager qui est responsable d'annoter ses propres activités en spécifiant les moments de début et fin de chaque activité. Ce travail a été adopté par plusieurs chercheurs à savoir Liao et al.[72], Tapia et al. [124], et Philipose et al. [102]. Cependant, cette méthode rend l'usager moins libre dans le sens où il ne doit pas oublier d'annoter une activité particulière. De plus, cette méthode n'est pas pratique, plus spécifiquement lorsque l'usager est une personne qui présente des déficits cognitifs ou qui souffre de la maladie d'Alzheimer. Ces personnes souvent oublient même quoi faire et requièrent une assistance continue pour pouvoir accomplir leur tâches quotidiennes.

Une troisième méthode implique l'expérimentateur qui indique à l'usager l'activité qu'il doit réaliser. Les activités sont alors annotées par l'expérimentateur avant même que les états de capteurs soient collectés. C'est le cas notamment des travaux de Cook et al.[28], Gu et al. [39], Maurer et al. [88].

Dans le cas de notre travail, nous avons besoin des données annotées pour évaluer notre méthode. Ainsi, les données issues du laboratoire Domus ont été annotées selon la première méthode. Nous avons analysé les données de capteurs en s'appuyant sur

les vidéos collectées lors des expérimentations, afin que l'annotation soit la plus précise possible.

Comme nous pouvons le constater à travers cette présentation des différentes méthodes existantes, l'annotation des données est une tâche longue qui est sujette aux erreurs. Par conséquent, les approches qui adoptent ces méthodes sont difficilement déployables dans des environnements intelligents réels. Par conséquent, pour créer un système de reconnaissance pratique et utilisable, notre approche tente de surmonter le problème d'annotation des données grâce à une méthode qui permet de découvrir les différentes activités de façon non supervisée.

2.4.4 Patrons fréquents et activités

L'utilisation du principe de patrons fréquents est devenue une pratique importante vu les résultats obtenus par les approches basées sur ce principe dans différents domaines. Dans le domaine des activités de la vie quotidiennes, les patrons fréquents représentent une vérité apparente qui se traduit par des tâches répétitives lors de la réalisation des activités. C'est cette répétition qui fait en sorte que les patrons fréquents paraissent comme une solution prometteuse pour représenter les activités ainsi que les tâches qui les composent. La figure 2.4 présente un exemple d'un patron répétitif dans une séquence d'événements.

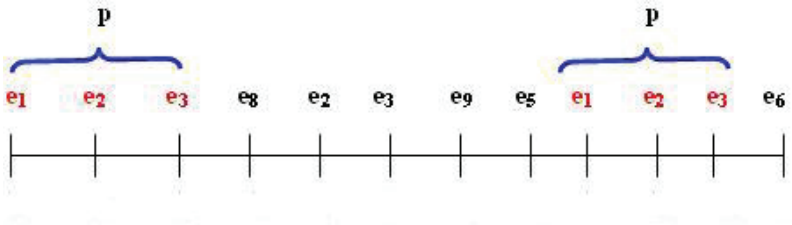


figure 2.4 – Exemple d'un patron répétitif dans une séquence d'événements

Dans notre travail, les patrons sont utilisés pour représenter les activités. Une activité peut être composée d'un ensemble de patrons. Chaque activité possède un certain nombre de patrons qui sont importants. La détection de ces patrons intéressants pour chaque activité et les relations qui lient ces patrons aux activités permettent non seulement de reconnaître les activités, mais aussi de construire un modèle pour

chaque activité. Nous allons voir à travers la présentation de notre approche comment détecter les patrons intéressants pour chaque activité et comment ces patrons peuvent aussi être utilisés pour dériver la sémantique reliée aux patrons.

2.4.5 Découverte des activités potentielles à l'aide du modèle LDA

Notre approche de reconnaissance d'activités est composée principalement de deux étapes importantes :

1. La première étape consiste, à la base du modèle LDA, à découvrir les activités potentielles dans les séquences. Cette étape elle même requiert la recherche des patrons fréquents dans les séquences. Nous utilisons le terme "activité potentielle" pour signifier les activités qui sont découvertes mais qui n'ont pas encore été validées comme activités réelles.
2. La deuxième étape consiste à reconnaître les activités découvertes.

La figure 2.5 présente de manière graphique les différentes étapes de notre approche.

Dans cette section nous allons introduire la première étape qui correspond à la découverte des activités potentielles dans les séquences.

Pour pouvoir aborder cette problématique, nous aurons tout d'abord besoin d'extraire les patrons fréquents à partir de la base de données des séquences. Pour ce faire, nous avons opté pour l'algorithme Apriori vu ses avantages par rapport à l'algorithme FP-Growth. Notons que notre objectif principal n'est pas de développer un algorithme de recherche de patrons fréquents dans les bases de données, mais plutôt de savoir ce que l'utilisation des patrons fréquents peut nous apporter dans le contexte de la reconnaissance des activités. Par conséquent, l'extraction des patrons fréquents à l'aide de l'algorithme Apriori ne constitue qu'un choix, et ne peut en aucun cas affecter notre approche de reconnaissance d'activités. Les autres algorithmes d'extraction de patrons fréquents peuvent aussi être utilisés à ce stade.

Comme nous l'avons mentionné auparavant, nous traitons le problème de découverte des activités comme un problème d'optimisation. Nous explorons le principe du modèle LDA afin de modéliser les activités dans les séquences comme la distribution des thèmes dans un document. Par conséquent, notre approche modélise les séquences

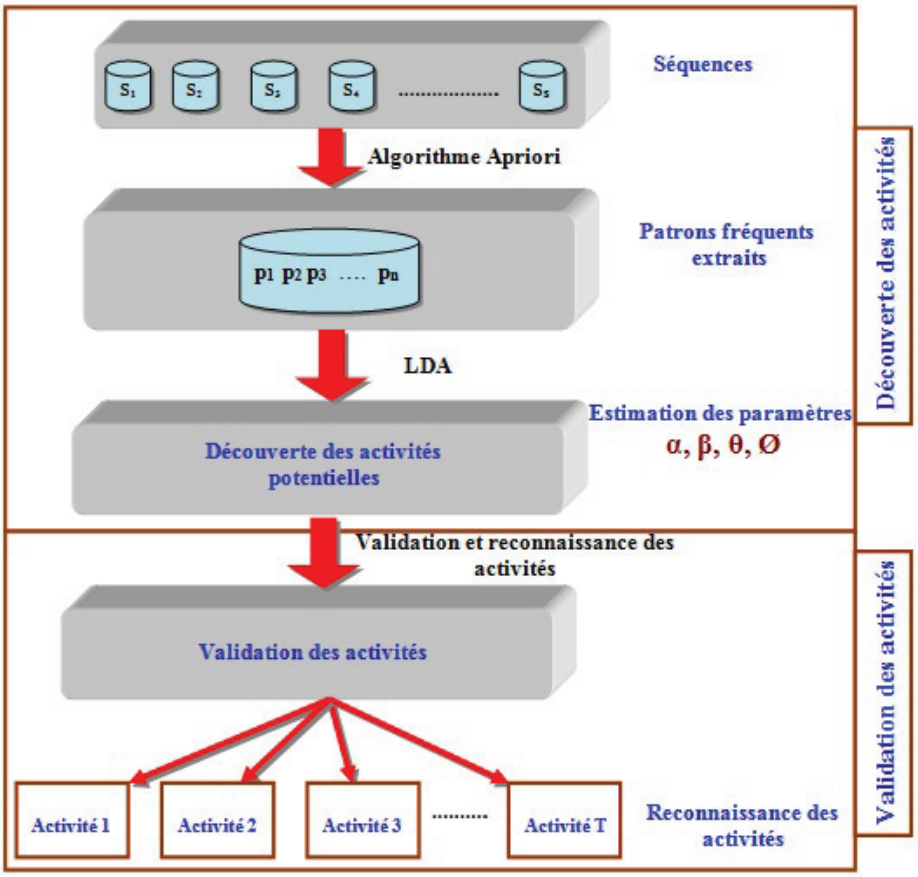


figure 2.5 – Les étapes de notre approche

comme étant des distributions d'activités, et les activités à leur tour sont modélisées comme étant des distributions de patrons fréquents. Cette façon de faire permet de résoudre l'un des grands défis dans l'analyse des séquences qui est la découverte des patrons significatifs [43, 126, 138]. La figure 2.6 présente de façon graphique notre modèle LDA pour la découverte des activités.

Pour cela nous aurons besoin de définir quelques notations. Soit $E = \{p_1, p_2, ..., p_V\}$ un ensemble de V patrons fréquents extraits par l'algorithme Apriori dans l'étape précédente, où p_n représente le $n^{ème}$ patron fréquent. Soit $D = s_1, s_2, ..., s_M$ un ensemble de M séquences dans la base de données. Dans ce qui suit, le terme "activité" désigne une activité potentielle sauf si le contraire est indiqué explicitement. Les notations suivantes seront utilisées dans notre modèle.

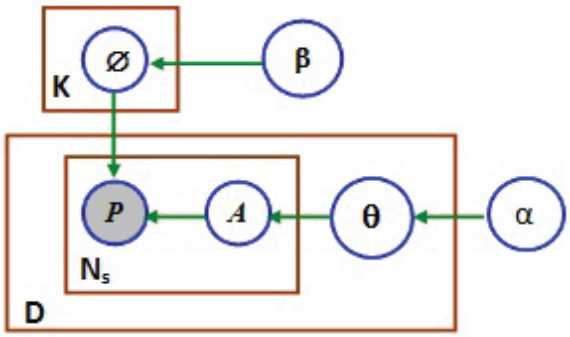


figure 2.6 – Notre modèle LDA

- V est le nombre total de patrons.
- K est le nombre d'activités potentielles.
- α est un paramètre de la distribution Dirichlet a priori de la proportion des activités potentielles dans la séquence (K-vecteur).
- β est un paramètre de la distribution Dirichlet a priori de la proportion des patrons dans une activité (V-vecteur).
- θ_s désigne la distribution des activités potentielles dans la séquence s, noté souvent $p(A|s = m)$. Pour toutes les séquences dans le corpus, on note : $\Theta = \{\theta_s\}_{s=1}^{M}$ une matrice $(M \times K)$.
- ϕ_k désigne la distribution des patrons dans l'activité potentielle k, noté souvent $p(t|A = k)$. Pour toutes les activités, on note : $\Phi = \{\phi_k\}_{k=1}^{K}$ une matrice $(K \times V)$.
- une séquence est composée de N_s patrons notée par $s = (p_1, p_2, ..., p_{N_s})$.
- un corpus est un ensemble de M séquences noté $D = s_1, s_2, ..., s_M$.
- A désigne une activité potentielle.

Comme nous pouvons le constater, notre formalisation est similaire à celle utilisée dans la description du modèle LDA de base. Cependant, le défi majeur de notre approche est qu'il n'y a pas de connaissances apriori sur les patrons fréquents, leur signification, et leur importance pour chaque activité. De plus, ces patrons ne sont pas directement observables, mais nous les découvrons tout d'abord pour qu'ils deviennent utilisables. En outre, nous n'avons aucune information sur les activités et leur proportion dans les séquences. Dès lors, notre approche généralise le modèle LDA pour qu'il soit fonctionnel sur les données séquentielles. Cela constitue une contribution impor-

tante qui mérite d'être mise en évidence. Cette contribution aura des retombées non seulement dans le domaine des habitats intelligents, mais aussi dans tous les domaines caractérisés par des données séquentielles à savoir le domaine de la biologie, les réseaux sociaux, le domaine médical, etc. Par conséquent, les formules seront adaptées selon nos propres notations et contexte. Nous supposons que le nombre d'activités potentielles est connu d'avance. En fait, le nombre d'activités potentielles dans notre cas correspond au nombre de thèmes dans le modèle LDA.

Donc, fixer le nombre d'activités potentielles dans notre modèle correspond à fixer le nombre de clusters dans un algorithme de clustering tel que le k-means. Idéalement, nous voulons que ce nombre soit déterminé de façon automatique. Toutefois, cette étape nécessite un processus d'essai-erreur doté d'un indice de validation comme les travaux présentés dans [7, 16]. Dans notre modèle qui est un modèle non supervisé, le nombre d'activités potentielles est un paramètre d'entrée. Nous allons discuter les méthodes qui peuvent être utilisées pour surmonter ce problème dans les travaux futurs.

Le problème d'optimisation consiste donc à trouver les paramètres α et β des distributions de Dirichlet qui maximisent la vraisemblance des données comme suit :

$$p(D|\alpha,\beta) = \prod_{s=1}^{M} \int\int p(\theta_s|\alpha).p(\Phi|\beta). \prod_{n=1}^{N_s} p(p_{s,n}|\theta_s,\Phi) d\Phi d\theta_s \tag{2.18}$$

Comme nous pouvons le constater, les seuls paramètres observés dans notre modèle sont les patrons fréquents. Tous les autres paramètres sont cachés. Pour estimer les paramètres de l'équation (2.18), nous utilisons l'échantillonnage de Gibbs tel que présenté auparavant. Donc, la probabilité d'une activité potentielle sachant les paramètres observés et cachés peut être calculée de la façon suivante :

$$P(A_i = j|\mathbf{A}_{-i}, D) \propto \frac{n_{-i,j}^{(p_i)} + \beta}{n_{-i,j}^{(.)} + V\beta} \frac{n_{-i,j}^{(s)} + \alpha}{n_{-i,.}^{(s)} + K\alpha} \tag{2.19}$$

où $\mathbf{A}$ désigne l'ensemble des activités et $\mathbf{A}_{-i}$ désigne l'ensemble des activités exceptée la $i^{ème}$ activité. Donc, dans cette formule, nous aurons besoin de quatre variables de comptage :

1. le compteur séquence-activité potentielle $n_{-i,j}^{(s)}$

2. la somme séquence-activité potentielle $n^{(s)}_{-i,\cdot}$
3. le compteur activité potentielle-patron fréquent $n^{(p_i)}_{-i,j}$
4. la somme activité potentielle-patron fréquent $n^{(.)}_{-i,j}$

Une fois **A** connu, nous pouvons estimer les paramètres θ et ϕ comme suit :

$$\phi_{j,p} = \frac{n_p^{(j)} + \beta}{\sum_{p=1}^{V} n_p^{(j)} + V\beta} \tag{2.20}$$

$$\theta_j^{(s)} = \frac{n_j^{(s)} + \alpha}{\sum_{A=1}^{K} n_A^{(s)} + K\alpha} \tag{2.21}$$

Où $n_p^{(j)}$ est le nombre de fois que le patron p est assigné à l'activité potentielle j, et $n_A^{(s)}$ est le nombre de patrons assignés à l'activité potentielle A.

2.5 Validation

Dans cette section, nous allons décrire les expérimentations que nous avons effectuées pour valider notre approche en utilisant des données réelles issues des habitats intelligents réels. Nous effectuons ces expérimentations pour répondre aux questions suivantes :

1. Les activités potentielles découvertes par notre modèle sont-elles consistantes avec les activités réelles ?
2. Les patrons significatifs découverts sont-ils suffisamment expressifs selon les points de vues sémantique et statistique pour représenter les activités ?
3. Le modèle que nous avons proposé est-il efficace comparé aux modèles existants ?

2.5.1 Jeux de données

Avant de présenter les expérimentations, nous devons tout d'abord présenter les données que nous avons utilisées afin de mener nos expérimentations. Nous allons présenter de façon plus détaillée la nature des données, les capteurs utilisés, les activités réalisées, le nombre d'usagers qui ont participé aux expérimentations, ainsi que la

durée des expérimentations. Dans nos expérimentations, nous avons utilisé plusieurs jeux de données provenant de plusieurs habitats intelligents. Notre objectif est de s'assurer que notre modèle fonctionne quelque soit le type de capteurs utilisés et la complexité des activités réalisées. Cela nous permet aussi de valider la généricité de notre modèle. Le tableau 2.1 présente les détails de chaque ensemble de données.

tableau 2.1 – Détails des données utilisées

Données	Nombre de séquences	Longueur min	Longueur max	Activities	Types de capteurs	Nombre des usagers	Période (jours)
Domus Série 1	58	100	470	Wake up, Bathe, Prepare breakfast, Have breakfast	Infrarouge, Électromagnétique, détecteur de pression, interrupteur, contact de portes et cabinets, Débimètre	6	10
Domus Série 2	30	210	680	Wake up, Bathe, Prepare breakfast, Have breakfast, Prepare tea	Les mêmes capteurs que Domus série 1	6	5
CASAS 1	501	19	1561	Wake up, Groom, Breakfast, Watch TV, Prepare dinner, Wash bathtub, Work at computer, Sleep, Prepare lunch, Clean, Work at dining room table	Infrarouge, capteurs RFID, Contact portes, Capteur eau chaude Capteur eau froide, Température, Électricité	2	90
CASAS 2	120	16	216	Make a phone call, Wash hands, Cook, Eat, Clean	Infrarouge, capteurs RFID, Contact portes, Capteur eau chaude Capteur eau froide, Température, Électricité Capteur de téléphone	24	90
ISLab	23	16	140	Idle, Leave house, Use toilet Take shower, Sleep, Breakfast, Dinner, Drink	Capteurs de contacts, Température, Humidité	1	23
StarHome	109	14	943	Take medicine, Use toilet, Wash hands, Wash clothes, Make oatmeal, Brush teeth, Make coffee, Use computer, Prepare drink, Eat meal, Fry eggs, Ironning, Listen music	Capteur RFID	4	15

2.5.2 Les conditions d'expérimentation

Dans cette section, nous allons discuter les conditions sous lesquelles se déroulent nos expérimentations et les hypothèses que nous avons émises.

Pour des raisons de simplicité de notre modèle LDA, nous avons utilisé des paramètres symétriques de la loi Dirichlet. Cela constitue l'hypothèse apriori de notre modèle. La plupart des modèles LDA employés dans la littérature font cette hypothèse afin de simplifier le modèle [38, 45, 37]. Le but d'utiliser cette hypothèse est de donner la même importance à tous les patrons par rapport à chaque activité. De la même manière, nous voulons que toutes les activités aient la même importance pour chaque séquence avant d'observer les données. Typiquement, cette hypothèse est très pratique dans notre modèle étant donné que nous n'avons aucune information apriori sur les patrons et leur importance pour chaque activité. De même, nous n'avons aucune information sur la distribution des activités dans chaque séquence. Par contre, dans certains domaines l'importance des mots pourrait être connue pour un thème particulier. Nous avons utilisé les valeurs symétriques suivantes ($\alpha = 0.1$ et $\beta = 0.01$) obtenus par l'échantillonneur de Gibbs. Le choix de ces valeurs est motivé par les valeurs recommandées dans la littérature [45, 131].

De plus, dans les bases de données de Domus, CASAS et ISLab que nous avons utilisées, les activités sont relativement simples. Il en résulte que l'algorithme aprioi génère des patrons fréquents de longueur petite (2 et 3). Donc, nous avons utilisé des patrons de longueur 2 et 3 dans toutes les expérimentations que nous avons effectuées. En outre, pour valider notre approche avec des patron fréquents de longueur plus grande que 3, nous avons trouvé une autre base de données (StarHome) qui contient des activités relativement complexes et avec lesquelles nous avons pu extraire des patrons de longueur 4 et 5.

2.5.3 Validation des activités potentielles

Pour répondre à la première question des critères d'évaluation, cette section présente la démarche adoptée pour valider les activités potentielles découvertes en utilisant notre modèle. Nous allons aussi présenter comment ces activités validées pourraient être utilisées pour reconnaître les activités.

2.5. Validation

Comme nous l'avons mentionné précédemment, dans notre modèle les activités potentielles sont découvertes de façon non supervisée par le groupement des patrons fréquents. Il en résulte que chaque patron fréquent est lié à une activité potentielle par une relation exprimée sous forme de probabilité. Il existe plusieurs méthodes pour vérifier si les activités potentielles sont des activités réelles. À titre d'exemple, nous pouvons utiliser les noms des objets utilisés dans chaque activité pour faire la correspondance entre les activités potentielles et les activités réelles comme le travail présenté par [136]. Cependant, il n'est pas toujours garanti d'avoir les noms d'objets dans les séquences d'événements. À cet effet, nous proposons une autre démarche pour valider les activités potentielles. Les étapes de notre démarche de validation sont présentées ci-dessous :

- **Étape 1** : Tout d'abord, nous exécutons notre modèle pour découvrir les activités potentielles à partir de la base de données des séquences. Chaque activité potentielle est caractérisée par un ensemble de patron fréquents qui possèdent de plus grandes valeurs de probabilité. Ces patrons sont considérés significatifs pour cette activité potentielle. Ces patrons significatifs sont obtenus en utilisant un seuillage à la base des probabilités comme il sera décrit dans la section suivante.
- **Étape 2** : Nous comparons les patrons significatifs de l'activité potentielle avec les patrons fréquents des séquences annotées. L'activité annotée qui correspond au meilleur score sera identifiée comme étant l'activité réelle de l'activité potentielle. Le score est défini comme suit : $score = \frac{M}{G}$, où M est le nombre de patrons significatifs correctement appariés, et G représente le nombre total des patrons significatifs dans l'activité potentielle. Nous constatons que, plus la valeur du score est proche de 1, meilleur est la correspondance entre l'activité potentielle et l'activité réelle. Le score peut prendre des valeurs dans un intervalle de $[0, 1]$. Nous aurons donc besoin de définir un seuil à partir duquel nous pouvons considérer qu'une activité potentielle correspond à une activité réelle. Nous avons choisi un seuil de 0.5 dans notre travail. Si le score dépasse le seuil minimal alors l'activité potentielle correspond à l'activité réelle, sinon l'activité potentielle ne correspond pas à l'activité réelle.

Notons que dans la démarche que nous avons entreprise, nous avons utilisé des

données annotées. Ces données annotées ne sont utilisées qu'à ce stade à des fins de validation et uniquement de validation. De plus, elles ne sont utilisées que dans la phase de validation des activités potentielles. Mais dans la phase de reconnaissance d'activités nous n'aurons pas besoin de l'annotation des activités. Le tableau 2.2 présente le nombre d'activités potentielles découvertes pour chaque ensemble de données et le nombre d'activités réelles correspondantes.

tableau 2.2 – Nombre d'activités potentielles correspondant aux activités réelles

Données	Nombre d'activités potentielles	Nombre d'activités potentielles qui correspondent aux activités réelles
Domus série 1	7	5
Domus série 2	8	6
CASAS 1	5	5
CASAS 2	5	5
ISLab	8	7
StarHome	15	14

Les activités potentielles qui correspondent aux activités réelles constituent les modèles d'activités qui seront utilisés par la suite pour la reconnaissance d'activités.

2.5.4 Patrons significatifs

Dans le but de répondre à la deuxième question de validation relative aux patrons significatifs, dans cette section nous allons présenter comment les patrons significatifs sont sélectionnés, et comment associer de la sémantique à ces patrons.

Comme nous l'avons mentionné auparavant, chaque patron appartient à une activité potentielle avec une certaine probabilité. Les patrons significatifs qui sont pertinents pour une activité potentielle possèdent des grandes valeurs de probabilité par rapport aux autres patrons. Par conséquent, nous pouvons extraire les patrons significatifs pour chaque activité en sélectionnant ceux qui possèdent des valeurs de probabilités qui dépassent un certain seuil δ. Un patron p_i est significatif pour une activité potentielle A si $P(p_i|A) \geq \delta$. Le seuil δ est : soit spécifié par l'utilisateur et sera considéré comme un paramètre du modèle, soit calculé en utilisant une formule

de moyenne comme suit :

$$\delta = \frac{\sum_K \frac{1}{N} \sum_{i=1}^{N} Prob(p_i|A)}{K}, \tag{2.22}$$

où N représente le nombre de patrons fréquents et K représente le nombre d'activités potentielles.

Avec un seuil de faible amplitude, beaucoup de patrons significatifs risquent d'être recueillis pour chaque activité potentielle. Le tableau 2.3 présente un exemple de patrons significatifs détectés pour quelques activités dans la base de données de StarHome.

tableau 2.3 – Exemple de patrons significatifs pour quelques activités dans la base de données StarHome

Base de données StarHome			
Brosser les dents	Probabilités	Laver les mains	Probabilités
Toothbrush Toothbrush	0.565504241	Water Water	0.43625974
Toothbrush-cup Toothbrush-cup	0.103450395	Toilet-door Water	0.083116883
Water Water	0.058435438	Towel Water	0.061766234
Water Toothbrush	0.05122321	Water Toilet-door	0.041558442
Toothbrush Water	0.049174282	Toilet-door Toilet-door	0.037558442
Toothbrush-cup Water	0.030733926	Toilet-door Toothbrush-cup	0.025974026
Water Toothbrush-cup	0.030651969	Toilet doorToothbrush	0.025766234
Mouthwash Mouthwash	0.023419252	Soap Toilet-door	0.01974026

Nous pouvons constater, à partir du tableau 2.3, que les patrons significatifs possèdent des plus grandes valeurs de probabilités. Par exemple, pour l'activité "Brosser les dents", le patron "Toothbrush Toothbrush" se distingue avec une probabilité de 0.565504241. De la même façon, le patron "Water Water" a une probabilité de 0.43625974 ce qui lui permet d'être le patron le plus significatif pour l'activité "Laver les mains". Ce sont ces patrons significatifs qui vont désigner les modèles d'activités. En effet, l'ensemble des patrons significatifs pour chaque activité permet de construire un modèle pour cette activité étant donné que chaque patron significatif représente une tâche particulière. Le groupement de ces patrons constitue l'activité tout entière.

Dans cette section nous avons discuté l'aspect statistique des patrons significatifs pour chaque activité. Cependant, l'aspect statistique pris séparément n'est pas suffisant. Il aura plus d'importance si nous lui incorporons un aspect sémantique des patrons. Découvrir la sémantique des patrons est un défi majeur non seulement dans le domaine des habitats intelligents, plus spécifiquement dans le problème de recon-

naissance d'activités, mais aussi un défi dans le domaine du forage des données. Dans notre travail nous proposons deux méthodes différentes pour extraire la sémantique des patrons.

- La première méthode consiste à analyser les patrons en regardant les noms des objets contenus dans chaque patron. Toutefois, cette méthode n'est applicable uniquement si les informations sur les objets sont disponibles. Ces informations sont généralement obtenues en utilisant des capteurs de type RFID qui sont placés sur tout objet utilisé dans les expérimentations. Par exemple, dans le tableau 2.3, le patron "Toothbrush Toothbrush" contient le nom de l'objet "Toothbrush" qui signifie que l'usager a pris l'objet "brosse à dent" lié à l'activité "Brosser les dents". Donc, à partir de ces objets nous pouvons extraire la sémantique des objets afin de rendre les patrons compréhensibles et interprétables.
- La deuxième méthode représente une alternative de la première dans le cas où les noms des objets ne sont pas disponibles dans la base de données. Cette méthode propose d'exploiter les positions des capteurs dans l'habitat intelligent. En effet, chaque capteur est caractérisé par sa position dans l'habitat intelligent. Cette information est très importante pour pouvoir inférer la sémantique des objets. Par exemple, si le capteur qui est installé sur la porte de la salle de bain est déclenché, nous pouvons déduire que l'usager est en train de faire une activité liée à la salle de bain, et toutes les autres activités seront écartées comme celles de la cuisine par exemple. Cette méthode permet aussi de suivre les déplacements de l'usager dans l'habitat intelligent en suivant les déclenchements des capteurs et leurs positions.

Dans cette section nous avons discuté comment les patrons significatifs sont extraits et comment associer la sémantique aux patrons. Pour répondre à cette dernière question, nous avons proposé deux méthodes différentes d'extraction de la sémantique des patrons. Avec ces méthodes, nous avons pu proposer des solutions pratiques pour l'énigme de la sémantique des patrons dans le domaine du forage des données.

2.5.5 Reconnaissance des activités

Dans cette section, nous allons valider notre modèle en terme de reconnaissance d'activités. Notons qu'avec la petite quantité des données disponibles, nous allons utiliser une technique du forage de données qui est la validation croisée [67] pour faire face au problème de manque des données. Les petites quantités de données ne nous permettent pas de créer deux ensembles différents représentatifs des données pour faire l'apprentissage et le test. Dans la validation croisée, la base de données est divisée en k échantillons. Nous choisissons un des k échantillons comme un ensemble de test et les autres $(k-1)$ pour faire l'apprentissage. Cette étape est répétée pour toutes les parties dans la base de données. Le résultat global sera la moyenne des résultats partiels obtenus dans chaque étape. De cette façon, nous pouvons garantir que chaque échantillon ait été utilisé une seule fois comme ensemble de test. Dans nos expérimentations, nous avons utilisé la technique de validation croisée "leave one out". Dans cette technique, chaque ensemble de test contient un jour d'expérimentation, les autres jours seront utilisés pour l'apprentissage.

Notre algorithme de reconnaissance d'activités peut être résumé dans les étapes suivantes :

- **Étape 1** : construire les modèles d'activités à partir des ensembles d'apprentissage tel que présenté précédemment dans la section 2.4.5, puis valider ces modèles comme discuté dans la section 2.5.3.
- **Étape 2** : sélectionner les patrons significatifs pour chaque activité en utilisant le seuil calculé automatiquement ou choisi par l'utilisateur comme nous l'avons discuté auparavant dans la section 2.5.4.
- **Étape 3** : comparer les patrons significatifs de chaque activité avec les patrons fréquents extraits à partir de l'ensemble de test. Le processus de la reconnaissance d'activités consiste à chercher les patrons des activités validées dans l'ensemble de test.

Nous avons utilisé le concept de précision pour mesurer la performance de notre modèle. Pour reconnaître une activité, les patrons de chaque modèle d'activité sont comparés avec les patrons dans l'ensemble de test, et l'activité est sélectionnée si le nombre de comparaisons correctes dépasse le seuil de reconnaissance (0.5) que nous avons défini pour que l'activité soit reconnue. Pour s'assurer que notre reconnaissance

est valide, nous les comparons avec les données annotées. La précision est calculée comme le rapport entre le nombre de reconnaissances correctes et le nombre total de reconnaissances. Les résultats de la reconnaissance d'activités dans toutes les bases de données sont présentés dans le tableau 2.4. Les figures 2.7 et 2.8 présentent graphiquement les résultats de la reconnaissance en utilisant respectivement des patrons de longueur 2 et 3.

Nous pouvons constater à partir de ces résultats, que la reconnaissance d'activités en utilisant des patrons de longueur 2 est relativement meilleure par rapport aux patrons de longueur 3. La seule exception se trouve dans la base de données ISLAB où les résultats sont très proches (80.67% et 81.78%) respectivement pour des patrons de longueur 2 et 3. Cela peut être exprimé par le fait que les activités utilisées dans les bases de données sont des activités relativement simples. Ce qui signifie que certaines activités sont assez courtes que nous ne puissions pas générer des patrons de longueur plus grande que 2. Par exemple, l'activité "Wake Up" (se lever) ne requiert pas dans la pratique beaucoup de capteurs pour qu'elle soit détectée. Un capteur de pression sur le lit et un capteur de contact sur la porte de la chambre à coucher suffiront pour détecter cette activité et la reconnaître. Dans ce cas de figure, nous pouvons générer un patron de longueur 2 par exemple "lit porte-chambre", mais probablement nous ne serons pas capables de générer des patrons de longueur plus grand que 2. Cela démontre les résultats obtenus dans nos expérimentations. Comme nous pouvons le constater dans le tableau 2.4, le taux de reconnaissance de l'activité "Wake Up" dans la base de données de Domus série 1 et 2 est de 100% avec des patrons de longueur 2, mais uniquement 55.8% avec des patrons de longueur 3. La même observation dans la base de données Domus série 2, où le taux diminue de 60% à 23.1%.

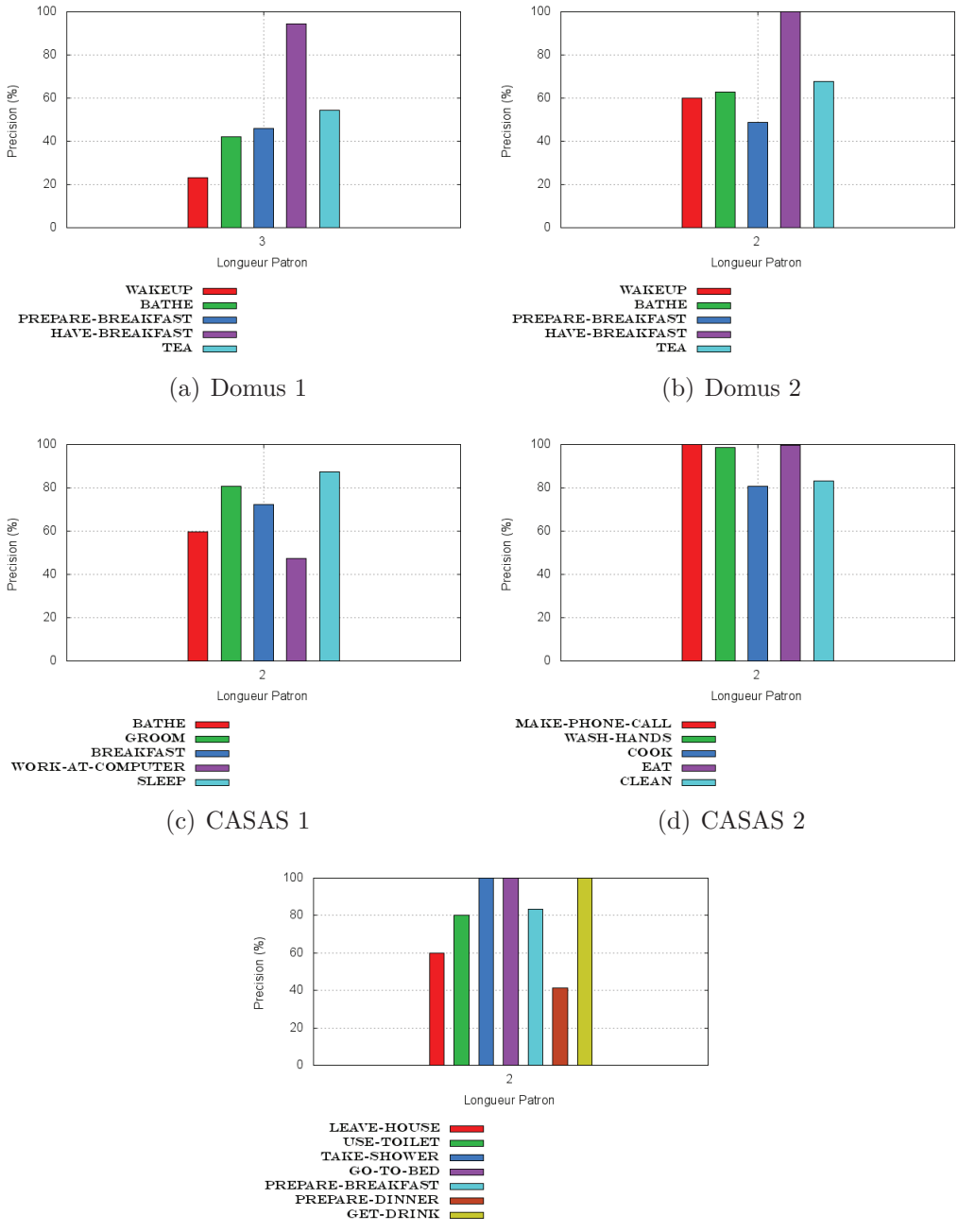
Par ailleurs, nous pouvons aussi constater que dans le cas des activités relativement complexes, i.e. des activités qui entraînent l'incorporation de plusieurs objets dans leur réalisation, la reconnaissance en utilisant des patrons de longueur 3 est généralement meilleure comparativement à celle basée sur des patrons de longueur 2. Comme présenté dans le tableau 2.4, les activités "Bathe", "Groom", et "Work at computer" dans la base de données CASAS 1 et qui sont relativement complexes, la reconnaissance en utilisant des patrons de longueur 3 est nettement meilleure que dans le cas de la reconnaissance avec des patrons de longueur 2 (59.78% vs 70.6%), (80.59%

tableau 2.4 – Résultats de la reconnaissance obtenus par notre modèle. PL : Patron de Longueur

Activité Domus Série 1	Précision (%) PL = 2	PL = 3	Activité Domus Série 2	Précision (%) PL = 2	PL = 3
Wake Up	100	55.8	Wake Up	60	23.1
Bathe	100	56.1	Bathe	62.72	42.2
Prepare-breakfast	54.04	91.4	Prepare-breakfast	48.6	46.1
Have-breakfast	50	41	Have-breakfast	100	94.5
			Prepare-tea	67.56	54.4
Moyenne	76.01	61.07	Moyenne	67.77	52.06

Activité CASAS 1	Précision (%) PL = 2	PL = 3	Activité CASAS 2	Précision (%) PL = 2	PL = 3
Sleep	87.48	70	Make a phone call	100	62.8
Bathe	59.78	70.6	Wash hands	98.63	79
Groom	80.59	100	Cook	80.85	76.2
Breakfast	72.41	30	Eat	99.82	88.3
Work at computer	47.47	100	Clean	83.19	90.9
Moyenne	69.54	74.12	Moyenne	92.49	79.44

Activité ISLab	Précision (%) PL = 2	PL = 3
Leave-house	60	100
Use-toilet	80	100
Take-shower	100	100
Got-to-bed	100	22.5
Prepare-breakfast	83.33	50
Prepare-dinner	41.42	100
Get-drink	100	100
Moyenne	80.67	81.78



(a) Domus 1

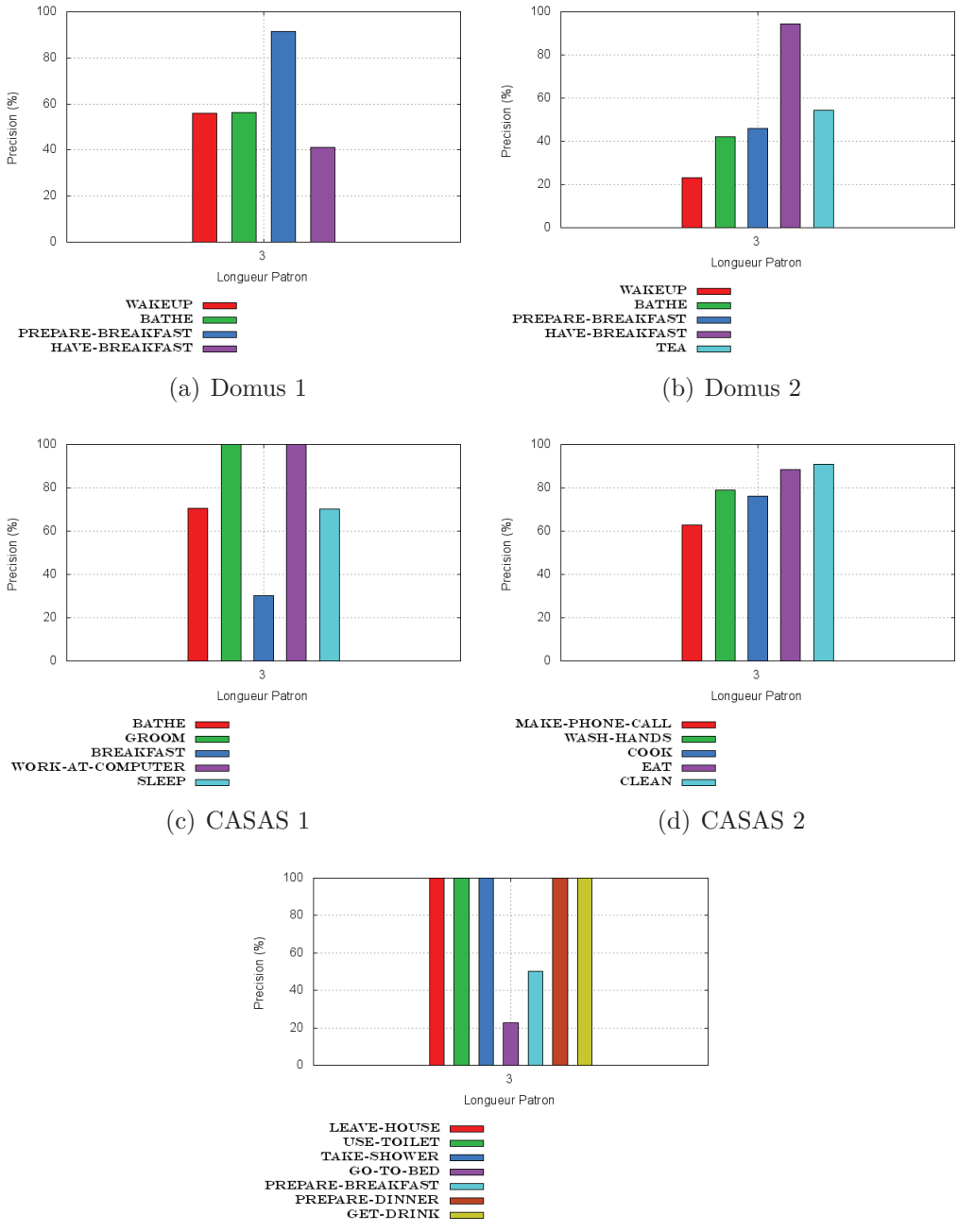
(b) Domus 2

(c) CASAS 1

(d) CASAS 2

(e) ISLAB

figure 2.7 – Résultats de reconnaissance pour chaque activité dans toutes les bases de données avec patrons de longueur 2



(a) Domus 1

(b) Domus 2

(c) CASAS 1

(d) CASAS 2

(e) ISLAB

figure 2.8 – Résultats de reconnaissance pour chaque activité dans toutes les bases de données avec patrons de longueur 3

vs 100%) et (47.47% vs 100%) respectivement. La même interprétation pourrait être généralisée pour les bases de données CASAS 2 et ISLAB.

Un autre point important qui doit être élaboré, et qui pourrait influencer les résultats de la reconnaissance, est le problème de partage des objets entre activités. Par exemple, les activités "Brosser les dents" et "Laver les mains" présentées dans le tableau 2.3 partagent le patron "Water Water". Ce qui veut dire que le patron "Water Water" est significatif dans deux activités différentes. Bien que le patron soit significatif dans les deux activités, il possède des probabilités d'appartenance différentes pour chaque activité. Par exemple, la probabilité du patron "Water Water" dans l'activité "Laver les mains" est 0.43625974, où il est considéré comme le patron le plus significatif. Mais le même patron possède une probabilité nettement inférieure dans l'activité "Brosser les dents" qui est de 0.058435438 pour laquelle le patron est considéré comme le troisième patron significatif pour l'activité. Ce genre de partage pourrait influencer les résultats de reconnaissance en sélectionnant ainsi la mauvaise activité au lieu de la bonne.

Pour valider la généricité de notre approche et son fonctionnement avec des patrons de longueur plus grand que 3, nous avons utilisé une autre base de données appelée "StarHome" qui contient des activités relativement complexes dont nous avons pu extraire des patrons de longueur 2,3,4 et 5. Les résultats de la reconnaissance en utilisant la base de données StarHome sont présentés dans le tableau 2.5.

Comme indiqué dans le tableau 2.5, notre approche donne des meilleurs résultats aussi bien pour des patrons de longueur 2 et 3 qu'avec des patrons de longueur plus grand. Cela expliquerait bien la performance de notre approche dans la reconnaissance des activités complexes. En outre, nous observons que la reconnaissance en utilisant des patrons de longueur 5 (84.9%) est moins bonne que celle basée sur les patrons de longueur 2, 3 et 4 qui est 85.9%, 91.45% et 92.62% respectivement. Cela pourrait être expliqué par le fait que certaines activités ne sont pas assez longues pour que nous puissions générer des patrons de longueur plus grande que 4. C'est pour cette raison que le taux de reconnaissance est moins significatif que les autres. Nous observons aussi que le taux de reconnaissance avec des patrons de longueur 4 est meilleur que celui avec des patrons de longueur 3, qui est lui même meilleur que celui avec des patrons de longueur 2.

tableau 2.5 – Résultats de reconnaissance pour chaque activité dans la base de données StarHome

Activité	**Précision (%)**			
	PL = 2	PL = 3	PL = 4	PL = 5
Take medicine	100	100	100	100
Use toilet	100	100	100	100
Wash clothes	100	100	100	100
Wash hands	30.76	100	100	100
Make oatmeal	100	83.38	97	55.07
Brush teeth	37.5	51.42	8.31	50
Make coffee	98.64	99.8	100	100
Use computer	100	97.9	98.8	100
Make drink	99.9	56.4	100	49.9
Eat meal	100	100	100	50
Fry eggs	100	100	100	100
Do ironing	100	100	100	100
Listen to music	50	100	100	100
Moyenne	85.9	91.45	92.62	84.9

2.5.6 Comparaison avec les méthodes existantes

Dans cette section, nous allons comparer notre approche avec les méthodes les plus connues dans la littérature pour la reconnaissance d'activités. Notre objectif est de positionner notre approche par rapport à des méthodes existantes.

Puisque notre approche est basée sur un modèle probabiliste, nous avons choisi les deux méthodes probabilistes les plus connues dans la littérature que sont les chaînes de Markov cachées (HMM) et le modèle CRF (Conditional Random Fields) [128]. Nous avons aussi adapté notre approche pour qu'elle soit opérationnelle directement sur des séquences d'événements. Nous avons appelé ce modèle "LDA-Événement". Cela veut dire que dans ce cas de figure, les événements dans les séquences sont considérés comme étant des patrons de longueur 1. Dans ce cas, il n'est pas nécessaire d'extraire des patrons fréquents. Donc la seule différence entre notre modèle de reconnaissance et notre modèle basé sur les événements réside dans la longueur des patrons. Avant de présenter les résultats de la comparaison, nous allons tout d'abord présenter les résultats de la reconnaissance en utilisant notre modèle basé sur les événements. Le tableau 2.6 présente les résultats de la reconnaissance d'activités en utilisant notre

modèle basé sur les événements dans les bases de données Domus série 1, Domus série 2, CASAS 1, CASAS 2, ISLAB. La figure 2.9 présente sous un format graphique

tableau 2.6 – Les résultats de reconnaissance dans chaque base de données avec le modèle LDA-Événement

Activité Domus Série 1	Précision (%) LDA-Événement	Activité Domus Série 2	Précision (%) LDA-Événement
Wake Up	36.09	Wake Up	46.24
Bathe	98.58	Bathe	44.77
Prepare-breakfast	54.7	Prepare-breakfast	59.22
Have-breakfast	28.89	Have-breakfast	25.79
		Prepare-tea	6.34
Moyenne	54.56	Moyenne	36,47

Activité CASAS 2	Précision (%) LDA-Événement	Activité CASAS 1	Précision (%) LDA-Événement
Sleep	16.55	Make a phone call	14.13
Bathe	16.52	Wash hands	4.13
Groom	20	Cook	77.93
Breakfast	37.33	Eat	3.79
Work at computer	9.58	Clean	0
Moyenne	20	Moyenne	24.99

Activité ISLab	Précision (%) LDA-Événement
Leave-house	100
Use-toilet	95
Take-shower	100
Got-to-bed	40
Prepare-breakfast	37.5
Prepare-dinner	47.5
Get-drink	42.85
Moyenne	66.12

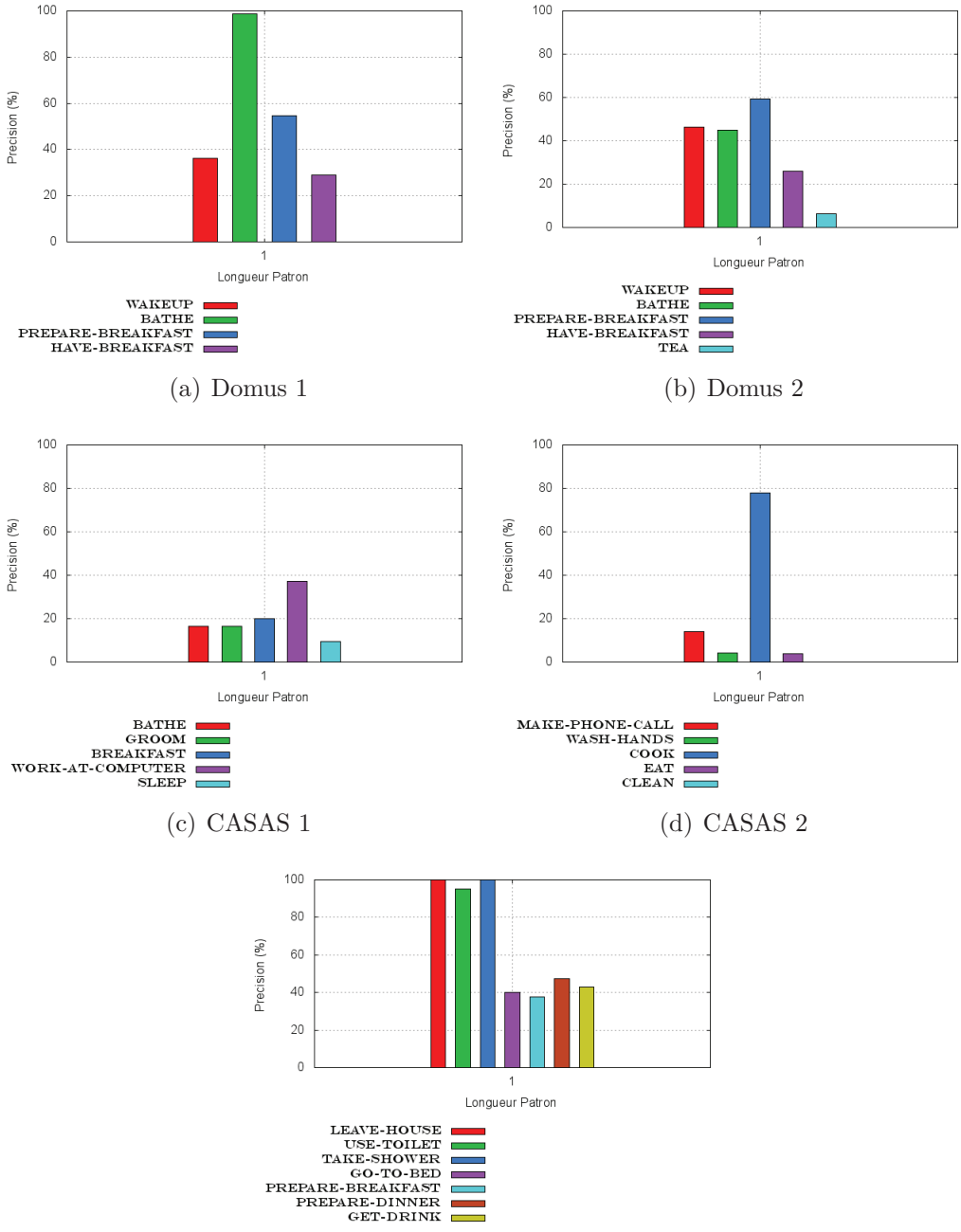
les résultats de reconnaissance obtenus pour chaque activité dans chaque base de

données. Nous avons également effectué des expérimentations avec la base de données StarHome. Le tableau 2.7 présente les résultats de reconnaissance pour chaque activité dans la base de données StarHome en utilisant le modèle LDA-Événement.

tableau 2.7 – Les résultats de reconnaissance dans la base de données StarHome avec le modèle LDA-Événement

Activité	Précision (%) LDA-Événement	Activité	Précision (%) LDA-Événement
Take medicine	50	Make coffee	86.34
Use toilet	97.3	Use computer	59.05
Wash clothes	100	Make drink	93.83
Wash hands	7.77	Eat meal	82.24
Make oatmeal	93.17	Fry eggs	98.87
Brush teeth	96.11	Make ironing	50
		Listen music	99.2
Moyenne			77.99

Afin de comparer objectivement les résultats de notre modèle avec les méthodes existantes, nous avons réalisé trois expérimentations différentes avec les modèles HMM et CRF. Typiquement, les modèles HMM et CRF utilisent des séquences d'événements pour l'apprentissage. Pour cela, nous avons effectué des expérimentations en utilisant directement les séquences d'événements, i.e. les modèles HMM et CRF sont entraînés sur des séquences d'événements. Dans la deuxième expérimentation, nous avons entraîné les modèles HMM et CRF sur des patrons de longueur 2, et finalement dans la troisième expérimentation nous avons entraîné les modèles HMM et CRF sur des patrons de longueur 3. Toutes ces expérimentations permettent à ces approches d'être entraînées sur les mêmes données. Pour les modèles HMM et CRF, nous avons utilisé les outils Matlab [128]. Le tableau 2.8 présente les résultats de la comparaison de notre approche avec les autres approches existantes. La figure 2.10 présente les mêmes résultats dans un format graphique.



(a) Domus 1

(b) Domus 2

(c) CASAS 1

(d) CASAS 2

(e) ISLAB

figure 2.9 – Résultats de reconnaissance pour chaque activité dans toutes les bases de données avec le modèle LDA-Événement

tableau 2.8 – Comparaison des résultats de reconnaissance de tous les modèles pour chaque base de données

Bases de données	Précision (%)								
	Notre modèle		LDA-	HMM			CRF		
	PL = 2	PL = 3	Événement	Événement	PL = 2	PL = 3	Événement	PL = 2	SL = 3
Domus série 1	76.01	61.07	54.56	56.8	60.79	60.05	63.8	59.86	60.84
Domus série 2	67.77	52.06	36.47	54	64	60.83	55.4	63.01	64.42
CASAS 1	69.54	74.12	20	29.99	52.59	33.11	16.86	67.54	52.29
CASAS 2	92.49	79.44	24.99	20	24	20	40	12	52.38
ISLab	80.67	81.78	66.12	66.7	55.11	44.58	60.07	43.62	42.41

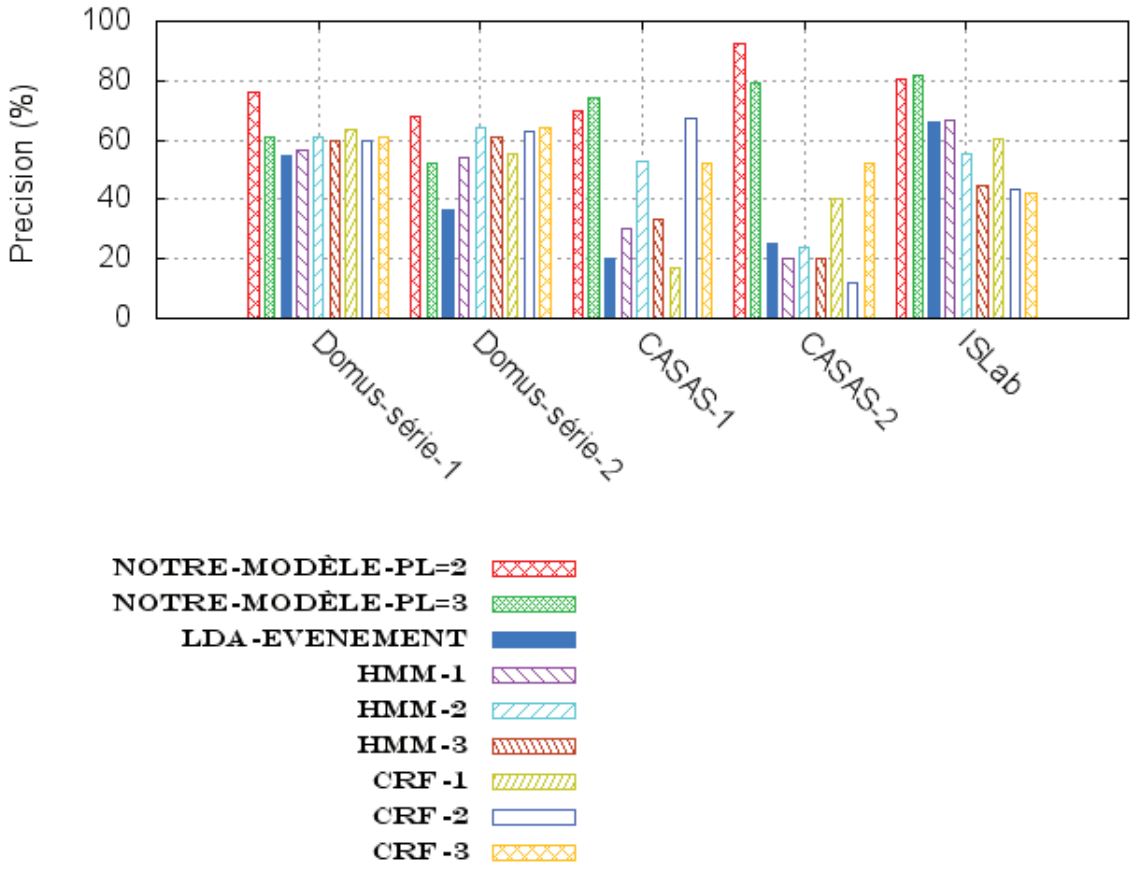


figure 2.10 – Comparaison des résultats de reconnaissance de tous les modèles pour chaque base de données

Les résultats de comparaison entre les différents modèles pour la base de données StarHome sont présentés dans le tableau 2.9. La figure 2.11 présente les mêmes résultats dans un format graphique.

tableau 2.9 – Comparaison des résultats de reconnaissance de tous les modèles pour la base de données StarHome

Base de données	Précision (%)						
	Notre modèle				LDA-Événement	HMM	CRF
	PL = 2	PL = 3	PL = 4	PL = 5			
StarHome	85.9	91.45	92.62	84.9	54.56	56.8	63.8

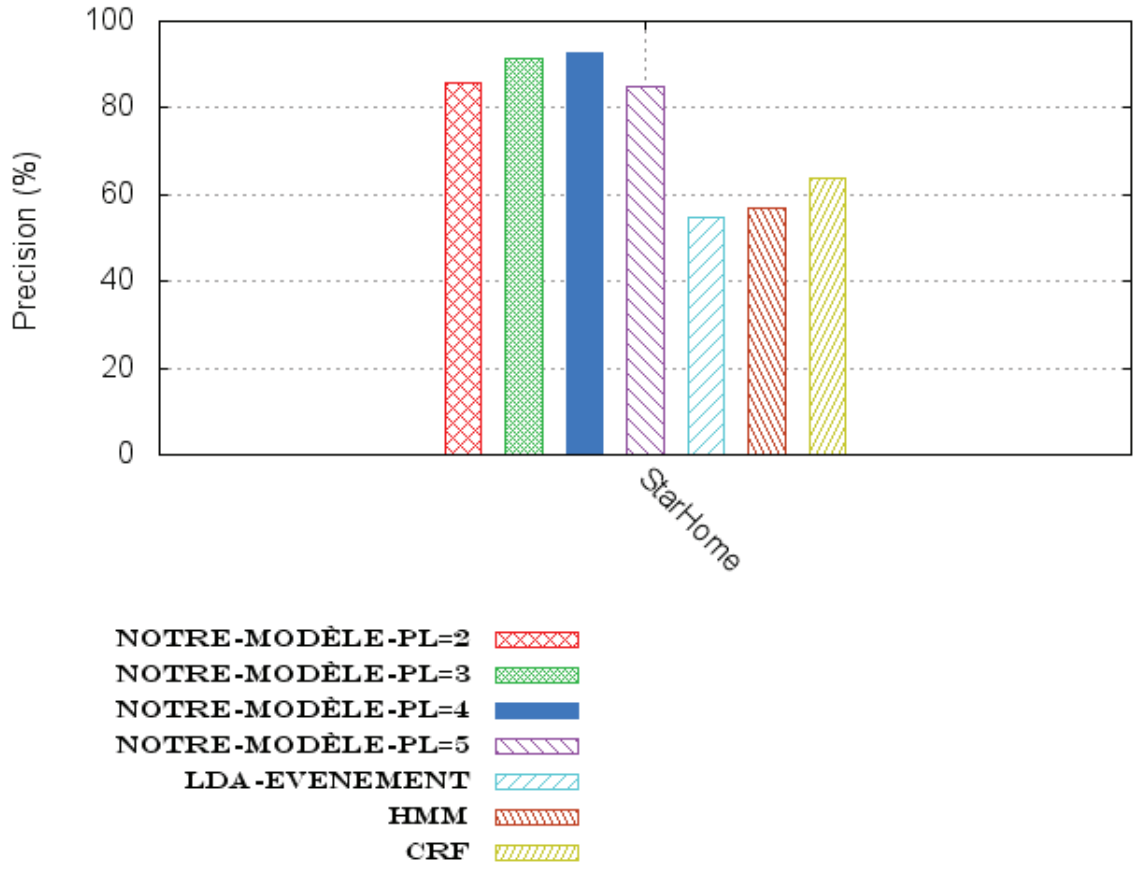


figure 2.11 – Comparaison des résultats de reconnaissance de tous les modèles pour la base de données StarHome

Les tableaux 2.8 et 2.9 montrent que les résultats obtenus avec notre modèle sont meilleurs en comparaison avec les résultats obtenus avec les autres modèles, et cela pratiquement dans toutes les bases de données à l'exception de la base de données Domus série 2 où les modèles HMM et CRF donnent des résultats meilleurs que les résultats de notre modèle avec des patrons de longueur 3. La même observation est valide pour le modèle CRF dans la base de données Domus série 1. Cela pourrait être expliqué par la petite quantité de données utilisées dans ces bases de données (5 séquences par usager dans le Domus série 1, et 10 séquences par usager dans Domus série 2).

Comme mentionné dans les tableaux 2.8 et 2.9, notre modèle appliqué directement

aux événements ne donne pas des résultats assez satisfaisants en comparaison avec les autres modèles plus spécifiquement avec notre modèle. La seule exception est dans la base de données StarHome où notre modèle basé sur les événements donne des bons résultats par rapport aux modèles HMM et CRF. En effet, dans notre modèle appliqué directement aux événements il n'y a pas de relation d'ordre entre les événements, ce qui n'est pas le cas avec les patrons de longueur 2 et plus où l'ordre entre les événements est important. De plus, les événements peuvent véhiculer une quantité du bruit ce qui pourrait influencer sur les performances du processus de la reconnaissance et par conséquent sur les résultats.

En ce qui concerne les modèles HMM et CRF, nous constatons que les résultats de façon générale changent selon le format des données utilisé. Il existe des résultats qui sont améliorés avec le changement dans le format de données en passant des patrons de longueur 1 et 2 jusqu'à 3. Par exemple, dans le tableau 2.8, les résultats du modèle HMM sont grimpés de 56.8% à 60.79% à 60.05% dans la base de données Domus série 1. De façon similaire, les résultats sont passés de 54% à 64% puis diminués à 60.83% avec des patrons de longueur 3 dans la base de données domus série 2. La même observation pourrait être faite pour le modèle CRF. Contrairement à ces observations, les résultats des deux modèles HMM et CRF sont baissés dans la base de données StarHome et ils passent de 66.7% à 55.11% à 44.58% pour le modèle HMM, et ils passent de 60.07% à 43.62% à 42.41% pour le modèle CRF. Globalement, notre modèle est plus performant que les autres modèles.

2.6 Discussion

Dans cette partie, nous avons concrétisé notre modèle proposé par des expérimentations et comparaisons avec les méthodes existantes. Nous avons analysé les résultats obtenus selon différents points de vues. Nous nous sommes concentrés particulièrement aux interprétations statistiques des relations entres les patrons et les activités, et des interprétations sémantiques des patrons eux mêmes extraits à partir des bases de données. Nous avons évalué la performance de notre modèle en utilisant différentes bases de données avec des activités de différents niveaux de complexité. Nous avons également utilisé des patrons fréquents de différentes longueurs variant de 1 à 5 selon

les bases de données disponibles.

De façon générale, notre modèle est plus performant que les autres modèles comme nous l'avons démontré expérimentalement dans cette partie. En outre, nous avons pu constater que les résultats de notre modèle avec des patrons de longueur 2 sont relativement meilleurs que ceux obtenus avec des patrons de longueur 3. Il serait intéressant d'incorporer un nouveau mécanisme permettant de sélectionner automatiquement la longueur optimale des patrons fréquents. Avec un tel mécanisme, nous pouvons effectuer une seule expérimentation pour valider notre modèle en utilisant la longueur optimale, ce qui nous permet de réduire significativement le temps des expérimentations. Notons que dans la littérature, tous les algorithmes de recherche de patrons fréquents génèrent des patrons de longueurs différentes. Mais ils ne spécifient pas la longueur optimale pour laquelle les patrons ont une meilleure représentation des données. À cet effet, l'une des étapes importantes à incorporer dans notre modèle est l'étape de sélection de la longueur optimale des patrons. Cela constitue l'une des étapes de notre travail futur que nous voudrions envisager.

La distribution Dirichlet utilisée par le modèle LDA, possède plusieurs avantages. Entre autres, elle est la distribution conjuguée de la distribution multinomiale, ce qui permet de faciliter l'étape d'estimation des paramètres [11]. Cependant, nous devons porter une attention particulière pour le choix de ces paramètres. Ce choix doit prendre en considération les cas extrêmes. Par exemple, plus α augmente, plus un patron sera associé à toutes les activités. Mais, si α est trop petit, un très petit nombre d'activités ou probablement aucune activité ne sera associée à un patron fréquent. Cette explication est valide aussi pour le paramètre β. Une grande valeur de β signifie que plusieurs patrons fréquents seront associés à la même activité, et vice versa. Cela montre l'influence de ces paramètres sur les performances du système de reconnaissance et le besoin de définir ces paramètres de façon automatique.

Parmi les points les plus importants dans l'analyse des séquences, nous trouverons le problème de comparaison des séquences. Dans le domaine des habitats intelligents, comparer les séquences permet de suivre les activités des résidents, de détecter les changements dans la réalisation des activités, et même de prédire les situations inhabituelles où les résidents oublient de faire certaines activités, ce qui permet aux intervenants d'intervenir en cas d'urgence. Comparer les séquences à la base d'évé-

nements n'est pas pratique et peut causer des erreurs pour la simple raison que les événements pris séparément ne donnent aucune information sur la structure des séquences, et peuvent être bruités et incertains. Par conséquent, notre approche permet de comparer les séquences en se basant sur les activités qu'elles contiennent. De cette façon, nous pouvons surmonter le problème du bruit d'un coté, et de tirer profit de l'aspect structurel des activités pour comparer les séquences. En effet, les séquences sont similaires si elles partagent les mêmes activités. Par conséquent, la similarité entre deux séquences s_1 et s_2 peut être calculée en calculant la similarité entre les distributions des activités dans chaque séquence θ^{s_1} et θ^{s_2}. La similarité entre deux distributions de probabilités peut être calculée en utilisant la divergence de Kullback Leibler (KL) [66] de la façon suivante :

$$D_{KL}(\theta^{s_1}, \theta^{s_2}) = \sum_{j=1}^{K} \theta_j^{s_1} \log_2 \frac{\theta_j^{s_1}}{\theta_j^{s_2}}. \tag{2.23}$$

Les séquences sont beaucoup plus similaires si la divergence de Kullback Leibler est petite, et inversement.

Un autre point important qui mérite d'être évoqué à ce stade est la comparaison théorique entre notre modèle et les modèles existants. Les modèles Markoviens utilisés dans la littérature pour la reconnaissance d'activités sont généralement des modèles d'ordre 1. Ils supposent que l'état courant ne dépend que de l'état précédent. Cela veut dire, l'utilisation directe des séquences d'événements ou des patrons de longueur 1. Par ailleurs, notre modèle basé sur les patrons fréquents de longueur 2 peut être considéré comme un modèle Markovien d'ordre 1, étant donné que le patron est composé uniquement de deux événements. De la même façon, notre modèle basé sur les patrons fréquents de longueur 3 peut être considéré comme un modèle Markovien d'ordre 2, et ainsi de suite. Cependant, notre modèle est performant quelque soit la longueur du patron comme nous l'avons présenté auparavant. Mais ce n'est pas le cas pour les modèles Markoviens d'ordre supérieur à 2 qui sont très complexes et moins performants. Cela est prouvé dans le travail de [34] où la complexité des modèles Markoviens augmente avec l'augmentation de l'ordre du modèle. Cela est tout à fait normal vu que le nombre de paramètres dans les modèles Markoviens augmente avec l'augmentation de l'ordre du modèle. Cela n'est pas le cas dans notre approche où

nous pouvons utilisé des patrons de longueurs différentes avec le même nombre de paramètres.

2.7 Conclusion

Nous avons présenté, dans ce chapitre, les détails de notre approche de reconnaissance d'activités, ainsi que la validation auprès de différentes données issues des habitats intelligents. Pour ce faire, nous avons tout d'abord présenté le contexte général de notre travail, ainsi que les différents aspects théoriques liés à notre travail à savoir le principe de forage de patrons fréquents, la notion d'activité et le modèle théorique de LDA. De cette façon, nous avons pu regrouper les éléments essentiels à la mise en oeuvre de notre approche aussi bien sur l'aspect théorique que sur l'aspect pratique. Pour clarifier notre approche, nous avons présenté les algorithmes de notre approche de façon mathématique et graphique.

Nous avons ensuite présenté les détails concernant la phase d'expérimentation. Nous avons débuté cette section par la présentation de notre démarche expérimentale, en spécifiant les critères d'évaluation, les objectifs visés au départ, les bases de données utilisées ainsi que les conditions d'expérimentation. Nous avons effectué trois types d'expérimentations pour répondre aux questions posées au départ. Une première expérimentation vise à tester les performances de notre approche dans la reconnaissance des différents types d'activités. La deuxième expérimentation vise à tester la capacité de notre approche à extraire des patrons significatifs pour chaque activité. Finalement, la troisième expérimentation vise à comparer notre approche avec les approches les plus connues dans la littérature. À la lumière de ces résultats, nous avons pu mettre en évidence les forces de notre modèle ainsi que les travaux futurs que nous envisagions. Nous avons aussi identifié quelques solutions permettant de guider l'amélioration future de notre modèle.

Chapitre 3

Construction du profil usager en utilisant l'analyse causale

3.1 Introduction

L'étude de la personne humaine constitue l'une des premières préoccupations des chercheurs dans les domaines des sciences humaines et psychologie. Vu la diversité des contextes de vie des personnes et la divergence de leurs comportements, l'étude de la personne humaine dans de telles conditions constitue un grand défi.

L'étude de la personne humaine a pris une nouvelle orientation avec le développement technologique. Plus spécifiquement, avec l'apparition des ordinateurs, les chercheurs ont créé un nouveau domaine de recherche qui vise l'étude de l'interaction homme-machine. L'objectif fondamental de ce nouveau domaine est d'étudier l'interaction de la personne avec la machine afin de faciliter l'utilisation de ces machines, et d'améliorer leur conception pour qu'elles soient adaptables aux personnes. Ce domaine de recherche a rapidement progressé en commençant tout d'abord par l'étude de l'interaction physique [12], en passant par la perception visuelle [15] jusqu'à arriver à l'étude des processus cognitifs impliqués lors de l'interaction homme-machine [25].

L'étude du profil de la personne, est une analyse détaillée des différents comportements permettant d'en savoir davantage sur la personne et ses habitudes de vie. Par exemple, dans le cas de l'interaction homme machine, l'étude du profil consiste

à analyser les différents comportements liés à l'utilisation de la souris et du clavier ainsi que les commandes utilisées. Cela nous permet de mieux comprendre la façon dont la personne interagit avec l'ordinateur. En outre, les informations sociologiques de la personne ainsi que ses préférences rentrent également dans l'étude du profil de la personne.

Définition 8. Profil : Selon Godoy et al. [32], le profil peut être défini comme la description des caractéristiques, des intérêts et des préférences de la personne qui peuvent être obtenus de façon statique en utilisant des questionnaires et des interviews, ou de façon dynamique en utilisant des approches de forage de données.

Le profil de la personne peut être classé en deux classes importantes. 1) l'étude du profil basé sur les connaissances, et 2) l'étude du profil basé sur le comportement. Le profil basé sur les connaissances s'intéresse à l'étude des caractéristiques, des préférences et des intérêts de la personne. Ces caractéristiques peuvent être statiques comme les informations sociologiques de la personne, comme elles peuvent être dynamiques et changent avec le temps comme les préférences et intérêts de la personne. Par ailleurs, le profil basé sur le comportement ou profil comportemental s'intéresse à l'étude des différents comportements de la personne lors de la réalisation de ses activités quotidiennes. Bien que ces deux classes de profil sont complémentaires, les chercheurs dans ce domaine [63, 148] donnent beaucoup plus d'importance au profil comportemental qu'au profil basé sur les connaissances. Ceci démontre en quelque sorte l'importance de ce type de profil non seulement dans le domaine de l'interaction homme machine, mais aussi dans tous les domaines où le comportement de la personne pourrait jouer un rôle déterminant dans l'acceptation, l'amélioration ou la prédiction des situations menant à l'augmentation des performances d'un système.

Dans ce chapitre, nous allons introduire une nouvelle approche que nous avons proposée pour la construction du profil usager. Nous allons donc nous attaquer à une problématique assez importante dans le domaine des habitats intelligents de façon particulière, et dans le domaine de la modélisation des usagers d'une manière générale. Nous allons porter une attention particulière au profil comportemental vu son importance dans le domaine des habitats intelligents. Le profil comportemental dans le domaine des habitats intelligents permet entre autres d'étudier les différents comportements de l'usager et de développer un modèle permettant de caractériser ces

comportements, et de prédire les comportements futurs de l'usager afin de personnaliser l'assistance et de donner une aide appropriée dans les moments opportuns.

Nous allons présenter de façon panoramique les différents travaux qui portent sur la construction du profil comportemental de la personne dans des différents domaines d'application. Cela nous permettra d'un coté de positionner notre travail par rapport à ces travaux, et de l'autre de mettre en valeur nos contributions dans ce domaine.

3.2 État de l'art

Dans cette section, nous allons discuter les différentes approches proposées dans la littérature pour la construction du profil usager. Vu l'importance du profil comportemental de la personne, nous allons porter une attention particulière à ce type de profil dans notre revue de la littérature.

Avant de commencer notre revue de la littérature, nous allons tout d'abort discuter les différentes méthodes utilisées pour collecter et recueillir les informations nécessaires pour la construction du profil usager. Il existe principalement deux méthodes pour la collecte des données relatives aux usagers qui nous permettent de construire le profil usager.

1. **Méthode statique** : dans cette méthode, les données sont collectées par le biais des questionnaires ou des interviews avec les usagers. Ces questionnaires comportent entre autres des questions relatives à la façon dont les usagers se comportent avec le système afin d'atteindre leurs objectifs. Cependant, cette méthode nécessite beaucoup de temps afin de rassembler tous les questionnaires réalisés avec les d'usagers. De plus, ce genre de questionnaire est généralement réalisé avec un échantillon d'usagers et non avec tous les usagers vu que ce travail est pénible et prend un temps considérable pour l'accomplir. De plus, cette méthode est inappropriée dans beaucoup de domaines, plus spécifiquement pour collecter les données sur internet où le nombre d'usagers est très grand. Cela rend cette méthode non pratique et assujettie aux erreurs.
2. **Méthode dynamique** : dans cette méthode, la collecte des données se fait de manière automatique par l'application des méthodes automatiques d'enregistrement et d'analyse des données comme les méthodes de forage des données.

À l'aide de cette méthode, nous pouvons collecter autant d'informations que nous pouvons auprès de nombreux usagers. Cela peut être fait dans un temps beaucoup plus rapide que la méthode statique.

Les méthodes dynamiques sont devenues une pratique inévitable dans beaucoup de domaines et plus spécifiquement dans le Web. Dans notre travail, nous utilisons des bases de données obtenues à l'aide des méthodes dynamiques dans le domaine des habitats intelligents.

Il existe beaucoup de défis et exigences pour les approches proposées pour la modélisation des profils usagers. Selon Geoffrey et al. [135], les exigences que les systèmes de modélisation des profils usagers doivent prendre en compte sont l'acquisition et l'annotation des données, la complexité du système, la personnalisation, l'adaptabilité, l'extensibilité, et l'interopérabilité.

- Acquisition et l'annotation des données : l'acquisition des données concerne les méthodes utilisées pour collecter les données. L'annotation des données est discutée en détail dans le chapitre 1.
- Complexité : les systèmes décrivant le profil usager doivent être efficaces et capables de satisfaire les besoins des utilisateurs dans un temps raisonnable.
- Personnalisation : le système de profil doit être en mesure de prendre en compte les préférences et les besoins personnels des usagers. Cela implique que l'usager exprime explicitement ses préférences, ou bien ses préférences seront inférées à partir des données.
- Adaptabilité : le système doit être capable de s'adapter aux changements des préférences des usagers dans le temps. Donc, le système devra s'ajuster en fonction des changements des préférences.
- Extensibilité : le système doit être capable d'incorporer des nouvelles données et préférences des usagers de façon simple et efficace.
- Interopérabilité : le système doit être capable de gérer des données provenant des sources hétérogènes. De plus, le système doit être fonctionnel sous différentes conditions.

Dans ce qui suit, nous allons discuter les différentes approches proposées dans la littérature pour la construction du profil comportemental usager.

3.2.1 Approches Statistiques

Les approches statistiques sont largement utilisées dans différents domaines avec une variété d'applications. Comme nous l'avons mentionné dans le chapitre 1, les approches statistiques ont pris de l'ampleur dans le domaine de la reconnaissance d'activité. De la même façon, ces approches sont aussi largement utilisées dans la construction du profil comportemental usager. Le principe de base de ces approches est toujours le même que celui présenté dans le chapitre 1. La seule différence réside dans le domaine d'application et la nature du problème traité.

Réseaux Bayésiens

Parmi les approches statistiques les plus utilisées pour la construction du profil usager, nous trouverons les réseaux Bayésiens. La puissance de ce modèle statistique est tirée de son aptitude à modéliser des phénomènes aléatoires complexes. Les réseaux Bayésiens ont été largement utilisés dans la modélisation du profil usager. Par exemple, Kritikou et al. [63] propose un modèle basé sur un réseau Bayésien pour la prédiction des préférences des usagers dans un système d'apprentissage en ligne. L'objectif de leur travail consiste à développer une nouvelle plateforme capable d'accélérer le processus d'apprentissage et de le rendre beaucoup plus efficace. Dans leur modèle, le réseau Bayésien est principalement utilisé pour encoder, apprendre et faire du raisonnement sur les relations probabilistes entre les différentes variables du domaine. Le système proposé comporte plusieurs composantes principales à savoir le domaine d'étude, l'interface usager et le profil usager. La figure 3.1 présente l'exemple du réseau Bayésien construit pour modéliser le profil usager.

Comme nous pouvons le constater dans la figure 3.1, le réseau Bayésien permet de modéliser les relations entre les différents paramètres à savoir la durée de la leçon, la difficulté du contenu, la durée du test, la performance, etc. Nous constatons par exemple que le paramètre "difficulté du contenu" est directement lié aux paramètres d'évaluation tels que "durée de la leçon", "durée du test" et "performance". Par conséquent, nous pouvons prédire la difficulté du contenu à partir de ces trois paramètres. La même observation pourrait être appliquée aux autres paramètres du modèle. Les réseaux Bayésiens ont prouvé leur performance dans la modélisation des systèmes

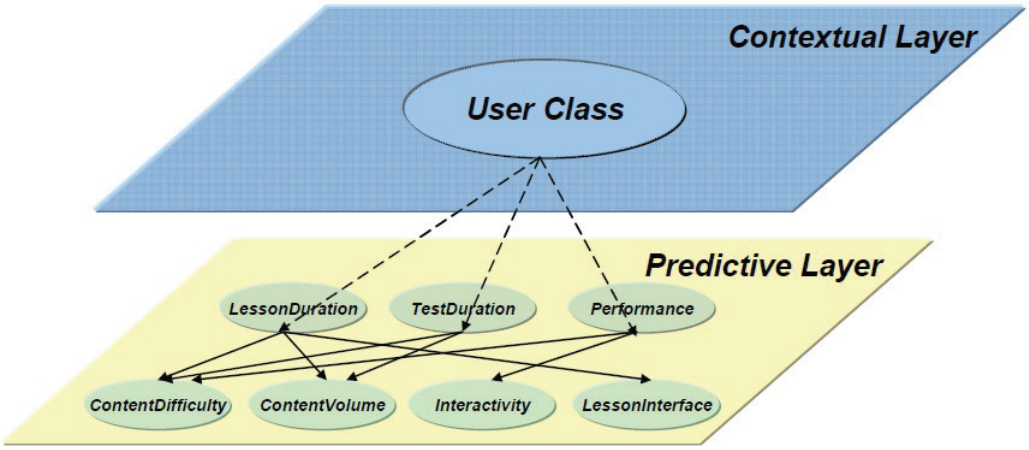


figure 3.1 – Exemple d'un réseau Bayésien construit pour modéliser le profil usager. Figure tirée de [63]

complexes et la prédiction. Cependant, ils sont basés sur l'hypothèse d'indépendance des variables qui n'est pas toujours le cas dans la pratique.

Schiaffino et al. [116] combinent les réseaux Bayésiens et le raisonnement basé sur le cas pour modéliser le profil usager lors de l'interaction avec une base de données. L'interaction avec une base de données est caractérisée par l'envoi des requêtes et la réception des réponses. Le raisonnement basé sur le cas permet de résoudre des nouveaux problèmes en se rappelant des situations similaires précédentes et en utilisant les informations et connaissances de cette situation [4]. Le raisonnement basé sur les cas est utilisé dans ce modèle du profil afin de stocker les informations relatives au comportement de l'usager lors de l'envoi de ses requêtes. Ces informations, qui représentent des mots clés employés par l'usager ainsi que des informations sur la date et heure de la requête, sont stockées dans la base de données sous forme de cas. Les requêtes sont généralement stockées à base de leur similarité avec des requêtes stockées précédemment dans la base de données. Le réseau Bayésien est utilisé pour modéliser les relations entre les différents attributs.

Les informations stockées dans la base des cas et le réseau Bayésien sont utilisées pour construire le profil d'un usager. Le profil dans ce cas là contient des informations sur les types de requêtes souvent utilisées par un usager, et les situations dans lesquelles sont exprimées ces requêtes. Un exemple de requête exprimée sous forme

de cas est présenté dans la figure 3.2.

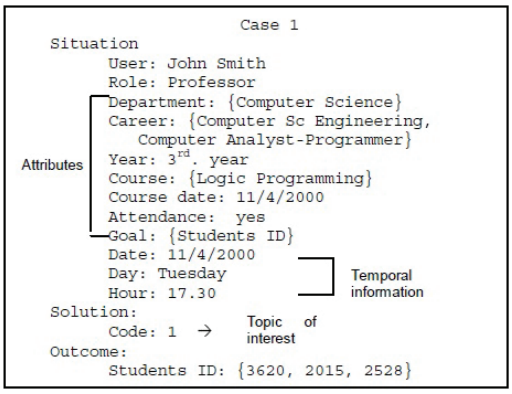


figure 3.2 – Exemple de requête exprimée sous forme de cas. Figure tirée du [116]

L'objectif de ce modèle est de construire des profils pour les usagers afin de leur suggérer prochainement des requêtes en se basant sur leurs profils déjà construits.

Les modèles Bayésiens hiérarchiques ont été utilisés par Zigoris et al. [150] pour construire le profil comportemental usager, afin de résoudre le problème du démarrage à froid dans les systèmes de recommandation. Le démarrage à froid correspond à l'ajout des nouveaux items ou usagers dans la base de données du système de recommandation. Dans ce modèle, les informations implicites et explicites de l'usager sont utilisées dans la construction de son profil. Les informations explicites représentent des avis exprimés explicitement par les usagers dans le système de recommandation. Par contre les informations implicites représentent les évaluations (rating) des usagers exprimées sur des items. Comme nous pouvons le constater à partir de la figure 3.3, chaque évaluation (Y) sur un document (X) est conditionnée par le modèle du profil (f) de l'usager. Les usagers partagent des informations sur leur modèle via le paramètre (θ).

Cependant, dans des systèmes de recommandations, les informations explicites ne sont pas toujours disponibles, de même pour les informations implicites vu que les usagers évaluent souvent un petit nombre d'items. De plus, les informations implicites ne reflètent pas toujours les mêmes préférences des usagers, ce qui pourrait influencer le partage de profil dans ce modèle et limite l'application de ce modèle dans la pratique.

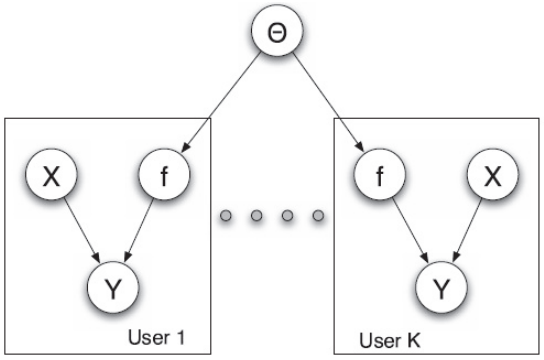


figure 3.3 – Exemple de dépendances entre les variables dans un modèle Bayésien Hiérarchique. Figure tirée de [150]

Zhang [148] propose un un modèle graphique pour la construction du profil usager lors de l'interaction avec un ordinateur. Ce modèle est basé sur les relations d'indépendance entre les différentes variables du modèle et produit un modèle graphique qui ressemble à un réseau Bayésien. L'hypothèse stipule que l'usager expose le même comportement durant la réalisation de ses activités. Pourtant, l'usager pourrait exhiber des comportements différents lors de la réalisation de la même activité. Étudier les relations entre ces comportements représente aussi notre objectif dans notre travail, cependant notre travail permet de prendre en compte les différents comportements des usagers et d'extraire les relations causales entre ces différents comportements afin de construire un profil comportemental des usagers.

Modèle Markovien

Les modèles Markoviens sont également utilisés afin de modéliser le profil usager. Beaucoup de travaux ont été réalisés pour la modélisation du profil usager en adoptant des modèles Markoviens cachés ou d'autres variantes du modèle Markovien comme les modèles hiérarchiques par exemple.

Manavoglu et al. [84] proposent un mélange de modèles Markoviens pour modéliser le profil usager. Le modèle Markovien est utilisé dans ce travail afin de modéliser l'historique des actions des usagers. Un modèle Markovien d'ordre 1 est utilisé où la prédiction de la prochaine action ne dépend que de l'action courante. Le modèle proposé consiste donc à prédire $P(A^{prochaine}|H(U), données)$ où $A^{prochaine}$ représente

l'action future, et $H(U)$ représente l'historique de l'usager. Conceptuellement, ce modèle ressemble à notre modèle proposé dans la détection des comportements pour chaque usager et pourrait être utilisé comme une base dans notre modèle pour la détection des relations causales. Cependant, l'hypothèse de Markov employée dans ce travail rend le modèle beaucoup plus limité en supposant que l'action future pourrait être prédite uniquement à la base de l'action courante. L'hypothèse d'indépendance conditionnelle des variables constitue une hypothèse sévère de ce modèle.

Un autre travail qui utilise les modèles Markoviens couplés (une variante du modèle HMM) pour l'analyse des comportements d'achat des usagers a été proposé par Longbing et al. [24]. Le modèle Markovien couplé est utilisé pour analyser les comportements partagés entre les usagers dans les données transactionnelles. Le comportement dans leur travail représente une opération d'achat qui comporte l'item acheté, la quantité achetée, la date et heure d'achat. Les auteurs ont étudié le couplage entre deux comportements qui pourraient conduire à des problèmes commerciaux. Ces comportements sont appelés des comportements anormaux.

Allocation Dirichlet Latente

Pour modéliser les comportements usagers dans des bases de données de haute dimensionnalité, Ahmed et al. [2] proposent un modèle statistique basé sur l'allocation Dirichlet latente (LDA). L'objectif de leur modèle est de personnaliser et cibler les publicités en ligne.

Pour ce faire, le modèle LDA a été utilisé afin d'exploiter sa puissance dans la détection des aspects sémantiques dans les données. De plus, ces aspects peuvent être détectés dans des groupes de données catégorisés de façon non supervisée à l'aide du modèle LDA. Chaque cluster contient des comportements similaires exprimés par les usagers. Dans ce modèle, les usagers (qui représentent les documents) sont modélisés comme des distributions de leurs préférences. Donc, les préférences représentent les thèmes dans leur modèle. Malgré sa complexité, leur modèle ressemble à notre travail relatif à la construction du profil comportemental du point de vue de la détection des comportements similaires. Donc, leur modèle pourrait être utilisé comme une base pour notre modèle sur laquelle nous pouvons bâtir notre approche permettant de découvrir les relations causales entre les différents comportements.

3.2.2 Résumé

Dans cette section du chapitre, nous avons présenté brièvement les différentes approches utilisées dans la littérature pour la modélisation du profil comportemental usager. Nous avons pu constater que les approches statistiques ont pris de l'ampleur et sont largement utilisées dans ce domaine. Cependant, ce que nous avons pu conclure est que ces approches ne permettent pas modéliser les relations bidirectionnelles. Une relation bidirectionnelle entre deux variables signifie que chaque variable possède une relation avec l'autre variable. Les réseaux Bayésiens ou les modèles Markoviens ne permettent pas de modéliser ce type de relations qui peuvent être très importantes dans la pratique. En outre, la vaste majorité des approches statistiques proposées dans la littérature se basent sur l'hypothèse d'indépendance des variables qui pourrait influencer sur les résultats des modèles développés. Pour surmonter ces problèmes, nous proposons une approche permettant la construction du profil comportemental usager en se basant sur l'analyse causale. Avec l'analyse causale, nous pouvons modéliser la bidirectionnalité entre les variables, et nous pouvons également suspendre l'hypothèse d'indépendance entre les variables. Les détails de notre approche seront présentés dans les sections suivantes.

3.3 Positionnement de notre travail par rapport aux travaux existants

Nous constatons à travers notre présentation panoramique des travaux existants que la vaste majorité des travaux existants souffrent des problèmes comme la variabilité du comportement usager, la découverte des relations causales entre les différents comportements des usagers, et la prise en compte de la bidirectionnalité entre les variables.

Dans cette thèse, et comparativement aux approches existantes, nous proposons une nouvelle approche pour la découverte des différents comportements usagers de même que des relations causales entre ces différents comportements. Comme nous avons pu le constater à travers l'état de l'art, il existe peu de travaux qui traitent le problème de la découverte des comportements de façon non supervisée. Dans ce

contexte, l'approche que nous proposons définit un cadre théorique innovant qui tire ses bases à partir de deux importantes approches de forage de données : 1) le forage de patrons fréquents en utilisant les arbres probabilistes des suffixes, et 2) le clustering basé sur les patrons fréquents. La combinaison de ces approches permet de découvrir les comportements des usagers, et d'exploiter également les relations entre ces différents comportements en adoptant une approche basée sur l'analyse causale. Enfin, les relations causales jouent un rôle très important dans l'identification des usagers et la prédiction des activités.

En outre, nous avons aussi pu constater que l'usager peut exhiber des comportements différents lors de la réalisation de ses activités. L'apport de notre approche dans ce contexte est qu'elle permet de découvrir les différents comportements des usagers de façon automatique. Ces comportements sont découverts à l'aide du modèle probabiliste basé sur les arbres probabilistes des suffixes appliqués sur les séquences de comportements. Un point très important qui en découle est que les patrons gardés dans l'arbre sont tous statistiquement significatifs selon une définition absolue de signification ou une définition bien spécifique au domaine d'application. Par conséquent, notre approche permet de découvrir ces différents comportements et permet aussi de les catégoriser de telle sorte que les comportements similaires appartiennent à la même catégorie (cluster). L'objectif derrière cette catégorisation est que les séquences similaires possèdent les mêmes relations causales. Donc, la catégorisation des séquences, dans notre cas, permet de faciliter la recherche des relations causales entre les comportements. Chaque catégorie (cluster) représente un profil comportemental de l'usager. Par conséquent, notre travail aborde une problématique triple : 1) la découverte des différents comportements à partir des séquences, 2) la catégorisation des séquences en se basant sur les comportements, et 3) la découverte des relations causales entre les différents comportements.

Le potentiel de découvrir ces relations causales réside dans le fait que ces relations peuvent être utilisées pour développer beaucoup d'applications à savoir l'identification des usagers selon leurs profils, et la prédiction des activités. Selon notre connaissance dans ce domaine, cela constitue un problème nouveau qui n'a jamais été soulevé par les approches existantes.

3.4 Construction du profil usager en utilisant l'analyse causale

Dans cette section, nous allons présenter notre approche de construction du profil comportemental usager. Notre approche comporte deux étapes importantes : 1) la découverte des patrons significatifs, et 2) la découverte des relations causales entre les patrons significatifs. Avant de présenter plus en détail ces deux étapes, nous allons tout d'abord présenter les points les plus importants qui nous ont motivé pour proposer une approche basée sur l'analyse causale pour la construction du profil usager.

3.4.1 Motivations pour l'analyse causale

Étudier les relations entre les différents comportements d'un usager permet de concevoir beaucoup d'applications liées à la personnalisation, l'adaptation et les applications centrées sur l'usager. Comme nous l'avons vu dans la revue de la littérature, ces relations prennent différentes formes telles que les relations probabilistes, les relations de corrélations, les relations d'indépendances. Ces relations peuvent être étudiées dans différents domaines avec différents niveaux de complexité. Cependant, dans des situations, certaines relations ne peuvent pas être détectées ou bien elles sont difficiles à chercher par les méthodes usuelles. Pour illustrer cette limitation, prenons les scénarios suivants :

"Alex a pris la route entre Montréal et Sherbrooke en conduisant sa voiture, il arrive à Magog, la voiture tombe en panne. Alex appelle son ami Marc qui habite à Sherbrooke pour venir le dépanner. Marc prend sa voiture et roule à une vitesse de 190 km/h, avant d'arriver à Magog, Marc n'a pas respecté le feu rouge sur la route et à l'intersection il rentre en collision mortelle avec une autre voiture." La question qui se pose maintenant : quelle est la cause réelle du décès de Marc ?

L'appel téléphonique d'Alex, l'excès de vitesse ou le non respect du feu rouge pouvaient être la cause du décès. Toutes ces propositions seront probables et peuvent être correctes. Toutefois, déterminer la cause réelle de décès reste un point difficile à résoudre. Les relations usuelles dans ce contexte peuvent ne pas donner d'interprétations correctes puisque les relations peuvent être cachées et donc ne pas être détectées

en utilisant les relations classiques comme une simple corrélation.

Les relations causales possèdent un potentiel important dans la détection des causes réelles en observant les effets, ou bien prédire les effets en observant les causes. Étant donné que les relations causales ne puissent être extraites que si les variables possèdent des relations de corrélation, il faut donc tout d'abord chercher les corrélations entre les variables pour chercher les relations causales. En effet, selon [120, 99], la causalité implique la corrélation, mais la corrélation n'implique pas nécessairement la causalité. Par exemple, l'emphysème est corrélé avec les maladies cardiaques, et toutes les deux sont causées par le tabagisme, mais nous ne pouvons pas dire que l'emphysème cause les maladies cardiaques ou inversement. De la même façon, la guerre peut entraîner des décès, donc les décès sont corrélés avec la guerre, mais nous ne pouvons pas dire que la cause réelle des décès est toujours la guerre. Dans cet exemple, nous constatons qu'il existe des relations dans les données qui ne peuvent pas être extraites par des méthodes simples de corrélation comme les méthodes Bayésiennes ou Markoviennes.

Un autre point important c'est la relation bidirectionnelle entre les variables. Par exemple, les personnes souffrant du diabète risquent d'avoir des symptômes de dépression. De même, les personnes présentant des symptômes de dépression risquent de développer un diabète [1]. Dans le même contexte, le manque de sommeil peut être considéré comme la cause du stress, et inversement, le stress peut être à son tour considéré comme la cause du manque du sommeil [2]. Comme un ultime exemple, des chercheurs Taïwanais ont pu découvrir des relations bidirectionnelles entre la schizophrénie et l'épilepsie [141]. En fait, les chercheurs ont conclu que les patients souffrant d'épilepsie sont susceptibles de développer la schizophrénie, et les patients souffrant de schizophrénie sont susceptibles de développer une épilepsie. Tous ces exemples montrent une bidirectionnalité entre des variables ou objets. Ces relations ne peuvent pas être modélisées en utilisant les modèles Markoviens ou réseaux Bayésiens dont les relations sont exprimées dans un seul sens. Ces points importants nous motivent à proposer une nouvelle approche permettant de modéliser ce type de relations en

1. http ://www.thefreelibrary.com/Diabetes+and+depression+show+bidirectional+association-a0184660928
2. http ://www.sciencedaily.com/releases/2009/06/090610091236.htm

tirant avantage de la théorie de causalité.

Une autre motivation de notre approche réside dans le fait que les réseaux Bayésiens sont incapables de représenter les relations causales. Comme l'illustre Pearl [60] via le paradoxe de Simpson. Ce paradox montre que les contributions des groupes pris séparément semblent s'inverser lorsqu'ils sont pris ensemble [79].

Dans le domaine d'assistance des personnes présentant des déficits cognitifs ou des personnes âgées, un système de reconnaissance d'activités isolé est insuffisant. En effet, les comportements variables de ces personnes et les situations imprévues peuvent survenir à tout moment et peuvent nuire à leur état de santé. Par exemple, les personnes atteintes de la maladie d'Alzheimer réagissent parfois d'une façon agressive. Cette agressivité découle souvent de la maladie plutôt que du caractère de la personne. Dans ces situations, les systèmes de reconnaissance d'activités ne pourront pas prédire ce type de situations. Cependant, si l'habitat intelligent est doté d'un système qui permet d'étudier et d'analyser les différents comportements de l'usager dans les différentes situations, nous serons capables, selon une situation donnée et des comportements observés, de prédire avec un certain degré de plausibilité ce qui va se passer dans les moments à venir. Par exemple, la frustration et l'anxiété sont des raisons possibles qui expliquent l'agressivité des personnes atteintes de la maladie d'Alzheimer. Par conséquent, nous pourrons prédire l'agressivité de la personnes si les signes de frustration ou d'anxiété apparaissent. Cela permet d'agir plus vite et de personnaliser et adapter les services nécessaires afin d'assurer le bien être de la personne. Par conséquent, un système de profillage est très important dans ce cas pour assurer une meilleure assistance des personnes tout en conservant leur bien être.

3.4.2 Formalisation du problème

Dans cette section nous allons introduire l'aspect formel du problème de recherche des relations causales entre les comportements usagers. Pour cela, nous aurons besoin de définir quelques notions et concepts qui nous permettent de comprendre le problème de l'analyse causale. Puisque nous abordons le problème du comportement de l'usager, nos définitions doivent être en accord avec les définitions utilisées dans les domaines de la psychologie et des sciences comportementales à savoir les travaux de

[100, 58, 70].

Le comportement humain est généralement considéré comme le processus permettant d'accomplir des activités mentales, physiques, émotionnelles et sociales d'un être humain. Dans ce travail, nous portons une attention particulière aux comportements permettant d'accomplir des activités de la vie quotidiennes comme manger, dormir, se réveiller, se déshabiller, se laver, travailler, voyager, marcher, nettoyer. Selon Wayne [70], le comportement humain est caractérisé par un ensemble de patrons d'actions. Par conséquent, l'action est considérée comme l'unité de base ou une composante du comportement. Dans ce qui suit, un patron d'action ou une action va correspondre à un patron de comportement ou appelé simplement un patron.

Soit $\mathcal{D} = \{S_1, S_2, S_3, ..., S_D\}$ une base de données contenant des séquences de comportements définies par des états de capteurs, des traces d'utilisation, des séquences biologiques, ou des données transactionnelles. Supposons que toutes les séquences correspondent à un usager particulier u. Soit $\xi = \{s_1, s_2, ..., s_n\}$ l'ensemble de tous les symboles dans la base de données, appelé aussi "alphabet". Étant donnée une base de données de séquences, notre objectif fondamental est de découvrir les relations causales entre les patrons de comportements. Cependant, comme nous l'avons mentionné précédemment, les relations causales sont généralement évaluées entre les variables significatives du domaine. À cet effet, notre travail comporte deux défis majeurs :

- Découvrir les patrons de comportements significatifs à partir de séquences.
- Découvrir les relations causales entre ces différents patrons de comportements significatifs.

Avant de présenter les méthodes que nous avons proposées pour résoudre ces défis, nous allons introduire tout d'abord quelques définitions et notations.

Notion du patron significatif

Dans la littérature, il existe beaucoup de définitions pour le concept du patron significatif. Dans certaines définitions, un patron fréquent peut être considéré comme un patron significatif [43]. La définition d'un patron fréquent, appelé aussi dans notre cas un patron de comportement, est donnée comme suit :

Définition 9. Patron de comportement : Un patron de comportement p est une

sous séquence de symboles statistiquement significative ($s_1 s_2 ... s_t$). Une sous séquence de symboles est statistiquement significative si elle apparaît dans la base de données un nombre de fois supérieur à un certain seuil défini par l'utilisateur.

Formellement, un patron est significatif s'il respecte la définition suivante :

Définition 10. Patron significatif : Soit p_α un patron dans $\mathcal{D}$. Notons l'ensemble de séquences dans lesquelles p_α apparaît comme $S_\alpha = \{S_i | p_\alpha \in S_i, S_i \in \mathcal{D}\}$. Un patron p_α est significatif dans la base de données $\mathcal{D}$, si $\frac{|S_\alpha|}{|\mathcal{D}|} \geq \sigma$, où σ est un seuil défini par l'utilisateur, et $\frac{|S_\alpha|}{|\mathcal{D}|}$ est appelé le support de p_α.

Cependant, le patron significatif véhicule plus de signification que la simple répétition de ce patron. Selon Han et al. [40], un patron est significatif (important ou intéressant) s'il est facile à comprendre par les utilisateurs (les humains en général), valide sur des nouvelles données ou des données de test avec un certain degré de certitude, potentiellement utile, ou s'il peut valider certaines hypothèses que les utilisateurs cherchent à confirmer [3]. Les mesures utilisées pour la détection des patrons significatifs peuvent varier d'un domaine d'application à un autre, mais l'importance du patron et sa définition reste toujours valable quelque soit le domaine étudié.

Il existe plusieurs méthodes pour l'extraction des patrons significatifs. La méthode la plus connue consiste à extraire les patrons fréquents, puis à appliquer une méthode de filtrage sur ces patrons pour ne garder que les patrons les plus intéressants (méthode a posteriori). D'autres méthodes utilisent ce qu'on appelle une optimisation sur les requêtes sans passer par les méthodes d'extraction des patrons fréquents (méthode a priori). Ces requêtes possèdent des contraintes sur la nature des patrons à extraire (constraint-based mining). Ces méthodes itératives s'arrêtent lorsque les critères de signification de patrons sont satisfaits. Des mesures ont été introduites pour détecter les patrons fréquents intéressants. Ces mesures sont divisées en deux classes : mesures objectives et mesures subjectives.

- Les mesures objectives sont basées sur les statistiques (comme la corrélation) et les structures des patrons, par exemple, le support, la confiance, la capacité de discrimination, etc.
- Les mesures subjectives sont basées sur les croyances et les observations de

3. Cette définition est la traduction de celle présentée par [40]

l'utilisateur sur les données, par exemple, la nouveauté, l'influence, etc.

Plusieurs algorithmes ont été développés pour le forage des patrons compressés ou approximatifs dans le but de réduire le nombre important de patrons fréquents. Par exemple, Wang et al.[134] ont introduit un algorithme pour extraire les k patrons fermés [133] les plus fréquents. L'algorithme incrémente graduellement le seuil de support et élague l'arbre FP-Tree durant et après la construction. Pei et al.[96] ont proposé une approche basée sur les patrons condensés. Dans cette approche les patrons sont partitionnés selon leur support, et les patrons les plus représentatifs seront détectés dans chaque groupe. Patil et al.[105] ont appliqué une mesure de poids sur les patrons fréquents pour extraire les patrons intéressants pour prédire des attaques cardiaques. Cambouropoulos [18] a utilisé la notion de période et de couverture pour extraire les patrons intéressants à partir des données musicales. Selon l'auteur, ces types de patrons sont aussi de grande importance lors du traitement des données biologiques. Pour les données musicales l'importance d'un patron peut être déterminée par la répétition immédiate d'un patron (deux répétitions consécutives ou plus). Tseng et al.[126] ont introduit la notion de patrons de grande utilité. Une mesure d'utilité est introduite dans ce travail afin de calculer l'utilité de chaque patron fréquent. Les patrons fréquents dont l'utilité est inférieure à un seuil défini par l'utilisateur ne seront pas considérés comme importants. Xin et al.[139] ont utilisé une mesure permettant d'extraire les patrons significatifs. Cette mesure calcule la différence entre la fréquence observée d'un patron et la fréquence attendue (expected). Ici, la fréquence désigne la proportion de transactions contenant le patron.

Afin de faciliter la découverte des relations causales entre les différents patrons de comportements, nous avons introduit une nouvelle méthode permettant la découverte des patrons significatifs en utilisant les arbres probabilistes des suffixes. En effet, la signification statistique est une définition absolue pour toutes les applications dans tous les domaines. Cependant, dans notre travail, cette définition est nécessaire mais elle est insuffisante pour décrire la causalité. À cet effet, nous avons incorporé le concept de la corrélation afin que la définition soit nécessaire pour le problème que nous sommes entrain de traiter.

Notion de causalité

La notion de causalité tire ses origines de la philosophie. La causalité est une relation entre un événement nommé une cause et un autre événement nommé l'effet. L'effet est toujours considéré comme la conséquence de la cause. Selon Heylighen [47], la causalité est une relation qui relie la cause à l'effet. Il existe plusieurs types de causalité à savoir la causalité probabiliste, la causalité contrefactuelle, la causalité systémique. Dans ce qui suit, nous allons discuter de la causalité probabiliste qui est la plus utilisée dans la littérature [120, 99, 60].

Définition 11. ***Causalité :*** Soient X et Y deux variables aléatoires. La causalité entre X et Y, i.e. X causant Y exige les trois conditions suivantes :

- X précède Y.
- P(X) $\neq 0$.
- $P(Y|X) > P(Y)$.

Il existe différentes mesures qui permettent de détecter la causalité entre les variables. Nous allons introduire uniquement deux mesures qui sont largement utilisées dans la littérature.

La causalité de Granger

La causalité de Granger a été introduite par le chercheur Clive W. J. Granger [36], un économiste qui a reçu le prix Nobel en économie en 2003. Dans ce type de causalité, une série temporelle est utilisée pour prédire une autre série temporelle. Formellement, soient $X = \{(X(t))\}$ et $Y = \{(Y(t))\}$ deux séries temporelles stationnaires dont nous voulons étudier la relation de causalité. L'approche classique de Granger est basée sur la régression linéaire. La régression stipule que, X cause Y si les valeurs passées de X aident à prédire les valeurs futures de Y. Supposons les deux régressions suivantes :

$$Y(t) = \sum_{l=1}^{L} a_l Y(t-l) + \epsilon_1 \tag{3.1}$$

$$Y(t) = \sum_{l=1}^{L} a_l Y(t-l) + \sum_{l=1}^{L} b_l X(t-l) + \epsilon_2 \tag{3.2}$$

où L est le lag temporel maximum (le lag temporel correspond à une fenêtre temporelle dont la longueur est généralement égale à 1), et ϵ_1 et ϵ_2 représentent des bruits. Si l'équation 3.2 décrit mieux le modèle que l'équation 3.1 alors nous disons que X cause Y. Pour ce faire, nous calculons généralement les variances des deux modèles, puis utilisons un test statistique de signification comme le test de vraisemblance [85].

Transfert d'entropie

Introduit par le physicien Thomas Schreiber [118] en 2002, le principe du transfert d'entropie est utilisé afin de détecter la causalité entre deux systèmes dynamiques (appelés aussi processus). Le principe de cette mesure de causalité est de mesurer les dynamiques d'informations partagées entre deux processus. La mesure de l'information mutuelle, qui est une mesure symétrique, ne prend pas en compte la dynamique de l'information partagée entre les deux processus. Contrairement à la mesure de l'information mutuelle, le transfert d'entropie est une mesure asymétrique qui requiert d'être calculée dans les deux sens entre deux processus.

La mesure du transfert d'entropie est basée sur le concept d'entropie conditionnelle entre deux processus. Cette mesure peut s'écrire sous la forme suivante entre deux processus X et Y : $H_{X|Y} = \sum P(x,y)log(x|y) = H_{XY} - H_Y$. Pour incorporer la dynamique de la structure, nous pouvons exploiter les probabilités de transition. Pour ce faire, les processus sont approximés par des processus Markoviens stationnaires d'ordre k. Cela veut dire que, la probabilité que le processus X soit à l'état x_{n+1} à l'instant $n+1$ peut être exprimée sous la forme : $P(x_{n+1}|x_n, ..., x_{n-k+1}) = P(x_{n+1}|x_n, ..., x_{n-k})$.

Les bits nécessaires pour encoder un état additionnel du processus si tous les états précédents sont connus peuvent être calculés en utilisant le taux d'entropie comme suit :

$$h_X = -\sum_{x_n} P(x_{n+1}, x_n^{(k)}) log P(x_{n+1}|x_n^{(k)}) \tag{3.3}$$

Où $x_n^{(k)} = x_n, ..., x_{n-k+1}$. Cette équation n'est que la différence entre les entropies de Shannon des processus donnés par les vecteurs de $k+1$ et k dimensions respectivement.

Pour étudier la dynamique des informations partagées entre plusieurs processus, Schreiber propose de généraliser le taux d'entropie. Cette généralisation est appelée

le transfert d'entropie et peut s'écrire sous la forme :

$$T_{Y\to X} = \sum_{x_{n+1},x_n,y_n} P(x_{n+1}, x_n^{(k)}, y_n^{(l)}) log \frac{P(x_{n+1}|x_n^{(k)}, y_n^{(l)})}{P(x_{n+1}|x_n^{(k)})} \tag{3.4}$$

Les paramètres k et l prennent généralement la valeur de 1 pour des raisons de simplicité du système. Nous aurons donc des processus Markoviens d'ordre 1. Cette généralisation permet de mesurer le degré d'indépendance des deux processus X et Y par rapport à la propriété de Markov vue précédemment. La formule 3.4 présente le transfert d'entropie entre les deux processus X et Y. Notons que cette mesure est asymétrique et requiert d'être calculée dans les deux sens, c'est à dire de $X \to Y$ et de $Y \to X$. Si la valeur du transfert de $Y \to X$ est plus grande que celle de $X \to Y$, nous disons que le processus Y influence (cause) le processus X et inversement.

3.4.3 Patrons significatifs

Dans cette section, nous allons introduire notre approche pour la découverte des patrons significatifs dans les bases de données séquentielles.

La découverte des relations causales est généralement effectuée entre les variables significatives. Dans notre travail, l'objectif est de découvrir les relations causales entre les patrons de comportements. Ces patrons de comportements doivent satisfaire notre définition relative à la signification et l'importance en terme de signification statistique et de corrélation. Par conséquent, nous devons tout d'abord chercher les patrons statistiquement significatifs, puis par la suite nous pourrons évaluer la corrélation entre les patrons statistiquement significatifs.

Comme mentionné auparavant, il existe une multitude de méthodes qui permettent de découvrir les patrons significatifs dans les bases de données séquentielles. Mais, toutes ces techniques ne permettent pas d'extraire des informations riches qui accompagnent l'extraction des patrons comme les statistiques sur les patrons, les probabilités conditionnelles d'apparition, et ainsi de suite. Ces informations jouent un rôle important dans la recherche des relations causales. Pour cette raisons, nous proposons une nouvelle méthode permettant la découverte des patrons statistiquement significatifs avec des informations comme les probabilités conditionnelles d'apparition, la fréquence des patrons. Notre méthode exploite les arbres probabilistes des suffixes

qui ont montré de bonnes performances dans plusieurs domaines plus particulièrement dans le domaine de la biologie et l'analyse des séquences [62, 17, 117]. Dans ce qui suit, nous allons introduire le concept de l'arbre probabiliste des suffixes, ensuite nous présenterons comment extraire les patrons significatifs.

3.4.4 Arbre probabiliste des suffixes

Un arbre probabiliste des suffixes (PST), appelé aussi arbre des suffixes probabiliste, est un arbre enraciné. Un noeud est étiqueté par une sous séquence (patron) σ. Chaque noeud de l'arbre est associé à un vecteur de probabilité qui stocke les distributions des probabilités conditionnelles des prochains symboles possibles qui peuvent s'écrire sous la forme $P(s_{s\in\xi}|\sigma)$. Cette probabilité dénote la distribution de probabilité conditionnelle de l'occurrence du prochain symbole de ξ sachant la sous séquence précédente σ. La construction d'un PST peut se résumer comme suit :

- Un PST est un arbre de degré n, où n est la taille de l'alphabet.
- Chaque noeud excepté la racine est étiqueté par la paire (k, θ_i^{kj}), où k est une chaîne associée au parcours commençant du noeud jusqu'à la racine de l'arbre, θ_i^{kj} dénote la probabilité d'observer le symbole j après avoir observé la chaîne k dans la séquence S_i. θ_i^{kj} peut être calculée de la façon suivante : $\theta_i^{kj} = \frac{\eta_i^{kj}}{\eta_i^{k*}}$, où η_i^{kj} représente le nombre d'occurrences du symbole j après avoir observé la chaîne k dans la séquence S_i, et η_i^{k*} représente le nombre d'occurrences de n'importe quel symbole dans ξ après avoir observé la chaîne k dans la séquence S_i.
- Un PST garde un vecteur de distribution (θ^N) de probabilité pour chaque noeud N, et chaque parent d'un noeud est un suffixe pour ce noeud.

La figure 3.4 présente un exemple d'un PST où toutes les séquences sont composées de deux symboles 'a' et 'b'. Parcourir l'arbre de la racine à un noeud interne nous donne un patron inversé qui sert comme étiquette pour ce noeud. Dans cet arbre, le parcours de la racine en suivant le chemin $\Sigma \rightarrow b \rightarrow a$ montre que le patron 'ab' apparaît 185 fois dans la base de données des séquences, avec $p(b|ab) = 0.764$ et $p(a|ab) = 0.236$.

Par exemple, la probabilité d'une sous séquence 'aab' basée sur le PST de la figure 3.4 peut être calculée ainsi :

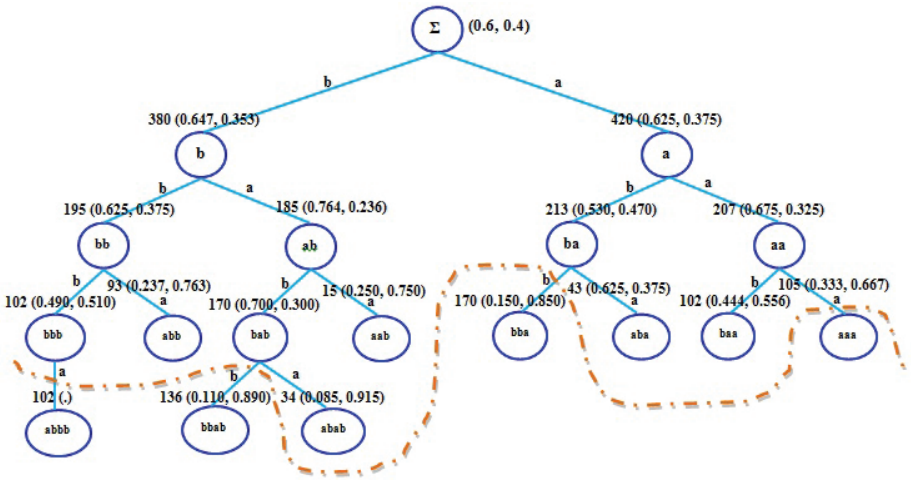


figure 3.4 – Exemple d'un arbre probabiliste des suffixes

$P(aab) = P(a) \times P(a|a) \times P(b|aa)$

$P(aab) = 0.4 \times 0.375 \times 0.675$

$P(aab) = 0.10125.$

Le PST permet un calcul très efficace des probabilités conditionnelles à l'intérieur de l'arbre comme indiqué dans [117].

Les PST représentent aussi un bon moyen pour calculer la similarité d'une séquence et un ensemble de séquences représentées sous forme d'un PST. Par exemple, la similarité de la sous séquence $\sigma = s_1 s_2 ... s_l$ en utilisant le PST peut être calculée de la façon suivante :

$$Sim(\sigma, PST) = \frac{P_{PST}(s_1) \times P_{PST}(s_2|s_1) \times P_{PST}(s_l|s_1 ... s_{l-1})}{P(s_1)P(s_2)...P(s_l)} \tag{3.5}$$

où $P_{PST}(.)$ représente la probabilité déduite à partir de l'arbre PST, et $P(.)$ représente la probabilité d'observation du symbole dans la sous séquence σ. Si la similarité dépasse un certain seuil, nous pouvons catégoriser cette sous séquence avec l'ensemble des séquences représentées sous forme de PST. Cette similarité joue un rôle très important dans le développement des algorithmes de classification non supervisée où la catégorisation des séquences similaires est de grande importance afin de découvrir des propriétés communes entre ces séquences.

3.4.5 Découverte des patrons significatifs

Dans notre travail, nous avons employé les PST afin de découvrir les patrons statistiquement significatifs. En effet, les PST permettent de découvrir ces patrons statistiquement significatifs par l'intermédiaire de statistiques que ce soit les fréquences ou les distributions des probabilités conditionnelles. Cependant, l'un des défis que présentent les PST est la profondeur de l'arbre. En fait, plus les patrons sont longs, plus le PST devient profond, et plus les calculs deviendront complexes et prennent beaucoup de temps. À cet effet, des mesures d'élagage de l'arbre ont été introduites afin de diminuer la profondeur, de contrôler la taille de l'arbre, et de garder ainsi uniquement les patrons les plus significatifs. Bejerano et al. [17] ont utilisé un mécanisme basé sur deux étapes pour l'élagage de l'arbre. Dans la première étape, un seuil des probabilités empiriques P_{min} est utilisé pour décider si le noeud fils sera augmenté (étendu) ou pas. Par exemple, dans le noeud étiqueté 'aa' dans l'arbre 3.4, si $P(baa) \geq P_{min}$, alors le noeud avec l'étiquette 'baa' sera ajouté dans l'arbre, sinon le noeud lui même et ses descendants seront ignorés. Dans la deuxième étape, un seuil de profondeur est employé pour couper le PST. Si la longueur du patron dans un noeud dépasse ce seuil, les descendants de ce noeud seront élagués.

Yang et al. [142] ont employé la notion de fréquence minimale min_{count} pour l'élagage de l'arbre. Si le nombre d'occurrence d'un patron est inférieur à min_{count}, alors le noeud et ses descendants seront élagués. Un travail similaire est utilisé par Xiong et al. [140]. Sun et al. [117] ont aussi utilisé la notion de fréquence minimale min_{count}.

Dans notre travail, nous avons proposé une autre méthode basée sur la divergence de Kullback Leibler (KL) [66]. Cette méthode tire avantage des distributions de probabilité stockées dans l'arbre et les utilise dans le calcul des divergences entre les distributions des noeuds qui se situent sur le même chemin (noeud père-fils). La divergence KL entre deux distributions θ^{N_1} et θ^{N_2} est calculée de la façon suivante :

$$D_{KL}(\theta^{N_1}, \theta^{N_2}) = \sum_{j=1}^{n} \theta_j^{N_1} \log_2 \frac{\theta_j^{N_1}}{\theta_j^{N_2}}. \tag{3.6}$$

La divergence KL est une mesure asymétrique qui requiert d'être évaluée dans les

deux sens. Afin d'obtenir une mesure de similitude, nous utilisons la formule suivante :

$$Sim(\theta^{N_1}, \theta^{N_2}) = \exp^{-(D_{KL}(\theta^{N_1}, \theta^{N_2}) + D_{KL}(\theta^{N_2}), \theta^{N_1}))}. \tag{3.7}$$

Par conséquent, la divergence KL est calculée pour chaque paire de patrons dans l'arbre. Si la mesure de similitude entre deux noeuds consécutifs est inférieure à un seuil Γ spécifié par l'utilisateur, le noeud fils et ses descendants seront élagués, sinon les deux noeuds seront gardés dans l'arbre. De cette façon, nous pouvons déterminer la profondeur optimale de l'arbre. Par exemple, supposons que $\Gamma = 0.5$. La divergence KL entre les distributions de la sous séquence 'ab' (qui peut être obtenue suivant le chemin $\Sigma \rightarrow b \rightarrow a$ dans le PST) de la figure 3.4 peut être calculée comme suit :

$D_{KL}(\theta^{ab}, \theta^{b}) = \sum_{j=1}^{2} \theta_j^{ab} \log_2 \frac{\theta_j^{ab}}{\theta_j^{b}}$

$D_{KL}(\theta^{ab}, \theta^{b}) = 0.764 \times \log_2 \frac{0.764}{0.647} + 0.236 \times \log_2 \frac{0.236}{0.353}$

$D_{KL}(\theta^{ab}, \theta^{b}) = 0.013885$

De la même façon, la divergence KL entre θ^b et θ^{ab} peut être calculée comme suit :

$D_{KL}(\theta^{b}, \theta^{ab}) = \sum_{j=1}^{2} \theta_j^{b} \log_2 \frac{\theta_j^{b}}{\theta_j^{ab}}$

$D_{KL}(\theta^{b}, \theta^{ab}) = 0.647 \times \log_2 \frac{0.647}{0.764} + 0.353 \times \log_2 \frac{0.353}{0.236}$

$D_{KL}(\theta^{b}, \theta^{ab}) = 0.01502$

En utilisant la formule de similarité de l'équation 3.7, nous aurons :

$Sim(\theta^{b}, \theta^{ab}) = \exp^{-(0.013885+0.01502)} = 0.971508.$

La valeur de similitude obtenue est très proche de 1, ce qui signifie que les deux distributions sont très proches. Par conséquent, les deux noeuds seront conservés, et le noeud fils pourra être augmenté par des suffixes 'a' ou 'b' sous les mêmes conditions de similarité citées précédemment. De plus, dans notre méthode, les noeuds qui ne possèdent pas de noeuds frères seront également élagués de même que leurs descendants. Cela permet de surmonter le problème de dégénérescence des sous arbres (qui se présente lorsque chaque noeud parent possède un seul noeud fils). Dans ce cas de figure, les probabilités des noeuds fils seront approximées par les probabilités de leurs noeuds parents. Par exemple, dans la figure 3.4, la probabilité $p(a|abbb) \approx p(a|bbb)$. Le processus d'élagage de l'arbre permet de garder uniquement les patrons significatifs qui seront utilisés par la suite pour étudier la corrélation entre eux.

3.4.6 Extraction des relations de corrélations entre les patrons

Comme nous l'avons mentionné auparavant, les relations causales sont généralement évaluées entre les variables corrélées. Selon [120, 99], la causalité implique nécessairement la corrélation, mais la corrélation n'implique pas nécessairement la causalité. Cela veut dire qu'il est primordial d'étudier les relations de corrélations entre les différents patrons significatifs afin de pouvoir évaluer les relations causales.

Par ailleurs, selon Liu et al. [75], les séquences similaires partagent les mêmes structures (relations) causales. À cet effet, nous proposons une méthode qui permet tout d'abord de catégoriser les séquences afin de rassembler les séquences similaires et de les mettre dans les mêmes groupes. Ensuite, nous étudions les relations de corrélations entre les différents patrons significatifs extraits dans chaque groupe de séquences (cluster). La catégorisation des séquences à ce stade possède beaucoup de potentiels. Par exemple, en plus de la facilité qui découle de cette catégorisation pour l'extraction des relations causales, les autres potentiels de la catégorisation (clustering) des séquences peuvent être résumés dans les points suivants :

- Un patron pourrait être significatif dans un ensemble de séquences, mais il n'est pas nécessairement significatif par rapport à toute la base de données des séquences.
- Le clustering des séquences permet de regrouper les séquences qui partagent les mêmes comportements. Cela permet d'extraire des patrons partagés par les différents comportements observés. Cela aussi pourrait être appliqué dans la recherche des patrons partagés par un groupe d'usagers.
- Le clustering des séquences peut potentiellement révéler des nouveaux groupes qui pourraient conduire à une meilleure compréhension du comportement usager.

Notons que les activités sont réalisées par les usagers de façon libre, sans contrainte sur la façon de réaliser les activités. De plus, il n'y a aucune condition sur la façon de faire de ces activités. Cela nous motive à étudier les profils des usagers lors de la réalisation de leurs activités. En effet, les usagers accomplissent leurs activités de façon différente. De plus, un usager peut accomplir la même activité de différentes

manières. Toutes ces constatations démontrent l'intérêt d'étudier ces différentes façons de faire, et les comportements des usagers qui y sont associés. Cependant, pour prendre en considération la variabilité du comportement usager, il est intéressant de regrouper les comportements similaires dans un même groupe pour essayer d'extraire des informations communes et partagées par ces comportements. Ce regroupement facilite d'un coté, l'étude des corrélations entre les différents patrons significatifs extraits pour chaque cluster, et facilite également la découverte des relations causales entre les patrons corrélés de l'autre coté. Par conséquent, nous proposons une méthode simple pour faire le clustering des séquences d'activités. Notre méthode tire profit des propriétés statistiques des arbres probabilistes des suffixes afin d'extraire les patrons significatifs qui seront employés dans le processus du clustering.

Il existe une multitude de méthodes de clustering des séquences dans la littérature. Toutefois, chaque méthode requiert des préconditions et paramètres à fixer a priori à savoir la mesure de similitude utilisée, la distribution de probabilité utilisée, et les seuils employés [143, 83]. Dans notre méthode, le clustering est basé sur les patrons significatifs extraits à partir de l'arbre probabiliste des suffixes. Il existe des travaux qui utilisent les patrons fréquents pour effectuer le clustering tel que le travail de [91] dans lequel les patrons de différentes longueurs sont utilisés. Contrairement au travail de [91], notre méthode utilise uniquement les patrons significatifs les plus longs et non mutuellement inclusifs dans les séquences. Les patrons non mutuellement inclusifs sont des patrons qui n'apparaissent pas nécessairement en même temps. L'avantage de notre méthode est que ces patrons significatifs existent déjà dans l'arbre probabiliste des suffixes ce qui permet d'accélérer le processus de clustering et d'éviter l'étape d'extraction de ces patrons à partir des séquences.

Notre méthode est basée sur la mesure de Jaccard pour mesurer la distance entre les séquences. Notre idée dans le clustering est basée sur le fait que les séquences similaires possèdent des patrons en commun. De plus, comme nous l'avons mentionné précédemment, notre méthode emploie les patrons les plus longs dans les séquences, ce qui permet de réduire significativement le temps de calcul contrairement aux autres travaux [91]. Les patrons les plus longs sont des patrons dont la longueur dépasse un certain seuil minimal. Dans notre méthode, nous avons choisi '3' comme longueur minimale des patrons dans notre algorithme de clustering.

Formellement, soit S_1 et S_2 deux séquences dans $\mathcal{D}$. Soit $P_{S_1} = \{p_{11}, p_{12}, ..., p_{1h}\}$ et $P_{S_2} = \{p_{21}, p_{22}, ..., p_{2m}\}$ les ensembles des plus longs patrons significatifs non mutuellement inclusifs des séquences S_1 et S_2 respectivement extraits à partir de l'arbre probabiliste des suffixes. La similarité de Jaccard entre les deux séquences S_1 et S_2 peut être calculée de la façon suivante :

$$Sim(S_1, S_2) = \frac{|P_{S_1} \bigcap P_{S_2}|}{|P_{S_1} \bigcup P_{S_2}|} \tag{3.8}$$

Les séquences S_1 et S_2 sont considérées similaires si la similarité $Sim(S_1, S_2) \geq \lambda$, où λ est un seuil spécifié par l'utilisateur. Plus la similarité est proche de 1, plus les séquences sont similaires et inversement. Notre algorithme de clustering est présenté dans l'algorithme 1.

Algorithme 1 Algorithme de clustering

Entrée :

- Base de données des séquences $\mathcal{D}$
- L'arbre probabiliste des suffixes (PST)
- Le seuil de similarité λ

Sortie :

- Le nombre de clusters T tel que : $C_i \bigcap C_j = \emptyset$ and $\bigcup_i C_i = \mathcal{D}$

1 : - Initialiser n clusters C_i, chaque cluster contient une sequence ;
2 : - Extraire les patrons les plus longs P_{C_i} pour chaque cluster C_i à partir du PST ;
3 : **pour chaque** paire de clusters C_i et C_j **faire**
4 : - calculer la similarité de Jaccard en utilisant la formule 3.8 ;
5 : **fin**
6 : **tant que** $sim(C_i, C_j) \geq \lambda$ **faire**
7 : - regrouper les clusters C_i et C_j dans un nouveau cluster C_u ;
8 : - mettre à jour les centres des clusters tel que l'ensemble des patrons les plus longs égale à $P_{C_i} \bigcap P_{C_j}$;
9 : **pour chaque** paire de clusters C_u et C_v $(u \neq v)$ **faire**
10 : - calculer la similarité de Jaccard $sim(C_u, C_v)$;
11 : **fin**
12 : **fin**
13 : **retourner les clusters**

Comme nous pouvons le constater dans l'algorithme 1, les séquences sont regroupées de façon itérative en se basant sur la mesure de similarité la plus grande. L'algorithme prendra fin lorsque la mesure de similarité sera inférieure au seuil prédéfini λ. Le résultat de cet algorithme est l'ensemble des différents clusters créé lors du processus de clustering.

Une fois l'algorithme de clustering fini, nous aurons des clusters contenant des séquences similaires et nous serons en mesure d'étudier les relations de corrélations entre les patrons significatifs extraits à partir de chaque cluster. À cet effet, nous construisons un arbre probabiliste des suffixes pour chaque cluster, pour extraire les patrons significatifs pour chaque cluster. Chaque arbre doit représenter un profil de l'usager. Notre méthode prend en considération des patrons avec différentes longueurs.

Une fois les PST construits pour chaque cluster, et les patrons significatifs déterminés à l'aide du PST, nous analysons les relations de corrélations entre ces patrons. Il existe plusieurs mesures de corrélations dont l'information mutuelle, les coefficients de corrélations. Dans notre travail, nous avons adapté la mesure de l'information mutuelle pour mesurer les relations de corrélations entre les patrons significatifs extraits à partir de chaque cluster. L'information mutuelle notée MI (Mutual Information) est l'une des mesures les plus utilisées dans la théorie d'information. La mesure MI est une mesure compréhensive, interprétable et facile à implémenter et qui a démontré ses forces dans différents domaines à savoir la médecine, la biologie, les télécommunications, la neuroscience, et la psychologie [118, 60]. Dans notre travail, nous adaptons la mesure MI afin de quantifier la force d'association entre les patrons significatifs dans le même cluster en incorporant les probabilités joints d'occurrence des patrons.

Formellement, soient p_α et p_β deux patrons significatifs, et soient $X_\alpha = \{0, 1\}$ et $X_\beta = \{0, 1\}$ deux variables aléatoires pour les occurrences de p_α et p_β respectivement dans le même cluster. Notons $\mathbf{S}_{C_i}$ l'ensemble des séquences dans le cluster C_i, $\mathbf{S}_{p_\alpha}$ l'ensemble des séquences dans le cluster C_i dans lequel le patron p_α apparaît, et $\mathbf{S}_{p_\beta}$ l'ensemble des séquences dans le cluster C_i dans lequel le patron p_β apparaît. Nous utilisons une version adaptée de la mesure de MI utilisée dans le travail de [91]. Par conséquent, la mesure $MI(X_\alpha, X_\beta)$ entre les deux variables aléatoires X_α et X_β est calculée comme suit :

$$MI(X_\alpha, X_\beta) = \sum_{x \in X_\alpha} \sum_{y \in X_\beta} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right), \tag{3.9}$$

où $p(x=1,y=1) = \frac{|\mathbf{S}_{p_\alpha} \cap \mathbf{S}_{p_\beta}|}{|\mathbf{S}_{C_i}|}$, $p(x=0,y=1) = \frac{|\mathbf{S}_{p_\beta}| - |\mathbf{S}_{p_\alpha} \cap \mathbf{S}_{p_\beta}|}{|\mathbf{S}_{C_i}|}$, $p(x=1,y=0) = \frac{|\mathbf{S}_{p_\alpha}| - |\mathbf{S}_{p_\alpha} \cap \mathbf{S}_{p_\beta}|}{|\mathbf{S}_{C_i}|}$, $p(x=0,y=0) = \frac{|\mathbf{S}_{C_i}| - |\mathbf{S}_{p_\alpha} \cup \mathbf{S}_{p_\beta}|}{|\mathbf{S}_{C_i}|}$.

Notons que les probabilités marginales et jointes $p(x)$, $p(y)$ et $p(x,y)$ sont calculées par rapport à un cluster, et les relations de corrélations sont évaluées par rapport à un cluster.

La mesure de MI telle que définie dans l'équation 3.9 ne possède pas de borne supérieure ni inférieure. À cet effet, nous allons normaliser cette mesure pour qu'elle ait des valeurs dans l'intervalle $[0,1]$. La valeur 1 correspond à une corrélation parfaite (co-occurrence complète), et la valeur 0, l'indépendance. La formule de l'information mutuelle normalisée (NMI) est présentée comme suit [13] :

$$NMI(X_\alpha, X_\beta) = \frac{MI(X_\alpha, X_\beta)}{-\sum_{x \in X_\alpha} \sum_{y \in X_\beta} p(x,y) \log_2 p(x,y)} \tag{3.10}$$

À ce stade, les relations de corrélation seront identifiées et extraites pour chaque cluster en observant les valeurs de NMI. Si $NMI(X_\alpha, X_\beta) \geq \pi$, où π est un seuil défini par l'utilisateur (nous avons choisi $\pi = 0.5$ dans notre travail), alors l'association entre les deux variables aléatoires X_α et X_β est jugée intéressante, et les deux patrons p_α et p_β possèdent une relation de corrélation dont nous ignorons la direction. L'avantage de cette étape est double. D'un coté, les relations de corrélations sont identifiées en choisissant celles dont la mesure NMI satisfait le seuil minimal. Cette étape permet d'élaguer les relations non intéressantes et de ne garder que les relations intéressantes qui peuvent éventuellement avoir des relations de causalité. Cela permet aussi de réduire l'espace de calcul lors de la découverte des relations causales. L'algorithme 2 présente la procédure de calcul des relations de corrélations entre les patrons telles que mentionnées précédemment.

Algorithme 2 Extraction des relations de corrélation

Entrée :

- Les clusters $C_1, C_2, ..., C_T$

Sortie :

- Les graphes des relations de corrélations $G_1, G_2, ..., G_T$

1 : - Initialiser : $G_1 = (\mathbf{V}_1, \mathbf{E}_1)$, $G_2 = (\mathbf{V}_2, \mathbf{E}_2)$,..., $G_T = (\mathbf{V}_T, \mathbf{E}_T)$
2 : **pour** $i = 1$ à T **faire** (pour chaque cluster)
3 : **pour chaque** paire de patrons (p_α, p_β) dans un cluster **faire**
4 : - calculer NMI à l'aide de la formule 3.10 ;
5 : **si** p_α et $p_\beta \notin \mathbf{V}_i$ et $NMI \geq \pi$
6 : - ajouter p_α, p_β à $\mathbf{V}_i$ et (p_α, p_β) à $\mathbf{E}_i$;
7 : **fin**
8 : **fin**
9 : - retourner G_1, $G_2, ..., G_T$;

Comme nous pouvons le constater, l'algorithme 2 retourne des structures graphiques de corrélations entre les patrons où les noeuds représentent les patrons et les arrêtes représentent les relations de corrélations, dont nous ignorons la direction. Une fois les patrons significatifs découverts et les relations de corrélations identifiées sous forme de structures graphiques, nous pouvons évaluer les relations causales entre ces différents patrons. La prochaine section présentera en détail le processus de découverte des relations causales.

3.4.7 Découverte des Relations Causales

La découverte des relations causales à partir des séquences présente un défi majeur. Les relations causales sont généralement évaluées entre des variables significatives qui possèdent des relations de corrélations. Cela veut dire, que la découverte des patrons significatifs dans notre travail et l'extraction des relations de corrélations constituent des étapes préparatoires pour cette ultime étape de la découverte des relations causales.

Comme nous l'avons mentionné auparavant, le MI est une mesure de la force d'association entre les patrons. Elle ne permet pas de fournir des informations sur

les relations causales et leurs directionnalités, si les relations entre les patrons sont des relations mutuelles, c'est à dire que chaque patron possède une force d'influence sur l'autre, ou bien des relations unidirectionnelles, c'est à dire, un patron influence l'autre (une relation dans un seul sens).

Les règles d'association sont des algorithmes qui sont largement utilisés dans la littérature pour analyser les patrons et mesurer éventuellement les forces d'association entre les patrons. Cependant, les règles d'association ne permettent pas de prendre en considération les informations séquentielles disponibles dans les données. Les règles d'association indiquent que les items s_1 et s_2 apparaissent généralement ensemble, mais n'indiquent pas que l'item s_2 apparaît toujours immédiatement après l'item s_1. De plus, les algorithmes des règles d'association ne sont pas capables de gérer le problème de bruit dans les données ni le problème de données manquantes. Par conséquent, les mesures de MI et les règles d'associations sont inadéquates pour la découverte des relations causales et l'influence entre les patrons. Toutes ces limitations nous motivent à faire recours à la théorie de la causalité qui est une théorie très puissante pour la découverte des relations causales et de la dynamique de l'information.

Il existe plusieurs mesures qui permettent de découvrir les relations causales comme indiquée précédemment. Dans ce travail, nous utilisons la mesure de transfert d'entropie [118]. La mesure de transfert d'entropie (TE), qui est une mesure asymétrique, possède beaucoup d'avantages par rapport à toutes les autres mesures de causalité. En effet, contrairement aux autres mesures, le TE prend en compte les informations partagées en incorporant l'historique commune des variables. Ceci peut être effectué en utilisant les probabilités conditionnelles de transition. Lungarella et al. [76] ont effectué une étude de comparaison entre les deux mesures, la mesure de Granger et le TE, et ils ont constaté que le TE est plus stable et plus précis que la mesure de Granger sur des données séquentielles temporelles. De plus, la mesure de TE ne requiert aucune hypothèse de départ sur les distributions des données contrairement à celle de Granger. L'emploi de la mesure de TE est donc plus indiquée pour la découverte des relations causales.

Dans notre méthode, nous appliquerons donc la mesure de TE uniquement entre les patrons significatifs possédant des relations de corrélation découvertes par les algo-

rithmes présentés précédemment. Cela réduira considérablement l'espace de recherche et par conséquent le temps de calcul.

Formellement, soient p_α et p_β deux patrons significatifs, et soient $\mathbf{X}_\alpha = \{x_1, x_2, ..., x_n\}$ et $\mathbf{X}_\beta = \{y_1, y_2, ..., y_m\}$ deux variables aléatoires pour les occurrences des événements qui composent les patrons p_α et p_β respectivement dans le même cluster. Le transfert d'entropie entre les deux variables aléatoires $\mathbf{X}_\alpha$ et $\mathbf{X}_\beta$ est calculé comme suit [118] :

$$TE_{\mathbf{X}_\beta \to \mathbf{X}_\alpha} = \sum_{x_{i+1}, x_i \in \mathbf{X}_\alpha, y_j \in \mathbf{X}_\beta} P(x_{i+1}, x_i, y_j) \log \frac{p(x_{i+1}|x_i, y_j)}{p(x_{i+1}|x_i)} \tag{3.11}$$

Puisque la mesure de TE est une mesure asymétrique, le calcul de la direction inverse est défini de la même façon comme suit :

$$TE_{\mathbf{X}_\alpha \to \mathbf{X}_\beta} = \sum_{y_{j+1}, y_j \in \mathbf{X}_\beta, x_i \in \mathbf{X}_\alpha} P(y_{j+1}, y_j, x_i) \log \frac{p(y_{j+1}|y_j, x_i)}{p(y_{j+1}|y_j)} \tag{3.12}$$

où x_{i+1} et y_{j+1} représentent les prochaines valeurs que peuvent prendre les deux variables aléatoires $\mathbf{X}_\alpha$ and $\mathbf{X}_\beta$ respectivement, et x_i and y_j représentent les valeurs passées (ou bien l'historique).

Etant donné que : $p(x_{i+1}|x_i, y_j) = \frac{p(x_{i+1}, x_i, y_j)}{p(x_i, y_j)}$, et $p(x_{i+1}|x_i) = \frac{p(x_{i+1}, x_i)}{p(x_i)}$, les équations 3.11 et 3.12, se réécrivent :

$$TE_{\mathbf{X}_\beta \to \mathbf{X}_\alpha} = \sum_{x_{i+1}, x_i \in \mathbf{X}_\alpha, y_j \in \mathbf{X}_\beta} P(x_{i+1}, x_i, y_j) \log \frac{p(x_{i+1}, x_i, y_j).p(x_i)}{p(x_i, y_j).p(x_{i+1}, x_i)} \tag{3.13}$$

$$TE_{\mathbf{X}_\alpha \to \mathbf{X}_\beta} = \sum_{y_{j+1}, y_j \in \mathbf{X}_\beta, x_i \in \mathbf{X}_\alpha} P(y_{j+1}, y_j, x_i) \log \frac{p(y_{j+1}, y_j, x_i).p(y_j)}{p(x_i, y_j).p(y_{j+1}, y_j)} \tag{3.14}$$

Selon les formules 3.13 et 3.14, la mesure de TE représente l'information sur la prochaine valeur x_{i+1} que peut prendre la variable aléatoire $\mathbf{X}_\alpha$, obtenue en utilisant les observations passées x_i et y_j simultanément, et en écartant l'information sur la prochaine valeur x_{i+1} que peut prendre la variable aléatoire $\mathbf{X}_\alpha$, obtenue en utilisant seulement les observations passées x_i. L'algorithme 3 présente la procédure de calcul

de la mesure TE et la découverte des graphes représentant des relations causales entre les patrons.

Algorithme 3 Découverte des graphes des relations causales

Entrée :

- Graphes des relations d'associations $G_1, G_2, ..., G_T$

Sortie :

- Graphes des relations causales $CG_1, CG_2, ..., CG_T$

1 : **pour** $i = 1$ à T **faire** (pour chaque graphe)
2 : **pour** chaque arrête (p_α, p_β) dans le graphe **faire**
3 : - calculer $TE_{\mathbf{X}_\alpha \to \mathbf{X}_\beta}$ à l'aide de la formule 3.14 ;
4 : - calculer $TE_{\mathbf{X}_\beta \to \mathbf{X}_\alpha}$ à l'aide de la formule 3.13 ;
5 : **si** $TE_{\mathbf{X}_\alpha \to \mathbf{X}_\beta} > TE_{\mathbf{X}_\beta \to \mathbf{X}_\alpha}$
6 : - remplacer l'arrête (p_α, p_β) par l'arc (p_α, p_β) ;
7 : **sinon si** $TE_{\mathbf{X}_\alpha \to \mathbf{X}_\beta} < TE_{\mathbf{X}_\beta \to \mathbf{X}_\alpha}$
8 : - remplacer l'arrête (p_α, p_β) par l'arc (p_β, p_α) ;
9 : **sinon si** $(TE_{\mathbf{X}_\alpha \to \mathbf{X}_\beta} = TE_{\mathbf{X}_\beta \to \mathbf{X}_\alpha})$ et $(TE_{\mathbf{X}_\alpha \to \mathbf{X}_\beta} \neq 0)$
10 : - remplacer l'arrête (p_α, p_β) par l'arc (p_α, p_β) et l'arc (p_β, p_α) ;
11 : **fin**
12 : **fin**
13 : - retourner les graphes de relations causales $CG_1, CG_2, ..., CG_T$;

Comme nous pouvons le constater à partir de l'algorithme 3, l'algorithme retourne T graphes, avec des noeuds représentant des patrons significatifs et des arcs représentant des relations causales entre ces patrons significatifs. Chaque arc est étiqueté par la valeur de TE entre les deux patrons significatifs représentant les deux extrémités de l'arc. Dans le cas où $TE_{\mathbf{X}_\alpha \to \mathbf{X}_\beta} = TE_{\mathbf{X}_\beta \to \mathbf{X}_\alpha}$, nous disons que les deux patrons possèdent une causalité réciproque ou (feedback) [119]. Dans ce cas, un patron cause l'autre et inversement. Ce phénomène motive l'utilisation des graphes de causalité. Contrairement aux graphes orientés acycliques, comme les réseaux Bayésiens, les graphes de causalité peuvent avoir des cycles et deviennent des graphes orientés cycliques. Dans ce cas, les graphes de causalité peuvent être considérés comme étant

une extension des réseaux Bayésiens.

Le phénomène de la causalité réciproque est supporté par des études dans la littérature. Par exemple, selon Spirtes et al. [119], dans les sciences cognitives contemporaines, certains modèles de calcul des humains suivent des graphes orientés cycliques, où les noeuds représentent des variables aléatoires qui peuvent prendre dans certaines situations des valeurs discrètes. Un autre exemple qui montre ce phénomène, est la découverte des relations bidirectionnelles entre la schizophrénie et l'épilepsie [141] comme mentionné précédemment. Les patients souffrant d'épilepsie sont susceptibles de développer la schizophrénie, et les patients souffrant de la schizophrénie sont susceptibles de développer une épilepsie. Tous ces exemples et cette théorie augmentent la fiabilité de notre modèle, de la représentation mathématique d'une part, et de sa concordance avec l'interprétation cognitive de l'autre.

Une fois les graphes des relations causales découverts, le profil comportemental de l'usager peut maintenant être construit en regroupant tous les graphes des relations causales. Ce profil peut par conséquent être utilisé dans différentes applications à savoir la prédiction des activités et l'identification des usagers parmi un groupe d'usagers. Dans la prochaine section, nous allons valider notre approche par différentes applications pour démontrer l'importance de notre approche dans les applications pratiques.

3.5 Validation

Dans cette section, nous allons décrire les expérimentations que nous avons effectuées pour valider notre approche en utilisant des données réelles issues des habitats intelligents réels (Domus, ISLab, CASAS, MIT) et d'autres données (données de GPS). Nous effectuons ces expérimentations pour répondre aux questions suivantes :

1. Les profils découverts par notre approche sont-ils consistants avec les données ?
2. Les profils découverts par notre approche sont-ils génériques et applicables dans différents domaines et différents types de bases de données ?
3. Les relations causales découvertes sont-elles sémantiquement et statistiquement expressives pour représenter, interpréter et distinguer les usagers et leurs acti-

vités ?

Si la réponse à ces question est positive, alors l'approche que nous avons proposée est utile, utilisable, et pratique. Nous avons effectué deux sortes de validation dans cette étude expérimentale.

- Validation subjective : dans cette validation, nous évaluons subjectivement les relations causales découvertes et comment ces relations sont représentatives du comportement usager.
- Validation objective : dans cette validation, nous évaluons notre approche de l'analyse causale pour deux applications importantes dans les habitats intelligents : la prédiction des activités et l'identification des usagers en se basant sur leur profil.

3.5.1 Jeux de données

Avant de présenter les expérimentations, nous devons tout d'abord présenter les données que nous avons utilisées afin de mener à bien nous expérimentations. Dans nos expérimentations, nous avons utilisé plusieurs jeux de données provenant de plusieurs habitats intelligents. Aux jeux de données utilisés dans la validation de notre modèle de reconnaissance d'activités dans le chapitre précédent, nous avons rajouté des données sur les modes de transport. Notre objectif est de s'assurer que notre modèle fonctionne quelque soit le type de capteurs utilisés et la complexité des activités réalisées. Cela nous permet aussi de valider la généricité de notre modèle. Le tableau 3.1 présente les détails de chaque ensemble de données utilisé dans notre étude expérimentale.

tableau 3.1 – Détails des données utilisées

Données	Nombre de séquences	Longueur min	Longueur max	Activities	Types de capteurs	Nombre des usagers	Période (jours)
Domus Série 1	58	100	470	Wake up, Bathe, Prepare breakfast, Have breakfast	Infrarouge, Electromagnétique, détecteur de pression, interuupteur, contact de portes et cabinets, Débitmètre	6	10
Domus Série 2	30	210	680	Wake up, Bathe, Prepare breakfast, Have breakfast, Prepare tea	Les même capteurs que Domus série 1	6	5
CASAS 1	501	19	1561	Wake up, Groom, Breakfast, Watch TV, Prepare dinner, Wash bathtub, Work at computer, Sleep, Prepare lunch, Clean, Work at dining room table	Infrarouge, capteurs RFID, Contact portes, Capteur eau chaude Capteur eau froide, Température, Électricité	2	90
CASAS 2	120	16	216	Make a phone call, Wash hands, Cook, Eat, Clean	Infrarouge, capteurs RFID, Contact portes, Capteur eau chaude Capteur eau froide, Température, Électricité Capteur de télephone	24	90
ISLab	23	16	140	Idle, Leave house, Use toilet Take shower, Sleep, Breakfast, Dinner, Drink	Capteurs de contcats, Température, Humidité	1	23
MITPlaceLab	506	2	185	Use toilet, Wash dishes, Prepare drink, Eat, Watch TV, Clean, Groom, etc.	Infra red, Electromagnetic, Pressure detector, Switch contacts, Door and closet contacts, Flow sensors, etc.	2	15
GeoLife	32	22	3168	Walk, Bike, Take car, Take bus, Take Train, Take subway, Take plane	GPS	32	730

La figure 3.5 présente les distributions des activités dans chaque base de données. Ces distributions nous donne une idée très claire sur la fréquence de chaque activité dans chaque base de données, et comment cela pourrait être observé lors de l'extraction des relations causales.

3.5.2 Les conditions d'experimentation

Dans cette section, nous allons discuter les conditions sous lesquelles se déroulent nos expérimentations et les hypothèses que nous avons émises.

Tout d'abord, pour des raisons de simplicité, nous avons utilisé $T = 2$ dans toutes nos expérimentations. Cela veut dire, que l'algorithme de clustering génère deux clusters chacun correspond à un profil usager pour une activité particulière. Notons que la détermination du nombre optimal de clusters de façon automatique n'est pas traité dans notre approche. Il sera considéré dans un travail futur.

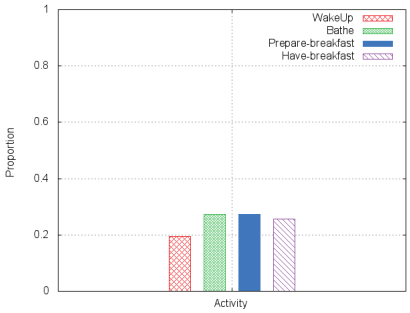
En ce qui concerne le seuil minimal d'optimisation de l'arbre probabiliste des suffixes Γ, nous avons choisi expérimentalement la valeur de $\Gamma = 0.4$. C'est à dire, si la valeur de similitude entre les deux distributions de probabilité dépasse ce seuil, alors les deux distributions sont similaires. Bien entendu, la valeur de Γ dépend de la nature des données comme la longueur des séquences, la fréquence des items dans les séquences.

Comme nous l'avons mentionné dans la section 3.4.6 relative aux relations de corrélations, nous avons fixé le seuil $\pi = 0.5$. C'est à dire, que les patrons qui ont une valeur de corrélation supérieure à 0.5 seront considérés corrélés, sinon, ils ne sont pas corrélés.

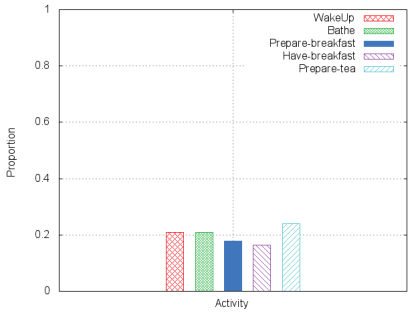
Finalement, toutes les bases de données que nous avons utilisées sont annotées, c'est à dire, les activités dans chaque séquence sont connues. La figure 3.6 présente un exemple des activités annotées dans une séquence ainsi que les patrons significatifs d'une activité.

3.5.3 Validation subjective

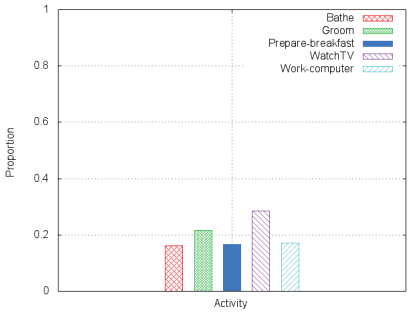
Pour répondre aux deux premiers critères de validation, cette section consiste à évaluer la capacité de notre approche à construire des profils comportementaux



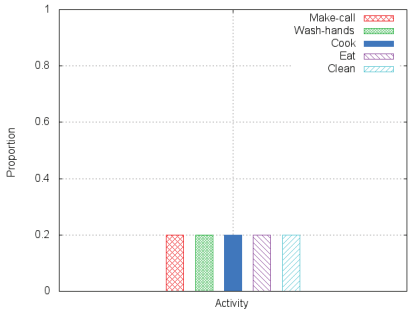
(a) Domus 1



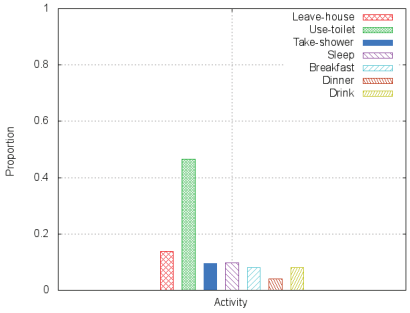
(b) Domus 2



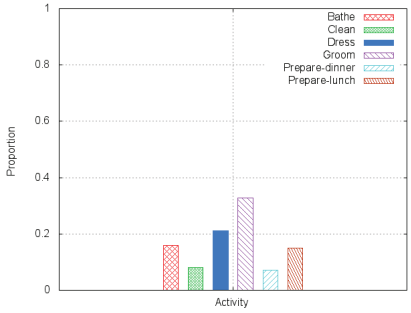
(c) CASAS 1



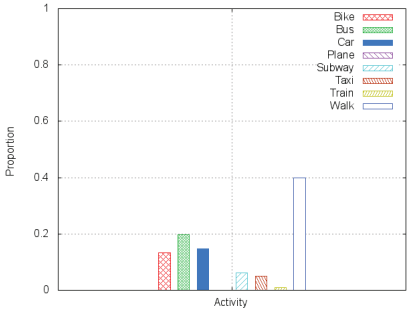
(d) CASAS 2



(e) ISLab



(f) MITPlaceLab



(g) GeoLife

figure 3.5 – Distributions des activities dans chaque base de données

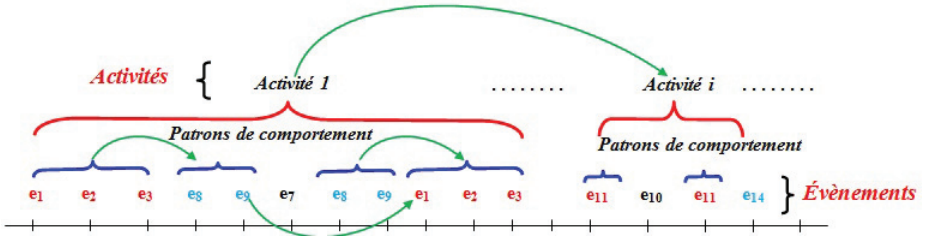


figure 3.6 – Exemple d'une séquence annotée

des usagers dans les habitats intelligents, et à extraire des profils communs pour un groupe d'usagers à l'extérieur des habitats intelligents. L'importance de construire des profils individuels est de pouvoir dresser un plan des différents comportements de l'usager, ce qui permettra d'un coté de les assister en adaptant et personnalisant les services appropriés, et d'un autre coté prédire les prochains comportements pour éviter d'éventuelles situations anormales. En ce qui concerne l'extraction du profil commun pour un groupe d'usagers, cela nous permet de découvrir les comportements impliqués lors de la réalisation des activités et ensuite de bâtir un modèle de ces activités en se basant sur ces profils.

Construction du profil usager dans les habitats intelligents

Dans cette section, nous allons évaluer la capacité de notre approche à construire des profils usagers dans les habitats intelligents. Pour ce faire, nous allons utiliser uniquement les bases de données issues des habitats intelligents. Dans ces bases de données, les usagers accomplissent leurs activités de la vie quotidienne. Notons que les profils sont construits pour chaque activité, et chaque usager est traité séparément. Dans ce qui suit et pour des contraintes d'espace, nous allons présenter des exemples de relations causales extraites pour quelques activités dans une seule base de données (ISLab par exemple). La figure 3.7 présente un exemple de relations causales extraites pour quelques activités dans la base de données ISLab. Les figures (3.7(a)) et (3.7(b)) représentent deux comportements de l'usager lors de la réalisation des activités 'Préparer une collation' (Get drink), (3.7(c)) et (3.7(d)) pour l'activité 'Préparer le dinner' (Prepare dinner), tandis que (3.7(e)) et (3.7(f)) représentent deux comportements de l'usager lors de la réalisation de l'activité 'Utiliser la salle de bain' (Use bathroom).

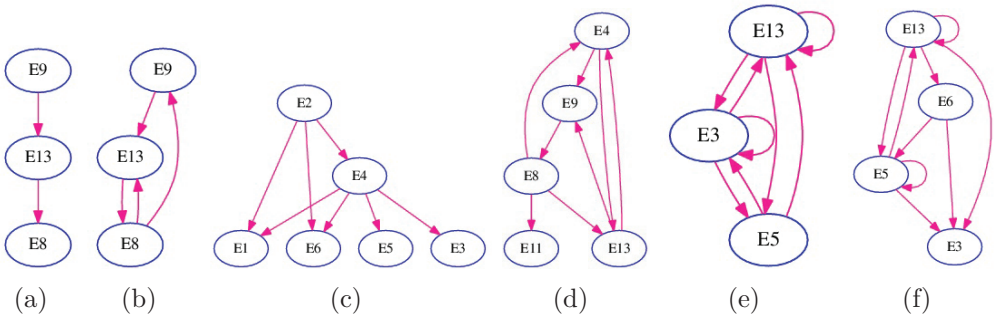


figure 3.7 – Exemple de graphes de relations causales pour certaines activités dans la base de données ISLab.

Comme nous pouvons le constater à partir de la figure 3.7, les relations causales découvertes à partir des deux clusters peuvent être différentes pour chaque activité. Notons que les noeuds dans les graphes de relations causales représentent des états de capteurs disséminés dans l'habitat intelligent. Par exemple, dans les graphes de relations causales de l'activité 'Préparer une collation', la relation causale ($E8 \rightarrow E9$) apparaît explicitement dans le graphe 3.7(a), mais elle n'apparaît pas explicitement dans le graphe 3.7(b). La même observation pour la relation causale ($E8 \rightarrow E13$). La différence observée dans ces graphes démontre la variabilité du comportement usager lors de la réalisation de ses activités, où l'usager accomplit la même activité de différentes manières. De plus, pour 19 instances collectées pour l'activité 'Préparer une collation', 15 instances partagent les relations causales présentées dans le graphe 3.7(a), et uniquement 4 instances partagent les relations causales présentées dans le graphe 3.7(b). Cela signifie que le graphe des relations causales présenté dans la figure 3.7(a) représente un comportement usuel et commun de l'usager, par contre le graphe présenté dans la figure 3.7(b) représente un comportement moins usuel. Dans le même sens, pour 24 instances recueillies pour l'activité 'Utiliser la salle de bain', 18 instances partagent les relations causales présentées dans le graphe 3.7(e), et uniquement 6 instances partagent les relations causales présentées dans le graphe 3.7(d). Ces constatations pourraient être très intéressantes dans l'optimisation du nombre de capteurs dans l'habitat intelligent ainsi que leur déploiement.

Les relations causales réciproques observées dans les graphes 3.7(b) et 3.7(d) signifient que l'activation d'un capteur entraîne l'activation de l'autre et vice versa. Le

potentiel de ce genre de relations est de pouvoir prédire le prochain capteur qui va être activé. Cette relation causale est assez importante pour les systèmes d'assistance dans les habitats intelligents afin d'adapter les services nécessaires selon le comportement prédit.

Dans les exemples que nous avons présentés, les noeuds des graphes représentent des patrons significatifs simples de longueur 1, c'est à dire des états de capteurs. Dans notre approche, nous avons également des graphes de relations causales entre des patrons plus complexes et plus longs. Par exemple, la figure 3.8 présente un exemple de graphes de relations causales pour certaines activités dans différentes bases de données.

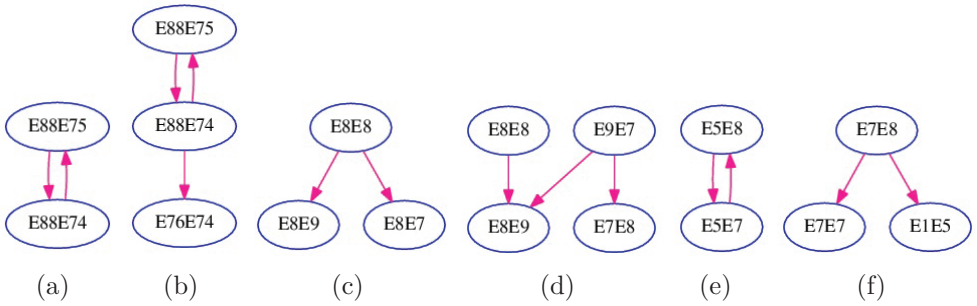


figure 3.8 – Exemple de graphes de relations causales avec des patrons plus complexes pour certaines activités.

Les figures (3.8(a)) et (3.8(b)) représentent deux comportements d'un usager lors de la réalisation de l'activité 'Se laver les mains' (Wash hands) dans la base de données CASAS 2, (3.8(c)) et (3.8(d)) pour l'activité 'Préparer le petit déjeuner' (Prepare breakfast) dans la base de données ISLab, tandis que (3.8(e)) et (3.8(f)) représentent deux comportements d'un usager lors de la réalisation de l'activité 'Utiliser la salle de bain' (Use bathroom) dans la base de données Domus 1.

Comme indiqué dans la figure 3.8, les noeuds dans les graphes représentent des patrons de longueur plus grande que 1. Cela démontre que notre approche est fonctionnelle quelque soit la longueur de patron. Notons que les activités dans les bases de données que nous avons utilisées sont des activités simples, et il en résulte des patrons de longueur petite. Les graphes construits sont simples ce qui facilite grandement leur utilisation dans différentes applications faisant appel à des recherches

dans les graphes et de comparaison entre les graphes. Ces applications peuvent être exécutées dans un temps raisonnable.

Nous avons présenté dans les figures 3.7 et 3.8 des relations causales entre des patrons d'activités. Notre approche possède également le potentiel de découvrir des relations causales entre les activités comme indiqué dans la figure 3.6. Dans ce cas, nous traitons des séquences d'activités et non pas des séquences d'événements. C'est à dire, l'activité de haut niveau est traité comme un événement. Découvrir des relations causales entre les activités joue un rôle très important dans les habitats intelligents, et plus spécifiquement dans les systèmes d'assistance. En effet, une relation causale entre deux activités veut dire qu'une activité cause l'apparition de l'autre. Cela nous permettra de prédire la prochaine activité à être réalisée par l'usager dans l'habitat intelligent. Ceci nous facilitera grandement la tâche d'adaptation et de personnalisation des services appropriés afin de fournir une assistance adéquate selon l'activité prévue. La figure 3.9 présente un exemple de relations causales entre quelques activités dans la base de données ISLab.

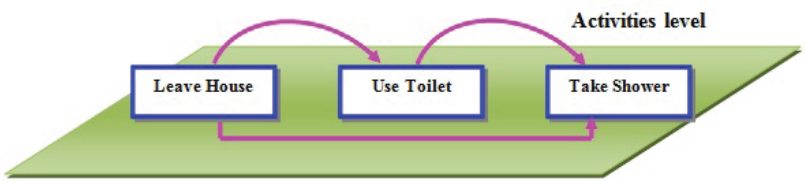


figure 3.9 – Exemple de relations causales entre des activités dans la base de données ISLab.

Ces relations causales reflètent la réalité de la réalisation des activités par l'usager dans la base de données ISLab. En effet, l'activité 'Quitter la maison' (Leave house) cause les deux activités 'Utiliser la toilette' (Use toilet) et l'activité 'Prendre une douche' (Take shower). Ce comportement est observé toujours lorsque l'usager, après avoir quitté la maison, il revient à la maison, il utilise la toilette puis il prend sa douche. Cette information est importante et permet par exemple de préparer la salle de bain dès le retour de l'usager à la maison.

Construction du profil usager à l'extérieur des habitats intelligents

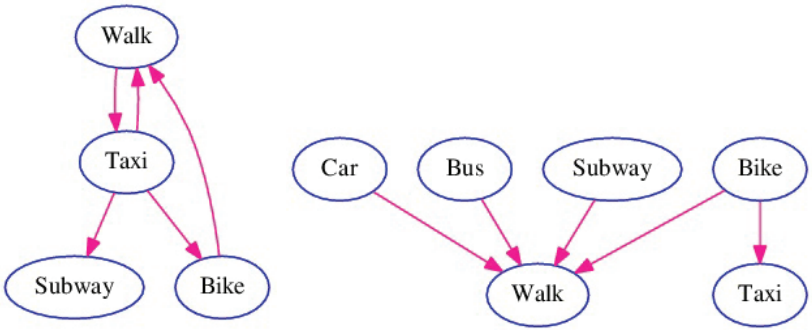
Dans cette section, nous allons évaluer la capacité de notre approche à construire des profils usagers à l'extérieur des habitats intelligents. Pour ce faire, nous allons utiliser la base de données GeoLife [149].

La construction du profil usager à l'extérieur des habitats intelligents est motivée par le fait que plusieurs personnes nécessitent d'être assistées à l'extérieur, particulièrement dans les voyages, magasinages, loisirs, et ainsi de suite. Cette assistance est beaucoup plus compliquée et pose beaucoup de défis étant donné les paramètres de l'environnement extérieur et leur variabilité. L'assistance des personnes à l'extérieur requiert des systèmes de navigation utilisables à l'extérieur comme par exemple le système de navigation GPS. Ces systèmes collectent des informations pertinentes à savoir l'emplacement de la personne et sa localisation. À l'aide de ces informations, nous serons capables de fournir l'aide à la personne en proposant la direction à prendre ou bien les stations de bus ou de trains les plus proches selon sa position. Dans la base de données GeoLife, les séquences collectées pour chaque usager représentent des séquences des modes de transport utilisés par l'usager, dont une exemple est fourni dans le tableau 3.2.

tableau 3.2 – Exemple de séquence d'activités des modes de transport dans la base de donnée GeoLife

De		Vers		Activité
Date	Temps	Date	Temps	
2007/8/4	14 :28 :36	2007/8/4	14 :29 :41	walk
2007/8/4	14 :29 :42	2007/8/4	14 :49 :49	bus
2007/8/4	14 :49 :50	2007/8/4	14 :55 :02	walk
2007/8/4	16 :20 :31	2007/8/4	16 :46 :45	subway
2007/8/4	16 :51 :35	2007/8/4	16 :52 :41	walk
2007/8/4	16 :52 :42	2007/8/4	17 :06 :06	Car
2007/8/4	17 :06 :07	2007/8/4	18 :11 :46	walk
2007/8/4	18 :11 :47	2007/8/4	18 :48 :20	taxi
2007/8/4	19 :53 :03	2007/8/4	21 :11 :29	walk

Dans les séquences collectées pour chaque usager, chaque mode de transport est traité comme un événement. La figure 3.10 présente un exemple de profil de l'usager sur l'utilisation des transports.



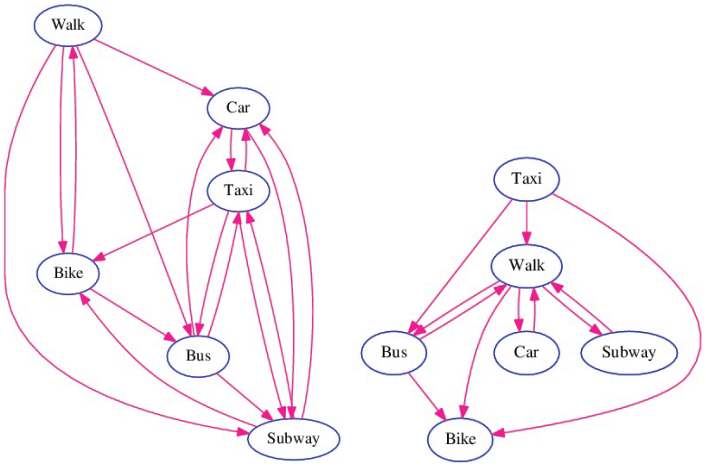
(a) Premier profil (b) Deuxième profil

figure 3.10 – Exemple de profil d'un usager lors de l'utilisation de transport

Selon la figure 3.10, l'usager possède deux comportements différents lors de l'utilisation des modes de transport. Par exemple, dans le graphe 3.10(a), uniquement quatre modes de transport sont impliqués dans les relations causales ce qui résume le premier comportement. Par contre, deux autres modes de transport (Car) et (Bus) sont impliqués dans le graphe 3.10(b). Le deuxième profil présenté dans le graphe 3.10(b) est différent de celui présenté dans le graphe 3.10(a). Cela explique comment le comportement usager est variable et peut changer pour le même usager.

Une des applications importantes à la fois dans le domaine des habitats intelligents et dans d'autres domaines, comme les réseaux sociaux, les systèmes de recommandation, et les applications du commerce électronique, est la découverte des profils communs et partagés par un groupe d'usagers. Dans ce cas de figure, le profil partagé peut exprimer des préférences et intérêts partagés par des usagers. Cela permettra d'effectuer des recommandations à un groupe d'usagers et de pouvoir personnaliser des services selon leurs préférences. Notre approche est capable de construire des profils communs à un groupe d'usagers. À cet effet, nous avons regroupé toutes les séquences de tous les usagers ensemble dans la base de données GeoLife afin de construire deux profils partagés par ces usagers. La figure 3.11 présente les graphes de relations causales représentant deux profils communs différents pour tous les usagers de la base de

données GeoLife.



(a) Profil 1 (b) Profil 2

figure 3.11 – Examples of causal graphs representing common profiles for all users discovered from the GeoLife dataset.

Comme nous pouvons le constater dans la figure 3.11, les modes de transport les moins utilisés ne figurent pas dans les graphes de relations causales comme le train (train) et l'avion 'plane', qui possèdent les valeurs les plus petites dans la distributions des probabilités comme présenté dans la figure 3.5(g) comparativement aux autres modes de transport. L'avantage de la découverte des relations causales entre les modes de transport est de pouvoir prédire les embouteillages et de proposer, en conséquence, des services personnalisés aux usagers.

L'un des avantages majeurs de l'utilisation du GPS, est de fournir des informations pertinentes et riches sur le contexte géographique des usagers. Ceci permet la détection de la localisation des usagers pour fournir une aide appropriée selon l'emplacement des usagers et de leurs besoins. La figure 3.12 présente des relations causales entre activités représentant des modes de transport enrichies par des informations géographiques. Ces relations causales peuvent être utilisées pour interpréter, analyser et résoudre beaucoup de problèmes liés à des situations pratiques dans la gestion des embouteillages. À titre d'exemple, si nous prenons connaissance de la relation causale (Subway $\rightarrow$ Taxi), alors il serait intéressant de proposer aux usagers une autre route plus rapide pour les voitures afin d'éviter la congestion lors de la sortie des personnes

du (Subway), particulièrement durant les heures de pointes. Cela permettra d'un coté de gagner du temps pour les personnes en évitant la circulation, et d'éviter d'éventuels accidents de l'autre. Selon nos connaissances, la découverte des relations causales entre les différents modes de transport est nouvelle. Cela constitue une contribution importante dans ce domaine.

(a) Région 1

(b) Région 2

figure 3.12 – Exemple de relations causales entre les modes de transport avec des information géographiques dans quelques régions de Pékin.

3.5.4 Validation objective

Pour répondre au troisième critère de validation, cette section explique comment notre approche de causalité aide à développer des algorithmes simples pour la prédiction des activités et l'identification des usagers. Bien que les deux problématiques soient assez complexes dans la littérature, nous allons démontrer comment notre approche permet de développer des algorithmes permettant de résoudre ces deux problématiques. Nous avons choisi ces deux applications vu leur importance dans les habitats intelligents où la prédiction des activités est une tâche nécessaire dans tout système d'assistance.

Prédiction des activités des usagers

La découverte des activités causales joue un rôle fondamental dans les systèmes de personnalisation et d'adaptation de services, où la prédiction de la prochaine activité est de grande importance pour améliorer la satisfaction des usagers et d'assurer leur bien être. Dans ces expérimentations, nous allons démontrer comment notre approche aide à développer un algorithme simple pour la prédiction des activités. Notre algorithme de prédiction des activités est résumé dans les points suivants :

- Les données de l'algorithme sont les séquences d'activités pour chaque usager.
- Pour effectuer la prédiction des activités pour un usager, nous devons tout d'abord construire les profils usagers pour chaque activité. Notons que les activités sont annotées dans les bases de données utilisées. Les profils construits pour chaque activité représentent des relations causales entre les patrons significatifs découverts dans chaque activité.
- Pour une séquence donnée de test , pour prédire l'activité au t+1, nous cherchons les patrons significatifs découverts dans la séquence de test jusqu'à l'instant t.
- Nous comparons ensuite les patrons identifiés dans la séquence de test avec les profils de l'usager construits pour chaque activité. Nous choisissons donc les graphes de relations causales qui correspondent le mieux aux patrons identifiés. Si plusieurs correspondances existent pour un patron p_α, par exemple : $p_\alpha \rightarrow p_\beta$ et $p_\alpha \rightarrow p_\gamma$, à ce moment là, nous allons choisir la relation causale avec la plus grande valeur de transfert d'entropie.

Ce processus est répété pour chaque usager dans la base de données, et le résultat de prédiction est la moyenne des résultats pour chaque usager.

Puisque notre approche permet de prédire la prochaine activité sachant les activités précédentes, par conséquent, pour que la comparaison soit la plus objective possible avec des approches existantes, il faut que ces approches adoptent le même principe à savoir les modèles Markoviens et les modèles Bayésiens. Nous avons comparé notre approche avec le modèle HMM [128], le modèle CRF [128], le réseau Bayésien (BN) [102], le modèle CRF hiérarchique (HCRF) [73], le modèle dynamique latent CRF (LDCRF) [86], et le modèle Bayésien naif (NB) [124].

Nous avons utilisé la validation croisée pour évaluer la performance de notre approche. Nous avons également utilisé la métrique F-mesure ($F - mesure = 2 \times \frac{Précision \times Rappel}{Précision + Rappel}$) pour évaluer la performance de notre approche. Le tableau 3.3 présente la moyenne des résultats de prédiction obtenus dans toutes les bases de données.

tableau 3.3 – Résultats de prédiction obtenus dans toutes les bases de données

Approches	Bases de données					
	DOMUS 1	DOMUS 2	CASAS 1	CASAS 2	ISLab	MITPlaceLab
Notre approche	0.791	0.775	0.733	0.638	0.925	0.75
HMM	0.568	0.54	0.3	0.2	0.197	0.667
CRF	0.638	0.554	0.168	0.4	0.23	0.6
CRF hiérarchique	0.4	0.5	0.6	0.5	0.6	0.5
BN	0.506	0.306	0.813	0.3	0.854	0.845
NB	0.398	0.248	0.792	0.391	0.476	0.831
Latent Dynamic CRF	0.75	0.5	0.65	0.5	0.6	0.5

Comme nous pouvons l'observer à partir du tableau 3.3, notre approche dépasse significativement les approches existantes pour la prédiction des activités. Les autres approches donnent des meilleurs résultats uniquement dans deux bases de données : la base de donnée ISLab (avec F-mesure ≥ 0.6), et la base de données CASAS 1 (avec F-mesure ≥ 0.79) avec respectivement le modèle des réseaux Bayésiens et le modèle Bayésien naif. En fait, dans la base de données CASAS 1, chaque activité possède plusieurs assistances, ce qui rend les résultats de prédictions meilleurs. En revanche, comme mentionné dans le tableau 3.3, toutes les approches sont peu performantes y compris notre approche en utilisant la base de données CASAS 2. Plus spécifiquement,

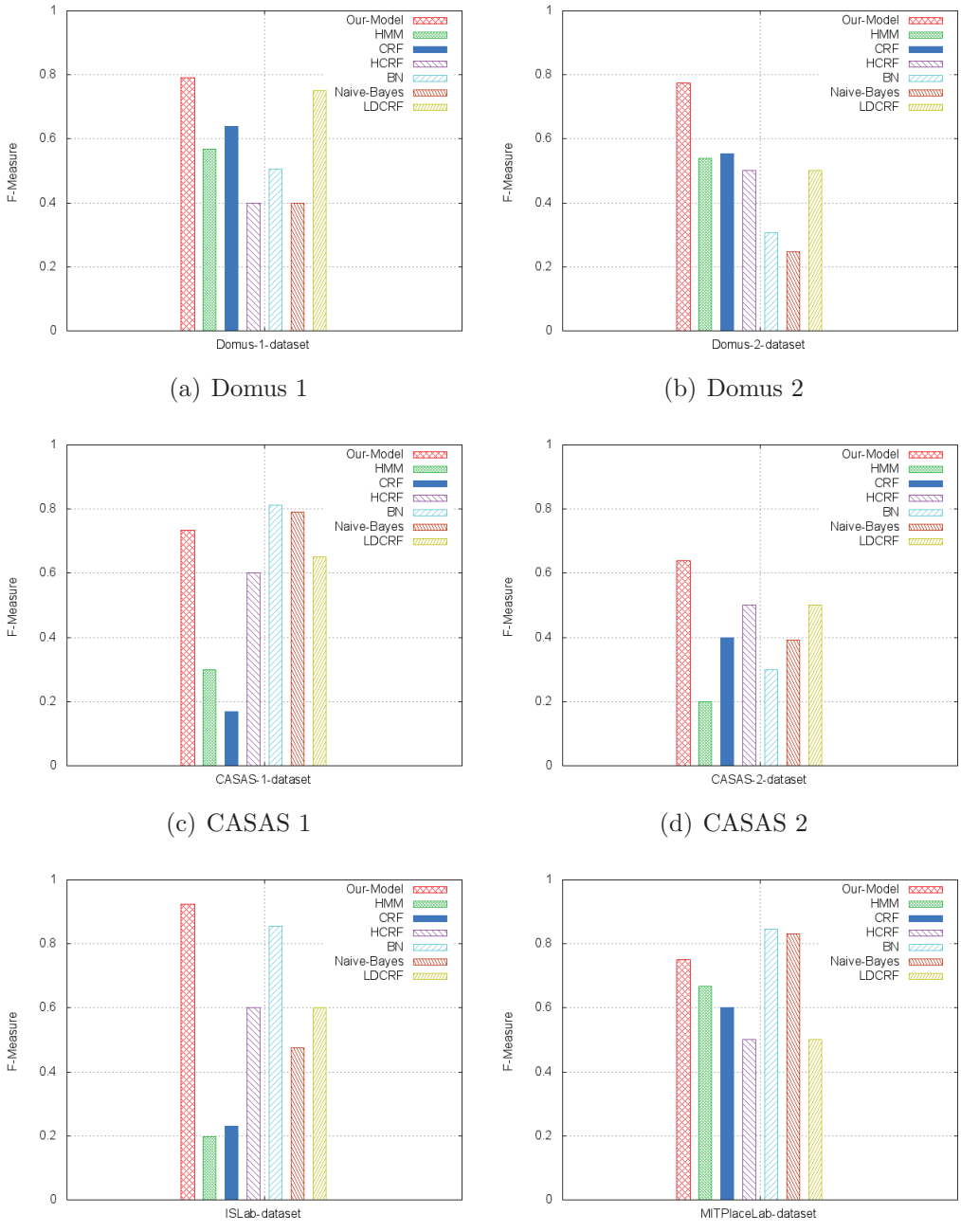
les approches HMM, le réseau Bayésien et le Bayésien naif. Cela peut être expliqué par le petit nombre d'instances pour chaque activité dans cette base de données. En effet, comme nous l'avons décrit dans la partie de description des bases de données utilisées, dans cette base de données, chaque usager accomplit cinq activités une seule fois, contrairement aux autres bases de données où les usagers accomplissent plusieurs fois la même activité, ce qui permet de mieux comprendre le comportement des usagers lors de la réalisation de leurs activités.

Dans le cas de la base de données ISLab, chaque activité possède ses propres capteurs, ce qui rend facile la distinction entre les différentes activités, contrairement aux autres bases de données où les activités peuvent partager plusieurs capteurs et objets ce qui rend difficile le processus de prédiction. Par conséquent, les résultats obtenus démontrent l'importance du profil usager et comment il peut être utilisé pour représenter et expliquer les comportements usagers. La figure 3.13 présente graphiquement les résultats de prédiction obtenus dans toutes les bases de données.

Le potentiel de notre approche dans la prédiction des activités est la capacité de planifier les prochaines activités afin d'adapter les services appropriés en conséquence. Par exemple, si l'activité 'Se réveiller' (Wake up) cause l'activité 'Prendre une douche' (Take shower), cela pourrait être très important dans un système d'assistance dans les habitats intelligents de préparer la salle de bain pour l'usager en allumant le chauffage par exemple. De plus, la découverte des relations causales entre les patrons significatifs dans la même activité a une importance majeure dans l'assistance des usagers à accomplir correctement leurs activités de la vie quotidienne et de pouvoir détecter des situations anormales plus particulièrement pour les usagers présentant des déficits cognitifs.

Identification des usagers

L'une des applications importantes dans les habitats intelligents est l'identification des usagers dans un environnement multi usagers. Le but de cette application est d'arriver à distinguer entre les usagers afin de pouvoir fournir l'aide adéquate à un usager particulier parmi plusieurs. Dans cette expérimentation, nous allons présenter comment notre approche permet de développer un algorithme simple pour l'identification des usagers dans un environnement multi usagers.



(a) Domus 1

(b) Domus 2

(c) CASAS 1

(d) CASAS 2

(e) ISLab

(f) MITPlaceLab

figure 3.13 – Comparaison des résultats de prédiction obtenus dans toutes les bases de données

L'identification des usagers est une problématique qui a été abordée dans la littérature en utilisant différentes approches. Nous pouvons citer par exemple l'approche basée sur l'image [1, 19], l'approche biométrique [59], l'approche non invasive [61, 26], et autres. Cependant, dans les environnements intelligents où l'intimité des gens est une priorité fondamentale, l'utilisation des caméras est peu recommandée. À cet effet, notre approche est non invasive et permet de distinguer entre les usagers sans avoir nécessairement recours aux caméras.

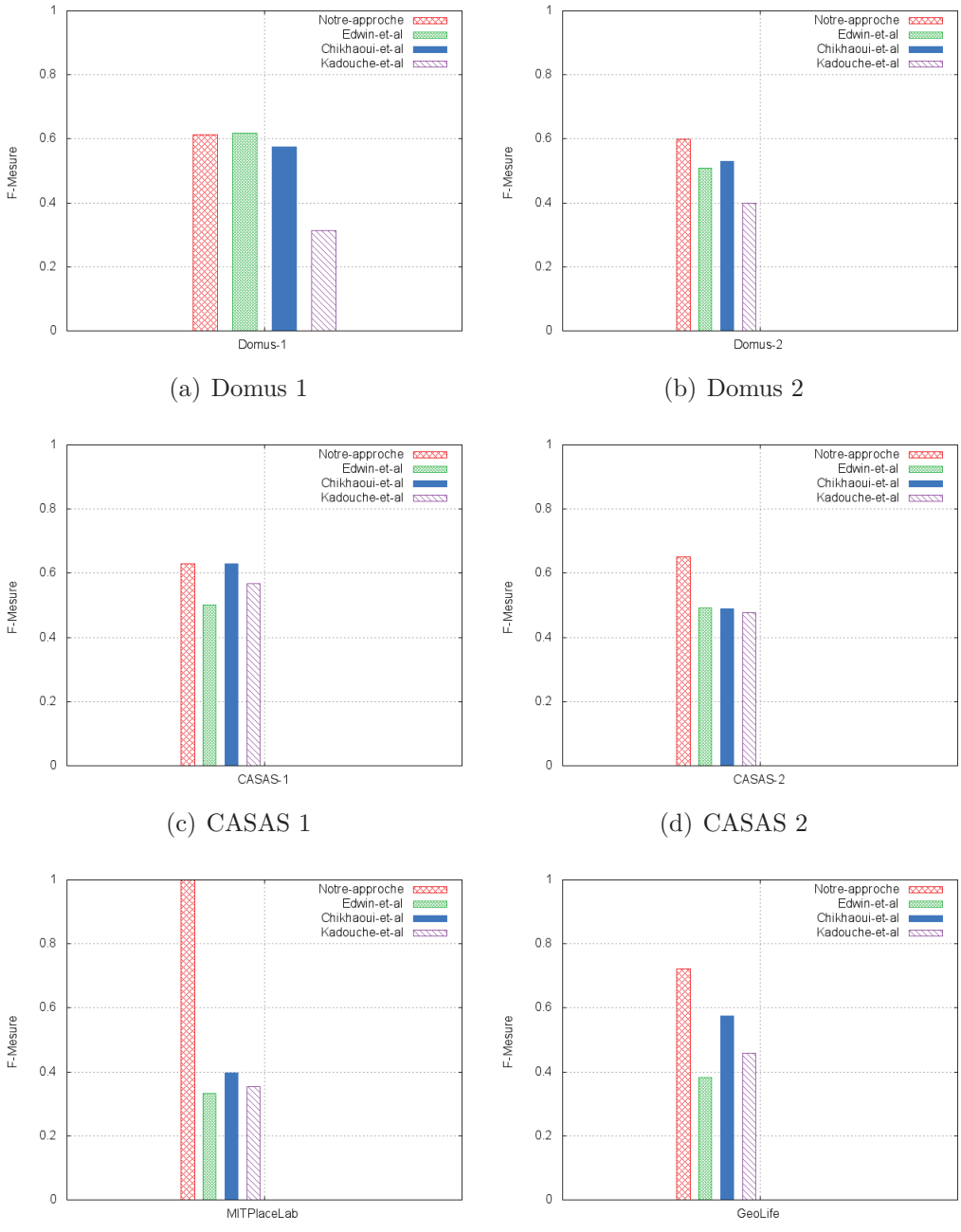
Notre approche d'identification des usagers se base sur les profils découverts pour chaque usager lors de la réalisation de ses activités de la vie quotidienne. Par conséquent, le concept clé dans l'identification des usagers est le profil. Notre algorithme d'identification est présenté dans les points suivants :

- Les données de l'algorithme sont les séquences d'activités pour chaque usager.
- Pour identifier un usager, nous aurons besoin tout d'abord de construire les profils de cet usager selon notre approche décrite précédemment. Les profils construits représentent des relations causales entre les patrons significatifs.
- Pour une séquence donnée de test, nous cherchons les patrons significatifs découverts dans la séquence de test.
- Nous comparons ensuite les patrons identifiés dans la séquence de test avec les profils usager construits. Nous choisissons donc les graphes de relations causales qui correspondent le mieux aux patrons identifiés. Lorsque plusieurs correspondances existent pour un patron p_α, par exemple : $p_\alpha \rightarrow p_\beta$ et $p_\alpha \rightarrow p_\gamma$, nous choisissons la relation causale avec la plus grande valeur de transfert d'entropie.

Nous avons comparé notre approche d'identification des usagers avec trois approches existantes à savoir l'approche de Edwin et al. [56], l'approche de Kadouche et al. [61] et l'approche de Chikhaoui et al. [26]. Nous avons utilisé la validation croisée pour évaluer la performance de notre approche. Nous avons également utilisé la métrique F-mesure pour évaluer la performance de notre approche. Le tableau 3.4 présente la moyenne des résultats d'identification obtenus dans toutes les bases de données.

La figure 3.14 présente graphiquement les résultats d'identification obtenus dans toutes les bases de données.

Comme nous pouvons l'observer à partir du tableau 3.4, notre approche donne



(a) Domus 1

(b) Domus 2

(c) CASAS 1

(d) CASAS 2

(e) MITPlaceLab

(f) GeoLife

figure 3.14 – Comparison des résultats d'identification des usagers dans toutes les bases de données.

tableau 3.4 – Résultats d'identification des usagers dans toutes les bases de données.

Approches	Datasets					
	Domus 1	Domus 2	CASAS 1	CASAS 2	MITPlaceLab	GeoLife
Notre approche	0.612	**0.598**	0.629	**0.65**	**1**	**0.723**
Edwin et al. [56]	**0.618**	0.508	0.5	0.491	0.333	0.383
Chikhaoui et al. [26]	0.574	0.53	**0.63**	0.489	0.397	0.574
Kadouche et al. [61]	0.314	0.4	0.567	0.478	0.354	0.458

de meilleurs résultats d'identification d'usagers en se basant sur leurs comportements comparativement aux autres approches. En effet, notre approche prend en compte les relations, exprimées sous formes de relations causales, entre les différents patrons significatifs, comparativement aux autres approches où les relations entre les variables ne sont pas explicitement identifiées.

D'un point de vue théorique, les modèles de relations causales sont beaucoup plus appropriés pour la tâche d'identification des usagers par rapport aux autres modèles comme les réseaux Bayésiens et les modèles Markoviens. En effet, les modèles Markoviens ou les réseaux Bayésiens sont basés sur les probabilités conditionnelles qui sont calculées sous l'hypothèse d'indépendance entre les variables. Contrairement à ces modèles, il n'y a pas d'hypothèses utilisées dans notre approche. De plus, les autres modèles sont basés sur les séquences d'événements qui peuvent être bruitées, ce qui influe sur les résultats d'identification. Notre approche emploie le concept de patron significatif, ce qui représente une autre façon de représenter le comportement usager. Cette nouvelle façon de représenter le comportement usager, permet de réduire significativement la quantité du bruit dans les données. Finalement, le concept de bidirectionnalité observé dans le modèle de relations causales développé pour chaque usager pourrait ajouter des nouvelles connaissances sur les comportements des usagers et sur les relations entre ces comportements, ce qui n'est pas le cas pour les autres modèles où les relations sont exprimées uniquement dans un seul sens.

3.6 Discussion

Dans la section précédente, nous avons concrétisé notre modèle proposé par des expérimentations et comparaisons avec les méthodes existantes. Nous avons analysé

les résultats obtenus selon différents points de vues. Nous nous sommes concentrés particulièrement aux interprétations statistiques des relations entres les patrons significatifs. Nous avons évalué la performance de notre modèle en utilisant différentes bases de données avec des activités de différents niveaux de complexité. Nous avons également utilisé différents types de bases de données collectées à l'intérieur et à l'extérieur des habitats intelligents afin d'évaluer la performance de notre modèle dans différentes situations. Dans cette section, nous allons discuter les étapes abordées dans notre approche, les limites rencontrées ainsi que les perspectives d'amélioration et les pistes de solutions qui pourraient être envisagées afin de surmonter ces limites.

De façon générale, notre modèle est plus performant que les autres modèles comme nous l'avons démontré expérimentalement dans la section précédente. En outre, nous avons pu constater l'importance des arbres probabilistes des suffixes dans la découvertes des patrons significatifs et leurs statistiques. Les performances des arbres probabilistes des suffixes se trouvent à la fois au niveau de la découverte des patrons et au niveau de la facilité des calculs probabilistes impliqués lors de l'extraction des patrons.

Pour optimiser la profondeur de l'arbre probabiliste des suffixes, nous avons introduit une nouvelle méthode permettant de calculer la divergence entre les distributions de probabilités des noeuds. Cette méthode est basée sur la mesure de KL. Notre méthode constitue également une contribution importante qui mérite d'être mise en évidence. Contrairement aux méthodes existantes dans lesquelles un seul de probabilité ou de fréquence est déterminé à l'avance, notre méthode calcule tout d'abord la divergence entre les distributions, ensuite, elle fixe un seuil en se basant sur les résultats de la mesure KL. Ceci dit, il serait intéressant d'incorporer un nouveau mécanisme permettant de calculer le seuil d'optimisation de façon automatique. Notons que dans la littérature, tous les algorithmes basés sur les arbres probabilistes des suffixes utilisent des seuils spécifiés par les utilisateurs. L'incorporation d'un tel mécanisme permettrait d'éliminer définitivement l'intervention de l'utilisateur dans l'optimisation de l'arbre probabiliste des suffixes.

L'une des contributions importantes de notre approche est l'algorithme de clustering que nous avons développé. En effet, notre algorithme de clustering se base sur les patrons significatifs découverts à l'aide de l'arbre probabiliste des suffixes. De plus,

notre algorithme sélectionne uniquement les patrons significatifs les plus longs afin de réduire la complexité de l'algorithme. La complexité de notre algorithme de clustering dans le pire cas est de l'ordre de $O(2|\mathcal{D}|^2)$. Après avoir construit l'arbre probabiliste des suffixes dans les meilleurs cas cela prend un temps de $O(nlogn)$, où n est la longueur de la séquence. Il y a lieu de mentionner que tout algorithme de clustering des séquences pourrait être utilisé à ce stade comme les algorithmes proposés par [140, 143, 83], et le développement d'un nouvel algorithme de clustering n'est pas une priorité importante pour répondre à nos objectifs fixés au départ. Notre objectif avec le développement de notre algorithme de clustering est de tirer profit des avantages de l'arbre probabiliste des suffixes, et de proposer un nouvel algorithme de clustering permettant de catégoriser les séquences tout en minimisant la complexité de l'algorithme et facilitant l'étude de la corrélation entre les patrons. Dans notre algorithme de clustering, nous avons utilisé les plus longs patrons non mutuellement inclusifs. Selon les bases de données que nous avons utilisées, nous avons fixé un seuil minimal égale à 2 pour la longueur des patrons. En outre, plus la longueur des patrons est grande, plus la similarité entre les séquences aura de sens, et plus petit sera le nombre de patrons. Cela aidera à réduire le nombre de patrons impliqués dans le processus de clustering, et permettra également de réduire la complexité de l'algorithme de clustering. Ceci dit il serait intéressant d'incorporer un nouveau mécanisme permettant de calculer la longueur optimale des patrons avec laquelle nous pouvons effectuer le clustering sans fixer une longueur particulière des patrons.

L'une des étapes requises pour étudier la causalité consiste à analyser la corrélation entre les variables. Dans notre approche, la corrélation entre les patrons est étudiée en adaptant la mesure de l'information mutuelle. À l'aide de cette mesure, nous avons pu extraire les relations de corrélation qui existent entre les différents patrons. Notons que les relations de corrélation ont été analysées par rapport aux séquences similaires, c'est à dire les séquences appartenant au même cluster. Cela nous permet d'éviter de chercher de telles relations dans toutes les séquences de la base de données. La complexité de cette étape dans le pire cas est de l'ordre de $O(TK^2)$, où T est le nombre de clusters et K est le nombre total des patrons significatifs. L'importance de cette étape est, tout d'abord, l'étude de la corrélation entre les patrons afin de faciliter par la suite l'examen des relations causales qui ne peuvent exister qu'entre

des patrons corrélés. De plus, cette étape permet d'éliminer tous les patrons qui ne possèdent pas de relations de corrélations, ce qui permet de réduire l'espace de recherche lors de l'étude des relations causales. Nous avons fixé dans cette étape un seuil de corrélation à 0.5. Il serait intéressant dans le futur d'incorporer un nouveau mécanisme permettant de calculer ce seuil de façon automatique.

Finalement, la notion de causalité que nous avons introduite dans notre approche constitue une principale contribution de notre approche. Selon nos connaissances de l'état de l'art, notre travail constitue le premier travail faisant appel à la théorie de la causalité probabiliste pour modéliser les relations causales entre les différents comportements des usagers observés lors de la réalisation des différentes activités de la vie quotidienne. Le potentiel de notre approche est qu'elle est capable de modéliser des phénomènes auxquelles les approches existantes ne peuvent faire face, à savoir le problème de la bidirectionnalité. Nous avons choisi l'approche de transfert d'entropie, qui est une approche simple, rapide et efficace, pour la découverte des relations causales entre les différents patrons. En revanche, d'un point de vue statistique, la causalité aura beaucoup plus de sens si le nombre de séquences dans la base de données est important. Cela permettra d'établir des interprétations statistiques sur les relations causales détectées et d'en tirer des conclusions sur leurs effet et importance. Dans certaines bases de données utilisées dans nos expérimentations, le nombre de séquences est petit comme celles de Domus 1 et 2. Cela pourrait ne pas donner suffisamment d'information statistique et rendre l'interprétation des résultats beaucoup plus délicate. Ceci dit que, plus le nombre de séquences augmente dans les bases de données, plus notre approche aura de potentiels et plus les résultats seront facilement interprétables. Dans le cas des activités de vie quotidienne, plus la séquence d'activité est longue, plus le nombre de patrons augmente. Cela facilitera l'étude des relations de corrélations de même que les relations causales. Notons que cette limite n'est pas liée uniquement à notre approche, mais elle est rencontrée par la vaste majorité des approches statistiques. Il serait intéressant dans le futur d'incorporer un nouveau mécanisme permettant de faire face au problème des données de petite taille.

En addition aux points sus cités, notre approche permet d'étudier les relations causales de façon statique. C'est à dire, notre approche ne prend pas en compte l'évolution temporelle ni le changement des patrons au cours du temps. L'évolution

temporelle des patrons est un phénomène lié au changement ou à l'évolution du comportement usager au cours du temps. Par exemple, l'usager pourrait changer ses préférences et intérêts au cours du temps. Si l'usager avait l'habitude de conduire une voiture de marque Japonaise 'Nissan', après un certain temps, il veut changer la marque de la voiture en achetant une voiture allemande 'Audi' par exemple. Ceci dit qu'il serait très intéressant dans le futur d'incorporer un nouveau mécanisme permettant de prendre en compte les nouvelles préférences des usagers et de mettre à jour leurs profils en conséquence.

3.7 Conclusion

Nous avons présenté dans ce chapitre, une nouvelle approche de modélisation du profil comportemental des usagers, ainsi qu'une validation qui nous a permis de concrétiser l'approche présentée dans ce chapitre. Pour ce faire, nous avons d'abord présenté les travaux existants dans ce domaine afin de dégager les limites de ces travaux et d'en proposer une approche permettant de surmonter ces limites. Ensuite, nous avons présenté le contexte général notre travail, ainsi que les différents aspects théoriques liés à notre travail à savoir la construction des arbres probabilistes des suffixes, le clustering des séquences, la corrélation entre les patrons significatifs et la la notion de causalité à travers le principe de transfert d'entropie. De cette façon, nous avons pu regrouper les éléments essentiels à la mise en oeuvre de notre approche aussi bien sur l'aspect théorique que sur l'aspect pratique. Pour clarifier notre approche, nous avons présenté les différentes composantes architecturales de notre approche de façon formelle.

Nous avons ensuite présenté les détails concernant la phase d'expérimentation. Nous avons débuté cette section par la présentation de notre démarche expérimentale, en spécifiant les critères d'évaluation, les objectifs visés au départ, les bases de données utilisées ainsi que les conditions d'expérimentation. Nous avons effectué deux types de validation pour répondre aux questions posées au départ. Une validation subjective qui vise à évaluer la capacité de notre approche à construire des profils comportementaux des usagers à l'intérieur des habitats intelligents et l'extérieur. De plus cette validation a également évalué la capacité de notre approche à

construire des profils communs pour un groupe d'usagers. La deuxième validation est la validation objective dans laquelle nous avons évalué la capacité de notre approche à prédire les activités des usagers et à identifier les usagers dans un environnement multi usagers. Dans cette deuxième validation, nous avons également comparé notre approche avec les approches les plus connues dans la littérature pour pouvoir positionner notre approche par rapport à l'état de l'art du domaine de la prédiction d'activités et l'identification des usagers. À la lumière de ces résultats, nous avons pu mettre en évidence les forces de notre modèle ainsi que les travaux futurs que nous envisagions réaliser. Nous avons aussi identifié quelques solutions permettant de guider l'amélioration future de notre approche.

Conclusion

Le projet de recherche, mené à terme par l'achèvement de cette thèse de doctorat, se veut une réponse à deux problématiques différentes et en même temps complémentaires énoncées en introduction. Ces deux problématiques portent respectivement sur la découverte et la reconnaissance des activités humaines dans les habitats intelligents, et sur la modélisation du profil usager dans les habitats intelligents. Nous avons pu constater à travers nos états de l'art présentés dans les chapitres 1 et 3, qu'il existait très peu d'approches permettant de surmonter le problème d'annotation des données. De même pour la deuxième problématique abordée dans cette thèse où les approches existantes ne permettent pas de modéliser certains phénomènes à savoir la bidirectionnalité entre les variables comme nous l'avons mentionné dans le chapitre 4. Pour faire face à ces limites, nous avons proposé dans cette thèse deux modèles statistiques permettant de surmonter les limites existantes. Le premier modèle que nous avons proposé est un modèle non supervisé basé sur le modèle LDA, et qui permet de découvrir les activités dans les séquences de façon non supervisée sans faire nécessairement appel à l'annotation des données. Cela constitue la première et importante contribution de cette thèse. Le deuxième modèle statistique que nous avons proposé permet d'étudier les relations causales entre les différents comportements des usagers afin de construire leurs profils comportementaux. L'introduction de la théorie de la causalité dans la modélisation du profil usager constitue la deuxième contribution importante de cette thèse.

La première étape de notre projet a permis de faire des investigations en profondeur concernant les deux problématiques abordées dans cette thèse. Cette étape a servi à faire une étude ciblée des différents travaux existants pour chacune des problématiques afin de bien comprendre les caractéristiques de chaque approche ainsi que

les limites de ces approches. Cela nous a permis de mieux cerner les besoins inhérents à notre contexte et de dégager les limites des approches existantes.

La deuxième étape de notre projet consistait, à la lumière des investigations du premier volet, de proposer des pistes de solutions afin de faire face aux limites soulevées dans la première étape. Cette étape a permis de mettre de l'avant nos hypothèses afin de définir formellement les modèles répondant aux besoins identifiés pour la découverte et la reconnaissance des activités, et pour la construction du profil comportemental usager. Nous avons donc opté pour le développement des approches statistiques qui s'adaptent mieux à notre contexte de reconnaissance d'activités et de la modélisation du profil usager. Ensuite, nous avons formalisé les approches proposées en s'appuyant sur le modèle statistique de LDA jumelé avec le principe de forage de patrons fréquents pour la problématique de la reconnaissance d'activités, et en s'appuyant sur la notion de causalité entre les patrons pour la problématique de la modélisation du profil usager. Cette étape nous a donc permis de poser les fondements théoriques de chaque approche.

La dernière étape de notre projet consistait de valider les modèles théoriques développés dans l'étape précédente, de façon à vérifier leurs fonctionnalités et performances dans un contexte réel avec des données réelles issues des habitats intelligents réels. Pour ce faire, nous avons utilisé des bases de données contenant différents types d'activités et provenant de plusieurs habitats intelligents. Nous avons spécifié des critères de validation au début de chaque validation afin de pouvoir tester nos approches selon différents aspects et de comparer celles ci avec les travaux connexes existants à savoir [128, 124, 121] ayant procédé à des expérimentations similaires dans un contexte similaire. Cette étape finale de notre projet nous a permis d'identifier les forces et limites des modèles que nous avons développés, et d'entrevoir les pistes de développement futurs pour l'amélioration et l'extension de nos modèles.

Cette thèse résulte en plusieurs contributions permettant l'avancement du domaine des habitats intelligents en général, et les domaines de la reconnaissance d'activités et la modélisation du profil des usagers en particulier. Nous avons montré, dans notre contribution dans le domaine de la reconnaissance d'activités, comment le principe du forage des patrons fréquents combiné avec le modèle statistique de LDA permettait de définir un modèle formel capable de résoudre deux problèmes épineux

qui sont la découverte et la reconnaissance d'activités. De plus, nous avons exploité la puissance de notre modèle afin de résoudre d'autres défis comme la découverte de patrons significatifs pour chaque activité. Cela constitue une autre importante contribution dans le domaine de la reconnaissance d'activités. De façon similaire, nous avons montré, dans notre contribution dans le domaine de la modélisation du profil usager, comment la notion de la causalité combiné avec les patrons fréquents permettait de définir un modèle capable de découvrir les comportements significatifs et de découvrir également les relations causales entre ces différents comportements. La découverte des comportements significatifs nous a permis d'exploiter la puissance des arbres probabilistes des suffixes et d'en proposer une méthode pour optimiser la profondeur de l'arbre. Ceci constitue une contribution non seulement dans le domaine de la modélisation du profil de l'usager, mais aussi dans le domaine du forage de données de façon générale. De plus, la notion de causalité entre les comportements, que nous avons introduite dans cette thèse, nous a permis de résoudre certains problèmes dans la littérature à savoir le problème de la bidirectionnalité entre les variables ce qui constitue une contribution majeure dans ce domaine, et une solution permettant de surmonter les limites des approches existantes pour la modélisation du profil usager à savoir les modèles Markoviens ou Bayésiens.

Enfin, les contributions citées en haut sont des contributions originales de cette thèse, qui permettent de fournir des apports importants aux problématiques de la reconnaissance d'activités et de la modélisation du profil usager. Les approches que nous avons proposées constituent des briques dans une importante plateforme d'assistance des personnes dans les habitats intelligents du laboratoire Domus. Nos approches fournissent des outils, des librairies, des documents et des articles à des fins de recherches, d'applications et également des fins pédagogiques.

Les approches que nous avons proposées comportent, bien entendu, certaines limitations identifiées lors des différentes étapes de validation. D'abord, pour l'approche de la reconnaissance d'activités, même si notre approche permettait de reconnaître les différents types d'activités avec différents niveaux de complexité, certaines activités pouvaient être difficilement reconnues. Cette limitation est induite principalement par la simplicité de l'activité en tant que telle. Une activité très simple pourrait être identifiée par un petit nombre de capteurs, et dans certains cas un seul capteur suffira

pour identifier une activités. Dans ce cas de figure, il serait difficile d'identifier des patrons d'activités de longueur supérieure à 2 ou 3, ce qui influe par conséquent sur le processus de la reconnaissance.

D'autre part, une autre limite de notre approche est induite par le postulat de départ qui suppose que les activités sont exécutées l'une après l'autre de façon séquentielle. Cela exclut le fait que les activités dans la pratique peuvent être exécutées de façon parallèle ou de façon entrelacée. Cependant, pour exploiter toutes les forces du modèle statistique LDA afin de prendre en compte ces types d'exécution des activités, il est nécessaire de prendre en considération les relations temporelles entre les différentes activités et la durée de chaque activité. Cela pourrait, par exemple, être considéré en augmentant le modèle courant par d'autres distributions de probabilité permettant de relâcher la contrainte de la séquentialité.

Pour l'approche de la modélisation du profil usager, bien que cette approche permette de découvrir les comportements significatifs et les relations causales entre ces comportements, on constate qu'elle souffre aussi du problème des activités très simples et du petit nombre d'instances de chaque activité. En effet, cette réalité découle du fait que, du point de vue statistique, effectuer le clustering sur un petit nombre d'instances pour une activité donnée, ne donne pas d'informations valides statistiquement. Ceci pourrait influer sur les résultats obtenus et sur la qualité de l'interprétation qui pourrait en être tirée. D'autre part, les activités simples, comme nous l'avons mentionnées précédemment, ne permettent pas d'identifier des patrons d'activités sur lesquels nous pouvons bâtir des modèles et tirer des conclusions.

Par ailleurs, dans les deux approches que nous avons proposées, comme c'est d'ailleurs le cas pour les approches existantes, nous avons fixé des seuils et utilisé des paramètres selon nos besoins dans différents contextes lors du processus de validation. Ces suppositions, bien que moins rigides dans la plupart des cas par rapport à celles utilisées dans la littérature, constituent des limites des approches proposées et ne correspondent pas vraiment à la réalité observées dans la vie quotidienne. comme par exemple les paramètres de la loi Dirichlet que nous avons supposé symétriques. Il serait intéressant d'incorporer de nouveaux mécanismes permettant de faire face à ces limites en calculant ces seuils et paramètres de façon automatique. Bien que les approches proposées dans cette thèse souffrent de certaines limites, elles doivent

être considérées comme des outils contribuant de manière significative à l'évolution des domaines de la reconnaissance d'activités et de la modélisation du profil usager, et permettant de surmonter beaucoup de limites observées dans les approches existantes dans la littérature. Dans cette optique, les contributions originales de cette thèse permettent d'envisager plusieurs perspectives de développement intéressantes, non seulement dans le domaine des habitats intelligents, mais aussi dans d'autres domaines connexes comme la médecine, la biologie, la sécurité, les réseaux sociaux. Nos contributions permettent de guider les prochains travaux découlant de nos approches et d'ouvrir des nouvelles voies de développement et de recherche dans différents domaines.

Annexe A

Modèles Markoviens

Dans cet annexe, nous allons présenter brièvement les techniques proposées pour résoudre les trois problèmes liés à un modèle Markovien caché.

Dans le problème d'évaluation, nous devons calculer la probabilité $P(Y|\lambda)$. Pour une séquence d'états S, nous pouvons calculer de façon directe cette probabilité comme suite : $P(Y|S,\lambda) = \prod_{t=1}^{T} P(y_t|s_t,\lambda)$. Cependant, la séquence d'états S n'est pas toujours connue en pratique. Pour cela, un algorithme d'avant-arrière (forward-backward) a été proposé pour résoudre le problème d'évaluation. Cet algorithme est composé de deux étapes de calcul. Dans l'étape avant, des variables $\alpha_t(i)$ sont introduites pour calculer les probabilités d'observation de la séquence partielle du début jusqu'au temps t soit $\{y_1, y_2, ..., y_t\}$. La technique de la programmation dynamique est utilisée en stockant les résultats intermédiaires pour éviter le calcul répétitif. Dans l'étape arrière qui est symétrique à l'étape avant, des variables $\beta_t(i)$ sont introduites pour calculer les probabilités d'observation de la séquence partielle $\{y_{t+1}, y_{t+2}, ..., y_T\}$.

Dans le problème de décodage, la séquence d'états $S = \{s_1, s_2, ..., s_T\}$ la plus probable est calculée étant donnée une séquence d'observations $Y = \{y_1, y_2, ..., y_T\}$. Pour ce faire, l'algorithme de programmation dynamique de Viterbi dédié aux modèles de Markov cachés est utilisé [108]. Dans cet algorithme, nous cherchons à maximiser la probabilité :

$\delta_t(i) = argmax_{s_1,s_2,...,s_{t-1}} P(\{y_1, y_2, ..., y_t\}, \{s_1, s_2, ..., s_{t-1}\}, s_t = i|\lambda)$

De la même façon, une technique de programmation dynamique est requise dans ce contexte pour optimiser le temps de calcul. La probabilité $\delta_t(i)$ est calculée récur-

sivement pour les différentes valeurs de t ($t = 1, 2, ..., T$).

Pour le problème d'apprentissage, la tâche consiste à trouver le modèle λ qui décrit le mieux les données $\mathcal{D}$. Trouver le modèle λ signifie trouver les paramètres $\lambda = (\pi, A, B)$ à partir d'un ensemble de données d'observation $\mathcal{D} = \{Y^i\}_{i=1}^k$. Cela revient à résoudre la maximisation suivante : $\lambda^* = argmax_\lambda P(\mathcal{D}|\lambda)$. La solution de cette maximisation peut être atteinte en utilisant l'algorithme Baum-Welch qui est similaire à l'algorithme EM présenté précédemment. Cet algorithme est composé de deux étapes :

- **Étape E** : cette étape consiste à estimer les probabilités d'être dans un certain état ($\theta_t(i)$) et de faire une certaine transition ($\Gamma_t(i, j)$) à partir d'un modèle λ. Donc, $\theta_t(i) = P(s_t = S_i|Y, \lambda)$.
- **Étape M** : cette étape vise à estimer les paramètres du modèle λ à partir des probabilités calculées dans l'étape E, i.e. à partir des probabilités d'être dans un état ($\theta_t(i)$) et de faire une transition ($\Gamma_t(i, j)$). Donc, $\Gamma_t(i, j) = P(s_t = S_i, s_{t+1} = S_j|Y, \lambda)$.

Exemple 4. Prenons un exemple des conditions météorologiques[1]. Nous pouvons avoir les trois états suivants : $S = \{sun, cloud, rain\}$, donc $|S| = 3$. Nous observons le climat sur quelques jours, par exemple : $\{y_1 = s_{sun}, y_2 = s_{cloud}, y_3 = s_{cloud}, y_4 = s_{rain}, y_5 = s_{cloud}\}$ avec $T = 5$. Les états observés du climat représentent la sortie d'un processus aléatoire dans le temps. Nous pouvons avoir les transitions suivantes (matrice de transition) entre les différents états :

$$A = \begin{pmatrix} & s_0 & s_{sun} & s_{cloud} & s_{rain} \\ s_0 & 0 & 0.33 & 0.33 & 0.33 \\ s_{sun} & 0 & 0.8 & 0.1 & 0.1 \\ s_{cloud} & 0 & 0.2 & 0.6 & 0.2 \\ s_{rain} & 0 & 0.1 & 0.2 & 0.7 \end{pmatrix}$$

La figure A.1 représente graphiquement le HMM relatif aux conditions climatiques.

L'état s_0 constitue l'état initial du processus. Il représente la distribution de probabilité initiale sur les états au temps 0. Notons que cette distribution initiale est

1. Exemple tiré de [109]

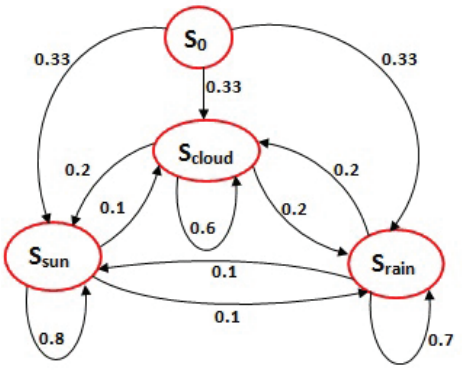


figure A.1 – Exemple d'un HMM

appelée π (ou la loi initiale d'un HMM). Comme nous pouvons le constater dans la matrice A, l'état initial présente une loi de probabilité uniforme pour transiter à chaque état parmi les trois états du processus. Nous constatons aussi à partir de la diagonale de la matrice A que le climat est auto-corrélé, c-à-d, si le climat est dans un état 'sun' il tend à rester dans le même état 'sun'. La même remarque s'applique pour les autres états. Ceci est une observation valable dans plusieurs modèle Markoviens [109].

Maintenant, la probabilité d'observer la séquence $\{y_1 = s_{sun}, y_2 = s_{cloud}, y_3 = s_{cloud}, y_4 = s_{rain}, y_5 = s_{cloud}\}$ peut être calculée comme suit :

$P(y_1 = s_{sun}, y_2 = s_{cloud}, y_3 = s_{cloud}, y_4 = s_{rain}, y_5 = s_{cloud})$, qui peut être factorisée comme suit :

$P(s_{sun}|s_0).P(s_{cloud}|s_{sun}).P(s_{cloud}|s_{cloud}).P(s_{rain}|s_{cloud}).P(s_{cloud}|s_{rain})$.

$= 0.33 \times 0.1 \times 0.6 \times 0.2 \times 0.2$

$= 0.00079$.

Nous pouvons aussi poser la question suivante : sachant que le climat aujourd'hui est 'sun', quelle est la probabilité que le climat sera 'cloud' demain et 'rain' après demain ?

Donc, fondamentalement, nous voulons déterminer la probabilité $P(y_2 = s_{cloud}, y_3 = s_{rain}|y_1 = s_{sun})$.

$= P(y_3 = s_{rain}|y_2 = s_{cloud}, y_1 = s_{sun}).P(y_2 = s_{cloud}|y_1 = s_{sun})$

$= P(y_3 = s_{rain}|y_2 = s_{cloud}).P(y_2 = s_{cloud}|y_1 = s_{sun})$

$= 0.2 \times 0.1 = 0.02$

Annexe B

Analyse des séquences

Dans cette annexe Plusieurs algorithmes ont été proposés dans la littérature pour extraire des patrons fréquents. Ces algorithmes peuvent être divisés en deux classes principales : extraction de patrons avec génération de candidats et extraction de patrons sans génération de candidats. L'extraction de patrons avec génération de candidats permet de générer des patrons appelés 'candidats', à partir des patrons existants, puis une étape de recherche dans la base de données permet de valider si ces candidats existent réellement dans la base de données ou pas. Les candidats existants seront utilisés pour générer d'autres candidats de longueur plus grande, et ainsi de suite. Cette stratégie n'est pas adoptée par les algorithmes sans extraction de candidats. De plus, ces classes peuvent être encore divisées en deux catégories selon le format de la base de données. Les bases de données qui adoptent le format horizontal sont représentées comme suit : (TID, Itemset), où le TID est l'identifiant de la transaction dans la base de données, et l'Itemset est l'ensemble des itemsets dans la transaction TID. Par contre les bases de données qui adoptent le format vertical prennent la forme de (Item, TID-set) où le TID-set représente l'ensemble des transactions contenant l'Item. Ces deux représentations sont illustrées dans les tableau B.1 et B.2.

Nous allons présenter brièvement deux algorithmes représentant les deux différentes classes d'algorithmes d'extraction de patrons fréquents à partir des données transactionnelles.

tableau B.1 – Exemple de base de données transactionnelle sous format (TID, Itemset)

Numéro de transaction	Itemset
100	a b c e
200	b d
300	b c
400	a b d
500	a b c
600	b c e
700	a c
800	a b e
900	a b c
1000	b c

tableau B.2 – Exemple de base de données transactionnelle sous format (Item, TID-set)

Item	Transactions
a	100, 400, 500, 700, 800, 900
b	100, 200, 300, 400, 500, 600, 800, 900, 1000
c	100, 300, 500, 600, 700, 900, 1000
d	200, 400
e	100, 600, 800

B.1 Algorithme Apriori

L'algorithme Apriori, introduit par Agrawal [5], est un algorithme clé pour l'extraction des règles d'association car la vaste majorité des algorithmes proposés pour l'extraction des patrons fréquents et la recherche des règles d'association se basent sur le principe de l'algorithme Apriori. Le nom de l'algorithme Apriori est tiré de l'heuristique qui utilise l'information connue a priori sur la fréquence des items dans la base de données. Cette heuristique peut être traduite formellement comme suit : si le support d'un item $I \in E$ (E est un sous-ensemble d'items de l'ensemble $\mathcal{E}$), est inférieur au support minimal, alors l'item I ne peut pas être engagé dans une règle d'association avec un autre item $J \notin E$ de l'ensemble $\mathcal{E}$.

L'algorithme Apriori est un algorithme itératif basé sur deux étapes importantes : 1) génération des candidats, et 2) élagage. Ces deux étapes s'exécutent l'une après

l'autre. C'est à dire, l'étape de génération de candidats et toujours suivie par l'étape d'élagage après un balayage de la base de données.

B.1.1 Génération des candidats

Supposons, sans perte de généralité, que la base de données contient des données transactionnelles. Pour chercher les patrons fréquents, l'algorithme effectue plusieurs balayages de la base de données. Dans le premier balayage, l'algorithme identifie les patrons candidats ainsi que la fréquence des items pour calculer leur support. Les patrons qui possèdent un support plus grand que la valeur σ prédéterminée par l'usager seront conservés et considérés comme patrons fréquents de longueur 1. Ces patrons fréquents seront par la suite utilisés pour générer des candidats de longueur 2 pour le deuxième balayage. Ce processus continue pour les patrons de longueur 3, 4 et plus jusqu'à ce qu'il n'y ait plus de candidats à générer.

B.1.2 Élagage

L'étape d'élagage est une étape très importante pour éliminer les patrons non fréquents et réduire ainsi l'espace de recherche. Une fois les candidats générés, le support de tous les candidats est calculé par un balayage de la base de données. Les candidats dont le support ne dépasse pas le support minimal défini par l'utilisateur seront éliminés. En se basant sur le principe d'Apriori, un candidat qui a été éliminé à l'étape k ne sera pas considéré à l'étape $k + 1$.

Exemple 5. Voici un exemple illustrant le fonctionnement de l'algorithme Apriori sur la base de données transactionnelle présentée dans le tableau B.1. Les étapes de l'algorithme sont détaillées comme suit :

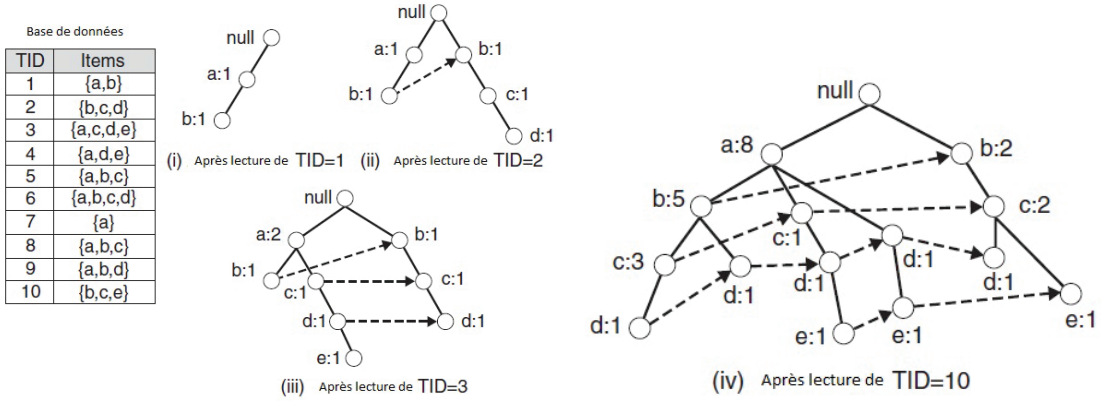
- la première étape de l'algorithme consiste à parcourir la base de données pour compter le support de chaque 1-itemset et former l'ensemble des candidats. Supposons que le support minimal égal à 30 % ou fréquence de 3.
- les itemsets ayant un support supérieur ou égal à celui du support minimal seront conservés et les autres seront éliminés (premier élagage). Dans notre cas : les 1-itemsets fréquents sont {a : 60 %}{b : 90 %}{c : 70 %}{e : 30 %}.

- les 1-itemsets fréquents trouvés dans la première étape seront utilisés pour générer des candidats. Cette génération est effectuée en liant l'ensemble des 1-itemsets fréquents avec lui même. Le nombre de candidats générés dans cette étape est le nombre de combinaisons possibles entre les 1-itemsets qui est $\frac{n(n-1)}{2}$, où n est le nombre des 1-itemsets. Dans notre cas, six $((4 \times 3)/2 = 6)$ candidats sont générés qui sont {a, b}{a, c}{a, e}{b, c} {b, e}{c, e}.
- après avoir généré les candidats (2-itemsets), l'algorithme effectue un deuxième balayage de la base de données afin de calculer le support respectif des candidats. Il en résulte de cette étape les patrons fréquents suivants : {a, b : 5}{a, c : 4}{b, c : 6} {b, e : 3}.
- les patrons fréquent de l'étape précédente seront utilisés pour former des candidats 3-itemsets suivants : {a, b, c}{a, b, e}{b, c, e}. L'algorithme effectue un nouveau balayage pour calculer le support respectif de ces candidats. Le résultat de cette étape est le patron fréquent {a, b, c : 3} et les deux autres candidats sont éliminés. À ce stade, ne pouvant plus générer de candidats de longueur plus grand que 3 (4-itemsets), l'algorithme s'arrête.

B.2 Algorithme FP-growth

L'algorithme FP-growth, introduit par Han et al. [53], utilise une structure de données compacte appelée Frequent-Pattern tree (FP-Tree) et qui apporte une solution au problème de la fouille de patrons fréquents dans une grande base de données transactionnelle. Cet algorithme, contrairement à l'algorithme Apriori, permet d'éviter le balayage répétitif de la base de données en stockant les patrons fréquents dans une structure compacte. De plus, pour des fins de performance, l'algorithme procède à un tri des items dans la structure compacte, ce qui accélère la recherche des patrons dans la base de données.

Le FP-Tree est construit en parcourant la base de données une transaction à la fois. Cette transaction correspond à un chemin dans l'arbre FP-Tree. Plus les chemins se chevauchent, plus l'arbre devient compact minimisant ainsi l'espace mémoire. La figure B.1 présente les différentes étapes de construction d'un arbre FP-Tree à partir d'une base de données transactionnelle (cet exemple est tiré du livre [125]).



(a) Lecture de TID 1, TID 2 et TID 3 (b) Lecture de TID 10

figure B.1 – Exemple de construction d'un FP-Tree.

Initialement, le FP-Tree contient uniquement le noeud racine représenté par le symbole *null*. Par la suite, le FP-Tree commence à s'élargir au fur et à mesure de la lecture des transactions dans la base de données. Ces étapes sont détaillées comme suit [1] :

- la base de données est parcourue une fois pour calculer le support de chaque item. Les items non fréquents sont éliminés et les autres sont triés dans un ordre décroissant de leur support.
- la base de données est parcourue une autre fois afin de construire l'arbre FP-Tree. Après la lecture de la première transaction, {a, b} dans notre cas, les noeuds a et b seront crées, et le chemin $null \rightarrow a \rightarrow b$ sera crée.
- après la lecture de la deuxième transaction, i.e. {b, c, d}, les noeuds b, c et d seront crées dans l'arbre. Bien que les deux transactions ont l'item b en commun, elles ne partagent pas le même chemin, parce que tout simplement ces transactions ne possèdent pas le même préfixe.
- la troisième transaction {a, c, d, e} possède un préfixe commun avec la première transaction. Par conséquent, leurs chemins se chevauchent dans l'item a. À cause de ce chevauchement, la fréquence de a s'incrémente.
- ce processus continue jusqu'à ce que toutes les transactions soient lues et les noeuds correspondant soient crées dans l'arbre.

1. Ces étapes sont extraites de [125]

Les deux algorithmes que nous avons présentés sont les algorithmes les plus utilisés dans la littérature. Ces algorithmes adoptent le format horizontal de la base de données. Il existe d'autres algorithmes qui sont basés sur ces deux algorithmes, et qui ont été proposés pour la recherche de patrons fréquents dans les bases de données transactionnelle ou séquentielle. On peur citer par exemple : BIDE [133], PrefixSpan [104], Free-span [52], Eclat [145], GSP [115], SPADE [146], CHARM [147], UP-growth [126].

B.3 Différence entre l'extraction des patrons séquentiels et patrons fréquents

Dans la section précédente, nous avons présenté deux algorithmes d'extraction de patrons fréquents à partir des bases de données transactionnelles. Cependant, le problème de recherche de patrons séquentiels dans les bases de données séquentielles est un peu différent. En effet, la recherche des patrons séquentiels vise à chercher des patrons inter transactions ou séquences, comme par exemple une séquence d'items achetés suivie par l'achat d'un autre item dans une transaction avec une contrainte temporelle. La contrainte temporelle joue un rôle très important dans la détermination des patrons séquentiels. Nous constatons d'après cet exemple, que la recherche des patrons séquentiels vise en fait à chercher les enchaînements fréquents dans les bases de données en prenant en compte les contraintes temporelles. Ces enchaînements peuvent être entre des ensembles d'items, un ensemble d'items avec un seul autre item, entre deux items, et ainsi de suite.

Nous pouvons différencier l'extraction des patrons séquentiels et patrons fréquents selon deux plans essentiels : une différence sur le plan théorique, et une autre sur le plan pratique. La différence théorique réside dans les contraintes temporelles employées dans le cas des données séquentielles. Par exemple, le patron $ABCD$ est considéré comme un patron fréquent parce que les items contenus dans le patron apparaissent ensemble fréquemment. Par contre, le patron $A(BC)D$ est considéré comme un patron fréquent séquentiel et signifie que l'item A a été acheté, suivi des items BC, suivis de l'item D. Donc, la contrainte temporelle est bel est bien apparente

dans cet exemple.

La différence pratique importante que nous pouvons constater réside dans le fait que les items dans un itemset sont uniques, et les répétitions ne sont pas tolérées. Par contre dans un patron séquentiel, nous pouvons avoir des répétitions d'items. Nous aurons donc besoin de prendre en compte ces propriétés pour qu'un algorithme d'extraction d'itemsets soit adapté à l'extraction de patrons séquentiels.

B.4 Comparaison entre les algorithmes Apriori et FP-growth

Bien que les deux familles d'algorithmes Apriori et FP-Growth soient très utilisées dans la littérature, ceux-ci possèdent des avantages et des limites. Dans ce qui suit, nous allons comparer les deux familles d'algorithmes afin d'élucider les points forts de chaque algorithme.

Nous avons mentionné auparavant que l'algorithme Apriori effectue plusieurs balayages de la base de données pour extraire les patrons fréquents, contrairement à l'algorithme FP-growth qui effectue uniquement deux parcours de la base de données. La complexité de l'algorithme Apriori est de l'ordre de ($O(Nw)$ (génération des items de longueur 1)+ $\sum_{k=2}^{w}(k-2)|C_k|$ (génération de candidats)+ $\sum_{k=2}^{w} k(k-2)|C_k|$ (élagage)+ $O(N\sum_k \binom{w}{k}\alpha_k)$ (comptage de support)), où w est la longueur moyenne des séquences, N est le nombre total des transactions dans la base de données, C_k est l'ensemble des itemsets de longueur k, et α_k est le coût de mise à jour du support d'un itemset [125]. La complexité de l'algorithme FP-growth est de l'ordre de $O(|DB|)$ (création de l'arbre) + $O(H^2.D)$ (recherche dans l'arbre), où $|DB|$ est la taille de la base de données, H est la taille de la table d'entête créée au début par l'algorithme, et D est la profondeur de l'arbre FP-Tree. Cela démontre un avantage de l'algorithme FP-growth par rapport à l'algorithme Apriori. En outre, Antunes et al. [3] ont effectué une étude comparative entre Apriori et FP-growth et ils ont conclu que l'algorithme FP-growth est plus performant lorsque le support minimum est plus petit. Par ailleurs, l'algorithme FP-growth souffre du problème d'espace mémoire dû à la taille de l'arbre FP-tree, et les bases de données projetées dans le cas de l'algorithme

PrefixSpan par exemple.

Selon Zaki et al. [147], l'algorithme Apriori est plus performant sur des données éparses, comme dans les données transactionnelles où les patrons fréquents ne sont pas longs. Par contre, sur des données denses comme les données de télécommunications où nous pouvons avoir beaucoup de patrons fréquents très longs, la performance de l'algorithme Apriori se dégrade.

B.5 Conclusion

Dans ce chapitre, nous avons introduit la notion de forage de patrons fréquents et comment ils sont découverts dans les base de données. Nous avons présenté deux exemples d'algorithme largement utilisés dans la littérature et qui adoptent le format horizontal de la base de données. Chaque algorithme possède des avantages et des inconvénients. Par exemple, l'algorithme Apriori est très performant, mais il souffre de deux problèmes fondamentaux : 1) si le nombre de patrons fréquents est grand, et 2) le balayage répétitif de la base de données. Par ailleurs, l'algorithme FP-growth permet d'éviter le balayage répétitif de la base de données en créant une représentation compacte sous forme d'arbre, mais requiert un large espace mémoire si la base de données est large.

L'extraction de patrons fréquents dans le comportement des personnes doit prendre en compte que la chance de recueillir des séquences de capteurs identiques est faible dans un habitat intelligent. En effet, une activité est réalisable de différentes manières à cause de l'ordre choisi pour réaliser les sous tâches qui peuvent s'intercaler lors de la réalisation d'une activité. Par conséquent, ces parties identiques, que nous appelons patrons, auront généralement une longueur petite, ce qui rend la recherche de ces patrons dans les bases de données des séquences très rapide. Par conséquent, nous avons choisi l'algorithme Apriori dans notre travail pour la recherche des patrons fréquents dans les bases de données. De plus, les bases de données des séquences que nous avons utilisées contiennent un grand nombre d'items (capteurs) ce qui rend l'algorithme FP-growth inapproprié à cause du problème d'espace mémoire. Par conséquent, le choix de l'algorithme Apriori dans notre travail nous semble approprié et plus judicieux dans ce contexte.

B.5. Conclusion

Dans le chapitre suivant, nous allons introduire notre approche proposée pour la découverte et la reconnaissance des activités.

Annexe C

Algorithme de l'échantillonnage de Gibbs pour LDA

Cette annexe présente en détail l'algorithme de l'échantillonneur de Gibbs pour LDA. L'implémentation de l'algorithme requiert des variables de comptage qui sont initialisées aléatoirement. Puis, une loupe est exécutée un certain nombre de fois. Dans chaque itération, un thème est sélectionné pour chaque instance de mot dans le corpus. Suivant ces itérations, les compteurs sont utilisés pour calculer les distributions latentes Θ et Φ.

Les compteurs impliqués dans l'algorithme sont : le compteur $n_m^{(k)}$ qui correspond au nombre de thèmes dans un document, le compteur $n_k^{(t)}$ qui correspond au nombre de mots associés à un thème, le compteur n_m qui correspond au nombre de fois un thème apparaît dans un document m, et le compteur n_k qui correspond au nombre de fois un mot quelconque est associé au thème k. L'algorithme global de l'échantillonnage de Gibbs pour LDA est présenté ci-dessous[1].

1. Cet algorithme est tiré du [45].

Algorithme 4 Algorithme de l'échantillonnage de Gibbs pour LDA

//Initialisation

- Mettre à zero tous les compteurs : $n_m^{(k)}, n_m, n_k^{(t)}$ et n_k

Pour tous les documents $m \in [1..M]$ dans le corpus **faire**

Pour tous les mots $n \in [1..N_m]$ du document m **faire**

- Échantillonner l'index de thème $z_{m,n} = k \sim Mult(1/k)$
- Incrémenter le compteur document-thème : $n_m^{(k)} + 1$
- Incrémenter la somme document-thème : $n_m + 1$
- Incrémenter le compteur thème-terme : $n_k^{(t)} + 1$
- Incrémenter la somme thème-terme : $n_k + 1$

fin pour

fin pour

// Échantillonnage

Tant que il n'est pas terminé **faire**

Pour tous les documents $m \in [1..M]$ **faire**

Pour tous les mots $n \in [1..N_m]$ du document m **faire**

//Pour l'attribution courante de k au terme t du mot $w_{m,n}$:

- Décrémenter les compteurs et sommes : $n_m^{(k)} - 1$; $n_m - 1$; $n_k^{(t)} - 1$; et $n_k - 1$

//Échantillonnage multinomiale

- Échantillonner l'index de thème $\hat{k} \sim P(z_i | Z_{-i}, D)$

// Utiliser le nouvel assignement de $z_{m,n}$ au terme t du mot $w_{m,n}$ pour :

- Incémenter les compteurs et sommes : $n_m^{(k)} + 1$; $n_m + 1$; $n_k^{(t)} + 1$; et $n_k + 1$

fin pour

fin pour

// Vérifier la convergence et estimer les paramètres

Si convergé **faire**

// les valeurs des paramètres sont en moyenne

- Estimer le paramètre Φ selon l'équation 2.16
- Estimer le paramètre Θ selon l'équation 2.17

fin Si

fin tant que

Bibliographie

[1] C. B. Abdelkader, « Motion-Based Recognition of People in EigenGait Space, » dans *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, série FGR '02. Washington, DC, USA : IEEE Computer Society, 2002, pp. 267–. Disponible à http://dl.acm.org/citation.cfm?id=874061.875420

[2] A. Ahmed, Y. Low, M. Aly, V. Josifovski, et A. J. Smola, « Scalable distributed inference of dynamic user interests for behavioral targeting, » dans *KDD*, 2011, pp. 114–122.

[3] C. Antunes et A. L. Oliveira, « Generalization of Pattern-Growth Methods for Sequential Pattern Mining with Gap Constraints, » dans *MLDM*, 2003, pp. 239–251.

[4] A. Aamodt et E. Plaza, « Case-based reasoning : foundational issues, methodological variations, and system approaches, » *AI Commun.*, vol. 7, no. 1, pp. 39–59, mars 1994. Disponible à http://dl.acm.org/citation.cfm?id=196108.196115

[5] R. Agrawal et R. Srikant, « Fast Algorithms for Mining Association Rules in Large Databases, » dans *Proc. VLDB 1994*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1994, pp. 487–499.

[6] R. Agrawal et R. Srikant, « Mining sequential patterns, » dans *Proc. ICDE 1995*. IEEE Computer Society, 1995, pp. 3–14.

[7] R. Arun, V. Suresh, C. E. V. Madhavan, et M. N. N. Murthy, « On Finding the Natural Number of Topics with Latent Dirichlet Allocation : Some Observations, » dans *PAKDD (1)*, 2010, pp. 391–402.

[8] M. F. A. bin Abdullah, A. F. P. Negara, M. S. Sayeed, D.-J. Choi, et K. S. Muthu, « Classification Algorithms in Human Activity Recognition using Smartphones, » *International Journal of Computer and Information Engineering*, vol. 6, pp. 77–84, 2012.

[9] J. L. Boyd-Graber, D. M. Blei, et X. Zhu, « A Topic Model for Word Sense Disambiguation, » dans *EMNLP-CoNLL*, 2007, pp. 1024–1033.

[10] L. Bao et S. S. Intille, « Activity Recognition from User-Annotated Acceleration Data, » dans *Proc. Pervasive 2004*, 2004, pp. 1–17.

[11] D. M. Blei, A. Y. Ng, et M. I. Jordan, « Latent dirichlet allocation, » *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, March 2003.

[12] R. M. Baecker et Others, *Readings in Human Computer Interaction : Toward the Year 2000*, 2nd édition. Morgan Kaufmann, 1995, ch. 2 : Design and Evaluation, pp. 73–186.

[13] G. Bouma, « Normalized (pointwise) mutual information in collocation extraction, » dans *Proceedings of Biennial GSCL Conference*, 2009, pp. 31–40.

[14] M. Buettner, R. Prasad, M. Philipose, et D. Wetherall, « Recognizing daily activities with RFID-based sensors, » dans *Proc. Ubicomp 2009*. New York, NY, USA : ACM, 2009, pp. 51–60.

[15] P. Biswas et P. Robinson, « Modelling perception using image processing algorithms, » dans *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers : Celebrating People and Technology*, série BCS-HCI '09. Swinton, UK, UK : British Computer Society, 2009, pp. 494–503. Disponible à http://dl.acm.org/citation.cfm?id=1671011.1671075

[16] M. Bouguessa, S. Wang, et H. Sun, « An objective approach to cluster validation, » *Pattern Recognition Letters*, vol. 27, no. 13, pp. 1419–1430, 2006.

[17] G. Bejerano et G. Yona, « Modeling protein families using probabilistic suffix trees, » dans *Proceedings of the third annual international*

conference on Computational molecular biology, série RECOMB '99. New York, NY, USA : ACM, 1999, pp. 15–24. Disponible à http://doi.acm.org/10.1145/299432.299445

[18] E. Cambouropoulos, « Extracting Significant Patterns from Musical Strings : Some Interesting Problems, » dans *London String Days workshop*, 2000.

[19] C. Cedras, « Motion-based recognition a survey, » *Image and Vision Computing*, vol. 13, no. 2, pp. 129–155, mars 1995. Disponible à http://dx.doi.org/10.1016/0262-8856(95)93154-k

[20] G. F. Cooper et E. Herskovits, « A Bayesian Method for the Induction of Probabilistic Networks from Data, » *Mach. Learn.*, vol. 9, no. 4, pp. 309–347, octobre 1992. Disponible à http://dx.doi.org/10.1023/A:1022649401552

[21] D. M. Chickering, « Optimal structure identification with greedy search, » *J. Mach. Learn. Res.*, vol. 3, pp. 507–554, mars 2003. Disponible à http://dx.doi.org/10.1162/153244303321897717

[22] X. Chen, X. Hu, T. Y. Lim, X. Shen, E. K. Park, et G. L. Rosen, « Exploiting the Functional and Taxonomic Structure of Genomic Data by Probabilistic Topic Modeling, » *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 9, no. 4, pp. 980–991, juillet 2012. Disponible à http://dx.doi.org/10.1109/TCBB.2011.113

[23] R. G. Cowell, S. L. Lauritzen, A. P. David, et D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems*, 1st édition, V. Nair, J. Lawless, et M. Jordan, éditeurs. Secaucus, NJ, USA : Springer-Verlag New York, Inc., 1999.

[24] L. Cao, Y. Ou, et P. S. Yu, « Coupled Behavior Analysis with Applications, » *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 8, pp. 1378–1392, 2012.

[25] B. Chikhaoui et H. Pigot, « Towards analytical evaluation of human machine interfaces developed in the context of smart homes, » *Interacting with Computers*, vol. 22, no. 6, pp. 449–464, 2010.

[26] B. Chikhaoui, S. Wang, et H. Pigot, « A New Algorithm Based On Sequential Pattern Mining For Person Identification In Ubiquitous Environments, » dans *Proceedings of SensorKDD*, 2010, pp. 19–27.

[27] B. Chikhaoui, S. Wang, et H. Pigot, « A Frequent Pattern Mining Approach for ADLs Recognition in Smart Environments, » dans *Proceedings of the 25th international conference on Advanced Information Networking and Applications*, 2011.

[28] M. S.-E. D.J. Cook, « Assessing the Quality of Activities in a Smart Environment, » *Methods of Information in Medicine*, vol. 48, no. 5, pp. 480–485, 2009.

[29] A. P. Dempster, N. M. Laird, et D. B. Rubin, « Maximum Likelihood from Incomplete Data via the EM Algorithm, » *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. pp. 1–38, 1977.

[30] R. des Nations Unies, « Journée des personnes âgées : promouvoir leurs droits et leur participation pleine et entière, » http ://www.unmultimedia.org/radio/french/detail/6487.html, 2008.

[31] S. M. et Fronteau A, « Les conditions du maintien à domicile des personnes âgées dépendantes, » Centre de recherche pour l'étude et l'observation des conditions de vie, France, Rapport de recherche, 1996.

[32] D. Godoy et A. Amandi, « User profiling in personal information agents : a survey, » *Knowl. Eng. Rev.*, vol. 20, no. 4, pp. 329–361, décembre 2005.

[33] Z. Ghahramani, « Hidden Markov models. » River Edge, NJ, USA : World Scientific Publishing Co., Inc., 2002, ch. An introduction to hidden Markov models and Bayesian networks, pp. 9–42. Disponible à http://dl.acm.org/citation.cfm?id=505741.505743

[34] M. Girolami et A. Kabán, « Sequential Activity Profiling : Latent Dirichlet Allocation of Markov Chains, » *Data Min. Knowl. Discov.*, vol. 10, no. 3, pp. 175–196, 2005.

[35] T. Gu, H. K. Pung, D. Q. Zhang, H. K. Pung, et D. Q. Zhang, « A Bayesian approach for dealing with uncertain contexts, » dans *Proceedings of Advances in Pervasive Computing in Pervasive' 04*, 2004, pp. 205–210.

[36] C. W. J. Granger, « Investigating causal relations by econometric models and cross-spectral methods, » *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[37] T. Griffiths, « Gibbs Sampling in the Generative Model of Latent Dirichlet Allocation, » Rapport technique, 2002.

[38] T. L. Griffiths et M. Steyvers, « Finding scientific topics, » *Proceedings of the National Academy of Sciences*, vol. 101, no. Suppl. 1, pp. 5228–5235, April 2004.

[39] T. Gu, Z. Wu, X. Tao, H. K. Pung, et J. Lu, « epSICAR : An Emerging Patterns based approach to sequential, interleaved and Concurrent Activity Recognition, » dans *Proc. PERCOM 2009*. Los Alamitos, CA, USA : IEEE Computer Society, 2009, pp. 1–9.

[40] J. Han, *Data Mining : Concepts and Techniques*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 2005.

[41] T. Huỳnh, U. Blanke, et B. Schiele, « Scalable recognition of daily activities with wearable sensors, » dans *Proceedings of the 3rd international conference on Location-and context-awareness*, série LoCA'07. Berlin, Heidelberg : Springer-Verlag, 2007, pp. 50–67. Disponible à http://portal.acm.org/citation.cfm?id=1777235.1777241

[42] A. Helal, D. J. Cook, et M. Schmalz, « Smart Home-Based Health Platform for Behavioral Monitoring and Alteration of Diabetes Patients, » *Journal of Diabetes science and technology*, vol. 3, no. 1, pp. 141Ű–148, 2009.

[43] J. Han, H. Cheng, D. Xin, et X. Yan, « Frequent pattern mining : current status and future directions, » *Data Min. Knowl. Discov.*, vol. 15, pp. 55–86, August 2007. Disponible à http://portal.acm.org/citation.cfm?id=1275092.1275097

[44] D. Heckerman, « Learning in graphical models, » M. I. Jordan, éditeur. Cambridge, MA, USA : MIT Press, 1999, ch. A tutorial on learning with Bayesian networks, pp. 301–354. Disponible à http://dl.acm.org/citation.cfm?id=308574.308676

[45] G. Heinrich, « Parameter estimation for text analysis, » University of Leipzig, Rapport technique, 2008.

[46] K. Henderson et T. Eliassi-Rad, « Applying latent dirichlet allocation to group discovery in large graphs, » dans *Proceedings of the 2009 ACM symposium on Applied Computing*, série SAC '09. New York, NY, USA : ACM, 2009, pp. 1456–1461. Disponible à http://doi.acm.org/10.1145/1529282.1529607

[47] F. Heylighen, « Causality as distinction conservation. A theory of predictability, reversibility, and time order, » *Cybernetics and Systems*, vol. 20, no. 5, pp. 361–384, 1989.

[48] T. Huynh, M. Fritz, et B. Schiele, « Discovery of activity patterns using topic models, » dans *Proc. UbiComp 2008*. New York, NY, USA : ACM, 2008, pp. 10–19.

[49] D. Heckerman, D. Geiger, et D. M. Chickering, « Learning Bayesian Networks : The Combination of Knowledge and Statistical Data, » *Machine Learning*, vol. 20, no. 3, pp. 197–243, 1995.

[50] R. Helaoui, M. Niepert, et H. Stuckenschmidt, « Recognizing Interleaved and Concurrent Activities : A Statistical-Relational Approach, » dans *PerCom 2011*, 2011, pp. 1–9.

[51] T. Hofmann, « Probabilistic Latent Semantic Indexing, » dans *SIGIR*, 1999, pp. 50–57.

[52] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, et M.-C. Hsu, « FreeSpan : frequent pattern-projected sequential pattern mining, » dans *Proc. KDD 2000*. New York, NY, USA : ACM, 2000, pp. 355–359.

[53] J. Han, J. Pei, Y. Yin, et R. Mao, « Mining Frequent Patterns without Candidate Generation : A Frequent-Pattern Tree Approach, » *Data*

Min. Knowl. Discov., vol. 8, no. 1, pp. 53–87, janvier 2004. Disponible à http://dx.doi.org/10.1023/B:DAMI.0000005258.31418.83

[54] T. Huynh et B. Schiele, « Unsupervised Discovery of Structure in Activity Data Using Multiple Eigenspaces. » dans *LoCA'06*, 2006, pp. 151–167.

[55] M. H. Hansen et B. Yu, « Model Selection and the Principle of Minimum Description Length, » Bell Labs, Technical Memorandum, 1998.

[56] E. O. H. III et D. J. Cook, « Improving Home Automation by Discovering Regularly Occurring Device Usage Patterns, » dans *ICDM*, 2003, pp. 537–540.

[57] M. I. Jordan, éditeur, *Learning in graphical models*. Cambridge, MA, USA : MIT Press, 1999.

[58] R. W. B. Jr, *Patterns of Behavior*. Chicago University Press, 2005.

[59] A. Jain, A. Ross, et S. Prabhakar, « An introduction to biometric recognition, » *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, no. 1, pp. 4 – 20, jan. 2004.

[60] P. Judea, *Causality : Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[61] R. Kadouche, B. Chikhaoui, et B. Abdulrazak, « User's behavior study for smart houses occupant prediction, » *Annales des Télécommunications*, vol. 65, no. 9-10, pp. 539–543, 2010.

[62] C. Kermorvant et P. Dupont, « Improved Smoothing for Probabilistic Suffix Trees Seen as Variable Order Markov Chains, » dans *Proceedings of the 13th European Conference on Machine Learning*, série ECML '02. London, UK, UK : Springer-Verlag, 2002, pp. 185–194. Disponible à http://dl.acm.org/citation.cfm?id=645329.650045

[63] Y. Kritikou, P. Demestichas, E. F. Adamopoulou, K. P. Demestichas, M. E. Theologou, et M. Paradia, « User Profile Modeling in the context of web-based learning management systems, » *J. Network and Computer Applications*, vol. 31, no. 4, pp. 603–627, 2008.

[64] S. Katz, A. B. Ford, R. W. Moskowitz, B. A. Jackson, et M. W. Jaffe, « Studies of illness in the aged. The index of adl : a standardized measure of biological and psychosocial function, » *JAMA*, vol. 185, pp. 914–919, 1963.

[65] E. Kim, S. Helal, et D. Cook, « Human Activity Recognition and Pattern Discovery, » *IEEE Pervasive Computing*, vol. 9, pp. 48–53, 2010.

[66] S. Kullback et R. A. Leibler, « On information and sufficiency, » *Annals of Mathematical Statistics*, vol. 22, pp. 49–86, 1951.

[67] R. Kohavi, « A study of cross-validation and bootstrap for accuracy estimation and model selection, » dans *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.

[68] R. Kadouche, H. Pigot, B. Abdulrazak, et S. Giroux, « Support Vector Machines for Inhabitant Identification in Smart Houses, » dans *UIC*, 2010, pp. 83–95.

[69] P. J. Krause, « Learning probabilistic networks, » *Knowledge Engineering Review*, vol. 13, no. 4, pp. 321–351, février 1999.

[70] W. Lee, *Decision Theory and Human Behavior*, first edition édition. John Wiley & Sons Inc, 1971.

[71] P. LERAY, « Réseaux bayésiens : apprentissage et modélisation de systèmes complexes, » Département ASI, INSA de Rouen, Université de Rouen, Rapport de recherche, 2006.

[72] L. Liao, D. Fox, et H. A. Kautz, « Location-Based Activity Recognition using Relational Markov Networks, » dans *Proc. IJCAI 2005*, 2005, pp. 773–778.

[73] L. Liao, D. Fox, et H. Kautz, « Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields, » *Int. J. Rob. Res.*, vol. 26, no. 1, pp. 119–134, janvier 2007.

[74] B. E. Lawton MP, « Assessment of older people : self-maintaining and instrumental activities of daily living, » *Gerontologist*, vol. 9, no. 3, pp. 179–186, 1969.

[75] Y. Liu, A. Niculescu-Mizil, A. C. Lozano, et Y. Lu, « Learning Temporal Causal Graphs for Relational Time-Series Analysis, » dans *Proceedings of ICML*, 2010, pp. 687–694.

[76] M. Lungarella, T. Pegors, D. Bulwinkle, et O. Sporns, « Methods for Quantifying the Informational Structure of Sensory and Motor Data, » *Neuroinformatics*, vol. 3, no. 3, pp. 243–262, 2005.

[77] S. Lühr, G. West, et S. Venkatesh, « Recognition of emergent human behaviour in a smart home : A data mining approach, » *Pervasive Mob. Comput.*, vol. 3, no. 2, pp. 95–116, mars 2007. Disponible à http://dx.doi.org/10.1016/j.pmcj.2006.08.002

[78] X. Liu, L. Zhang, M. Li, H. Zhang, et D. Wang, « Boosting image classification with LDA-based feature combination for digital photograph management, » *Pattern Recogn.*, vol. 38, no. 6, pp. 887–901, juin 2005. Disponible à http://dx.doi.org/10.1016/j.patcog.2004.11.008

[79] A. Mahmood, « Structure Learning of Causal Bayesian Networks : A Survey, » Department of Computing Science, University of Alberta, Edmonton, Canada, Rapport technique, 2011.

[80] Z. J. M. et T. Barbara, « Event structure in perception and conception, » *Psychological bulletin*, vol. 127, no. 1, pp. 3–21, 2001.

[81] J. Modayil, T. Bai, et H. Kautz, « Improving the recognition of interleaved activities, » dans *Proceedings of the 10th international conference on Ubiquitous computing*, série UbiComp '08. New York, NY, USA : ACM, 2008, pp. 40–43. Disponible à http://doi.acm.org/10.1145/1409635.1409641

[82] T. P. Minka, « Estimating a Dirichlet distribution, » MIT, PhD Thesis, 2000.

[83] V. Miele, S. Penel, et L. Duret, « Ultra-fast sequence clustering from similarity networks with SiLiX, » *BMC Bioinformatics*, vol. 12, p. 116, 2011.

[84] E. Manavoglu, D. Pavlov, et C. L. Giles, « Probabilistic User Behavior Models, » dans *Proceedings of ICDM*, 2003, pp. 203–210.

[85] D. Marinazzo, M. Pellicoro, et S. Stramaglia, « Kernel-Granger causality and the analysis of dynamical networks. » *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 77, no. 5-2, 2008.

[86] L.-P. Morency, A. Quattoni, et T. Darrell, « Latent-Dynamic Discriminative Models for Continuous Gesture Recognition, » dans *CVPR*, 2007.

[87] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, et E. Teller, « Equation of State Calculations by Fast Computing Machines, » *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[88] U. Maurer, A. Smailagic, D. P. Siewiorek, et M. Deisher, « Activity Recognition and Monitoring Using Multiple Sensors on Different Body Positions, » dans *Proc. BSN 2006*. Washington, DC, USA : IEEE Computer Society, 2006, pp. 113–116.

[89] H. Mannila, H. Toivonen, et A. I. Verkamo, « Discovery of Frequent Episodes in Event Sequences, » *Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 259–289, 1997.

[90] K. P. Murphy, « Dynamic Bayesian Networks : Representation, Inference and Learning, » UNIVERSITY OF CALIFORNIA, BERKELEY, PhD Thesis, 2002.

[91] Q. Mei, D. Xin, H. Cheng, J. Han, et C. Zhai, « Semantic annotation of frequent patterns, » *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 3, décembre 2007.

[92] R. M. Neal et G. E. Hinton, « Learning in graphical models, » M. I. Jordan, éditeur. Cambridge, MA, USA : MIT Press,

1999, ch. A view of the EM algorithm that justifies incremental, sparse, and other variants, pp. 355–368. Disponible à http://dl.acm.org/citation.cfm?id=308574.308679

[93] D. Niedermayer, « An Introduction to Bayesian Networks and their Contemporary Applications, » http ://www.niedermayer.ca, Rapport de recherche, 1998.

[94] N. T. Nguyen, D. Q. Phung, S. Venkatesh, et H. Bui, « Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Models, » dans *Proc. CVPR 2005*. Washington, DC, USA : IEEE Computer Society, 2005, pp. 955–960.

[95] N. Oliver, E. Horvitz, et A. Garg, « Layered Representations for Human Activity Recognition, » dans *Proc. ICMI 2002*. Washington, DC, USA : IEEE Computer Society, 2002.

[96] J. Pei, G. Dong, W. Zou, et J. Han, « On computing condensed frequent pattern bases, » dans *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, 2002, pp. 378 – 385.

[97] O. PARENT et J. EUSTACHE, « Les Réseaux Bayésiens, » Université Claude Bernard Lyon 1, Rapport de recherche, 2007.

[98] J. Pearl, *Probabilistic reasoning in intelligent systems : networks of plausible inference*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1988.

[99] R. PERRY, « Correlation versus Causality : Further Thoughts on the Law Review/Law School Liaison, » *CONNECTICUT LAW REVIEW*, vol. 39, no. 1, pp. 77–99, 2006.

[100] A. V. Petrovsky, *Psychology*. Prosvesheny, 1986.

[101] M. Philipose, K. Fishkin, M. Perkowitz, D. Patterson, D. Fox, H. Kautz, et D. Hahnel, « Inferring activities from interactions with objects, » *Pervasive Computing, IEEE*, vol. 3, no. 4, pp. 50–57, Oct.-Dec. 2004.

[102] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, et D. Hahnel, « Inferring Activities from Interactions with

Objects, » *IEEE Pervasive Computing*, vol. 3, no. 4, pp. 50–57, octobre 2004. Disponible à http://dx.doi.org/10.1109/MPRV.2004.7

[103] H. Pigot, S. Giroux, P. Mabilleau, et F. Bouchard, « L'assistance cognitive dans les habitats intelligents pour favoriser le maintien à domicile, » dans *CRIR publications, PUB-003 -Défis technologiques*, 2007, pp. 14–25.

[104] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, et M. Hsu, « Mining Sequential Patterns by Pattern-Growth : The PrefixSpan Approach, » *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1424–1440, 2004.

[105] S. B. Patil et Y. Kumaraswamy, « Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction, » *International Journal of Computer Science and Network Security*, vol. 9, no. 2, pp. 228–235, 2009.

[106] D. J. Patterson, L. Liao, D. Fox, et H. A. Kautz, « Inferring High-Level Behavior from Low-Level Sensors, » dans *Proc. Ubicomp 2003*, 2003, pp. 73–89.

[107] J. A. Palmer, D. P. Wipf, K. Kreutz-Delgado, et B. D. Rao, « Variational EM Algorithms for Non-Gaussian Latent Variable Models, » dans *NIPS*, 2005.

[108] L. R. Rabiner, « A tutorial on hidden markov models and selected applications in speech recognition, » *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[109] D. Ramage, « Hidden Markov Models Fundamentals, » Computer Science Department, University of Stanford, Rapport technique, 2007.

[110] S. P. Rao et D. J. Cook, « PREDIcting inhabitant action using action and task models with application to smart homes, » *International Journal on Artificial Intelligence Tools*, vol. 13, no. 1, pp. 81–99, 2004.

[111] P. Rashidi et D. J. Cook, « Mining Sensor Streams for Discovering Human Activity Patterns over Time, » dans *Proceedings of ICDM*, 2010, pp. 431–440.

[112] P. Rashidi, D. J. Cook, L. B. Holder, et M. Schmitter-Edgecombe, « Discovering Activities to Recognize and Track in a Smart Environment, » *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 4, pp. 527–539, 2011.

[113] L. R. Rabiner et B. H. Juang, « An introduction to hidden Markov models, » *IEEE ASSP Magazine*, pp. 4–15, January 1986.

[114] W. A. Rogers, B. Meyer, N. Walker, et A. D. Fisk, « Functional limitations to daily living tasks in the aged : a focus group analysis, » *Hum Factors*, vol. 40, no. 1, pp. 111–25, 1998.

[115] R. Srikant et R. Agrawal, « Mining Sequential Patterns : Generalizations and Performance Improvements, » dans *Proc. EDBT 1996*, P. M. G. Apers, M. Bouzeghoub, et G. Gardarin, éditeurs, vol. 1057. Springer-Verlag, 1996, pp. 3–17.

[116] S. N. Schiaffino et A. Amandi, « User profiling with Case-Based Reasoning and Bayesian Networks, » dans *IBERAMIA-SBIA 2000 Open Discussion Track*, 2000, pp. 12–21.

[117] P. Sun, S. Chawla, et B. Arunasalam, « Mining for Outliers in Sequential Databases, » dans *SIAM International Conference on Data Mining*, 2006, pp. 94–106.

[118] T. Schreiber, « Measuring Information Transfer, » *Phys. Rev. Lett.*, vol. 85, pp. 461–464, Jul 2000.

[119] P. Spirtes, C. Glymour, et R. Scheines, *Causation, Prediction, and Search, Second Edition*, 2nd édition. The MIT Press, janvier 2001.

[120] B. Shipley, « Testing Causal Explanations in Organismal Biology : Causation, Correlation and Structural Equation Modelling, » *Oikos*, vol. 86, no. 2, pp. 374–382, 1999.

[121] D. Sánchez, M. Tentori, et J. Favela, « Hidden Markov Models for Activity Recognition in Ambient Intelligence Environments, » dans *Proc. ENC 2007*, 2007, pp. 33–40.

[122] D. Sánchez, M. Tentori, et J. Favela, « Activity Recognition for the Smart Hospital, » *IEEE Intelligent Systems*, vol. 23, no. 2, pp. 50–57, 2008.

[123] O. Sarah, M. Victoria, et M. Sridhar, « Learning Hierarchical Models of Activity, » dans *Proc. IEEE/RSJ IROS 2004*, Sendai, Japan, 2004.

[124] E. M. Tapia, S. S. Intille, et K. Larson, « Activity Recognition in the Home Using Simple and Ubiquitous Sensors, » dans *Proc. Pervasive 2004*, 2004, pp. 158–175.

[125] P.-N. Tan, M. Steinbach, et V. Kumar, *Introduction to Data Mining, (First Edition)*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc., 2005.

[126] V. S. Tseng, C.-W. Wu, B.-E. Shie, et P. S. Yu, « UP-Growth : an efficient algorithm for high utility itemset mining, » dans *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, série KDD '10. New York, NY, USA : ACM, 2010, pp. 253–262. Disponible à http://doi.acm.org/10.1145/1835804.1835839

[127] T. van Kasteren et B. Krose, « Bayesian activity recognition in residence for elders, » *IET Conference Publications*, vol. 2007, no. CP531, pp. 209–212, 2007.

[128] T. van Kasteren, A. K. Noulas, G. Englebienne, et B. J. A. Kröse, « Accurate activity recognition in a home setting, » dans *Proc. UbiComp 2008*, 2008, pp. 1–9.

[129] L. T. Vinh, S. Lee, H. X. Le, H. Q. Ngo, H. I. Kim, M. Han, et Y.-K. Lee, « Semi-Markov conditional random fields for accelerometer-based activity recognition, » *Applied Intelligence*, vol. 35, no. 2, pp. 226–241, octobre 2011. Disponible à http://dx.doi.org/10.1007/s10489-010-0216-5

[130] D. L. Vail, M. M. Veloso, et J. D. Lafferty, « Conditional random fields for activity recognition, » dans *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, série AAMAS '07. New York, NY, USA : ACM, 2007, pp. 235 :1–235 :8. Disponible à http://doi.acm.org/10.1145/1329125.1329409

[131] Y. Wang, H. Bai, M. Stanton, W.-Y. Chen, et E. Y. Chang, « PLDA : Parallel Latent Dirichlet Allocation for Large-Scale Applications, » dans *AAIM*, 2009, pp. 301–314.

[132] L. Wang, T. Gu, X. Tao, H. Chen, et J. Lu, « Recognizing multi-user activities using wearable sensors in a smart home, » *Pervasive Mob. Comput.*, vol. 7, no. 3, pp. 287–298, juin 2011. Disponible à http://dx.doi.org/10.1016/j.pmcj.2010.11.008

[133] J. Wang et J. Han, « BIDE : efficient mining of frequent closed sequences, » dans *Proc. ICDE 2004*, March-2 April 2004, pp. 79–90.

[134] J. Wang, J. Han, Y. Lu, et P. Tzvetkov, « TFP : an efficient algorithm for mining top-k frequent closed itemsets, » *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 5, pp. 652 – 663, may 2005.

[135] G. I. Webb, M. J. Pazzani, et D. Billsus, « Machine Learning for User Modeling, » *User Modeling and User-Adapted Interaction*, vol. 11, pp. 19–29, March 2001.

[136] D. Wyatt, M. Philipose, et T. Choudhury, « Unsupervised activity recognition using automatically mined common sense, » dans *Proc. AAAI 2005*. AAAI Press, 2005, pp. 21–27.

[137] C. R. Wren et E. M. Tapia, « Toward scalable activity recognition for sensor networks, » dans *Proceedings of the Second international conference on Location- and Context-Awareness*, série LoCA'06. Berlin, Heidelberg : Springer-Verlag, 2006, pp. 168–185. Disponible à http://dx.doi.org/10.1007/11752967_12

[138] D. Xin, X. Shen, Q. Mei, et J. Han, « Discovering interesting patterns through user's interactive feedback, » dans *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, série KDD '06. ACM, 2006, pp. 773–778.

[139] D. Xin, X. Shen, Q. Mei, et J. Han, « Discovering interesting patterns through user's interactive feedback, » dans *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and*

data mining, série KDD '06. New York, NY, USA : ACM, 2006, pp. 773–778. Disponible à http://doi.acm.org/10.1145/1150402.1150502

[140] T. Xiong, S. Wang, Q. Jiang, et J. Z. Huang, « A New Markov Model for Clustering Categorical Sequences, » dans *ICDM*, 2011, pp. 854–863.

[141] I.-J. T. F.-C. S. Z.-N. C. H.-T. K. C.-H. T. I.-C. C. Yu-Tzu Chang, Pei-Chun Chen, « Bidirectional Relation Between Schizophrenia and Epilepsy : A Population-Based Retrospective Cohort Study, » *Epilepsia*, vol. DOI : 10.1111/j.1528-1167.2011.03268.x, 2011.

[142] J. Yang et W. Wang, « Towards Automatic Clustering of Protein Sequences, » dans *Proceedings of the IEEE Computer Society Conference on Bioinformatics*, 2002, pp. 175–186.

[143] J. Yang et W. Wang, « CLUSEQ : efficient and effective sequence clustering, » dans *Data Engineering, 2003. Proceedings. 19th International Conference on*, march 2003, pp. 101 – 112.

[144] J.-Y. Yang, J.-S. Wang, et Y.-P. Chen, « Using acceleration measurements for activity recognition : An effective learning algorithm for constructing neural classifiers, » *Pattern Recognition Letters*, vol. 29, no. 16, pp. 2213–2220, 2008.

[145] M. J. Zaki, « Scalable Algorithms for Association Mining, » *IEEE Trans. Knowl. Data Eng.*, vol. 12, no. 3, pp. 372–390, 2000.

[146] M. J. Zaki, « SPADE : An efficient algorithm for mining frequent sequences, » *Machine Learning*, vol. 42, no. 1/2, pp. 31–60, 2001.

[147] M. J. Zaki et C.-J. Hsiao, « Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure, » *IEEE Trans. on Knowl. and Data Eng.*, vol. 17, no. 4, pp. 462–478, avril 2005. Disponible à http://dx.doi.org/10.1109/TKDE.2005.60

[148] Y. Zhang, « Complex adaptive filtering user profile using graphical models, » *Inf. Process. Manage.*, vol. 44, pp. 1886–1900, November 2008. Disponible à http://dl.acm.org/citation.cfm?id=1453256.1453388

[149] Y. Zheng, L. Liu, L. Wang, et X. Xie, « Learning transportation mode from raw gps data for geographic applications on the web, » dans *Proceedings of WWW*, 2008, pp. 247–256.

[150] P. Zigoris et Y. Zhang, « Bayesian adaptive user profiling with explicit &amp ; implicit feedback, » dans *Proceedings of CIKM*, 2006, pp. 397–404.

[151] X. Zhang, X. Zhou, H. Huang, S. Chen, et B. Liu, « A hierarchical symptom-herb topic model for analyzing traditional Chinese medicine clinical diabetic data, » dans *Biomedical Engineering and Informatics (BMEI), 2010 3rd International Conference on*, vol. 6, oct. 2010, pp. 2246 –2249.