De la capture de trajectoires de visiteurs vers l'analyse interactive de comportement après enrichissement sémantique

Jérémy Richard

▶ **To cite this version:**

Jérémy Richard. De la capture de trajectoires de visiteurs vers l'analyse interactive de comportement après enrichissement sémantique. Base de données [cs.DB]. Université de La Rochelle, 2023. Français. NNT : 2023LAROS012 . tel-04313594

HAL Id: tel-04313594
https://theses.hal.science/tel-04313594

Submitted on 29 Nov 2023

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

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

UNIVERSITÉ DE LA ROCHELLE

ÉCOLE DOCTORALE EUCLIDE

LABORATOIRE : L3I

THÈSE présentée par :

Jérémy Richard

soutenue le : **22 mai 2023**

pour obtenir le grade de : **Docteur de l'université de La Rochelle**

Discipline : **Informatique et applications**

De la capture de trajectoires de visiteurs vers l'analyse interactive de comportement après enrichissement sémantique

---

Rapporteurs	Marianne Huchard	Professeure des universités	LIRMM, Université de Montpellier
	Thomas Guyet	Chargé de recherche HDR	Inria, Lyon,
Examinateurs	Peggy Cellier	Maîtresse de conférences HDR	Irisa, INSA Rennes
	Jérôme Gensel	Professeur des universités	IMAG, Université Grenoble-Alpes
	Sergei Kuznetsov	Professeur des universités	ILISSA, HSE University, Moscou
	Jean-Loup Guillaume	Professeur des universités	L3i, La Rochelle Université
Invités	Christophe Demko	Maître de conférences	L3i, La Rochelle Université
	Christophe Bortolaso	Ingénieur R&D	Entreprise Berger-Levrault, Toulouse
Direction	Karell Bertet	Professeure des universités	L3i, La Rochelle Université
	Cyril Faucher	Maître de conférences	L3i, La Rochelle Université

Remerciement

Je tiens à prendre un moment pour exprimer ma reconnaissance et ma gratitude envers toutes les personnes qui ont contribué à la réussite de cette thèse et m'ont soutenu tout au long de ce parcours.

Je commence par exprimer ma profonde gratitude envers ma directrice de thèse, Karell Bertet. Sa patience, son expertise, son dévouement et sa connaissance scientifique ont été une source constante de motivation et de soutien pour moi. Grâce à ses précieux conseils, sa disponibilité et sa passion pour la recherche, j'ai pu accomplir cette étape importante de ma vie académique.

Je voudrais également remercier Cyril Faucher pour sa contribution, ses conseils éclairés et son expertise sur les diverses questions administratives auxquelles je me heurte trop souvent.

Je tiens à exprimer ma gratitude envers Christophe Demko pour son soutien tout au long de mon parcours, même s'il n'a pas été directement impliqué dans l'encadrement de ma thèse. Ses conseils avisés et ses discussions stimulantes ont contribué à nourrir mes réflexions et à approfondir ma compréhension des concepts fondamentaux.

De plus, je souhaite adresser ma sincère reconnaissance aux rapporteurs de ma thèse, qui ont consacré leur temps et leurs compétences à évaluer mon travail de recherche. Je suis reconnaissant de l'attention qu'ils ont portée à mon travail et je suis honoré d'avoir pu bénéficier de leurs précieux conseils.

Je tiens également à remercier chaleureusement toute l'équipe GALACTIC pour leur collaboration et leur esprit d'équipe. Leur expertise, leurs idées novatrices et leur soutien mutuel ont créé un environnement de recherche dynamique et stimulant.

Je tiens également à souligner l'importance de toutes les personnes associées au projet DA3T pour leur contribution et leur bonne humeur. Leurs compétences complémentaires et pluridisciplinaires m'ont permis d'explorer de nouveaux horizons et d'ouvrir de nouvelles pistes de recherche, ce qui a grandement enrichi ma thèse.

Je tiens également à exprimer ma gratitude envers le laboratoire L3I pour avoir rendu possible cette expérience de recherche. Leur soutien financier et logistique a été essentiel pour la réussite de ce projet.

Mes collègues de laboratoire ont été une source d'inspiration constante. Leurs échanges intellectuels, leur collaboration et leur camaraderie ont rendu cette expérience de recherche plus enrichissante et agréable.

Je tiens à exprimer ma profonde gratitude envers l'entreprise Berger-Levrault pour leur soutien financier. Leur contribution a permis de financer les ressources, les équipements et les déplacements nécessaires à la réussite

de cette étape importante de ma carrière.

Je suis également reconnaissant envers mes amis et ma famille, en particulier Marina et Juliana, pour leur soutien inconditionnel, leurs encouragements constants et leur amour durant les moments difficiles. Cette thèse est pour toi, maman.

Enfin, je souhaite remercier tous les chercheurs, les auteurs et les professionnels dont les travaux ont nourri ma réflexion et ont façonné ma compréhension du sujet de ma thèse.

Chacune de ces personnes a contribué de manière significative à ma réussite et je leur suis profondément reconnaissant pour leur soutien indéfectible.

Merci du fond du cœur à tous ceux qui ont fait partie de cette aventure de recherche avec moi et m'ont permis d'accomplir cette étape importante de ma vie.

De la capture de trajectoires de visiteurs vers l'analyse interactive de comportement après enrichissement sémantique

Richard Jérémy

Mars 2023

Table des matières

1 Introduction **11**
1.1 Introduction générale 12
1.2 Contexte de la thèse 14
1.2.1 Le projet DA3T 14
1.2.2 Le projet JPeuxPasJMusée 15
1.3 Problématiques et contributions 15
1.3.1 Contribution 1 : Reconstruction de trajectoire en environnement indoor contraint 16
1.3.2 Contribution 2 : Modèle multi-aspects générique d'enrichissement sémantique 16
1.3.3 Contribution 3 : Analyse multi-séquences et hétérogène 18
1.3.4 Contribution 4 : Vers une analyse interactive avec "ReducedContextCompletion" 19
1.4 Plan du manuscrit 19

I Etat de l'art **23**

2 Etat de l'art : Des données de capteurs aux trajectoires sémantiques **24**
2.1 Introduction 26
2.2 Des capteurs aux trajectoires : reconstruction de trajectoire en intérieur 27
2.2.1 Les différents systèmes d'indoor positionning 27
2.2.2 Les algorithmes de reconstruction de trajectoires 29
2.2.3 Les travaux notables qui s'appliquent aux musées . . . 32
2.2.4 Conclusion 33
2.3 Des trajectoires aux trajectoires sémantiques 33
2.3.1 Les informations contextuelles 33
2.3.2 Les modèles de trajectoires sémantiques 34
2.3.3 Conclusion 37

2.4 Des trajectoires sémantiques aux comportements 37
2.4.1 Introduction 37
2.4.2 La fouille de sous-séquences fréquentes : Le *sequence mining* 38
2.4.3 Fouille de séquences temporelles 40
2.4.4 Fouille de séquences d'intervalles temporels 42
2.4.5 Conclusion 43
2.5 Conclusion 44

3 Etat de l'art : Pattern mining et Analyse Formelle de Concepts 47
3.1 Introduction 49
3.2 *Pattern mining* 50
3.2.1 Notations et vocabulaire 50
3.2.2 La fouille de motifs fréquents 51
3.2.3 Les motifs fermés 52
3.3 L'Analyse Formelle de Concepts 53
3.3.1 Contexte et concepts 54
3.3.2 Treillis des concepts 54
3.3.3 Les structures de motifs 55
3.4 Théorie des treillis 57
3.4.1 Ensemble partiellement ordonné 57
3.4.2 La structure de treillis 58
3.4.3 Les familles de Moore 59
3.4.4 Treillis des fermés 60
3.4.5 La table binaire d'un treillis 61
3.5 L'algorithme NextPriorityConcept 62
3.5.1 Algorithme NextPriorityConcept 63
3.6 Conclusion 70

II Contributions 72

4 Présentation des jeux de données 74
4.1 Introduction 75
4.2 Museum d'histoire naturelle de La Rochelle 75
4.2.1 Notations : 75
4.2.2 Le contexte des expérimentations ***Museum_ 1*** et ***Museum_ 2*** 77
4.2.3 Dataset ***Museum_ 1*** - Expérimentation mai 2019 : . . 78
4.2.4 Dataset ***Museum_ 2*** - Expérimentation janvier 2021 : 79

4.2.5 Dataset ***Museum_ 3*** - Museum d'histoire naturelle : septembre 2021 . . . 81
4.2.6 Le système de micro-localisation . . . 81
4.3 Dataset : ***Geoluciole*** - Collecte de trajectoires touristiques . . 85
4.4 Dataset : ***La Cité du Vin*** . . . 86
4.5 Synthèse de la problématique pour chaque jeu de données . . . 90

5 Contribution 1 : Reconstruction de trajectoires en environnement indoor contraint 92
5.1 Introduction . . . 93
5.2 Traitement des données indoor . . . 95
5.2.1 Lissage et filtrage . . . 96
5.3 L'algorithme GraphPositionning : pour une reconstruction de la trajectoire à gros grain . . . 97
5.3.1 Le principe de l'algorithme . . . 98
5.3.2 Visualisation des visites capturées . . . 100
5.3.3 Conclusion . . . 101
5.4 L'algorithme Minmal Zone Searching (MZS) : pour une reconstruction plus fine . . . 102
5.4.1 Calcul des distances . . . 102
5.4.2 Le principe de l'algorithme . . . 103
5.4.3 Expérimentations . . . 106
5.4.4 Conclusion . . . 112
5.5 Conclusion . . . 113

6 Contribution 2 : Modèle multi-aspects générique d'enrichissement sémantique 115
6.1 L'enrichissement de trajectoires . . . 116
6.2 Un modèle d'enrichissement générique multi-aspects . . . 117
6.2.1 Les aspects . . . 117
6.2.2 Hiérarchie de trajectoires sémantiques . . . 118
6.3 Modélisation et visualisation de trajectoires enrichies : Application sur des données réelles . . . 121
6.3.1 Enrichissement du dataset **Geoluciole** . . . 121
6.3.2 Enrichissement du dataset **Geoluciole** par le biais d'un entretien . . . 124
6.3.3 Enrichissement du dataset **Museum_3** avec des données d'application . . . 129
6.4 Conclusion . . . 133

7 Contribution 3 : Analyse multi-séquences et hétérogène 137
7.1 Introduction . . . 138
7.2 Analyse des trajectoires avec **GALACTIC** . . . 139
7.2.1 Impact des stratégies sur le déluge de motifs . . . 143
7.3 Expérimentations . . . 146
7.3.1 Détection de comportements particuliers dans les trajectoires de visites des musées . . . 148
7.3.2 Détection de comportements particuliers dans les visites de la ville : Le choix des données . . . 152
7.4 Conclusion . . . 156

8 Contribution 4 : Vers une analyse interactive avec Reduced Context Completion 158
8.1 Introduction . . . 159
8.2 Algorithme de complétion d'un contexte réduit . . . 160
8.2.1 Formalisation de la problématique de complétion . . . 160
8.2.2 Description de l'algorithme . . . 162
8.2.3 Exemple . . . 167
8.3 Preuve de l'algorithme . . . 169
8.4 Expérimentations . . . 171
8.4.1 Construction itérative du treillis des concepts de la cité du vin . . . 171
8.4.2 Comparaison avec l'algorithme NextPriorityConcept . . . 174
8.4.3 Impact de l'ordre d'insertion . . . 177
8.5 Conclusion et perspectives . . . 181

9 Conclusion 183
9.1 Bilan . . . 184
9.1.1 Reconstruction de trajectoires en environnement contraint 184
9.1.2 Enrichissement sémantique des trajectoires . . . 185
9.1.3 Analyse de trajectoires sémantiques . . . 185
9.1.4 Analyse interactive . . . 186
9.2 Perspectives . . . 187

Table des figures

1.1 Schéma de la chaîne de traitement proposée dans ce manuscrit 22

2.1 Signaux récupérés par un émetteur durant une déambulation . 28
2.2 Système d'indoor positionning 28
2.3 Schéma représentant le principe de la triangulation [Wang et al., 2013] 30
2.4 Schéma représentant le principe de la trilatération [Pradhan et al., 2019] 31
2.5 Trajectoire sémantique sous la forme d'épisodes de *stop* et de *move* 35
2.6 Exemple de fouille de motifs temporels par épisode, Ft_1 42
2.7 Exemple de fouille de motifs temporels par épisode, Ft_2 42
2.8 Les 7 relations de base d'Allen, entre deux intervalles A et B . 44

3.1 Motifs de D du tableau 3.1 sous forme hiérarchique. La valeur de support est indiquée au-dessus des noeuds 52
3.2 Diagramme de Hasse du treillis de concepts $L = (C, \leq)$ de l'exemple 3.3.2 56
3.3 Diagramme de Hasse du Treillis L. 59
3.4 Diagramme de Hasse du treillis L de la famille de Moore $\mathcal{F} = \mathcal{P}(S)$ sur l'ensemble S = {a,b,c,d} 60
3.5 Exemple : Un treillis L - join-irréductible ■ meet-irréductible ■ 61
3.6 Contexte réduit du treillis de la figure 3.5 et son treillis de concepts associé 62
3.7 Le processus de l'algorithme NextPriorityConcept [Demko et al., jun 2022] 67
3.8 Treillis généré avec la stratégie s^1 68
3.9 Treillis généré avec les deux stratégies s^1 et s^2 69
3.10 Schéma récapitulatif des contributions et agencement des chapitres 73

4.1 Architecture technique de collecte de déplacements 78
4.2 Entretien avec un visiteur durant la nuit des musées 79
4.3 Placement des raspberries au sein du muséum durant les expérimentations de janvier 2019 et 2021 80
4.4 Processus d'envoi des données pendant l'expérimentation . . . 82
4.5 Plan du placement des balises dans le muséum 83
4.6 Vue accueil 84
4.7 Vue plan 84
4.8 Vue détail 84
4.9 Logo de l'application Geoluciole 85
4.10 Échantillon de traces GPS récoltées durant l'expérimentation . 86
4.11 La Cité du Vin, musée phare de Bordeaux 87
4.12 Le plan de la Cité du Vin, Bordeaux 87

5.1 Processus de pré-traitement de données indoor 95
5.2 L'impact de la moyenne glissante sur les données brutes récoltées 96
5.3 Le graphe des possibles au sein du muséum où chaque noeud est un émetteur 99
5.4 instantané de la trajectoire d'un visiteur (1) 101
5.5 instantané de la trajectoire du visiteur (2) 101
5.6 instantané de la trajectoire du visiteur (3) 101
5.7 instantané de la trajectoire du visiteur (4) 101
5.8 Recalibrage du point calculé en fonction de la limite de la pièce 103
5.9 Enveloppe convexe des points fixes actifs à n=0 105
5.10 Évolution de MZS à chaque n exécution 107
5.11 L'algorithm MZS appliqué à **DS1** 108
5.12 Comparaison entre MZS et la triangulation 110
5.13 Propriétés des zones calculées par les deux méthodes 111
5.14 Comparaison visuelle entre la triangulation et MZS , avec les données du badge numéro 8 sur ***Museeum_2***. 112

6.1 Représentation de trajectoires sémantiques 118
6.2 Diagramme du modèle d'enrichissement sémantique multi-aspects 119
6.3 Représentation de la hiérarchie de trajectoires sémantiques . . 120
6.4 Exemple de trajectoire sémantique avec ses aspects 122
6.5 Modèle sous la forme d'un diagramme de classe. 123
6.6 Visualisation d'une partie de séquence d'épisodes à différents niveaux 127
6.7 Représentation des 4 niveaux de la séquence d'épisodes 127
6.8 Modélisation de la trajectoire sémantique au sein du muséum . 129

6.9 Données d'application mises en relation avec les données de déplacement avec le temps 129
6.10 Visualisation du déplacement d'un visiteur enrichie des données de l'application mobile 131
6.11 Instantané de la trajectoire d'un visiteur (1) 132
6.12 Instantané de la trajectoire du visiteur (2) 132
6.13 Instantané de la trajectoire du visiteur (3) 132
6.14 Instantané de la trajectoire du visiteur (4) 132
6.15 Exemple des différents aspects d'une trajectoire de **Geoluciole** 135

7.1 Treillis L avec la description numérique SND et la stratégie σ_Q SNS 144
7.2 Treillis obtenu pour le dataset D (table 3.3) avec la description d'intervalle temporel δ_{SCMAX}, la description numériqueSND, la stratégie σ_Q SNS et la stratégie σ_{CMA} 145
7.3 Nombre de concepts générés en fonction des attributs ajoutés successivement 147
7.4 Prédicats de 3 concepts du treillis obtenu à partir du dataset **Géoluciole** avec les 5 aspects 148
7.5 Zoom sur le treillis des concepts avec le dataset de la **Cité du Vin** 149
7.6 Arbre de décision construit à partir du treillis des concepts de la figure 7.5 150
7.7 Treillis des concepts du dataset **Museum_3** 153
7.8 Zoom sur la partie du treillis montrant l'impact de la pluie sur les trajets 154
7.9 Échantillon des prédicats montrant les quartiers de résidence touristique de séjour en fonction du type de statut 156

8.1 Ajout de l'élément (c) qui fait apparaître a comme le nouvel élément $\top$ 165
8.2 Ajout de l'élément (h) qui fait apparaître i comme le nouvel élément $\bot$ 165
8.3 Illustration de l'exécution de l'algorithme *ReducedContextCompletion* 168
8.4 Insertion d'un visiteur de la Cité du Vin 172
8.5 Insertion d'un deuxième visiteur 172
8.6 Insertion d'un troisième visiteur 173
8.7 Insertion d'un quatrième visiteur 174
8.8 Exemple de treillis de nombre entier 175
8.9 Description du jeu de données des iris 178

8.10 Insertion de singletons classe par classe 179
8.11 Insertion de singletons dans un ordre aléatoire 180
8.12 Temps d'exécution de chaque boucle à l'insertion d'un élément suivant le mode d'insertion 181
8.13 Temps d'exécution total suivant le mode d'insertion 181

9.1 Les problématiques abordées dans ce manuscrit 183
9.2 Architecture de la plateforme **GALACTIC** 206

Liste des tableaux

3.1 Ensemble de données D de la figure 3.1 51
3.2 Contexte (G, M, I) 56
3.3 Exemple de données catégorielles et numériques 64

4.1 Exemple d'un dataset d'une expérimentation au muséum . . . 77
4.2 Description de ***Museum_2***. 80
4.3 Description des individus de ***Museum_2***. 81
4.4 Extrait des données récoltées dans le dataset **Museum_3** pour un visiteur 85
4.5 Formalisation des données pour chaque dataset 90
4.6 Tableau comparatif des dataset 91

5.1 Description de **DS1**. 107
5.2 Valeurs de précision entre l'algorithme MZS et la triangulation appliqué à **DS1** 109
5.3 Nombres d'images calculés par MZS et la triangulation. 110

6.1 Ontologie du modèle de représentation de trajectoire sémantique 126
6.2 Aspects du dataset **Geoluciole** 135
6.3 Aspects du dataset **Museum_3** 136

7.1 Descriptions et stratégies utilisées pour chaque type de données 143
7.2 Dataset D 144
7.3 Échantillon des concepts du treillis de la Fig 7.7 151
7.4 Échantillon des prédicats montrant l'impact de la météo sur les déplacements à la plage 154

8.1 Liste des irréductibles, du maximum et du minimum de chaque sous-treillis de L pour $X = \{d, e, b, c\}$ 167
8.2 Comparaison des temps de calcul en secondes de NextPriorityConcept et ReducedContextCompletion 177
8.3 Tableau des descriptions et caractéristiques utilisées 178

Chapitre 1

Introduction

Table des matières

1.1 Introduction générale 12
1.2 Contexte de la thèse 14
1.2.1 Le projet DA3T 14
1.2.2 Le projet JPeuxPasJMusée 15
1.3 Problématiques et contributions 15
1.3.1 Contribution 1 : Reconstruction de trajectoire en environnement indoor contraint 16
1.3.2 Contribution 2 : Modèle multi-aspects générique d'enrichissement sémantique 16
1.3.3 Contribution 3 : Analyse multi-séquences et hétérogène 18
1.3.4 Contribution 4 : Vers une analyse interactive avec "ReducedContextCompletion" 19
1.4 Plan du manuscrit 19

1.1 Introduction générale

Après avoir posé la question *"Pourquoi êtes-vous resté aussi longtemps dans cette salle ? Elle n'est pas si grande."*, la personne interrogée lors d'une expérimentation de visualisation de déambulation au muséum d'histoire naturelle de La Rochelle - répondit *"J'y ai rencontré un ami, on a pas mal discuté"*. Dans ce type d'expérience, l'objectif est de retracer le parcours d'un individu dans un espace, ici dans un musée, afin d'identifier des comportements représentatifs des visiteurs. L'exploitation de ces résultats pourra ensuite servir de base à la modification de la topologie d'un lieu en fonction des zones d'influence détectées ; cette détection a l'ambition de retranscrire efficacement les dynamiques et utilisations diverses intrinsèques à un lieu. Mais les déplacements seuls permettent-ils de réellement comprendre le comportement d'un visiteur ? Peut-on tirer une quelconque conclusion en se basant uniquement sur des données de mobilité ? Comme peut l'illustrer cet exemple, les données de mobilité, sans aucun contexte ou expertise, peinent à représenter efficacement le réel. Une zone où l'on passe du temps ne signifie

pas toujours qu'il s'agit d'une zone intéressante pour un usager.

Forts de cette observation, des chercheurs [Ilarri et al., 2015] [Abdelnasser et al., 2015] ont utilisé la trajectoire sémantique qui permet d'enrichir une trajectoire brute avec des informations supplémentaires relatives à un comportement ou toute sorte de données extérieures apportant un éclaircissement sur une trajectoire [Parent et al., 2013]. L'émergence des recherches autour de la trajectoire sémantique a mis en exergue deux sujets fondateurs dans les études de mobilité : la volonté de croiser plusieurs types de données afin de caractériser d'autant plus les trajectoires et l'importance de l'expertise sur le jeu de données manipulé. Avec l'explosion des bases de données en ligne et l'accessibilité grandissante des moyens de récolte d'informations, un des enjeux clés des travaux de recherche en géographie consistera ainsi à coupler des données externes avec des traces de mobilité afin de préciser celles-ci. Notre travail s'inscrit ainsi dans une démarche de développement de méthodes et d'outils permettant aux experts géographes de pouvoir facilement traiter et analyser ces formats de données différents, ainsi que de réaliser des analyses croisées de trajectoires humaines. Ainsi, un problème reste en suspens, comment analyser ces trajectoires complexes ? Comment, par le biais d'une analyse, prendre en compte des données de localisation couplées avec des données externes dans le but d'extraire des comportements significatifs ? Comment permettre à l'analyste, ici l'expert en géographie ou le gestionnaire de musée, de comprendre les comportements identifiés et d'interroger ces données pour les affiner ou les préciser ?

Dans ces travaux de recherche, nous nous intéressons aux trajectoires touristiques et à la façon dont les usagers utilisent l'espace qui leur est proposé. Par trajectoires touristiques, nous entendons tous les types de pratiques liés à l'activité touristique, aussi bien la visite de la ville, que celle d'un musée. Nous proposerons à travers nos travaux des approches pour étudier et analyser les comportements de mobilité des usagers. Nous aurons ici une approche horizontale, traversant plusieurs domaines de recherche bien différents afin de proposer un processus complet autour de la notion de trajectoire, de la reconstruction de celle-ci jusqu'à l'extraction interactive de comportements significatifs.

Notre approche se veut générale et traitera à la fois des trajectoires dites *"classiques"* en extérieur, via des coordonnées GPS, mais aussi en intérieur où les signaux GPS ne sont pas utilisables. Cependant, les trajectoires des individus en intérieur devront d'abord être retrouvées. Pour ce faire, nous devrons reconstruire la trajectoire des visiteurs via des techniques "*d'indoor*

positionning" [Li, 2008] (ie. positionnement en intérieur). Cette reconstruction de trajectoire sera appuyée par des expérimentations sur le terrain, permettant de constituer notre propre jeu de données de trajectoires. Par la suite, nous enrichirons ces trajectoires avec des informations contextuelles, comme expliqué ci-avant, afin de proposer une représentation de la trajectoire plus proche de la réalité et ne se reposant pas seulement sur un aspect spatial, mais aussi sur d'autres aspects. Ici, nous appelons aspects toutes les autres informations de contexte que l'on peut ajouter à la trajectoire brute. En effet, nos travaux étant axés sur les approches et méthodes informatiques pour l'étude des comportements touristiques, le contexte où s'effectuent ces pratiques est tout aussi important et doit être pris en compte au même titre que la mobilité. Nous proposerons en bout de chaîne une méthode d'analyse de ces traces enrichies qui apporte une explicabilité et qui peut s'envisager dans une approche interactive.

Dans ce chapitre, nous développerons le contexte de la réalisation de cette thèse, qui se positionne dans le projet régional DA3T. Nous détaillerons ensuite toutes les problématiques traitées dans ce manuscrit et en présenterons le plan.

1.2 Contexte de la thèse

1.2.1 Le projet DA3T

Cette thèse est menée dans le cadre du projet régional Nouvelle-Aquitaine **DA3T** (**D**ispositif d'**A**nalyse des **T**races numériques pour la valorisation des **T**erritoires **T**ouristiques) [1]. Comme l'indique son nom, l'objectif du projet est de proposer un dispositif d'analyse des traces de mobilité dans le but d'aider les aménageurs et décideurs locaux dans la gestion et la valorisation des territoires touristiques en Nouvelle-Aquitaine que ce soit au niveau de la ville ou des bâtiments comme les musées. Il s'agit d'un projet pluridisciplinaire réunissant informaticiens et géographes dans le but de produire des outils et des méthodes de traitement et d'analyse de traces de mobilité. Le projet s'articule autour de trois thèses, à savoir :

— Une thèse en géographie, soutenue par Mélanie Mondo en mars 2022, sur l'apport des traces de mobilité couplées à des entretiens de touristes volontaires dans l'étude et la compréhension des comportements touristiques [Mondo, 2022] ;

1. Site DA3T : https ://lienss.univ-larochelle.fr/DA3T

— une thèse en informatique, soutenue par Cécile Cayèré en novembre 2022, sur la conception d'outils dédiés au traitement des traces de mobilité [Cayèré, 2022] ;

— Une seconde thèse en informatique, réalisée par moi-même, visant à fournir des outils et une méthodologie d'analyses génériques permettant de travailler simultanément en intérieur et en extérieur, par rapport aux deux thèses précédemment mentionnées.

1.2.2 Le projet JPeuxPasJMusée

"J'peux pas, j'ai musée !" est un consortium regroupant des partenaires institutionnels du monde culturel et universitaire (enseignement supérieur et recherche). Celui-ci vise à concevoir des produits numériques pour les musées en tirant parti des compétences respectives de chaque acteur. Ce processus de co-construction collective s'inscrit dans une démarche pédagogique et professionnalisante pour les étudiants (notamment de l'IUT de La Rochelle) et stagiaires qui y contribuent chaque année depuis 2014. Il en ressort une expérience concrète de développement de produits et de travail en équipe, en répondant aux besoins réels des musées impliqués tout au long du processus de création. Dans le cadre de ce projet, une application mobile "Visite Musée" a été conçue par les étudiants de la licence professionnelle Développement Mobile de l'IUT de La Rochelle, dans le but d'accompagner les visiteurs lors de leur visite des musées équipés de la région Nouvelle-Aquitaine. Cette application permet aux visiteurs d'avoir accès à des informations complémentaires sur les œuvres exposées, les artistes, les techniques utilisées et l'histoire des musées, tout en offrant une expérience de visite interactive et personnalisée. Le laboratoire L3i participe activement à ce consortium depuis sa création, notamment dans le cadre d'expérimentations pour le suivi à la fois des déplacements et de l'activité de visiteurs lors de leur déambulation dans un musée.

1.3 Problématiques et contributions

Dans ce manuscrit, nous répondrons à un certain nombre de problématiques afin de proposer un processus global de traitement de trajectoires, allant de la reconstruction de celles-ci jusqu'à la proposition d'un outil d'analyse de trajectoires enrichies orienté utilisateur, s'appuyant sur l'analyse formelle de concept et la théorie des treillis. Dans cette section, nous faisons un état des lieux des thématiques abordées durant nos travaux.

1.3.1 Contribution 1 : Reconstruction de trajectoire en environnement indoor contraint

Notre première étape sera de reconstruire une trajectoire par le biais de données provenant d'un système de localisation en intérieur où les signaux GPS sont détériorés voir absents. Il est cependant possible de capter des déplacements à l'aide d'émetteurs et de capteurs adéquats. Pour ce faire, l'équipement de localisation en intérieur devra être conçu en prenant en compte les contraintes liées à l'environnement. Les musées représentent un terrain d'expérimentation idéal pour cela, avec des contraintes liées au positionnement des œuvres présentées, une déambulation des visiteurs lente et sur une longue durée, ainsi que des gestionnaires et visiteurs impliqués. Cela implique de minimiser les perturbations sur l'expérience de visite des visiteurs. Cela peut avoir un impact négatif sur les méthodes de localisation classiques. Le système d'indoor positioning doit donc être adapté spécifiquement aux musées, en considérant les potentiels problèmes associés, et la méthode de suivi des visiteurs doit être applicable à ces environnements contraints de manière générique.

Nous proposons deux méthodes de reconstruction de trajectoires de visiteurs dans un musée :

— Une première méthode dite à “gros grain" permet de reconstruire toute la visite d'un individu au sein de cet espace en utilisant un graphe des déplacements possibles.

— Une seconde méthode avec une granularité plus fine pour proposer une estimation plus précise des zones de présence d'un individu au cours de son déplacement.

Nous présenterons une visualisation pour chaque méthode qui permettra de représenter la trajectoire du visiteur à l'intérieur de cet espace.

Les trajectoires relatant du déplacement dans la ville ont été récupérées quant à elles sous forme de signaux GPS par le biais d'une application et la reconstruction de celles-ci sera abordée chapitre 4 *“Présentation des jeux de données"*.

1.3.2 Contribution 2 : Modèle multi-aspects générique d'enrichissement sémantique

Une fois la trajectoire reconstruite, nous nous intéresserons aux méthodes d'enrichissement d'une trajectoire brute en différentes trajectoires séman-

tiques. Ceci afin de compléter les trajectoires reconstruites avec des informations diverses, telles que les quartiers visités, la météo en extérieur, les salles traversées, ou encore des données d'application mobile pour un musée en intérieur. Comme notre méthode est générique elle s'inscrit parfaitement dans le projet DA3T et est suffisamment versatile pour être appliquée aussi bien sur des études en intérieur qu'en extérieur. Cette méthode d'enrichissement pourra ainsi être appliquée à différents types de trajectoires. Notre problématique ici est de proposer un modèle de structure de données pour les trajectoires sémantiques qui permette de représenter des trajectoires enrichies avec des données valables à la fois en intérieur et en extérieur. Ces trajectoires peuvent être définies sur plusieurs aspects d'enrichissement, dont un aspect classique est le positionnement spatial, qui peut correspondre à des salles dans un musée, à la proximité d'une œuvre, à des quartiers dans une ville, à un POI (Point d'Intérêt), à une plage, un parc, etc.

Nous nous intéressons ainsi ici à plusieurs jeux de données : des trajectoires touristiques dans la ville de La Rochelle et trois datasets retraçant des visites dans les musées. Chacun de ces jeux de données a ses propres spécificités qui seront décrites dans le chapitre 4 " Présentation des jeux de données ".

— Nous nous intéresserons à l'enrichissement des trajectoires reconstituées dans le musée avec des données de logs d'une application mobile, "Visite musée", afin de proposer un véritable journal de bord de l'activité du visiteur durant la visite dans le musée.

— Dans le cadre d'un travail interdisciplinaire alliant informatique et géographie, nous appliquerons ce modèle d'enrichissement sémantique aux trajectoires de déplacements de touristes à La Rochelle. Pour ce faire, nous mènerons une campagne d'entretiens semi-directifs pour recueillir le discours des personnes ayant réalisé ces trajectoires, qui constituera notre jeu de données. Ainsi enrichies, ces trajectoires seront étudiées sous un angle plus qualitatif et permettront d'approfondir notre compréhension des parcours touristiques à La Rochelle.

— Enfin, nous viendrons compléter le jeu de données des trajectoires de La Rochelle par le biais d'enrichissements spatiaux des quartiers, mais aussi de données provenant de bases en ligne afin d'inclure des aspects tels que la météo, les marées ainsi que plusieurs autres données provenant d'une application mobile appelée "Géoluciole", qui nous renseigne sur le profil de la personne (habituation à la ville ou non, moyen de locomotion, visite en famille ou seul, etc).

1.3.3 Contribution 3 : Analyse multi-séquences et hétérogène

Dans notre étude sur le comportement des individus, nous introduisons une méthodologie pour analyser les trajectoires enrichies. Ces dernières sont complexes à étudier en raison de la multitude d'informations qu'elles renferment, composées toutes de plusieurs trajectoires enrichies avec des informations temporelles. Nous définissons le comportement comme étant des sous-groupes d'individus partageant un même déplacement ou un même comportement. Cela peut être vu par exemple, à travers l'analyse de sous-groupes de personnes qui se rendent à la plage en même temps et l'étude de leurs comportements communs en termes temporalité (heure, durée) ou en les combinant avec d'autres aspects tels que la météo ou la marée pour comprendre les interactions et influences entre ces différents facteurs.

Tous les aspects ajoutés à la trajectoire dans la contribution précédente peuvent être exprimés sous forme de séquences temporelles d'intervalles. Dans cette représentation, chaque événement qui compose l'aspect est caractérisé par des informations sémantiques temporelles telles qu'une date de début et une date de fin, ainsi que l'information proprement dite. Le problème de l'analyse du comportement est étroitement lié à la découverte de sous-groupes d'individus ayant des attributs communs. Ce comportement commun se décrit par une sous-séquence commune de déplacement (par exemple : aller à la plage), mais aussi par une sous-séquence commune d'un autre aspect de la trajectoire enrichie (par exemple : la météo, la marée). Une telle analyse relève de l'analyse de données non supervisée et plus précisément du clustering hiérarchique, où chaque cluster correspond à un groupe de données décrit par un motif commun qui apporte une explicabilité à l'analyste. Parmi les approches de clustering, le sequence-mining s'intéresse aux données séquentielles mais ne permet pas de considérer plusieurs séquences simultanément pour chaque individu. Nous avons choisi la plateforme d'analyse formelle de concept **GALACTIC** qui permet l'analyse des données complexes et hétérogènes telles que ces trajectoires enrichies.

Après l'enrichissement de trajectoires, nous proposerons plusieurs expérimentations portant à la fois sur les déplacements des individus et leurs données contextuelles associées. Celles-ci porteront aussi bien sur des données obtenues dans les musées que dans la ville avec une approche générique de la problématique.

1.3.4 Contribution 4 : Vers une analyse interactive avec "ReducedContextCompletion"

Les travaux présentés ici abordent un enjeu commun pour les informaticiens et les géographes : l'analyse de trajectoires. Cette question réunit les deux domaines pour trouver une solution qui soit compréhensible pour les géographes tout en utilisant les méthodes avancées proposées par les informaticiens. Certaines méthodes proposées par les informaticiens peuvent sembler complexes pour les géographes qui sont mieux à même de comprendre la sémantique des données. L'objectif final est de développer un outil d'analyse facile à utiliser pour l'analyste des données, qui permettra de choisir les critères sur lesquels il souhaite se concentrer. Les trajectoires enrichies sont des structures de données complexes comprenant des dimensions spatiales, temporelles et contextuelles différentes. Le but est de proposer une méthode interactive d'exploration sémantique des données pour permettre à l'analyste de sélectionner la sémantique qui l'intéresse. Par exemple, introduire la "marée" n'a de sens que pour les sous-groupes allant à la "plage". Cette méthode permettra également de limiter la volumétrie des ressources au niveau des calculs mais aussi des données, en ne prenant en compte que les données contextuelles pertinentes pour un sous-groupe particulier, ce qui s'inscrit dans une démarche de numérique responsable.

L'analyse de ce type de données et l'extraction de comportements significatifs nécessitent une expertise fine des données. Notre travail consistera à personnaliser le traitement pour l'analyste des données afin qu'il puisse orienter les axes d'analyse en utilisant les aspects contextuels qui l'intéressent.

Dans le cadre de la création d'un outil d'analyse interactif permettant une sélection des données de contexte en fonction des objectifs d'analyse de l'utilisateur, nous avons conçu un algorithme, *ReducedContextCompletion*, basé sur les principes fondamentaux de la théorie des treillis, sur le même modèle que **GALACTIC**, mais offrant une fonctionnalité supplémentaire qui n'est pas disponible avec ce dernier. Ce projet s'inscrit dans une démarche visant à offrir une navigation interactive au sein d'un treillis de concepts et n'a pour l'heure pas d'équivalent parmi les algorithmes existants.

1.4 Plan du manuscrit

Ce manuscrit sera organisé en deux parties :

Partie 1. La première partie se focalisera sur un état de l'art des divers domaines dans lesquels notre travail s'inscrit. Cet état de l'art sera lui-même divisé en deux parties distinctes.

— Le chapitre 2 est un état de l'art du domaine des trajectoires. Premièrement, nous étudierons la façon dont il est possible de reconstruire une trajectoire en intérieur. Ensuite, la façon dont ces trajectoires peuvent être enrichies avec une introduction à la notion de trajectoire sémantique et de séquence temporelle. Enfin, nous analyserons comment il est possible d'extraire des comportements significatifs à partir de ces trajectoires, ainsi que divers outils de fouille de séquences.

— Le chapitre 3 est un état de l'art centré sur la thématique de *"pattern mining"* (ou fouille de motifs). Cet état de l'art, plus détaché des trajectoires, nous permettra d'appréhender les techniques disponibles dans le domaine de l'extraction de motifs communs correspondant à des comportements communs dans un jeu de données. Dans ce chapitre, nous présenterons aussi en détail la plateforme **GALACTIC** que nous avons choisi d'utiliser et l'algorithme NextPriorityConcept sur laquelle elle repose, un algorithme issu de la théorie des treillis, notion que nous aborderons également.

Partie 2. La deuxième partie de ce mémoire présentera les quatre contributions majeures effectuées durant ces 3 années de thèse. Le chapitre 4 fera office d'introduction aux contributions, explicitant les jeux de données utilisés dans nos travaux et les problématiques qu'ils soulèvent. Les jeux de données utilisés dans ce manuscrit sont tous issus de données réelles collectées sur des zones touristiques, allant des musées aux rues de la ville de La Rochelle. L'ordre des contributions suivra celui des problématiques présentées précédemment dans un processus global abordant différents domaines représenté dans la figure 3.10.

— Le chapitre 5 - *Contribution 1 : Reconstruction de trajectoires en environnement indoor contraint* - porte sur la reconstruction de trajectoires (bloc (a) figure 3.10).

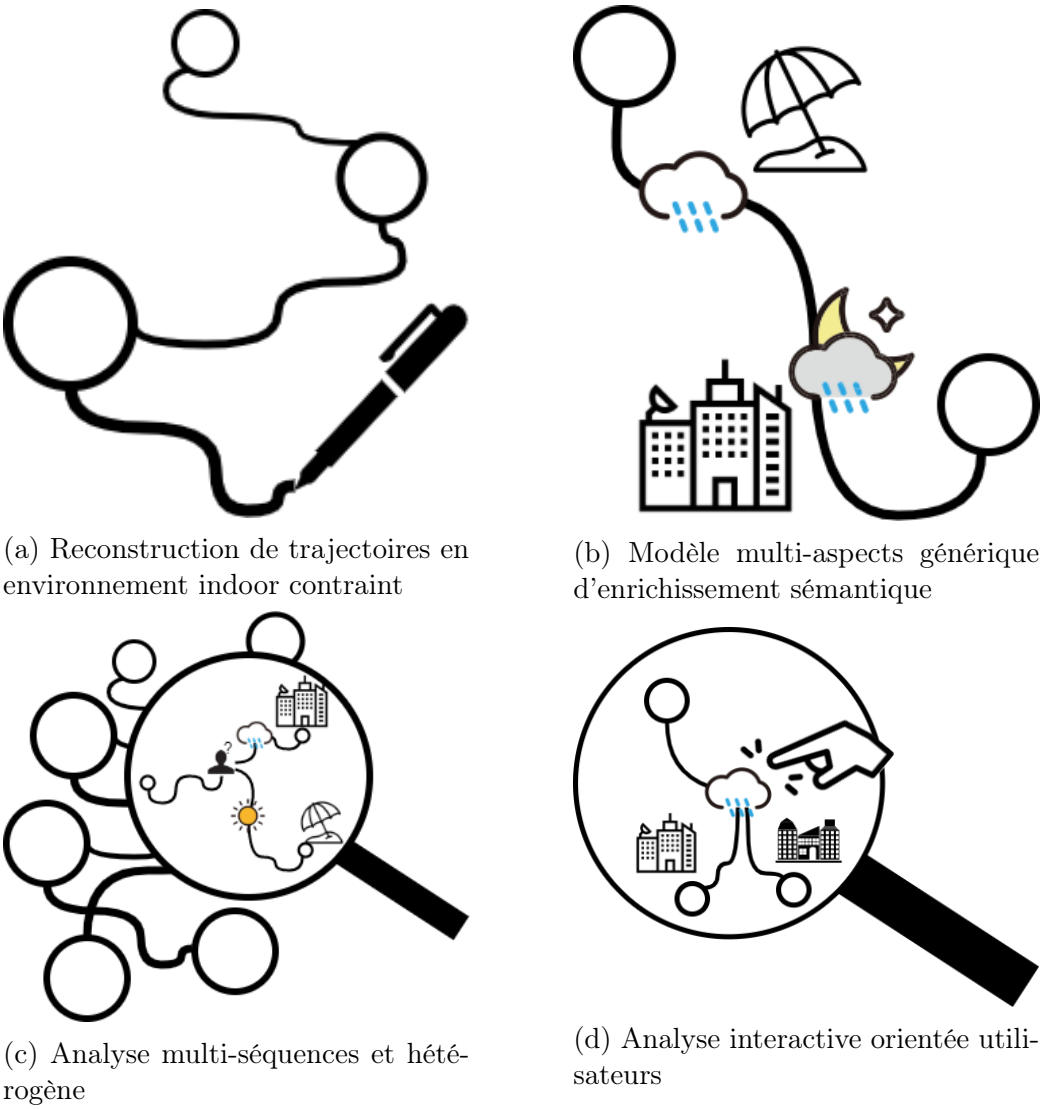
— Nous viendrons enrichir avec des données contextuelles ces trajectoires au chapitre 6 - *Contribution 2 : Modèle multi-aspects générique d'enrichissement sémantique* (bloc (b) figure 3.10).

— Puis nous analyserons les comportements liés à ces trajectoires en utilisant cet enrichissement au chapitre 7 - *Contribution 3 : Analyse multi-séquences et hétérogène* (bloc (c) figure 3.10).

— Pour enfin terminer avec un chapitre 8 - *Contribution 4 : Vers une analyse interactive avec* Reduced Context Completion - portant sur

une proposition d'analyse orientée par l'utilisateur (bloc (d) figure 1.1).

La figure 1.1 est une représentation du processus global décrit dans ce manuscrit, et les connexions entre les différents domaines abordés dans ces travaux de recherche.



(a) Reconstruction de trajectoires en environnement indoor contraint

(b) Modèle multi-aspects générique d'enrichissement sémantique

(c) Analyse multi-séquences et hétérogène

(d) Analyse interactive orientée utilisateurs

FIGURE 1.1 – Schéma de la chaîne de traitement proposée dans ce manuscrit

Première partie

Etat de l'art

Chapitre 2

Etat de l'art : Des données de capteurs aux trajectoires sémantiques

Table des matières

2.1 Introduction 26
2.2 Des capteurs aux trajectoires : reconstruction de trajectoire en intérieur 27
2.2.1 Les différents systèmes d'indoor positionning 27
2.2.2 Les algorithmes de reconstruction de trajectoires 29
Algorithmes de reconstruction par recoupement du signal 29
Algorithme d'apprentissage appliqué à l'indoor positioning 31
2.2.3 Les travaux notables qui s'appliquent aux musées . . . 32
2.2.4 Conclusion 33
2.3 Des trajectoires aux trajectoires sémantiques 33
2.3.1 Les informations contextuelles 33
2.3.2 Les modèles de trajectoires sémantiques 34
2.3.3 Conclusion 37
2.4 Des trajectoires sémantiques aux comportements 37
2.4.1 Introduction 37
2.4.2 La fouille de sous-séquences fréquentes : Le *sequence mining* 38
2.4.3 Fouille de séquences temporelles 40
Les épisodes 41
2.4.4 Fouille de séquences d'intervalles temporels 42
2.4.5 Conclusion 43
2.5 Conclusion 44

2.1 Introduction

Une trajectoire au sens large, ou données de mobilité, est une structure de donnée décrivant le déplacement d'un objet dans l'espace. Ainsi on note une trajectoire brute T comme étant une structure ordonnée telle que $T = \langle t_i, (x_i, y_i)\rangle_{0\leq i}$ où t_i est une donnée temporelle (timestamp, date, etc), avec $t_i < t_{i+1}$ et (x_i, y_i) une donnée relative à la spatialité qui peut être des coordonnées GPS ou encore des données de capteurs [Parent et al., 2013]. Dans ce manuscrit, nos données sont des trajectoires provenant de différentes bases de données et de différentes natures. Les problématiques présentées ici traverseront plusieurs étapes, allant de la reconstruction de trajectoires jusqu'à leur analyse, afin d'extraire des comportements significatifs.

Dans ce chapitre, nous nous intéresserons à différentes facettes de la trajectoire afin de définir et d'expliciter notre objet d'étude et les travaux existants.

Dans la section 2.2, nous explorerons la manière dont il est possible de reconstruire une trajectoire. En effet, une trajectoire de déplacement n'est tout d'abord pas toujours située en extérieur, même si la technologie du GPS et ses dérivées sont la manière la plus populaire de reconstruire le déplacement d'un objet dans le temps et l'espace, certains lieux (notamment en intérieur) nécessitent d'utiliser des technologies différentes amenant leurs propres problématiques.

Dans la section 2.3, nous étudierons la façon dont les chercheurs du domaine ont pu enrichir des trajectoires. Dans l'étude de la mobilité, il n'est souvent pas pertinent de se reposer uniquement sur l'aspect spatio-temporel de tels jeux de données. Nous parlons ainsi de la notion de trajectoire sémantique, où l'objectif sera d'enrichir des trajectoires brutes $T = \langle t_i, (x_i, y_i)\rangle_{0\leq i}$ en des trajectoires $T = \langle t_i, x_i, y_i, D_i\rangle_{0\leq i}$ où D_i est une donnée de contexte expliquant et précisant le contexte de la mobilité, comme les quartiers d'une ville, la météo ou la vitesse. Ces trajectoires complexes permettent de mieux représenter la trajectoire d'un individu en y ajoutant de l'explicabilité pour comprendre certains comportements de mobilité.

Dans la section 2.4, nous exposerons les différents types d'analyse permettant d'extraire des comportements à partir de ces trajectoires sémantiques. Le but, une fois la trajectoire enrichie de données de contexte, sera de les analyser et les traiter pour extraire des groupes d'individus ou de trajectoires comportant des similitudes. Ce faisant, cela permet aux chercheurs d'extraire différents comportements chez les individus étudiés, différents types de pra-

tiques, etc.

2.2 Des capteurs aux trajectoires : reconstruction de trajectoire en intérieur

Alors que le positionnement en extérieur est rendu possible grâce aux signaux GPS lorsque la zone est couverte, le positionnement en intérieur reste une problématique actuelle car les signaux des satellites ne peuvent pas toujours traverser les murs des bâtiments. Reconstruire une trajectoire dans ces lieux nécessite un système de positionnement en intérieur. Celui-ci est composé d'un ensemble d'émetteurs envoyant un signal périodique et un ensemble de récepteurs le capturant. De nombreux types de signaux sont utilisés tels que les radio-fréquences [Piccinni et al., 2016] [Piccinni et al., 2020], les ultrasons et l'infrarouge [Farid et al., 2013] [Liu et al., 2007] [Luo and Hsiao, 2018]. En fonction de la force du signal mesurée entre un émetteur et un récepteur (appelée RSSI pour Received Signal Strength Indication), l'objectif est de déterminer la position de l'utilisateur dans un espace fermé. Parmi tous les signaux utilisés dans ces systèmes, le BLE (Bluetooth Low Energy) est devenu de plus en plus populaire depuis 2013, avec la technologie iBeacon conçue spécifiquement pour la localisation en intérieur, que l'on retrouve dans de nombreuses applications comme BlueSentinel [Conte et al., 2014], où les auteurs proposent une méthode pour capturer l'activité d'individus à l'intérieur d'un bâtiment.

Dans cette section, nous étudierons la façon dont les chercheurs du domaine ont pu passer d'un format de données sous la forme $\langle RSSI_i, t_i \rangle_{0 \leq i}$ où $RSSI_i$ est un relevé de capteur à un instant t_i, $t_i < t_{i+1}$ vers un format de type $T = \langle t_i, (x_i, y_i) \rangle_{0 \leq i}$.

2.2.1 Les différents systèmes d'indoor positionning

Un système de localisation en intérieur (ou indoor positionning) fonctionne via une série d'émetteurs et de récepteurs disposés dans un espace clos. Le signal envoyé par les émetteurs et reçu par les récepteurs servira de base aux traitements effectués afin de déterminer la position d'un individu au sein de l'environnement clos. S'exerçant au sein de bâtiments, la localisation à l'intérieur s'effectuera généralement sur plusieurs plans à deux dimensions représentant les étages de celui-ci s'il y en a [Li, 2008]. La figure 2.2 représente le schéma d'un système d'indoor positionning d'une façon globale et la figure 2.1 représente l'évolution du signal RSSI de plusieurs capteurs durant une expérimentation.

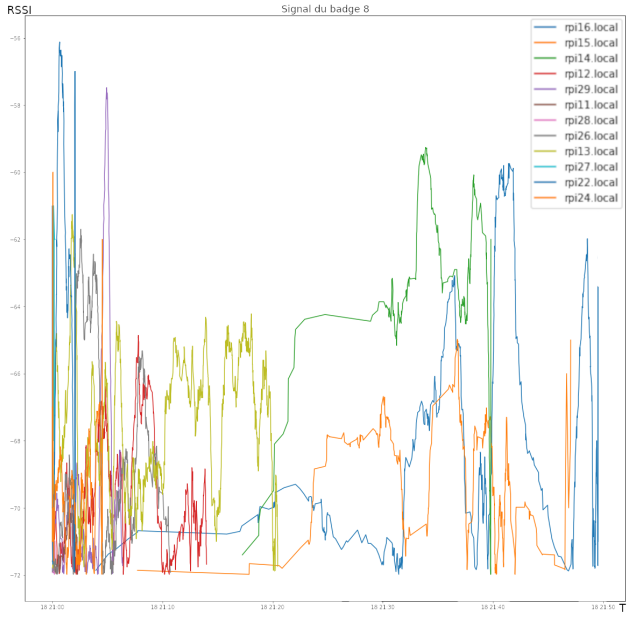


FIGURE 2.1 – Signaux récupérés par un émetteur durant une déambulation

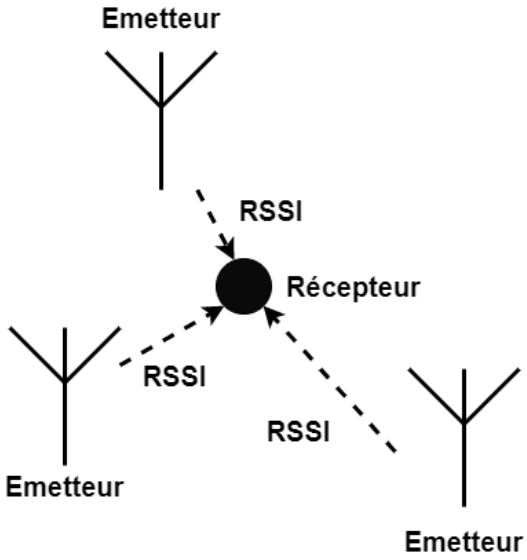


FIGURE 2.2 – Système d'indoor positionning

La représentation que l'on donnera des déplacements d'un individu au sein de cet espace, que ce soit sa trajectoire ou la topologie des lieux s'appelle un modèle de localisation [Leonhardt, 1998] [Afyouni et al., 2012]. Ce modèle de localisation peut varier dans sa forme et ainsi s'exprimer de plusieurs façons. D'abord d'une manière géométrique, dans le sens où la trajectoire d'un individu sera représentée par une suite de points dans l'espace sous la forme $T = \langle t_i, (x_i, y_i)\rangle_{0 \leq i}$ où (x_i, y_i) sont des coordonnées sur un plan [Berkovich, 2014] [Liu et al., 2020]. Ensuite, ce modèle peut prendre une forme symbolique, moins *"précise"*, où la trajectoire se formalisera de telle sorte que $T = \langle t_i, P_i\rangle_{0 \leq i}$ où P_i est une information sémantique de type "Salle A", "Salle B" [Jensen et al., 2009] etc .. La forme que prendra le modèle de localisation dépendra entièrement du type de topologie et du but recherché : plus le degré de granularité recherché est élevé, plus les coûts matériels et de calculs le seront, il s'agit ainsi de trouver le bon équilibre entre les deux.

Dans ce manuscrit, nous contribuerons aux deux types de modèles de localisation, à la fois géométrique et symbolique.

2.2.2 Les algorithmes de reconstruction de trajectoires

Cette sous-section décrit les deux principaux algorithmes de reconstruction de trajectoires, à savoir la triangulation et la trilatération nécessitant un minimum de 3 appareils de captation ou d'émission, ainsi que des algorithmes d'apprentissage nécessitant une vérité terrain. Nous poursuivrons ensuite en présentant les travaux de reconstruction réalisés dans les musées, qui soulignent les limites de ces algorithmes en termes d'adaptation.

Algorithmes de reconstruction par recoupement du signal

La figure 2.3 représente le principe de la triangulation comme défini dans l'article de Y. Wang et al. [Wang et al., 2013].

La conversion des relevés de capteurs en mètre se fait en utilisant le "path-loss model" à distance logarithmique, adoptée dans plusieurs études telles que [Li et al., 2018] [Dong and Dargie, 2012].

$$L_s = (-10n) log_{10}(RSSI_s) + L_0 \tag{2.1}$$

Où L_s est la distance en mètre basée sur la valeur $RSSI_s$ du capteur s, n est la valeur de la puissance de diffusion et L_0 est la puissance RSSI mesurée à 1 mètre. Il faut cependant garder à l'esprit que cette valeur peut changer radicalement en fonction de l'environnement (comme la taille de la pièce étudiée) et donc cette constante doit être mesurée dans tous les nouveaux

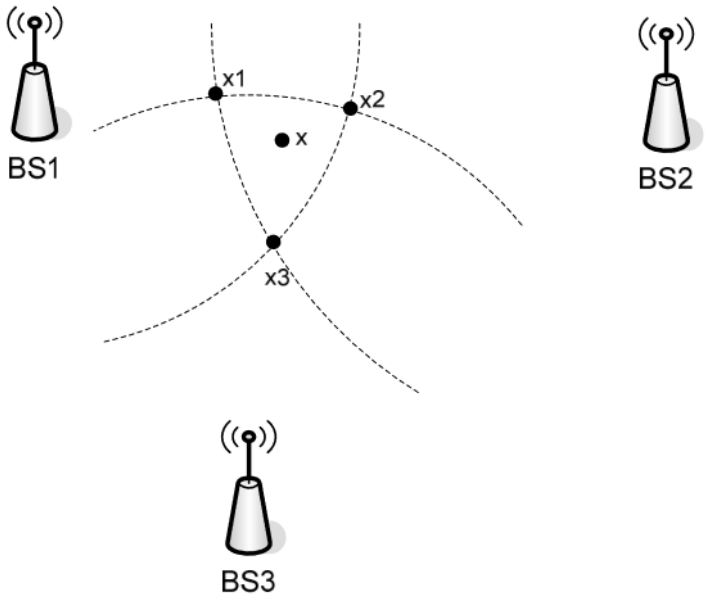


FIGURE 2.3 – Schéma représentant le principe de la triangulation [Wang et al., 2013]

environnements.

Une fois la valeur de L_s récupérée, l'algorithme de triangulation cherche l'intersection des cercles de distance pour chaque capteur, puis récupère le centroïde de toutes les intersections, comme le montre la figure 2.3.

L'algorithme de trilatération expliqué dans [Pradhan et al., 2019] et souvent utilisé dans les travaux de positionnement intérieur [Cantón Paterna et al., 2017] [Dinh et al., 2020] [He et al., 2015]. La figure 2.4 représente le concept d'intersection de lignes.

Le principe de la trilateration est similaire à la triangulation, dans le sens où l'algorithme cherche les points d'intersection entre les cercles de distance pour chaque capteur après la conversion du signal RSSI en mètre. La différence notable c'est le traçage de lignes entre les points d'intersections. L'intersection d'un cercle C_1 avec un cercle C_2 créera deux points A et A', l'intersection entre un cercle C_1 et C_3 créera deux points B et B'. L'algorithme trace ensuite une ligne entre A et A', B et B' etc. C'est l'intersection de toutes les lignes qui donnera une estimation de la position de l'individu dans la pièce.

Cependant, la principale limite de ces algorithmes réside dans leur dépen-

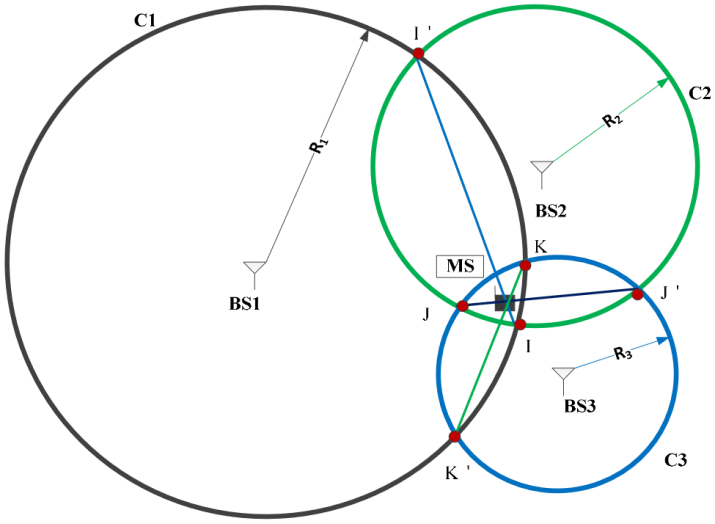


FIGURE 2.4 – Schéma représentant le principe de la trilatération [Pradhan et al., 2019]

dance à un positionnement précis des balises pour encadrer efficacement le déplacement du visiteur. En effet, l'emplacement des dispositifs de positionnement en intérieur est crucial et un mauvais positionnement dans la pièce peut réduire drastiquement les performances de ces deux algorithmes [Rezazadeh et al., 2018].

Algorithme d'apprentissage appliqué à l'indoor positioning

La fluctuation de la puissance du signal reçu dans les systèmes d'indoor positionning ainsi que les lieux contraints où on ne peut placer les balises comme on le souhaite sont un véritable problème qui a poussé les chercheurs du domaine à utiliser d'autres méthodes d'apprentissages [Munadhil et al., 2020] [Nessa et al., 2020]. Les algorithmes de classification sont en effet adaptés pour ce genre de problématiques et on trouvera ainsi des méthodes utilisant K-NN [Hoang et al., 2018], Support-Vector [Bottou and Lin, 2007], Random Forest etc.

Plusieurs études appliquant de l'apprentissage dans la reconstruction de déplacements en intérieur en se basant sur le signal reçu, comme expliqué par [Belmonte-Hernández et al., 2019], offrent une très bonne précision et globalement de très bons résultats.

Les algorithmes d'apprentissage appliqués à la localisation en intérieur ont pu montrer leurs fruits dans plusieurs domaines et notamment dans la robotique [Károly et al., 2020] comme dans les travaux de J. Xiang et al. [Xiang

et al., 2019], où des méthodes d'apprentissage par renforcement ont été utilisés afin d'aider la navigation d'un robot en intérieur.

Même si les algorithmes d'apprentissage pour la localisation en intérieur restent en voie de développement et montrent des résultats prometteurs, ils souffrent cependant d'importantes limites [Zafari et al., 2019]. La plupart des applications de localisation en intérieur sont déployées sur des applications mobiles ou des appareils facilement transportables (comme des raspberries), où les capacités de calcul et de stockage sont largement limitées. Ainsi, ces algorithmes complexes et coûteux sont difficilement compatibles avec ce type de matériel. De plus, la nécessité d'avoir une étape de labélisation pour les algorithmes d'apprentissage supervisés rendent le processus coûteux, long et uniquement viable pour une localisation donnée. Une phase d'apprentissage sera ainsi nécessaire pour chaque nouveau lieu que l'on équipe d'un tel système.

2.2.3 Les travaux notables qui s'appliquent aux musées

Ces dernières années, un certain nombre d'études de positionnement en intérieur visent à proposer des technologies de suivi aux responsables des musées [Yoshimura et al., 2014] [Spachos and Plataniotis, 2020] [Giuliano et al., 2020]. En plus de permettre aux gestionnaires de musées d'avoir une meilleure vue d'ensemble de la fréquentation de leurs lieux et donc d'avoir un meilleur retour sur les visites afin de distinguer les œuvres populaires de celles qui sont délaissées, cela permet également d'étudier le comportement des visiteurs dans de tels environnements [Kontarinis et al., 2017] [Juniarta et al., 2018b]. Par ailleurs, les responsables de musées sont toujours ravis d'organiser ce type d'événements qui offrent l'opportunité de proposer au public de participer à une telle expérimentation tout en permettant de populariser des recherches scientifiques.

Sur un plan plus technique, les musées offrent un terrain d'expérimentation parfait pour les technologies de positionnement en intérieur. En effet, les visiteurs restent longtemps dans une même salle et marchent lentement, ce qui nous permet de mieux capter leur parcours. Les musées apportent également des contraintes qui soulèvent des problématiques auxquelles les chercheurs en localisation en intérieur doivent faire face. Tout d'abord, les approches classiques d'approximation de la position restent limitées. Par ailleurs, bien que certaines études appliquent des algorithmes d'apprentissage automatique pour les systèmes de localisation intérieur, comme expliqué dans [Belmonte-Hernández et al., 2019], il est très difficile cependant de les utiliser dans les musées. En effet il faut beaucoup de ressources et de temps

pour collecter des données afin de les entraîner et elles ne fonctionneront que dans un seul musée, l'architecture et la disposition des salles n'étant pas les mêmes d'un musée à l'autre. Deuxièmement, les musées offrent un environnement contraint où nous sommes limités dans le nombre et l'emplacement des balises pour nos systèmes de positionnement intérieur. En effet, nos balises ne doivent pas déranger les visiteurs pendant leur visite (nous ne pouvons donc pas les placer au milieu d'un chemin de parcours). La topologie des lieux n'ayant pas été conçue pour accueillir un système de positionnement intérieur, il est nécessaire de contourner les œuvres et de jouer avec l'espace disponible. Cette contrainte limite donc les performances d'algorithmes tels que la triangulation et la trilatération.

2.2.4 Conclusion

Dans cette section nous avons décrit les problématiques de reconstructions de trajectoires indoor. Il n'y a pas de consensus réellement établi sur la manière de procéder, et la façon dont nous pouvons récupérer la trajectoire de visiteurs à l'intérieur d'un musée spécifiquement reste un problème ouvert. Dans la contribution 1 de ce manuscrit, nous apportons deux façons de procéder que nous avons utilisées afin de reconstruire de la façon la plus efficace possible, selon nos besoins, la déambulation de visiteurs à l'intérieur de musées avec un positionnement des balises restreint par la topologie des lieux.

2.3 Des trajectoires aux trajectoires sémantiques

Au fur et à mesure des études de mobilité, les chercheurs du domaine ont commencé à ajouter un troisième élément aux tuples $\langle t_i, (x_i, y_i)\rangle$ composant la trajectoire : une donnée d_i qui permet d'annoter et de caractériser des portions de trajectoires et ainsi les enrichir avec par exemple des noms de lieux que la personne traverse, l'altitude, la vitesse de déplacement etc.

Dans cette section nous parlerons de la trajectoire au sens large du terme et de la notion de trajectoire sémantique permettant d'intégrer des données contextuelles, connaissances apportant une nouvelle information sur les déplacements.

2.3.1 Les informations contextuelles

La connaissance contextuelle peut être définie comme toute information permettant de mieux décrire un événement, une personne ou un objet. Ces

informations apportent aux trajectoires un autre point de vue en évitant de se baser uniquement sur les données de déplacement. Les conditions météorologiques ou toutes sortes d'informations de contexte peuvent avoir un impact sur le comportement d'un individu et ainsi refaçonner sa trajectoire. Les informations contextuelles sont ainsi des éléments cruciaux et apportent un nouveau regard sur les données de déplacement. Ces données peuvent être ajoutées aux trajectoires. Celles-ci sont segmentées via des données temporelles sous la forme d'un épisode. Un épisode est un intervalle de temps $(A, \underline{t}, \overline{t})$ où $\underline{t}$, $\overline{t}$ sont des horodatages et A une annotation sémantique. L'utilisation d'une succession d'épisodes telle que la séquence de quartiers traversée par une trajectoire peut être interprétée comme une "discrétisation" de celle-ci. On appelle ainsi trajectoire brute une trajectoire ne contenant que les informations spatiales, et trajectoire sémantique lorsqu'il y a un ou des ajouts d'informations directement sur ces traces de mobilités.

La trajectoire sémantique représente une part importante dans le domaine de la mobilité, et un certain nombre de travaux s'intéresse à la formalisation du concept de trajectoire sémantique. Nous pouvons citer les travaux tel que [Yan et al., 2008] [Parent et al., 2013] [Bogorny et al., 2014a] [Flouvat, 2019] [Mello et al., 2019] [Noureddine et al., 2022] travaillant sur cette notion de telle sorte que nous proposons la définition suivante afin de caractériser la trajectoire sémantique :

Definition 1. *Une trajectoire sémantique est une trajectoire brute enrichie à l'aide de connaissances par le biais d'annotations, segmentant la trajectoire en épisodes où un épisode correspond à une annotation. Nous pouvons proposer de définir une trajectoire sémantique T par :*

$$T = \langle (\underline{t_i}, \overline{t_i}), P_i, s_i \rangle_{0 \leq i}$$

avec $\underline{t_i}$ et $\overline{t_i}$ des dates, $(\underline{t_i}, \overline{t_i})$ l'intervalle de temps d'un épisode ($(\overline{t_i} \leq \underline{t_{i+1}})$ et $\overline{t_i} < \underline{t_{i+1}}$) et P_i la liste des positions géographiques affectées par l'annotation s_i.

2.3.2 Les modèles de trajectoires sémantiques

La trajectoire sémantique est définie par *Parent et al. 2013* [Parent et al., 2013] comme : *"... une trajectoire brute qui a été enrichie par des annotations et/ou une ou plusieurs segmentations complémentaires."*. L'ajout de connaissances sémantiques directement à une trajectoire brute permet aux chercheurs de mieux interpréter ces données. Un ensemble de coordonnées GPS peut devenir "rue A", et une série de radiofréquences provenant de

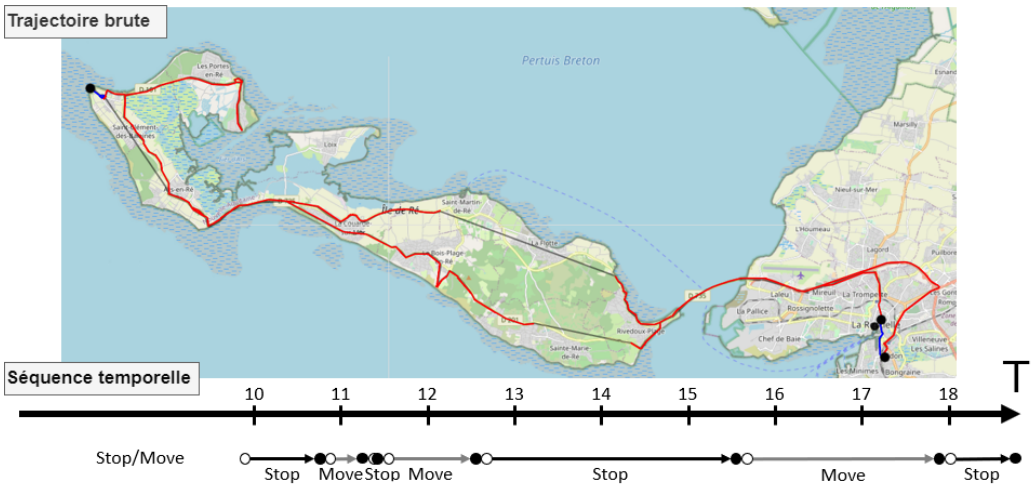


FIGURE 2.5 – Trajectoire sémantique sous la forme d'épisodes de *stop* et de *move*

multiples capteurs dans un système de localisation intérieur peut devenir "pièce A". La définition et la formalisation de la trajectoire sémantique seront différentes d'un modèle de données à l'autre, mais l'idée principale est de fournir un cadre suffisamment générique pour être utilisé dans de nombreux contextes. Comme [Noureddine et al., 2020], qui propose un modèle de trajectoires pour les espaces intérieurs et extérieurs.

Pour ce faire, ces modèles sont généralement structurés comme une séquence d'épisodes selon un paramétrage défini par le scientifique des données [Mountain and Raper, 2001]. Ce paramétrage peut être un POI (Place Of Interest), un comportement ou tout autre attribut sur lequel le chercheur choisit de focaliser son modèle, et qui prendra la forme d'un alphabet. Le premier modèle de trajectoires sémantiques est le modèle "stop and move" [Spaccapietra et al., 2008a]. En annotant sémantiquement le moment où un individu s'arrête et se déplace à nouveau, S. Spaccapietra et al. (avec L. O. Alvares, également sur la base de stop and move [Alvares et al., 2007]) ont construit leurs trajectoires sémantiques comme une succession de *stop* et *move*. Un *stop* est un intervalle de temps non vide $[\underline{t}, \overline{t}]$ où l'objet en déplacement ne bouge pas et un *move* est un intervalle de temps non vide $[\underline{t}, \overline{t}]$ où l'objet en déplacement est en mouvement. Dans ce modèle, *stop* et *move* sont des épisodes. Ce travail constitue le premier pas dans le domaine de la trajectoire sémantique inspirant plusieurs travaux tels que [Bogorny et al., 2014b] [Ruback et al., 2016] [Yan et al., 2011] où un cadre d'annotation et d'enrichissement sémantique est proposé. Peu de temps après, les trajectoires sémantiques ont commencé à utiliser des modèles basés sur les ontologies tels

que [Yan et al., 2008] [Baglioni et al., 2009a], toujours basés sur le premier modèle *stop and move*. Le formalisme ontologique est alors utilisé dans ces modèles comme un moyen de convertir la trajectoire brute en une représentation sémantique de haut niveau en utilisant des thésaurus ou des taxonomies.

Le modèle "stop" et "move" peut être vu comme une discrétisation de la vitesse tandis qu'un "*district*" est une discrétisation de l'espace en fonction des quartiers. Dans l'étude [Baglioni et al., 2009b], les auteurs ont construit un modèle basé sur le langage OWL (Web Ontology Language) pour l'enrichissement sémantique de données brutes à partir de connaissances géographiques afin de trouver des modèles communs de comportement dans un contexte touristique.

Ces modèles de données basés sur les ontologies pour les trajectoires sémantiques sont seulement "interrogeables" et ne peuvent pas être utilisés dans les processus d'analyse. En fait, l'analyse de ces modèles de données est un problème majeur dans le domaine des trajectoires sémantiques. Néanmoins, nous pouvons citer des travaux tels que [Andrienko et al., 2011] où les auteurs proposent un modèle conceptuel pour les trajectoires sémantiques considérant les données de mouvements et les connaissances contextuelles comme une composition d'événements spatio-temporels et où ils peuvent calculer des relations temporelles telles que la distance, l'ordre, etc.

Dans [Van Hage et al., 2012], les auteurs ont utilisé le modèle SEM (Simple Event Model) basé sur les ontologies définies dans [van Hage et al., 2011] pour structurer et améliorer les données brutes des trajectoires de navires. Le SEM est une ontologie déjà en place qui est ensuite adaptée pour travailler avec des données géographiques et intégrer en même temps des connaissances provenant de différentes sources sur le web à partir de sources hétérogènes. En étendant le travail de [Spaccapietra et al., 2008a], le modèle SEM a été capable de s'adapter en considérant l'arrêt et le déplacement comme des types d'événements spatio-temporels spéciaux ; les ontologies sont alors utilisées pour détailler de tels événements en ajoutant des sous-événements, des types d'événements, des lieux, etc. Récemment, plusieurs modèles de trajectoires sémantiques tendent à diverger du modèle *stop and move*. En effet, les épisodes tels que *stop* et *move* sont souvent insuffisants pour recréer fidèlement le passage d'un objet en mouvement dans l'espace. En travaillant avec des données de faible qualité, il est difficile de dire quand l'individu s'est arrêté, et encore plus lorsque la trajectoire brute provient des métadonnées d'applications comme Instagram ou Flicker. Dans [Fileto et al., 2015], les auteurs définissent le modèle Baquara[2] pour l'enrichissement sémantique avec une segmentation basée sur des épisodes personnalisables, au lieu de la méthode

classique de "*stop et move*".

2.3.3 Conclusion

La problématique d'enrichissement des trajectoires brutes dans le but d'obtenir des trajectoires sémantiques est un domaine encore largement étudié par les chercheurs en géomatique, et nécessite la plupart du temps la création d'un tout nouveau modèle afin de correspondre efficacement au sujet traité. La limite d'une telle démarche est que plus un modèle de trajectoire sémantique est complexe, plus il sera difficile à analyser, et à exporter de façon générique vers d'autres sujets de recherche. Dans la deuxième contribution de cette thèse, nous abordons cette problématique en proposant un modèle générique de trajectoire sémantique. Ce modèle est conçu de manière à pouvoir être utilisé sans contrainte dans différents types de mobilités humaines et de pratiques, que ce soit en intérieur ou en extérieur. Plus précisément en utilisant les séquences temporelles dans l'objectif de récupérer des comportements similaires présents dans notre base de données. Car au-delà de l'aspect descriptif d'un modèle de trajectoire sémantique, notre ambition est de pouvoir extraire de notre base de données des comportements distinctifs. Pour ce faire, nous étudierons les travaux partageant cet objectif de l'état de l'art dans la prochaine section.

2.4 Des trajectoires sémantiques aux comportements

2.4.1 Introduction

Identifier des comportements revient à extraire des sous-groupes ou *cluster* d'individus ayant un même "comportement". Ainsi, la plupart des méthodes d'analyse de données de mouvement sont proposées dans le domaine du data mining. Dans [Pujari, 2001], l'auteur définit le data mining comme : "... l'exploration et l'analyse de grands ensembles de données afin de découvrir des modèles et des règles significatifs". Dans les études de mobilité humaine, la recherche de "règles" ou de "modèles significatifs" peut revêtir plusieurs aspects. Dans [Zheng, 2015] l'auteur recense quatre grands types de modèles de mobilité : *Moving Together pattern*, *Periodical pattern*, *Sequential patterns* et *Trajectory clustering patterns*.

Dans cette optique, il est possible d'utiliser des méthodes dites "classiques" de clustering, à condition de pouvoir définir une mesure de similarité entre les trajectoires. Pour ce faire, les modèles de mobilité tendent à vectoriser les

trajectoires GPS brutes, comme les modèles de regroupement de trajectoires où l'objectif est de déterminer les trajectoires communes afin de regrouper les trajectoires similaires. Dans [Pelekis et al., 2007] [Trajcevski et al., 2007] ou [Toohey and Duckham, 2015] les auteurs proposent un tour d'horizon de mesures de similarité. Plusieurs études ont proposé des algorithmes de clustering de trajectoires tels que [Giannotti et al., 2007] [Camargo et al., 2007] ou [Lee et al., 2007] ; où les auteurs ont utilisé ces métriques pour comparer des sous-trajectoires communes. Néanmoins, ces études ont utilisé des trajectoires GPS brutes et les mesures de similarité peuvent être difficiles à adapter à d'autres types de trajectoires telles que les trajectoires sémantiques en intérieur.

Plusieurs inconvénients à ces méthodes sont cependant à relever. Premièrement, ces analyses manquent d'explicabilitées dans le sens où il n'existe pas de description des comportements communs à chaque sous-groupe. Ensuite, ces méthodes sont difficilement extensibles aux trajectoires sémantiques que nous avons abordées à la section précédente. Pour ces raisons, nous nous intéresserons ici aux méthodes dites de séquence / pattern mining dans le traitement des trajectoires. Ces méthodes reposent sur les aspects sémantiques des données et sont applicables aux trajectoires sémantiques. Celles-ci permettront une approche plus générique de l'analyse des données de mobilités que ce soit en extérieur ou en intérieur.

2.4.2 La fouille de sous-séquences fréquentes : Le *sequence mining*

Pour l'analyse des trajectoires, des méthodes de fouille de séquences ont été employées afin d'étudier les comportements fréquents au sein d'une base de données de trajectoires Giannotti et al. [2007]. Dans cette section, nous examinerons les algorithmes de fouille de séquences applicables aux trajectoires segmentées de manière sémantique.

Le *sequence mining*, aussi appelé la fouille de motifs séquentiels fréquents, vise à extraire des comportements fréquents extraits d'une base de données de séquence. Le premier algorithme de *sequence mining* à avoir été développé est celui d'Agrawal et Srikant. Celui-ci se veut comme une extension de l'algorithme Apriori pour des données séquentielles, GSP [Srikant and Agrawal, 1996] (pour *Generalized Sequential Pattern* algorithm) avec un système d'utilisation de contraintes temporelles afin de limiter le nombre de sous-séquences générées. GSP reprend la même logique qu'Apriori où les groupes sont décrit par leurs attributs communs, ici les itemsets. Ces itemsets deviennent des

sous-séquences communes qui décrivent le comportement commun des individus de chaque cluster. L'algorithme effectue plusieurs itérations sur un même jeu de données jusqu'à ce qu'aucune autre sous-séquence commune ne soit générée, toujours en ajoutant d'autres éléments aux sous-ensembles - ici sous-séquences - dont le support est supérieur au seuil entré par l'utilisateur. Tout comme les algorithmes de fouille de motifs fréquents se basant sur les *itemsets*, ces algorithmes se basent sur un seuil de support minimum défini par l'utilisateur où le support est la fréquence d'apparition de sous-séquences communes dans le jeu de données. Tout comme GSP, la plupart des algorithmes de sequence mining se basent sur l'algorithme Apriori, telles que AprioriAll, PSP [Masseglia et al., 1998], SPADE [Zaki, 2001], SPAM [Ayres et al., 2002]. Cependant, d'autres algorithmes tel que PrefixSpan [Han et al., 2001] ou FreeSpan [Han et al., 2000] utilisant une représentation en forme d'ensemble de données projetées. On peut aussi noter plusieurs améliorations de l'algorithme GSP notamment l'algorithme MFS [Li et al., 2011], MSPS [Luo and Chung, 2004] ou PSP (Prefix Tree for Sequential Pattern) [Masseglia et al., 1998] utilisant une arborescence afin de stocker et d'indexer les sous-séquences retenues.

Toutefois, les algorithmes de fouilles de motifs séquentiels se confrontent eux aussi au problème du déluge de patterns, où les "patterns" sont des sous-séquences, qui en rend l'exploitation difficile. Certaines sous-séquences communes générées sont incluses dans d'autres super-séquences de même support provoquant ainsi une redondance d'informations. Ainsi, pour y répondre, la voie de la fouille de motifs séquentiels a suivi la même direction que la fouille de motifs simple : Travailler sur une récupération et une analyse de motifs séquentiels fermés. Réduire le nombre de sous-séquences communes afin de renvoyer la même quantité d'information en retour d'algorithme fut la motivation principale afin de proposer des algorithmes tel que CloSpan [Yan et al., 2003], BIDE [Wang and Han, 2004] ou ClaSP [Gomariz et al., 2013].

Nous définissons une séquence simple S par $S = \langle s_i \rangle_{0 \leq i}$. Cette notion de séquence permet de traiter un grand nombre de données tel que du texte, l'ADN où des traces d'exécutions. Cette structure de données peut avoir certaines limites dans des cas très spécifiques. L'impossibilité d'avoir un indicateur pour formaliser un espacement entre les éléments d'une séquence peut être une limite quand on traite des données médicales, ou l'espacement (ie. la temporalité) entre les différents symptômes (qui peuvent être des éléments de notre séquence), est une donnée tout aussi cruciale à prendre en compte. En outre, étant donné que le grand nombre de motifs peut constituer une limite pour les algorithmes de fouille de motifs, prendre en compte

la temporalité permet également de limiter cette contrainte.

Dans une étude de regroupement de trajectoires très proche de nos travaux [Juniarta et al., 2018a], les auteurs expérimentent l'utilisation de l'exploration de motifs séquentiels pour analyser le parcours des visiteurs dans un musée. En façonnant les trajectoires sous forme de séquences, l'article combine l'exploration de séquences avec des algorithmes tels que MFCS (pour "**M**ining **F**requent **C**ontiguous **S**ubsequences") et MRGS (pour "**M**ining **R**are **G**eneral **S**ubsequences") avec l'analyse des données de mouvement ; les auteurs ont pu identifier quatre comportements de visite. L'utilisation de techniques d'exploration de motifs pour traiter les trajectoires [Shaw and Gopalan, 2014] montre des résultats prometteurs pour l'analyse et le regroupement de données de mouvement avec des trajectoires sémantiques [Chen et al., 2019]. Ces travaux sont étroitement liés à ce qui a été fait en matière d'analyse de motifs séquentiels. Il s'agit d'une question émergente dans le domaine de l'analyse des données de mouvement, mais nous pouvons tout de même citer [Zhang et al., 2014] où les auteurs ont proposé Splitter, un modèle capable d'exploiter efficacement les motifs séquentiels dans les trajectoires sémantiques avec comme approche la classification. Dans [Cao et al., 2005], C. Huiping et al. ont proposé d'utiliser un algorithme de type *Apriori* pour résoudre ce problème dans un ensemble de données de trajectoires.

2.4.3 Fouille de séquences temporelles

Dans ce manuscrit, nous nous intéressons tout particulièrement aux séquences de déplacements d'individus dans un espace. Dans ce contexte, l'information temporelle est un élément à prendre en compte pour plusieurs raisons. D'abord, visiter un lieu n'a pas la même signification suivant l'heure à laquelle on s'y rend - on parlera ainsi de pratiques différentes. Ensuite, la notion de blanc des cartes, cette instant situé dans le temps où aucune activité n'a pu être relevée (par un défaut de capteur par exemple) est une notion avec laquelle nous nous devons de composer dans les études géographiques. Ainsi, un enchaînement d'un *lieu A* vers un *lieu B* n'a pas la même signification s'il se fait dans un espace temporel restreint par rapport à une temporalité plus grande. L'analyse de ces motifs a une dimension supplémentaire de temporalité ajoutant de ce fait une profondeur et une pertinence non négligeable dans les résultats de traitement de ces données.

Dans un premier temps, les chercheurs du domaine ont formalisé la séquence temporelle en ajoutant une donnée temporelle aux éléments d'une séquence simple. Ainsi, on note une séquence temporelle $S = \langle (t_i, s_i) \rangle_{0 \leq i}$ où

chaque élément s_i est associé à une donnée temporelle (timestamp, date ..) avec $t_i < t_{i+1}$ afin de respecter la notion d'ordre intrinsèque à la structure en séquence. On retrouvera ainsi des algorithmes de fouille temporelle avec des contraintes temporelles [Masseglia et al., 2009] [Bonchi et al., 2006] [Zaki, 2000] [Orlando et al., 2004], d'épisodes fréquents [Mannila et al., 1997] ou l'intervention de la notion de chroniques [Dousson and Duong, 1999].

L'élément commun entre tous les travaux sur les séquences temporelles est la notion de contraintes temporelles. Une contrainte temporelle [Pei et al., 2007] vérifie l'écart de temporalité entre les éléments d'un ensemble donné. Ainsi, lors de l'extraction d'un motif par un algorithme de fouille de motifs, celui-ci pourrait ne pas être généré si l'écart temporel des éléments qui le composent dépasse un certain seuil de temps défini par l'analyste.

Les épisodes

La fouille d'épisode fréquents répond en un sens à cette notion de contrainte temporelle en se focalisant sur l'extraction d'une collection d'événements temporels lorsque ceux-ci sont proches [Manila, 1995]. Cependant, cette notion d'épisode spécifique au domaine du pattern mining temporel (à ne pas confondre avec l'épisode d'une trajectoire sémantique) est défini comme des ensembles partiellement ordonnés dans un intervalle de temps correspondant à la taille de la fenêtre temporelle spécifiée par l'utilisateur. On retrouvera ainsi des types d'épisodes qui peuvent différer, les épisodes en série tout d'abord qui sont comparables à des sous-séquences communes d'événements temporels, et les épisodes parallèles qui ne prennent pas en compte l'ordre des éléments se rapprochant ainsi des fouilles de motifs classiques.

Exemple 2.4.1. *Pour une séquence temporelle* $ST = \langle (a,1),(b,2),(d,3),(a,5)\rangle$ *et une fenêtre temporelle* Ft *de taille* $Ft_{size} = 2$*, les figures 2.6 et 2.7 représentent le déplacement de la fenêtre temporelle FtFt sur nos données temporelles. Ainsi on compte les épisodes* $Ft_1 = \{a,b\}$ *et* $Ft_2 = \{a,b,d\}$ *pouvant être traités au choix comme des épisodes en série (sous la forme d'une séquence) ou en parallèle (sous la forme d'itemset).*

De ces épisodes, les chercheurs du domaine proposent plusieurs façons de les analyser. Premièrement avec l'algorithme Winepi [Manila, 1995] appliquant Apriori pour trouver des épisodes fréquents en utilisant la fenêtre glissante comme plusieurs séquences d'événements. Celui-ci a ensuite été étendu avec Minepi [Mannila et al., 1997] en se basant cette fois-ci sur les occur-

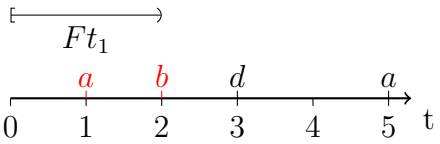


FIGURE 2.6 – Exemple de fouille de motifs temporels par épisode, Ft_1

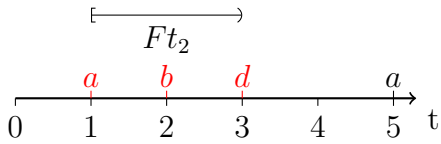


FIGURE 2.7 – Exemple de fouille de motifs temporels par épisode, Ft_2

rences minimales d'événements au sein des séquences.

Ces méthodes de pattern mining sont efficaces lorsque les données que nous traitons sont des événements "spontannés", des points singuliers dans le temps. Cependant, le suivi d'individus avec une trajectoire sous forme de successions d'étapes requière très probablement une structure de données plus adéquate. En effet, certains événements persistent dans le temps, ce qui est le cas lors de la visite d'un lieu, sur une certaine durée, il ne s'agit pas d'un événement ponctuel dans le temps. Symboliser une trajectoire touristique sous la forme d'une séquence temporelle simple, plaçant les visites de quartiers ou les salles de musées en tant qu'évènements temporels ne permet pas de prendre en compte la durée et la fin de la visite d'un lieu. Avec une séquence temporelle simple, il est impossible de détecter les "blancs des cartes". Ainsi, cette modélisation n'est pas suffisamment appropriée pour notre problème.

2.4.4 Fouille de séquences d'intervalles temporels

Afin de répondre à cette problématique, les chercheurs du domaine ont conçu une représentation pour une séquence d'intervalles. Une séquence d'intervalles est définie par $S = \langle([\underline{t_i}, \overline{t_i}], s_i)\rangle$, où s_i un élément d'un dictionnaire Σ et $\underline{t_i}$ et $\overline{t_i}$ des dates où on note $\underline{t_i} \leq \overline{t_i}$ et $\overline{t_i} \leq \underline{t_{i+1}}$. De nombreux algorithmes travaillent avec des séquences d'intervalles tels que [Guyet and Quiniou, 2011a] [Winarko and Roddick, 2007] [Kam and Fu, 2000] [Moskovitch and Shahar, 2015]. Là où les séquences temporelles simples utilisaient une fenêtre temporelle afin de "glisser" le long de nos données formées d'événements sur une frise chronologique, ici, les éléments de nos séquences temporelles intègrent déjà une forme de fenêtre temporelle ou d'intervalle temporel. Les différentes relations possibles entre ces intervalles ont été introduites par Allen, [Allen, 1981]. Dans toute forme de processus, réside en réalité une emprise forte avec le temps, et les actions de celui-ci couvre une période de temps plus ou moins grande et obéissant à des règles définies. Ces règles ont abouti à une algèbre appelée *les relations d'Allen* qui sont au nombre de 13.

La figure 2.8 décrit les 7 relations de base (sans les inverses). Il est à noter que dans le cas des trajectoires sémantiques où $\overline{t_i} \leq \underline{t_{i+1}}$, les relations qu'entretiennent les épisodes entre eux correspondent aux relations de *Precedes* et *Meets*.

De part ce vocabulaire et la formalisation de ces relations, cette puissante structure de données permet de représenter bon nombre d'événements du monde réel. Cela permet entre autres, l'extraction de motifs / règles complexes et significatives pour l'analyste tel que "A précède l'événement B et finit l'événement C". C'est en tout cas ce que proposent des travaux tels que [Höppner and Klawonn, 2001] [Kam and Fu, 2000]. Nous pouvons aussi citer d'autres algorithmes travaillant et prenant en compte les séquences d'intervalles temporels comme les travaux de T. Guyet avec QTempIntMiner [Guyet and Quiniou, 2008] et QTIPrefixSpan [Guyet and Quiniou, 2011a] ou même QTPSpan [Nakagaito et al., 2009] et QFIminer [Washio et al., 2007],

Tous les algorithmes cités précédemment suivent l'algèbre d'Allen et ses relations. Il est toutefois à noter que l'algèbre d'Allen a pu connaître des extensions dans la littérature, notamment Freksa [Freksa, 1992] qui dès 1992 complète cette algèbre par ce qu'on appelle des "intervalles flous", ou demi-intervalle permettant de représenter des données manquant de précision : "Je sais que l'évènement A se passe avant l'évènement B, mais je ne sais pas quand il se termine". Ce faisant, Freksa introduit 11 nouvelles relations venant directement compléter celles d'Allen et introduisant par la même occasion une *mesure d'interlude* quant au manque de précision de la donnée. Ainsi, on peut trouver dans la littérature des algorithmes de sequence mining prenant en compte cette extension de vocabulaire dans des travaux tels que [Lattner et al., 2005] où les auteurs utilisent cette logique dans le pilotage d'un robot.

2.4.5 Conclusion

Dans cette section, nous avons passé en revue plusieurs approches avec des objectifs bien différents. Malgré tout, une limite commune de ces travaux est l'impossibilité de prendre en compte plusieurs trajectoires sémantiques en même temps, elles-mêmes pouvant évoluer sur plusieurs dimensions. Peu de méthodes de sequence mining prennent en compte la sémantique des données et les informations contextuelles. Il y a ici une difficulté à integrer également d'autres informations contextuelles telles que l'âge du visiteur par exemple. Les comportements communs ne sont pas caractérisés avec autant de précision que la trajectoire sémantique peut permettre de l'envisager.

Cette problématique sera traitée dans la contribution 3 de ce manuscrit

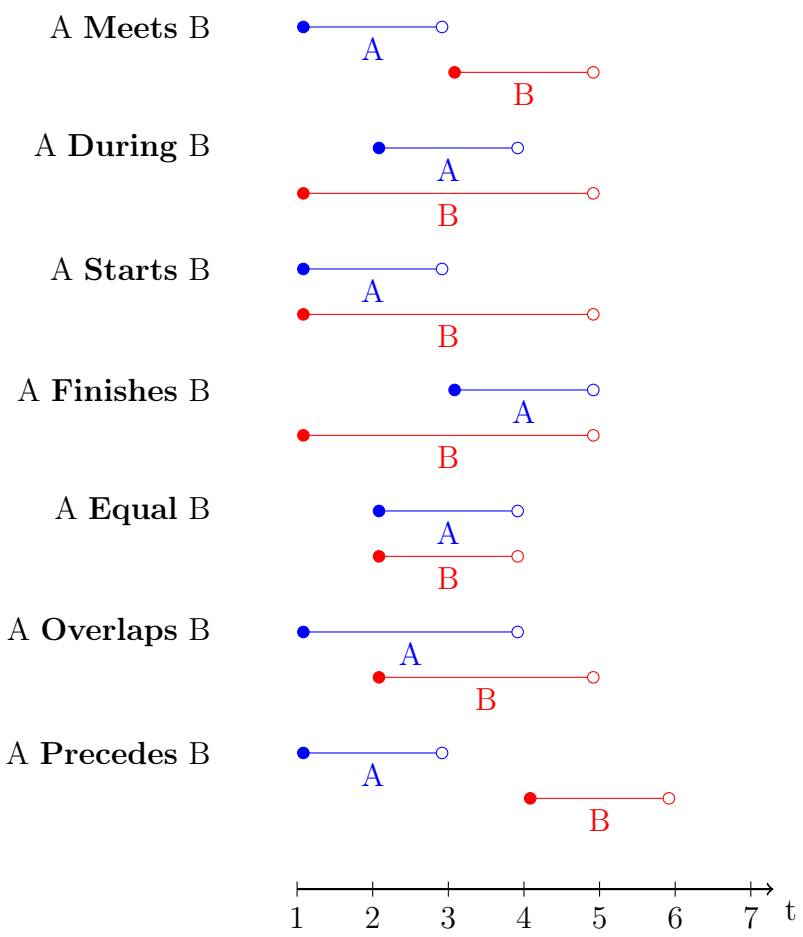


FIGURE 2.8 – Les 7 relations de base d'Allen, entre deux intervalles A et B

où nous explorons la possibilité d'analyser le comportement d'individus à la fois sur un plan spatial mais aussi contextuel (comme des comportements selon la météo, ou selon d'autres facteurs extérieurs). Cette approche sera ensuite complétée par la contribution 4 qui propose d'intégrer l'analyste de données directement dans cette boucle d'analyse de façon à ce qu'il puisse interagir avec les données au cours de l'analyse.

2.5 Conclusion

Dans ce chapitre, nous avons présenté des travaux d'état de l'art abordant différents aspects de la trajectoire. Nous y avons vu une première problématique de reconstruction dans un environnement où il n'est pas aisé d'obtenir de telles trajectoires. En intérieur, les GPS ne fonctionnant pas, nous avons

vu comment il est possible, avec différents degrés de granularité, d'obtenir une trajectoire - i.e une séquence de positions spatio-temporalisées relatant du passage d'un objet de type (t_i, x_i, y_i) avec t_i une information temporelle et (x_i, y_i) des coordonnées géographiques. La mise en place de l'équipement d'indoor positionning dans un environnement contraint constitue une première limite des méthodes existantes. L'absence de vérité terrain en est une seconde. Ces deux contraintes auront un impact significatif sur les performances des différentes méthodes citées. La contribution 1 de ce manuscrit portera ainsi sur deux méthodes permettant de limiter la baisse de performance associée au placement contraint du matériel.

À partir d'une trajectoire, il est ensuite possible de l'enrichir afin de caractériser et de compléter ces données de déplacement par de l'information contextuelle. Notre deuxième contribution consiste à créer un modèle de trajectoire sémantique qui formalisera toutes les données et enrichira la trajectoire brute en y ajoutant des épisodes. Ce processus d'enrichissement nous permettra de structurer les trajectoires sémantiques à la fois en intérieur (par exemple dans les musées) et en extérieur (par exemple les trajets touristiques à La Rochelle). Nous détaillerons également la façon dont nous visualiserons ces trajectoires sémantiques.

Malgré tout, ces modèles de trajectoires sémantiques sont difficiles à analyser tant la sémantique des données est importante et complexe. Ces modèles de trajectoires peuvent être formalisés sous la forme d'une séquence temporelle et ainsi utiliser des méthodes de sequence-mining temporel sur celle-ci afin de pouvoir extraire des comportements communs à un sous-groupe d'individus. Ainsi, nous avons passé en revue les différentes façons d'analyser les trajectoires, mais aussi les séquences et les séquences temporelles qui peuvent être applicables aux trajectoires sémantiques. Cependant, il n'existe pour le moment que peu de méthodes permettant d'analyser plusieurs trajectoires sémantiques en même temps ou d'intégrer d'autres informations contextuelles (comme le profil d'un individu) dans le processus. En contribution 3, nous nous intéresserons à l'analyse conjointe de ces mêmes trajectoires sémantiques.

Dans le chapitre suivant, nous présenterons l'algorithme NextPriorityConcept qui est intégré à la plateforme **GALACTIC**. Nous avons choisi d'utiliser cette plateforme car elle offre la possibilité d'analyser plusieurs séquences couplées à d'autres données de nature différente. **GALACTIC** peut être utilisé pour l'analyse des trajectoires sémantiques, mais aussi pour limiter le déluge de motifs engendré par l'utilisation d'algorithmes classiques de

séquence et de fouille de motifs. **GALACTIC** repose sur l'analyse formelle de concepts et la théorie des treillis, une structure de données permettant de former une hiérarchie de groupes d'individus ayant des comportements communs parmi nos trajectoires sémantiques. Nous détaillerons ces méthodes dans le chapitre suivant.

Chapitre 3

Etat de l'art : Pattern mining et Analyse Formelle de Concepts

Table des matières

3.1 Introduction 49
3.2 *Pattern mining* 50
3.2.1 Notations et vocabulaire 50
3.2.2 La fouille de motifs fréquents 51
3.2.3 Les motifs fermés 52
3.3 L'Analyse Formelle de Concepts 53
3.3.1 Contexte et concepts 54
3.3.2 Treillis des concepts 54
3.3.3 Les structures de motifs 55
3.4 Théorie des treillis 57
3.4.1 Ensemble partiellement ordonné 57
3.4.2 La structure de treillis 58
3.4.3 Les familles de Moore 59
3.4.4 Treillis des fermés 60
3.4.5 La table binaire d'un treillis 61
3.5 L'algorithme NextPriorityConcept 62
3.5.1 Algorithme NextPriorityConcept 63
Principe de l'algorithme 63
Descriptions génériques par des prédicats 64
Stratégies génériques par des prédicats 65
Description de l'algorithme 66
3.6 Conclusion 70

3.1 Introduction

De nos jours, avec le rapide développement de l'informatique et de ses implications, la collecte de données s'est largement imposée comme une composante essentielle de n'importe quel outil numérique. Les données récoltées se sont ainsi naturellement glissées dans la grande majorité des recherches contemporaines toutes disciplines confondues. Ces bases de données volumineuses sont devenues un moyen à part entière de mener à bien des études comportementales. Cependant, cette masse de données doit être "fouillée", c'est à dire analysée par des algorithmes spécifiques afin d'en retirer le maximum de connaissances pertinentes. Dans ce manuscrit, nous nous concentrons sur des données de type séquence avec une temporalité représentant une trajectoire d'un utilisateur.

Ainsi, la fouille de motifs séquentiels consiste à extraire des informations communes à un sous-groupe d'individus. Par motifs communs (ou *pattern* en anglais), nous entendons des suites d'éléments d'une séquence retrouvée dans un certain nombre de trajectoires (sémantiques) des individus : en anglais, des *patterns*. Le plus généralement, cela consiste à extraire dans un ensemble de séquences des sous-séquences communes. Celles-ci donnent une description d'un comportement ou une information sur sous-groupe d'individus de notre jeu de données comme introduit dans le chapitre précédent. Il existe plusieurs types de séquences bien spécifiques, les séquences dites simples sous forme d'une succession d'information sémantique, et les séquences dites temporelles, où chaque élément de cette séquence est relié à une information temporelle sous forme d'un instant ou d'un intervalle temporel.

Dans nos travaux, nous manipulons des données de type différent. Dans le chapitre précédent nous avons introduit la notion de trajectoire sémantique, des trajectoires enrichies de données de contexte, pouvant être de type différent. Comme nous l'avons vu précédemment, les trajectoires sémantiques comportent plus que de l'information spatiale et nous proposerons dans ce manuscrit l'utilisation de l'algorithme NextPriorityConcept [Demko et al., 2020] et de la plateforme **GALACTIC** comme outils d'analyse. Cette plateforme propose une extension de l'Analyse Formelle de Concepts (AFC) pour des traitements de données complexes (comme les séquences) et hétérogènes que nous présenterons dans ce chapitre.

Nous avons choisi d'utiliser **GALACTIC** pour pouvoir analyser plusieurs séquences à la fois, qu'elles soient simples ou temporelles ainsi que d'autres types de données (numériques, catégorielles, etc.) décrivant le profil

du visiteur. Celui-ci est le seul outil disponible actuellement permettant de traiter plusieurs séquences à la fois. **GALACTIC** calcule une hiérarchie de sous-groupes et repose à la fois sur les approches de pattern mining et sur l'analyse formelle de concept. Dans ce chapitre, nous revenons plus en détail sur les algorithmes de pattern mining et sequence mining traité au chapitre précèdent afin d'introduire les fondements de l'Analyse Formelle de Concepts qui reposent sur la théorie des treillis. Nous présenterons enfin la plateforme **GALACTIC** d'analyse de données utilisée dans les contributions de ce manuscrit.

3.2 *Pattern mining*

3.2.1 Notations et vocabulaire

On note Σ un ensemble d'éléments qui seront présents dans un ensemble de données D. Chaque donnée est décrite par un ensemble d'attributs, motifs ou *itemsets* $T \subseteq \Sigma$.

Il est également possible de représenter ces données sous forme d'une relation binaire, où la table binaire dispose des données en lignes et des attributs en colonnes. Comme énoncé dans l'introduction, l'objectif de l'analyse de motifs fréquents dans un dataset réside avant tout dans la capacité de l'algorithme à extraire un ensemble de motifs fréquents. Par motifs fréquents, nous entendons des *itemsets* communs partagés par suffisamment des données de notre dataset. On appelle le support d'un motif $B \subseteq \Sigma$, sa fréquence d'apparition parmi nos données :

$$support(B) = \frac{|\{T \in D | B \subseteq T\}|}{|\{D\}|} \tag{3.1}$$

L'élément $|\{T \in D | B \subseteq T\}|$ de l'équation 3.1 correspondant au nombre d'*itemsets* contenant l'ensemble B dans D.

Un motif sera dit fréquent si son support est supérieur à un seuil donné par l'utilisateur. Alors qu'un *itemset* est un ensemble d'éléments dans un ordre aléatoire - on ne cherchera ici que de savoir si un sous-ensemble est compris dans un autre avec la relation d'inclusion - une séquence simple est quant à elle définie pour une séquence $S = \langle a_i \rangle_{i \leq n}$ comme une liste ordonnée d'éléments a_i. On note que pour tout élément a_i de $S \in D$, on a $a_i \in \Sigma$.

3.2.2 La fouille de motifs fréquents

Le premier algorithme proposé pour adresser le problème de la fouille de motifs fréquents est l'algorithme Apriori [Agrawal and Srikant, 1995]. Cet algorithme génère l'ensemble des *itemsets* fréquents à partir d'un ensemble de données D en fonction d'un seuil minimal de support renseigné par l'utilisateur, où chaque donnée est décrite par un ensemble d'attributs.

L'algorithme Apriori présente la particularité de fonctionner par niveaux successifs. La génération commence par l'ensemble vide (*itemset* $\emptyset$) de support 1, puis les *itemsets* de cardinalité 1, puis 2, et ainsi de suite, sont générés où chaque cardinalité correspondant à un niveau. À chaque étape, seuls les *itemsets* fréquents (c'est-à-dire ceux dont le support est supérieur au seuil) sont conservés. Cette méthode de génération par niveaux est rendue possible grâce à la fonction de support, qui est une fonction monotone décroissante à partir de l'ensemble vide $\emptyset$ vers des *itemsets* de plus en plus grands. C'est-à-dire que si un ensemble est fréquent c'est que tous ses sous-ensembles sont fréquents (si l'ensemble $\{a, b\}$ est fréquent dans D, $\{a\}$ et $\{b\}$ le sont aussi).

Exemple 3.2.1. *On note un dictionnaire $\Sigma = \langle a, b, c, d \rangle$, un support minimum $s = \frac{2}{4}$ et un ensemble d'itemset D correspondant au tableau 3.1 :*

La figure 3.1 montre une représentation de tous les itemsets sur Σ reliés par inclusion. Les noeuds en rouge correspondent aux sous-ensembles fréquents retenus par l'algorithme Apriori pour un seuil de support à 2/4, les autres sous-ensembles ne sont pas retenus.

Individu	***itemset***
1	{ a, c, d}
2	{ a, d}
3	{ a, c, d}
4	{ a, b, c, d}

TABLEAU 3.1 – Ensemble de données D de la figure 3.1

Des améliorations du principe de base de l'algorithme Apriori sont rapidement apparues dans la littérature avec des travaux permettant notamment de réduire le temps de calcul [Zaki, 2000] [Borgelt, 2012] [Zhang et al., 2013] [Vo et al., 2016]. L'algorithme Apriori dessine la base de ce que deviendra le pattern mining, ou fouille de motifs fréquents en français.

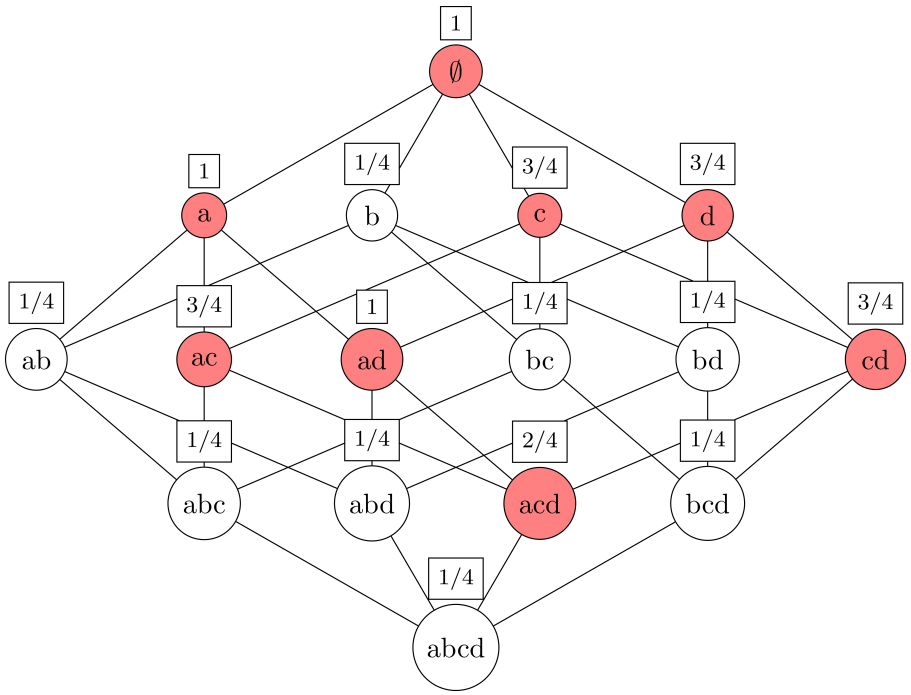


FIGURE 3.1 – Motifs de D du tableau 3.1 sous forme hiérarchique. La valeur de support est indiquée au-dessus des noeuds

3.2.3 Les motifs fermés

Dans l'exemple précédent on peut observer que l'algorithme a généré plusieurs *itemsets* détectés comme fréquents : { c } avec un support 3, { d } avec un support de 3 et { c, d } lui aussi avec un support de 3. Puisque c et d n'apparaissent qu'ensemble dans le jeu de données 3.1, générer les *itemsets* {c}, {d} et {c,d} tous les trois avec un support 3 n'apportent pas plus d'informations que de générer seulement l'*itemset* {c,d}.

Lorsque plusieurs *itemsets* sont partagés par les mêmes objets, et donc ont le même support, ils décrivent la même information et les conserver est redondant. Il se trouve que chaque regroupement possède un *itemset* maximal qui synthétise cette information commune. Cet *itemset* maximal est dit fermé. L'objectif sera ainsi de ne générer que les *itemsets* **fermés** pour réduire le nombre de motifs générés.

Sur cette même logique, de nombreux chercheurs ont travaillé sur des algorithmes dont l'objectif est de ne générer que des *itemsets* fermés. Le premier

algorithme générant des *itemsets* fermés est l'algorithme *A-Close* [Pasquier et al., 1999] et a rapidement été suivi d'algorithmes comme Closet [Pei et al., 2000] ou CHARM [Zaki and Hsiao, 2002]. Cette notion d'itemset fermé est formalisée en Analyse Formelle de Concepts et en théorie des treillis à l'aide d'un opérateur de fermeture, que nous présenterons dans la section suivante. La structure hiérarchique des seuls fermés possède les propriétés de treillis, aussi identifiée en Analyse Formelle de Concepts.

3.3 L'Analyse Formelle de Concepts

C'est en 1982 que la première mention d'Analyse Formelle de Concepts, ou AFC, apparaît dans la littérature, introduite et formalisée par [Wille, 1982]. L'ouvrage de référence de [Ganter and Wille, 1999b] est d'ailleurs régulièrement cité comme étant la fondation de l'AFC. On définit un concept formellement par son intension et son extension. L'extension d'un concept est un sous-ensemble d'objets qui partagent tous la même intension ou description. Nous parlons d'analyse formelle dans le sens où ce type d'analyse se base sur des fondamentaux définie de façon théorique sous forme algébrique.

Exemple 3.3.1. *L'intension du nombre 2 peut être "Nombre pair et nombre premier". "Nombre pair" est une intension s'appliquant à 2 comme à 4,6,8 etc ..*

L'AFC vise à extraire des concepts à partir d'un contexte. On entend par contexte une relation binaire entre un ensemble d'objets et un ensemble d'attributs. Ainsi, nous pouvons définir plus formellement un concept comme étant un sous-ensemble d'objets partageant le même sous-ensemble d'attributs. Cette relation binaire entre objets et attributs peut se formaliser sous la forme d'une table binaire ou encore par l'ensemble des attributs reliés à chaque objet tel qu'introduit par l'algorithme Apriori où chaque objet est associé à un itemset.

La connexion entre les objets et leurs attributs est appelé une connexion de Galois, introduits pour la première fois en 1944 [Ore, 1944]. Dans l'analyse formelle de concepts, nous utilisons la propriété de connexion de Galois afin de générer le treillis de concepts qui correspond à la hiérarchie des concepts extraits de notre ensemble de données et d'attributs.

Dans cette section nous traiterons des fondements mathématiques sur lesquels repose l'Analyse Formelle de Concepts et plus particulièrement des treillis de concepts issus de nos données.

3.3.1 Contexte et concepts

Definition 2. *On appelle contexte un triplet (G,M,I) avec G un ensemble d'objets ($G \neq \emptyset$), M un ensemble d'attributs ($M \neq \emptyset$) et I une relation binaire entre les objets et leurs attributs.*

Avec un couple (a, x), $a \in G$ et $x \in M$ en relation par I est noté $(a, x) \in I$ ou a I x. Le contexte est ici une autre façon de définir des données en tableau de type binaire et exprime que l'objet a possède l'attribut x.

À partie de la relation I , on définit deux opérateurs α et β associant les objets avec leurs attributs. On les appelle des opérateurs de dérivation. Ces opérateurs forment une connexion de Galois (α, β).

Nous définissons ainsi $\beta : 2^G \to 2^M$ qui est l'application associant un sous-ensemble d'attributs $A \subseteq G$, à un sous ensemble d'objets $B \subseteq M$ comme étant l'ensemble des objets partageant le sous-ensemble d'attributs B.

$$\beta(B) = \{a | a \in G \text{ et } \forall b \in B, (a, b) \in I\} \tag{3.2}$$

À l'inverse, nous définissons $\alpha : 2^M \to 2^G$ comme l'application associant un sous-ensemble d'objets $B \subseteq M$ au sous ensemble d'attributs $A \subseteq G$ comme étant l'ensemble des attributs communs à un sous-ensemble d'objets A.

$$\alpha(A) = \{b | b \in M \text{ et } \forall a \in A, (a, b) :\in I\} \tag{3.3}$$

3.3.2 Treillis des concepts

À partir d'un contexte (G, M, I), nous définissons la notion de concept et de treillis des concepts :

Definition 3. *Un concept est un couple (A, B), $A \subseteq G$ et $B \subseteq M$, associant un ensemble maximal d'objets avec leurs attributs communs défini par $A = \beta(B)$ et $B = \alpha(A)$. A est appelé l'extension et B l'intension du concept.*

Il est possible d'obtenir plusieurs concepts à partir d'un seul contexte qui s'organisent sous la forme d'un treillis.

Pour définir un treillis de concepts, il est nécessaire de définir la relation d'ordre entre les concepts appelée la relation de généralisation/spécialisation.

Definition 4. *La relation de généralisation / spécialisation est une relation d'ordre sur un ensemble de concepts C.*

$$(A_2, B_2) < (A_1, B_1) \iff A_2 \subset A_1, B_1 \subset B_2 \tag{3.4}$$

La relation de généralisation/spécialisation est similaire à une relation d'inclusion d'ensembles, avec deux concepts $c_1 \leq c_2 \in C$ où $c_1 = (A_1, B_1)$ et $c_2 = (A_2, B_2)$. Ici, c_2 est appelé un sous-concept de c_1 (à l'inverse, c_1 est un super-concept de c_2). On parle ainsi de relation de généralisation / spécialisation, car c_1 est un concept plus général dans le sens où il regroupe un plus grand nombre d'objets ($A_2 \subset A_1$) mais un plus petit nombre d'attributs ($B_1 \subset B_2$). À l'inverse, c_2 est un concept renforçant la "*spécialisation*" dans le sens où il comporte un plus petit nombre d'objets mais un plus grand nombre d'attributs. Nous pouvons interpréter cette relation comme une "double relation d'inclusion" sur les objets et les attributs.

Definition 5. *On appelle un treillis de concepts $L = (C, \leq)$ associé à l'ensemble de tous les concepts C doté de cette relation de généralisation / spécialisation $\leq$.*

Les deux principaux algorithmes de génération d'un treillis de concepts sont l'algorithme de Bordat [Bordat, 1986] et l'algorithme NextClosure [Ganter and Wille, 1999a]. Ces algorithmes fonctionnent selon une approche de génération niveau par niveau, similaire à Apriori et A-close, mais ils diffèrent dans la façon dont les concepts sont générés et ordonnés dans le treillis. L'algorithme de Bordat utilise une approche bottom-up et génère les concepts en se basant sur des opérations de fusion, tandis que l'algorithme NextClosure utilise une approche top-down et génère les concepts en se basant sur des opérations de clôture. Ces algorithmes sont largement utilisés dans le domaine de la fouille de données pour la classification, la visualisation et l'analyse de données complexes.

Un treillis de concepts est ainsi composé d'objets et de leurs plus petites descriptions communes avec la relation d'ordre de généralisation / spécialisation. De ces treillis, nous pouvons extraire des règles d'implication et des descriptions de groupes d'objets. On ne parlera pas des règles d'implication ici, mais de nombreux travaux sont disponibles sur ce sujet spécifiquement [Hu et al., 1999] [Pasquier et al., 1999] [Bertet and Monjardet, 2010] [Shemis and Mohammed, 2021]

Exemple 3.3.2. *La figure 3.2 présente le diagramme de Hasse du treillis de concepts $L = (C, \leq)$ généré à partir du contexte (G, M, I) avec $G = 1, 2, 3$ et $M = a, b, c$ où C représente l'ensemble des concepts décrits par le tableau 3.2.*

3.3.3 Les structures de motifs

L'AFC propose des algorithmes efficaces, mais ils ne peuvent s'appliquer qu'à des contextes ou des données binaires. Les structures de motifs (ou

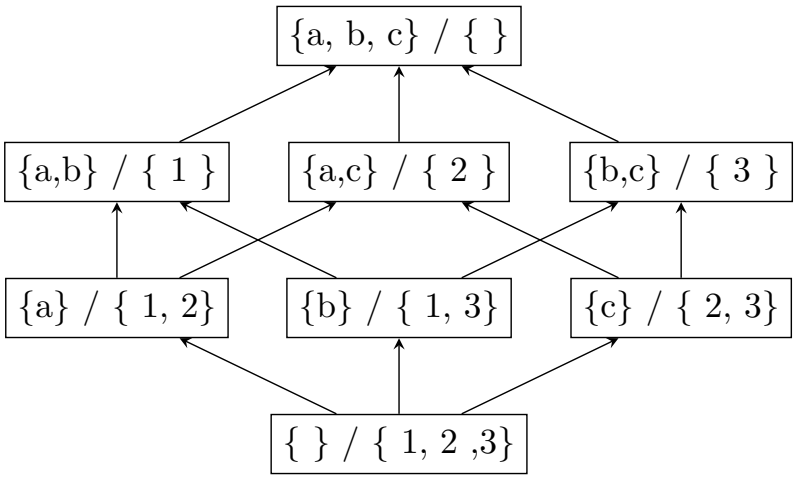


FIGURE 3.2 – Diagramme de Hasse du treillis de concepts $L = (C, \leq)$ de l'exemple 3.3.2

Attributs / Objets	a	b	c
1	X	X	
2	X		X
3		X	X

TABLEAU 3.2 – Contexte (G, M, I)

pattern structure) [Ganter and Kuznetsov, 2001] sont une extension de l'AFC où les données sont décrites sur un espace de description quelconque défini de façon générique par une structure de motifs. Une structure de motif est un triplet $(G, (D, \sqcap), \delta)$ où G est un ensemble d'objets, $(D, \sqcap)$ est un espace de description muni d'un opérateur $\sqcap$. De façon similaire à l'AFC, on définit les opérateurs α et β qui associent des objets à leur description commune :

— $\alpha_D : 2^G \to D$ est une application qui associe une description d pour chaque sous-ensemble $A \subseteq G$ tel que :

$$\alpha_D(A) = \sqcap_{g \in A} \delta(g) \tag{3.5}$$

$\alpha_D(A)$ est la description commune de l'ensemble d'objets A.

— $\beta_D : D \to 2^G$ est une application qui associe un sous-ensemble $A \subseteq G$ pour chaque description $d \subseteq D$ tel que :

$$\beta_D(d) = \{g \in G \mid d \sqsubseteq \delta(d)\} \tag{3.6}$$

$\beta_D(d)$ est l'ensemble des objets qui partagent la description d.

Ainsi, on peut définir la notion de concept et de treillis des concepts en utilisant α_D et β_D. Les concepts sont de la forme (A, d), $A \subseteq G$, $d \in D$ tel que $\alpha_D(A) = d$ et $A = \beta_D(d)$. La relation d'ordre est définie entre les concepts par :

$$(A_2, d_2) < (A_1, d_1) \Longleftrightarrow A_2 \subset A_1 (\Longleftrightarrow d_1 \sqsubset d_2) \tag{3.7}$$

L'ensemble des concepts doté de cette relation forme un treillis de motifs (ou pattern lattice).

3.4 Théorie des treillis

Le treillis des concepts introduit en AFC repose sur la structure de treillis définie par des propriétés algébriques en théorie des treillis. Cette section est consacrée à la structure de treillis et à ses propriétés qui impactent sur la compréhension et la manipulation algorithmique du treillis des concepts.

3.4.1 Ensemble partiellement ordonné

Un treillis est une relation d'ordre spécifique. Définissons tout d'abord ce qu'est une relation d'ordre.

Definition 6. *Une relation binaire $\leq$ sur un ensemble S est une relation d'ordre si :*

- $\leq$ *est réflexive : si* $\forall x \in S, x \leq x$
- $\leq$ *est antisymétrique : si* $\forall x, y \in S$, *si* $x \leq y$ *et* $y \leq x$ *alors on a* $x = y$
- $\leq$ *est transitive : si* $\forall x, y, z \in S$, $x \leq y$ *et* $y \leq z$ *alors on trouve* $x \leq z$

Lorsque $x \leq y$, on dit que x est un prédécesseur de y, ou encore que y est un successeur de x. Il est à noter que tous les éléments $x, y \in S$ peuvent ne pas être comparables. C'est-à-dire, si à la fois $x \not\leq y$ et $x \not\geq y$, alors x et y sont incomparables.

Une relation d'ordre classique est la relation d'inclusion $\subseteq$ définie sur l'ensemble des parties d'un ensemble.

3.4.2 La structure de treillis

Dans la littérature, il existe deux définitions pour formaliser ce qu'est un treillis, une définition ordinale et une définition algébrique.

Definition 7. *On note une relation d'ordre $L = (S, \leq)$. L'ensemble L est un treillis si, pour toutes les paires d'éléments $x, y \in S$:*

- *La borne inférieure, ou infimum, de deux éléments $x, y \in S$ existe et se note $x \wedge y$, il s'agit de l'unique élément maximal de l'ensemble des prédécesseurs communs à x et y.*
- *La borne supérieure, ou supremum, de deux éléments $x, y \in S$ existe et se note $x \vee y$, il s'agit de l'unique élément minimal de l'ensemble des successeurs communs à de x et y.*

Definition 8. *Un treillis est un triplet $L = (S, \vee, \wedge)$ où $\wedge$ et $\vee$ sont deux opérateurs binaires sur l'ensemble S qui satisfont les propriétés suivantes :*

- *L'associativité, où pour tout $x, y, z \in S$ on a $(x \vee y) \vee z = x \vee (y \vee z)$*
- *La commutativité, où pour tout $x, y \in S$, on a $x \vee y = y \vee x$ et $x \wedge y = y \wedge x$*
- *l'idempotence, où pour tout $x \in S$, on a $x \wedge x = x = x \vee x$*
- *La loi de l'absorption, où pour tout $x, y \in S$ on a $x(x \wedge y) = x = x \vee (x \vee z)$*

Cette définition fut introduite par Birkhoff en 1940 [Birkhoff, 1940] et n'utilise que les opérateurs algébriques $\wedge$ et $\vee$.

On peut étendre l'infimum et le supremum à un ensemble $A \subseteq S$ avec $\wedge A$ et $\vee A$ comme étant respectivement le résultat de la borne inférieure et supérieure de tous ses éléments.

On note ainsi qu'un treillis contient toujours un minimum global que l'on appelle le *bottom* noté $\perp$, défini par $\perp = \bigwedge \{y \in S\} = \bigwedge S$. À l'inverse, tout treillis possède un maximum global que l'on note $\top$ et qui est le supremum de L défini par $\top = \bigvee \{y \in S\} = \bigvee S$.

Le diagramme de Hasse est une représentation de la relation d'ordre où les relations de réflexivité (qui correspondent aux boucles) et de transitivité (qui correspondent à des chemins) ne sont pas représentés car elles se déduisent facilement, cela participe à une meilleure lisibilité dans la visualisation des treillis. Dans ce manuscrit, les treillis seront représentés sous la forme d'un diagramme de Hasse.

La figure 3.3 montre le diagramme de Hasse d'un treillis $L = (S, \leq)$ défini sur $S = \{a, b, c, d, e, f, g, i\}$. Pour l'exemple, dans le cas de ce treillis L, on trouve $\top$ = a, $\bot$ = i, la borne supérieure de d et e est $d \vee e = b$ et la borne inférieure des éléments d et f est $d \wedge f = i$.

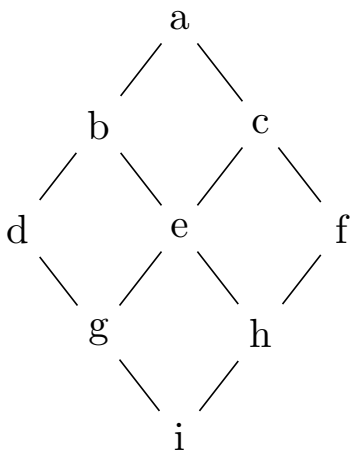


FIGURE 3.3 – Diagramme de Hasse du Treillis L.

3.4.3 Les familles de Moore

Une famille $\mathcal{F} \in \mathcal{P}(S)$ sur un ensemble S muni de la relation d'ordre $\subseteq$ est un ensemble partiellement ordonné. Si cet ensemble est stable par intersection ($\forall F, F' \in \mathcal{F}, F \cup F' \in \mathcal{F}$) nous appelons cet ensemble une famille de Moore, [Moore, 1909] et celle-ci garantit ainsi une structure de treillis (figure 3.4), avec $L = (\mathcal{P}(S), \subseteq)$. On retrouve des exemples de familles de Moore dans les hiérarchies, où une hiérarchie est une famille $\mathcal{F}$ sur un ensemble S se contenant lui-même ainsi que l'ensemble vide $\emptyset$. Il est également à noter que les *itemsets* générés par Apriori forment une famille de Moore.

Exemple 3.4.1. *Soit un ensemble $S = \{a, b, c, d\}$, la famille de Moore $\mathcal{F} = \mathcal{P}(S)$ correspondant à toutes les parties de S est $\mathcal{F} = \langle \emptyset$, a, b, c, d, ab, ac, ad, bc, bd, cd, abc, abd, acd, bcd, abcd $\rangle$. Puisque $\mathcal{F}$ est stable par intersection, $L = (\mathcal{F}, \subseteq)$ est un treillis. Il est représenté figure 3.4*

On peut noter que le treillis des concepts correspond à la combinaison de deux familles de Moore : l'une est définie sur l'ensemble des attributs avec la relation d'inclusion $\subseteq$, l'autre est définie sur l'ensemble des objets avec la relation d'inclusion inversée $\supseteq$.

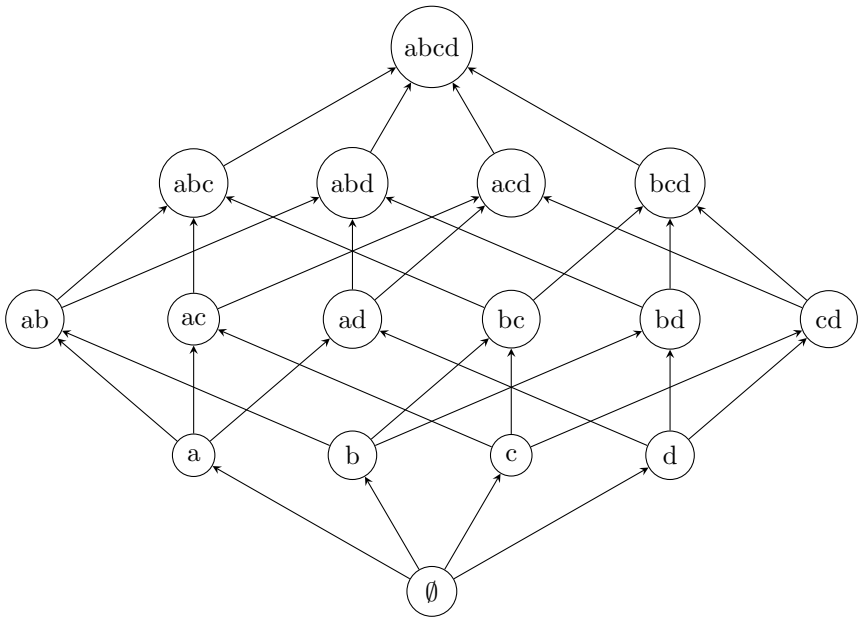


FIGURE 3.4 – Diagramme de Hasse du treillis L de la famille de Moore $\mathcal{F} = \mathcal{P}(S)$ sur l'ensemble S = {a,b,c,d}

3.4.4 Treillis des fermés

On obtient également une famille de Moore, et donc un treillis, à l'aide d'un opérateur de fermeture.

Definition 9. *Un opérateur de fermeture φ est une application sur un ensemble S qui garantit les propriétés suivantes :*

- ***L'extensivité*** *: $\forall X \subseteq S$, alors on a $X \subseteq \varphi(X)$*
- ***L'isotonie*** *: $\forall X, Y \subseteq S$ avec $X \subseteq Y$, alors on a $\varphi(X) \subseteq \varphi(Y)$*
- ***L'idempotence*** *: $\forall X \subseteq S$, on a $\varphi(\varphi(X)) = \varphi(X)$*

Un ensemble $X \subseteq S$ tel que $X = \varphi(X)$ est appelé un ensemble fermé. L'ensemble $\mathcal{F} = \{X \subseteq S : \varphi(X) = X\}$ de tous les fermés forme une famille de Moore, et $\{\varphi, \subseteq\}$ est donc un treillis.

En AFC, la composition $\varphi = \alpha \circ \beta$ des deux opérateurs α et β forme un opérateur de fermeture sur les attributs, et la composition $\beta \circ \alpha$ forme un opérateur de fermeture sur les objets. Ces deux opérateurs de fermeture permettent de générer des fermés qui correspondent aux deux familles de Moore imbriquées qui composent un treillis des concepts. Il est également

à noter que les *itemsets* fermés introduits par l'algorithme A-Close décrit dans la section 3.2.3 correspondent aux ensembles fermés de l'opérateur de fermeture $\alpha \circ \beta$.

3.4.5 La table binaire d'un treillis

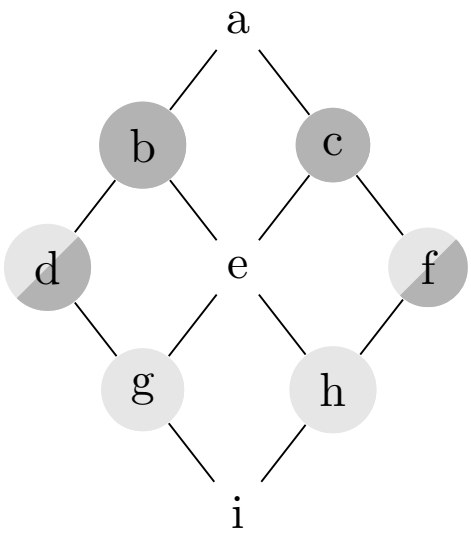


FIGURE 3.5 – Exemple : Un treillis L - join-irréductible ■ meet-irréductible ■

Dans un treillis $L = (S, \leq)$, on distingue des éléments particuliers dit irréductibles, qui ne sont pas une borne supérieure ou inférieure :

— Un sup-irréductible est un élément $j \in S$ qui n'est pas la borne supérieure d'un ensemble ne le contenant pas. Formellement, si $j = \bigvee X$ pour $X \subseteq S$ et $j \notin X$, alors $j \in J$ où J est l'ensemble des sup-irréductibles de L. La caractérisation d'un sup-irreductible est qu'il ne possède qu'un seul prédécesseur immédiat.

— Un inf-irréductible est un élément $m \in S$ qui n'est pas la borne supérieure d'un ensemble ne le contenant pas. Formellement, si $m = \bigwedge X$ pour $X \subseteq S$ et $m \notin X$, alors $m \in M$ où M est l'ensemble des inf-irréductibles de L. La caractérisation d'un inf-irreductible est qu'il ne possède qu'un seul successeur immédiat.

On appelle la table binaire ou le contexte réduit d'un treillis $L = (C, \leq)$, la table binaire $(J, M, \leq)$ composée des sup-irreductibles, des inf-irrreductibles

du treillis L et de la relation binaire $\leq$ du treillis. En posant les sup-irréductible en colonne et les inf-irréductible en ligne, on obtient la table binaire de ce treillis. Le théorème fondamental de la théorie des treillis [Barbut and Monjardet, 1970] est le suivant :

Théorème 1. *Chaque treillis est isomorphe au treillis des concepts de sa table binaire.*

Dans les faits, cela signifie qu'avec le contexte réduit d'un treillis, il est possible de le reconstruire. La figure 3.6 montre la table binaire du treillis de la figure 3.5, avec ses sup-irréductibles en ligne et inf-irréductibles en colonne et le treillis des concepts de cette table binaire.

	d	b	c	f
d	x	x		
f			x	x
g	x	x	x	
h		x	x	x

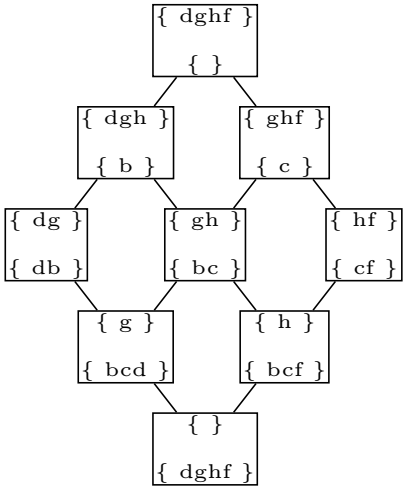


FIGURE 3.6 – Contexte réduit du treillis de la figure 3.5 et son treillis de concepts associé

Clairement, les deux treillis (figure 3.5, figure 3.6) sont isomorphes. Les éléments non-irréductibles peuvent ainsi être recréés à partir de combinaisons de borne supérieures et borne inférieures d'éléments irréductibles.

3.5 L'algorithme NextPriorityConcept

Fort des enseignements et des concepts théoriques de l'Analyse Formelle de Concepts et de la théorie des treillis, l'algorithme NextPriorityConcept fut introduit par Christophe Demko et Karell Bertet, [Demko et al., 2020] et propose un moyen d'utiliser l'Analyse Formelle de Concepts guidée par l'utilisateur pour des données complexes et hétérogènes. Autour

d'un écosystème en constante évolution, l'algorithme est directement intégré dans la plateforme **GALACTIC** (pour **GA**lois **LA**ttices, **C**oncept **T**heory, **I**mplicational systems and **C**losures). Avec cette plateforme, l'utilisateur choisit des *caractéristiques* permettant de représenter le type de données à traiter, des *descriptions* qui décrivent un ensemble de données et des *stratégies* décrivant la manière dont les données seront analysées.

La particularité de l'algorithme NextpriorityConcept est qu'il intègre la notion de *prédicats* monadiques de façon générique, utilisés par les descriptions et les stratégies. Un utilisateur pourra ainsi utiliser des descriptions et des stratégies différentes pour analyser des données hétérogènes d'un même jeu de données. En résultent des concepts intégrant plusieurs types de données et permettant l'analyse de données complexes et hétérogènes.

Dans cette partie nous décrirons le fonctionnement de l'algorithme NextPriorityConcept.

3.5.1 Algorithme NextPriorityConcept

Principe de l'algorithme

L'algorithme NextPriorityConcept se concentre sur les objets s'inspire de l'algorithme de Bordat. Il considère l'ensemble entier G d'objets dès le départ. Ensuite, pour chaque concept (A, B), il ne teste pas un nouvel objet potentiel, mais un nouvel attribut potentiel $b \in M/B$ décrivant un sous-ensemble de A. De cette manière, A diminue tandis que B augmente, ce qui correspond à la relation de prédécesseur. Ce processus correspond à la version duale du théorème de Bordat affirmant que les prédécesseurs immédiats de (A, B) sont les sous-ensembles d'inclusion maximale de la famille suivante sur les attributs M :

$$\mathcal{FP}_{(A,B)} = \{\beta(b) \cap A : b \in M/B\} \tag{3.8}$$

L'algorithme NextPriorityConcept est une nouvelle version de l'algorithme de Bordat où la récursion est remplacée par une file de priorité utilisant le support des concepts. À chaque itération, le concept (A, B) de support maximal est produit, puis ses prédécesseurs immédiats sont calculés, qui correspondent aux ensembles d'inclusion maximale de $\mathcal{FP}_{(A,B)}$, puis ils sont stockés dans la file de priorité. Par conséquent, les concepts sont générés niveau par niveau, en commençant par le concept supérieur $(G, \alpha(G))$, et chaque concept est généré avant ses prédécesseurs.

Descriptions génériques par des prédicats

Pour représenter les données, NextPriorityConcept utilise des *descriptions* δ. δ est une application permettant de fournir le plus petit ensemble de prédicats décrivant un ensemble d'objets $A \subseteq G$, sur la base de leurs caractéristiques. Une caractéristique décrit nos objets et peut être numérique, discrète, sémantique, temporelle, etc. Un concept $(A, \delta(A))$ est composé d'un sous-ensemble d'objets de $A \subseteq G$ et d'un ensemble de prédicats $\delta(A)$ décrivant les objets. Un exemple de prédicat peut être "*is lesser than c*", pour une donnée numérique où c est le maximum des valeurs numériques des objets de A ou "*match subsequence s*" où s est une sous-séquence commune maximale des caractéristiques séquences de A. Les algorithmes de génération sont différents selon le type de données et renvoient des prédicats $\delta'_i(A)$ spécifiques à chaque type de caractéristiques. La *description* finale δ est l'union de tous ces prédicats de telle sorte que :

$$\delta(A) = \bigcup_{i \in 1...n} \delta_i(A) \tag{3.9}$$

Exemple 3.5.1. *L'exemple du tableau 3.3 présente un ensemble G composé de cinq objets (pokémons) avec leurs types d'éléments (feu, vol ou électrique), leurs niveaux et leurs identifiants. Dans cet ensemble de données nous trouvons deux types de caractéristiques différentes :* $C^1 = \{Type\}$ *et* $C^2 = \{Niveau\}$

id	Nom	Type	Niveau
1	Pikachu	{ Électrique }	12
2	Salamèche	{ Feu }	10
3	Roucoups	{ Vol }	6
4	Dracaufeu	{ Feu, Vol }	36
5	Électhor	{ Électrique, Vol }	50

Tableau 3.3 – Exemple de données catégorielles et numériques

Ainsi la première caractéristique C^1 *est un ensemble d'éléments catégoriels, qui seront traités par des prédicats de type "has elements in common with X". La caractéristique* C^2 *décrite par des prédicats de type "is greater than/ is lesser than X".*

La description du sous ensemble d'objets $A = \{3, 4, 5\} \subseteq G$ *est la suivante :*

— *Une première description* δ^1 *obtenue pour la caractéristique* $C^1 = Type$ *tel que* $\delta^1 = \{match\ Type\ [Vol]\}$. *L'élément "Vol" étant commun aux individus 3, 4 et 5.*

— *La seconde description* δ^2 *obtenue pour la caractéristique* $C^2 = Niveau$ *produit deux prédicats :* $\delta^2 = \{Niveau\ \geq 6, Niveau\ \leq 50\}$.

La description finale du sous ensemble A est donc :

$$\delta(A) = \delta^1 \cup \delta^2 = \{match\ Type\ [Vol], Niveau\ \geq 6, Niveau\ \leq 50\}$$

Le concept $(A, \delta(A))$ *exprime donc l'information que dans le groupe des individus 3, 4 et 5, tous sont de type "Vol" et leurs niveaux sont compris entre 6 et 50.*

Stratégies génériques par des prédicats

À partir d'un concept $(A, \delta(A))$, une stratégie σ permet de générer des prédicats $\sigma(A)$, chaque prédicat étant vérifié par un sous-ensemble non vide $A' \subset A$. Ainsi, une stratégie permet de proposer des sous-ensembles $A' \subset A$ candidats pour être des prédécesseurs immédiats du concept $(A, \delta(A))$ dans le treillis. Une stratégie σ est donc une application $\sigma : 2^G \to 2^P$ où P est l'ensemble de tous les prédicats possibles.

Tout comme les descriptions, une ou plusieurs stratégies σ peuvent être utilisées pour traiter un ensemble de données hétérogènes. La stratégie finale utilisée sera donc l'union de toutes les stratégies utilisées :

$$\sigma(A) = \bigcup_{i \in 1...n} \sigma_i(A) \tag{3.10}$$

Plusieurs stratégies sont disponibles pour proposer des prédécesseurs d'un concept. Les stratégies naïves permettent de considérer tous les prédécesseurs possibles d'un concept, alors que des stratégies non-naïves permettent de réduire le nombre de prédécesseurs. Ainsi, la façon dont seront sélectionnés les sous-ensembles A' à chaque niveau dépend des stratégies entrées par l'utilisateur.

Alors que les stratégies naïves permettent de générer tous les sous-ensembles d'un concept $(A, \delta(A))$ candidats pour le niveau suivant, des stratégies plus élaborées vont se focaliser sur une particularité précise des caractéristiques. Par exemple, la stratégie **PMS** [Boukhetta et al., 2020a] (**P**refix **M**atch **S**trategy) pour des séquences va se focaliser sur le début des séquences. Les stratégies non-naïves sont donc un moyen de réduire le déluge de motifs tout en garantissant la structure de treillis. Les stratégies sont un moyen de réduire

le déluge de motifs, problème intrinsèque au domaine du pattern mining et sequence mining dont nous avons discuté précédemment, en se focalisant sur une particularité précise d'une caractéristique.

Description de l'algorithme

L'algorithme NextPriorityConcept est un algorithme de génération de concepts niveau par niveau. À partir de l'élement top du treillis $(G, \delta(G))$, l'algorithme générera les prédécesseurs $(A, \delta(A))$ avec $A \subset G$ puis réitère sur les concepts ainsi générés. Le retour de l'algorithme est un treillis des concepts $L = (G, P, \leq)$ avec G l'ensemble des objets de l'ensemble de données de départ ; P l'ensemble des prédicats de description des concepts ; la relation d'ordre $\leq$ est la relation d'ordre de généralisation / spécialisation définie ci-avant.

Le premier objectif de l'algorithme NextPriorityConcept est de garantir la structure du treillis renvoyé en retour d'algorithme. Puisque c'est un algorithme de génération par le haut, il faut garantir l'existence de l'opérateur de borne inférieure. Pour maintenir les opérateurs du treillis au moment de la génération des sous-concepts, l'algorithme NextPriorityConcept introduit un mécanisme de propagation de contraintes entre concepts garantissant l'existence de leurs bornes inférieures et aux potentiels sous-prédécesseurs de ceux-ci.

Definition 10. *Le mécanisme de propagation de contraintes est défini comme une application $\mathcal{C}[A]$ définie pour l'intension d'un concept (A,D) vers 2^P par $C[A] = C_{residual} \cup C_{cross}$ avec :*

— *$C_{residual}$ est l'ensemble des contraintes résiduelles issues du concept (A', D') qui a généré (A, D) :*

$$C_{residual} = C[A']D$$

— *C_{cross} est l'ensemble des contraintes issues des autres prédécesseurs immédiats (A_i, D_i) de (A, D) :*

$$C_{cross} = (\bigcup_i A_i \cap \sigma(A'))D$$

— *et $C[G] = \emptyset$*

Théorème 2. *Si $\delta(A) \sqsubseteq \delta(A')$ alors l'algorithme* NextPriorityConcept *génère le treillis des concepts de (G, P, I) où P est l'ensemble des prédicats fournis par les descriptions et $I = \{(a, p) : a \in G, p \in P$ et $p(a)\}$, $p(a)$ signifiant que l'objet a vérifie le prédicat p.*

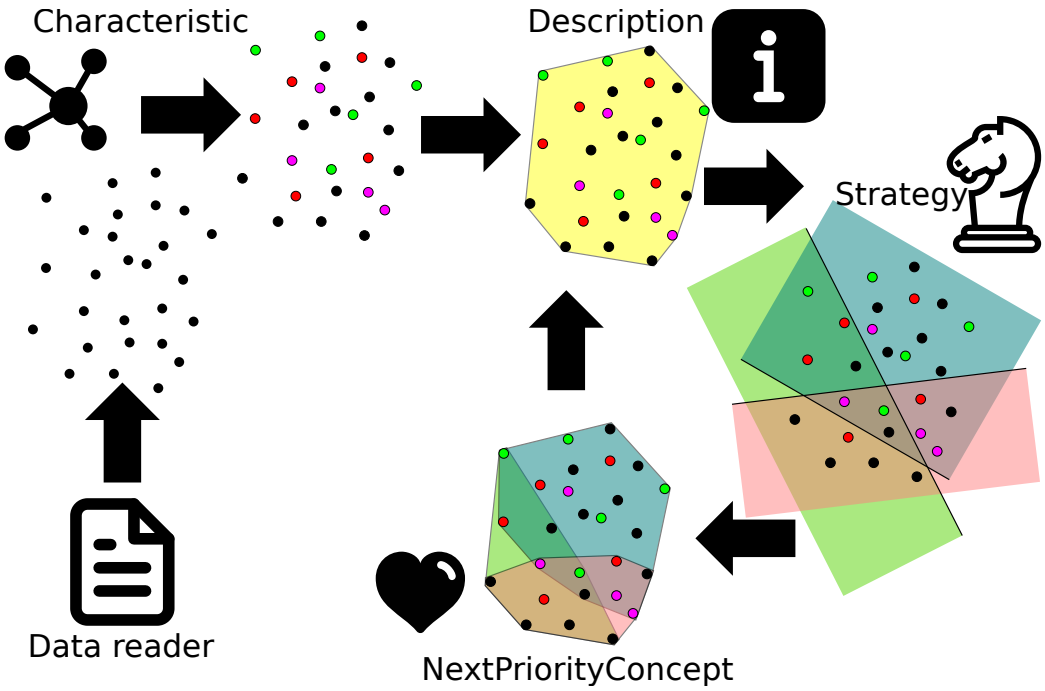


FIGURE 3.7 – Le processus de l'algorithme NEXTPRIORITYCONCEPT [Demko et al., jun 2022]

Ainsi, un sous-concept $(A', \delta(A'))$ issu d'un concept $(A, \delta(A))$ est généré suivant les descriptions de $(A, \delta(A))$ mais aussi des contraintes de A, les sélecteurs sont donc issus de $\sigma(A) \cup C[A]$. De plus, d'après une extension du théorème de Bordat les prédécesseurs immédiats de $(A, \delta(A))$ sont les sous-ensembles maximaux par inclusion de :

$$\mathcal{FD}_{(A,D)} = \{\{a \in A \mid p(a)\} | p \in (\sigma(A) \cup \mathcal{C}[A])\}$$

La figure 3.7 montre le processus de traitement des données de l'algorithme NEXTPRIORITYCONCEPT.

Exemple 3.5.2. *Reprenons l'exemple du jeu de données du tableau 3.3. Nous allons, à l'aide des descriptions vues précédemment et de stratégies, générer le treillis de ce jeu de données en se basant sur les deux types de caractéristiques*

présentes c'est à dire C^1 = *Niveau de type numérique et* C^2 = *Type de type Catégoriel.*

Une première stratégie s^1 *à appliquer sur le jeu de données est la stratégie CompleteMatchStrategy sur la caractéristique* C_1 *afin de dresser un premier treillis.*

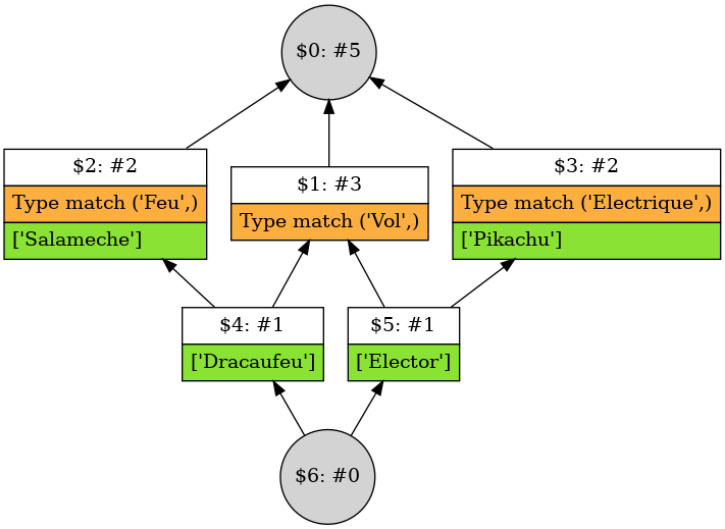


FIGURE 3.8 – Treillis généré avec la stratégie s^1

La figure 3.8 montre un treillis généré avec la seule stratégie s^1. *La stratégie CompleteMatchStrategy crée des prédicats à partir des sous-ensembles d'éléments communs dans nos données. Ainsi, le concept* \$1 *contient trois individus : "Roucoups", "Dracaufeu" et "Électhor" qui sont tous de type "Vol". Le concept* \$2 *contient deux individus : "Salaméche" et "Dracaufeu" partageant le type "Feu". Enfin le concept* \$1 *contenant deux éléments : "Pikachu" et "Électhor" de type Electrique. La stratégie va ensuite ajouter un attribut pour chaque concept* $\$1, \$2, \$3$ *qui va se manifester en ajoutant un élément aux éléments communs des concepts. De cette façon,* \$4 *est obtenu par l'ajout de l'élément "Vol" sur la description du concept* \$2 *: seul le pokémon "Dracaufeu" possède les types correspondants à { "Feu", "Vol" }. Celui-ci est aussi généré en ajoutant le type "Feu" à la description du concept* \$1 *d'où les deux arcs issus de ce concept ; Cela montre que ce concept est à l'intersection du concept* \$2 *et* \$1, *autrement dit la borne inférieure de ceux-ci (on note alors* $\$4 = \$1 \wedge \$2$*). Le concept* \$5 *est lui à l'intersection des concepts* \$1 *et* \$3 *et est généré de la même façon : seul le pokémon "Électhor" possède les types { "Vol", "Electrique" }.*

On introduit maintenant la stratégie s^2 *qui va permettre de classer les pokémons par quantiles de niveaux.*

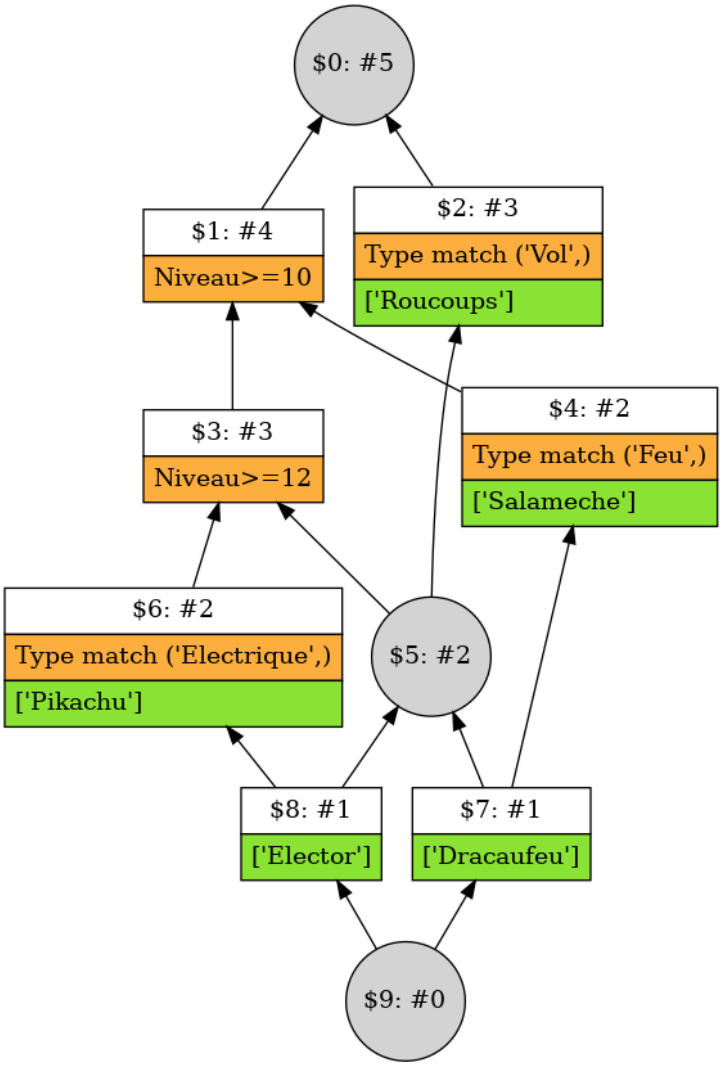


FIGURE 3.9 – Treillis généré avec les deux stratégies s^1 et s^2

La figure 3.9 montre le treillis formé de l'union des deux précédentes stratégies, s^1 et s^2. Ses concepts sont formés de plusieurs descriptions formant ainsi des concepts plus complexes et hétérogènes, tels que le concept \$5 *dont la description est : "Niveau supérieur à 12 et de type Vol" comportant "Dracaufeu" et "Électhor".*

3.6 Conclusion

Dans ce chapitre, nous avons présenté différents algorithmes issus des bases posées par l'algorithme Apriori. Cet algorithme est le premier à proposer une méthode de fouille de motifs fréquents dans des données binaires. Par motif nous entendons les attributs communs à un groupe d'objets. La particularité de cet algorithme est qu'il fonctionne niveau par niveau, ajoutant de plus en plus de données à chaque niveau, décrivant ainsi des sous-groupes plus petits. Les évolutions de l'algorithme se sont en partie concentrées sur la génération de sous-groupes fermés afin de réduire le déluge de motif, le surplus d'informations en sortie d'algorithme. Ces sous-groupes fermés ont entre autres la particularité de posséder la propriété de treillis des fermés.

L'algoritme NextPriorityConcept repose sur l'Analyse Formelle de Concepts qui se concentre sur l'analyse des concepts à partir d'un contexte, où le contexte est initialement défini comme la relation binaire entre objets et attributs puis étendu à d'autres descriptions. On définit ainsi un concept comme un sous-ensemble d'objets qui partagent le même sous-ensemble d'attributs. Ces sous-groupes sont organisés en hiérarchie pour représenter notre jeu de données. Nous avons présenté les théorèmes et les fondements de la théorie des treillis, en particulier la notion de table binaire d'un treillis qui y est fondamentale. En utilisant les inf-irreductibles et les sup-irreductibles à partir desquels le treillis peut se reconstruire par application des opérateurs de borne sup et inf, nous proposons dans la contribution 4 de ce manuscrit un algorithme, "ReducedContextCompletion", qui permet à l'utilisateur d'injecter manuellement des concepts et de générer progressivement un treillis en utilisant exclusivement le contexte réduit de ce dernier. Ainsi, nous fournissons un outil d'analyse personnalisable pour l'utilisateur.

Dans ce chapitre, nous avons également présenté NextPriorityConcept qui génère des concepts composés d'un sous-ensemble d'objets avec leur description commune correspondant à un motif. Nous avons décrit son fonctionnement et avons étudié comment il peut combiner différentes descriptions pour traiter des données hétérogènes. À ces descriptions sont associées des stratégies naïves qui permettent de générer tous les concepts possibles, mais aussi d'autres stratégies permettant d'en réduire le nombre et ainsi limiter le déluge de motifs. Cela nous permet d'analyser et de traiter des données complexes telles que les trajectoires sémantiques qui ont été présentées dans le chapitre précédent. Nous mettons en application les connaissances et les travaux abordés dans ce chapitre dans la contribution 3 de ce manuscrit. Nous utilisons un ensemble de stratégies et de descriptions pour extraire de

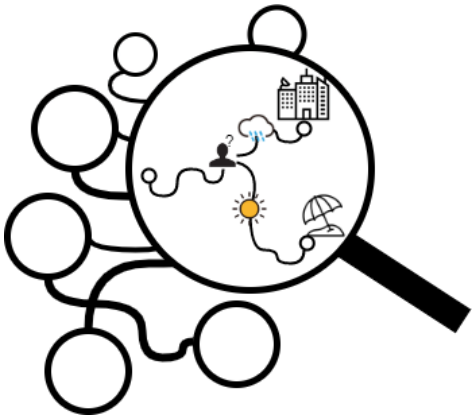
la base de données de trajectoires sémantiques des comportements similaires sur plusieurs aspects différents tout en limitant la surabondance de motifs grâce à des stratégies non-naïves.

Deuxième partie

Contributions

(a) Reconstruction de trajectoires en environnement indoor contraint

(b) Modèle multi-aspects générique d'enrichissement sémantique



(c) Analyse multi-séquences et hétérogène

(d) Optimiser et favoriser l'exploration des résultats d'analyses fondées sur l'AFC : "ReducedContextCompletion"

FIGURE 3.10 – Schéma récapitulatif des contributions et agencement des chapitres

Chapitre 4

Présentation des jeux de données

Table des matières

4.1 Introduction 75
4.2 Museum d'histoire naturelle de La Rochelle 75
4.2.1 Notations : 75
4.2.2 Le contexte des expérimentations ***Museum_ 1*** et ***Museum_ 2*** 77
4.2.3 Dataset ***Museum_ 1*** - Expérimentation mai 2019 : . . 78
4.2.4 Dataset ***Museum_ 2*** - Expérimentation janvier 2021 : 79
4.2.5 Dataset ***Museum_ 3*** - Museum d'histoire naturelle : septembre 2021 81
4.2.6 Le système de micro-localisation 81
L'application "Visite Musée" 83
4.3 Dataset : ***Geoluciole*** - Collecte de trajectoires touristiques . . 85
4.4 Dataset : ***La Cité du Vin*** 86
4.5 Synthèse de la problématique pour chaque jeu de données . . . 90

4.1 Introduction

Dans le cadre de cette thèse, nous avons utilisé un certain nombre de datasets que nous allons expliciter dans ce chapitre. Ceux-ci ont été constitués par notre équipe dans la plupart des cas, suivant une méthodologie particulière. Chacun des datasets comporte des trajectoires d'individus, en intérieur ou en extérieur, avec une précision sur les déplacements variable suivant la méthode de suivi utilisée.

4.2 Museum d'histoire naturelle de La Rochelle

4.2.1 Notations :

Pour le reste de ce manuscrit, nous considérons les notations suivantes pour les expérimentations de positionnement en intérieur :

Émetteurs E est l'ensemble des émetteurs, pour chaque élément $e \in E$ représente l'emplacement de l'émetteur avec $e = (x, y)$ lorsqu'il s'agit d'un émetteur fixe.

Récepteur Nous notons R l'ensemble des récepteurs, avec $r \in R$ et $r = (x, y)$

Signal $RSSI_i$ est le signal brut RSSI capté à un instant t_i envoyé pour un émetteur $e \in E$ et reçu par un récepteur $r \in R$.

Visiteurs V est l'ensemble de tous les visiteurs avec $v \in V$ un visiteur.

Selon le système de localisation en intérieur utilisé, nous disposons soit des émetteurs soit des récepteurs pour localiser un visiteur que nous appelons des points fixes. La méthodologie reste la même, ce qui nous intéresse, ce sont leurs coordonnées (x, y) lorsqu'elles ne changent pas :

— Dans le cas où les émetteurs sont fixes et les récepteurs mobiles, les visiteurs correspondent aux récepteurs ($V = R$) et les traces obtenues sont :

$$T = \{v = r, \{e = (x, y), \langle t_i, RSSI_i \rangle_i\}_{0 \leq i} : e \in E, v \in V\}$$

— Dans le cas où les récepteurs sont fixes et les émetteurs mobiles, les visiteurs correspondent aux émetteurs ($V = E$) et les traces obtenues sont :

$$T = \{v = e, \{r = (x, y), \langle t_i, RSSI_i \rangle_i\}_{0 \leq i} : r \in R, v \in V\}$$

Exemple 4.2.1. *Prenons le cas d'une expérimentation dans laquelle nous utilisons des émetteurs e pour localiser un visiteur dans le musée. Pour chaque visiteur v, nous leur attribuons un récepteur r qui les accompagnera pendant leur déambulation. Le tableau 4.1 représente la structure du jeu de données que nous recueillons à la fin de cette expérimentation.*

Visiteur	**Émetteur**		
	$e_1 = (18, 52)$	$e_2 = (15, 45)$	$e_3 = (10, 23)$
$Visiteur_1$	1686564345 : -82 1686564373 : -75 1686564423 : -62	1686564352 : -72 1686564384 : -81 1686564436 : -89	1686564367 : -92 1686564399 : -105 1686564447 : -110
$Visiteur_2$	1686564362 : -92 1686564374 : -99 1686564422 : -105	1686564365 : -80 1686564381 : -89 1686564423 : -94	1686564367 : -85 1686564380 : -76 1686564421 : -65

TABLEAU 4.1 – Exemple d'un dataset d'une expérimentation au muséum

4.2.2 Le contexte des expérimentations *Museum_1* et *Museum_2*

Deux expérimentations ont été menées au muséum d'histoire naturelle de La Rochelle en mai 2019 et janvier 2021, celles-ci sont comparables en plusieurs points. D'abord, toutes deux utilisent le même système d'indoor positionning, construit à partir d'une technologie bluetooth low energy (BLE). Des raspberries pi configurées pour réceptionner le signal BLE émis par des émetteurs sous forme de badges à porter autour du cou ont été utilisés. Ce système se devait d'être le plus simple possible à mettre en place, que ce soit du côté de l'installation (rapidement opérationnelle) que du côté des visiteurs (afin de limiter leur investissement). La figure 4.1 représente notre architecture de collecte des déplacements en intérieur, utilisée lors de deux expérimentations.

Les raspeberries étaient contrôlées par un ordinateur (master node sur la figure 4.1), communiquant en permanence avec elles, sur un modèle d'architecure dit "master / slave". Quand la séance de collecte est lancée, toutes les raspberries renvoient directement au master node les trames bluetooth captées. Ces trames sont constituées du timestamp de la collecte, du UUID du badge (Universal Unique IDentifier), de l'identifiant de la raspberry ainsi que du RSSI (Recived Signal Strength Indication), indiquant la puissance du signal reçu. Les données sont directement insérées dans une base de données no-sql ElasticSearch. Ces informations se formalisent sous la forme d'un ensemble $T = \{v = e\{r = (x, y), \langle t_i, RSSI \rangle\}\}$ pour chaque visiteur v et chaque récepteur BLE r. À partir de ces informations il est possible de position-

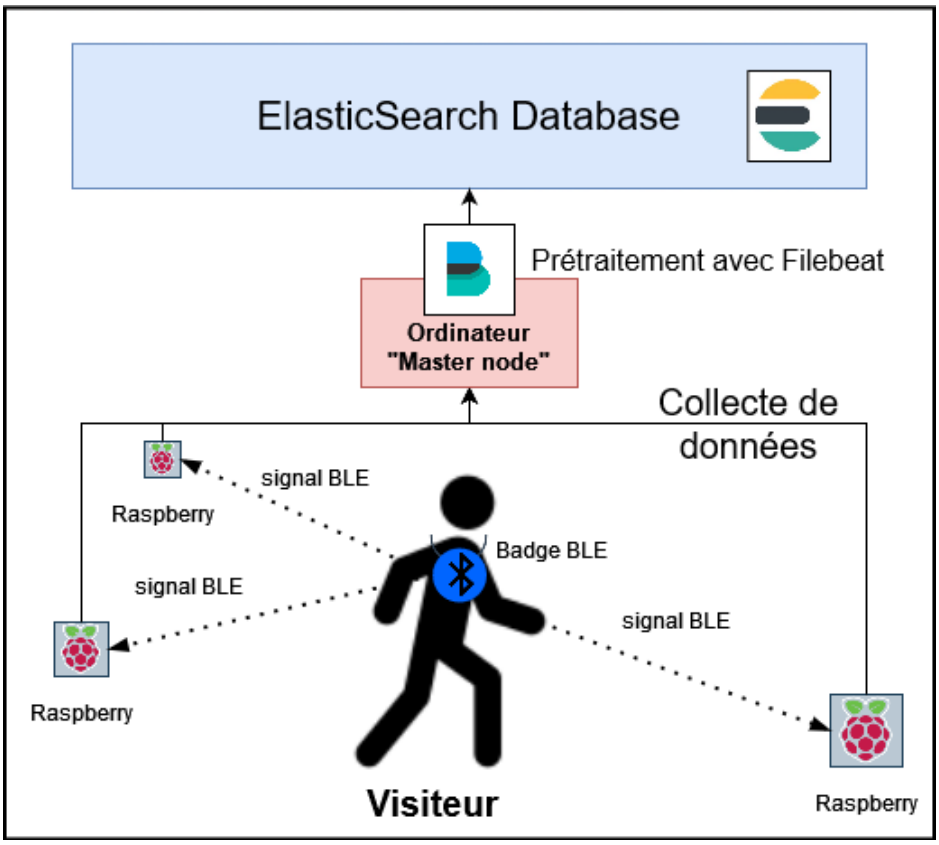


FIGURE 4.1 – Architecture technique de collecte de déplacements

ner l'utilisateur au sein d'un espace sous forme d'une trajectoire $\langle t_i, (x_i, y_i)\rangle$ pour chaque visiteur v par reconstruction de sa trajectoire. L'extrait de fichier JSON listing 1 montre un exemple de trame qui a pu être récupérée par ce système d'indoor positionning

4.2.3 Dataset *Museum_1* - Expérimentation mai 2019 :

La première expérimentation d'indoor positionning que nous avons réalisée s'est tenue durant un événement appelé "la Nuit des Musées" au sein du muséum d'histoire naturelle de La Rochelle, au cours duquel nous avons tenu un atelier de vulgarisation scientifique. Après avoir préalablement installé le système décrit ci-avant, nous proposions aux visiteurs de prendre part à l'atelier (figure 4.2). Celui-ci consistait à confier un badge émettant un signal bluetooth BLE à des visiteurs volontaires. À la fin de leur visite, nous pouvions alors discuter avec eux des traces de positions capturées durant la visite, tout en vulgarisant cette technologie émergente (figure 4.1).

```
{
    ''timestamp'' : 2021-01-29T09:07:15.927Z
    ''badge_number'': 8,
    ''beacon_id'' : rpi15.local,
    ''rssi'' : -68
}
```

Listing 1 – Exemple de trame récupérée par notre système d'indoor positionning

Au total 13 parcours de visiteurs ont été collectés sur une durée de 4 heures. L'objectif de cette expérimentation était de tester la précision maximale que nous pouvions atteindre avec ce genre de système. Ainsi, un seul étage du muséum (le plus grand) a été équipé. L'emplacement de chaque raspberry est décrit figure 4.3.

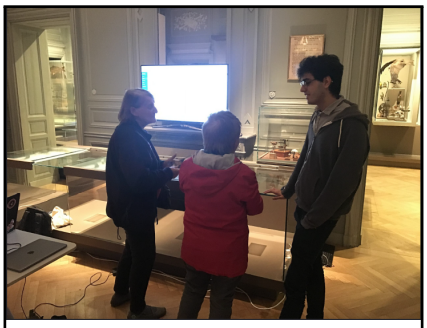


FIGURE 4.2 – Entretien avec un visiteur durant la nuit des musées

4.2.4 Dataset *Museum_2* - Expérimentation janvier 2021 :

La deuxième expérimentation d'indoor positionning s'est déroulée avec une classe de terminale d'un lycée rochelais pendant le confinement. Le but de cette deuxième expérimentation était, comme la première, de récupérer un dataset permettant de pousser la précision d'une localisation en intérieur au maximum avec ce type d'équipement. Ainsi, seule la plus grande salle a été équipée afin de capturer les traces des visiteurs déambulant à l'intérieur de celle-ci.

La figure 4.3 détaille le placement des raspberries au sein du musée qui est similaire au placement des raspberries au point de collecte dans l'expérimentation 1.

Durant cette expérimentation, l'idée était d'offrir une représentation des déplacements des participants visitant le musée librement en temps réel. Les détails ainsi que les visuels de cette expérimentation sont disponibles dans le chapitre "Contribution 2". Au total 3 trajectoires ont été récupérées en totalité durant cette expérimentation.

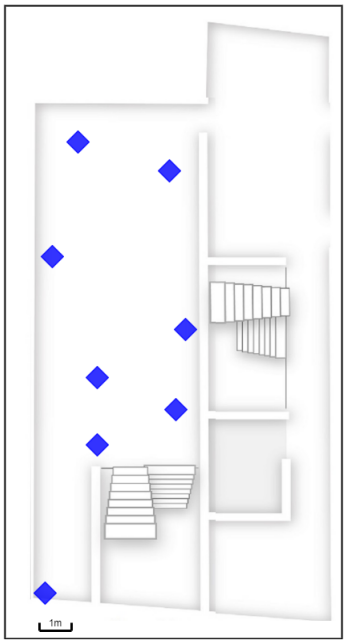


FIGURE 4.3 – Placement des raspberries au sein du muséum durant les expérimentations de janvier 2019 et 2021

Les traces extraites de cette manière sont de la même forme que celles de **Museum_1** :

$$T_{M1} = T_{M2} = \{v, \{r = (x, y), \langle t_i, RSSI_i \rangle\}\} \tag{4.1}$$

TABLEAU 4.2 – Description de ***Museum_2***.

Nb capteurs	**Nb instances**	**Placement**	**Nb individus**
8	82,630	Désorganisé	3

Le tableau 4.2 présente le nombre de capteurs utilisés, le nombre total d'instances de capture de signal, le type de placement et le nombre de sujets. Le tableau 4.3 présente une description de ces trois sujets de notre expérimentation. Nous devons noter que le nombre d'instances n'est pas un indicateur du temps passé dans les salles étudiées, comme on le voit avec l'individu 3 où sur 24,414 instances captées, nous n'avons obtenu que 3 min de données exploitables (3 capteurs envoyant une valeur RSSI supérieure à un seuil de signal de $-90dB$ dans la même fenêtre temporelle).

TABLEAU 4.3 – Description des individus de ***Museum_2***.

Numéro du badge	**Nb d'instances**	**Durée de passage**
5	26,701	16 min
8	31,403	22 min
3	24,414	3 min

4.2.5 Dataset *Museum_3* - Museum d'histoire naturelle : septembre 2021

4.2.6 Le système de micro-localisation

Dans cette expérimentation, le processus de collecte a été inversé. Cette fois-ci, les points de collecte Bluetooth deviennent des balises émettrices d'un signal Bluetooth qui sera récupéré par une application mobile : "Visite musée".

Le système d'indoor positionning (micro-localisation) utilise des balises qui sont émettrices d'un signal bluetooth BLE qui sera ensuite capté par l'application visite musée. Ces données sont ensuite envoyées sur une base de données "Parse Server" (Technologie No-SQL de stockage de données) sur un serveur localisé à l'Université de La Rochelle. L'application enregistre les actions de l'utilisateur par le biais des changements de vues dans l'application, et sont envoyées de la même façon dans cette base de données. La figure 4.4 montre une représentation du fonctionnement de l'application "Visite musée" pour la micro-localisation.

Nous avons cette fois-ci fait le choix d'équiper l'intégralité du musée en plaçant au moins un émetteur e dans chaque salle afin de récupérer la trajectoire globale de la visite pour chaque visiteur. Les balises ont été placées de telle sorte que les déplacements des visiteurs puissent être suivis durant toute sa déambulation.

Cette expérimentation a été réalisée sur une journée entière, durant laquelle nous avons pu récupérer 33 parcours de visites complètes, ainsi que les logs d'activités sur l'application "Visite musée".

Nous formalisons les données de déplacements récoltées lors de cette expérimentations pour le jeu de données **Museum_3** pour un ensemble de traces pour chaque émetteur $e = (x, y)$ et chaque visiteur v :

$$T_{M3} = \{v, \{e = (x, y), \langle t_i, RSSI_i \rangle\}\}$$

Les deux précédentes expérimentations d'indoor positionning avaient pour

FIGURE 4.4 – Processus d'envoi des données pendant l'expérimentation

objectif de collecter un dataset de logs de déplacements de visiteurs, sur un espace restreint, avec suffisamment de points de collecte afin de déterminer une position précise à l'intérieur de cet espace. Dans cette expérimentation, plus longue, plus ambitieuse, l'objectif était double :

— Récupérer toute la déambulation d'un visiteur au sein du muséum, du début de sa visite jusqu'à la fin.

— Capturer le comportement d'un visiteur sur une application mobile, "Visite Musée".

Afin d'atteindre ces objectifs, le placement des balises dans le muséum fut une étape primordiale et essentielle de l'expérimentation. La figure 4.5, montre le placement des balises au sein du musée. Ce placement a été pensé afin que nous ne perdions pas le visiteur au cours de sa déambulation à l'intérieur de cet espace. Chaque salle du musée a été équipée d'un émetteur, afin de savoir quand un visiteur entre et sort d'une pièce. Les escaliers, point clé dans le déplacement d'un visiteur ont eux aussi été balisés afin de prendre connaissance du changement d'étage à l'intérieur du muséum.

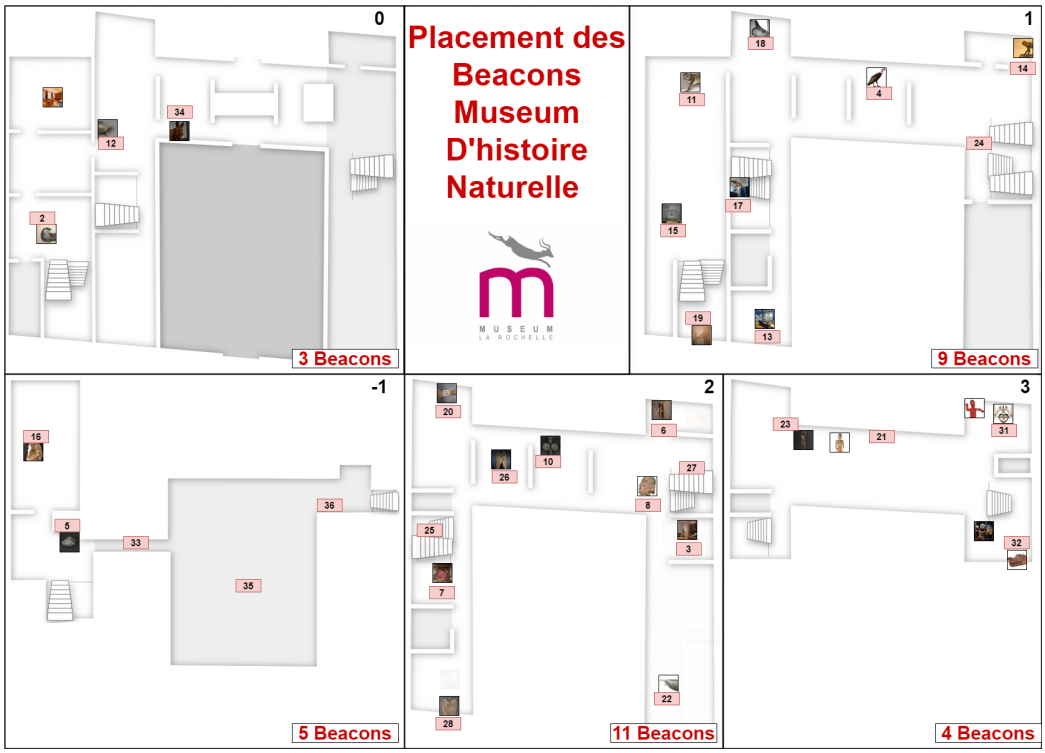


FIGURE 4.5 – Plan du placement des balises dans le muséum

L'application "Visite Musée"

Dans les salles, les balises ont été placées à proximité des œuvres présentes dans l'application mobile "Visite musée". Ce faisant, nous voulions savoir si une personne utilise le descriptif de l'œuvre dans l'application quand elle est à proximité de l'œuvre en question. Les œuvres détaillées sur l'application sont présentes sur la figure 4.5 directement sur le plan.

L'application "Visite Musée" est une application qui a été développée par la licence professionnelle - Parcours Développement Mobile - spécialisée dans le développement d'applications mobiles de l'IUT de La Rochelle. Celle-ci se veut agir comme un compagnon de visite, accompagnant le visiteur lors de son parcours du musée. Elle est actuellement utilisée dans bon nombre de musées de la région Nouvelle-Aquitaine dans le cadre du projet JPPasJMusée. Cette application se compose de trois vues principales, montrées figure 4.6, 4.8 et 4.7.

Vue accueil Il s'agit de la vue principale de l'application. Elle permet de choisir le parcours du musée et de naviguer vers les autres vues.

Vue plan Elle permet de localiser les œuvres à l'intérieur du musée.

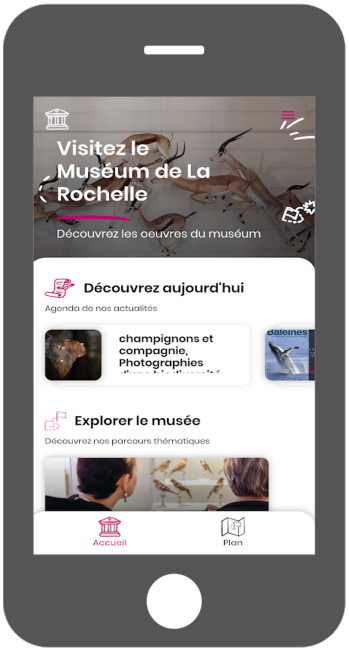


FIGURE 4.6 – Vue accueil

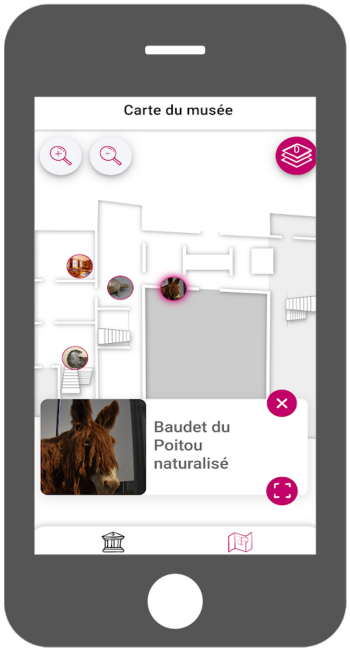


FIGURE 4.7 – Vue plan

FIGURE 4.8 – Vue détail

Vue détail Elle permet d'avoir plus d'informations sur une œuvre enregistrée dans l'application avec une description du contexte de l'œuvre, de son auteur et de son histoire.

L'application "Visite musée" utilise le récepteur à signaux bluetooth BLE du téléphone qui servira ici afin de localiser le parcours de l'individu à travers l'espace de visite en captant les signaux des émetteurs. Nous formalisons les données récoltées de vues de l'application pour un visiteur v par $\{v, \langle t_i, \text{ vue }_i\rangle_{i\leq n}\}$ où la "vue" est une information sémantique. Nous notons ainsi les données de ce dataset comme suit :

$$T_{M3} = \{v, \{e = (x, y), \langle t_i, RSSI_i\rangle\}, \langle t_k, \text{vue}_k\rangle\}$$

Le tableau 4.4 présente un extrait des données que nous avons recueillies lors de l'expérimentation "Museum_3".

Visiteur	**Émetteur**		**Vues**
	e_1	e_2	
$Visiteur_1$	09 :47 :40 : -72 09 :47 :41 : -81 09 :47 :42 : -89 ...	09 :47 :34 : -92 09 :47 :38 : -105 09 :47 :41 : -110 ...	09 :47 :29 : “Plan" 09 :47 :37 : “Accueil" 09 :47 :42 : “Baudet" ...

TABLEAU 4.4 – Extrait des données récoltées dans le dataset **Museum_3** pour un visiteur

4.3 Dataset : *Geoluciole* - Collecte de trajectoires touristiques

La problématique du projet DA3T étant l'étude des traces touristiques au sein de la Nouvelle-Aquitaine, il était ainsi nécessaire de construire un jeu de données de traces touristiques dites “outdoor”, dans une ville, correspondant à une pratique touristique. Pour cela, nous avons lancé une enquête durant l'été 2020, de juin à août, dans la ville de La Rochelle en Nouvelle-Aquitaine.

Afin de participer à notre enquête, les visiteurs installent sur leurs téléphones une application, développée dans le cadre du projet, permettant d'enregistrer leurs déplacements. Cette application nommée “Géoluciole” a été développée par des étudiants de Master informatique de l'université de La Rochelle, que nous avons supervisés, afin de correspondre aux besoins de l'étude (le logo de l'application est présenté figure 4.9). Un échantillon des traces récoltées est montré sur la figure 4.10.

FIGURE 4.9 – Logo de l'application Geoluciole

En complément de ses données de géolocalisation, le visiteur complète un questionnaire permettant de récupérer des informations sur son séjour. Ce questionnaire a été créé par Mélanie Mondo dans le cadre de sa thèse en géographie, travaillant au sein du projet DA3T et permet de récolter :

— Les dates de séjours
— Le type de séjour (familial, entre amis)
— Le moyen d'arrivée à La Rochelle (train, voiture, vélo)

Toutes ces informations seront utilisées pour analyser les comportements

des personnes dans nos contributions 2 et 3.

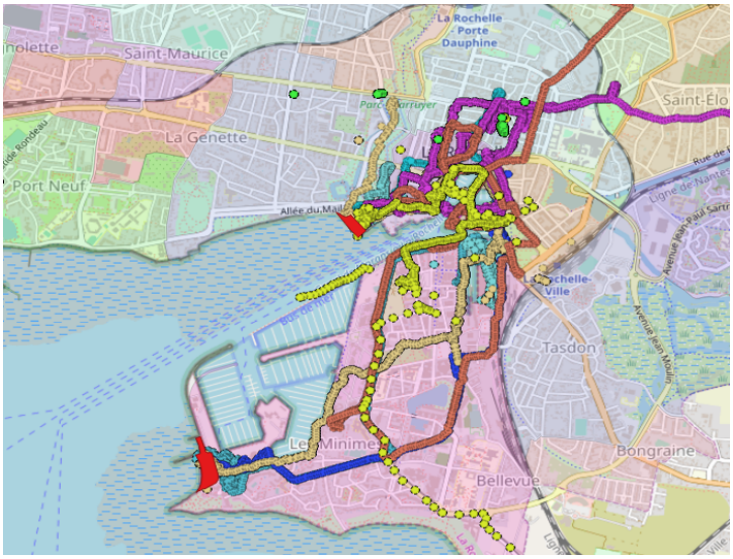


FIGURE 4.10 – Échantillon de traces GPS récoltées durant l'expérimentation

Dans un second temps, Mélanie a contacté les touristes pour leur proposer un entretien afin de commenter leurs traces de localisation en vue de collecter de la donnée qualitative. Ces données viennent enrichir les traces collectées.

Les données sont stockées sur une base MongoDB localisée sur un serveur de l'université de La Rochelle. Celles-ci ne sont accessibles que depuis le réseau de l'établissement et sont chiffrées afin de sécuriser au maximum les données personnelles. Durant cette expérimentation, nous avons récupéré 122 trajectoires touristiques. Parmi celles-ci, 12 ont été commentées et annotées par le biais d'un entretien.

Les données de déplacement d'un visiteur v sont au format GPS. Les coordonnées GPS permettent de récolter les traces suivantes pour un visiteur :

$$T_{GL} = \{v, D_v, \langle t_i, x_i, y_i \rangle : v \in V\}$$

où D_v représente les données supplémentaires et globales recueillies par le biais du questionnaire et de l'entretien.

4.4 Dataset : *La Cité du Vin*

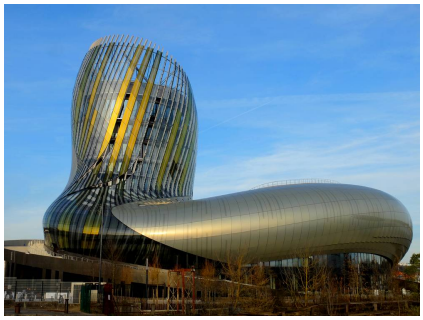


FIGURE 4.11 – La Cité du Vin, musée phare de Bordeaux

Ce jeu de données a été collecté au musée "La Cité du Vin" à Bordeaux, à partir des visites sur une période d'un an (mai 2016 à mai 2017), représentant près d'un million de visites. La Cité du Vin est un musée iconique de la ville de Bordeaux, et retrace l'histoire du vin à travers l'histoire et les cultures jusqu'à nos jours (cf. figure 4.11).

Le musée est un grand espace ouvert (*open-space*), où les visiteurs sont libres d'explorer le musée comme ils le souhaitent sans parcours prédéterminé. Lorsqu'ils arrivent au musée, ils reçoivent un petit dispositif personnel, un "*compagnon de visite*". Ce dispositif permet d'écouter des explications quand l'utilisateur passe devant les lieux d'animations. Le musée est composé de plusieurs espaces dits modules. Chaque module est composé de plusieurs sous-modules, activant pistes sonores, vidéos explicatives etc. Dans ce musée, il n'y a pas d'œuvre à proprement parler, tout passe par le compagnon de visite afin de pouvoir profiter de ce que cet espace a à offrir. La Figure 4.11 représente le plan du musée.

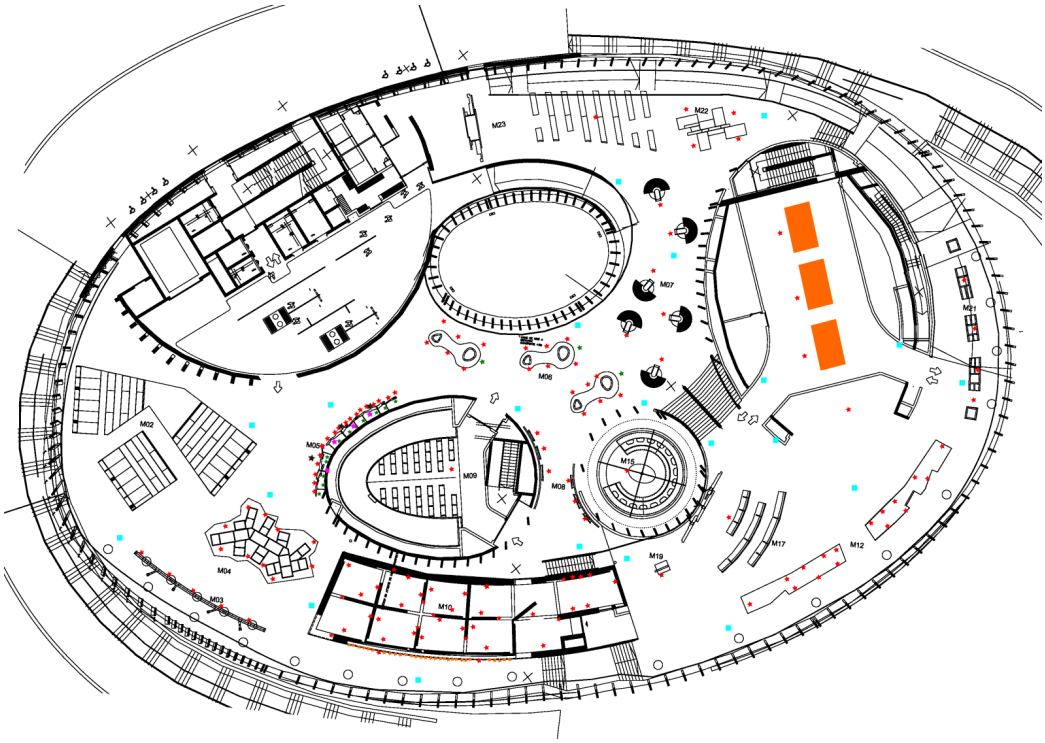


FIGURE 4.12 – Le plan de la Cité du Vin, Bordeaux

```
{
    ''ParcoursDuVisiteur_0'': [
        {
            ''visitor_id'': 0,
            ''timestamp'': ''2016-06-01T09:23:39Z'',
            ''module_nb'': ''M02'',
            ''sequence_nb'': ''M02.0 Totem'',
            ''language'': ''Francais'',
            ''age'': ''adulte'',
        }
        {
            ''visitor_id'': 0,
            ''timestamp'': ''2016-06-01T09:25:51Z'',
            ''module_nb'': ''M04'',
            ''sequence_nb'': ''M04.10'',
            ''language'': ''Francais'',
            ''age'': ''adulte'',
        }
}
```

Listing 2 – Exemple d'activation d'un module, dans le dataset de la Cité du Vin

En extrayant les séquences d'activation de chaque module, nous obtenons une trace de visite pour chaque visiteur du musée. Ce jeu de données, à l'opposé des trois autres ci-avant que nous avons construit, nous a été confié cette fois-ci par l'entreprise Berger-Levrault, partenaire du projet DA3T. L'extraction d'une partie de fichier JSON listing 2 montre l'activation d'un module par un visiteur de la Cité du Vin. Sur ces traces, nous récupérons le temps d'activation de l'activité ($sequence\_nb$) associée à son module plus général ($module\_nb$), sa langue ($language$) et sa tranche d'âge (age). Le plan de la Cité du Vin est montré figure 4.12

La formalisation du dataset de ***La Cité du Vin*** est notée pour un visiteur :

$$T_{CV} = \{v, D_v, \langle t_i, M_i \rangle_{0 \leq i}\}$$

avec v un visiteur, M_i un module activé à l'instant t_i, et $D_v = (\text{langue}, \text{ age})$ les données contextuelles du visiteur v. Le dataset ***La Cité du Vin*** comporte

100 000 trajectoires de visiteurs durant l'année 2016.

4.5 Synthèse de la problématique pour chaque jeu de données

Nous avons ainsi cinq datasets qui varient selon le lieu (indoor / outdoor), le type de capture, la taille, les possibilités d'enrichissement etc . La richesse de ces différents datasets nous permet d'aborder plusieurs problématiques propres aux trajectoires de visite. La formalisation de chaque jeu de données est décrite dans le tableau 4.5.

Dataset	Environement	Formalisation	Nb trajectoires
Museum_1	indoor	$T_{M1} = \{v, \{r = (x, y), \langle t_i, RSSI_i \rangle\}\}$	30
Museum_2	indoor	$T_{M2} = \{v, \{r = (x, y), \langle t_i, RSSI_i \rangle\}\}$	3
Museum_3	indoor	$T_{M3} = \{v, \{e = (x, y), \langle t_i, RSSI_i \rangle\}, \langle t_k, \text{vue}_k \rangle\}$	33
Geoluciole	outdoor	$T_{GL} = \{v, D, \langle t_i, x_i, y_i \rangle\}$	122
La Cité du Vin	indoor	$T_{CL} = \{v, D_v, \langle t_i, M_i = (x, y) \rangle\}$	100k

TABLEAU 4.5 – Formalisation des données pour chaque dataset

Une première problématique de reconstitution est étudiée dans la contribution 1. En se basant sur les datasets **Museum_1**, **Museum_2** et **Museum_3** qui sont des traces de capteurs des visiteurs, nous reconstruisons leurs trajectoires en intérieur. Ensuite, nous montrerons l'enrichissement de trajectoires spatialisées par le biais de données contextuelles, n'ayant pas de lien direct à la spatialisation d'un individu en utilisant les datasets **Museum_3** et Geoluciole dans la contribution 2. Enfin, nous utiliserons les datasets **Museum_3**, **Geoluciole** et de **La Cité du Vin** afin de les analyser pour en ressortir des comportements communs au sein de ceux-ci en prenant en compte le caractère spatialisé des trajectoires d'une part et de leurs données contextuelles associées d'autre part. Le tableau 4.6 recense les datasets avec leurs problématiques associées qui seront traitées dans les chapitres de contribution.

Dataset	Capture	Reconstitution	Enrichissement	Analyse
Museum_1	X	X		
Museum_2	X	X		
Museum_3	X	X	X	X
Geoluciole	X		X	X
La Cité du Vin			X	X

TABLEAU 4.6 – Tableau comparatif des dataset

Chapitre 5

Contribution 1 : Reconstruction de trajectoires en environnement indoor contraint

Table des matières

5.1 Introduction 93
5.2 Traitement des données indoor 95
5.2.1 Lissage et filtrage 96
5.3 L'algorithme GraphPositionning : pour une reconstruction de la trajectoire à gros grain 97
5.3.1 Le principe de l'algorithme 98
5.3.2 Visualisation des visites capturées 100
5.3.3 Conclusion 101
5.4 L'algorithme Minmal Zone Searching (MZS) : pour une reconstruction plus fine 102
5.4.1 Calcul des distances 102
5.4.2 Le principe de l'algorithme 103
5.4.3 Expérimentations 106
Jeux de données utilisés 106
Expérimentation sur **DS1** : 108
Expérimentation sur ***Museum_2*** : 109
Discussion : 111
5.4.4 Conclusion 112
5.5 Conclusion 113

5.1 Introduction

Les musées sont de parfait terrains d'expérimentation pour les technologies d'indoor positionning. En effet, les gestionnaires de ces lieux sont toujours à la recherche d'activités innovantes pour mettre en valeur leurs espaces, et il n'est pas difficile de trouver un public curieux prêt à accepter de participer à ce genre d'expérimentations. De plus, la nature de la visite - une marche souvent lente afin de pouvoir profiter des oeuvres - permet une meilleure capture des déplacements.

Cependant, la capture en musée pose un certain nombre de contraintes intrinsèques à ces lieux, que nous devons prendre en compte. Étant donné que

l'architecture des musées peut varier considérablement, il est peu probable que l'entraînement d'un modèle et l'étiquetage des données pour une classification supervisée puissent être universels pour tous les sites. Enfin, nous sommes limités et contraints dans le nombre et le positionnement d'équipements que nous pouvons placer dans un environnement de type musée. En effet, notre équipement ne doit pas perturber le parcours des visiteurs et les capteurs ou émetteurs ne pourront ainsi être placés qu'à des endroits stratégiques, alors qu'idéalement ils devraient être positionnés au milieu du cheminement du visiteur. Ainsi, le placement de l'équipement doit être en accord avec une architecture qui peut parfois ne pas être propice à la pose d'un tel système. Dû à la contrainte de placement de notre équipement, l'efficacité des techniques de triangulation et de trilatération sera bien amoindrie. C'est pourquoi nous proposons de nouveaux algorithmes plus adaptés dans des lieux contraints afin de reconstruire la trajectoire d'un visiteur.

Dans ce chapitre, nous présentons deux algorithmes de reconstruction de trajectoires qui opèrent sans avoir recours à une vérité terrain, tout en tenant compte des contraintes liées au positionnement de notre équipement. L'objectif ici est de reconstruire la trajectoire de visite d'un individu en se basant sur les traces récoltées par le biais d'un système d'indoor positionning comme expliqué dans la description des datasets **Museum_1**, **Museum_2** et **Museum_3**. En contraste avec d'autres études d'indoor positionning, nous nous intéresserons ici à la récolte de la trajectoire et de l'activité pour chaque visiteur spécifiquement dans les musées et non au processus de localisation exacte à tout instant.

Dans ce chapitre, nous verrons en première section le pré-traitement nécessaire afin de traiter efficacement les données de localisations. En section 2, nous aborderons la reconstruction de la trajectoire d'un visiteur à "gros grain", à l'échelle de la salle avec l'algorithme GraphPositionning. Avec le système d'indoor positionning défini dans la description du dataset **Museum_3**, l'objectif a été de récupérer la trajectoire de visite d'un individu de l'entrée dans le muséum jusqu'à sa sortie. Dans un second temps en section 3, nous verrons une autre stratégie, une approche plus "précise" où la volonté première était de travailler sur une granularité fine du déplacement du visiteur dans une pièce. Pour ce faire, nous avons mis au point l'algorithme MZS, pour Minimal Zone Searching. Cet algorithme renvoi une "zone probable d'activité" délimitant l'espace minimal où le visiteur pourrait se situer.

5.2 Traitement des données indoor

Pour permettre la micro-localisation indoor, une série de pré-traitements des données doit être effectuée. La figure 5.1 illustre la séquence des traitements pouvant être réalisés en temps réel. Cette chaîne de traitement utilise un signal BLE pour un individu v, qui peut être émis par un émetteur e fixe, ou encore capté par un récepteur r fixe suivant si v est un émetteur ou un récepteur. Ainsi, la trace d'un visiteur est représentée sous la forme suivante :

$$T[v = r] = \{e = (x, y), \langle t_i, RSSI_i \rangle\} \text{ ou } T[v = e] = \{r = (x, y), \langle t_i, RSSI_i \rangle\}$$

Dans les deux cas, la trace d'un visiteur est de la forme

$$T[v] = \{(x, y), \langle t_i, RSSI_i \rangle\} \tag{5.1}$$

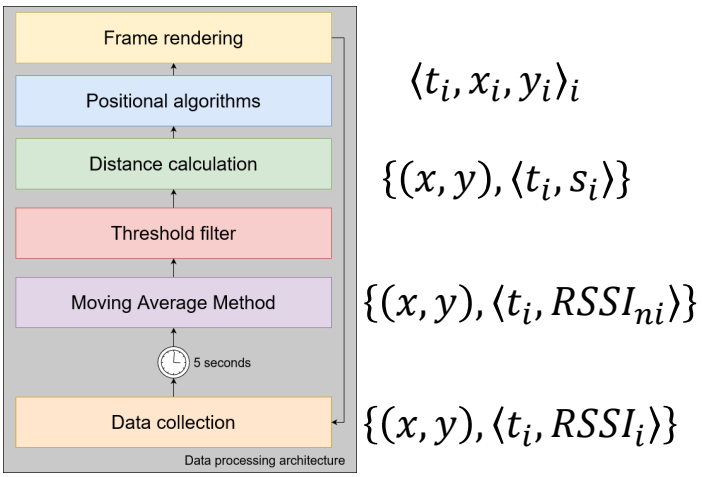


FIGURE 5.1 – Processus de pré-traitement de données indoor

La position de l'individu est calculée toutes les n secondes, où n est une valeur saisie par l'utilisateur. En utilisant un signal BLE $\langle t_i, RSSI_i \rangle$ pour notre système d'indoor positioning, nous sommes soumis à des imprécisions et des valeurs aberrantes, pouvant venir de signaux qui rebondissent ou traversant un corps humain. La fenêtre temporelle n permet de récolter suffisamment de données afin de lisser les valeurs en un signal $\langle t_i, RSSI_{ni} \rangle$ Il faut faire attention cependant à ne pas mettre une valeur de temps trop longue, sinon les résultats n'auront pas de sens puisque la personne peut se déplacer.

Nous avons déterminé une valeur de fenêtre temporelle de 5 secondes après avoir réalisé une analyse approfondie de nos exigences en matière de

collecte de données et de précision temporelle pour notre étude. Une fenêtre temporelle plus courte aurait entraîné une quantité insuffisante de points de données, rendant ainsi plus complexe le lissage du signal temporel. D'autre part, une fenêtre temporelle plus longue aurait pu compromettre la précision de la capture des déplacements. Par conséquent, la valeur de 5 secondes a été considérée comme un compromis adéquat, permettant d'obtenir à la fois une quantité de données raisonnable et une précision temporelle appropriée, conformément à nos besoins de recherche.

5.2.1 Lissage et filtrage

Il s'agit ensuite de lisser chaque signal $\langle t_i, RSSI_{ni} \rangle$ en un signal $\langle t_i, s_i \rangle$. La méthode de la moyenne glissante est une technique fréquemment utilisée en indoor positionning [Dong and Dargie, 2012] [Subedi et al., 2016] [Choi and Jang, 2017], harmonisant les valeurs prises en compte et les données de signal le rendant plus facile à manipuler.

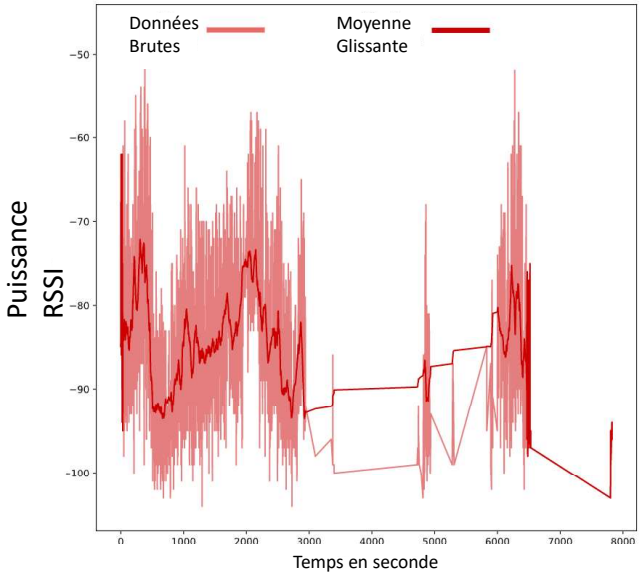


FIGURE 5.2 – L'impact de la moyenne glissante sur les données brutes récoltées

La figure 5.2 montre l'impact de la moyenne glissante sur les données

brutes récoltées par un seul récepteur. La forme du signal est erratique oscillant entre les valeurs extrêmes. La moyenne glissante lisse complètement le signal et limite les disparités dans les valeurs de signaux. On peut ainsi constater que le signal s'harmonise et réduit l'impact des valeurs aberrantes sur le signal final. Cette étape est suivie d'une phase de filtrage consistant à calculer la valeur moyenne de toutes les données RSSI après la sortie de la moyenne glissante, puis à décider si le récepteur est pertinent ou non. En effet, dans notre cas, le musée a plusieurs étages et il arrive que le signal BLE envoyé passe à travers le plafond. Dans ce travail, nous ne prendrons pas en compte une valeur RSSI inférieure à un seuil. C'est-à-dire que nous considérons que le visiteur n'est pas dans la pièce étudiée si la valeur est inférieure à la valeur du seuil. Le résultat de cette étape est une liste $\langle t_i, s_i \rangle$ de signaux capteurs / émetteurs (points fixes) pour chaque individu v. Cette liste est obtenue à partir des signaux $\{(x, y), \langle t_i, RSSI_{ni} \rangle\}$ où $s_i = max(RSSI_{ni})$ si $s_i >$ seuil. Après pré-traitement, on note pour chaque visiteur :

$$T[v] = \{(x, y), \langle t_i, s_i \rangle_{0 \leq i}\}$$

L'objectif est ensuite de construire une trajectoire pour chaque visiteur v à partir de ces signaux. Dans les sections suivantes, nous présentons deux algorithmes de reconstruction :

— GraphPositionning qui reconstruit une trajectoire :

$$\text{Traj}[v] = \langle t_i, (x_{v_i}, y_{v_i}) \rangle_{0 \leq i}$$

— MinimalZoneSearching (MZS) qui reconstruit une trajectoire de zones minimales de présence :

$$\text{Traj}[v] = \langle t_i, \text{zone}_i \rangle_{0 \leq i}$$

5.3 L'algorithme GraphPositionning : pour une reconstruction de la trajectoire à gros grain

Dans cette section, nous décrirons le fonctionnement de l'algorithme GraphPositionning prenant en compte les données remontées par les émetteurs que nous avons positionnés dans le muséum. Ensuite, nous avons associé la sortie des données de déplacement de l'algorithme GraphPositionning aux données issues de l'application mobile "Visite Musée" par le biais d'un modèle structurant ces données par utilisateur. Enfin, nous avons proposé

une visualisation de chaque visite, de déplacement et de l'utilisation de l'application sur le même plan par le biais d'infographies synthétisant les visites collectées.

5.3.1 Le principe de l'algorithme

Le muséum d'histoire naturelle de La Rochelle porte des contraintes de terrain auxquelles nous avons dû faire face. Mettre en place une expérimentation permettant de récupérer la trajectoire complète d'un individu n'a rien d'aisé. En effet, la principale contrainte vient du comportement des signaux BLE au sein de l'infrastructure même du bâtiment. Le muséum est un bâtiment historique de La Rochelle, datant du XVIIIe siècle et, ce faisant, ne possède pas d'isolation entre les étages : ceux-ci ne sont séparés que par un mince plancher de bois. Les signaux envoyés par nos émetteurs Bluetooth traversent ainsi aisément les étages. Cela se traduit directement dans la puissance des signaux que nous récoltons avec l'application. On constate alors que les émetteurs situés à côté d'un utilisateur - au même niveau - portent la même puissance de signal qu'un émetteur juste à l'étage au-dessus. Ceci signifie que dorénavant, nous ne pouvons plus uniquement nous fier aux signaux RSSI capturés, et que nous devons prendre en compte la topologie des lieux et les possibilités de déplacement afin de pouvoir reconstituer les déplacements des visiteurs. De façon plus générale, dans les expérimentations d'indoor positionning, la valeur retournée par le système n'est pas fiable à 100%, et une connaissance accrue des lieux est nécessaire afin de construire un tel système.

Pour cela, nous avons utilisé la notion de graphe des possibles décrivant sur la base de notre connaissance, l'agencement des salles et des chemins possibles que le visiteur peut emprunter. La figure 5.3 montre un graphe des possibles du muséum d'histoire naturelle de La Rochelle, c'est-à-dire l'enchaînement des salles autorisé durant la période de déplacement de l'individu. Chaque noeud du graphe représente un point de passage entre les pièces surveillé par un émetteur dans le muséum et les arcs représentent un lien de déplacement possible entre celles-ci. On note (x_0, y_0) le point de départ de la visite à l'entrée du musée à un instant t_0.

L'objectif de l'algorithme GraphPositionning est de construire la trajectoire de visite de l'utilisateur, en fonction des données que nous avons saisies pendant l'expérimentation. L'utilisateur ne peut se "déplacer" qu'au sein de ce graphe des possibles.

Le principe de l'algorithme GraphPositionning est le suivant. Il accepte en entrée une fenêtre temporelle w ; la trace $T[v]$ d'un visiteur v ; un

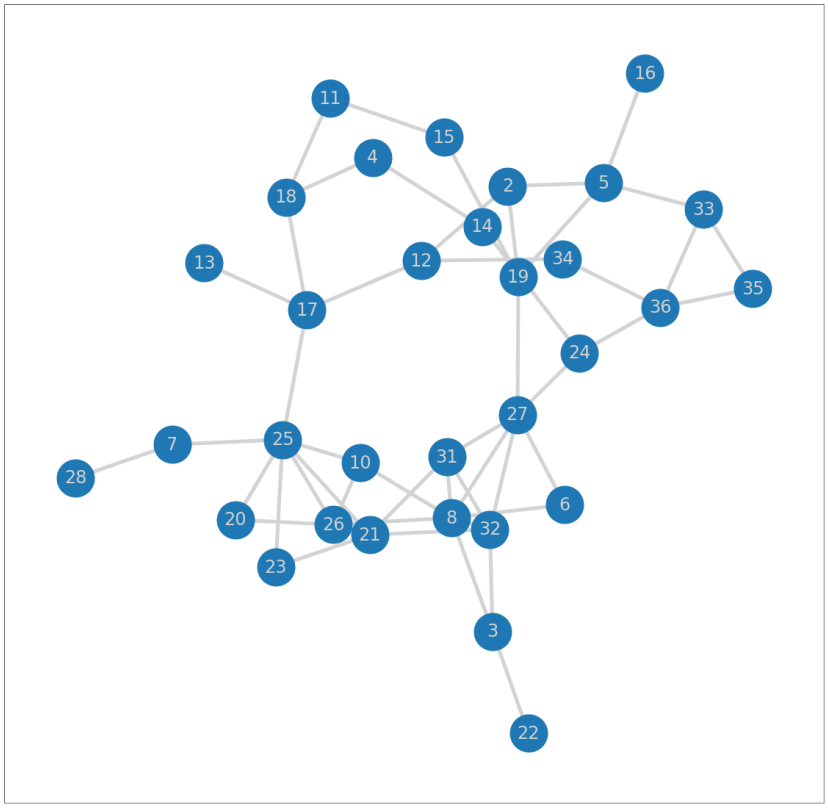


FIGURE 5.3 – Le graphe des possibles au sein du muséum où chaque noeud est un émetteur

graphe des possibles $G = (E, \text{possibles})$ ou $G = (R, \text{possibles})$ donné par la relation des possibles sur l'ensemble des points fixes (x, y) qui peuvent être des récepteurs comme des émetteurs. La trame de signal $T[v] = \{(x, y), \langle t_i, s_i \rangle\}$ est traitée tout en parcourant le graphe $T[v] = \{(x, y), s_i\}$ à partir du point d'entrée (x_0, y_0). À chaque étape t_i, l'ensemble des relevés des signaux $T[v, t_i] = \{(x, y), s_i\} \subset T[v] = \{(x, y), \langle t_i, s_i \rangle\}$ est considéré. Les points fixes voisins de l'état courant dans le graphe des possibles sont conservés. Nous comparons ensuite les données des voisins avec la position actuelle. Après cela, nous changeons l'état en fonction de la plus forte valeur de s_i parmi les points fixes conservés. Le processus recommence jusqu'à épuisement des données récupérées durant l'expérimentation. L'algorithme 1 décrit notre algorithme GRAPHPOSITIONNING.

L'efficacité de cet algorithme repose clairement sur le graphe des possibles qui doit prendre en compte le point fixe de départ (x_0, y_0) du musée,

Algorithm 1: GraphPositionning

Input:
- un visiteur v
- ses traces captées $T[v] = \{e = (x, y), \langle t_i, s_i \rangle \}$
- le graphe des possibles G
- le point d'entrée (x_0, y_0)
- une fenêtre temporelle w

Output: La trajectoire de v

```
courant = (x_0, y_0)
Traj[v] = ⟨t_0, x_0, y_0⟩
forall t_i; w; ∈ [t_0, t_n] do
    max = 0
    forall e = (x, y), s_i ∈ possibles(courant) do
        if s_i > max then
            suivant = e
            max = s_i
    Traj[v]+ = ⟨t, (x, y)⟩
    courant = suivant
return Traj[v]
```

mais aussi des points fixes aux différents points d'intersection des différents parcours de visite afin qu'il n'y ait pas d'angles morts. Une expérimentation plus poussée avec une vérité terrain serait pertinente avec une phase de tests et d'expérimentations afin de prouver la validité d'une telle approche. Nous nous sommes contentés ici d'une vérification par questionnement des visiteurs en raison des difficultés d'expérimentation liées aux périodes de confinement.

5.3.2 Visualisation des visites capturées

Avec l'algorithme GraphPositionning nous avons pu reconstruire les trajectoires des visiteurs ayant déambulé dans le musée sous la forme d'une séquence des salles visitées. Nous avons ainsi pu recréer le trajet d'un visiteur, comme le montrent les figures 5.4, 5.5, 5.6 et 5.7. Les visualisations que nous avons conçues lors de cette expérimentation étaient sous la forme d'une animation, ces figures en sont une capture. L'animation a permis de retranscrire tout le cheminement de la visite. Tous les noeuds du graphe ont été placés sur la carte suivant la position des émetteurs associés dans le musée. Durant l'animation de la trajectoire de l'individu, nous rendons visible les noeuds et les arcs empruntés. Ainsi, le graphe se découvre peu à peu en fonction des espaces qui ont été visités. La figure 5.7 montre par où le visiteur est passé

durant l'intégralité de sa visite.

FIGURE 5.4 – instantané de la trajectoire d'un visiteur (1)

FIGURE 5.5 – instantané de la trajectoire du visiteur (2)

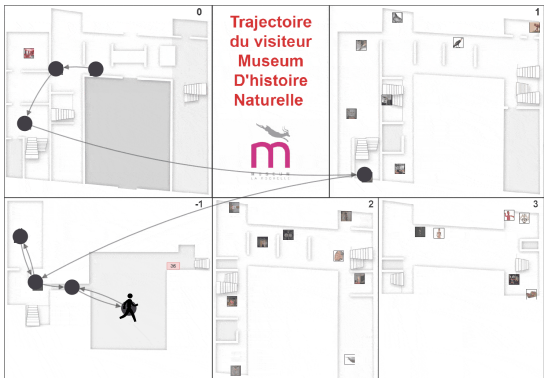


FIGURE 5.6 – instantané de la trajectoire du visiteur (3)

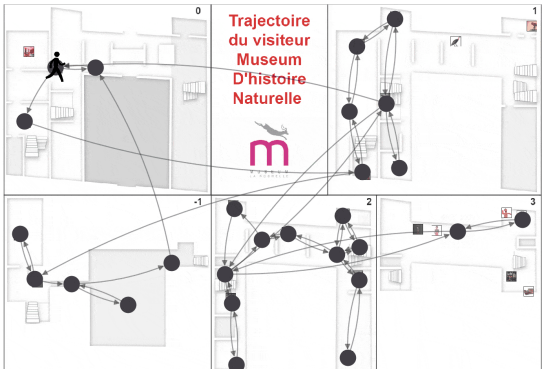


FIGURE 5.7 – instantané de la trajectoire du visiteur (4)

5.3.3 Conclusion

L'algorithme GRAPHPOSITIONNING, permet de transformer les signaux $\{(x, y), \langle t_i, s_i \rangle\}$ en une trajectoire $\langle t_i, x_i, y_i \rangle$ d'un individu, via la séquence des salles qu'il a traversées durant sa visite. En se basant sur la topologie du musée, nous avons montré qu'il était possible de reconstruire une trajectoire en équipant des points de passage dans le musée. Ceci est un exemple de reconstruction que nous appelons à "gros grain", la trajectoire finale n'étant ainsi pas précise et nous ne connaissons pas les déplacements de l'individu à l'intérieur de ces dites salles.

Dans la section suivante, nous expliciterons nos travaux sur une méthode cherchant à renforcer la précision afin de connaître en détails les déplacements

des visiteurs à l'intérieur des salles des musées.

5.4 L'algorithme Minmal Zone Searching (MZS) : pour une reconstruction plus fine

Dans cette section, nous décrirons les étapes mises en place pour traiter les données de position laissées par les visiteurs afin de retourner une estimation de sa position. De plus, cet algorithme peut être utilisé en temps réel. Ici, nous utilisons des points fixes récupérant un signal Bluetooth envoyé par des badges portés par les visiteurs. Le pré-traitement des données reste le même. Nous expliciterons notre méthodologie de calcul des distances et présenterons notre algorithme d'estimation de position, MZS [Richard et al., 2021], pour *Minimal Zone Searching*. Nous montrerons la particularité d'un tel algorithme, retournant la position d'un individu et ce même si le placement des points fixes n'est pas optimal. En contraste avec l'algorithme Graph-Positionning, l'algorithme MZS nécessite lui plusieurs points de collecte proches (plus de 3) afin d'affiner la position.

5.4.1 Calcul des distances

MZS considère en entrée des traces $T[v] = \{(x, y), \langle t_i, L_{s_i} \rangle\}$ obtenues à partir des traces $T[v] = \{(x, y), \langle t_i, s_i \rangle\}$ par un calcul de distances par la méthode du "path loss-model" décrite eq 2.1 du chapitre 2 de l'état de l'art où L_0 la puissance du signal à 1 mètre :

$$L_s = (-10n) log_{10}(s_i) + L_0$$

Nous avons utilisé cette méthode de recalibration commune à tous les algorithmes de positionnement. Lorsqu'un point calculé tel que l'intersection du cercle avec la triangulation ou un point Γ_s (i.e. la partie clé du MZS que nous expliquerons dans la section suivante) est situé en dehors de la pièce étudiée, nous le posons à la limite de celle-ci comme décrit dans la figure 5.8.

La ligne d'angle de la figure 5.8 relie un point de référence au point situé à l'extérieur de la pièce afin de déterminer l'angle de l'intersection avec la limite de la pièce. Ce point de référence est le deuxième point d'intersection dans l'algorithme de triangulation et le point Γ_s correspondant pour le capteur s dans l'algorithme proposé.

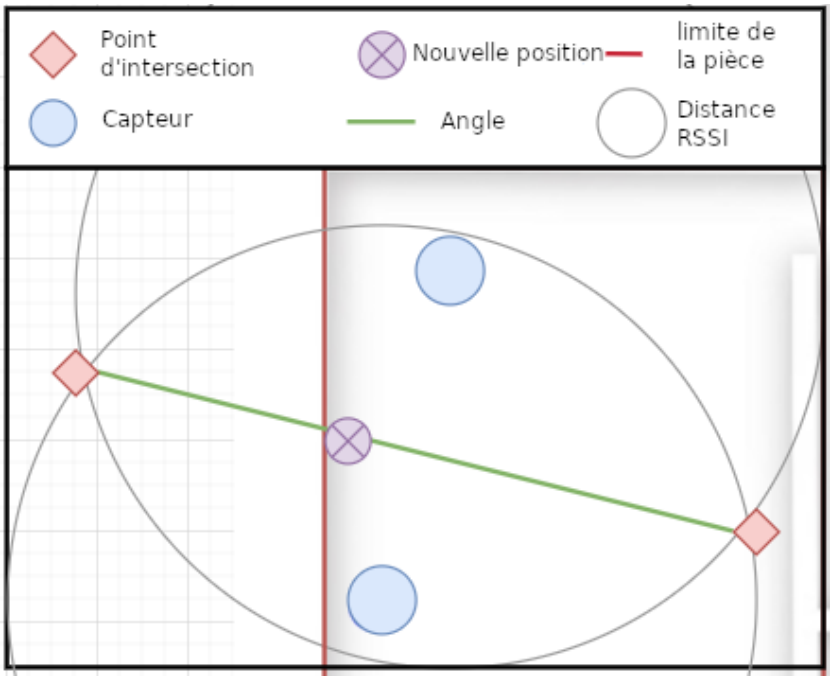


FIGURE 5.8 – Recalibrage du point calculé en fonction de la limite de la pièce

5.4.2 Le principe de l'algorithme

L'objectif de l'algorithme proposé est de trouver la plus petite zone possible où le visiteur est susceptible de se trouver à un certain moment.

La plupart des approches dans les systèmes de positionnement en intérieur, comme la trilatération ou la triangulation, ne considèrent que trois à quatre capteurs ou émetteurs qui ont enregistré la valeur maximale de RSSI. Avec notre approche, nous prenons en compte tous les points fixes de la pièce. L'algorithme considère l'ensemble des traces $T[v,t] = \{(x,y), L_s : L_s \neq 0\}$ captées/émises par un point fixe dans la pièce (i.e L_s non nul) pour un visiteur v et une unité de temps t. À chaque instant t, l'algorithme calcule une zone minimale zone_t comme la représentation des positions possibles de l'individu. En itérant l'algorithme MZS sur chaque unité de temps t_i, on peut donc reconstruire une trajectoire $Traj[v] = \langle t_i, \text{zone}_{t_i} \rangle$ à partir de l'ensemble des traces $T[v] = \{(x,y), \langle t_i, L_{s_i} \rangle\}$. Ainsi, Pour chaque unité de temps t, MZS calcule zone_t par affinage successifs de zones :

$$\text{zone}_t^0, \text{zone}_t^1, \ldots, \text{zone}_t^n = \text{zone}_t$$

Initialisation : Le traitement est initialisé avec la première zone $zone_0$

définie comme l'enveloppe convexe des points fixes (x, y) dans la pièce (Équation (5.2)) ayant une distance L_s non nulle, le badge étant nécessairement dans cette $zone_0$.

$$zone_0 = Hull(x, y : (x, y) \in T[v, t]) \tag{5.2}$$

Affinage des zones : L'algorithme itère jusqu'à ce que la zone optimale soit atteinte. À chaque itération k, une nouvelle zone $zone_t^k$ est calculée à partir du centroïde de la zone issue de l'itération précédente et des valeurs L_s de chaque point fixe $(x, y) \in T[v, t]$. Puis cette zone est affinée en prenant en compte la distance L_s et l'enveloppe convexe de l'ensemble des points de distance, où un point de distance Γ_s est calculé à partir de la valeur de L_s du point fixe $(x, y) \in T[v, t]$ et selon le centroïde $c = (x_c, y_c)$ de l'enveloppe convexe précédente (Équations (5.3) et (5.4)).

$$\Gamma_s = \begin{bmatrix} x_c + L_s cos(\text{atan2}((y_c - y), (x_c - x))) \\ \\ y_c + L_s sin(\text{atan2}((y_c - y), (x_c - x))) \end{bmatrix} \tag{5.3}$$

Dans ce contexte, la fonction atan2 est utilisée pour calculer l'angle entre les coordonnées du centroïde c de la zone précédente et les coordonnées d'un point fixe $(x, y) \in T[v, t]$. Cette fonction est également appelée "tangente inverse de deux arguments" et elle prend en compte la polarité de chaque argument pour déterminer le quadrant dans lequel se trouve le point correspondant sur le cercle trigonométrique. Dans ce cas, le cercle trigonométrique est déterminé par la puissance du signal L_s relevé sur le point fixe $(x, y) \in T[v, t]$.

A chaque itération k, on considère l'enveloppe convexe de tous les points Γ_s et le nouveau centroïde C_k. L'itération du calcul de la zone est basée sur l'équation suivante :

$$zone_t^k = Hull(\{\Gamma_s : s = (x, y) \in T[v, t]\}) \tag{5.4}$$

Une nouvelle zone est alors obtenue en calculant l'enveloppe convexe de ces points de distance. Le traitement est réitéré afin d'affiner cette nouvelle zone.

L'algorithme s'arrête lorsque la taille et la position de $zone_t^{k-1}$ et de $zone_t^k$ sont comparables, c'est-à-dire lorsque l'on obtient une faible distance entre leurs centroïdes consécutifs ou si l'on atteint le nombre maximal d'itérations. En utilisant le jeu de données de notre expérimentation et un seuil de 0.1m, nous avons observé que l'algorithme a besoin en moyenne de 6 itérations et de 0.2s de temps de calcul pour trouver un centre "stable", ce qui convient à une exécution en temps réel. La figure 5.9 montre une représentation visuelle

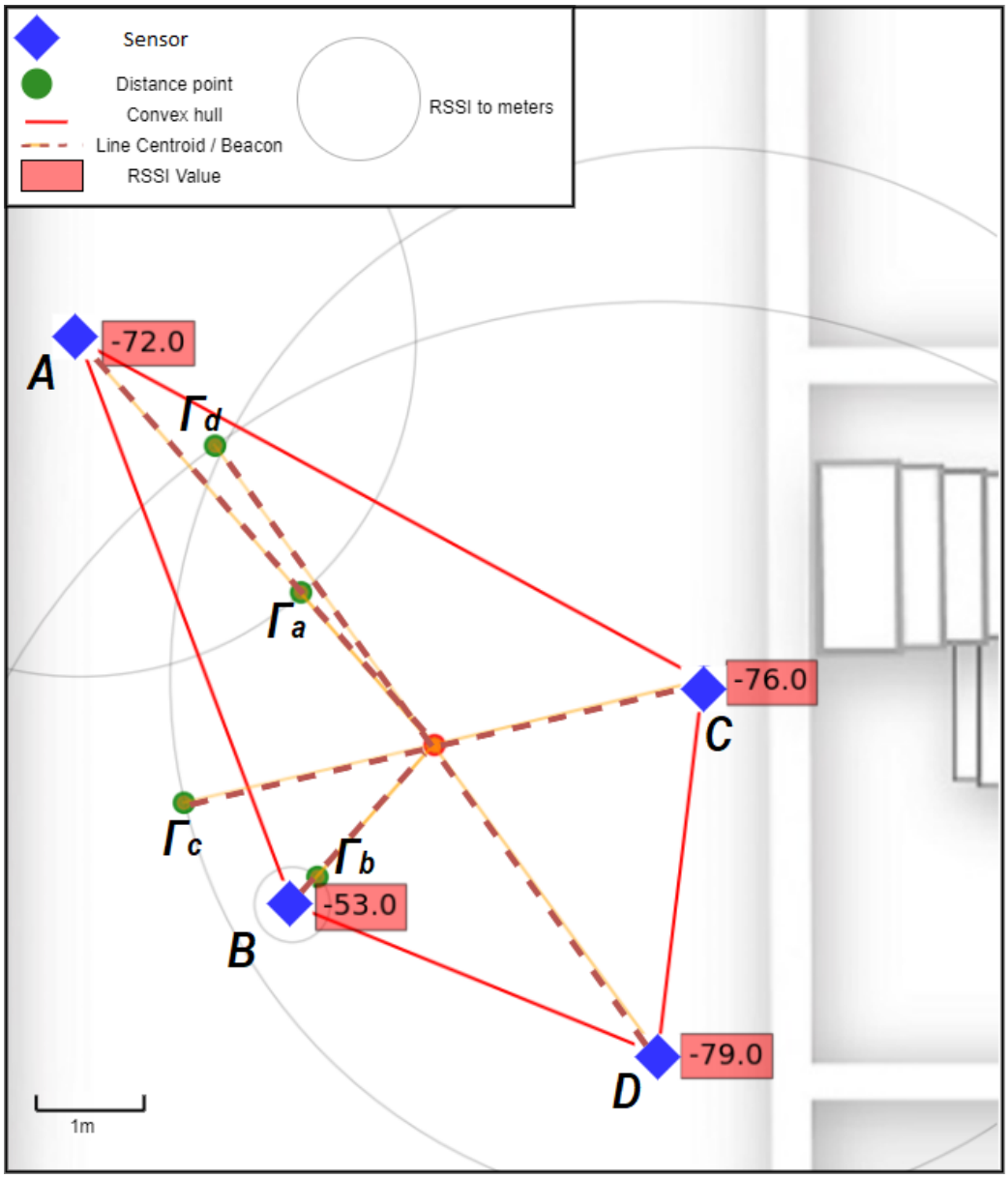


FIGURE 5.9 – Enveloppe convexe des points fixes actifs à n=0

Algorithm 2: MINIMAL ZONE SEARCHING

Input: - un visiteur $v \in V$
- un timestamp t
- les distances $T[v,t] = \{(x,y), L_s\}$ entre v et le signal des points fixes à l'instant t.
- un seuil de distance seuil_b
- un seuil du nombre d'itération seuil_iter

Output: Une zone minimale d'activité ; Zone

init
- Zone $\leftarrow Hull(\{(x,y) \in T[v,t]\})$;
- $centroid \leftarrow get\_centroid(Zone)$;
- $nb\_iter \leftarrow 0$;

do
- nb_iter++ ;
- $new\_centroid \leftarrow centroid$;
- $li\_DP \leftarrow []$;
- **forall** $((x,y), L_s) \in T[v,t]$ **do**
  - $\Gamma_s = DistancePoints(L_s, centroid)$;
  - $li\_DP.push(\Gamma_s)$;
- $Zone \leftarrow Hull(li\_DP)$;
- $centroid \leftarrow get\_centroid(Zone)$;

while $Distance(centroid, new\_centroid) > seuil\_d$ *and* $nb\_iter < seuil\_iter$;

return *Zone*

d'une étape d'exécution, où Γ_a, Γ_b, Γ_c et Γ_d sont respectivement les points de distance des points fixes A, B, C et D. La figure montre le pas entre l'itération $n = 0$ et $n = 1$ trouvé dans la Figure 5.10.

5.4.3 Expérimentations

Jeux de données utilisés

Pour ces expérimentations, nous avons utilisé le jeu de données **Museum_2** ainsi qu'un jeu de données **DS1** qui possède une vérité terrain, nous permettant ainsi de valider notre approche. Pour chacun de ces jeux de données, nous avons utilisé la même architecture de traitement des données. comme expliqué dans la Section 2.2.2.

Le jeu de données est un jeu de données en accès libre réalisé par *Mehdi*

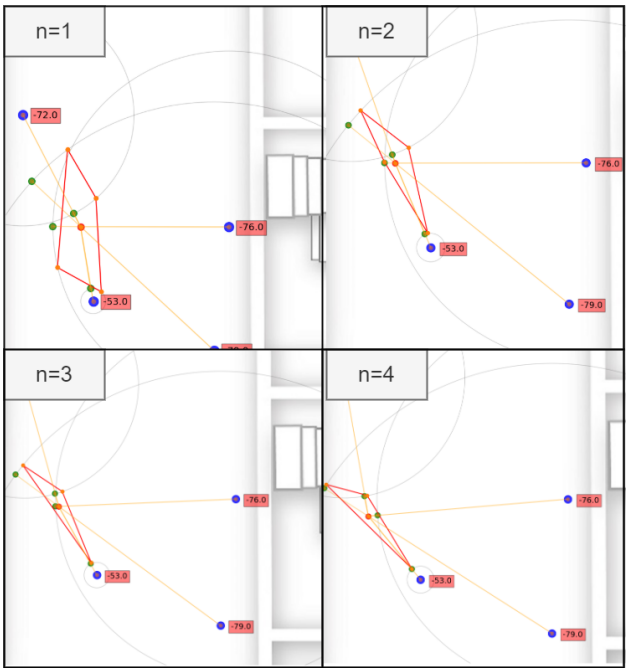


FIGURE 5.10 – Évolution de MZS à chaque n exécution

Mohammadi et al., 2017 [Mohammadi et al., 2017]. Ce jeu de données est composé des valeurs BLE RSSI émises par 13 iBeacons et de la position réelle de l'utilisateur. Comme ce jeu de données possède une vérité terrain, il nous permet de calculer la valeur de précision de notre méthode. Et ainsi valider notre approche.

DS1 a un placement "grille". Le tableau 5.1 est une description du jeu de données. Le nombre d'instances est le nombre de valeurs RSSI pour tous les points fixes.

TABLEAU 5.1 – Description de **DS1**.

Nb points fixes	**Nb instances**	**Placement**	**Durée**	**Frames**
13	1420	Grille	8m33	33

Le nombre limité d'instances de ce jeu de données nous oblige à utiliser une taille de fenêtre temporelle de 15s. En effet, pour valider notre approche, nous avons besoin d'au moins 3 points fixes retournant une valeur RSSI entre ces deux intervalles de temps. Cela nous laisse 8 minutes et 33 secondes de données à traiter, ce qui est limité, mais la vérité terrain est une donnée précieuse, surtout si elle est issue d'un site en libre accès.

Expérimentation sur DS1 :

Représentation visuelle de MZS et de la vérité terrain : La figure 5.11 montre une représentation visuelle de l'algorithme MZS sur une "petite portion" de **DS1**. Cette "petite portion" est composée de 15 images fusionnées (4 min), afin de mieux observer les résultats calculés par l'algorithme MZS. La trajectoire quant à elle est directement donnée par le créateur de ce dataset. Dans cette figure, nous retrouvons en bleu les points fixes, en vert la vérité terrain, en rouge les zones calculées par l'algorithme MZS. Les zones en sortie d'algorithme correspondent ainsi peu ou prou aux déplacements de l'individu.

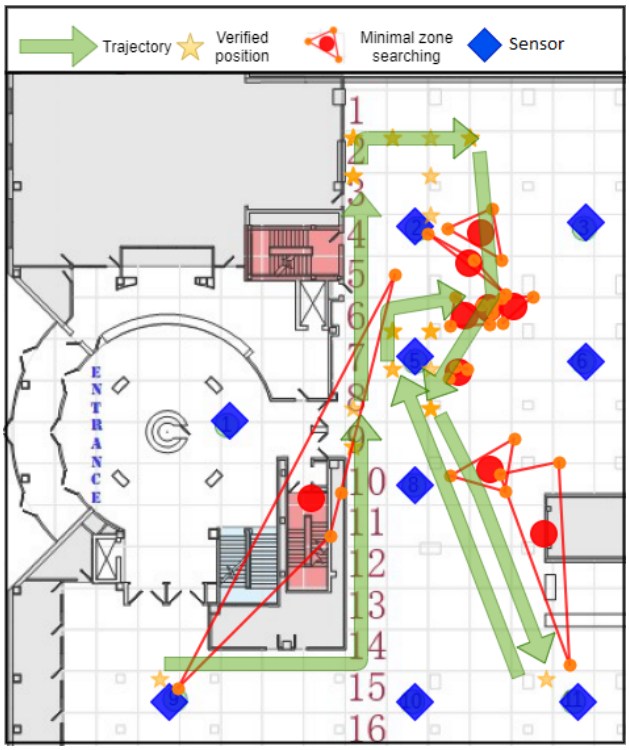


FIGURE 5.11 – L'algorithm MZS appliqué à **DS1**

Comparaison entre MZS et Triangulation : La vérité terrain de **DS1** nous permet de calculer un indicateur de précision. Le paramètre de précision est calculé en utilisant la distance euclidienne de la zone retournée avec tous les points de position vérifiés toutes les 15 s, car nous pouvons avoir plusieurs points de ce type par intervalle de temps. Si la zone inclue le point, la précision est fixée à 0.

Le tableau 5.2 montre le paramètre de précision calculé pour la triangulation et l'algorithme MZS, la moyenne et l'écart-type de la taille des zones calculées. La colonne "images calculées" du tableau 5.2 fait référence au nombre d'estimations de positions qui ont pu être calculées.

Ainsi, avec cette expérimentation, nous pouvons observer que l'algorithme MZS donne des résultats proches de la triangulation et ainsi validant notre approche. Néanmoins, la triangulation garde de meilleurs résultats que MZS. Ce résultat s'explique de par le positionnement en "grille" favorable à cette méthode.

TABLEAU 5.2 – Valeurs de précision entre l'algorithme MZS et la triangulation appliqué à **DS1**

Méthode	**Précision**	**Moyenne zone**	**Écart-type zone**	**Images calculées**
MSZ	0.74 m	0.49 m^2	0.85	25
Triangulation	0.53 m	1.21 m^2	0.95	19

Expérimentation sur *Museum_ 2* :

La seconde expérimentation avec le dataset, ***Museum_ 2*** décrit dans le chapitre 4, s'effectue en conditions réelles avec un placement devant s'adapter à une topologie particulière : celle d'un musée. Nous avons comparé la triangulation et notre algorithme MZS sur le jeu de données qui, rappelons-le, ne possède pas de vérité terrain.

Importance de la position des points fixes : La Figure 5.12 montre une situation où la méthode de triangulation ne donne pas le meilleur résultat. La valeur RSSI est affichée à côté du capteur et le cercle qui les entoure représente la valeur de L_s pour chaque point fixe. C'est un bon exemple de la façon dont le capteur avec la plus grande valeur de RSSI, (et donc indiquant qu'une personne se trouve à proximité) n'intersecte aucun cercle, ce qui rend la méthode de triangulation difficile. L'algorithme MZS a, comme il ne cherche pas les intersections entre les cercles, une meilleure précision dans les cas où la position des points fixes n'est pas optimale.

Comparaison des zones calculées : La Figure 5.13 montre la taille des zones de MZS avec trois métriques, la taille moyenne de la zone calculée, la médiane et l'écart-type (SD). Le Tableau 5.3 montre sur combien de trames

les deux méthodes auraient pu être calculées. Une trame est un intervalle de temps de 5s, où au moins 2 points fixes renvoient une valeur RSSI, donc une personne se trouve dans la salle équipée. Une image représente ainsi un retour d'algorithme. Pour le badge 5 par exemple, MZS retourne 181 zones, la triangulation retourne 135 zones sur un total de 194 exécutions des deux algorithmes. Cela signifie que parfois un algorithme réussi à renvoyer une position quand l'autre n'y arrive pas et inversement.

TABLEAU 5.3 – Nombres d'images calculés par MZS et la triangulation.

Badge	**MZS**	**Triangulation**	**Nombre d'images**
5	181	135	194
8	254	200	263
3	24	5	31

Dans le Tableau 5.3, nous voyons qu'avec le badge numéro 5, 194 images remplissent ces conditions, ce qui signifie que ce visiteur a passé approximativement au moins 16 min dans la pièce. Le badge numéro 8 obtient 263 images exploitables (22 min) et le badge numéro 3 en obtient 31, ce qui signifie que le visiteur est probablement passé rapidement dans cette pièce.

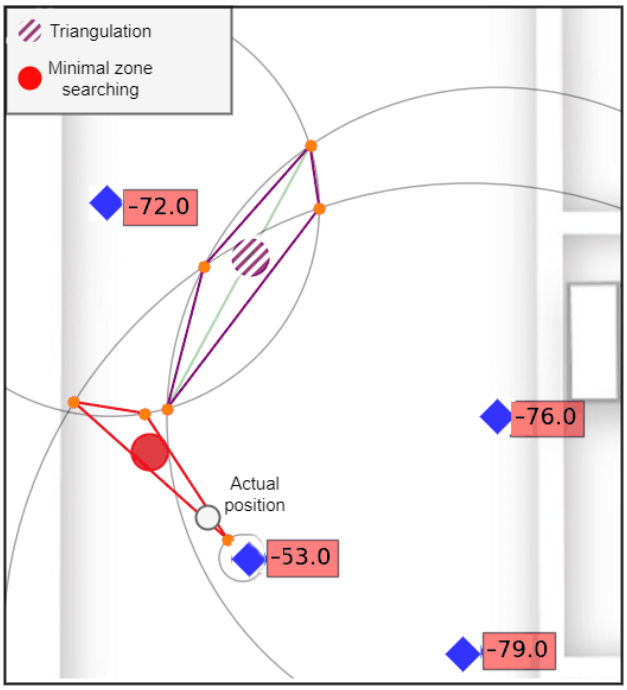


FIGURE 5.12 – Comparaison entre MZS et la triangulation

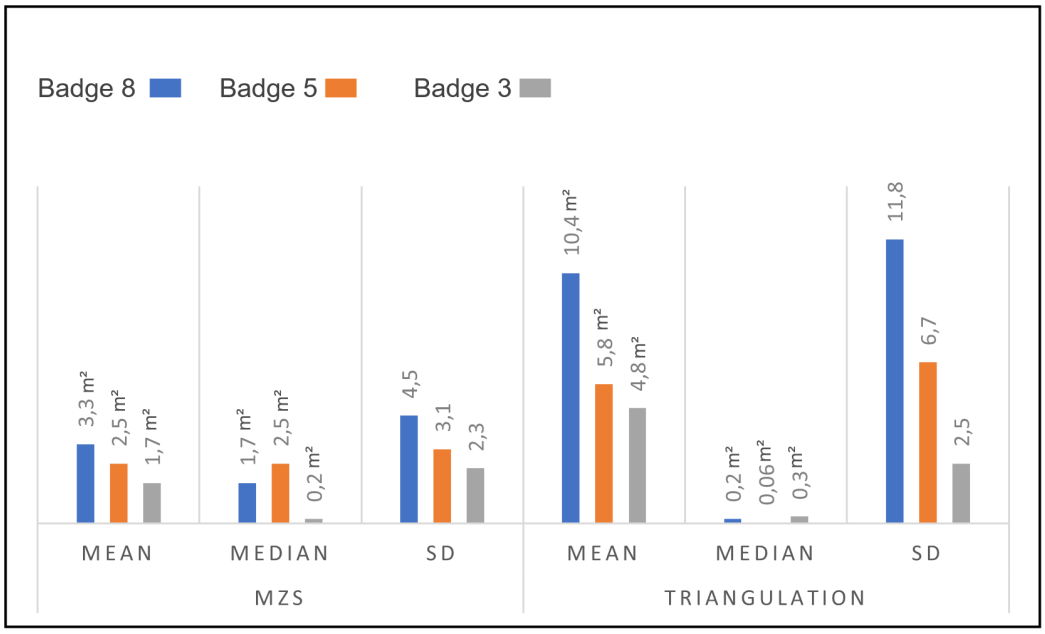


FIGURE 5.13 – Propriétés des zones calculées par les deux méthodes

Discussion :

En gardant ces informations à l'esprit, le tableau 5.3 montre que l'algorithme MZS est capable de retourner une zone de position dans la plupart des images observées, soit davantage que la technique de triangulation. De plus, MZS donne un résultat plus stable, avec une valeur de variation standard plus faible (cf. tableau 5.2). Cependant, la technique de triangulation renvoie une zone plus petite en comparaison lorsque nous prenons la médiane. La technique de triangulation vise certainement une position plus précise de l'utilisateur lorsque la cible est stationnaire et ne peut pas fonctionner correctement lorsque le sujet est en mouvement pendant les intervalles de cinq secondes. De plus, la position des points fixes dans la pièce n'est pas optimisée pour l'algorithme de triangulation et peut jouer un rôle important dans ces résultats car les situations décrites dans la Figure 5.12 conduisent à de mauvaises performances.

La figure 5.14 montre une représentation visuelle des surfaces calculées par les deux méthodes et tenues comme une illustration du Tableau 5.3 et des données du badge numéro 8. Toutes les aires calculées par ces deux algorithmes sont projetées ensemble sur la même image.

L'algorithme MZS montre sa force sur cette figure. La localisation du visiteur au centre de la pièce est similaire avec les deux méthodes, mais on

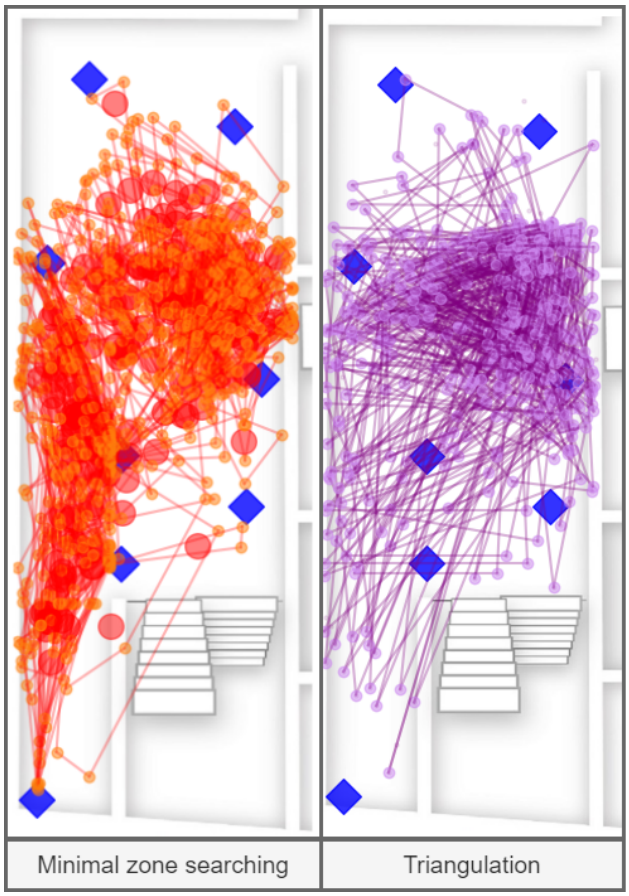


FIGURE 5.14 – Comparaison visuelle entre la triangulation et MZS , avec les données du badge numéro 8 sur ***Museeum_2***.

observe une différence vers le bas à gauche de la pièce, qui est un espace restreint. Nous constatons que la méthode de triangulation donne de mauvais résultats dans les espaces restreints, où seuls quelques points fixes sont accessibles. Et donc, les résultats retournés par l'algorithme MZS sont plus exploitables et donnent une vue plus significative de la visite du sujet.

5.4.4 Conclusion

Les contraintes architecturales des musées ont un impact dans les placements possibles des points fixes, comme nous l'avons vu et rend parfois les algorithmes tels que la triangulation moins efficaces. L'algorithme MZS, donne des résultats prometteurs car la position des points fixes semble avoir moins d'impact le rendant ainsi plus robuste aux espaces contraints et de petite taille. Ce type de technique de positionnement en intérieur est encore à

un stade précoce de développement et des améliorations peuvent être apportées. Le placement des points fixes a un impact sur les différentes techniques d'indoor positionning [Rezazadeh et al., 2018] [Sharma and Badarla, 2018] et l'algorithme MZS est une piste afin de remédier aux terrains difficiles.

5.5 Conclusion

Les systèmes d'indoor positionning appliqués aux musées offrent des moyens nouveaux et prometteurs d'étudier les comportements des visiteurs. Ils peuvent aider à la gestion de telles installations, en permettant de suivre le parcours des visiteurs et de détecter les pièces les plus visitées et inversement les moins visitées. En effet, les gestionnaires de musées sont toujours à la recherche de nouvelles façons d'améliorer l'expérience de visite afin de mieux comprendre l'exploration de ces espaces.

Dans ce chapitre nous avons pu décrire deux algorithmes de calcul de la position d'un utilisateur dans un contexte de visite dans un musée. Nous avons exposé les problématiques et avons répondu avec deux méthodes différentes.

D'une part, nous avons montré qu'il était possible d'approcher un positionnement précis de l'utilisateur avec un algorithme simple à mettre en place, possible en temps réel, et un système ne nécessitant pas de moyens importants. D'autre part, nous avons montré qu'il était aussi possible de travailler à gros grains, en plaçant les points fixes dans des endroits stratégiques du musée en utilisant l'architecture du terrain que nous allons utiliser - la topologie des lieux. Celui-ci donne des résultats satisfaisants pour nos besoins et permet de calculer la trajectoire de l'utilisateur.

Néanmoins, un problème persiste. En effet, la trajectoire de déplacement d'un visiteur ne suffit pas pour interpréter les choix de son comportement au sein d'un musée, ni pour en déduire un intérêt ou un désintérêt des oeuvres présentées.

Durant une expérimentation, nous nous sommes intéressés à l'activité des visiteurs sur l'application mobile "Visite musée" décrite dans le chapitre d'introduction aux contributions. En effet, nous souhaitions faire correspondre l'utilisation de cette application pour quantifier le degré d'intérêt des visiteurs. Bon nombre d'individus ont en effet passé un certain temps dans des salles, non pas pour observer les oeuvres (et être dans une démarche de visite

active) mais pour trouver leur chemin à l'aide du plan (changeant de fait le type de pratique). Ceci illustre pour nous le besoin de ne pas se reposer uniquement sur les déplacements et la trajectoire des individus, mais de trouver un moyen de coupler ces données avec d'autres sources de données afin de dresser un état complet de l'activité au sens large des personnes suivies.

De plus, dans le dataset **Museum_3** nous avons accès à l'activité de l'utilisateur sur son application. Celle-ci pourrait nous donner plus d'informations sur l'activité de celui-ci durant sa visite, et de différencier les oeuvres qu'il a préférées, de celles pouvant moins lui correspondre. L'utilisation de données autres que des données de déplacement est cependant un problème plus large que nous aborderons dans le chapitre suivant. Nous avons choisi d'étudier la manière dont nous pouvions coupler le déplacement avec des données extérieures à celui-ci, que l'on appelle *données contextuelles*. Afin d'étudier la façon dont nous pouvons procéder, nous nous sommes intéressés à des concepts spatiaux. Le chapitre suivant décrira comment "*enrichir*" d'une manière générale ces trajectoires, que ce soit des trajectoires dans les musées (indoor) ou des trajectoires GPS (outdoor).

Chapitre 6

Contribution 2 : Modèle multi-aspects générique d'enrichissement sémantique

Table des matières

6.1 L'enrichissement de trajectoires 116
6.2 Un modèle d'enrichissement générique multi-aspects 117
6.2.1 Les aspects 117
6.2.2 Hiérarchie de trajectoires sémantiques 118
Exemples : 120
6.3 Modélisation et visualisation de trajectoires enrichies : Application sur des données réelles 121
6.3.1 Enrichissement du dataset **Geoluciole** 121
6.3.2 Enrichissement du dataset **Geoluciole** par le biais d'un entretien 124
6.3.3 Enrichissement du dataset **Museum_3** avec des données d'application 129
6.4 Conclusion 133

6.1 L'enrichissement de trajectoires

Une trajectoire enrichie de diverses informations contextuelles est appelée une trajectoire sémantique. La trajectoire sémantique a la particularité de combiner les dimensions temporelle, sémantique et spatiale. Une trajectoire sémantique régulièrement citée, introduite en 2008 Spaccapietra et al. [2008b], est le modèle "Stop and Move" qui vise à enrichir la trajectoire avec des informations liées à la mobilité d'un objet. Elle associe à un symbole (tel un lieu ou un type de mobilité) une donnée temporelle et cette succession de symboles prend ainsi la forme d'une séquence temporelle. Le modèle APM (Activity, POI et Move model) [Moreau et al., 2019], représente ici une trajectoire sémantique sous la forme d'une séquence d'épisodes. Un épisode est défini comme étant un couple $(s, [l, u])$ où s est un symbole et [l,u] est un intervalle avec l est la date de début et u la date de fin. Ainsi que mentionné dans le chapitre précédent, les traces de log d'une application mobile de visite dans un musée apportent une information qualitative du comportement du

visiteur que la trajectoire seule de déplacement ne peut pas fournir.

Notre objectif est d'enrichir les données quantitatives avec des données qualitatives en nous appuyant sur la dimension temporelle des données. Il s'agit d'un processus d'annotation de la trajectoire brute afin de lier les données qualitatives aux données quantitatives et de leur donner une sémantique. Une annotation est une donnée additionnelle associée à la trajectoire, à une sous-trajectoire ou à un point de la trajectoire [Parent et al., 2013]. La trajectoire ainsi enrichie devient une trajectoire sémantique. Par le biais de la trajectoire sémantique, nous pouvons dès lors mieux comprendre le comportement d'un individu de par ses déplacements, mais aussi ses agissements. Le modèle APM propose ainsi une ontologie afin de décrire une trajectoire par le biais d'une séquence d'épisodes.

Dans ce chapitre, nous verrons comment structurer une trajectoire sémantique répondant à nos besoins : un modèle suffisamment générique pour pouvoir prendre en compte l'intégralité des datasets présentés dans le chapitre 4 d'introduction aux contributions. Ensuite, nous présenterons les jeux de données enrichis de données contextuelles : D'abord, le dataset $Museum\_3$ enrichie des données d'utilisation de l'application "Visite musée" durant l'expérimentation. Ensuite, nous verrons comment nous avons enrichi le dataset *Geoluciole* avec diverses informations (la météo, les marées mais aussi le questionnaire de l'application *Geoluciole*), mais aussi avec des données provenant directement d'un entretien avec les personnes de notre étude.

6.2 Un modèle d'enrichissement générique multi-aspects

6.2.1 Les aspects

Notre définition des aspects est largement inspirée par les travaux de [Mello et al., 2019] et leur modèle **MASTER** :

Un aspect est un phénomène du monde réel qui est pertinent pour l'analyse des trajectoires. Il est caractérisé par un type et par un ensemble d'attributs reliés à ce type. De plus, le type peut être un sous-type d'un type plus général.

Dans notre modèle, les aspects possèdent un ou plusieurs types (comme une information sémantique, numérique, etc).

Chaque aspect caractérise à sa façon une dimension de la trajectoire particulière avec ses propres spécificités. Ensemble, cela permet d'avoir plus d'in-

formations sur une trajectoire et d'étudier le comportement d'un individu en prenant en compte un maximum d'information disponible.

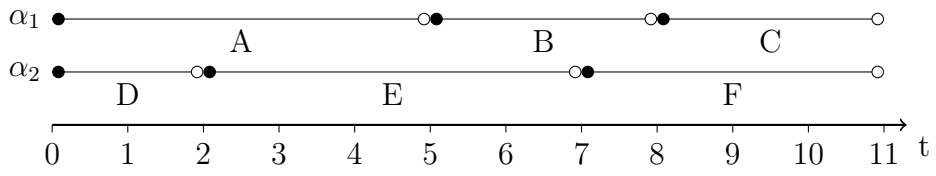


FIGURE 6.1 – Représentation de trajectoires sémantiques

La figure 6.1 montre une représentation des aspects où les séquences temporelles α_1 et α_2 évoluent sur la même temporalité. Dans nos travaux, les aspects proviennent de plusieurs sources de données qui peuvent être :

— De l'information spatiale seulement (quartiers, points d'intérêts, pièces des musées, œuvres et modules).

— De l'information temporelle seulement : la journée ou la nuit.

— De la combinaison des deux, comme la météo et les marées.

— Ou d'autres sources d'informations, comme des entretiens pour déterminer le profil de la personne.

Nous formalisons le modèle de trajectoire sémantique pour une trajectoire T comme suit : $T : \langle \alpha_1, \alpha_2 \dots \alpha_m \rangle$ où T est décrit par m aspects et chaque aspect α_i, avec $1 \leq i \leq m$ peut être soit une séquence temporelle $\alpha_i = (\langle a_k, \underline{t_k}, \overline{t_k} \rangle)_{1 \leq k}$, sur un alphabet Σ_i et $a_k \in \Sigma_i$, une valeur sémantique $\alpha_i = X$ sur un alphabet Σ_i et $X \in \Sigma_i$, ou une valeur numérique $\alpha_i = x$.

La figure 6.2 illustre le diagramme de notre modèle d'enrichissement sémantique, où chaque aspect α_i est une séquence d'épisodes liés par la relation de *Successeur*. En outre, il est possible qu'un aspect possède une structure hiérarchique avec des sous-aspects, qui sont définis à l'aide de la relation *Parent*.

6.2.2 Hiérarchie de trajectoires sémantiques

Afin de structurer les données qualitatives pour les coupler aux traces générales $\langle t_i, x_i, y_i \rangle$, le modèle de représentation doit être assez général pour pouvoir supporter tout type d'éléments. Nous utiliserons ici la notion de séquence temporelle, vue dans l'état de l'art, où une séquence temporelle est une séquence d'intervalles temporels. Nous reprendrons ici la définition définie par [Guyet and Quiniou, 2011b] comme étant des épisodes, afin de lisser

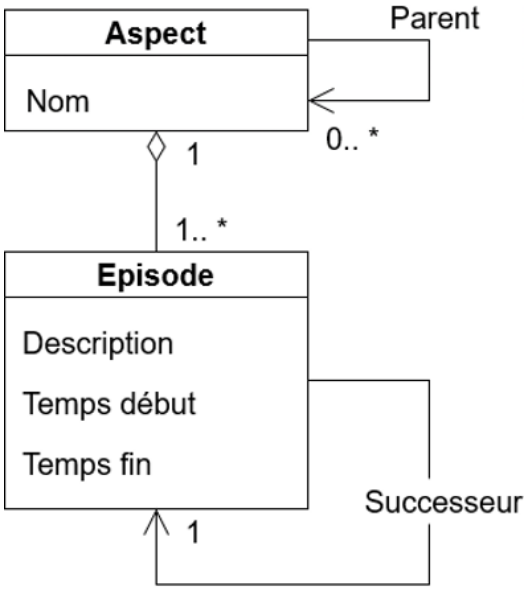


FIGURE 6.2 – Diagramme du modèle d'enrichissement sémantique multi-aspects

le vocabulaire entre les domaines utilisés dans ces travaux.

"Une séquence temporelle S est un ensemble ordonné d'événements, où un événement A = (A, [l, u]) est composé d'un symbole A et d'un intervalle non vide [l, u], où $l, u \in R, l < u$ *sont des dates."*

Nous proposons ainsi un modèle de structure de données reposant sur une succession ordonnée d'intervalles. Ce modèle vise avant tout à représenter de manière la plus exhaustive possible les informations contextuelles (données d'application, météo, etc). Cet objectif nous pousse à vouloir caractériser certains éléments d'une séquence, voire même dans certains cas, les généraliser en un seul et même épisode plus large, afin de pouvoir qualifier les épisodes relevant d'un circuit touristique ou au contraire, les moments d'égarements dans la ville. Ainsi, on peut vouloir considérer une suite d'épisodes A :

$$A = (Velo, [8h00, 9h30]), (Marche, [9h30, 11h00]), (Repas, [11h00, 13h00])$$

comme étant : *(Circuit touristique, [8h00, 13h00])* dans certains cas.

Nous considérons que certains épisodes vont être composés de plus petits épisodes, plus précis. Nous appelons ce genre d'abstraction, un niveau. Fondamentalement, on considère l'épisode "Parcours découverte", comme étant un épisode de niveau 1, et quant à la suite d'épisodes A, composant le "par-

cours découverte" nous le qualifions de niveau 2. Plus nous entrons dans les détails de ce que compose un épisode, plus nous augmentons la valeur du niveau.

Ainsi, nous proposons ici d'ajouter cette même logique de niveau à ce modèle. En outre, nous autorisons chaque élément $s_{i,n}$ de niveau n de S de contenir lui-même une séquence S_{n+1} d'intervalles temporels, où n est le niveau de cette séquence et $\sqsupseteq$ est la relation d'inclusion. Nous formalisons l'inclusion des épisodes des sous-aspects comme suit :

$$(s_{n,i}, [\underline{t_{n,i}}, \overline{t_{n,i}}]) \sqsupseteq (< s_{n+1,i}, [\underline{t_{n+1,i}}, \overline{t_{n+1,i}}] >)$$

Avec $\overline{t_i} \leq \underline{t_{i+1}}$. Chaque niveau correspond à une hiérarchie de trajectoire sémantique. Chacun possède un dictionnaire Σ composé de symboles. Par conséquent, chaque élément d'un niveau est décrits par un ou plusieurs symboles du dictionnaire associé.

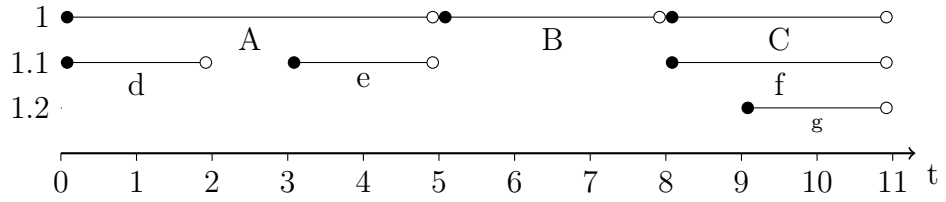


FIGURE 6.3 – Représentation de la hiérarchie de trajectoires sémantiques

La figure 6.3 montre une représentation de la hiérarchie en niveau des trajectoires sémantiques, où un épisode A contient les épisodes d et e tels que $(d, 0, 2) \sqsubset (A, 0, 5)$ et $(e, 3, 5) \sqsubset (A, 0, 5)$.

Exemples :

Ainsi, quelques exemples de hiérarchies :

Hiérarchie liée à l'information spatiale en extérieur :

ville $\sqsupseteq$ quartiers $\sqsupseteq$ plages
$\sqsupseteq$ POI
$\sqsupseteq$ parcs

Hiérarchie liée à l'information spatiale en intérieur

musée $\sqsupseteq$ pièce $\sqsupseteq$ œuvre

Hiérarchie liée à l'information temporelle

séjour $\sqsupseteq$ journée $\sqsupseteq$ matin
$\sqsupseteq$ aprés-midi
$\sqsupseteq$ nuit

6.3 Modélisation et visualisation de trajectoires enrichies : Application sur des données réelles

6.3.1 Enrichissement du dataset Geoluciole

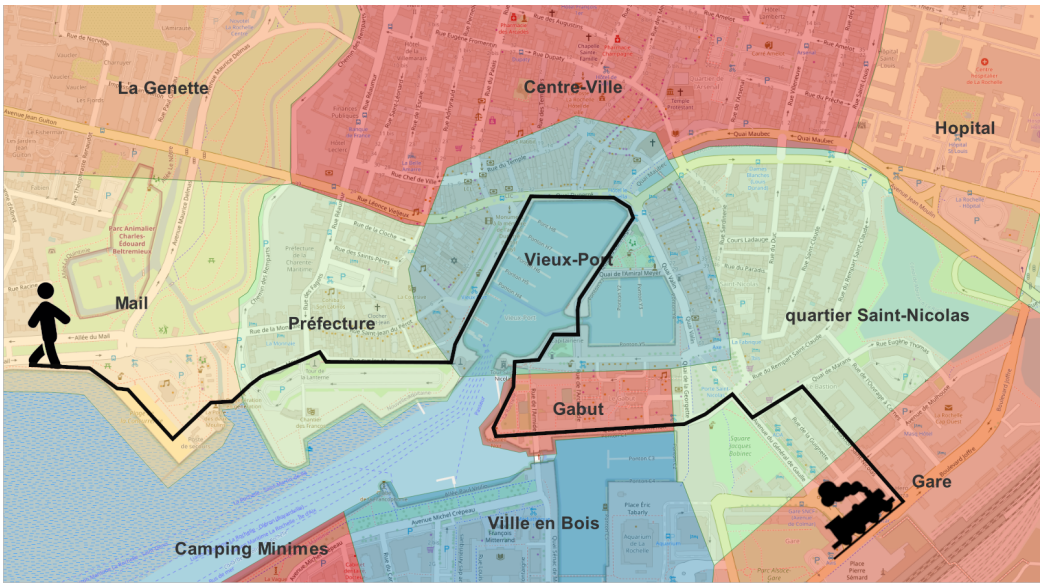
Dans cette sous-section, nous enrichirons le dataset **Geoluciole** avec des données directement récupérées automatiquement par le biais d'API. En faisant cela, nous avons pu ajouter des données météorologiques et de marées. En jouant sur la notion d'aspects définis ci-avant, nous avons construit des trajectoires complexes évoluant sur plusieurs dimensions à la fois.

Aspects d'enrichissement
aspects :
quartiers $\sqsupseteq$ plages
parcs
marée
météo
température
arrivé

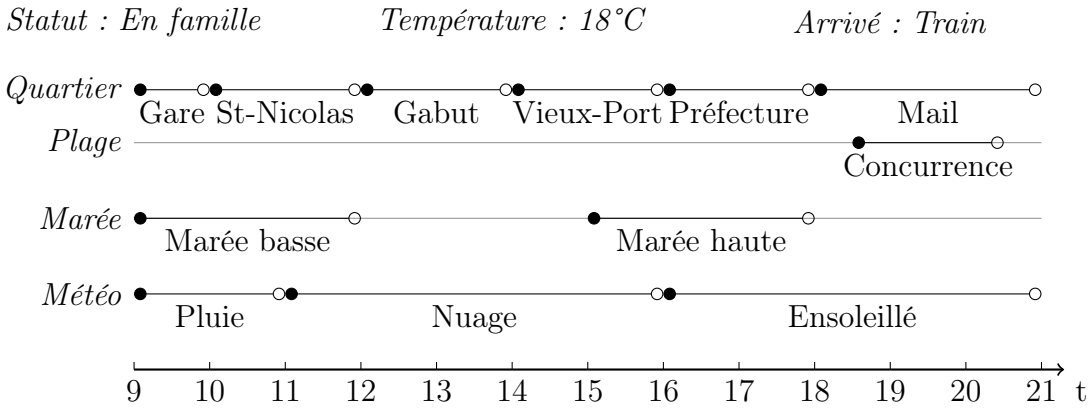
La figure 6.4 représente un exemple de trajectoire sémantique d'une personne. Celle-ci possède 5 aspects, décrit dans la sous-figure 6.4b :

- $\alpha_1 = Quartier$ représente les quartiers traversés par l'individu durant la trajectoire.
- $\alpha_{1.5} = Plage$, bien que ce soit un aspect lié à la géolocalisation de l'individu (comme l'aspect α_1), c'est une donnée que nous voulons voir apparaître si nous souhaitons faire ressortir cette information précise pour étudier les déplacements des touristes à la plage. La notion de plage est ainsi un sous-niveau de l'aspect $Quartier$. Il est à notée qu'un quartier peut avoir plusieurs plages à La Rochelle dans des quartiers différents.
- $\alpha_2 =$ Marée où le dictionnaire de α_2, $\Sigma_{\alpha_2} = [\text{Marée haute}, \text{Marée basse}]$ est une information sur le niveau de la mer selon le temps.
- $\alpha_3 =$ Météo où le dictionnaire de α_3, $\Sigma_{\alpha_3} = [Ensoleille, Nuage, Pluie]$ et représente la météo durant la trajectoire.
- $\alpha_4 =$ Statut où le dictionnaire de α_4, $\Sigma_{\alpha_3} =$ [Voyage entre amis, en famille].
- $\alpha_5 =$ Température, cet aspect est une donnée numérique et garde ainsi son type. Cet aspect représente la température moyenne pendant le déplacement de la personne.
- $\alpha_6 =$ Arrivée, cet aspect est une sémantique qui correspond au moyen d'arrivée dans la ville (Train, voiture, vélo etc.).

Les aspects spatio-temporels sont directement liés à la trajectoire en elle-même. Via un logiciel de **S**ystème d'**I**nformation **G**éographique (SIG), nous



(a) Trajectoire brute



(b) Enrichissement de la trajectoire avec aspects

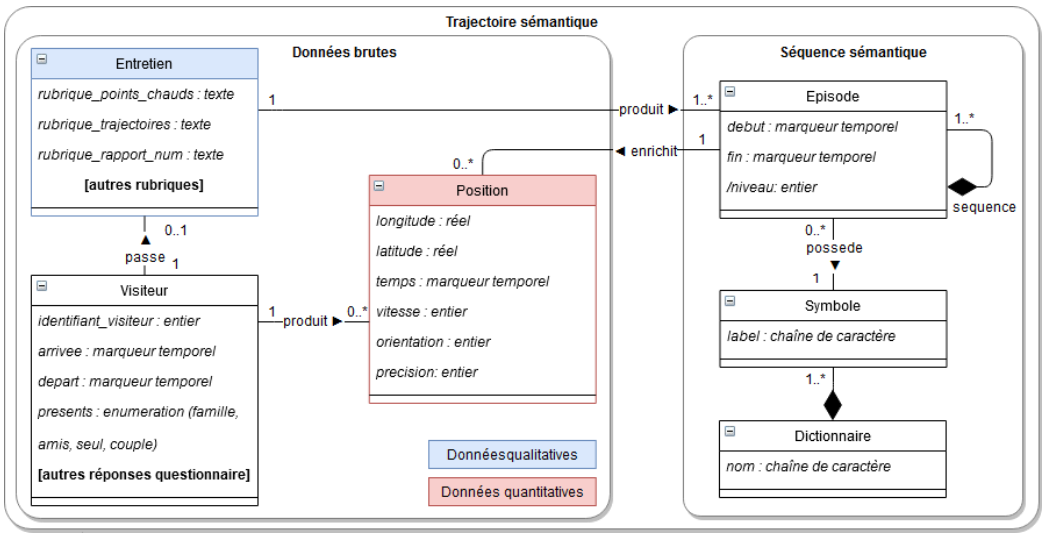
FIGURE 6.4 – Exemple de trajectoire sémantique avec ses aspects

avons pu récupérer une segmentation de la trajectoire brute en séquences d'épisodes suivant un découpage de la ville sur plusieurs aspects. Les aspects plus "contextuels" (α_2, α_3, α_4) ont quant à eux été directement récupérés sur internet via des bases de données ouvertes. Pour récupérer la météo par exemple , nous avons utilisé openweathermap [1]. Toutes ces informations ont donc été regroupées dans une même structure de données afin d'enrichir la trajectoire. Ce premier travail exploratoire permet dans un premier temps de proposer une visualisation plus complète des trajectoires touristiques. Le

1. https ://openweathermap.org/

modèle utilisé dans ces travaux est représenté par le diagramme de classes présenté en figure 6.5. Nous distinguons deux parties différentes dans ce diagramme. La partie gauche représente les données brutes et contient toutes les informations concernant le visiteur observé VISITEUR, ses déplacements et son entretien. Ces informations ont été décrites dans la partie 3.1. Le déplacement du visiteur est décrit par une suite de positions spatio-temporelles POSITION. L'entretien ENTRETIEN passé après le séjour est retranscrit. La partie droite du diagramme représente la partie sémantique de notre modèle. Les réponses du visiteur durant son entretien permettent de construire une séquence d'épisodes EPISODE afin d'enrichir les données brutes quantitatives. Cette séquence est multi-niveaux ; c'est-à-dire que chaque épisode peut être affiné par une nouvelle séquence d'épisodes.

FIGURE 6.5 – Modèle sous la forme d'un diagramme de classe.



6.3.2 Enrichissement du dataset Geoluciole par le biais d'un entretien

Aspects d'enrichissement

aspects :
pratique $\sqsupseteq$ discours
niveau 1.n $\sqsupseteq$ niveau 1.n+1

Dans cette section, nous allons explorer l'enrichissement des trajectoires du dataset **Geoluciole**, en particulier en utilisant les commentaires de certains individus qui font partie de notre dataset de trajectoires, comme expliqué dans le chapitre 4 "Présentation des jeux de données". Au total, une dizaine d'entretiens ont été réalisés. Ce discours a été récupéré après un entretien réalisé par Mélanie Mondo, une doctorante du projet DA3T.

L'entretien est en grande partie retranscrit sous la forme d'un texte. Afin de pouvoir utiliser des méthodes de traitement automatique du langage dans ce genre de travail il faut avant tout savoir ce que l'on cherche et donc définir un thésaurus. Ce travail a été réalisé en collaboration étroite avec des chercheurs dans le domaine de la géographie. Notre approche exploratoire décrite dans ce travail permet d'envisager le traitement automatique du langage naturel afin de créer un processus d'enrichissement semi-automatique dans des travaux futurs.

Extrait de discours d'entretien :

"Quels sont les lieux, les endroits que vous aviez prévu de voir/visiter/-fréquenter ?"
"En fait, on n'avait pas prévu ! Enfin, si uniquement l'Aquarium, c'est pour ça qu'on est venus à La Rochelle. [...] En arrivant, mon épouse a vu qu'avec le Covid il y avait des restrictions et qu'il fallait réserver un créneau donc on a décidé de prendre celui de 17h30, et comme il fallait combler la journée on s'est rendus directement à l'Office du Tourisme".
"Quels sont les lieux, les lieux/endroits marquants et/ou importants de son séjour ?"
"je connais pas le nom du quartier mais pas très loin de l'office du tourisme, il y a un endroit où les gens font des tags sur les murs pas loin du port. On s'est bien arrêté une bonne dizaine de minutes pour regarder ce qu'ils faisaient. Après on s'est dirigé vers la tour Saint-Nicolas, on s'est arrêté un petit moment là aussi à regarder. C'était marée basse et on se demandait ce qu'on voyait au fond (les blocs de béton). On a monté les marches mais on n'a pas pu aller plus loin vu que c'était fermé, c'était passé midi. On n'a pas

visité les tours. Puis on a passé du temps sur les écriteaux qui expliquaient ce qui s'est passé dans la ville avec Richelieu, etc. À chaque fois on s'arrêtait pour prendre des photos dès qu'on pouvait. On a cherché un petit restaurant mais on n'a pas trouvé de suite parce que tous les restos un peu vente rapide étaient bien remplis donc on a marché un petit peu et on s'est retrouvé dans des rue avec des sortes d'arcades [...]".

Nous avons mis en place le tableau 6.1 formalisé afin de remplir manuellement toutes les informations potentiellement intéressantes tirées de l'entretien. Ce tableau découpe l'information en niveaux de précision contraints par un dictionnaire. En effet, un enjeu de l'enrichissement sémantique est de pouvoir utiliser une information qualitative riche (ce que l'enquêté raconte de son séjour) afin d'enrichir la trace GPS associée, mais aussi de pouvoir réfléchir aux différentes pratiques touristiques évoquées par le visiteur. La méthode décrite vise à être générique mais les dictionnaires sont spécifiques à chaque domaine applicatif et doivent être construits par les experts du domaine. Ces dictionnaires peuvent être alimentés au fur et à mesure et peuvent varier dans la précision selon le degré de généralisation souhaité.
Dans le cas des pratiques touristiques, nous décidons de mobiliser les cinq dimensions des pratiques touristiques telles que définies par l'équipe MIT (2011) du laboratoire du LIENS de l'université de La Rochelle, partenaire du projet DA3T :

- Le niveau 1.1 correspond aux dimensions de notre modèle et représente l'activité de l'individu avec $\Sigma_{1.1} = \{\text{jeu}, \text{sociabilité}, \text{découverte}, \text{repos}, \text{shopping}\}$.
- Le niveau 1.2 nous permet de décomposer ces dimensions en actions de la pratique touristique avec $\Sigma_{1.2} = \{\text{promenenade}, \text{restauration}, \text{observation}, \textit{etc.}\}$.
- Le niveau 1.3 se concentre sur les lieux tels qu'ils sont nommés et associés à la pratique touristique avec $\Sigma_{1.3} =$ Ensemble des lieux.
- Le niveau 1.4 est celui qui se rapproche le plus du discours et correspond à des extraits d'entretiens. Ce niveau 1.4, rassemblant des données brutes, permet ainsi de qualifier les autres niveaux, à savoir le lieu, l'action et la dimension de la pratique touristique (Table 6.1). On note $\Sigma_{1.4} = \{\text{Le discours de l'individu}\}$
- Le niveau 1 correspond au niveau racine, à savoir l'ensemble du séjour et de l'entretien.

On peut ainsi envisager une structuration de l'entretien aprés un traitement TAL sous forme d'aspects hiérarchisés de trajectoires sémantiques

Type de vocabulaire	**Exemple**
Niveau racine	Entretien n°3
Typologies des pratiques touristiques	Soin de soi, découverte, ...
Catégories affinées de pratiques touristiques	Observer, visiter, se restaurer, ...
Lieux nommés	Aquarium, chocolatier, ...
Extraits de discours	Trouver un restaurant, lire les panneaux, ...

TABLEAU 6.1 – Ontologie du modèle de représentation de trajectoire sémantique

$$\text{niveau } 1 \sqsupseteq \text{ activité } \sqsupseteq \text{ pratique } \ldots \sqsupseteq \text{ niveau } 1.n.$$

Il s'agit d'une première proposition afin de vérifier l'opérationnalité de notre modèle, mais celle-ci pourra être rediscutée afin de prendre en compte les multiples dimensions d'une même pratique touristique. A terme ce niveau 1.4 pourrait être une autre dimension de la pratique touristique comme les émotions , le niveau marquant de la pratique touristique : "j'ai vraiment aimé ou détesté cet endroit".
Nous intégrons ainsi une logique fondée sur la composition d'intervalles, où le niveau 1 ou niveau racine, est la dénomination de la visite complète de l'individu. Cela nous permet de caractériser, les différents épisodes dont est composé le séjour d'un visiteur lors de ses vacances à La Rochelle. Ce type de modèle permet de filtrer les sujets qui nous intéressent, comme les promenades touristiques ou encore les moments de restauration.

Au moyen d'une approche entretien-centrée, les traces peuvent être subdivisées en multiples séquences qui encapsulent des niveaux d'information variés, englobant ainsi divers aspects de la pratique étudiée. Ces niveaux d'information sont intrinsèquement liés à des dimensions spécifiques de l'activité observée, à savoir l'activité en elle-même, la spatialité qui lui est associée, ainsi que le discours émanant de cette pratique.

Dans l'extrait présenté, plusieurs épisodes sont évoqués mais avec peu d'informations temporelles précises. Nous utilisons donc tant le discours que les traces associées pour retrouver le découpage temporel des épisodes racontés. La figure 6.6 montre la séquence d'épisodes sur plusieurs niveaux, correspondant à un découpage du discours de l'enquêté.

Ainsi, nous avons caractérisé les différents épisodes de notre extrait d'en-

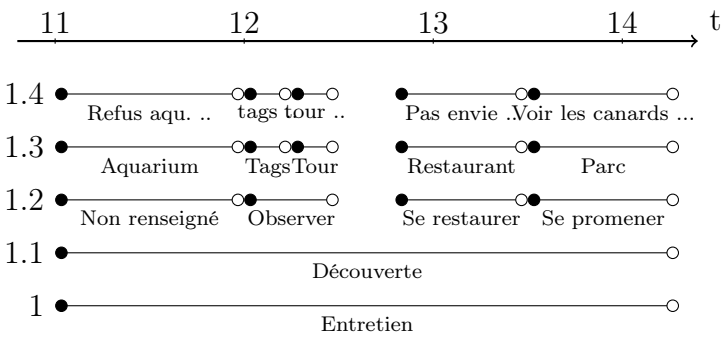


FIGURE 6.6 – Visualisation d'une partie de séquence d'épisodes à différents niveaux

tretien, permettant de visualiser notre trajectoire selon les différents niveaux / aspects.

L'annotation ainsi complétée depuis l'extrait d'entretien, il est possible de visualiser la trace enrichie sémantiquement de façon cartographique. Dans la figure 6.7, nous représentons ainsi notre parcours touristique selon les quatre niveaux vus précédemment.

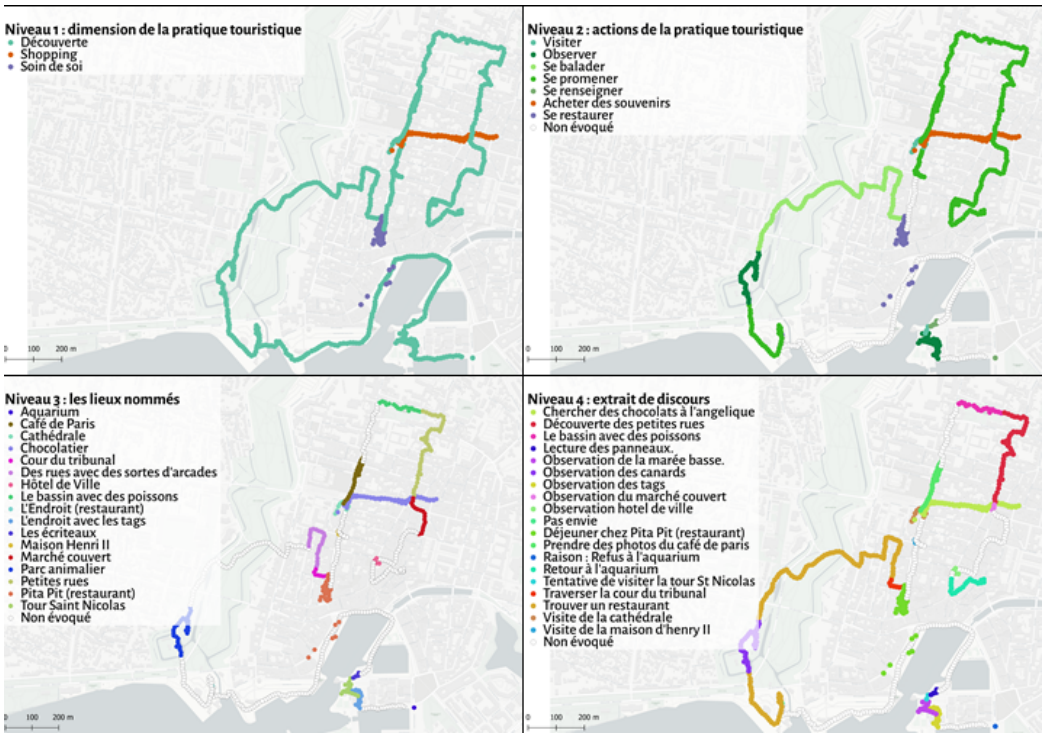


FIGURE 6.7 – Représentation des 4 niveaux de la séquence d'épisodes

La généralisation des deux premiers niveaux (figure 6.7, niveau 1.1 et 1.2) permet de visualiser la trajectoire selon soit la dimension principale de la pratique touristique, soit l'action principale. Le troisième niveau répertorie les lieux évoqués pendant l'entretien. Ainsi, "l'endroit avec tous les tags" (figure 6.7, niveau 1.3) correspondant au toponyme usuel la "friche du Gabut" [2] est considérée comme un lieu marquant et apprécié pendant le séjour. En utilisant l'information temporelle cela permet de retrouver ce lieu-dit sur la carte. Dans le même niveau, nous pouvons également voir que le lieu de restauration a des points dispersés : il s'agit là d'une imprécision du GPS qu'il sera donc possible de corriger. En dernier niveau (figure 6.7, niveau 1.4), nous retrouvons les extraits du discours de l'enquêté : cela permet d'affiner notre compréhension des niveaux précédents mais rend plus difficile une généralisation de la méthode.

Si la carte nous permet de visualiser les pratiques associées à la trace, elle met également en avant le blanc du discours. Ainsi, la balade autour du vieux-port n'est jamais évoquée alors qu'il s'agit d'un lieu emblématique de La Rochelle : cette absence d'informations se retrouve dans la modalité "non évoqué" de nos légendes. Il sera ainsi intéressant d'explorer cette catégorie afin d'en avoir une vue plus précise. A l'inverse, des informations contenues dans l'entretien ne sont pas visibles dans cette cartographie : le début du séjour où l'application n'était pas encore installée, mais également la fin de journée suite à un arrêt imprévu de l'application. Si la suite du séjour a été évoquée en entretien, il n'est pour le moment pas visible sur cette carte.

2. Cet endroit, au coeur de La Rochelle, est dénommé comme tel car il s'agit d'une ancienne friche remarquable dans le quartier du Gabut à La Rochelle, régulièrement occupée durant la période estivale et à vocation culturelle.

6.3.3 Enrichissement du dataset Museum_3 avec des données d'application

La modélisation et la visualisation de trajectoires enrichies par des données de contexte a été effectuée sur le dataset **Museum_3**, afin de pouvoir combiner les données de déplacement et les données d'utilisation d'application mobile durant la période de visite des volontaires à l'étude. Ces deux types de données sont dites hétérogènes, car elles ne sont pas du même type. Les données issues de l'application, retranscrivent sous une forme sémantique l'enchaînement des différentes vues au sein de celle-ci.

Aspects d'enrichissement
aspects : passages ⊒ depl_œuvre appli_œuvre

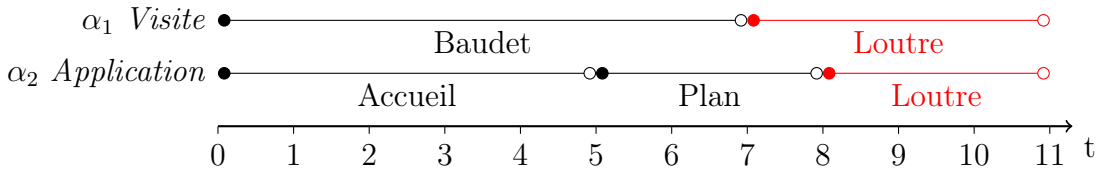


FIGURE 6.8 – Modélisation de la trajectoire sémantique au sein du muséum

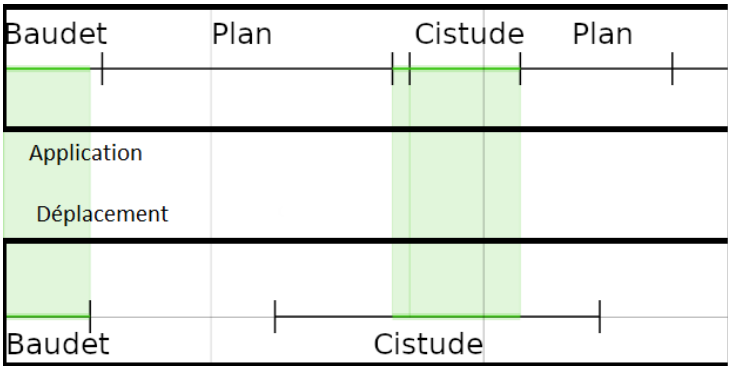


FIGURE 6.9 – Données d'application mises en relation avec les données de déplacement avec le temps

La figure 6.8 montre une mise en forme des données de déplacement et données issues de l'application sous la forme de deux séquences d'intervalles. La figure 6.9 met en relation les deux aspects, où les bandes vertes montrent les concordances en terme d'activités sur les deux aspects. Par exemple, si on est dans la salle de la tortue Cistude, et que l'on regarde les explications

liées à cette œuvre sur l'application, alors on estime qu'il y a concordance.

Ainsi, nous notons un aspect "*déplacement*" ainsi qu'un aspect "*Application*". Les données de déplacement sont directement issues de l'algorithme GraphPositionning du chapitre précédent, et mis sous la forme d'une séquence temporelle d'intervalles retraçant le parcours du visiteur. L'aspect "*Application*" est quant à lui formalisé de la même façon, en créant des intervalles temporels basés sur le temps passé sur une des vues de l'application.

En se basant sur le formalisme que nous avons vu précédemment, les figures 6.10 à 6.14 montrent une visualisation des deux types de données synchronisés.

Nous avons travaillé sur cette visualisation pour pouvoir montrer de façon cohérente plusieurs informations, dans l'objectif d'offrir un "journal de bord" de la visite d'un utilisateur à partir d'une trajectoire sémantiquement enrichie. Ces informations bien différentes se devaient d'être mises en avant et facilement distinguables en même temps :

— La trajectoire de la personne, dans le musée ;

— L'utilisation de l'application, avec le visuel de la vue ;

— Le rapport au temps, afin de dresser le lien entre les déplacements et l'utilisation de l'application ;

— L'animation du trajet de la personne afin d'illustrer son utilisation de l'espace muséal ;

L'indicateur significatif regroupant tous ces points est la concordance entre l'utilisation de l'application et les déplacements de l'utilisateur dans le musée. Ainsi, si l'activité sur "*Visite musée*" est directement reliée à la pièce où se trouve le visiteur, on peut considérer qu'il s'intéresse tout particulièrement à cette pièce du musée, qu'il n'y passe pas seulement, mais est activement engagé dans sa découverte de ce que peut lui proposer le muséum. Afin de mettre en valeur cette donnée, cela signifie que nous sommes en mesure de montrer : La trajectoire de la personne, l'utilisation de l'application, ainsi que la période de temps où il y a synchronisation.

La figure 6.10 montre la représentation que nous avons réalisée sur une trajectoire d'un visiteur, non seulement l'utilisation de l'espace, mais aussi le type de démarche à chaque pas temporel : Recherche d'œuvre à découvrir, découverte et simple déplacement. Le dictionnaire utilisé et les vues de l'application sont décrites dans le chapitre 4 d'introduction aux jeux de données. Les déplacements sont présents sur la partie supérieure de l'image, sous la

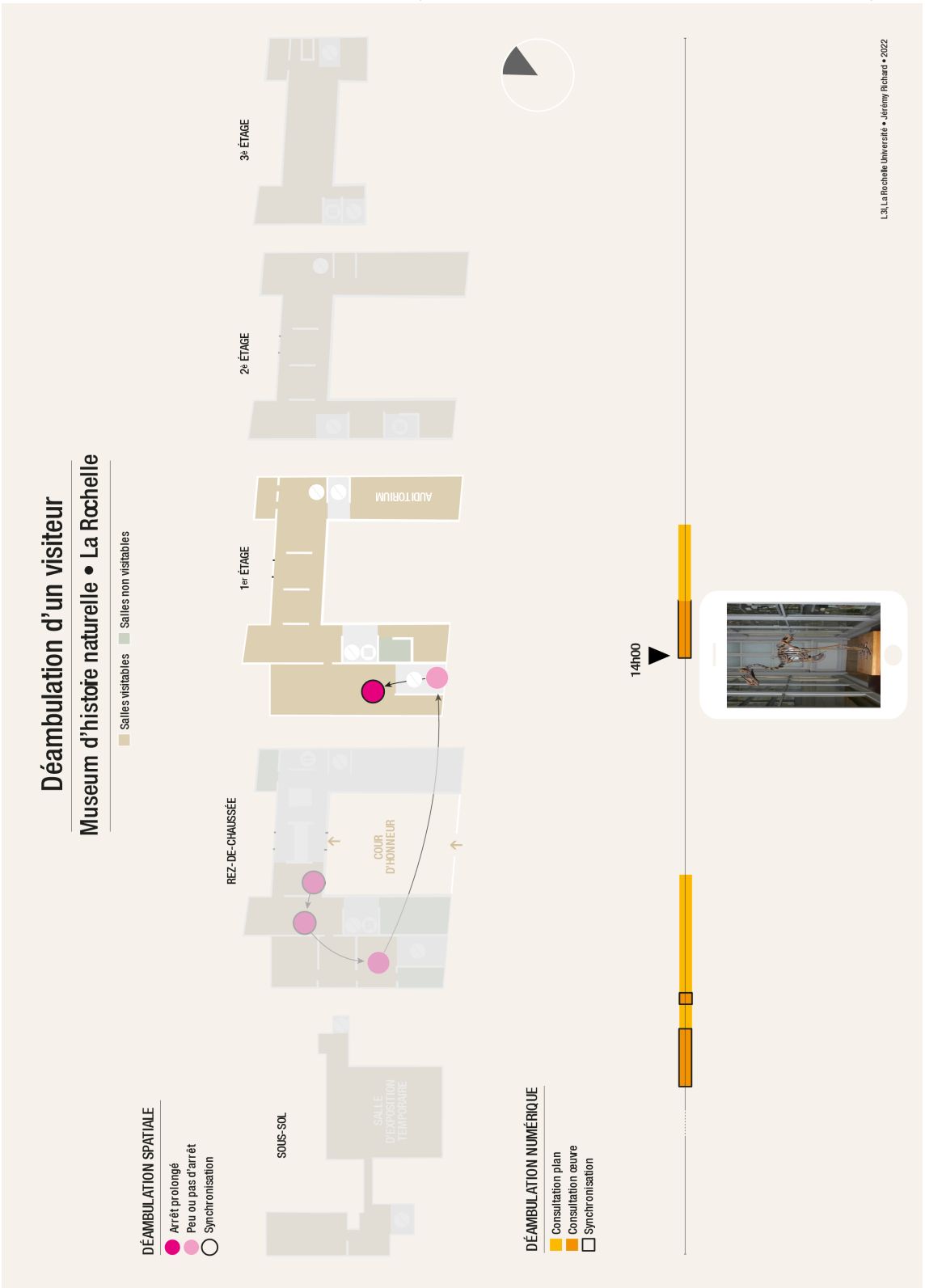


FIGURE 6.10 – Visualisation du déplacement d'un visiteur enrichie des données de l'application mobile

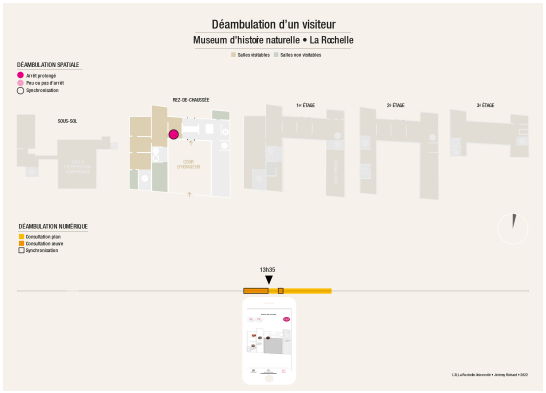


FIGURE 6.11 – Instantané de la trajectoire d'un visiteur (1)

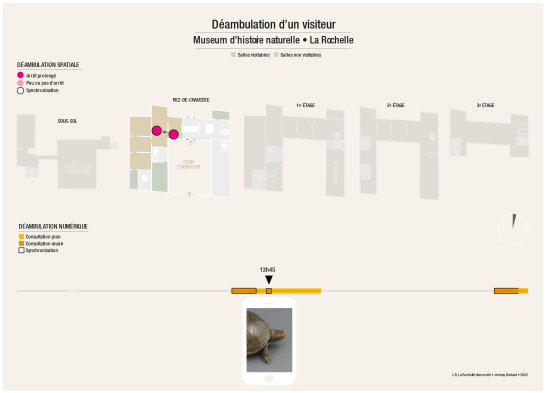


FIGURE 6.12 – Instantané de la trajectoire du visiteur (2)

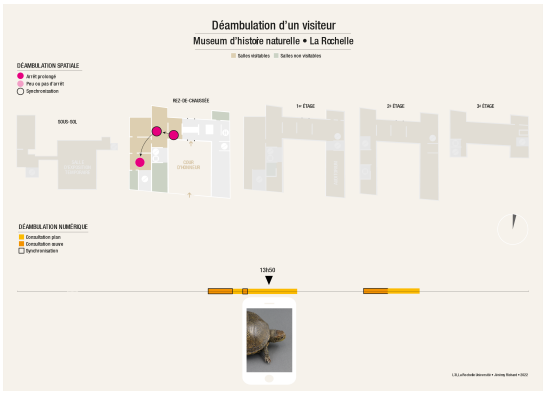


FIGURE 6.13 – Instantané de la trajectoire du visiteur (3)

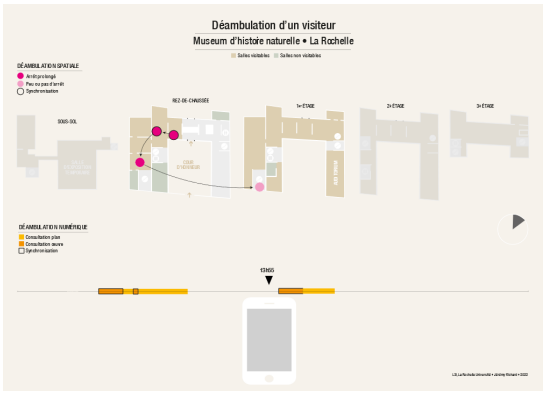


FIGURE 6.14 – Instantané de la trajectoire du visiteur (4)

forme d'un graphe, retour fidèle de l'algorithme GraphPositionning et l'utilisation de l'application se trouve elle en bas de l'image. On y retrouve une frise chronologique, vulgarisant le type de la donnée, la séquence d'intervalles temporels. Une image de la vue courante de l'application est un indicateur visuel de l'activité de la personne. Quand l'application et le déplacement sont concordants, nous le signifions en entourant le cercle de position d'une épaisse bordure noire. L'intensité en couleur du point est directement reliée au temps passé à la position indiquée. C'était une façon pour nous, de signifier le rapport au temps entretenu lors d'une visite par l'usager. Plus la couleur est foncée, plus le temps passé à côté de l'œuvre a été significatif. Du côté de la visualisation de l'application mobile, les intervalles défilent sur une ligne temporelle, changeant de couleur en fonction du type de pratique. La consultation du plan dans l'application mobile pour nous était significative dans l'expérience de visite, où l'on cherche la suite de notre parcours

(œuvre suivante qui intéresse le visiteur). Signifier la différence de pratique met en valeur et ajoute du contexte aux déplacements au-dessus. Lorsqu'il y a concordance entre le déplacement et l'application, encore, nous le signifions en ajoutant une épaisse bordure noire autour de l'intervalle, où le visiteur manifeste un intérêt certain pour cette œuvre.

Cette visualisation a été pensée pour un affichage dynamique du résumé de la visite comme le montrent les figures 6.11 à 5.7. On peut y voir le graphe de déplacement se construire peu à peu, et la frise d'utilisation de l'application défile en fonction du temps. Ajoutant une autre dimension contextuelle dans la visualisation d'une visite nous avons un regard plus "proche" de la réalité et plus fidèle à l'expérience de visite de l'individu.

6.4 Conclusion

Ce travail exploratoire nous a permis de valider l'intérêt de cette méthode d'enrichissement : en segmentant par le temps la trace de localisation à partir des informations recueillies de diverses manières (les pièces pour les musées et quartiers pour l'outdoor), nous avons pu ajouter une information sémantique à celle-ci et ainsi l'enrichir. Notre approche est suffisamment générale pour pouvoir être appliquée sur diverses expérimentations différentes. Ainsi, nous avons pu appliquer notre méthodologie à des trajectoires de visites dans un musée où les données d'enrichissements étaient des données d'application mobile. Cela nous a permis de préciser d'autant plus l'expérience de visite d'un individu, dans la continuation directe de la reconstruction de sa trajectoire au chapitre précédent.

Cette méthodologie nous a permis d'enrichir des trajectoires de touristes en extérieurs dans la ville de La Rochelle avec dans un premier temps des extraits d'entretiens pour un enrichissement "personnalisé", où l'information sémantique apportée provenait directement de l'entretien de l'individu. Puis, dans un second temps des données provenant d'API web qui a pu permettre un enrichissement avec des informations météorologiques, de température, etc.

En faisant cela, nous avons pu proposer une visualisation de chacune de ces expérimentations, que ce soit dans les musées ou en extérieur. Ces visualisations ont su préciser et apporter une tout autre dimension aux données de déplacement. La visualisation des données de déplacement a été une étape cruciale dans la compréhension de ces expérimentations, offrant une représen-

tation graphique claire et précise de ces données. Toutefois, il est important de noter que cette approche ne permet pas de rendre compte de manière exhaustive de l'expérience de chaque individu, ni des dynamiques collectives qui se déploient au sein d'un groupe. Ainsi, pour une analyse plus approfondie de l'enrichissement des personnes impliquées dans ces expérimentations, une étude approfondie de chaque aspect est nécessaire.

Pour ce faire, nous procéderons à une analyse plus approfondie des données de déplacement en utilisant le modèle de représentation mentionné précédemment, conjointement avec des techniques de fouille de motifs (pattern mining). Cette approche nous permettra de révéler les pratiques courantes chez les individus étudiés. La fouille de motif dans les séquences d'intervalles étant un sujet déjà exploré dans l'état de l'art de ce domaine. En effet, maintenant que nous avons enrichi les trajectoires, à la fois à l'échelle de la ville et à l'échelle des musées, l'enjeu est maintenant d'analyser celles-ci, en prenant en compte les différentes informations que nous leur avons ajoutées. Bien que cette méthode d'enrichissement de données offre de nombreuses possibilités, la principale limite est qu'elle augmente considérablement le volume de données à traiter. En effet, l'ajout d'aspects augmente le nombre d'informations à prendre en compte lors de l'analyse et a un impact important sur la masse de motifs extraits. Afin de traiter cette problématique, nous avons opté pour l'utilisation de la plateforme **GALACTIC**, qui offre la possibilité d'utiliser une approche stratégique permettant de cibler les aspects pertinents pour notre étude. La sous-section suivante expose les jeux de données qui ont été enrichis avec des informations sémantiques et qui présentent une taille adéquate pour l'analyse de sous-groupes. Ces données seront utilisées dans la contribution 3.

Enrichissement des datasets utilisés dans la contribution 3

Nous avons enrichi les trajectoires touristiques des datasets **Geoluciole**, **Museum_3** et de la **Cité du Vin** par plusieurs informations contextuelles permettant d'obtenir plusieurs aspects :

— **Geoluciole** : Les trajectoires du dataset **Geoluciole** ont été enrichies de plusieurs informations, formant un certain nombre d'aspects de types différents. Les aspects *district*, *weather*, *tide*, *green space* (parcs) et *beach* sont représentés sous forme de séquences d'intervalles temporels. Dans chaque aspect temporel d'une séquence, l'information temporelle représente l'heure du jour. Les aspects *district* et *green space* se

déduisent de l'information spatiale (x_i, y_i) des trajectoires, tandis que les aspects *beach*, *weather* et *tide* se déduisent de l'information temporelle t_i des trajectoires et de sites spécialisés. L'acquisition ainsi que le type de ces données sont décrits dans le tableau 6.2. Pour les besoins de cette étude, le nom des aspects a été traduit en anglais.

Aspects	Type	Dictionnaire Σ	Source
district	Seq. temporelle	Quartiers de La Rochelle	Trajectoire
weather	Seq. temporelle	{ Pluvieux, Nuageux, Soleil }	Internet
beach	Seq. temporelle	Plages de La Rochelle	Trajectoire
green space	Seq. temporelle	Parcs de La Rochelle	Trajectoire
tide	Seq. temporelle	{High, Low}	Internet
temperature	Numeric		Internet
stay status	Chain	{ family, friends }	Questionnaire

TABLEAU 6.2 – Aspects du dataset **Geoluciole**

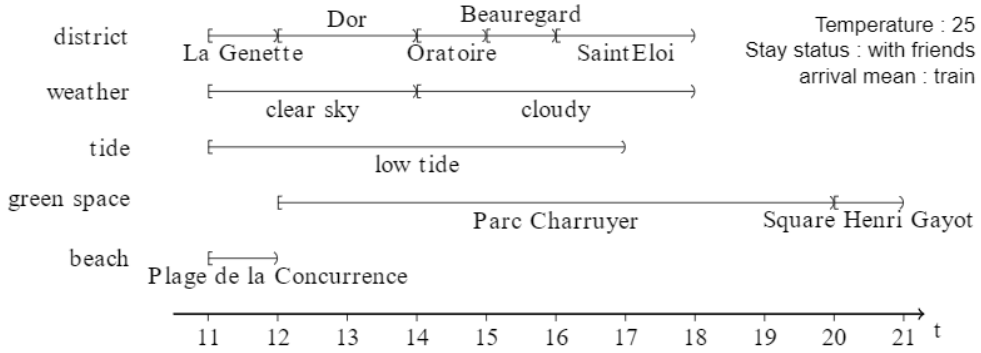


FIGURE 6.15 – Exemple des différents aspects d'une trajectoire de **Geoluciole**

— **Museum_3** : Ce dataset contient des trajectoires sémantiques à deux aspects : d'une part, les déplacements du visiteur, c'est-à-dire l'enchaînement des salles traversées pendant sa visite, et d'autre part, l'utilisation de l'application mobile "Visite Musée" du Muséum, notamment la consultation des différentes vues disponibles sur l'application, telles que le plan ou les détails d'une œuvre. Le tableau 6.3 décrit ces deux aspects. Les intervalles temporels des séquences temporelles sont des durées en minutes. Ne disposant que de 30 trajectoires, nous ne pouvions utiliser les heures de la journée. Ainsi une séquence telle que $A = \langle([0, 5], AneBaudet), ([5, 12], Cistude)\rangle$ correspond à un passage

devant l'âne baudet les cinq premières minutes de la visite, puis devant la Cistude pendant les 6 minutes suivantes. Ainsi nous pouvons comparer les différentes trajectoires en terme de durée.

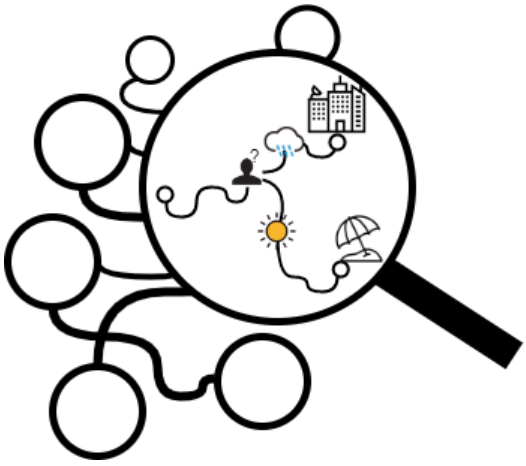
Aspects	Type	Dictionnaire
Visite	séquence temporelle	Σ = Ensemble des points de passage
Application	séquence temporelle	Σ = Ensemble des vues de l'application

TABLEAU 6.3 – Aspects du dataset **Museum_3**

— La Cité du Vin : Le dataset de la **Cité du Vin** est constitué de trajectoires sémantiques à un seul aspect de type séquence simple qui représente l'ensemble des modules activés pendant la visite d'un individu, ainsi que des informations décrivant le profil du visiteur telle que son âge ou sa langue. Il est à noter que le dataset de la **Cité du Vin** possède déjà un aspect "module" sous forme de séquence temporelle d'intervalles et ne nécessite pas d'opérations complémentaires.

Chapitre 7

Contribution 3 : Analyse multi-séquences et hétérogène



Ce travail a été présenté pour CLA2022 Richard et al. [jun 2022]

Table des matières

7.1 Introduction 138
7.2 Analyse des trajectoires avec **GALACTIC** 139
Données numérique : 140
Les données de séquences simples : 140
Les données de séquences temporelles : 141
7.2.1 Impact des stratégies sur le déluge de motifs 143
7.3 Expérimentations 146
Analyse Naïve utilisant des intervalles temporels 146
7.3.1 Détection de comportements particuliers dans les trajectoires de visites des musées 148
7.3.2 Détection de comportements particuliers dans les visites de la ville : Le choix des données 152
Les trajets à la plage en fonction de la météo 152
Les habitudes de séjour suivant le profil d'une personne 154
Conclusion de l'expérimentation 155
7.4 Conclusion 156

7.1 Introduction

Dans la précédente contribution, nous avons étudié l'enrichissement de trajectoires par des informations contextuelles à celles-ci avec la notion d'aspects (quartiers, météo, discours ..). Dans ce chapitre, nous proposons un moyen d'analyser ces trajectoires enrichies multi-aspects en utilisant les similarités entre les trajectoires sémantiques et les séquences temporelles. Non seulement la trajectoire "prenant en compte son contexte" est difficile à analyser, mais les dimensions temporelles d'une trajectoire sémantique sont rarement prises en compte et seules quelques études abordent directement ce problème. Les avancées récentes dans le domaine de la fouille de motifs et de l'analyse formelle de concepts (AFC) proposent des voies prometteuses pour analyser des données hétérogènes et complexes telles que les séquences que nous utilisons dans nos travaux. Le traitement de trajectoires sémantiques

complexes à l'aide du framework **GALACTIC** et de l'algorithme NextPriorityConcept [Demko et al., 2020] permet d'extraire de tels modèles de mobilité [Boukhetta, 2022]. Nous pouvons également modifier la stratégie d'exploration de type de données contextuelles au cours du processus, ce qui permet une analyse interactive où nous pouvons suivre la mobilité d'un individu en se focalisant sur les aspects qui nous intéressent dans la trajectoire de l'individu. De plus, le changement interactif de stratégies est un moyen d'éviter le déluge de motifs en se concentrant sur des comportements spécifiques. Nous proposons une méthode d'analyse des données de mouvement qui se concentre sur l'aspect sémantique des trajectoires et sur l'analyse interactive des différents aspects présents. A ce jour, il n'existe que très peu de méthodes d'analyse qui mixent un objet en mouvement et des informations contextuelles.

L'objectif de ces travaux sera de fournir des outils permettant à l'analyste de répondre à sa problématique de l'analyse comportementale d'individus, en se basant à la fois sur leurs données de déplacement et des différents aspects enrichissant ces traces. Ainsi, nous nous intéresserons à l'extraction de sous-groupes d'individus possédant un motif commun décrivant un comportement partagé pour chaque sous-groupe dans notre jeu de données. Puisque nous travaillons avec des données réelles sans la possibilité d'avoir une vérité terrain, notre approche sera non-supervisée en mettant l'emphase sur l'explicabilité et la lisibilité des motifs extraits.

Ce chapitre explorera une telle approche avec des expérimentations sur des données hétérogènes. Nous utilisons des trajectoires sémantiques réelles capturées par notre équipe et composées de parcours touristiques dans la ville de La Rochelle (France) et enrichies de connaissances contextuelles, avec le dataset ***Geoluciole***. Nous utiliserons également le dataset **Museum_3** ainsi que celui de ***La cité du vin***, afin d'appuyer la généricité d'utilisation de l'outil sur des trajectoires sémantiques, à la fois indoor et outdoor. Ces analyses auront pour but de détecter des comportements communs de façon lisible parmi les individus susceptibles d'intéresser l'analyste des données.

7.2 Analyse des trajectoires avec GALACTIC

Dans ce chapitre, nous utiliserons **GALACTIC** pour extraire des concepts de nos jeux de données **Geoluciole**. Afin d'utiliser les trajectoires sémantiques multi-aspects avec **GALACTIC**, nous notons notre jeu de données de trajectoires $\mathcal{D}$ qui est donné par $T_j : \langle \alpha_{1_j}, \alpha_{2_j}, \ldots, \alpha_{n_j} \rangle$. L'espace de description utilisé avec **GALACTIC** pour toute trajectoire $T_j \in \mathcal{D}$ est un

ensemble de prédicats définis par :

$$\delta(T_j) : \langle \delta_1(\alpha_{1_j}), \delta_2(\alpha_{2_j}) \dots \delta_m(\alpha_{n_j}) \rangle \tag{7.1}$$

Où $\delta_i(\alpha_i)$ est l'ensemble des prédicats pour chaque aspect α_i.

NextPriorityConcept utilise les descriptions δ pour définir un ensemble de prédicats décrivant les attributs. En tant que prédicats, les descriptions peuvent alors être vues comme des données/attributs binarisés associés aux données et le dataset comme une table binaire objets x prédicats. Nous utilisons également des stratégies pour affiner chaque concept en des concepts prédécesseurs plus petits à chaque itération.

Données numérique :

Nous décrivons un ensemble $A = \{x_1, ..x_k\}$ de valeurs numériques par la description SimpleNumericalDescription (SND) δ_S :

$$\delta_S(A) = \{\text{is greather than min(A), is lesser than max(A)}\}$$

La stratégie SimpleNumericalStrategy (SNS) σ_Q utilisée avec la description des quantiles $\sigma_Q(A, k)$ où k est le nombre de quantiles est définie telle que :

$$\sigma_Q(A) = \text{ is greater than } q_j \text{ ; is lesser than } q_j \text{ ; } q_j \text{ is a k-quantile} \tag{7.2}$$

Les données de séquences simples :

Parmi les descriptions et stratégies proposées par Salah Eddine Boukhetta, nous nous attarderons sur les descriptions et stratégies de fouille de séquences. Ces descriptions et stratégies manipulent des prédicats du type : "contient x comme sous-séquence / match x". Nous les définissons ici par des ensembles de sous-séquences, les prédicats s'en déduisant. Ils reposent sur la relation de sous-groupe définie comme suit :

Definition 11. *Une séquence $S = \langle s_1, s_2 \dots s_n \rangle$ est une sous-séquence d'une séquence $R = \langle r_1, r_2 \dots r_m \rangle$ et on écrit $S \sqsubseteq_s R$ s'il existe des entiers $1 \leq i_1 < i_2 < \dots < i_n \leq m$ tels que $s_j = r_{i_j}$ avec $j \leq n$. On dit aussi que R est une super-séquence de S ou R contient S.*

Il existe deux descriptions pour la fouille de séquence simple :

— La description de sous-séquences communes maximales SCM. Pour un ensemble de séquences $A \subseteq G$, la description SCM est définie par :

$$\begin{aligned}\delta_{SCM}(A) = \{X \in \Sigma^* | \forall S \in A, X \sqsubseteq_s S \text{ et si} \\ \exists X' \in \delta_{SCM}(A) \text{ tel que } X \sqsubseteq_s X' \text{ alors } X = X'\}\end{aligned} \tag{7.3}$$

— La description de sous-séquences communes préfixées SCP. Pour un ensemble de séquences $A \subseteq G$, la description SCP est définie par :

$$\begin{aligned}\delta_{SCP}(A) = \{X \in \Sigma^* | \forall S \in A, X \text{préfixe de } S \\ \text{et si } \exists X' \in \delta_{SCP}(A) \text{ tel que } X \sqsubseteq_s X' \text{ alors } X = X'\}\end{aligned} \tag{7.4}$$

La génération du préfixe commun maximal pour un ensemble de séquences A est réalisée par le biais de la description. Les prédicats qui en découlent sont formulés comme "S commence par la séquence X", où X représente tout élément de l'ensemble des préfixes communs de A, noté $\delta_{SCP}(A)$ et $S \in A$.

Deux stratégies sont associées aux descriptions SCM et SCP : la stratégie naïve SN et la stratégie augmentée SA :

— La stratégie naïve SN permet de sélectionner toutes les sous-séquences possibles, sans limite de support, de A avec SS(A) l'ensemble des sous-séquences de A et donc tous les sous-ensembles $A' \subset A$:

$$\sigma_{SN} = \{x | x \in SS(A), \text{support(x)} < 1\} \tag{7.5}$$

— La stratégie augmentée SA va générer des prédicats composés seulement des sous-séquences communes $\delta(A)$ augmentées d'un élément de l'alphabet :

$$\sigma_{SA} = \{x + a | x \in \delta(A), a \in \Sigma\} \tag{7.6}$$

Les données de séquences temporelles :

Pour des séquences d'intervalles temporelles nous utiliserons la description **S**ous-séquences d'intervalles **C**ommunes **MAX**imales SCMAX et la stratégie de cardinalité minimale augmentée CMA [Boukhetta et al., 2020b]. Nous notons qu'une séquence temporelle $(x_i, \underline{t_i}, \overline{t_i})$ peut aussi se définir sous la forme $(X = \{x_i\}, T = \{(\underline{t_i}, \overline{t_i})\}$. Ces descriptions δ et stratégies σ reposent sur la relation de sous-séquence définie comme suit :

Definition 12. *Pour une séquence d'intervalle S^1, celle-ci est une sous-séquence d'intervalles d'une autre séquence d'intervalle S^2 si pour tout $(T_i^1, X_i^1) \in S_1$ il existe $(T_j^2, X_j^2) \in S^2$ tel que $T_i^1 \preceq T_j^2$ et $X_i^1 \subseteq X_j^2$. On note ainsi que $S^1 \sqsubseteq_s S^2$.*

Avant d'expliciter les formules des descriptions et des stratégies, nous introduisons Φ, un opérateur de projection sur un intervalle T ou sur un ensemble d'éléments X. Cet opérateur renvoie tous les ensembles d'éléments d'une séquence d'intervalle S inclus dans T. Ainsi on note :

$$\Phi_T(S) = \{X | T' \preceq T \text{ et } (T', X') \in S\}$$

À l'inverse, cet opérateur s'utilise aussi sur un ensemble d'éléments X afin de sélectionner tous les intervalles où les éléments de X peuvent apparaître :

$$\Phi_X(S) = \{T' | X' \subseteq X \text{ et } (T', X') \in S\}$$

— La description SCMAX δ_{scmax} est formalisée pour un ensemble de séquences d'intervalles $A \subseteq G$, par :

$$\delta_{scmax}(A) = \{\langle (T, X) \rangle : \forall S \in A, X \subseteq \Phi_T(S)\}$$

SCMAX génère l'ensemble des sous-séquences d'intervalles communes pour un ensemble de séquences A. Ces sous-séquences d'intervalles communes pourront avoir la structure $\langle (T_1, X_1), (T_2, X_2) \rangle$ pour un ensemble de séquences d'intervalles dont T_1 et T_2 sont des intervalles communs et X_1, X_2 des ensembles d'éléments en commun.

— La stratégie CMA σ_{CMA} est formalisée pour un ensemble de séquences d'intervalles $A \subseteq G$ par :

$$\begin{aligned} \sigma_{CMA}(A) = \{\langle (T, X) \rangle | \forall S \in A, \Phi_T(S) \subseteq X \text{ et } \forall x \in X, \\ card(A, T, x) = |A| \text{ où } card(A, T, x) = card_{min}(A, T)\} \end{aligned} \tag{7.7}$$

Où $card(A, T, x)$ correspond au nombre de séquences d'un ensemble A possédant l'élément x durant l'intervalle T. La stratégie de CMA calcule tous les raffinements possibles d'un concept $(A, \delta_{SCMAX}(A))$ en ajoutant les éléments de cardinalité minimale $card_{min}(A, T)$ dans les intervalles des prédicats de SCMAX.

Tout d'abord, nous étudierons l'impact d'une analyse hétérogène sur le nombre de concepts (ie. sous-groupes) générés. Ensuite, nous nous concentrerons sur des parties spécifiques du jeu de données avec des informations

sémantiques sélectionnées afin de mieux identifier les comportements touristiques. La table 7.1 indique les descriptions et les stratégies utilisées pendant cette expérimentation. Le tableau 6.2 présente une description de chaque jeu de données.

	δ	σ
Séquence temporelle	SCMAX	CMA
Séquence simple	SCM, SCP	SN, SA
Numérique	SND	SNS

TABLEAU 7.1 – Descriptions et stratégies utilisées pour chaque type de données

7.2.1 Impact des stratégies sur le déluge de motifs

L'exécution d'une analyse hétérogène avec plusieurs types de descriptions et de stratégies peut réduire le nombre de concepts dans le treillis final. Alors qu'une stratégie "naïve" sélectionne tous les prédicats au sein d'une description δ, comme σ_{CM}, d'autres stratégies, comme σ_{AMC}, sélectionneront des prédicats spécifiques et en ignoreront d'autres afin de réduire le nombre de motifs générés. Cette sous-section est un exemple illustrant l'impact des stratégies sur le déluge de motifs.

Le tableau 7.2 présente un jeu de données D composé de cinq individus, caractérisés chacun par une valeur numérique et une séquence d'intervalles. Dans cet exemple, nous utiliserons la description numérique simple SND et la stratégie des quantiles SNS pour les aspects numériques et la description de l'intervalle commun maximal ainsi que la stratégie δ_{AMC} pour la séquence d'intervalles. Le résultat est présenté figure 7.2

En utilisant que la description numérique et la stratégie quantile σ_Q, nous obtenons le treillis de la figure 7.1. En ajoutant une description et une stratégie d'intervalles, on obtient le treillis L' avec un nombre inférieur de concepts.

Nous pouvons observer que les concepts \$1 ($x \leq 4$) et \$2 ($x \geq 2$) de la figure 7.2 de L n'apparaissent pas dans le treillis L' car ils n'ont pas été sélectionnés par la stratégie "minimal cardinality". Les concepts contenus dans L' sont les plus représentatifs des deux espaces de description.

Individus	Numérique	Interval
a	1	(8.3, 11) : “P”, (13, 15) : “M”, “H”
b	2	(10, 12) : “P”, ”H”, (14, 16) : “M”
c	3	(8.3, 12) : “P”, (14, 16) : “M”, “C”
d	4	(7, 9) : “P”, “H”, (12, 13) : “M”
e	5	(10, 11) : “P”, (12, 12) : “M”, “C”

TABLEAU 7.2 – Dataset D

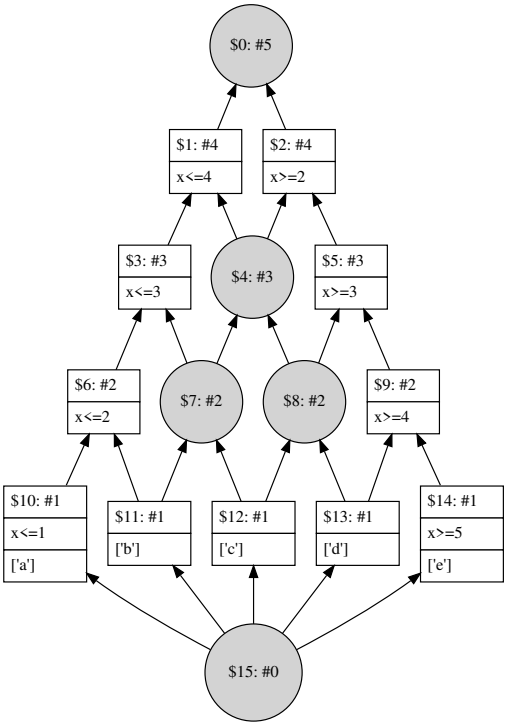


FIGURE 7.1 – Treillis L avec la description numérique SND et la stratégie σ_Q SNS

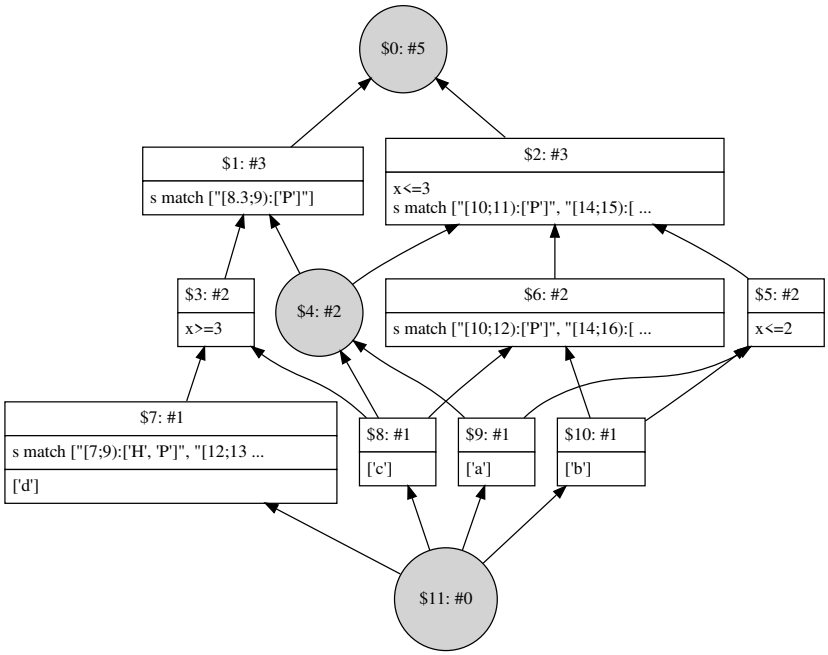


FIGURE 7.2 – Treillis obtenu pour le dataset D (table 3.3) avec la description d'intervalle temporel δ_{SCMAX}, la description numériqueSND, la stratégie σ_Q SNS et la stratégie σ_{CMA}

7.3 Expérimentations

Dans cette section, nous analyserons les trajectoires sémantiques multi-aspects des jeux de données de la **Cité du Vin**, de **Museum_3** ainsi que de **Geoluciole** afin de détecter des motifs communs sur plusieurs aspects à la fois. Nous commencerons par examiner l'impact des stratégies sur la quantité de motifs générés avec le dataset **Geoluciole**. Ensuite, nous présenterons les expérimentations en deux parties : les visites de musées et les trajectoires extérieures.

— **Dans les musées.** Nous commencerons par analyser simplement l'aspect spatial des trajectoires des visiteurs de la **Cité du Vin** pour détecter les différentes prises de décision au cours de leur visite, notamment en début de parcours. Nous poursuivrons ensuite avec l'analyse des données spatiales de **Museum_3** reconstruites dans la contribution 1 et couplées aux données de l'application "Visite musée", afin d'obtenir une analyse des comportements sur deux aspects différents.

— **En extérieur.** Dans cette étude, nous utiliserons le jeu de données **Geoluciole**, qui a été enrichi de plusieurs aspects dans la contribution 2. Nous examinerons tout d'abord l'impact de la météo sur les habitudes de séjour, puis nous étudierons les comportements communs en fonction du profil du touriste.

Analyse Naïve utilisant des intervalles temporels

En utilisant le jeu de données **Geoluciole**, nous considérerons dans un premier temps le seul aspect *districts* puisqu'il constitue une segmentation géographique des trajectoires en *districts* de la ville, avec la description SC-MAX et la stratégie CMA de séquences temporelles. En procédant ainsi, nous obtenons 416 concepts avec un faible support. Le grand nombre de concepts générés impliquant seulement une ou deux trajectoires est un indicateur qu'il n'y a pas de lieux significatifs où plusieurs individus se sont trouvés au même moment de la journée. Pour affiner notre recherche, nous avons ajouté d'autres aspects des trajectoires sémantiques afin de découvrir de nouveaux comportements significatifs dépendant de facteurs externes. La section suivante présente des expérimentations en prenant en compte plusieurs aspects des trajectoires sémantiques, tels que *weather* ou *statut de séjour*. Avec seulement *weather* et la même approche, nous obtenons 379 concepts.

Dans cette expérience, nous exploitons la capacité de **GALACTIC** à traiter des données hétérogènes en ajoutant successivement les autres attributs à l'attribut *weather* (qui génère le plus grand nombre de concepts). La

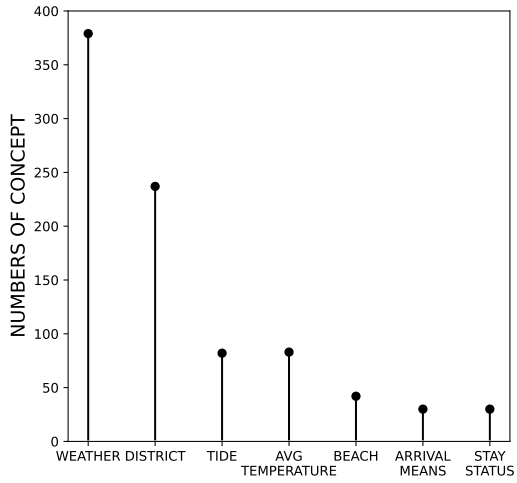


FIGURE 7.3 – Nombre de concepts générés en fonction des attributs ajoutés successivement

figure 7.3 indique le nombre de concepts obtenus par l'ajout de chaque attribut, ce qui diminue de 379 (avec le seul attribut *weather*) à 36 (avec tous les attributs). Le tableau 7.4 donne les descriptions par prédicats communs de trois concepts, dont le concept numéro \$0 qui correspond au concept $\top$ contenant toutes les données initiales pour tous les attributs.

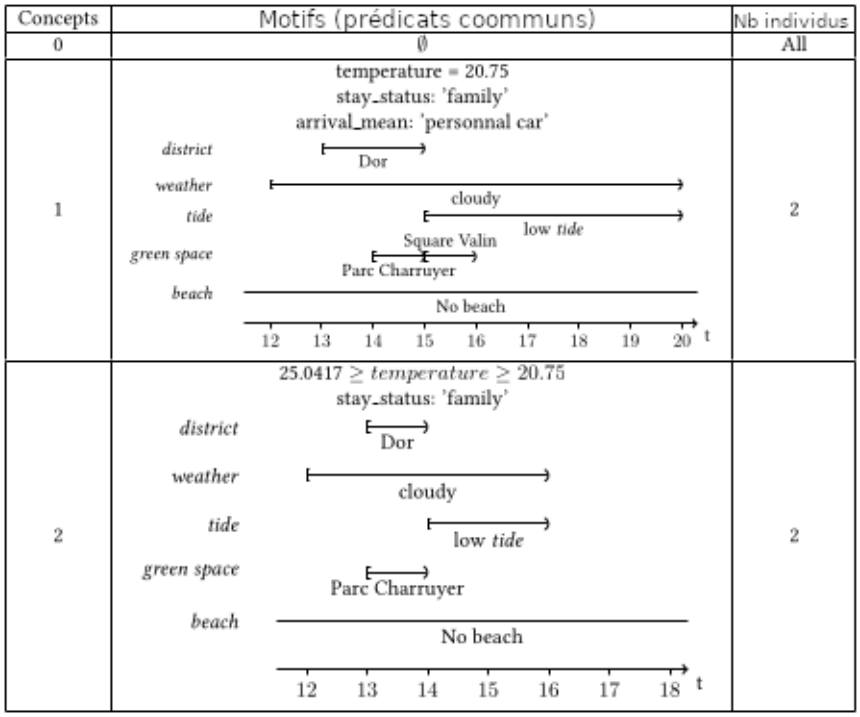


FIGURE 7.4 – Prédicats de 3 concepts du treillis obtenu à partir du dataset **Géoluciole** avec les 5 aspects

7.3.1 Détection de comportements particuliers dans les trajectoires de visites des musées

Expérimentation au musée de la Cité du Vin :

Dans cette expérimentation, nous nous intéressions à la façon dont les visiteurs abordent le musée et notamment les choix pris au tout début de la visite. Le musée de la cité du vin est un musée dont l'architecture en "open space" permet aux personnes de déambuler librement dans celui-ci. Notre objectif est ici de savoir si un motif de trajectoires se dessine malgré l'absence de guidage dans cet espace. Pour ce faire, nous avons utilisé la description δ de préfix SCP sur les séquences de trajectoires simples avec un attribut de type séquence simple sur les modules. En se concentrant seulement sur le début de ces séquences de trajectoires, il nous est apparu que des motifs récurrents se dessinent. La figure 7.5 montre un zoom sur le diagramme de Hasse d'un treillis des concepts, avec 32 concepts, qui est le résultat d'une analyse lancée sur le dataset de **la cité du vin** en utilisant 100 000 trajectoires de visiteurs.

On observe que le concept \$0 *"M00"* est présent dans la quasi intégralité des trajectoires, ce qui est normal puisqu'il s'agit du module correspondant au lieu de la récupération du *compagnon de visite* - celles qui ne commencent pas par *"M00"* sont probablement des erreurs. Le treillis montre très clairement un motif dans les modules choisis par les visiteurs quand ils entrent dans le musée : le module *"M03"* présent dans 25 000 trajectoires et le module *"M02"* présent dans 22 000. Ainsi, plus de 50% des trajectoires commencent soit par la sous-séquence commune *"M00 - M03"* ou *"M00 - M02"*.

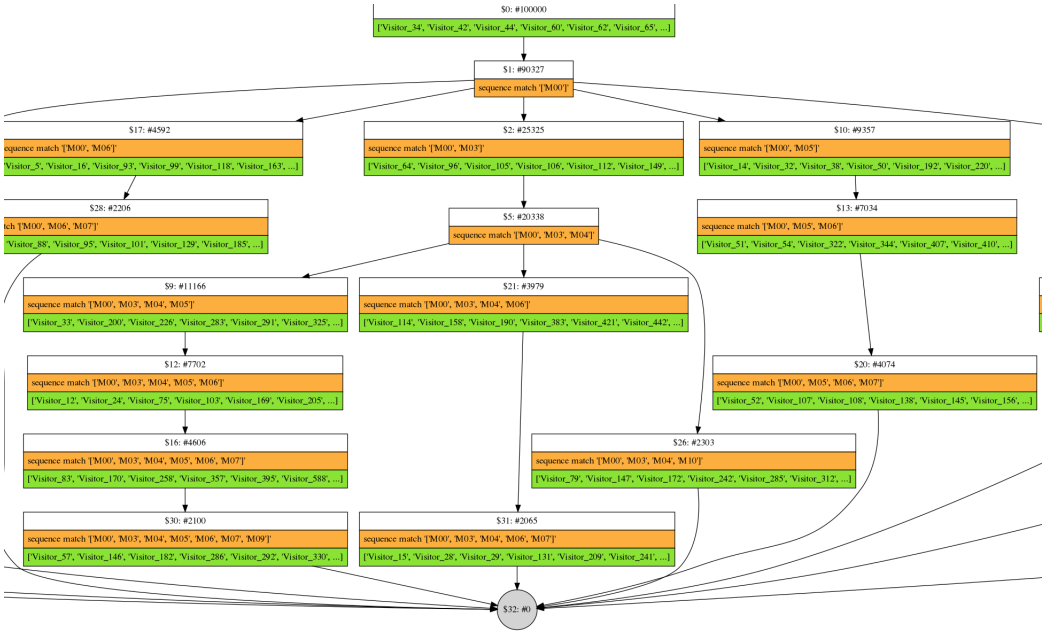


FIGURE 7.5 – Zoom sur le treillis des concepts avec le dataset de la **Cité du Vin**

À partir du treillis obtenu, on peut en déduire un arbre de décision présent figure 7.6 représentant les débuts de visite des personnes déambulant dans le musée. Ceci est dû au fait que la description **SCP** contient un seul prédicat qui est le préfixe maximal, puis la stratégie augmente ce préfixe maximal pour le niveau suivant. Nous obtenons ainsi un arbre de décision en supprimant l'élément $\top$ présent dans toutes nos trajectoires ("M00"). Basé sur les deux choix de départ, module *M02* ou module *M03*, cet arbre nous indique les choix des visiteurs au début de leurs trajectoires, avec le nombre de parcours concernés au centre des noeuds. Par le biais de cette visualisation, nous pouvons observer de longs débuts de trajectoires communs, comme $\langle M03, M04, M05, M06, M07, M09\rangle$ réalisés par 2100 individus (2% du dataset). Aussi, un détail important ressort par l'analyse de **Galactic** : Si un

visiteur "rate" un module durant sa visite, il n'y retournera pas. Ainsi, si un visiteur commence son parcours par le module *M03*, il ne retournera pas à *M02* et continuera à avancer, de même quand un visiteur passe du module *M02* à *M04*, il ne retournera pas sur *M03*. Il s'agit là d'une information intéressante pour le gestionnaire du musée.

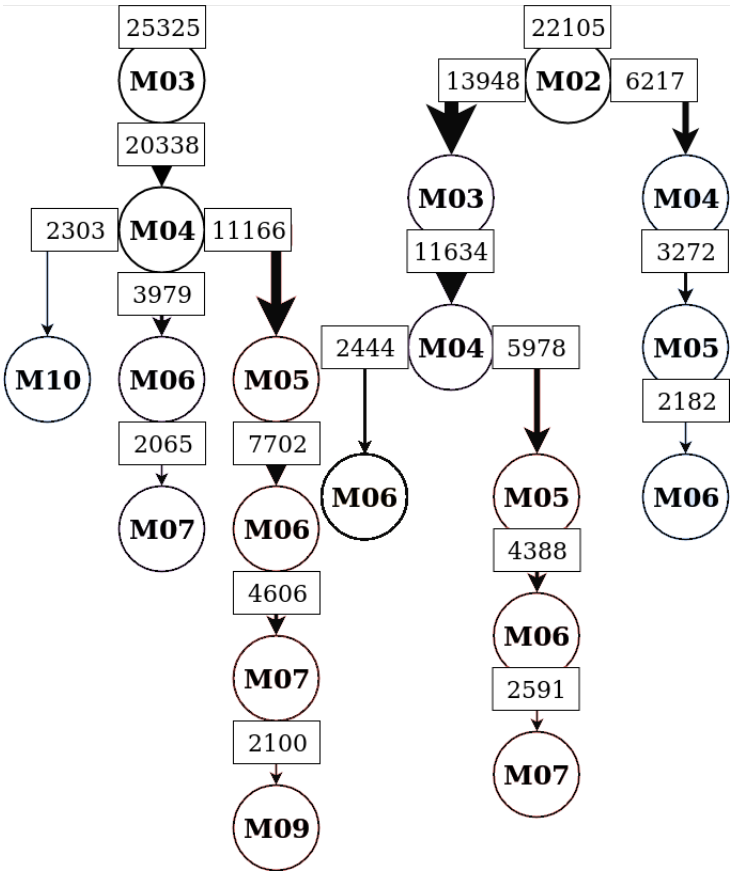


FIGURE 7.6 – Arbre de décision construit à partir du treillis des concepts de la figure 7.5

Expérimentation au museum d'histoire naturelle de La Rochelle (Museum_3) :

À partir des deux aspects *depl_oeuvre* et *appli_oeuvre*, on peut facilement calculer les durées et les concordances entre la visite et l'activité sur l'application mobile pour chaque œuvre, ce qui est une information intéressante pour les gestionnaires de musées. Avec **GALACTIC**, il est possible d'extraire toutes les concordances à l'aide d'une analyse multi-séquences sur les deux aspects, et d'en avoir une description précise. Dans cette expérimentation, nous avons utilisé la description SCMAX et la stratégie CMA pour les séquences temporelles afin d'extraire des sous-groupes d'individus ayant un comportement commun. Le treillis généré par **Galactic** est présent figure 7.7 et contient 49 concepts, le tableau 7.3 décrit trois des concepts générés. Le concept \$1 montre un lien entre la vue "Accueil de l'application" et la première salle de la visite où se situe l'âne Baudet. Au début d'une visite, 13 individus sont restés sur l'accueil de l'application durant au moins 12 minutes.

N°	Motif commun du sous-groupe	Nb Individus
1	*Visite* Baudet *Application* Accueil 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 t	13
2	*Visite* Baudet Cistude *Application* Accueil 0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 t	5
3	*Visite* Loutre *Application* Plan 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 t	3

TABLEAU 7.3 – Échantillon des concepts du treillis de la Fig 7.7

De ces 13 individus, le concept \$2 indique que 5 d'entre eux sont restés sur l'écran d'accueil de la visite et ont continué à explorer le musée puisqu'on peut les retrouver à la salle Cistude au milieu de leurs visites entre la 26e et la 28e minute. Cela veut probablement dire qu'ils n'ont tout simplement pas utilisé l'application et que celle-ci est restée sur la vue d'accueil.

Enfin, le concept \$3 quant à lui montre l'activation de la vue "plan" de l'application pendant quelques minutes après être passé devant la pièce dénommée "Loutre" que l'on retrouve dans trois des visites de notre dataset (sur 30 au total). Ce comportement est en effet en accord avec l'agencement

des salles du muséum : cette salle se positionne juste avant un escalier où il est possible de descendre ou de monter, et la consultation du plan pour savoir où l'on se rend prend tout son sens.

En définitive, cette analyse permet de dégager trois types de comportements en début de visite :

— Le sous-groupe A, les personnes restant longtemps à l'accueil avant de commencer la visite (au moins 7 minutes devant le baudet) et restant sur la vue "Accueil" de l'application.

— De ce sous-groupe A, se dégage le sous-groupe B. Celui-ci est constitué des personnes effectuant leur visite mais n'utilisant à priori pas l'application ou du moins, restant sur la vue "Accueil" de l'application les 30 premières minutes.

— Le troisième sous-groupe C est constitué des personnes consultant le plan afin de prendre la décision de monter à l'étage ou au contraire de descendre vers les salles au sous-sol.

En explorant ainsi tous les concepts générés, le gestionnaire est à même de mieux comprendre les déplacements des visiteurs dans son musée et de connaître les oeuvres qui les ont particulièrement intéressées, que ce soit au niveau du temps passé devant l'oeuvre, mais aussi sur l'application.

7.3.2 Détection de comportements particuliers dans les visites de la ville : Le choix des données

Une façon pour l'analyste de données d'analyser l'ensemble de données est de se concentrer sur des comportements particuliers. Pour illustrer ce type d'analyse, nous présentons deux études de comportement particulier :

— Un comportement lié à une information spatiale et une information contextuelle, avec les personnes allant à la plage

— Un comportement lié à une information temporelle, avec les lieux de séjour la nuit.

Les trajets à la plage en fonction de la météo

Dans un premier temps, nous ferons un focus sur les lieux de visite en prenant en compte les visites vers la plage. Afin de nous concentrer sur les visiteurs se rendant à la plage, nous choisissons d'analyser conjointement l'aspect *beach* avec l'aspect *weather* afin d'identifier les liens entre ces deux informations. Il est composé de 108 trajectoires, et génère 93 concepts en

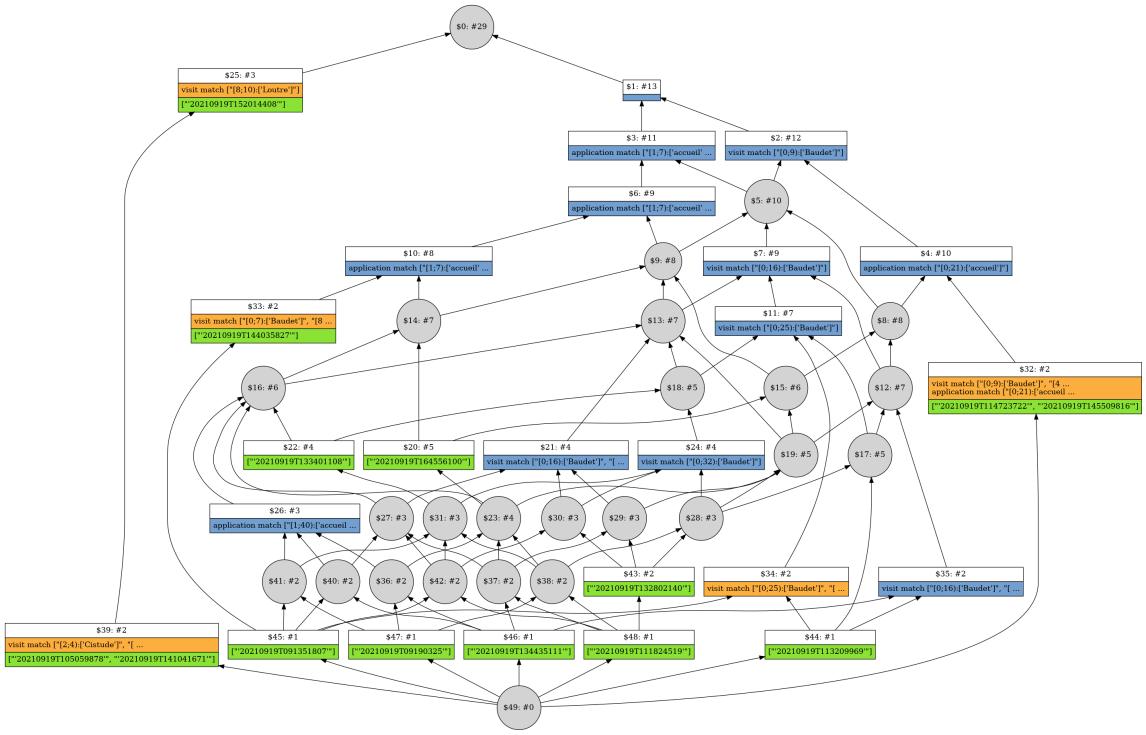


FIGURE 7.7 – Treillis des concepts du dataset **Museum_3**

utilisant la description SCMAX et la stratégie CMA pour l'aspect *beach* et *weather*. Le tableau 7.4 fournit un échantillon de certains concepts qui décrivent un comportement intéressant. La figure 7.8 montre l'extrait du treillis généré apportant cette observation. Le tableau 7.4 montre le support des prédicats parmi tous les prédicats dans la colonne support, et uniquement des prédicats contenant de la pluie pour la deuxieme colonne. La première ligne indique que le prédicat *Weather match [[11 :12) : raining], beach match [No Beach]* est présent dans 7% de nos données, et monte à 26% sur les trajectoires contenant de la pluie. On peut notamment observer une forte corrélation entre *"pluie"* et *"pas de plage"*, ce qui signifie que la majorité des touristes du dataset ne sont pas allés à la plage lorsqu'il pleuvait. Dans l'ensemble des données, 31 trajectoires contiennent la sous-séquence *pluie*.

Sur ces trajectoires, nous pouvons observer que 8 d'entre elles peuvent être décrites avec un prédicat qui correspond à "*pluie*" entre 11 et 12 heures et "*pas de plage*" dans séquence d'intervalle. D'autres prédicats sont également générés qui indiquent également que la sous-séquence *"pluie"* de l'aspect *weather* a un impact sur une visite à la plage, notamment le matin.

Motifs (prédicats communs)	Support	Support “raining"	Nb Individus
weather match$[(11;12)$: [raining]] beach match$[(0;24)$: [No beach]]	0.074	0.260	8
weather match$[(1;2)$: [raining]] beach match$[(0;24)$: [No beach]]	0.037	0.130	4
weather match$[(8;9)$: [raining]] beach match$[(0;24)$: [No beach]]	0.028	0.100	3
weather match$[(1;4)$: [raining]] beach match$[(0;24)$: [No beach]]	0.019	0.065	2

TABLEAU 7.4 – Échantillon des prédicats montrant l’impact de la météo sur les déplacements à la plage

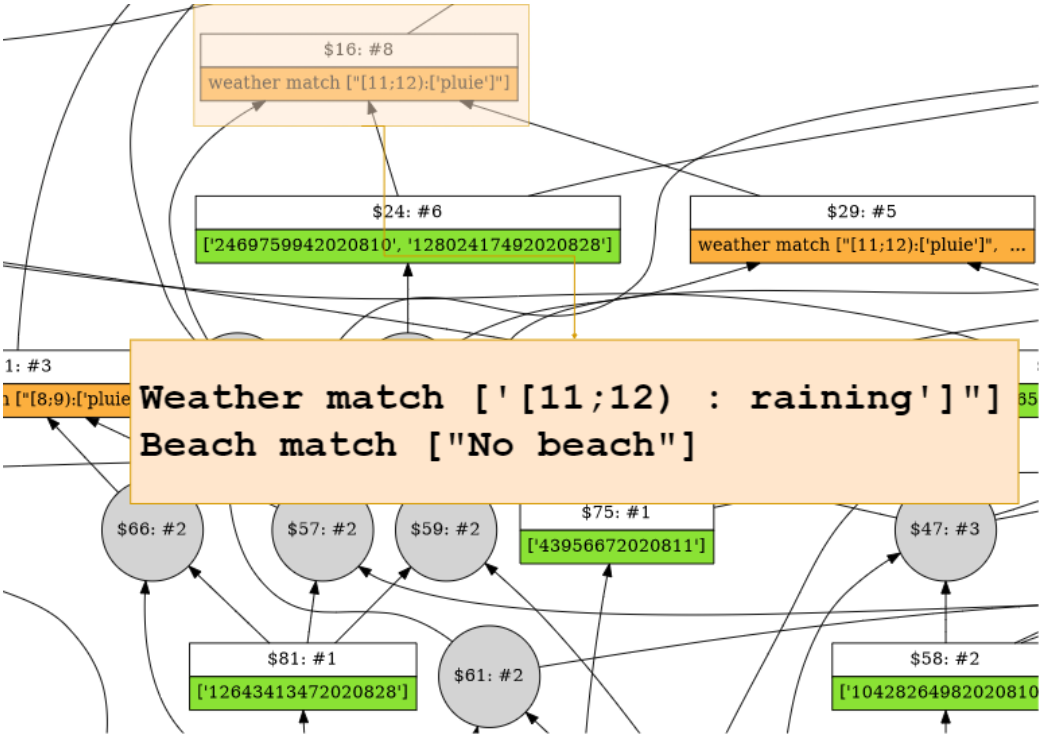


FIGURE 7.8 – Zoom sur la partie du treillis montrant l’impact de la pluie sur les trajets

Les habitudes de séjour suivant le profil d’une personne

Grâce aux informations temporelles, il est également possible d’analyser n’importe quel moment de la journée. Une analyse possible axée sur l’utilisateur pourrait être de savoir si le lieu de séjour la nuit dépend du *staying status*

des individus. Ici, nous effectuons un focus sur la temporalité, en étudiant les habitudes des touristes de nuit.

Pour ce faire, nous ferons une analyse avec deux attributs *staying status* qui est une chaîne et les lieux *district* qui sont des séquences temporelles. Pour ces deux aspects, nous avons utilisé la description SCMAX ainsi que la stratégie CMA sur l'aspect *district* et la description SCM ainsi que la stratégie SN pour l'aspect *staying status*. Nous obtenons au total 340 concepts, et 108 pour la période de sommeil.

Le tableau 7.9 décrit certains concepts obtenus. Comme on peut l'observer, des prédicats sont générés et sont soutenus par près de 10% des trajectoires "en famille" et plus de 10% des trajectoires "avec des amis" vers des sous-séquences de localisation, qui en les comparant avec d'autres prédicats de la super-séquence semblent représenter des lieux de séjour.

Conclusion de l'expérimentation

En conclusion de ces deux expérimentations, les données prises en compte dépendent de la nature du comportement que l'analyste souhaite mettre en avant. La prise en compte de la météo et des séjours à la plage ont permis de détecter un type de comportement. La recherche des lieux de séjours lors d'une analyse temporelle avec le statut de la personne a quant à elle permis d'extraire un autre comportement d'une nature différente. Ainsi, avec un même jeu de donnée, nous générons des analyses de différents types en sélectionnant les attributs que nous souhaitons étudier.

En choisissant ainsi les données qui peuvent être spatiales (la plage) ou temporelles (la nuit), l'analyste peut analyser un comportement particulier des visiteurs.

Motifs (prédicats communs)	Support	Support by status	Nb individus
'stay_status' match ('family',) *district* Oratoire 9 10 t	0.060	0.098	8
'stay_status' match ('family',) *district* Oratoire Oratoire 3 4 5 6 7 8 9 10 t	0.043	0.073	6
'stay_status' match ('family',) *district* Oratoire Oratoire 7 8 9 10 t	0.036	0.061	5
'stay_status' match ('family',) *district* Oratoire 9 10 11 12 13 t	0.036	0.061	5
'stay_status' match ('family',) *district* Oratoire Oratoire 1 2 3 4 5 6 7 8 9 10 t	0.029	0.049	4
'stay_status' match ('with friends',) *district* Fétilly 7 8 9 t	0.049	0.108	4

FIGURE 7.9 – Échantillon des prédicats montrant les quartiers de résidence touristique de séjour en fonction du type de statut

7.4 Conclusion

Dans ces travaux, nous proposons un moyen de traiter des données hétérogènes en utilisant l'algorithme NEXTPRIORITYCONCEPT en parallèle avec la plateforme **GALACTIC**. Dans les expériences, nous illustrons l'avantage de mélanger des données hétérogènes tout en conservant leur sémantique et leur lisibilité à l'aide de deux exemples. Tout d'abord, en réduisant le nombre de motifs en ajoutant des connaissances contextuelles sur les trajectoires et ensuite en se concentrant sur des aspects spécifiques des données pour détecter et identifier des comportements spécifiques. La spécificité de cette approche est de conserver les données brutes et d'éviter toute technique de vectorisation. De plus, elle nous permet d'enrichir le jeu de données avec des informations sémantiques provenant de l'espace et/ou des connaissances temporelles directement issues de la trajectoire (*districts*), des enquêtes remplies

par les individus (*staying status*), de bases de données en ligne (*weather*, *tide*) ou données d'application ("Visite Musée"). Le système de plugins de **GALACTIC** permet d'intégrer facilement des descriptions et des stratégies pour des données hétérogènes et offre un large éventail de types de données avec lesquelles travailler. NextPriorityConcept et sa capacité à exécuter des analyses interactives et hétérogènes est un grand pas dans la science des données et offre de nouvelles façons de traiter des structures de données complexes telles que les trajectoires sémantiques. Cependant, ce type de calcul et de processus n'est pas à la portée d'un utilisateur non-informaticien.

Dans le chapitre suivant nous présenterons une adaptation de l'algorithme NextPriorityConcept afin de créer un outil incrémental pour renforcer l'interactivité. L'objectif est de permettre à l'analyste des données de sélectionner l'aspect des données qu'il souhaite explorer, étant le seul qui a connaissance de la sémantique des données qu'il manipule. Par exemple, les informations *weather* et *marée* ne sont pas intéressantes pendant la nuit. Nous espérons qu'en donnant la possibilité à un expert d'explorer interactivement les données dans une approche orientée utilisateur, nous serons en mesure d'optimiser les processus d'exploration de données.

Chapitre 8

Contribution 4 : Vers une analyse interactive avec Reduced Context Completion

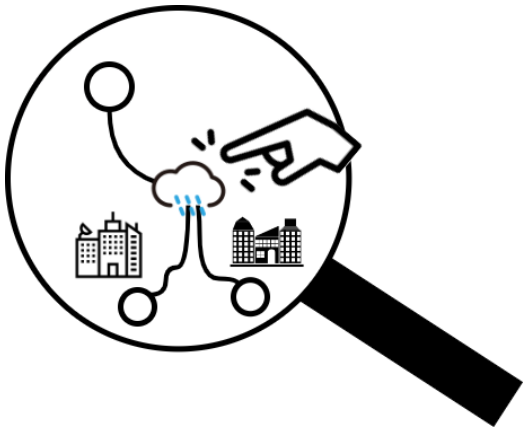


Table des matières

8.1 Introduction 159
8.2 Algorithme de complétion d'un contexte réduit 160
8.2.1 Formalisation de la problématique de complétion . . . 160
8.2.2 Description de l'algorithme 162
8.2.3 Exemple 167
8.3 Preuve de l'algorithme 169
8.4 Expérimentations 171
8.4.1 Construction itérative du treillis des concepts de la cité du vin 171
8.4.2 Comparaison avec l'algorithme NextPriorityConcept 174
8.4.3 Impact de l'ordre d'insertion 177
8.5 Conclusion et perspectives 181

8.1 Introduction

Dans l'objectif de repositionner les experts métiers au centre du processus d'analyse de leurs données, il est nécessaire de développer des solutions centrées sur l'expérience utilisateur ("user driven tools" en anglais). Ainsi, un utilisateur non-informaticien, ou non formé sur les domaines de l'AFC ou du pattern mining, doit tout de même pouvoir utiliser ces outils pour réaliser des analyses qui soient explicables. La motivation principale de ces travaux est de développer des outils permettant à l'utilisateur d'interagir directement pendant le processus d'analyse. Afin de répondre à cette problématique, nous avons conçu l'algorithme *"ReducedContextCompletion"*, qui sera l'objet de ce chapitre.

L'algorithme *"ReducedContextCompletion"* permet de construire un treillis de concepts en insérant des éléments de manière itérative. L'utilisateur peut ainsi choisir les concepts qu'il souhaite inclure dans son analyse, ce qui permet de guider le processus d'analyse. En maintenant seulement le contexte réduit d'un treillis, nous pouvons ainsi calculer l'impact d'une insertion de concept

dans celui-ci. Dans ce chapitre, nous allons nous placer dans un domaine beaucoup plus théorique que ce que nous avons vu jusqu'à maintenant, afin de garantir la validité d'une telle approche d'une part, et le bon fonctionnement d'une telle méthode d'autre part.

8.2 Algorithme de complétion d'un contexte réduit

8.2.1 Formalisation de la problématique de complétion

La validité de l'algorithme NextPriorityConcept repose sur des descriptions δ définies par l'utilisateur, qui correspondent à un opérateur de fermeture, et changer de description au cours de l'analyse impacterait la relation entre concepts et donc la structure même du treillis. Par contre, changer de stratégie en cours d'analyse n'impacte pas l'opérateur de fermeture et donc ne modifie pas les concepts préalablement générés.

L'utilisation de stratégies naïves sur toutes les données d'entrée génère le treillis de tous les concepts possibles, souvent volumineux, avec un grand nombre de concepts/motifs générés, ce qui positionne l'analyste devant la problématique du déluge de motifs. Des stratégies non naïves permettent de générer moins de concepts, et le treillis obtenu est alors un sous-treillis de ce treillis volumineux car l'opérateur de fermeture des descriptions est maintenu.

De façon plus formelle, considérons le treillis $L = (S, \varphi)$ correspondant au treillis volumineux pour des stratégies naïves, donné par un ensemble d'éléments S et un opérateur de fermeture φ. De façon équivalente, on peut considérer le treillis $L = (S, \vee, \wedge)$ donné par les opérateurs de bornes sup et bornes inf qui se déduisent de l'opérateur de fermeture.

Utiliser des stratégies non naïves consiste alors à construire le treillis $L_X = (X, \vee, \wedge)$ pour un sous-ensemble de concepts $X \subseteq S$ tout en maintenant les bornes sup et inf. Ce treillis $L_X = (X, \vee, \wedge)$ est alors un sous-treillis du treillis $L = (S, \vee, \wedge)$, et plus précisément le plus petit sous-treillis de L contenant X. Notre problématique est alors la suivante :

Problématique : Étant donné un ensemble $X \subseteq S$, calculer le plus petit sous-treillis L_X de $L = (S, \vee, \wedge)$ contenant X.

Il existe des algorithmes de génération incrémentale d'un treillis par ajout d'éléments dans le contexte dont il est issu, qu'ils soient des objets et/ou des attributs. Toutefois, à notre connaissance, il n'existe pas d'algorithme incré-

mental permettant d'ajouter successivement des concepts dans le treillis tout en maintenant ses bornes supérieures et inférieures.

On utilise le théorème fondamental de la théorie des treillis décrit dans la section 3.3.3 de l'état de l'art [Barbut and Monjardet, 1970] qui établit que tout treillis est isomorphe au treillis des concepts de sa table des irréductibles, encore appelée contexte réduit. Dans le cas non binaire, on étend ce théorème en prenant en compte les opérateurs join $\wedge$ et meet $\vee$ qui se déduisent de l'opérateur de fermeture. Nous aurons également besoin dans la suite des éléments top $\top$ et bottom $\bot$ du treillis. Nous considérons l'ensemble J_{L_X} des éléments join-irréductibles du plus petit treillis L_X contenant X, ainsi que l'ensemble M_{L_X} des éléments meet-irréductibles de ce même treillis. Nous notons par ailleurs $\top_{L_X}$ et $\bot_{L_X}$ les éléments top $\top$ et bottom $\bot$ de L_X. Cela nous permet de préciser notre problématique :

Problématique : Étant donné un ensemble $X \subseteq S$, calculer le contexte réduit $(J_{L_X}, M_{L_X}, \top_{L_X}, \bot_{L_X}, \vee, \wedge)$ du plus petit treillis L_X de $L = (S, \vee, \wedge)$ contenant X.

Le treillis L_X se retrouve alors par la construction du treillis des concepts :

$$L_X = CL(J_{L_X}, M_{L_X}, \top_{L_X}, \bot_{L_X}, \vee, \wedge)$$

L'enjeu de l'algorithme *"ReducedContextCompletion"* pour un treillis L et un sous-ensemble $X \subseteq S$, est de pouvoir reconstruire le treillis L_X qui est le plus petit sous-treillis de L contenant X. En ce sens, l'algorithme *"ReducedContextCompletion"* se rapproche des travaux de complétion d'ordres partiels.

Cependant, l'algorithme *"ReducedContextCompletion"* prend uniquement en compte le contexte réduit en maintenant les opérateurs de join $\vee$ et de meet $\wedge$ de L, réduisant ainsi drastiquement le nombre d'éléments à traiter. Dans ce chapitre nous définirons et formaliserons l'algorithme *"ReducedContextCompletion"* en section 1, puis nous détaillerons son fonctionnement et sa complexité en section 2. Ses performances et son comportement seront testés sur des jeux de données réels, discutés dans la troisième section et nous étudierons des cas d'utilisation d'un tel algorithme dans la section des expérimentations.

8.2.2 Description de l'algorithme

L'algorithme ReducedContextCompletion fonctionne de manière incrémentale pour chaque $x \in X$. Plus précisément, si on note $X = \{x_1, x_2, .., x_n\}$, l'algorithme calcule successivement le contexte réduit du plus petit treillis contenant x_1, puis du plus petit treillis contenant $\{x_1, x_2\}$ par ajout de x_2, puis de celui contenant $\{x_1, x_2, x_3\}$ par ajout de x_3, .. jusqu'au contexte réduit du plus petit treillis contenant $X = \{x_1, x_2, ...x_n\}$. Il s'agit donc, pour chaque x_i, de mettre à jour le contexte réduit calculé à l'étape précédente.

Ce processus s'initialise avec le treillis contenant un seul élément $\top = \bot = x_1$, et qui n'a pas d'irréductibles. Son contexte réduit, noté λ^{x_1}, est ainsi défini par :

$$\lambda^{x_1} = (\emptyset, \emptyset, \{x_1\}, \{x_1\}, \vee, \wedge)$$

L'algorithme 3 décrit l'algorithme ReducedContextCompletion qui prend en entrée un ensemble $X = \{x_1, x_2, \ldots, x_n\}$, ainsi que les opérateurs join $\wedge$ et meet $\vee$. Il renvoie le contexte réduit du plus petit treillis contenant X.

Les premières lignes d'initialisation avant la boucle principale calculent le contexte réduit λ^{x_1} : les ensembles J et M des sup-irréductibles et des inf-irréductibles sont initialisés à vide, le top $\top$ et le bottom $\bot$ sont initialisés avec x_1. À chaque itération i de la boucle, pour $2 \leq i \leq n$, l'algorithme calcule le contexte réduit $\langle \top_{x_i}, \bot_{x_i}, M_{x_i}, J_{x_i}, \wedge, \vee \rangle$ du plus petit treillis contenant $\{x_1, \ldots, x_n\}$ à partir du contexte réduit du plus petit treillis contenant $\{x_1, \ldots, x_{i-1}\}$, calculé lors de l'itération précédente. Pour cela, il met à jour les éléments irréductibles J_{x_i} et M_{x_i}, ainsi que les éléments $\top_{x_i}$ et $\bot_{x_i}$ calculés lors de l'itération précédente. Cette mise à jour se décompose en 4 parties, pour $x \in X/\{x_1\}$:

Élagage des éléments déjà présents dans le treillis [l. 1 - 3]

Les premières lignes sont là pour vérifier que x n'est pas déjà dans le treillis. Auquel cas, l'itération s'arrête.

Ajout d'un nouveau $\top$ ou d'un nouveau $\bot$ [l. 4 - 8]

Si $x > \top$, ceci signifie qu'il se place strictement au-dessus de $\top$ dans le treillis. L'insertion de x n'ajoutera donc pas de nouveaux irréductibles et n'engendrera pas d'autre élément que x dans le sous-treillis. Puisque l'insertion de x est strictement au-dessus de $\top$, x n'a qu'un seul prédécesseur et il est donc join-irréductible. Le $\top$ quant à lui devient meet-irréductible

Algorithm 3: REDUCEDCONTEXTCOMPLETION

Input: X un ensemble;
$\vee$ un opérateur de borne supérieure;
$\wedge$ un opérateur de borne inférieure;
Output: Le contexte réduit de L_X
$M \leftarrow \emptyset$ `// Les meet-irreducibles du treillis`
$J \leftarrow \emptyset$ `// Les join-irreducibles du treillis`
$\top \leftarrow x_1$
$\bot \leftarrow x_1$
for $x \in X$ **do**
  if x *in* J ***or*** x *in* M ***or*** $x == \top$ ***or*** $x == \bot$ **then**
    continue
  if $x > \top$ **then** `// Ajout d'un nouveau` $\top$
    $add(x)$ in J
    $add(\top)$ in M
    $\top \leftarrow x$
    continue
  if $x < \bot$ **then** `// Ajout d'un nouveau` $\bot$
    $add(x)$ in M
    $add(\bot)$ in J
    $\bot \leftarrow x$
    continue
  if $x \not\leq \top$ *or* $x \not\geq \top$ **then** `// Extension du maximum`
    add($\top$) in M
    $\top \leftarrow (\top \vee x)$
  if $x \not\leq \bot$ *or* $x \not\geq \bot$ **then** `// Extension du minimum`
    add($\bot$) in J
    $\bot \leftarrow (\bot \wedge x)$
  $add(x)$ in J
  $add(x)$ in M
  $C \leftarrow \{x\}$
  (`// liste des candidats irréductibles`)
  do
    INSERTIRREDUCIBLE($J, M, C.pop()$)
    UPDATEIRREDUCIBLE(J, M, C)
  while $C \neq \emptyset$
return $(J, M, \top, \bot, \vee, \wedge)$

Algorithm 4: InsertIrreducible

Input: J, M ;
x, provenant de $C.pop()$;
for e *in* $M \cup J$ **do**
    $r^- \leftarrow e \wedge x$
    $r^+ \leftarrow e \vee x$
    if $r^- \neq \bot$ **then**
        $C.add(r^-)$
    if $r^+ \neq \top$ **then**
        $C.add(r^+)$

Algorithm 5: UpdateIrreducible

Input: J, M, C
$M_{new} \leftarrow \emptyset$
$J_{new} \leftarrow \emptyset$
for e *in* $M \cup J \cup C$ **do**
    $e^+ = e$
    for $y \in M \cup J$ **do**
        if $y > e$ **then**
            $e^+ = e^+ \vee y$
    if $e \neq e^+$ **then**
        $M_{new}.add(e)$
        if e *not in* M **then**
            $C.add(e)$
    $e^- = e$
    for $y \in M \cup J$ **do**
        if $y < e$ **then**
            $e^- = e^- \wedge y$
    if $e \neq e^-$ **then**
        $J_{new}.add(e)$
        if e *not in* J **then**
            $C.add(e)$
$J \leftarrow J_{new}$
$M \leftarrow M_{new}$

car il possède maintenant x pour unique successeur. L'ajout d'un nouveau minimum $\bot$ suit cette même logique si x se place strictement en dessous de $\bot$ faisant de lui un nouveau join-irreducible et x est meet-irreducible. La recherche de nouveaux irréductibles n'est donc pas nécessaire et l'itération s'arrête.

Extension du maximum ou du minimum [l. 14 - 19]

Ici, nous savons que $x \not\geq \top$ et $x \not\leq \bot$, car sinon l'exécution s'arrêterait aux étapes précédentes. Si maintenant $x \not\leq \top$, cela signifie que x n'est pas comparable au maximum courant, c'est-à-dire qu'il n'est ni supérieur ni inférieur à celui-ci. Or, pour respecter la structure de treillis, celui-ci doit avoir un maximum. Le nouveau maximum $\top$ est alors la borne supérieure de x et $\top$. L'ancien maximum possède alors le nouveau maximum comme unique successeur et devient meet-irréductible. Ici, l'itération ne s'arrête pas, l'étape suivante est nécessaire. Si $x \not\geq \bot$, l'extension du minimum est similaire : le nouveau minimum $\bot$ est la borne inférieure de x et $\bot$, et l'ancien minimum possède alors le nouveau minimum comme unique prédécesseur et devient join-irréductible.

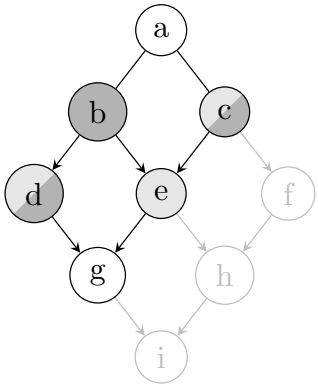


FIGURE 8.1 – Ajout de l'élément (c) qui fait apparaître a comme le nouvel élément $\top$

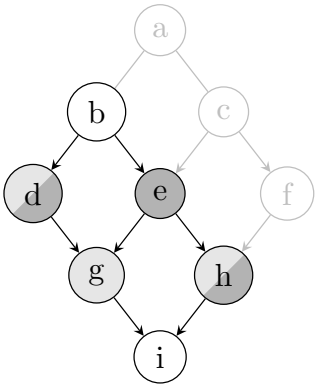


FIGURE 8.2 – Ajout de l'élément (h) qui fait apparaître i comme le nouvel élément $\bot$

Boucle d'exécution principale [l. 20 - 25]

Ici, le $\top$ et le $\bot$ ont été mis à jour et on sait que x est dans l'intervalle $[\bot, \top]$. Il s'agit alors de mettre à jour les ensembles d'irréductibles J et M.

On maintient pour cela une liste C de potentiels nouveaux irréductibles, initialisée avec x. Cette liste contiendra tous les nouveaux éléments détectés au cours des itérations qui sont de potentiels nouveaux irréductibles. Pour cela, tant que C n'est pas vide, les deux fonctions suivantes sont exécutées :

InsertIrreducible :

L'ajout de x dans le treillis fait apparaître de nouveaux éléments qui ne peuvent être que la borne supérieure entre x et un élément sup-irréductible, ou la borne inférieure entre x et un élément inf-irréductible, de par les propriétés des irréductibles dans un treillis. La fonction InsertIrreducible calcule la borne inférieure (r^-) et la borne supérieure (r^+) entre chaque élément irréductible e du treillis courant et x. Ainsi, x est dans l'intervalle $[e \wedge x, e \vee x]$. Les successeurs de $e \wedge x$ sont alors modifiés, ce qui en fait un potentiel nouveau inf-irréductible, sauf si $x \wedge e = \bot$. De façon similaire, $x \vee e$ est un potentiel nouveau irréductible, sauf si $x \vee e = \top$. La liste C est ainsi mise à jour avec $e \wedge x$ et $x \vee e$.

UpdateIrreducible :

La fonction UpdateIrreducible met à jour les irréductibles après ajout de x. Pour cela, elle initialise J_{new} et M_{new} à vide. Les nouveaux irréductibles peuvent être soit des anciens irrédutibles (donc dans $J \cup M$) soit de potentiels nouveaux irréductibles (donc dans C). On va tester chaque élément e de $J \cup M \cup C$ afin de déterminer s'il s'agit d'un nouveau meet irréductible et il sera alors ajouté à M_{new} ou d'un nouveau join irréductible et il sera alors ajouté à J_{new}. Il est nécessaire de mettre à jour les précédents éléments ajoutés par InsertIrreductible, pouvant perdre la priorité d'irréductible avec l'insertion de x. De cette façon, nous testons tous les éléments de $\{r_1^- \dots r_n^-\}$, $\{r^+ \dots r_n^+\}$, $\{x\}$, M et J afin de mettre à jour la table des irréductibles du treillis courant. À chaque fois qu'un nouvel irréductible est détecté, nous l'ajoutons également comme nouvel élément dans C.

Pour tester si e est dans M_{new} on teste si e était déjà dans le treillis. On a deux cas :

— soit e était dans M

— soit e était la borne inf de tous les éléments au-dessus de lui

On calcule pour cela $e^+ \leftarrow \bigvee\{y \mid y \in \{M \cup J\}\}$ et $y > e$:

— si $e = e^+$ alors e était une borne inf

— si $e \neq e^+$ et $e \in M$ alors e était un meet irréductible déjà présent dans le treillis, qui ne perd pas sa propriété d'irréductible sur cette itération car il a alors un seul successeur immédiat e^+.

— si $e \neq e^+$ et $e \notin M$, alors e est un nouvel élément donc ajouté à C comme potentiel élément irréductible et ajouté à M_{new} car il possède e^+ comme unique successeur.

Tant que C n'est pas vide et qu'aucun nouvel irréductible n'est découvert, nous recommençons les opérations ci-dessus. De nouveaux irréductibles peuvent aussi changer la structure du treillis.

L'algorithme de ReducedContextCompletion est directement implémenté au sein du framework **GALACTIC**. L'implémentation a été écrite en Python, et se situe dans la classe LATTICE, classe structurant le stockage des treillis. Celui-ci stocke les treillis en ne gardant que leurs contextes réduits.

8.2.3 Exemple

La figure 8.3 est une illustration de l'exécution de l'algorithme REDUCEDCONTEXTCOMPLETION. Le treillis initial $L = (S, \leq)$ y apparaît en arrière plan. REDUCEDCONTEXTCOMPLETION est exécuté pour $X \in \{b, e, d, c\} \in S$. Chaque itération de REDUCEDCONTEXTCOMPLETION considère successivement b,e,d puis c et calcule le contexte réduit du sous-treillis $L_{\{b\}}$ puis $L_{\{b,e\}}$ puis $L_{\{b,e,d\}}$ puis $L_{\{b,e,d,c\}}$. Chaque sous-treillis est représenté figure 8.3 et son contexte réduit est donnée dans le tableau 8.1.

Sous-treillis	J_L	M_L	$\top$	$\bot$
$L_{\{b\}}$	$\emptyset$	$\emptyset$	b	b
$L_{\{b,e\}}$	$\{b\}$	$\{e\}$	b	e
$L_{\{b,e,d\}}$	$\{d, e\}$	$\{d, e\}$	b	g
$L_{\{b,e,d,c\}}$	$\{d, e\}$	$\{d, c, b\}$	a	g

TABLEAU 8.1 – Liste des irréductibles, du maximum et du minimum de chaque sous-treillis de L pour $X = \{d, e, b, c\}$

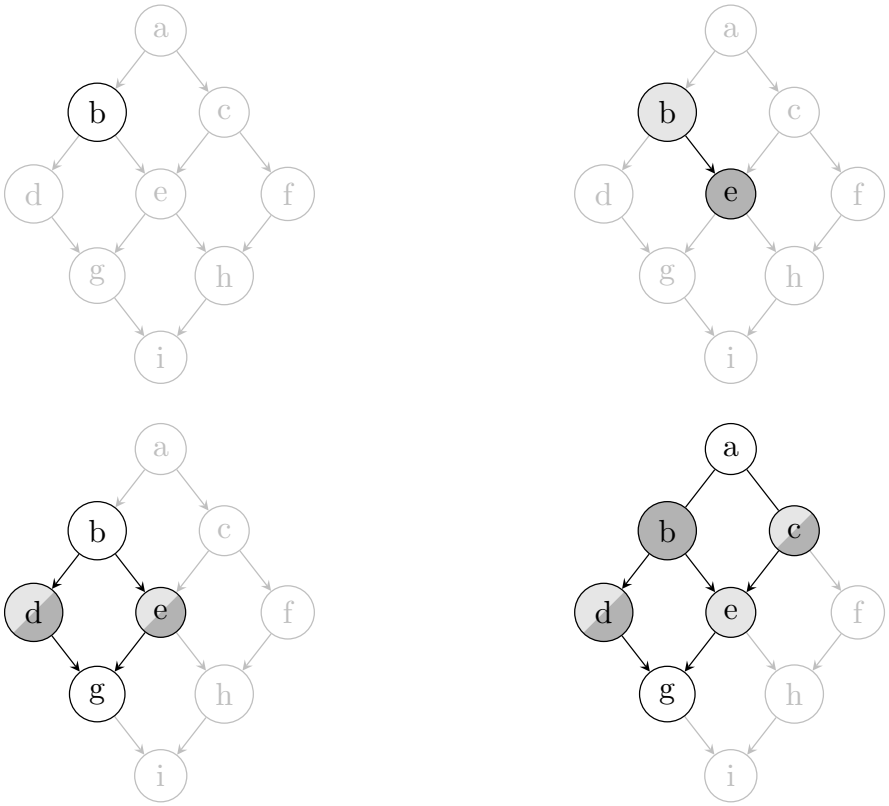


FIGURE 8.3 – Illustration de l'exécution de l'algorithme *ReducedContext-Completion*

8.3 Preuve de l'algorithme

Il s'agit de prouver le résultat suivant :

Théorème 3. *Soit un treillis $L = (S, \vee, \wedge)$ et un sous-ensemble $X \subseteq S$. L'algorithme* ReducedContextCompletion *calcule le contexte réduit du plus petit sous-treillis de L contenant X.*

Pour prouver ce théorème, on formalise le processus itératif de l'algorithme en introduisant la function κ^x qui, appliquée au contexte réduit d'un treillis L', renvoie le contexte réduit du treillis $L' + \{x\}$. Ainsi, l'algorithme ReducedContextCompletion calcule le contexte réduit du plus petit treillis contenant $X = \{x_1, x_2, .., x_n\}$ par itérations successives de la fonction κ :

$$\kappa^{x_n}(\kappa^{x_{n-1}}(\dots \kappa^{x_2}(\lambda^{x_1})))$$

où λ^{x_1} correspond au contexte réduit du treillis contenant le seul élément $\top = \bot = x_1$.

Nous introduisons également l'ensemble $\mathcal{FM}$ de tous les sous-treillis de $L = (S, \vee, \wedge)$. Cet ensemble est une famille de Moore, et peut donc s'organiser sous la forme d'un treillis (car toute famille de Moore est un treillis). Il s'agit du treillis de tous les sous-treillis reliés entre eux par inclusion de leur famille de Moore. L'élément maximal de ce treillis $(\mathcal{FM}, \subseteq)$ est le treillis L lui-même, et son élément minimal est le treillis vide. les atomes sont les treillis singletons λ^x pour tout $x \in S$. On définit $\alpha_{\mathcal{FM}}$ l'opérateur de fermeture de ce treillis $\mathcal{FM}$. Ainsi, le plus petit treillis contenant $X = \{x_1, x_2, .., x_n\}$ correspond à $\alpha_{\mathcal{FM}}(X)$.

Le Théorème 3 est ainsi une conséquence directe des deux lemmes suivants :

Lemme 1.

$$\mathcal{CL}(\kappa^{x_n}(\kappa^{x_{n-1}}(\dots \kappa^{x_2}(\lambda^{x_1})))) = \mathcal{CL}(J_{L_X}, M_{L_X}, \top_{L_X}, \bot_{L_X}, \wedge, \vee)$$

Démonstration. Appliqué pour $X = \{x_1, x_2, .., x_n\}$, il s'agit de montrer que κ calcule successivement, à chaque itération, le contexte réduit du plus petit sous-treillis de L contenant $\{x_1\}$, puis $\{x_1, x_2\}$, ... puis X. Ce qui permet de conclure que le contexte réduit obtenu est bien celui de $L_X = \alpha_{\mathcal{FM}}(X)$.
On note $X_i = \{x_1, x_2, .., x_i\}$ pour chaque itération x_i. On a alors $X_n = X$ à la dernière itération x_n. Soit $(J_{X_i}, M_{X_i}, \top_{X_i}, \bot_{X_i})$ le contexte réduit calculé à l'itération x_i. La preuve se fait par récurrence sur i.
Clairement, $(J_{X_1}, M_{X_1}, \top_{X_1}, \bot_{X_1})$ correspond à λ^{x_1} qui est le contexte réduit

du treillis singleton L_{x_1} contenant le seul élément $\top = \bot = x_1$.
Il s'agit donc de prouver que si $(J_{X_{i-1}}, M_{X_{i-1}}, \top_{X_{i-1}}, \bot_{X_{i-1}})$ est le contexte réduit de $L_{X_{i-1}}$, alors $(J_{X_i}, M_{X_i}, \top_{X_i}, \bot_{X_i})$ est le contexte réduit de L_{X_i}. Ce qui revient à prouver que κ calcule à l'itération x_i le plus petit treillis contenant $X_{i-1} \cup \{x_i\} = X_i$.
Clairement, les premières étapes de l'algorithme calculent bien les éléments $\top_{X_i}$ et $\bot_{X_i}$ de L_{X_i} suite à l'ajout de x_i dans $L_{X_{i-1}}$. Il s'agit donc de prouver que J_{X_i} et M_{X_i} correspondent bien aux irréductibles de L_{X_i}. A chaque itération, tous les nouveaux éléments détectés sont placés dans C, soit par ajout direct de x (il s'agit des éléments r^+ et r^-), soit par ajout indirect de x (il s'agit des éléments e^- et e^+).
Considérons alors $e \in M_{x_i}$ (la preuve est similaire pour $e \in J_{x_i}$). On a alors deux cas possibles :

— soit $e \notin M_{x_{i-1}}$: alors e est un nouvel élément détecté et placé dans C. Il s'agit alors de tester s'il s'agit d'un inf-irréductible, donc possédant un seul successeur immédiat. Pour cela, la borne inf e^+ de tous les irréductibles y qui lui sont supérieurs est calculée, car tout élément est une borne inf d'irréductibles, il suffit alors de tester si e^+ est différent de e, car dans ce cas e^+ est l'unique successeur de e et donc un inf-irréductible.

— soit $e \in J_{x_{i-1}}$: dans ce cas, e est testé de façon similaire.

Considérons maintenant par l'absurde qu'il existe un irréductible $e \notin J_{X_i}$. Alors e est un nouvel élément de L_{X_i} issu de l'ajout direct ou indirect de x_i. Il serait donc placé dans C puis testé, et potentiellement ajouté dans J_{X_i}. Donc une contradiction.
On conclut que $(J_{X_i}, M_{X_i}, \top_{X_i}, \bot_{X_i})$ est le contexte réduit du plus petit treillis contenant $X_{i-1} \cup \{x_i\} = X_i$, et donc que $\mathcal{CL}(X_{i-1} \cup \{x_i\} = X_i) = L_{X_i}$. □

Lemme 2.

$$\alpha_{\mathcal{FM}}(X = \{x_1, x_2, .., x_n\}) = \mathcal{CL}(\kappa^{x_n}(\kappa^{x_{n-1}}(\dots \kappa^{x_2}(\lambda^{x_1}))))$$

Démonstration. Ceci revient à montrer que κ agit ainsi comme un opérateur de fermeture sur les contextes réduits des sous-treillis de $\mathcal{FM}$, et donc que κ vérifie les propriétés d'un opérateur de fermeture, à savoir les trois propriétés suivantes qui sont des conséquences directes du lemme précédent, où on utilisera $\kappa(X)$ pour $\kappa^{x_n}(\kappa^{x_{n-1}}(\dots \kappa^{x_2}(\lambda^{x_1})))$:

— Opération extensive : avec $X \subseteq S$, on a clairement $X \subseteq \mathcal{CL}(\kappa(X))$. En effet, le treillis du contexte réduit $\kappa(X)$ contient nécessairement X.

— Opération idempotente : avec $X \subseteq S$, on a clairement $\mathcal{CL}(\kappa(\kappa(X))$ car réitérer deux fois la boucle principale sur le même contexte réduit n'apportera aucune modification.

— Opération isotone : avec $X \subseteq Y \subseteq S$, on a clairement $\mathcal{CL}(\kappa(X)) \subseteq \mathcal{CL}(\kappa(Y))$. En effet, en ajoutant progressivement les éléments de Y qui contient X, on obtient le contexte réduit de L_Y qui contient L_X.

□

8.4 Expérimentations

8.4.1 Construction itérative du treillis des concepts de la cité du vin

Nous allons utiliser l'algorithme de ReducedContextCompletion sur des trajectoires de déplacements de visiteurs de la **Cité du vin**. Pour ce faire, nous "injecterons" un à un les concepts / fermetures contenant les séquences de déplacements de visiteurs, afin d'observer la construction pas à pas du treillis. Nous utiliserons ici des séquences simples de déplacements, et la description δ_{SCM}. Le premier concept généré est le concept \$0 $(G, \delta(G))$ contenant tous les objets (51) et leurs sous-séquences communes avec δ_{SCM}, ici $\emptyset$.

Visitor_0 :

Nous commençons le traitement avec l'insertion de la fermeture du premier visiteur dont la séquence de déplacement est la suivante :

Visitor_0 : [M05, M06, M07, M09, M12, M14, M15, M20, M23, M99]

La figure 8.4 montre le concept \$1 du "*Visitor_0*", correspondant à sa fermeture. Il est à noter que ce concept contient deux individus, cela provient du fait qu'une autre séquence de visiteur est directement comprise dans la séquence de déplacement de "*Visitor_0*".

Visitor_1 :

La fermeture du deuxième visiteur est ensuite insérée dans le treillis de la figure 8.4. Celui-ci possède la séquence de déplacement suivante :

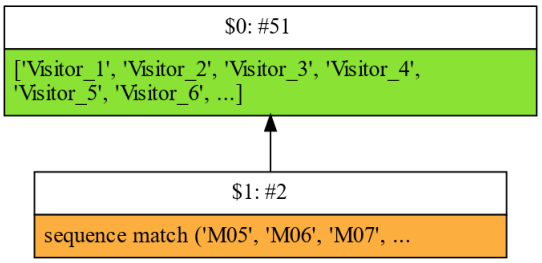


FIGURE 8.4 – Insertion d'un visiteur de la Cité du Vin

Visitor_1 : [M00, M04, M05, M06, M12, M14, M20, M23]

La figure 8.5 représentant le sous-treillis obtenu par REDUCEDCONTEXT-COMPLETION par l'insertion de ce visiteur, son concept est le concept $2. On voit apparaître deux nouveaux concepts : $3 dont la description est : *Séquence match [M05, M06, M12, M14, M20, M23]* qui correspond à la borne supérieure de $1 et $2 représentant les trajectoires de nos deux visiteurs - ie. leur sous-parcours commun et $4 = $1 ∨ $2.

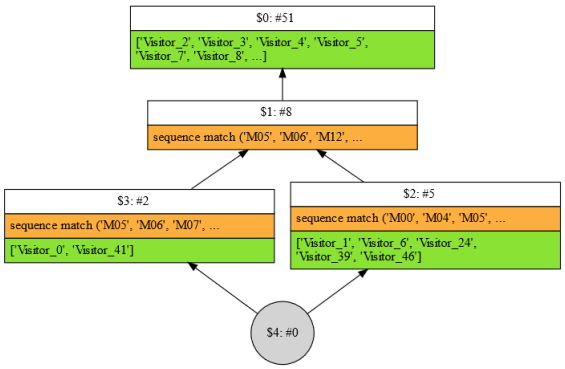


FIGURE 8.5 – Insertion d'un deuxième visiteur

Visitor_2 et Visitor_3 :

La figure 8.6 et la figure 8.7 montre les sous-treillis obtenus par l'algorithme après l'insertion de "Visitor_2" et "Visitor_3"

— *Visitor_2 : [M00, M02, M15, M99]*

— *Visitor_3 : [M00, M03, M04, M05, M07, M08, M09, M12, M15, M17, M19, M20, M23, M99]*

Les treillis générés font apparaître 9 puis 6 nouveaux concepts respectivement. L'émergence de concepts partagés par plusieurs individus est observée,

tels que*Chain match [M00]*, *Chain match [M15, M99]* ou *Chain match [M05, M12, M20, M23]*, qui sont des co-atomes.

Il est à noter que les concepts contenant un seul visiteur ce qui est le cas pour “Visitor_24”, “Visitor_39” ou “Visitor_41” sont des atomes. Cela signifie que les concepts insérés ont suffisamment d’informations pour pour faire apparaître le déplacement d’autres visiteurs présents dans notre espace de description.

Cette expérimentation illustre la construction itérative d’un treillis de concepts par injection itérative d’individus qui peuvent être choisis par l’utilisateur. Ces 4 treillis sont des sous-treillis du treillis généré par NextPriorityConcept sur l’ensemble des données, où le maintien des bornes supérieure et inférieure est rendu possible grâce à l’utilisation de l’opérateur de fermeture.

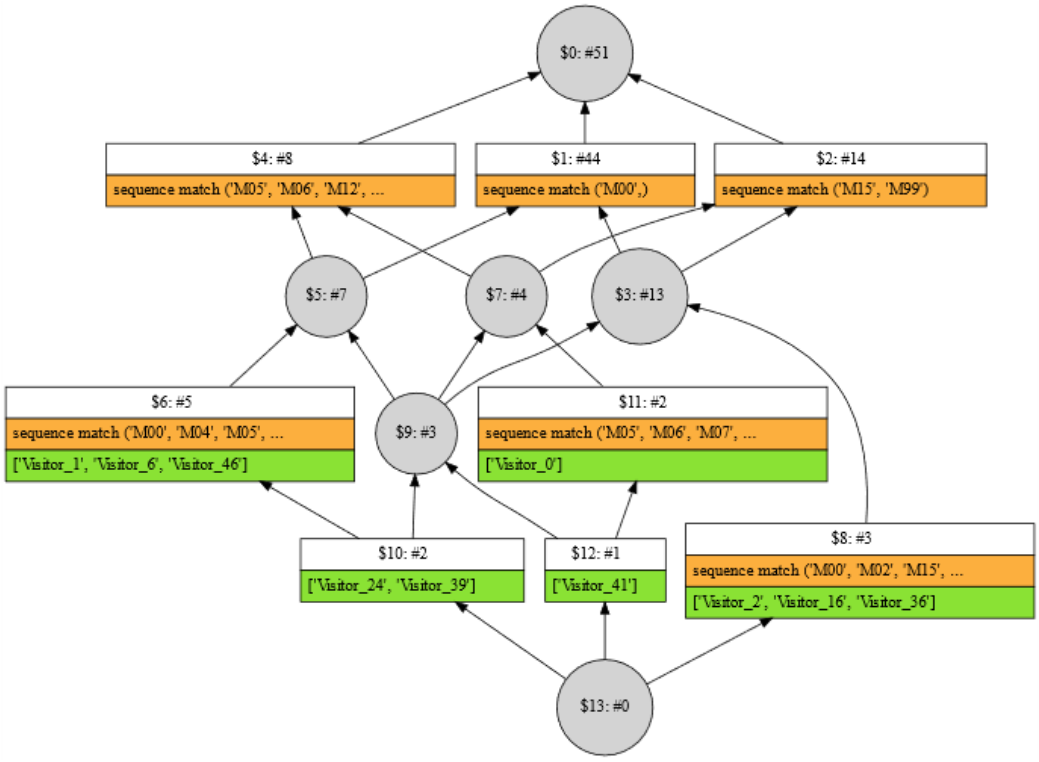


FIGURE 8.6 – Insertion d’un troisième visiteur

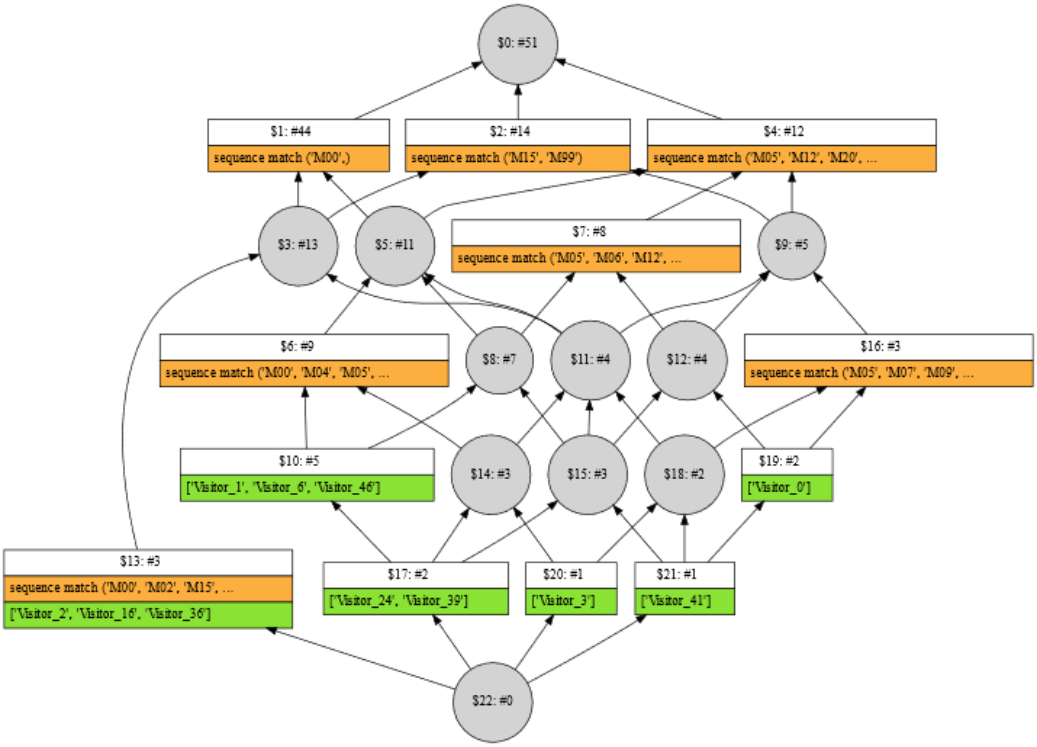


FIGURE 8.7 – Insertion d'un quatrième visiteur

8.4.2 Comparaison avec l'algorithme NEXTPRIORITYCONCEPT

Dans le but de démontrer l'avantage de la méthode consistant à conserver uniquement le contexte réduit d'un treillis, cette expérimentation vise à comparer les temps de calcul de l'algorithme REDUCEDCONTEXTCOMPLETION et de l'algorithme NEXTPRIORITYCONCEPT, sachant que le premier calcule le contexte réduit du treillis généré par le second. Des calculs seront effectués sur des paires de nombres entiers (x, y) générés aléatoirement (voir Listing 3).

Les deux algorithmes construiront un treillis en utilisant les descriptions DivisorDescription δ_{DD} et MultipleDescription δ_{DM} sur les deux caractéristiques x et y. La description δ_{DD} calcule les diviseurs communs d'un ensemble d'entiers, alors que la description δ_{DM} calcule leurs multiples communs.

NEXTPRIORITYCONCEPT utilise les stratégies DivisorStrategy σ_{DS} et MultipleStrategy σ_{MS} qui sont des stratégies naïves. La figure 8.8 présente un treillis de paires de nombres entiers obtenues avec ces descriptions et stra-

```
{
    ''x'': Nombre aléatoire [1 - 1000]
    ''y'' : Nombre aléatoire [1 - 1000]
}
```

Listing 3 – Exemple d'individu

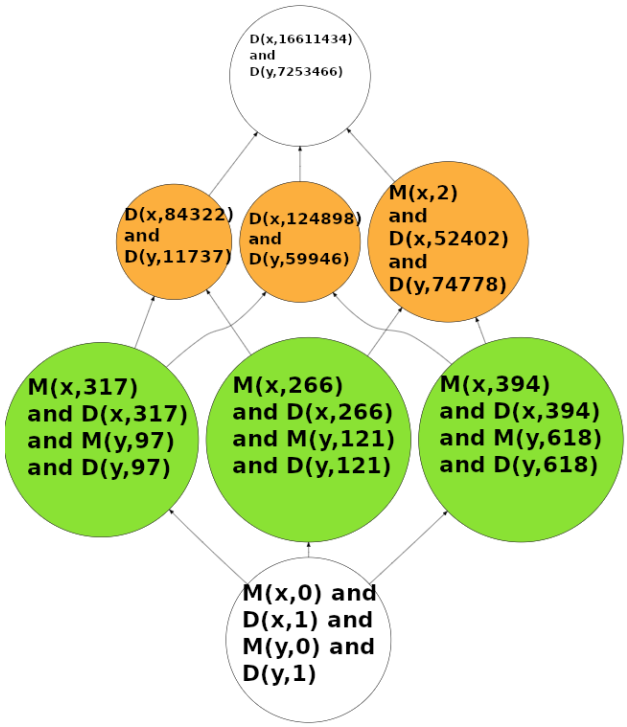


FIGURE 8.8 – Exemple de treillis de nombre entier

tégies.

Dans cet exemple, nous avons inséré les éléments $(x = 317, y = 97)$, $(x = 266, y = 121)$ et $(x = 394, y = 618)$. La borne inférieure de ces éléments est déterminée par un multiplicateur commun de 0 et un diviseur commun de 1. La borne supérieure entre les éléments $(x = 317, y = 97)$ et $(x = 266, y = 121)$ est définie par le diviseur commun des attributs x, soit 84322 qui est le diviseur commun de 317 et 121, ainsi que par le diviseur commun des attributs y, soit 11737 qui est le diviseur commun de 97 et 121. Cet élément ne possède pas de diviseur commun supplémentaire car ces deux nombres ne

sont pas divisibles, ce qui signifie qu'il est égal à 1 et n'est pas affiché.

La borne supérieure des éléments ($x = 266, y = 121$) et ($x = 394, y = 618$) est déterminée par le multiple commun de 2 pour l'attribut x, et le diviseur commun des attributs x est 52402, tandis que le diviseur commun des attributs y est 74778.

Enfin, l'élément $\top$ maximal correspond au diviseur commun 16611434 pour les attributs x, ce qui signifie qu'il est divisible par tous les nombres des attributs x inférieurs à lui, et le diviseur commun 7253466 pour les attributs y, ce qui signifie qu'il est divisible par tous les nombres des attributs y inférieurs à lui.

Le tableau 8.2 montre les temps de calcul de chaque algorithme suivant les paires de nombres entiers passées en entrée. L'algorithme NextPriorityConcept présente une évolution exponentielle des temps de calcul, tandis que ReducedContextCompletion maintient des valeurs stables qui peuvent varier en fonction du nombre de concepts ajoutés à chaque insertion de paires de nombres entiers. Cette évolution exponentielle du temps de calcul de NextPriorityConcept est directement liée au nombre de concepts générés par les descriptions et stratégies numériques, car il doit générer et prendre en compte l'entièreté des éléments dans la construction du treillis qui est de taille exponentielle en fonction des données. En revanche, ReducedContextCompletion se concentre sur une table binaire restreinte à un nombre limité de concepts (les éléments irréductibles), et cherche uniquement à générer le contexte réduit du plus petit treillis contenant la liste d'entiers passée en paramètre qui est linéaire en fonction des données. Cette différence explique les écarts de valeurs drastiques entre les deux approches. Il est à noter que l'algorithme ReducedContextCompletion utilise des opérateurs de calcul de bornes inférieures et bornes supérieures, dont la complexité est directement liée au type de la donnée. Ici, la nature spécifiquement numérique des individus permet un temps de traitement rapide dû à l'utilisation de fonctions à faible complexité dans l'espace de description associé, ce qui ne serait pas le cas pour des données plus complexes. Cependant, notre algorithme utilise une liste intermédiaire C d'éléments à traiter, qui, elle, peut contenir un nombre exponentiel d'éléments en cours de traitement.

NextPriorityConcept	ReducedContextCompletion	Nb
0.000575	0.000114	1
0.005807	0.002216	2
0.037843	0.020555	3
0.12473	0.079169	4
0.361836	0.19804	5
0.733962	0.431291	6
1.672523	0.758613	7
3.065295	1.330059	8
6.233383	2.141679	9
12.647243	5.457347	10
28.055211	3.490823	11
52.113587	3.118299	12
117.072309	4.55471	13
161.12263	8.75362	14
385.797701	2.510143	15
783.708289	2.714268	16
1443.523878	8.551827	17
3819.281459	2.833938	18
4255.70227	6.075568	19
16720.488369	3.079176	20
26594.559902	4.374208	21
58776.084291	3.276765	22
88377.481884	5.589516	23
98136.166157	4.699135	24

Tableau 8.2 – Comparaison des temps de calcul en secondes de NextPriorityConcept et ReducedContextCompletion

8.4.3 Impact de l'ordre d'insertion

Dans cette expérimentation nous utilisons le dataset "Iris flower data set"[1], un jeu de données populaire composé de 150 individus, décrits par cinq attributs : la longueur et la largeur des pétales, la longueur et largeur des sépals ainsi que l'espèce de la fleur (setosa, versicolor, virginica).

La longueur ainsi que la largeur des pétales / sépales sont des données numériques et l'espèce est une donnée de type catégoriel.

Nous utiliserons la description δ_{CH} Convex Hull qui s'applique non pas à

1. http ://archive.ics.uci.edu/ml/datasets/Iris

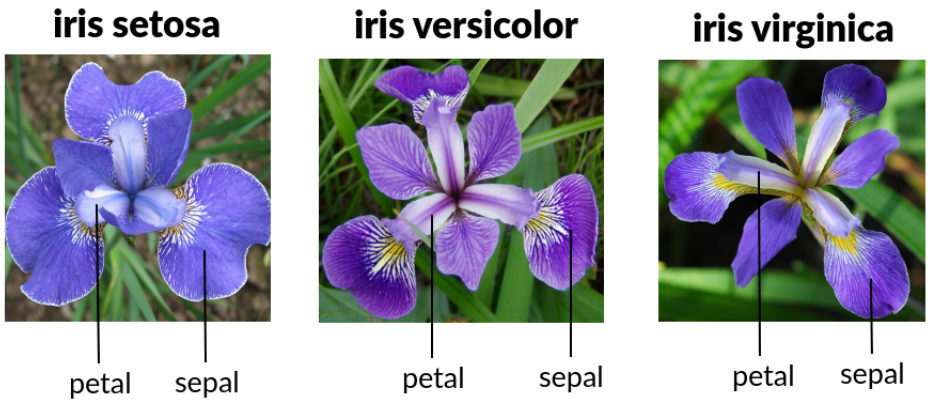


FIGURE 8.9 – Description du jeu de données des iris

Attribut	Caractéristique	Description
Largeur/Longueur pétale	Numérique	Numerical Hull
Largeur/Longueur sépale	Numérique	Numerical Hull
Espèce	Catégorielle	Categorical description

TABLEAU 8.3 – Tableau des descriptions et caractéristiques utilisées

une seule caractéristique mais à un ensemble de caractéristiques numériques. Elle calcule les bordures de leur enveloppe convexe. Les prédicats générés sont issus des droites représentant les bordures.

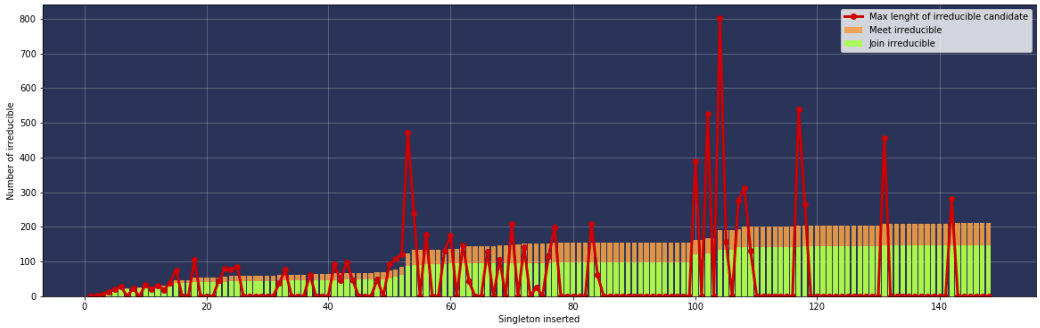
Ici, nous utilisons cette description δ_{CM} sur les deux caractéristiques numériques des pétales d'un côté, et sur les deux caractéristiques numériques des sépales de l'autre. Nous utilisons une description catégorielle pour l'espèce (cf Tableau 8.3).

Nos expérimentations portent sur l'importance de l'ordre d'insertion des éléments. Nous comparerons un ordre d'insertion classe par classe avec un ordre d'insertion aléatoire. On rappelle que les éléments insérés sont des concepts, donc des fermetures de singletons. Par singleton nous entendons une fermeture sur un seul élément. Ainsi, nous insérons une à une ces fermetures contenant un seul élément pour chaque individu de notre jeu de données, au nombre de 150. Il est à noter que certaines fermetures peuvent contenir plus d'un seul élément, c'est-à-dire qu'une fermeture sur un élément a pu aussi décrire un élément b (individu similaire). Nous avons mené deux expérimentations, une première sur l'évolution du nombre des irréductibles

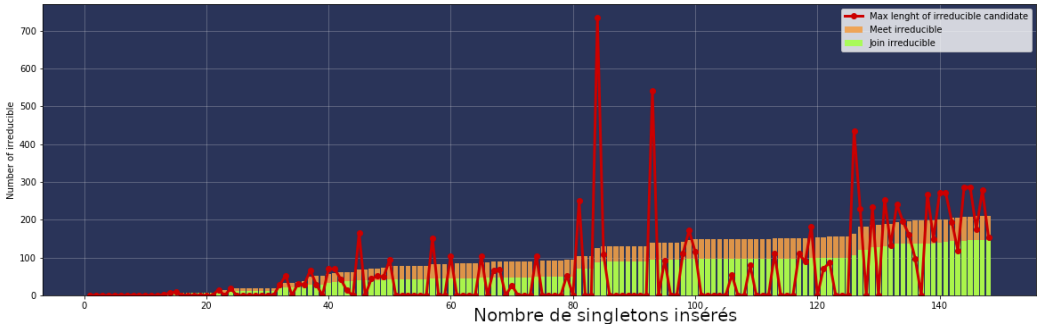
et des éléments intermédiaires de C et une deuxième sur les temps de calcul.

Évolution des irréductibles et des éléments intermédiaires

Pour cette première expérimentation, nous étudions l'impact de l'ordre d'insertion sur le nombre de meet, join et éléments candidats générés (C dans l'algorithme 3) quand un singleton est inséré.



(a) Ordre s'insertion : setosa - versicolor - virginica



(b) Ordre s'insertion : virginica - versitoclor - setosa

FIGURE 8.10 – Insertion de singletons classe par classe

La figure 8.10 montre deux expérimentations, où nous avons inséré les éléments classe par classe (setosa, versicolor puis virginica pour la figure 8.10a et inversement figure 8.10b). On remarque ainsi que le nombre d'irréductibles connaît une forte augmentation à trois reprises, ce qui correspond très probablement aux trois classes d'iris. En effet, l'insertion d'un individu d'une nouvelle classe augmente ainsi drastiquement le nombre d'éléments de la liste candidat.

À l'inverse, on observe dans les deux cas une progression lente des irréductibles "découverts" lors de la complétion de ce treillis, afin d'arriver au nombre de 211 meet irréductibles et 147 join irréductibles.

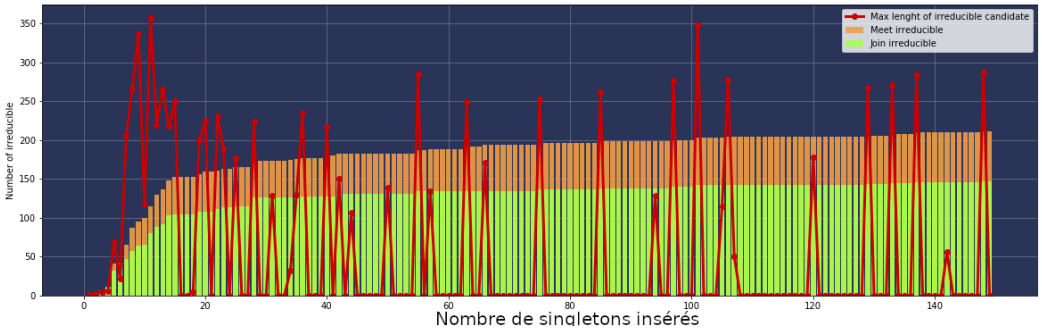


FIGURE 8.11 – Insertion de singletons dans un ordre aléatoire

La figure 8.11 décrit l'insertion de singletons dans un ordre "aléatoire". Ainsi, on peut observer que des irréductibles sont générés très tôt dans l'expérimentation. En effet, afin de décrire les nouveaux éléments qui n'ont pour la plupart rien de commun, il est nécessaire de construire un plus grand treillis dans lequel ils pourront exister. Ce faisant, on découvre rapidement un nombre d'irréductibles proche du nombre final dans le treillis complet des iris, limitant ainsi le nombre de fois où il est nécessaire de générer un nouveau sous-treillis contenant le nouvel élément.

En outre, cette expérimentation montre l'impact de l'information portée par un individu sur la structure d'un treillis. Un individu portant de nouvelles informations (comme la classe) qui n'ont pas encore été découvertes, générera un plus grand nombre d'éléments irréductibles afin de maintenir une structure de treillis capable de le décrire.

Le temps de calcul suivant l'ordre d'insertion :

Les courbes de la figure 8.12 montrent l'évolution du temps de calcul en fonction de l'ordre d'insertion des singletons à chaque itération. La courbe rouge correspond à une insertion classe par classe, quant à la courbe verte, elle correspond à une insertion aléatoire. La découverte du contexte réduit des iris complet est ainsi effectuée en 15 jours ($1.3 * 10^6$ secondes) avec une insertion classe par classe contre 11 ($9.8 * 10^5$ secondes) pour une insertion désordonnée. Ces courbes suivent celles du nombre d'éléments dans la liste candidat des figures 8.10 et 8.11.

Nous pouvons observer que l'évolution de ces courbes est similaire à celle du nombre d'éléments candidats traités par l'algorithme REDUCED-CONTEXTCOMPLETION, ce qui est cohérent avec le fait que la complexité de

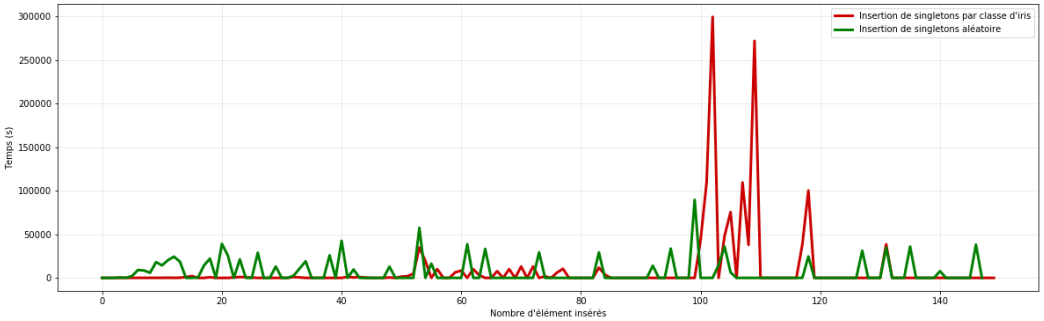


FIGURE 8.12 – Temps d'exécution de chaque boucle à l'insertion d'un élément suivant le mode d'insertion

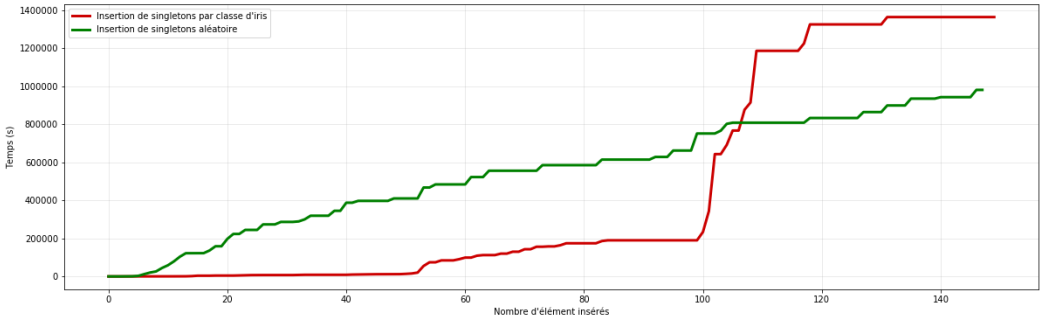


FIGURE 8.13 – Temps d'exécution total suivant le mode d'insertion

REDUCEDCONTEXTCOMPLETION dépend directement du nombre de candidats. Nous observons également 3 pics dans une insertion classe par classe, tandis que la courbe est plus régulière pour une insertion aléatoire, correspondant certainement aux 3 classes d'iris.

8.5 Conclusion et perspectives

À travers ce chapitre, nous avons exploré une façon de compléter le contexte réduit d'un treillis, à partir d'un ensemble d'individus. Nous avons d'abord décrit notre algorithme REDUCEDCONTEXTCOMPLETION et apporté la preuve de sa validité. La complétion est possible pour tout type de treillis et pas seulement pour des treillis de concepts.

Il est clairement plus avantageux de ne travailler que sur un contexte réduit en ne prenant en compte que les éléments irréductibles qui décrivent le treillis car leur taille reste linéaire en fonction des données, alors que celle du treillis est exponentielle. Le but de l'algorithme étant de découvrir de nou-

veaux éléments irréductibles en fonction de la donnée passée en entrée. Les expérimentations de ce chapitre confirment cette observation, et montrent l'impact de l'ordre d'insertion des éléments, l'ordre aléatoire étant à privilégier.

Le but de l'algorithme est de découvrir de nouveaux éléments irréductibles en fonction de la donnée passée en entrée, et les expérimentations de ce chapitre confirment qu'il est clairement plus avantageux de travailler sur un contexte réduit, en prenant en compte les seuls éléments irréductibles qui décrivent le treillis. La taille du contexte reste linéaire en fonction des données, alors que celle du treillis est exponentielle. De plus, ces expérimentations montrent l'impact de l'ordre d'insertion des éléments sur l'algorithme, l'ordre aléatoire étant à privilégier pour découvrir rapidement la plupart des éléments irréductibles du treillis final.

L'algorithme ReducedContextCompletion offre une nouvelle manière d'utiliser la théorie des treillis dans un processus d'analyse, plus proche de l'utilisateur. En effet, l'analyse formelle de concepts avait jusqu'alors pour but de générer un treillis, décrivant les relations et les informations communes dans un jeu de donnée de départ. Ici, nous ajoutons manuellement des concepts un à un, afin de dessiner progressivement les relations entre les concepts que l'on souhaite avant tout étudier.

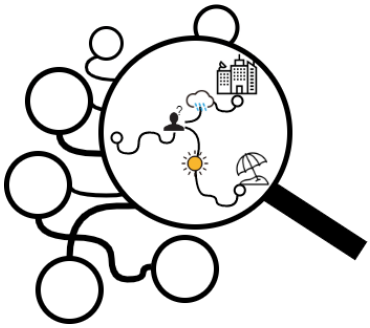
L'utilisation de l'algorithme montré dans la section 6, orientée vers une analyse guidée par l'utilisateur, ne révèle qu'une application partielle d'une telle méthode. En effet, la complétion du contexte réduit d'un treillis peut avoir une multitude d'applications pas seulement en analyse de déplacements. On peut envisager des applications : dans le domaine de la compression de bases de données, en stockant uniquement le contexte réduit du treillis et en le complétant lors d'une insertion ; dans la détection de concepts pertinents et représentatifs d'un jeu de données, en générant n treillis à partir d'échantillons aléatoires d'un dataset afin d'observer quels sont les concepts générés le plus fréquemment.

Chapitre 9

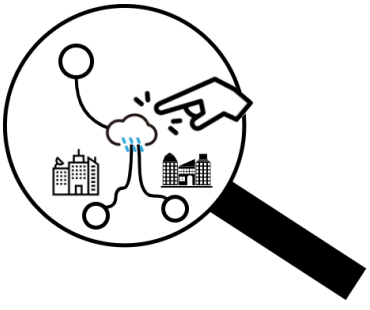
Conclusion

(a) Problématique de reconstruction de trajectoires en environnement contraint

(b) Problématique d'enrichissement sémantique des trajectoires



(c) Problématique d'analyse de trajectoires sémantiques hétérogènes



(d) Problématique d'interactivité dans l'analyse

FIGURE 9.1 – Les problématiques abordées dans ce manuscrit

9.1 Bilan

Au travers de nos travaux, nous avons construit une chaîne de traitement pour les trajectoires et proposé des solutions à plusieurs verrous scientifiques pour atteindre notre objectif, toutes représentées dans la figure 9.1 :

— **Problématique de reconstruction de trajectoires en environnement contraint.** (figure 9.1a)

— **Problématique d'enrichissement sémantique des trajectoires.** (figure 9.1b)

— **Problématique d'analyse de trajectoires sémantiques hétérogènes.** (figure 9.1c)

— **Problématique d'interactivité dans l'analyse.** (figure 9.1d)

Nous avons adopté une approche horizontale pour la reconstruction et l'analyse des trajectoires en intégrant des travaux provenant de différents domaines de recherche, en vue d'atteindre un objectif commun de reconstruction, d'enrichissement et d'analyse des déplacements des individus dans un objectif d'interactivité.

9.1.1 Reconstruction de trajectoires en environnement contraint

Afin de résoudre la question de la reconstruction de trajectoires en intérieur dans un environnement contraint comme les musées, nous avons mené une série d'expérimentations sur le terrain et effectué les évaluations sur des données réelles pour développer deux approches différentes. En effet, les dispositions idéales des capteurs ou émetteurs en grille par exemple sont souvent impossibles à mettre en place dans les environnements muséaux, et il était donc crucial de développer une approche qui puisse s'affranchir de ces contraintes. L'objectif de cette contribution scientifique était de fournir une méthode capable de surmonter les contraintes de placement de l'équipement de localisation en intérieur, qui est un défi majeur dans ce genre d'environnement. Nous avons proposer deux méthodes :

— Une première méthode à "gros grain" consiste à placer des balises dans chaque salle du musée, afin de pouvoir reconstruire toute la visite d'un individu, en fonction de l'architecture du musée et des transitions possibles entre les salles.

— La seconde, avec une granularité plus fine, calcule une estimation précise de la trajectoire d'un individu en prenant en compte le positionnement en intérieur et les contraintes de la topologie des lieux. Cette

estimation est calculée sous la forme d'une zone minimale d'activité à un instant t grâce à l'algorithme MZS.

9.1.2 Enrichissement sémantique des trajectoires

Afin de fournir une représentation plus complète et plus riche des trajectoires de visiteurs, nous avons proposé un modèle de trajectoires sémantiques qui intègre diverses sources de données de contexte, telles que les données provenant d'entretiens avec les individus, les données météorologiques et horaires des marées récupérées en ligne, ainsi que les données provenant d'une application mobile. Ce processus d'enrichissement a été appliqué sur des trajectoires réelles de parcours touristiques dans la ville de La Rochelle et de différentes visites au muséum d'histoire naturelle de La Rochelle.

Pour chacun des enrichissements, nous avons élaboré des visualisations afin de mettre en évidence les différents aspects de chaque trajectoire, permettant ainsi de comprendre de manière plus approfondie les différents contextes qui influencent la décision des visiteurs. En conclusion, ce modèle de trajectoires sémantiques offre une approche plus holistique pour la compréhension des parcours des visiteurs.

— Nous avons appliqué cette méthode à des séjours de touristes à La Rochelle dans un travail interdisciplinaire alliant informatique et géographie. L'enrichissement de trajectoires a permis, par le discours et par l'étude des **POI** aux alentours, de proposer une trajectoire complexe se reposant non plus seulement sur l'aspect spatial mais aussi contextuel et explicatif.

— Nous avons enrichi les trajectoires reconstituées de notre première contribution avec les données d'une application mobile, "Visite musée" afin de proposer un véritable journal de bord de l'activité réalisée durant la visite dans le musée, aussi bien des pièces visitées que des oeuvres qui ont pu marquer le visiteur, via l'utilisation de la vue de détails de l'application.

9.1.3 Analyse de trajectoires sémantiques

Nous avons utilisé l'Analyse Formelle de Concepts et l'algorithme NextPriorityConcept via la plateforme logicielle **GALACTIC** pour analyser les trajectoires enrichies décrites dans la contribution 2. Cette analyse a permis de découvrir des comportements similaires en créant une hiérarchie de

sous-groupes ayant des attributs communs. Les trajectoires utilisées dans cette contribution sont les trajectoires enrichies provenant de la contribution 2. L'approche hétérogène, qui combine plusieurs aspects des déplacements et des informations contextuelles, a montré la capacité de l'algorithme NextPriorityConcept à traiter efficacement des données complexes pour l'analyse de trajectoires sémantiques. En outre, nous avons effectué un certain nombre d'expérimentations sur plusieurs jeux de données, couvrant des déplacements à la fois en intérieur et en extérieur qui font émerger les comportements suivants :

— La fréquentation de la plage en fonction de la météo.

— Les habitudes de séjour selon la catégorie de touriste à La Rochelle.

— Les concordances de déplacement et d'utilisation de l'application des visiteurs du *Muséum d'Histoire Naturelle de La Rochelle.*

— Le début des visites communes du musée de la *Cité du Vin.*

9.1.4 Analyse interactive

Dans cette contribution, l'objectif était de développer un outil d'analyse facile à utiliser pour l'analyste des données, qui permettra de choisir les critères sur lesquels il souhaite focaliser son analyse. Le but est de proposer une méthode interactive d'exploration sémantique des données pour permettre à l'analyste de sélectionner la sémantique qui l'intéresse.

L'analyse de ce type de données et l'extraction de comportements significatifs nécessitent une expertise fine des données. L'idée consiste à personnaliser le traitement pour l'analyste des données afin qu'il puisse orienter les axes d'analyse en utilisant les aspects de la donnée qui l'intéressent pour un sous-groupe donné de visiteurs.

Pour ce faire, nous avons proposé un algorithme - ReducedContext-Completion - permettant de personnaliser la construction de cette même hiérarchie de sous-groupes de la contribution 3 en utilisant les notions fondamentales de la théorie des treillis. Cet algorithme met à jour la table binaire du treillis en fonction du sous-groupe inséré permettant une construction pas à pas du treillis en utilisant seulement la liste des meet-irreducibles et join-irreducibles. Ainsi, il est possible de personnaliser le traitement et de permettre à l'analyste des données d'orienter lui-même les axes d'analyse en utilisant les aspects des données qui l'intéressent. Axer les traitements en utilisant la sémantique même des données et permettre à l'utilisateur de choisir

l'angle d'approche du problème permet ainsi d'éviter des calculs souvent trop coûteux et parfois inutiles. En somme, notre travail vise à améliorer l'analyse de trajectoires en la rendant plus accessible et interactive pour l'analyste des données. Par ailleurs, permettre à l'analyste d'utiliser les données de façon interactive selon ses besoins s'inscrit dans une démarche de numérique responsable.

9.2 Perspectives

Les avancées réalisées dans les différents domaines de contribution permettent d'envisager des perspectives de recherche et applicatives. Nous présenterons ici les suites envisageables des différentes contributions de cette thèse.

Reconstruction de trajectoires et indoor positionning. Dans cette contribution, nous avons présenté nos travaux de recherche visant à estimer les itinéraires des visiteurs dans un musée. Ce travail est actuellement en cours d'intégration dans la conception d'un serious game où un robot sera utilisé en tant que maître du jeu. Les algorithmes développés dans ce projet serviront donc de base pour la réalisation du serious game. Par conséquent, notre étude représente une contribution importante pour la conception de systèmes interactifs ludiques basés sur des algorithmes d'estimation de la trajectoire.

Enrichissement sémantique. Il est envisageable de recourir à des techniques de traitement automatique du langage naturel pour fournir une amélioration automatisée basée sur les textes des entretiens réalisés, en utilisant notamment l'extraction de mots-clés. Cette approche permettrait de récupérer non seulement des informations sur les lieux visités lors des entretiens, mais également des éléments relatifs aux sentiments et aux avis des participants à l'égard de leur parcours, le tout de manière automatisée.

Analyse interactive pilotée par l'utilisateur. Les stratégies de NextPriorityConcept ne sont utilisées que pour réduire un concept afin de générer ses prédécesseurs, les sélecteurs qu'elles définissent ne sont pas conservés dans les descriptions des concepts, ni dans l'ensemble final des prédicats décrivant les données. Par conséquent, il est envisageable de choisir ou tester plusieurs stratégies pour chaque concept, dans une approche de découverte de motifs pilotée par l'utilisateur. Nous avons montré un premier axe permettant à un utilisateur de piloter l'analyse, mais cet axe de

recherche est actuellement en développement dans notre équipe afin de proposer une interface homme/machine facile d'utilisation pour la plateforme **GALACTIC**. L'utilisation de cette approche offre la flexibilité de modifier la stratégie d'analyse en cours, voire de se concentrer exclusivement sur un attribut particulier. Par exemple, il serait possible de se focaliser uniquement sur l'attribut "plage" lors de l'analyse des trajectoires touristiques enrichies. Cela permettrait d'approfondir l'exploration de cet aspect spécifique et d'obtenir des comportements plus précis sur les déplacements.

Complétion d'un contexte réduit d'un treillis. L'algorithme ReducedContextCompletion est un algorithme prometteur dont l'exploitation n'en est qu'à ses premiers pas. L'utilisation faite ici de la complétion du contexte réduit nous permet de travailler sur une analyse pilotée par l'utilisateur, mais ce principe peut être exporté dans d'autres domaines de recherche. Les propriétés d'un tel travail peuvent intéresser des chercheurs dans les bases de données par exemple, car il permet de ne stocker qu'une partie des données (les éléments irréductibles) pour pouvoir reconstruire le treillis final et supporte l'insertion d'autres données. Cet algorithme peut aussi encore être optimisé, la recherche autour de ce sujet reste à développer et n'est pas cantonnée à l'analyse de trajectoires. Ce travail fera l'objet d'une publication en cours.

Liste des publications

Reconstruction de trajectoires

RICHARD, Jérémy, KARELL, Bertet, et FAUCHER, Cyril, . Ble Based Indoor Positioning System and Minimal Zone Searching Algorithm (MZS) Applied to Visitor Trajectories within a Museum. Applied Sciences, 2021, vol. 11, no 13, p. 6107.

Enrichissement sémantique

CAYÈRÉ, Cécile, SALLABERRY, Christian, FAUCHER, Cyril, et al. Multilevel and multiple aspect semantic trajectory model : application to the tourism domain. ISPRS International Journal of Geo-Information, 2021, vol. 10, no 9, p. 592.

MONDO, Mélanie, CAYÈRÉ, Cécile and RICHARD Jérémy, Enrichir les traces GPS des visiteurs à partir de données qualitatives : enjeux et propositions, Atelier Inforsid 2021, Dijon France, jun 2021

Analyse multi-séquence

RICHARD, Jérémy, SAVARIT, Guillaume, BOUKHETTA, Salah Eddine, et al. Discover spatio-temporal cluster from trajectory data enhance by heterogeneous contextual knowledge using FCA and the NextPriorityConcept Algorithm, The 16th International Conference on Concept Lattices and Their Applications, Tallinn University of Technology, Estonia, jun 2022.

DEMKO, Christophe, BOUKHETTA, Salah Eddine, RICHARD, Jérémy, et al. GALACTIC : towards a generic and scalable platform for complex and heterogeneous data using Formal Concept Analysis, Workshop at the 16th International Conference on Concept Lattices and Their Applications, Tallinn University of Technology, Estonia, jun 2022.

BOUKHETTA, Salah Eddine, DEMKO, Christophe, BERTET, Karell, et al. Temporal sequence mining using fca and galactic. In : Graph-Based Representation and Reasoning : 26th International Conference on Conceptual Structures, ICCS 2021, Virtual Event, September 20–22, 2021, Proceedings 26. Springer International Publishing, 2021. p. 185-199.

BOUKHETTA, Salah Eddine, RICHARD, Jérémy, DEMKO, Christophe, et al. Interval-Based Sequence Mining Using FCA and the NextPriorityConcept Algorithm. In : FCA4AI@ ECAI. 2020. p. 91-102.

BERTET, Karell, DEMKO, Christophe, BOUKHETTA, Salah, et al. Analysis of Complex and Heterogeneous Data Using FCA and Monadic Predicates. In : Complex Data Analytics with Formal Concept Analysis. Cham : Springer International Publishing, 2021. p. 75-103.

Bibliographie

Heba Abdelnasser, Reham Mohamed, Ahmed Elgohary, Moustafa Farid Alzantot, He Wang, Souvik Sen, Romit Roy Choudhury, and Moustafa Youssef. Semanticslam : Using environment landmarks for unsupervised indoor localization. *IEEE Transactions on Mobile Computing*, 15(7) :1770–1782, 2015.

Imad Afyouni, Cyril Ray, and Claramunt Christophe. Spatial models for context-aware indoor navigation systems : A survey. *Journal of Spatial Information Science*, 1(4) :85–123, 2012.

Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14, 1995.

James F Allen. An interval-based representation of temporal knowledge. In *IJCAI*, volume 81, pages 221–226. Citeseer, 1981.

Luis Otavio Alvares, Vania Bogorny, Bart Kuijpers, Jose Antonio Fernandes de Macedo, Bart Moelans, and Alejandro Vaisman. A model for enriching trajectories with semantic geographical information. *Proc. of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems. ser. GIS '07. ACM, 2007*, page 22 :1–22 :8, 2007.

Gennady Andrienko, Natalia Andrienko, and Marco Heurich. An event-based conceptual model for context-aware movement analysis. *International Journal of Geographical Information Science*, 25(9) :1347–1370, 2011. doi : 10.1080/13658816.2011.556120. URL `https://doi.org/10.1080/13658816.2011.556120`.

Jay Ayres, Jason Flannick, Johannes Gehrke, and Tomi Yiu. Sequential pattern mining using a bitmap representation. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 429–435, 2002.

Miriam Baglioni, José Antônio Fernandes de Macêdo, Chiara Renso, Roberto Trasarti, and Monica Wachowicz. Towards semantic interpretation of movement behavior. In *Advances in GIScience : Proceedings of the 12th AGILE Conference*, pages 271–288, Berlin, Heidelberg, 2009a. Springer, Springer Berlin Heidelberg. ISBN 978-3-642-00318-9.

Miriam Baglioni, José Antônio Fernandes de Macêdo, Chiara Renso, Roberto Trasarti, and Monica Wachowicz. Towards semantic interpretation of movement behavior. In *Advances in GIScience : Proceedings of the 12th AGILE Conference*, pages 271–288, Berlin, Heidelberg, 2009b. Springer, Springer Berlin Heidelberg. ISBN 978-3-642-00318-9.

M Barbut and B Monjardet. Ordre et classification, algèbre et combinatoire, paris, hachette, 1970. *Zbl0267*, 6001, 1970.

Alberto Belmonte-Hernández, Gustavo Hernández-Peñaloza, David Martín Gutiérrez, and Federico Alvarez. Swiblux : Multi-sensor deep learning fingerprint for precise real-time indoor tracking. *IEEE Sensors Journal*, 19(9) :3473–3486, 2019.

Gennady Berkovich. Accurate and reliable real-time indoor positioning on commercial smartphones. In *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 670–677. IEEE, 2014.

Karell Bertet and Bernard Monjardet. The multiple facets of the canonical direct unit implicational basis. *Theoretical Computer Science*, 411(22-24) : 2155–2166, 2010.

Garrett Birkhoff. *Lattice theory*, volume 25. American Mathematical Soc., 1940.

Vania Bogorny, Chiara Renso, Artur Ribeiro de Aquino, Fernando de Lucca Siqueira, and Luis Otavio Alvares. Constant–a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS*, 18(1) :66–88, 2014a.

Vania Bogorny, Chiara Renso, Artur Ribeiro de Aquino, Fernando de Lucca Siqueira, and Luis Otavio Alvares. Constant – a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS*, 18(1) :66–88, 2014b. doi : https://doi.org/10.1111/tgis.12011. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12011`.

Francesco Bonchi, Fosca Giannotti, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Roberto Trasarti. Conquest : a constraint-based

querying system for exploratory pattern discovery. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 159–159. IEEE, 2006.

Jean-Paul Bordat. Calcul pratique du treillis de galois d'une correspondance. *Mathématiques et Sciences humaines*, 96 :31–47, 1986.

Christian Borgelt. Frequent item set mining. *Wiley interdisciplinary reviews : data mining and knowledge discovery*, 2(6) :437–456, 2012.

Léon Bottou and Chih-Jen Lin. Support vector machine solvers. *Large scale kernel machines*, 3(1) :301–320, 2007.

Salah Eddine Boukhetta. *Analyse de séquences avec GALACTIC – Approche générique combinant analyse formelle des concepts et fouille de motifs.* PhD thesis, La Rochelle Université, 2022.

Salah Eddine Boukhetta, Christophe Demko, Jérémy Richard, and Karell Bertet. Sequence mining using fca and the nextpriorityconcept algorithm. In *CLA*, pages 209–222, 2020a.

Salah Eddine Boukhetta, Jérémy Richard, Christophe Demko, and Karell Bertet. Interval-based sequence mining using fca and the nextpriorityconcept algorithm. In *FCA4AI@ ECAI*, pages 91–102, 2020b.

Suzana J Camargo, Andrew W Robertson, Scott J Gaffney, Padhraic Smyth, and Michael Ghil. Cluster analysis of typhoon tracks. part i : General properties. *Journal of Climate*, 20(14) :3635–3653, 2007.

Vicente Cantón Paterna, Anna Calveras Auge, Josep Paradells Aspas, and Maria Alejandra Perez Bullones. A bluetooth low energy indoor positioning system with channel diversity, weighted trilateration and kalman filtering. *Sensors*, 17(12) :2927, 2017.

Huiping Cao, Nikos Mamoulis, and David W Cheung. Mining frequent spatio-temporal sequential patterns. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE, 2005.

Cécile Cayèré. *Modélisation de trajectoires sémantiques et calcul de similarité intégrés à un ETL.* PhD thesis, La Rochelle Université, 2022.

Yi Chen, Peisen Yuan, Ming Qiu, and Dechang Pi. An indoor trajectory frequent pattern mining algorithm based on vague grid sequence. *Expert Systems with Applications*, 118 :614–624, 2019.

Min-Seok Choi and Beakcheol Jang. An accurate fingerprinting based indoor positioning algorithm. *Int. J. Appl. Eng. Res*, 12(1) :86–90, 2017.

Giorgio Conte, Massimo De Marchi, Alessandro Antonio Nacci, Vincenzo Rana, Donatella Sciuto, et al. Bluesentinel : a first approach using ibeacon for an energy efficient occupancy detection system. In *BuildSys@ SenSys*, pages 11–19. Citeseer, 2014.

Christophe Demko, Karell Bertet, Cyril Faucher, Jean-François Viaud, and Sergei O Kuznetsov. Nextpriorityconcept : A new and generic algorithm computing concepts from complex and heterogeneous data. *Theoretical Computer Science*, 845 :1–20, 2020.

Christophe Demko, Salah Eddine Boukhetta, Jérémy Richard, Guillaume Savarit, Karell Bertet, Cyril Faucher, and Damien Mondou. Galactic : towards a generic and scalable platform for complex and heterogeneous data using formal concept analysis. In *Workshop ETAFCA'2022 - Existing Tools and Applications for Formal Concept Analysis in conjunction with the 16th International Conference on Concept Lattices and Their Applications*, Tallinn University of Technology, Estonia, jun 2022.

Thai-Mai Thi Dinh, Ngoc-Son Duong, and Kumbesan Sandrasegaran. Smartphone-based indoor positioning using ble ibeacon and reliable lightweight fingerprint map. *IEEE Sensors Journal*, 20(17) :10283–10294, 2020.

Qian Dong and Waltenegus Dargie. Evaluation of the reliability of rssi for indoor localization. In *2012 International Conference on Wireless Communications in Underground and Confined Areas*, pages 1–6. IEEE, 2012.

Christophe Dousson and Thang Vu Duong. Discovering chronicles with numerical time constraints from alarm logs for monitoring dynamic systems. In *IJCAI*, volume 99, pages 620–626. Citeseer, 1999.

Zahid Farid, Rosdiadee Nordin, and Mahamod Ismail. Recent advances in wireless indoor localization techniques and system. *Journal of Computer Networks and Communications*, 2013, 2013.

Renato Fileto, Cleto May, Chiara Renso, Nikos Pelekis, Douglas Klein, and Yannis Theodoridis. The baquara2 knowledge-based framework for semantic enrichment and analysis of movement data. *Data and Knowledge Engineering*, 98 :104–122, 2015. ISSN 0169-023X. doi : https://doi.org/10.1016/j.datak.2015.07.010. URL `https://www.sciencedirect.com/science/article/pii/S0169023X15000555`. Research on conceptual modeling.

Frédéric Flouvat. *Extraction de motifs spatio-temporels : co-localisations, séquences et graphes dynamiques attribués*. PhD thesis, Université de la Nouvelle-Calédonie, 2019.

Christian Freksa. Temporal reasoning based on semi-intervals. *Artificial Intelligence*, 54(1-2) :199–227, March 1992. ISSN 00043702. doi : 10.1016/0004-3702(92)90090-K. URL `https://linkinghub.elsevier.com/retrieve/pii/000437029290090K`.

Bernhard Ganter and Sergei O Kuznetsov. Pattern structures and their projections. In *International conference on conceptual structures*, pages 129–142. Springer, 2001.

Bernhard Ganter and Rudolf Wille. Contextual attribute logic. In William M. Tepfenhart and Walling Cyre, editors, *Conceptual Structures : Standards and Practices*, pages 377–388, Berlin, Heidelberg, 1999a. Springer Berlin Heidelberg. ISBN 978-3-540-48659-6.

Bernhard Ganter and Rudolf Wille. *Formal concept analysis : mathematical foundations*. Springer Science Business Media, 1999b.

Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 330–339, 2007.

Romeo Giuliano, Gian Carlo Cardarilli, Carlo Cesarini, Luca Di Nunzio, Francesca Fallucchi, Rocco Fazzolari, Franco Mazzenga, Marco Re, and Alessandro Vizzarri. Indoor localization system based on bluetooth low energy for museum applications. *Electronics*, 9(6) :1055, 2020.

Antonio Gomariz, Manuel Campos, Roque Marin, and Bart Goethals. Clasp : An efficient algorithm for mining frequent closed sequences. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 50–61. Springer, 2013.

Thomas Guyet and René Quiniou. Mining temporal patterns with quantitative intervals. In *2008 IEEE International Conference on Data Mining Workshops*, pages 218–227. IEEE, 2008.

Thomas Guyet and René Quiniou. Extracting temporal patterns from interval-based sequences. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011a.

Thomas Guyet and René Quiniou. Extracting temporal patterns from interval-based sequences. In *International Join Conference on Artificial Intelligence (IJCAI)*, 2011b.

Jiawei Han, Jian Pei, Behzad Mortazavi-Asl, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. Freespan : frequent pattern-projected sequential pattern mining. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 355–359, 2000.

Jiawei Han, Jian Pei, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Meichun Hsu. Prefixspan : Mining sequential patterns efficiently by prefix-projected pattern growth. In *proceedings of the 17th international conference on data engineering*, pages 215–224. Citeseer, 2001.

Suining He, Tianyang Hu, and S-H Gary Chan. Contour-based trilateration for indoor fingerprinting localization. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, pages 225–238, 2015.

Minh Tu Hoang, Yizhou Zhu, Brosnan Yuen, Tyler Reese, Xiaodai Dong, Tao Lu, Robert Westendorp, and Michael Xie. A soft range limited k-nearest neighbors algorithm for indoor localization enhancement. *IEEE Sensors Journal*, 18(24) :10208–10216, 2018.

Frank Höppner and Frank Klawonn. Finding informative rules in interval sequences. In *International Symposium on Intelligent Data Analysis*, pages 125–134. Springer, 2001.

Keyun Hu, Yuchang Lu, Lizhu Zhou, and Chunyi Shi. Integrating classification and association rule mining : A concept lattice framework. In *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, pages 443–447. Springer, 1999.

Sergio Ilarri, Dragan Stojanovic, and Cyril Ray. Semantic management of moving objects : A vision towards smart mobility. *Expert Systems with Applications*, 42(3) :1418 – 1435, 2015. ISSN 0957-4174. doi : https://doi.org/10.1016/j.eswa.2014.08.057.

Christian S Jensen, Hua Lu, and Bin Yang. Graph model based indoor tracking. In *2009 Tenth International Conference on Mobile Data Management : Systems, Services and Middleware*, pages 122–131. IEEE, 2009.

Nyoman Juniarta, Miguel Couceiro, Amedeo Napoli, and Chedy Raïssi. Sequential pattern mining using fca and pattern structures for analyzing

visitor trajectories in a museum. In *CLA 2018-The 14th International Conference on Concept Lattices and Their Applications*, 2018a.

Nyoman Juniarta, Miguel Couceiro, Amedeo Napoli, and Chedy Raïssi. Sequential pattern mining using fca and pattern structures for analyzing visitor trajectories in a museum. In *CLA 2018 - The 14th International Conference on Concept Lattices and Their Applications*, Olomouc, Czech Republic, June 2018b. URL `https://hal.inria.fr/hal-01887914`.

Po-shan Kam and Ada Wai-Chee Fu. Discovering temporal patterns for interval-based events. In *International Conference on Data Warehousing and Knowledge discovery*, pages 317–326. Springer, 2000.

Artúr István Károly, Péter Galambos, József Kuti, and Imre J Rudas. Deep learning in robotics : Survey on model structures and training strategies. *IEEE Transactions on Systems, Man, and Cybernetics : Systems*, 51(1) : 266–279, 2020.

Alexandros Kontarinis, Claudia Marinica, Dan Vodislav, Karine Zeitouni, Anne Krebs, and Dimitris Kotzinos. Towards a better understanding of museum visitors' behavior through indoor trajectory analysis. In *Seventh International Conference on Digital Presentation and Preservation of Cultural and Scientific Heritage (DiPP2017)*, volume 7, pages 19–30, 2017.

Andreas D Lattner, Andrea Miene, Ubbo Visser, and Otthein Herzog. Sequential pattern mining for situation and behavior prediction in simulated robotic soccer. In *Robot Soccer World Cup*, pages 118–129. Springer, 2005.

Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. Trajectory clustering : a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 593–604, 2007.

Ulf Leonhardt. *Supporting location-awareness in open distributed systems*. PhD thesis, University of London London, 1998.

Guoquan Li, Enxu Geng, Zhouyang Ye, Yongjun Xu, Jinzhao Lin, and Yu Pang. Indoor positioning algorithm based on the improved rssi distance model. *Sensors*, 18(9) :2820, 2018.

Junhuai Li, Jinqin Wang, Lei Yu, and Jing Zhang. A novel frequent trajectory mining method based on gsp. In *International Conference on Web Information Systems and Mining*, pages 134–140. Springer, 2011.

Ki-Joune Li. Indoor space : A new notion of space. In *International symposium on web and wireless geographical information systems*, pages 1–3. Springer, 2008.

Hui Liu, Houshang Darabi, Pat Banerjee, and Jing Liu. Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(6) : 1067–1080, 2007.

Liu Liu, Bofeng Li, Ling Yang, and Tianxia Liu. Real-time indoor positioning approach using ibeacons and smartphone sensors. *Applied Sciences*, 10(6) : 2003, 2020.

Congnan Luo and Soon Myoung Chung. A scalable algorithm for mining maximal frequent sequences using sampling. In *16th IEEE International Conference on Tools with Artificial Intelligence*, pages 156–165. IEEE, 2004.

Ren C Luo and Tung Jung Hsiao. Dynamic wireless indoor localization incorporating with an autonomous mobile robot based on an adaptive signal model fingerprinting approach. *IEEE Transactions on Industrial Electronics*, 66(3) :1940–1951, 2018.

Heikki Manila. Discovering frequent episode in sequences. In *Proc. of the 1 st International Conference on Knowledge Discovery in Databases and Data Mining*, 1995.

Heikki Mannila, Hannu Toivonen, and A Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data mining and knowledge discovery*, 1(3) :259–289, 1997.

Florent Masseglia, Fabienne Cathala, and Pascal Poncelet. The psp approach for mining sequential patterns. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 176–184. Springer, 1998.

Florent Masseglia, Pascal Poncelet, and Maguelonne Teisseire. Efficient mining of sequential patterns with time constraints : Reducing the combinations. *Expert Systems with Applications*, 36(2) :2677–2690, 2009.

Ronaldo dos Santos Mello, Vania Bogorny, Luis Otavio Alvares, Luiz Henrique Zambom Santana, Carlos Andres Ferrero, Angelo Augusto Frozza, Geomar Andre Schreiner, and Chiara Renso. MASTER : A multiple aspect view on trajectories. *Transactions in GIS*, page tgis.12526, May 2019.

Mehdi Mohammadi, Ala Al-Fuqaha, Mohsen Guizani, and Jun-Seok Oh. Semisupervised deep reinforcement learning in support of iot and smart city services. *IEEE Internet of Things Journal*, 5(2) :624–635, 2017.

Mélanie Mondo. *Traces numériques et dimensions spatiales des pratiques de la ville*. PhD thesis, La Rochelle Université, 2022.

Eliakim Hastings Moore. *On a form of general analysis with applications to linear differential and integral equations*. Tipografia della R. Accademia dei Lincei, proprietà del cav. V. Salviucci, 1909.

Clément Moreau, Thomas Devogele, and Laurent Etienne. Calcul de similarité sémantique entre trajectoires. *Revue Internationale de Géomatique*, 29(1) :107–127, 2019. doi : 10.3166/rig.2019.00077.

Robert Moskovitch and Yuval Shahar. Fast time intervals mining using the transitivity of temporal relations. *Knowledge and Information Systems*, 42 (1) :21–48, 2015.

David Mountain and Jonathan Raper. 2000b) modelling human spatio-temporal behaviour : a challenge for location based services. In *University of Queensland*, 2001.

Zainab Munadhil, Sadik Kamel Gharghan, Ammar Hussein Mutlag, Ali Al-Naji, and Javaan Chahl. Neural network-based alzheimer's patient localization for wireless sensor network in an indoor environment. *IEEE Access*, 8 :150527–150538, 2020.

Fumiya Nakagaito, Tomonobu Ozaki, and Takenao Ohkawa. Discovery of quantitative sequential patterns from event sequences. In *2009 IEEE International Conference on Data Mining Workshops*, pages 31–36. IEEE, 2009.

Ahasanun Nessa, Bhagawat Adhikari, Fatima Hussain, and Xavier N Fernando. A survey of machine learning for indoor positioning. *IEEE access*, 8 :214945–214965, 2020.

Hassan Noureddine, Cyril Ray, and Christophe Claramunt. Semantic trajectory modelling in indoor and outdoor spaces. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 131–136. IEEE, 2020.

Hassan Noureddine, Cyril Ray, and Christophe Claramunt. A hierarchical indoor and outdoor model for semantic trajectories. *Transactions in GIS*, 26(1) :214–235, 2022.

Oystein Ore. Galois connexions. *Transactions of the American Mathematical Society*, 55(3) :493–513, 1944.

Salvatore Orlando, Raffaele Perego, and Claudio Silvestri. A new algorithm for gap constrained sequence mining. In *Proceedings of the 2004 ACM symposium on Applied computing*, pages 540–547, 2004.

Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, et al. Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4) :1–32, August 2013. ISSN 0360-0300, 1557-7341. doi : 10.1145/2501654.2501656. URL `https://dl.acm.org/doi/10.1145/2501654.2501656`.

Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *International Conference on Database Theory*, pages 398–416. Springer, 1999.

Jian Pei, Jiawei Han, Runying Mao, et al. Closet : An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD workshop on research issues in data mining and knowledge discovery*, volume 4, pages 21–30, 2000.

Jian Pei, Jiawei Han, and Wei Wang. Constraint-based sequential pattern mining : the pattern-growth methods. *Journal of Intelligent Information Systems*, 28 :133–160, 2007.

Nikos Pelekis, Ioannis Kopanakis, Gerasimos Marketos, Irene Ntoutsi, Gennady Andrienko, and Yannis Theodoridis. Similarity search in trajectory databases. In *14th International Symposium on Temporal Representation and Reasoning (TIME'07)*, pages 129–140, 2007. doi : 10.1109/TIME.2007.59.

G Piccinni, G Avitabile, and G Coviello. An improved technique based on zadoff-chu sequences for distance measurements. In *2016 IEEE Radio and Antenna Days of the Indian Ocean (RADIO)*, pages 1–2. IEEE, 2016.

Giovanni Piccinni, Gianfranco Avitabile, Giuseppe Coviello, and Claudio Talarico. Real-time distance evaluation system for wireless localization. *IEEE Transactions on Circuits and Systems I : Regular Papers*, 67(10) :3320–3330, 2020.

Sajina Pradhan, Youngchul Bae, Jae-Young Pyun, Nak Yong Ko, and Sukseung Hwang. Hybrid toa trilateration algorithm based on line intersection

and comparison approach of intersection distances. *Energies*, 12(9) :1668, 2019.

Arun K Pujari. *Data mining techniques*. Universities press, India, 2001. URL `https://books.google.ca/books?id=dH2KQhJboSYC`.

Willard V Quine. The problem of simplifying truth functions. *The American mathematical monthly*, 59(8) :521–531, 1952.

Javad Rezazadeh, Ramprasad Subramanian, Kumbesan Sandrasegaran, Xiaoying Kong, Marjan Moradi, and Farshad Khodamoradi. Novel ibeacon placement for indoor positioning in iot. *IEEE Sensors Journal*, 18(24) : 10240–10247, 2018.

Jérémy Richard, Bertet Karell, and Faucher Cyril. Ble based indoor positioning system and minimal zone searching algorithm (mzs) applied to visitor trajectories within a museum. *Applied Sciences*, 11(13) :6107, 2021.

Jérémy Richard, Guillaume Savarit, Salah Eddine Boukhetta, Cyril Faucher, Karell Bertet, and Christophe Demko. Discover spatio-temporal cluster from trajectory data enhance by heterogeneous contextual knowledge using fca and the nextpriorityconcept algorithm. In *The 16th International Conference on Concept Lattices and Their Applications*, Tallinn University of Technology, Estonia, jun 2022.

Livia Ruback, Marco Antonio Casanova, Alessandra Raffaetà, Chiara Renso, and Vania Vidal. Enriching mobility data with linked open data. In *Proceedings of the 20th International Database Engineering and Applications Symposium*, IDEAS '16, page 173–182, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341189. doi : 10.1145/2938503.2938550. URL `https://doi.org/10.1145/2938503.2938550`.

Ravi Sharma and Venkataramana Badarla. Geometrical optimization of a novel beacon placement strategy for 3d indoor localization. In *2018 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, pages 1–6. IEEE, 2018.

Arthur A Shaw and NP Gopalan. Finding frequent trajectories by clustering and sequential pattern mining. *Journal of Traffic and Transportation Engineering (English Edition)*, 1(6) :393–403, 2014.

Ebtesam Shemis and Ammar Mohammed. A comprehensive review on updating concept lattices and its application in updating association rules.

Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery, 11(2) :e1401, 2021.

Stefano Spaccapietra, Christine Parent, Maria Luisa Damiani, Jose Antonio de Macedo, Fabio Porto, and Christelle Vangenot. A conceptual view on trajectories. *Data and Knowledge Engineering*, 65 :126–146, 2008a. ISSN 0169-023X. doi : https://doi.org/10.1016/j.datak.2007.10.008.

Stefano Spaccapietra, Christine Parent, Maria Luisa Damiani, Jose Antonio de Macedo, Fabio Porto, and Christelle Vangenot. A conceptual view on trajectories. *Data and Knowledge Engineering*, 65(1) :126 – 146, 2008b. ISSN 0169-023X. doi : https://doi.org/10.1016/j.datak.2007.10.008.

Petros Spachos and Konstantinos N Plataniotis. Ble beacons for indoor positioning at an interactive iot-based smart museum. *IEEE Systems Journal*, 14(3) :3483–3493, 2020.

Ramakrishnan Srikant and Rakesh Agrawal. Mining sequential patterns : Generalizations and performance improvements. In *International conference on extending database technology*, pages 1–17. Springer, 1996.

Santosh Subedi, Goo-Rak Kwon, Seokjoo Shin, Suk-seung Hwang, and Jae-Young Pyun. Beacon based indoor positioning system using weighted centroid localization approach. In *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 1016–1019. IEEE, 2016.

Kevin Toohey and Matt Duckham. Trajectory similarity measures. *Sigspatial Special*, 7(1) :43–50, 2015.

Goce Trajcevski, Hui Ding, Peter Scheuermann, Roberto Tamassia, and Dennis Vaccaro. Dynamics-aware similarity of moving objects trajectories. GIS '07, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595939142. doi : 10.1145/1341012.1341027. URL `https://doi.org/10.1145/1341012.1341027`.

Marcel LJ van De Vel. *Theory of convex structures*. Elsevier, 1993.

Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. Design and use of the simple event model (sem). *Journal of Web Semantics*, 9(2) :128–136, 2011. ISSN 1570-8268. doi : https://doi.org/10.1016/j.websem.2011.03.003. URL `https://www.sciencedirect.com/science/article/pii/S1570826811000199`. Provenance in the Semantic Web.

Willem Robert Van Hage, Véronique Malaisé, Gerben KD de Vries, Guus Schreiber, and Maarten W van Someren. Abstracting and reasoning over ship trajectories and web data with the simple event model (sem). *Multimedia Tools and Applications*, 57 :175–197, 2012.

Bay Vo, Tuong Le, Frans Coenen, and Tzung-Pei Hong. Mining frequent itemsets using the n-list and subsume concepts. *International Journal of Machine Learning and Cybernetics*, 7(2) :253–265, 2016.

Jianyong Wang and Jiawei Han. Bide : Efficient mining of frequent closed sequences. In *Proceedings. 20th international conference on data engineering*, pages 79–90. IEEE, 2004.

Yapeng Wang, Xu Yang, Yutian Zhao, Yue Liu, and Laurie Cuthbert. Bluetooth positioning using rssi and triangulation methods. In *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*, pages 837–842. IEEE, 2013.

Takashi Washio, Koutarou Nakanishi, and Hiroshi Motoda. A classification method based on subspace clustering and association rules. *New Generation Computing*, 25(3) :235–245, 2007.

Rudolf Wille. Restructuring lattice theory : an approach based on hierarchies of concepts. In *International conference on formal concept analysis*, pages 314–339. Springer, 1982.

Edi Winarko and John F Roddick. Armada–an algorithm for discovering richer relative temporal association rules from interval-based data. *Data Knowledge Engineering*, 63(1) :76–90, 2007.

Jiaqi Xiang, Qingdong Li, Xiwang Dong, and Zhang Ren. Continuous control with deep reinforcement learning for mobile robot navigation. In *2019 Chinese Automation Congress (CAC)*, pages 1501–1506. IEEE, 2019.

Xifeng Yan, Jiawei Han, and Ramin Afshar. Clospan : Mining : Closed sequential patterns in large datasets. In *Proceedings of the 2003 SIAM international conference on data mining*, pages 166–177. SIAM, 2003.

Zhixian Yan, Jose Macedo, Christine Parent, and Stefano Spaccapietra. Trajectory ontologies and queries. *Transactions in GIS*, 12 :75–91, 2008.

Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. Semitri : A framework for semantic annotation of heterogeneous trajectories. In *Proceedings of the 14th International*

Conference on Extending Database Technology, EDBT/ICDT '11, page 259–270, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450305280. doi : 10.1145/1951365.1951398. URL https://doi.org/10.1145/1951365.1951398.

Yuji Yoshimura, Stanislav Sobolevsky, Carlo Ratti, Fabien Girardin, Juan Pablo Carrascal, Josep Blat, and Roberta Sinatra. An analysis of visitors' behavior in the louvre museum : A study using bluetooth data. *Environment and Planning B : Planning and Design*, 41(6) :1113–1131, 2014.

Faheem Zafari, Athanasios Gkelias, and Kin K Leung. A survey of indoor localization systems and technologies. *IEEE Communications Surveys Tutorials*, 21(3) :2568–2599, 2019.

Mohammed J Zaki. Sequence mining in categorical domains : incorporating constraints. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 422–429, 2000.

Mohammed J Zaki. Spade : An efficient algorithm for mining frequent sequences. *Machine learning*, 42(1) :31–60, 2001.

Mohammed J Zaki and Ching-Jui Hsiao. Charm : An efficient algorithm for closed itemset mining. In *Proceedings of the 2002 SIAM international conference on data mining*, pages 457–473. SIAM, 2002.

Chao Zhang, Jiawei Han, Lidan Shou, Jiajun Lu, and Thomas La Porta. Splitter : Mining fine-grained sequential patterns in semantic trajectories. *Proceedings of the VLDB Endowment*, 7(9) :769–780, 2014.

Fan Zhang, Yan Zhang, and Jason D Bakos. Accelerating frequent itemset mining on graphics processing units. *The Journal of Supercomputing*, 66 (1) :94–117, 2013.

Yu Zheng. Trajectory data mining : an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3) :1–41, 2015.

Annexe

La plateforme GALACTIC

La plateforme **GALACTIC** [1] (**GA**lois **LA**ttices, **C**oncept **T**heory **I**mplicational systems and **C**losures) intègre en son coeur l'algorithme NextPriorityConcept permettant de générer un treillis de concepts à partir de données complexes et hétérogènes. La plateforme est organisée comme illustré par la figure 9.2

- Le cœur de la plateforme (`GALACTIC Core`) qui contient l'implémentation de l'algorithme NextPriorityConcept et plusieurs outils pour la manipulation algébrique de treillis.
- Des extensions de *caractéristiques* (pétale `Characteristics`) qui définissent des types de données.
- Des extensions de *descriptions* (pétale `Descriptions`) qui définissent des descriptions par des prédicats pour chaque caractéristique.
- Des extensions de *stratégies* (pétale `Strategies`) qui définissent des stratégies pour la génération des sélecteurs pour chaque caractéristique. Deux méta-stratégies sont définies :
  - La méta-stratégie *SelectionFilter* permet de prendre les sélecteurs qui maximisent/minimisent une mesure.
  - La méta-stratégie *LimitFilter* permet de garder les sélecteurs qui génèrent des prédécesseurs respectant un seuil minimal d'une mesure.
- Des extensions de *mesures* (pétale `Measures`) qui définissent des mesures utilisées par les méta-stratégies. Trois extensions de mesures sont définies :
  - La mesure de la confiance.

1. https ://GALACTIC.univ-lr.fr

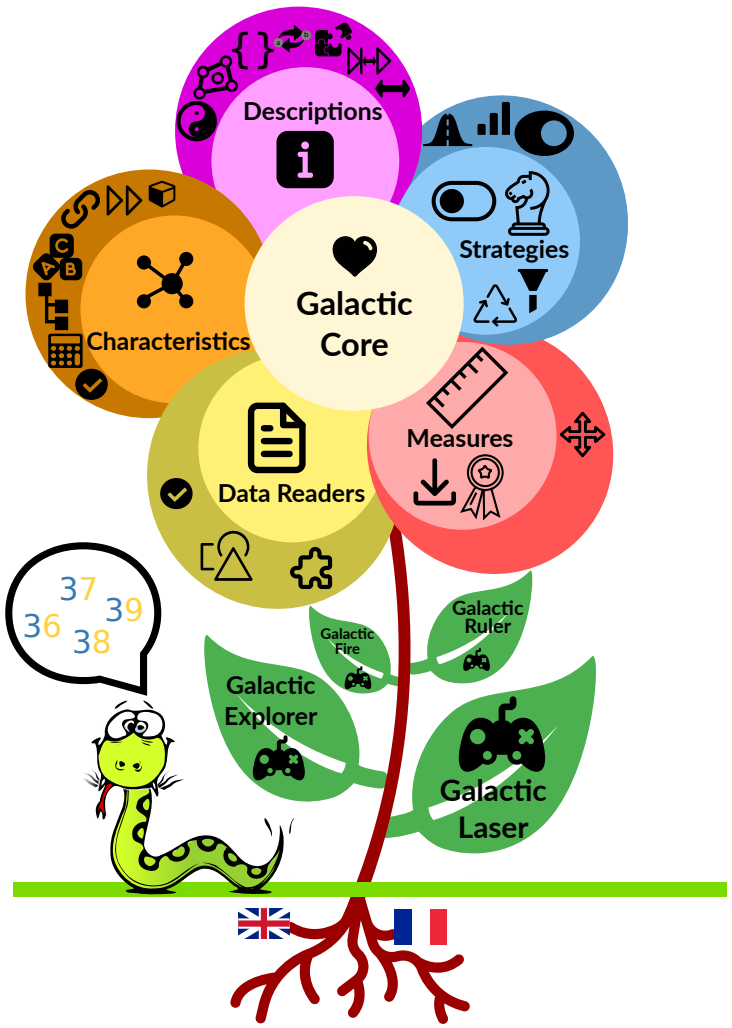


FIGURE 9.2 – Architecture de la plateforme **GALACTIC**

- La mesure de support.
- La mesure d'entropie.

Des extensions de lecture de données (pétale `Data Readers`) qui permettent la lecture de plusieurs types de fichiers de données :

- Extensions pour les données binaires (compatible avec les anciens formats de fichiers du l'AFC).
- Extensions pour les données hétérogènes (*CSV*, *TOML*, *INI*).
- Extensions pour les données complexes (*JSON*, *YAML*).

Des applications qui utilisent le cœur et les extensions (feuilles).

— `GALACTIC-laser` permet de dessiner un diagramme de Hasse d'un treillis de concepts généré à partir d'un fichier de stratégies et un fichier de données.

— `GALACTIC-ruler` permet de générer les règles d'associations.

Des extensions de localisation utilisées pour la traduction des applications en deux langues pour l'instant, anglais et français (racines).

Avec ce système d'extensions, la plateforme peut analyser plusieurs type de données. `GALACTIC` propose actuellement des extensions pour des données binaires, numériques et catégorielles :

Extensions pour des données binaires :

— une description classique décrivant un ensemble d'objets par un ensemble d'attributs communs.

— une stratégie consiste à générer des sélecteurs à l'aide de formules logiques impliquant un groupe de plusieurs attributs booléens et leurs négations [Quine, 1952].

Extensions pour des données numériques :

— une description numérique décrivant une collection d'objets par une enveloppe convexe [van De Vel, 1993] sur $\mathbb{R}^n$, c'est-à-dire un groupe de n caractéristiques numériques.

— deux types de stratégies possibles :

La stratégie *Normal* consiste à restreindre l'ensemble d'objets en utilisant $m \pm \alpha\sigma$ sur la composante principale (à l'aide de techniques d'analyse en composantes principales). m est la moyenne, σ est l'écart-type et α est un paramètre d'entrée de la stratégie.

La stratégie *Quantile* affine l'ensemble d'objets en utilisant un sous-groupe pour chaque quantile.

Extensions pour des données catégorielles :

— une description par le sous-ensemble minimal contenant les valeurs des catégories.

— une stratégie qui consiste à retirer une valeur de ce sous-ensemble permettant de sélectionner un ensemble d'objets plus petit.

Jérémy RICHARD

De la capture de trajectoires de visiteurs vers l'analyse interactive de comportement après enrichissement sémantique

Résumé :

Cette thèse porte sur l'étude comportementale de l'activité touristique en utilisant une approche d'analyse générique et interactive. Le processus d'analyse développé concerne la trajectoire touristique dans la ville et dans les musées en tant que terrain d'étude. Des expérimentations ont été menées pour collecter les données de déplacement dans la ville touristique en utilisant des signaux GPS, permettant ainsi l'obtention d'une trajectoire de déplacement. Toutefois, l'étude se focalise en premier lieu sur la reconstruction de la trajectoire d'un visiteur dans les musées à l'aide d'un équipement de positionnement intérieur, c'est-à-dire dans un environnement contraint. Ensuite, un modèle d'enrichissement sémantique multi-aspects générique est développé pour compléter la trajectoire d'un individu en utilisant plusieurs données de contexte telles que les noms des quartiers traversés par l'individu dans la ville, les salles des musées, la météo à l'extérieur et des données d'application mobile à l'intérieur. Les trajectoires enrichies, appelées trajectoires sémantiques, sont ensuite analysées à l'aide de l'analyse formelle de concept et de la plateforme **GALACTIC**, qui permet l'analyse de structures de données complexes et hétérogènes sous la forme d'une hiérarchie de sous-groupes d'individus partageant des comportements communs. Enfin, l'attention est portée sur l'algorithme "ReducedContextCompletion" qui permet la navigation interactive dans un treillis de concepts, ce qui permet à l'analyste de données de se concentrer sur les aspects de la donnée qu'il souhaite explorer.

Mots clés : Positionnement intérieur, musée intelligent, trajectoire, trajectoire sémantique, classification des données, analyse formelle des concepts, exploration des séquences, exploration des motifs, théorie du treillis, données hétérogènes

Jérémy RICHARD

Towards interactive analysis of visitor behavior through capturing and semantically enhancing trajectories

Summary :

This thesis focuses on the behavioral study of tourist activity using a generic and interactive analysis approach. The developed analytical process concerns the tourist trajectory in the city and museums as the study field. Experiments were conducted to collect movement data in the tourist city using GPS signals, thus enabling the acquisition of a movement trajectory. However, the study primarily focuses on reconstructing a visitor's trajectory in museums using indoor positioning equipment, i.e., in a constrained environment. Then, a generic multi-aspect semantic enrichment model is developed to supplement an individual's trajectory using multiple context data such as the names of neighborhoods the individual passed through in the city, museum rooms, weather outside, and indoor mobile application data. The enriched trajectories, called semantic trajectories, are then analyzed using formal concept analysis and the **GALACTIC** platform, which enables the analysis of complex and heterogeneous data structures as a hierarchy of subgroups of individuals sharing common behaviors. Finally, attention is paid to the "ReducedContextCompletion" algorithm that allows for interactive navigation in a lattice of concepts, allowing the data analyst to focus on the aspects of the data they wish to explore.

Keywords : Indoor positionning, smart museum, trajectory, semantic trajectory, clustering, formal concept analysis, sequence mining, pattern mining, lattice theory, heterogeneous data

Laboratoire L3i, Institut LUDI, Bâtiment Pascal, Avenue Michel Crépeau

17042 La Rochelle Cedex 1 - France