

FOUILLE DE DONNÉES DE SANTÉ

HABILITATION À DIRIGER DES RECHERCHES

Spécialité Informatique

UNIVERSITÉ DE MONTPELLIER

présentée et soutenue publiquement le 2 décembre 2015

par

Sandra BRINGAY

Rapporteurs : Diana INKPEN Professeur, Université de Ottawa, Canada
Patrick RUCH Professeur, Haute Ecole de Gestion de Genève, Suisse
Marie Christine JAULENT Directrice de recherche, INSERM U 872, France

Examineurs : Bruno CRÉMILLEUX Professeur, Université de Caen, France
Pascal PONCELET Professeur, Université de Montpellier, France

Mis en page avec la classe thloria.

TABLE DES MATIÈRES

Introduction	1
A Contexte général et thématiques des recherches	1
B Synthèse des travaux menés	2
C Organisation du manuscrit	2
<hr/>	
Chapitre 1 État de l’art : Méthodes d’Analyse appliquées aux Données de Santé	5
A Introduction	6
B Données et acteurs de la santé	7
C Taxonomie des méthodes d’analyse	8
D Le processus d’analyse de données	10
E Applications de l’analyse de données dans le domaine de la santé	11
F Challenges à venir pour l’application de l’analyse de données en santé	15
Chapitre 2 Formalisation de la connaissance médicale sous forme de motifs	19
A Introduction	21
B Des méthodes descriptives efficaces basées sur les motifs	22
1 Décrire des données de plus en plus riches sémantiquement	22
1.1 Motifs séquentiels	22
1.2 Motifs contextuels	24
1.3 Motifs spatio-temporels	27
1.4 Motifs partiellement ordonnés clos	29
2 Choisir, représenter et interpréter les meilleurs motifs	31
2.1 Mesures d’intérêt et leur combinaison	31

2.2	Visualisation de motifs	32
C	Vers des approches prédictives et prescriptives basées sur les motifs	38
1	Prédire le grade d'un cancer	38
2	Prédire la catégorie d'un patient et le prochain évènement	39
3	Prescrire en détectant une anomalie et en déclenchant une alerte	39
D	Conclusions et perspectives	40
1	Des motifs de plus en plus riches sémantiquement	41
2	Sélection et visualisation des motifs	41
3	De nouvelles applications aux motifs, utiles aux professionnels de la santé	42
Chapitre 3 Analyse de la santé via les médias sociaux		43
A	Introduction	44
B	Motivations et challenges liés à l'analyse des productions des e-patients dans les médias sociaux	45
C	Enrichissement de messages	48
1	De quoi ?	48
1.1	Prédiction supervisée de thèmes	48
1.2	Recherche d'informations	49
1.3	Détection non supervisée de thèmes	50
2	Comment ?	51
2.1	Sentiment : polarités et émotions	52
2.2	Constitution d'une ressource des émotions pour le français	54
2.3	Cibles et sources des sentiments	55
3	Qui ?	57
3.1	Détection de rôles médicaux	57
3.2	Confiance et réputation	59
4	Quand ?	60
D	Conclusions et perspectives	61
1	Des pré-traitements complexes	61
2	Vocabulaire patient	62
3	Ressources annotées	63
4	Et les motifs ?	63
5	Visualisations	63
Chapitre 4 Conclusion, perspectives		65
A	Résumé des contributions	66

	iii
B Perspectives	67
Bibliographie	71
Glossary	89
<hr/>	
Partie I Curriculum vitae	91
<hr/>	
Partie II Sélection de publications	107
Liste des Articles	109
<hr/>	
Partie III Annexes	137

INTRODUCTION

A Contexte général et thématiques des recherches

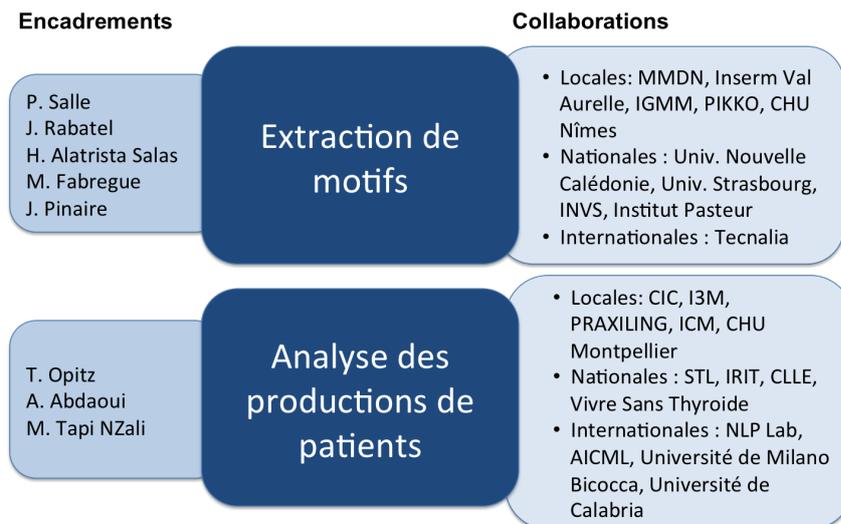


FIGURE 1 : Deux thématiques de recherche développées depuis 2007 avec les principaux encadrements et collaborations.

Les techniques d'analyse de données sont désormais utilisées par de nombreuses organisations (e.g. business, web). Dans le domaine de la santé, ces techniques sont de plus en plus populaires et se révèlent même indispensables pour gérer les gros volumes de données produits pour un patient et par le patient. Par exemple, les méthodes d'analyse de données ont été utilisées pour aider les professionnels de santé à identifier les traitements les plus efficaces pour certaines pathologies [Song et al., 2013] ou bien les meilleures pratiques permettant d'offrir les soins les plus adaptés tout en rationalisant les dépenses [Beresniak et al., 2012]. En effet, les données de santé générées sont parfois trop complexes et volumineuses pour être traitées par des méthodes traditionnelles. Les méthodes d'analyse fournissent alors des outils pour transformer ces données en informations utiles pour la prise de décision et en connaissances utiles pour la recherche bio-médicale. Ces méthodes s'adressent à tous les intervenants du système de

soin : les professionnels de santé qui interagissent avec le patient (e.g. médecin de ville, personnel hospitalier), les administrations (e.g. hôpitaux, réseaux de soin, mutuelles), les chercheurs, etc.

Dans le cadre de ma thèse, j'ai étudié les données produites dans le service de pédiatrie du **CHU** d'Amiens via la mise en place d'un dossier patient annoté de manière informelle. Depuis mon arrivée en 2007 dans l'équipe TATOO devenue **ADVANCE** (ADVanced Analytics for data SciencE)¹ du **LIRMM** (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier)², je m'intéresse aux méthodes d'analyse de données appliquées aux données de santé en général, qu'elles soient produites par les professionnels de la santé ou par les patients. Dans ce manuscrit d'Habilitation à Diriger des Recherches, je présente les deux thématiques résumées dans la Figure 1.2.

La première thématique porte sur la définition, la formalisation, l'implémentation et la validation de méthodes d'analyse permettant de décrire le contenu de bases de données médicales. Je me suis particulièrement intéressée aux données séquentielles en faisant évoluer la classique notion de motif séquentiel pour y intégrer des composantes contextuelles, spatiales, sur l'ordre potentiellement partiel des éléments composant les motifs. Ces nouvelles informations enrichissent la sémantique initiale de ces motifs.

La seconde thématique, débutée en 2013 pendant 6 mois de délégation suivis de 6 mois de Congés pour Recherches ou Conversions Thématiques (**CRCT**), se focalise sur l'analyse des productions et des interactions des patients au travers des médias sociaux. J'ai principalement travaillé sur des méthodes permettant d'analyser les productions narratives des patients selon leurs temporalités, leurs thématiques, les sentiments associés, le rôle et la réputation du locuteur s'étant exprimé dans les messages.

B Synthèse des travaux menés

Les travaux que je vais présenter dans ce manuscrit s'inscrivent dans ces deux thématiques et se déclinent dans le cadre de collaborations académiques, industrielles ainsi que dans le cadre d'encadrements de masters recherche, de thèses et de post-doctorats (cf. Tableau 1). D'autres collaborations et encadrements moins en rapport avec les thèmes de ce manuscrit ont également été menés depuis 2007. Ces derniers ne sont pas présents dans le tableau, ni dans le manuscrit mais indiqués dans le CV en fin de manuscrit. Des parties de ce tableau seront rappelées en en-tête de chaque chapitre et associées à des listes de publications dans des encadrés gris.

C Organisation du manuscrit

Ce manuscrit synthétise les deux thématiques selon un fil directeur détaillé ci-dessous.

Ce manuscrit débute par un Chapitre 1 d'état de l'art qui servira de cadre à la description des travaux présenté dans les chapitres 2 et 3. Après avoir introduit les données de santé de manière générale, je présente les différentes méthodes d'analyse : descriptive, prédictive et prescriptive, ainsi que le processus général d'analyse des données. Je reprends ensuite une typologie des applications de l'analyse des données dans le domaine de la santé, présentée dans plusieurs revues de la littérature et montre son incomplétude dans le cas des méthodes descriptives. Je conclus sur les challenges associés à l'analyse des données de santé en général et je montre comment mes travaux s'inscrivent au cœur de ces challenges.

1. <https://www.lirmm.fr/recherche/equipes/advance>

2. www.lirmm.fr

TABLE 1 : Résumé des projets et des encadrements. Légende : M-Master, T-Thèse, PostD-Post-doctorant.

Étudiant & M2/T	Co-encadrements	Projets & Collaborations	Thématiques
Extraction de motifs			
P. Salle (T 2007-2010)	M. Teisseire	MMDN, ANR Pradnet (2007-2011)	Motifs séquentiels, mesures d'intérêt
J. Rabatel (T 2008-2011)	P. Poncelet	Tecnalia (Espagne) (2008-2011)	Motifs contextuels, prédiction
H. Alatrística Salas (T 2010-2013)	F. Flouvat, N. Selmaoui et M. Teisseire	Univ. Nouvelle Calédonie, InVS, DASS, Institut Pasteur	Motifs spatio-séquentiels, mesures d'intérêt, visualisation
M. Fabregue (M2 2010-2011)	P. Poncelet	INSERM Val d'Aurelle (2010)	Prédiction
M. Fabregue (T 2011-2014)	A. Braud, F. Le Ber et M. Teisseire	Projet Fresqueau (2011-2014)	Motifs partiellement ordonnés clos, mesures d'intérêt, visualisation
A. Zine El Abidine (M2 2013-2014)	P. Poncelet et A. Sallaberry	IGMM (2010-2011)	Trajectoires, visualisation
J. Pinaire (T Depuis 2014)	J. Azé, P. Landais	CHU Nîmes (Depuis 2014)	Motifs contextuels, mesures d'intérêt
Analyse des productions des patients			
S. Melzi (M2 2012-2013)	P. Poncelet	Projet Patients Mind (Depuis 2013)	Analyse de sentiments
A. Abdaoui (T Depuis 2013)	J. Azé	Projet Patients Mind (Depuis 2013)	Rôle, Confiance, Réputation
T. Opitz (PostD 2013-2014)	C. Lavergne et C. Mollevi	ICM et I3M (Depuis 2014)	Qualité de vie, Enrichissement des messages
M. Tapi NZali (T Depuis 2014)	C. Lavergne et C. Mollevi	ICM et I3M (Depuis 2014)	Qualité de vie, Vocabulaire Patient

Dans le Chapitre 2, je présente les différents types de motifs sur lesquels nous avons été amenés à travailler. Dans le cadre d'une collaboration avec le MMDN (ANR Pradnet), nous avons appliqué les algorithmes classiques d'extraction de motifs séquentiels sur des données de puces à ADN (Thèse de P. Salle). L'originalité de ces travaux a été d'utiliser l'ordre entre les expressions des gènes à la place de l'ordre temporel classique. Nous avons ensuite étendu ces motifs pour prendre en compte des informations contextuelles (Thèse CIFRE de J. Rabatel avec la société Tecnalia) qui s'avèrent tout à fait pertinentes pour extraire des trajectoires de patients (Thèse de J. Pinaire en collaboration avec le CHU de Nîmes). Nous avons également étendu ces motifs pour intégrer des informations spatiales (Thèse de H. Alatrística Salas) qui permettent de suivre l'évolution des épidémies de dengue, dans le temps et l'espace, (Collaboration avec l'Université de la Nouvelle Calédonie, InVS, DASS, Institut Pasteur). Pour chacune de ces approches, nous avons défini des mesures d'intérêt pour la sélection des motifs et des visualisations qui permettent aux experts de s'approprier les résultats des méthodes d'analyse. Toutefois, ces méthodes génèrent de très nombreux résultats. Nous avons travaillé sur une représentation plus condensée des informations contenues dans les motifs via la définition des motifs partiellement ordonnés clos (Thèse de M. Fabregue) qui ont permis d'évaluer la qualité de l'eau des rivières (ANR Fresqueau). Ces différents motifs sont efficaces pour des tâches de prédiction comme la prédiction du grade du cancer (Master de M. Fabregue en collaboration avec l'INSERM Val d'Aurelle) ou la détection d'anomalies (Thèse CIFRE de J. Rabatel). Je conclus ce chapitre par des perspectives liées à la sémantique de ces motifs, à leur sélection, à leur visualisation et aux nouvelles applications que ces motifs permettent d'envisager.

Dans le Chapitre 3, je présente les travaux menés dans le cadre du projet Patients' Mind dont je suis le porteur sur l'analyse des productions des patients dans les médias sociaux. Après avoir présenté les

différents enjeux et challenges associés à l'analyse des médias sociaux, je présente des méthodes permettant d'enrichir sémantiquement les messages des patients selon quatre dimensions, temporelle (Thèse de M. Tapi NZali et Post-doctorat de T. Opitz), thématique (Post-doctorat de T. Opitz), liée au locuteur (Thèse de A. Abdaoui) et à la manière dont il s'exprime en nous focalisant notamment sur l'analyse de sentiments (Master 2 de S. Melzi et Thèse de A. Abdaoui). Je conclue ce chapitre en listant les limites inhérentes à ce type d'applications et les perspectives envisagées à savoir l'uniformisation des chaînes de pré-traitements complexes mais nécessaires dans ce type d'approche, le problème du vocabulaire utilisé par les patients qui est très différent de celui des médecins et qu'il faut acquérir pour augmenter les performances de nos méthodes, l'effort indispensable de construction de ressources annotées pour l'évaluation et le manque de réflexion sur des visualisations indispensables pour l'interprétation des résultats de nos méthodes. Je conclurai également sur les perspectives d'intégration des méthodes définies dans le chapitre 2 sur les chaînes de traitements du chapitre 3.

Dans le chapitre 4, je fais un bilan de l'ensemble des travaux décrits dans les chapitres 2 et 3 puis je présente les principales perspectives.

CHAPITRE

1

ÉTAT DE L'ART : MÉTHODES D'ANALYSE APPLIQUÉES AUX DONNÉES DE SANTÉ

Sommaire

- A Introduction
 - B Données et acteurs de la santé
 - C Taxonomie des méthodes d'analyse
 - D Le processus d'analyse de données
 - E Applications de l'analyse de données dans le domaine de la santé
 - F Challenges à venir pour l'application de l'analyse de données en santé
-

A Introduction

Les organisations produisent de plus en plus de données. Elles ont besoin de prendre les meilleures décisions, le plus rapidement possible, en se basant sur ces données. Dans ce contexte, l'*Analyse de Données* (DA Data Analytics) gagne en popularité [Dursun, 2014]. C'est un processus composé de plusieurs étapes qui vise à découvrir des connaissances utiles pour la prise de décision et qui peut aller jusqu'à suggérer des conclusions ou même des actions. L'analyse des données a été étudiée par plusieurs communautés (statistiques, informatique, sciences sociales, etc.). Il existe désormais de nombreuses méthodes et outils connus sous une grande variété de noms. En statistiques par exemple, on trouve l'*Analyse Exploratoire des Données* (EDA Exploratory Data Analysis - découverte de nouvelles caractéristiques). En informatique, on parle plutôt de *Fouille de Données* (DM Data Mining) et d'*Analyse Visuelle* (VA Visual Analytics) qui sont des étapes du *Processus de Découverte de Connaissances* (KDD Knowledge Discovery Process). Bien d'autres méthodes sont associées à la thématique de l'*Analyse de données*.

Dans la littérature, les auteurs opposent très souvent les méthodes d'analyse aux méthodes d'interrogation de bases de données [El-Sappagh et al., 2013]. Par exemple, un patient peut se demander *quel est le lien entre l'anorexie et le suicide ?* Juste en formulant cette requête, le patient fait l'hypothèse qu'il existe un lien entre l'anorexie et le suicide. Si certaines méthodes d'analyse ont besoin de telles hypothèses, d'autres au contraire vont traiter toutes les relations possibles existant dans les données sans un tel *a priori*. Elles vont inclure l'hypothèse formulée par le patient mais également identifier de nouvelles relations, "cachées", dans les données, que l'analyste n'aurait pas envisagées.

Dans le domaine de la santé, les méthodes d'analyse sont très utiles comme le montre les revues de la littérature [Wasan et al., 2006, Khajehei and Etemady, 2010, El-Sappagh et al., 2013]. Les très grandes quantités de données générées par le système de soins pour chaque patient sont trop complexes et volumineuses pour être traitées et analysées sans automatisation. L'analyse de données fournit alors des méthodes et des outils pour transformer ces gros volumes de données en informations utiles pour la prise de décision médicale. Tous les acteurs impliqués dans le système de soins peuvent bénéficier des applications de l'analyse de données : les médecins pour identifier les traitements les plus efficaces et de meilleures pratiques, les hôpitaux pour prendre des décisions de gestion et par exemple diminuer les coûts des soins, les patients pour identifier les soins les plus abordables, les assureurs pour détecter les fraudes et les abus, etc.

Dans ce chapitre, nous allons tout d'abord présenter dans la Section B le système de soins français, les acteurs pouvant être intéressés par des applications de l'analyse de données ainsi que les types de données sur lesquelles ces méthodes peuvent s'appliquer. Nous décrirons ensuite une taxonomie des différentes méthodes d'analyse de données [Dursun, 2014], devenue assez classique en *intelligence économique* (BI Business Intelligence) dans la Section C. Nous présenterons dans la Section D, le processus d'analyse. Nous utiliserons la taxonomie de la Section C pour explorer différentes applications dans le domaine de la santé dans la Section E et montrerons l'incomplétude de cette taxonomie. Pour finir, nous mettrons en évidence certaines limites et challenges associés à ces méthodes et nous positionnerons les travaux décrits dans les chapitres suivants dans la Section F.

B Données et acteurs de la santé

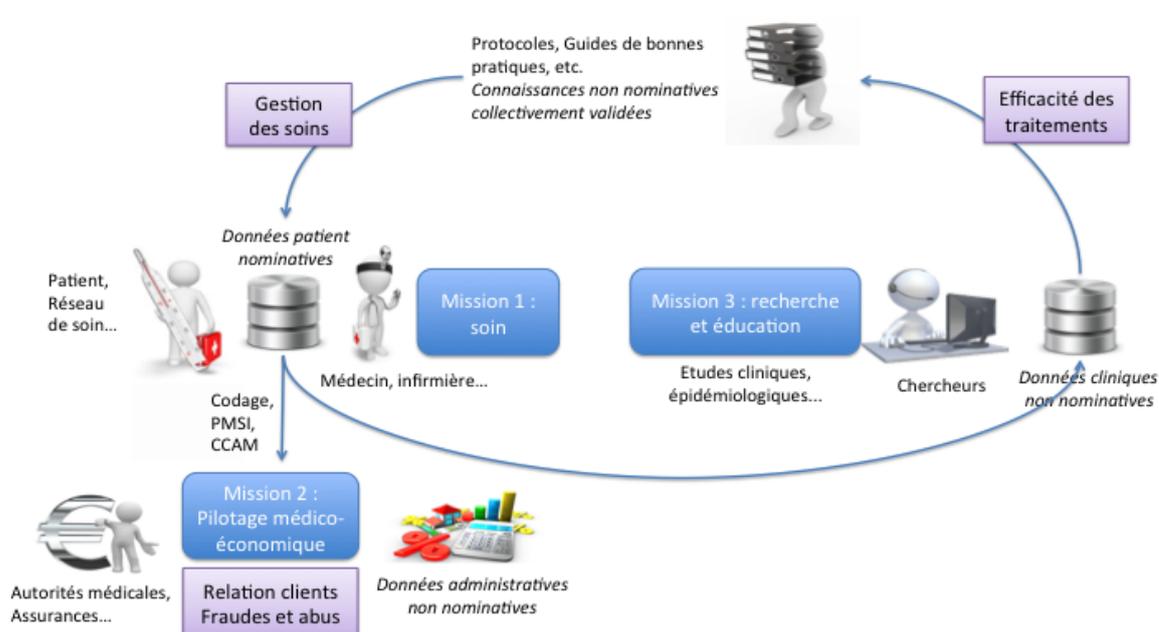


FIGURE 1.1 : Flux de connaissances médicales dans le système de soins français, schéma adapté de [Charlet, 2002].

La figure 1.1 représente les flux de connaissances dans le système de soins français. Cette figure est adaptée des travaux de [Charlet, 2002]. Nous avons placé au centre le dossier patient contenant toutes les informations nominatives produites à propos d'un patient (e.g. son poids, son âge) par les professionnels de la santé dans leur mission de soins (mission 1). Ces connaissances nominatives peuvent être présentées à d'autres acteurs (e.g. les patients, les réseaux de soins). Il est possible d'en extraire des connaissances non nominatives (e.g. le nombre de patients présentant un taux de cholestérol anormalement élevé, le nombre de lits occupés).

Dans leur mission de pilotage médico-économique (mission 2), les gestionnaires des établissements de soins utilisent ces informations non nominatives pour prendre des décisions afin d'améliorer le pilotage médico-économique des établissements (e.g. comptabilité analytique, contrôle des coûts). Par exemple, le Programme de Médicalisation des Systèmes d'Information (PMSI)³ oblige les établissements à évaluer leurs activités. Chaque patient est associé avec un Résumé de Sortie Standardisé (RSS), créé à la fin de l'hospitalisation et comprenant des données médicales, économiques et administratives. Le RSS classe les patients en groupes : les Groupements Homogènes de Maladies (GHM). Un coût est associé à chaque GHM et permet d'attribuer un budget à chaque établissement.

Dans leur mission de recherche et d'éducation (mission 3), les professionnels de la santé utilisent également des données non nominatives pour des études épidémiologiques, des études cliniques, etc. Ces études conduisent à de nouvelles connaissances scientifiques et de nouveaux savoir-faire. Ces études participent ainsi au développement de la *médecine à base de preuves* (EBM Evidence-Based Medicine), qui aide les professionnels de la santé à partager et appliquer des connaissances collectivement validées. Ces nouvelles connaissances sont décrites dans des articles, des protocoles, des guides de bonnes

3. <http://www.atih.sante.fr/mco/presentation>

pratiques, des thésaurus, etc. Ces éléments sont diffusés puis utilisés par les professionnels de la santé pendant les soins. Ces derniers doivent acquérir constamment ces nouvelles connaissances et savoir-faire, qu'ils doivent adapter au contexte, toujours particulier, des patients.

L'analyse de données peut être appliquée sur toutes les données décrites dans la figure 1.1 pour des applications très variées. Plusieurs revues de la littérature [Koh and Tan, 2005, El-Sappagh et al., 2013] ont montré l'apport des méthodes d'analyse pour les applications liées à la santé. Ces auteurs les décrivent selon 4 catégories que nous avons replacées dans le schéma précédent et décrivons ci-dessous :

Efficacité des traitements (Treatment Effectiveness). L'analyse de données peut être utilisée pour évaluer l'efficacité des traitements médicaux. En comparant les causes, les symptômes, les résultats des traitements et les effets secondaires, l'analyse peut montrer quels traitements s'avèrent efficaces et sur quelles populations [Milley, 2000] (mission 3).

Gestion des soins (Healthcare Management). Pour faciliter la gestion des soins, améliorer leur efficacité et réduire leur coût de manière générale, l'analyse de données peut être utile pour mieux appréhender les maladies chroniques, les patients à haut risque et leur impact sur les établissements de soins. Par exemple, il est possible d'optimiser la gestion des médicaments dans une pharmacie [Debruyne et al., 2015] (missions 1 et 2).

Gestion de la relation client (Customer Relationship Management). La gestion de la relation client est une approche très classique pour la gestion des interactions entre les organisations commerciales et leurs clients. Cette relation est de plus en plus étudiée afin d'évaluer les interactions des "patients-clients" avec les centres d'appels, les cabinets de médecins, les services de facturation, le milieu hospitalier, les établissements de soins ambulatoires, etc. L'analyse de données permet de déterminer les préférences de ces patients-clients, leur profil d'utilisation, leurs besoins actuels et futurs dans l'objectif de garantir un certain niveau de satisfaction (mission 2). Par exemple, [Wu et al., 2014] analyse les avis des patients à l'issue d'un séjour hospitalier.

Fraudes et abus (Fraud and Abuse). Pour détecter des fraudes et des abus, il faut tout d'abord définir des normes et identifier des comportements inhabituels ou anormaux par rapport à ces normes. Par exemple, il est possible de repérer des prescriptions inappropriées ou des réclamations de patients abusives (mission 2). Une revue de ces méthodes est décrite dans [Joudaki et al., 2014].

Dans la Section E, nous allons montrer que cette catégorisation reprise par différents auteurs [Koh and Tan, 2005, El-Sappagh et al., 2013] est loin d'être exhaustive et ne correspond pas à la réalité des recherches menées depuis les années 2000 en analyse de données dans le domaine de la santé. Nous présentons dans la Section C les méthodes d'analyse pouvant être appliquées aux données de santé.

C Taxonomie des méthodes d'analyse

Selon M. Wu⁴, il existe trois grandes familles de méthodes d'analyse permettant de répondre aux trois questions suivantes : l'analyse **descriptive** (*que s'est-il passé ?*), l'analyse **prédictive** (*que pourrait-il se passer ?*) et l'analyse **prescriptive** (*que devrions nous faire ?*). Si ces trois analyses partagent un objectif commun, celui de fournir aux organisations des éléments d'informations pour faciliter la prise de décision, chacune engendre des niveaux d'actions différents.

Analyse descriptive. L'analyse descriptive des données vise à résumer, condenser, les événements passés ou présents. En effet, la plupart du temps, les grands volumes de données brutes ne sont pas

4. <http://community.lithium.com/t5/Science-of-Social-blog/Big-Data-Reduction-3-From-Descriptive-to-Prescriptive/ba-p/81556>

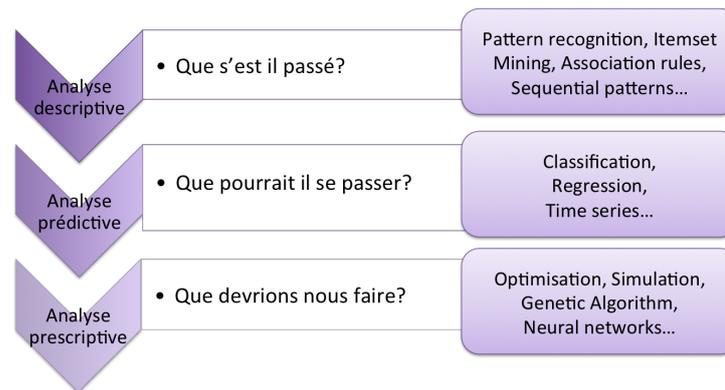


FIGURE 1.2 : Analyse descriptive, prédictive et prescriptive.

interprétables par les humains mais les informations dérivées de ces données grâce à l'analyse descriptive le sont. L'analyse descriptive donne alors le contexte nécessaire à l'analyste pour prendre une décision et réaliser des actions. Par exemple, les données récoltées par un électroencéphalogramme peuvent être résumées par certaines mesures non linéaires, précieuses pour étudier la maladie de Parkinson [Stam et al., 1995]. Le type le plus courant d'analyse réalisée est la découverte de motifs dans les données. Ces motifs sont généralement reportés sous forme d'indicateurs dans des tableaux de bord. Des alertes basées sur l'adéquation de nouvelles données à des motifs préalablement extraits peuvent être automatisées [Hou and Zhang, 2008]. Les motifs sont caractérisés par des mesures qui donnent des détails sur la répétition et la répartition de motifs dans les données telles que la fréquence des événements, la confiance, etc. Une autre approche descriptive classique est le clustering (ou partitionnement de données) qui vise à identifier dans un ensemble de données, des groupes homogènes partageant des caractéristiques communes [Jain et al., 1999]. Il existe également de nombreuses mesures pour évaluer la qualité d'un partitionnement comme la densité, l'inertie inter et intra cluster, etc [Lerman and Azé, 2007].

Analyse prédictive. Basée sur l'analyse descriptive, l'analyse prédictive vise à anticiper des scénarios, en planifiant à l'avance, plutôt qu'en réagissant à ce qui s'est déjà passé. L'analyse prédictive propose des modèles pour prédire de nouveaux événements et ainsi aider aux choix de futures actions. Les analyses prédictives sont de nature probabiliste (classification, régression, analyse de séries temporelles, etc.). Elles permettent de prévoir ce qu'il pourrait se passer à partir des données descriptives accumulées au fil du temps. Par exemple, la prédiction permet d'estimer les valeurs futures de certaines variables, comme le nombre de lits dans un hôpital, la durée du séjour [Azari et al., 2012], etc. Si la variable est catégorielle, on parle de classification et sinon de régression [Jacob and Ramani, 2012]. Si la variable prédite dépend du temps, on parle de prévision dans les séries temporelles [Jacob and Ramani, 2012]. Ces prévisions sont généralement caractérisées par des mesures qui donnent des indications sur la confiance que l'on peut avoir dans la prédiction.

Analyse prescriptive. L'analyse prescriptive va au-delà des analyses précédentes en explorant un ensemble d'actions possibles puis en recommandant une ou plusieurs actions et en montrant les conséquences probables de chaque action. Il ne s'agit plus de prédire un futur possible mais tous les futurs possibles et de les évaluer pour choisir le meilleur. Par exemple, [Liu et al., 2013] ont proposé un mécanisme de recommandation de musique pour prévenir la somnolence au volant. Si l'analyse prescriptive est l'échelon le plus haut de la typologie, les méthodes d'optimisation et de simulation sur lesquelles elle se base ne sont pas nouvelles. Elles prennent généralement en compte l'incertitude et recommandent des solutions afin de limiter les risques qui résultent de chaque action.

Le schéma 1.2, adapté de [Dursun, 2014], résume les trois types d'analyse que nous venons de décrire. Dans la Section D, nous présentons les étapes de ce processus communes aux trois types d'analyse puis dans la Section E, nous mettons en relation les applications décrites dans la Section B et la typologie de méthodes d'analyse que nous venons juste de décrire dans cette section.

D Le processus d'analyse de données

Dans la communauté *Extraction de Connaissances* (**KDD** Knowledge Discovery in Database), le processus décrit dans la Figure 1.3 par [Fayyad et al., 1996] est couramment référencé. Nous le généralisons ici à l'analyse de données. Ce processus a pour objectif de *transformer des données de bas niveau sous d'autres formes plus compactes, plus abstraites ou plus utiles. Il comprend un ensemble d'étapes interactives et itératives* [Fayyad et al., 1996].

Les données en entrée de ce processus sont diverses par nature : numériques, symboliques, booléennes, multi-dimensionnelles, multi-sources, etc. Elles peuvent également se distinguer par leur structure : données ensemblistes, arborescentes, séquentielles, sous la forme de graphes, etc. Elles peuvent être incomplètes, entachées d'erreurs. Elles peuvent être dynamiques, évoluer dans le temps, arriver en flots, etc.

Les attendus de ce processus peuvent prendre des formes diverses. L'utilisateur souhaitera, par exemple, faire face à un problème de classification, en associant une pathologie à un patient en fonction de ses symptômes. Dans un service hospitalier, on pourra chercher à prédire le nombre de lits occupés pour la semaine suivante afin de prévoir le personnel nécessaire. On pourra souhaiter détecter des anomalies dans le fonctionnement d'un équipement médical afin de déceler au plus vite les signes annonciateurs d'une panne.

Ce processus peut donc être très complexe et les étapes peuvent varier considérablement en fonction de la nature des données et des objectifs de l'application.

Sélection des données Pour sélectionner les données, il faut tout d'abord déterminer les sources d'informations qui pourront être utiles. Couramment effectuée à l'aide de requêtes, cette première étape consiste à sélectionner, dans une base ou un entrepôt de données, les informations relatives au problème pour lequel on souhaite construire de nouvelles connaissances.

Pré-traitements Les données sélectionnées sont souvent incomplètes, bruitées, de qualité hétérogène ou bien ne correspondent pas au format d'entrée des algorithmes de fouille. Elles sont donc nettoyées et formatées de façon à pouvoir appliquer une technique de fouille de données.

Analyse des données C'est l'étape centrale du processus. L'algorithme est choisi selon le type des données et la problématique applicative. Les données sélectionnées et pré-traitées sont explorées avec un ou plusieurs algorithmes. Ces algorithmes peuvent par exemple générer un ensemble de motifs, des règles ou un regroupement par classe. Même si l'étape d'analyse de données n'est qu'une partie du processus général, elle est celle qui suscite le plus de travaux dans la littérature.

Restitution Les informations extraites ne sont souvent pas directement interprétables. Cette phase consiste à traiter le format de sortie des algorithmes pour restituer les résultats, les rendre facilement visualisables et analysables par les utilisateurs. Une fois seulement les motifs validés, on obtient des connaissances.

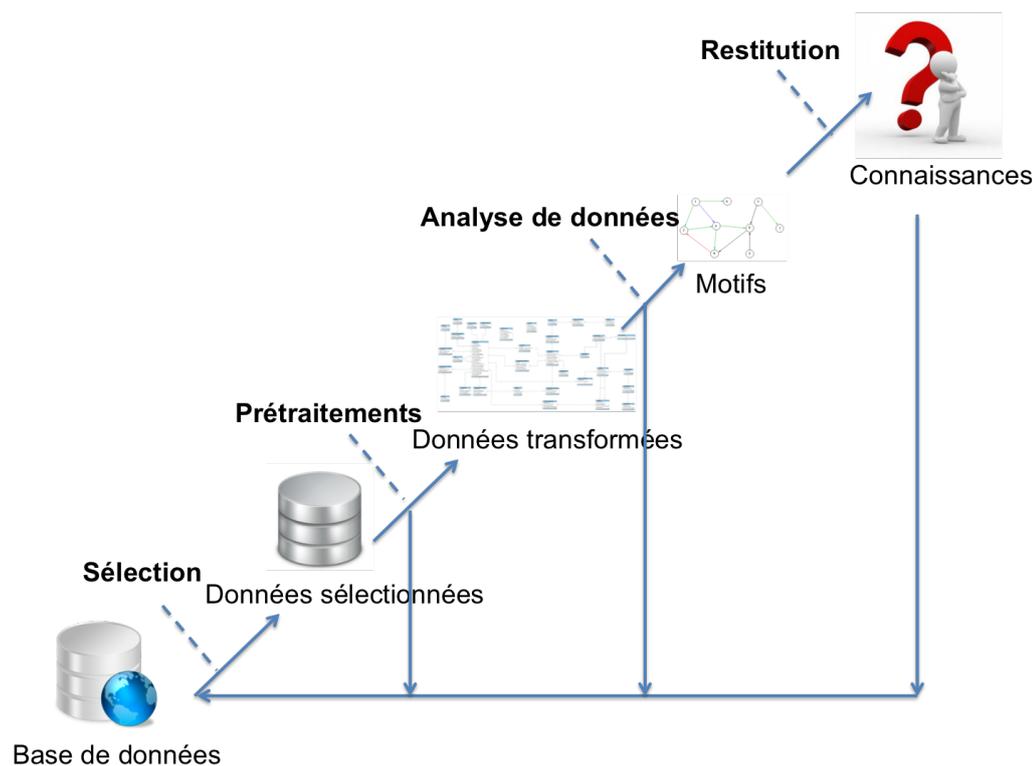


FIGURE 1.3 : Processus d'analyse de données, adapté de [Fayyad et al., 1996].

E Applications de l'analyse de données dans le domaine de la santé

Dans cette section, nous réalisons une revue des articles publiés sur le thème de l'analyse de données dans le domaine de la santé afin d'identifier les caractéristiques méthodologiques de ces études. Notre objectif n'est pas de réaliser une bibliographie exhaustive mais plutôt de donner au lecteur l'intuition de la richesse des méthodes d'analyse pour exploiter les données de santé. Concrètement, nous allons mettre en relation les types d'applications médicales identifiées dans la Section B et les types d'analyses réalisées décrites dans la Section C.

Pour constituer le corpus d'articles, nous avons utilisé les fonctionnalités avancées de l'API PUB-MED⁵. C'est le principal moteur de recherche de données bibliographiques de l'ensemble des domaines de spécialisation de la biologie et de la médecine. Il a été développé par le Centre américain pour les informations biotechnologiques (NCBI). Il donne accès à la base de données bibliographique MEDLINE qui contenait en juillet 2014 plus de 24 millions de citations publiées depuis 1950 dans environ 5 000 revues biomédicales⁶.

Dans un premier temps, nous avons cherché tous les articles dont les résumés/titres contenaient à la fois au moins un terme faisant référence à une des familles d'applications médicales telles que décrites dans la section B et à la fois au moins un terme faisant référence à une méthode d'analyse soit descriptive, soit prédictive, soit prescriptive. Nous avons sélectionné uniquement des articles publiés depuis 2000 en anglais à la date du 07 janvier 2015. Le tableau 1.1 liste les termes utilisés pour construire chacune des requêtes et un exemple de requête construite. Le tableau 1.2 reporte les résultats quantitatifs obtenus

5. <http://www.ncbi.nlm.nih.gov/pubmed>

6. <http://fr.wikipedia.org/wiki/PubMed>

avec ces requêtes. Les méthodes prédictives sont les plus utilisées (773 575), notamment les méthodes de classification (454 684) et de régression (365 500). Les deux thématiques médicales utilisant le plus des méthodes d'analyse sont l'efficacité des traitements (2 042) et la gestion des soins (2 181). Ces deux thématiques ne couvrent pas tous les articles associés aux méthodes d'analyse et nous pouvons en déduire que la catégorisation de référence décrite dans la littérature (voir Section B) ne semble pas appropriée.

TABLE 1.1 : Requetes utilisées pour réaliser la revue de la littérature.

Aspects médicaux
AP1 "Treatment effectiveness"
AP2 "Healthcare management" OR "Hospital Management"
AP3 "Customer relationship management"
AP4 "Fraud and abuse" OR "inappropriate prescriptions" OR "fraudulent insurance"
AP5 "genomic profiling" OR "genomics"
Aspects Analyse de données
DM1 "Descriptive analytics" OR "Association rules" OR "Sequential patterns" OR "Clustering" OR "Trajectory" OR "Correspondence analysis" OR "pattern recognition" OR "Itemset Mining"
DM2 "Predictive analytics" OR "Classification" OR "Regression" OR "Times series" OR "sequence discovery" OR "Ordinary least squares regression" OR "Logistic regression" OR "Neural networks" OR "Decision trees" OR "Memory-based reasoning" OR "Support vector machines" OR "Multi-adaptive regression splines" OR "Decision trees Clustering" OR "K means" OR "Neural networks" OR "Discriminant analysis" OR "Self-organizing maps" OR "Bagging and boosting ensembles" OR "Naïve Bayes classifiers" OR "Decision trees" OR "predictive modelling"
DM3 "Prescriptive analytics" OR "optimisation" OR "simulation" OR "Genetic Algorithm" OR "Neural networks" OR "nearest neighbor methods"
Exemple de requête combinant AP1 et DM3
("Treatment effectiveness") AND ("Prescriptive analytics" OR "optimisation" OR "simulation" OR "Genetic Algorithm" OR "Neural networks" OR "nearest neighbor methods")

TABLE 1.2 : Résultats quantitatifs des requêtes.

	Sans aspect analyse ↓	DM1	DM2	DM3	Total
Sans aspect médical →		98 425	773 575	226 168	1 098 168
AP1 Treatment effectiveness	2 042	7	233	57	297
AP2 Healthcare management	2 181	13	240	37	290
AP3 Customer relationship	45	1	4	1	6
AP4 Fraud and abuse	178	4	42	4	50
Total	4 446	26	519	98	643

Comme dans le Chapitre 2 de ce manuscrit, nous nous intéressons particulièrement aux méthodes d'analyse descriptive, nous nous focalisons dans ce paragraphe, sur la deuxième colonne du tableau 1.2 en gras et sur les 4 premières applications (AP1, AP2, AP3 et AP4), soit 26 articles. Pour chaque article, nous avons récupéré manuellement l'année de publication, le type d'article (e.g. cas d'étude, review), le thème (e.g. étude d'impacts environnementaux, diagnostic), le type de méthode appliquée (e.g. règles d'association) et le pays. Sur les articles (7) traitant de l'efficacité de traitement (AP1), tous sont des cas d'étude. La plupart ont été écrits récemment (6 depuis 2000) et traitent de l'impact des soins selon différentes maladies (e.g. impact des telecares - soins à distance sur la qualité de vie). L'analyse se limite pour la plupart à des pourcentages (2) et du clustering (5). Sur les articles (13) traitant de la gestion de soins (AP2), un a été exclu car hors sujet. Nous avons trouvé une seule revue de la littérature. Comme précédemment, les articles sont récents (12 depuis 2010) et traitent majoritairement de l'impact des soins (9). On trouve également 2 articles traitant de l'impact de l'environnement (e.g. les émissions des

transports sur l'asthme des enfants). Les méthodes de régression (6) et de clustering (4) sont les plus utilisées. On ne trouve qu'un seul article pour la catégorie Gestion du client (AP3) et un seul article pour la catégorie Fraude et abus (AP4). Les détails de ce pointage sont donnés en annexe.

On trouve peu d'articles à l'intersection des méthodes d'analyse et des 4 thématiques d'application identifiées dans la littérature. Ce constat s'explique par la difficulté de décrire les thématiques médicales avec quelques mots clés mais surtout par la non couverture par cette catégorisation pour l'ensemble des applications de l'analyse de données récemment étudiées. Nous avons donc exploré avec des méthodes sans *a priori* le contenu de ces articles. Pour cela, nous avons utilisé la librairie E-utilities fournie par le NCBI⁷ pour récupérer automatiquement tous les titres, résumés et années des articles traitant des méthodes d'analyse descriptive (DM1). Nous avons supprimé les mots outils et utilisé la méthode LDA [Blei et al., 2003] (voir Chapitre 3) pour extraire les 10 principaux thèmes associés à ces méthodes. Nous avons retenu les 30 mots les plus fréquents par thème que nous avons représenté par des nuages de mots (voir Figure 1.4).

Seul le premier thème est centré sur les patients, les liens entre facteurs de risque, traitements et mortalité. Les études portent essentiellement sur les maladies chroniques comme le cancer et le HIV. Les méthodes les plus utilisées sont l'identification de trajectoires et le clustering. Le deuxième thème porte sur le cerveau, la mémoire et son développement chez les enfants dans le cadre d'études longitudinales. Les méthodes les plus utilisées sont l'identification de trajectoires. Le troisième thème identifié porte sur l'analyse des mouvements du corps et les méthodes les plus utilisées sont la génération et le suivi des trajectoires. Le quatrième thème porte sur l'impact de l'environnement (e.g. climat, eau) sur le développement des communautés et des espèces. Les méthodes les plus fréquentes sont la reconnaissance de motifs. Dans le cinquième thème associé au domaine de la chimie, nous trouvons des méthodes d'analyse de spectrométrie de masse avec essentiellement du clustering hiérarchique. Avec les 3 thèmes suivants, nous entrons dans le domaine de la biologie. Le thème 6 porte sur la dynamique moléculaire des protéines, le thème 7 porte sur l'expression des gènes dans les cellules (*Cell signaling, pathways*) et le thème 8 porte sur l'analyse des séquences du génome, l'étude des espèces et des populations. Les méthodes les plus utilisées sont l'identification de trajectoires, le clustering et les simulations. Les deux derniers thèmes ont regroupés les éléments relatifs aux méthodes : modèle, clustering, classification, sélection d'attributs, trajectoire, espace, temps et apprentissage, etc.

Si l'étude réalisée dans cette section n'est pas exhaustive et devra être approfondie, elle donne déjà l'intuition de la richesse des méthodes d'analyse de données pour des applications médicales, bien plus étendue que celle annoncée dans les revues de la littérature présentées dans la Section B. Nous prévoyons de réaliser prochainement une étude systématique en appliquant la méthode PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [Moher et al., 2009] dont le but est d'aider les auteurs à rédiger des revues systématiques et des méta-analyses de la littérature. La méthodologie se compose d'une liste de 27 items à compléter et d'un processus en quatre phases à respecter. Les auteurs distinguent les revues systématiques et les méta-analyses. Une revue systématique correspond à la revue d'une question clairement formulée et qui utilise des méthodes systématiques et explicites pour identifier, sélectionner et évaluer de façon critique des articles de recherche répondant à cette question puis qui recueille et analyse les données extraites des publications décrivant les études incluses. Une méta-analyse peut être utilisée en complément d'une revue de la littérature pour analyser et résumer les résultats des études. Cette méta-analyse se base sur des techniques statistiques. Pour conclure, cette revue bibliographique, nous a permis d'identifier différents challenges détaillés dans la section suivante F, que nous allons mettre en perspective avec le travail de recherche et d'encadrement de la recherche mené dans l'équipe ADVANSE.

7. <http://www.ncbi.nlm.nih.gov/books/NBK25500/>



FIGURE 1.4 : Nuages de mots des 10 thèmes associés à l'application des méthodes descriptives sur les données de santé.

F Challenges à venir pour l'application de l'analyse de données en santé

Nous synthétisons et complétons dans cette section les limitations et challenges associés aux méthodes d'analyse identifiés dans plusieurs revues de la littérature [El-Sappagh et al., 2013, Koh and Tan, 2005, Yang and Parthasarathy, 2006]. Nous montrons également dans les encadrés de couleurs que les travaux réalisés et décrits dans la suite de ce manuscrit d'Habilitation à Diriger les Recherches sont au cœur des challenges actuels.

En 2001, un rapport de recherche du META Group (devenu Gartner) définit les enjeux inhérents à la croissance des données (Big Data) comme étant tri-dimensionnels⁸. Les méthodes d'analyses de données doivent se confronter aux "3V" : Volume, Vitesse et Variété. De nos jours, ce modèle est toujours utilisé pour décrire ce phénomène et s'applique tout à fait au cas des données de santé.

1. **Volume** : les bases de données cliniques sont très grandes avec parfois des centaines de tables et de champs et des millions d'enregistrements. Par exemple, le Système national d'information inter-régimes de l'assurance maladie (SNIRAM⁹), contient plus de 10 milliards d'informations sur les prescriptions de médicaments, les consultations, les tarifs, les maladies, etc. pour toute la population française, ce qui est quasi unique au monde. Cette base offre de grandes perspectives d'analyse notamment en pharmacovigilance. Les méthodes d'analyse doivent être efficaces pour traiter ces gros volumes de données en intégrant si besoin des aspects parallèles pour le passage à l'échelle [Simoes et al., 2013, Jackin et al., 2014].

Dans le Chapitre 2, nous allons décrire différentes méthodes d'analyse descriptive basées sur la reconnaissance de motifs et adaptées aux données séquentielles, qui permettent d'extraire des pépites de connaissances à partir de gros volumes de données médicales. En particulier, nous évoquerons dans la Section 1.2 des travaux en cours et qui visent à extraire des trajectoires de patients dans une grande base de données issue du PMSI. Nous utiliserons une approche descriptive originale basée sur l'extraction de motifs séquentiels contextuels qui assure entre autre, le passage à l'échelle.

2. **Variété** : les données de santé sont par nature extrêmement variées. On trouve les données structurées et catégorielles comme la plupart des données administratives, des données numériques comme les résultats de tests de laboratoire non toujours standardisés et dépendant des matériels médicaux utilisés par les établissements de soin, des données textuelles comme les compte rendus d'hospitalisation, les notes des médecins et certains dossiers cliniques, des images comme les radiographies, scanners, etc. L'analyse doit pouvoir tenir compte de toute cette hétérogénéité. Par ailleurs, l'efficacité des analyses repose sur la qualité de ces données. L'analyse doit rester efficace avec des données manquantes, bruitées, corrompues, inconsistantes, non standardisées, etc. Dans ce contexte, un challenge important est l'analyse des données textuelles. Si de nombreux travaux ont défini des vocabulaires contrôlés [Parès et al., 2014, Charlet et al., 2014] pour codifier les informations textuelles dans l'objectif de futures analyses, la fouille de textes libres reste difficile mais nécessaire par exemple pour exploiter dans les dossiers des patients les diagnostics ou résumer l'historique d'un patient.

8. <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data/-Volume-Velocity-and-Variety.pdf>

9. <http://www.ameli.fr/1-assurance-maladie/statistiques-et-publications/sniiram/finalites-du-sniiram.php>

Dans le cadre d'un travail portant sur l'analyse des médias sociaux, décrit dans le Chapitre 3 de ce manuscrit, nous nous intéressons à l'extraction de connaissances dans les textes libres rédigés par les patients. Il s'agit d'une tâche encore plus difficile par rapport à l'analyse des productions des professionnels de la santé du fait de la variabilité dans l'expression personnelle et intime de ces patients. Nous avons défini pour cela des méthodes de fouille de textes innovantes pour enrichir le contenu des messages. Nous les utilisons actuellement dans deux projets portant sur l'analyse de la qualité de vie des patientes atteintes d'un cancer du sein et la détection de personnes suicidaires dans les réseaux sociaux.

3. **Vélocité** : le système de soins est alimenté en temps réel par de très nombreuses informations. Un défi est de concevoir des outils d'analyse s'adaptant aux contraintes imposées par ces flux de données. Un outil d'analyse doit savoir apprendre au cours du temps, s'adapter aux évolutions de la population réelle, s'enrichir automatiquement avec ces nouvelles connaissances, mettre à jour en continu les modèles de connaissances pour correspondre aux évolutions du monde réel. L'analyse doit être active et notamment déclencher des alertes quand les données qui arrivent en flux diffèrent des données attendues. L'analyse doit rester semi-automatique car la prise de décision est de la responsabilité du professionnel de santé et si celui-ci doit intervenir à minima, l'outil ne peut que l'aider dans ce processus [Pitarch et al., 2010].

Dans ce contexte, nous avons travaillé sur la détection d'anomalies décrite dans la Section 1.2 du Chapitre 2 en appliquant une analyse descriptive basée sur l'extraction de motifs contextuels, pour identifier des comportements déviants à partir de données de capteurs arrivant en flots et en très grandes quantités.

D'autres problématiques, non liées aux big data, sont importantes à prendre en compte lorsque l'on traite des données de santé.

Les analyses doivent considérer des aspects longitudinaux, temporels et spatiaux. Par exemple, via le système de soins français, les chercheurs construisent des cohortes, qui consistent à suivre un groupe d'individus durant une période de temps déterminée. L'application de diverses mesures à des temps précis et stratégiques permet de mesurer l'évolution du phénomène étudié. Lors du recueil, il est possible d'inclure des informations temporelles et spatiales (comme la localisation des patients) qu'il est important de prendre en compte dans le processus d'analyse.

Nous décrivons dans la Section 1.3 du Chapitre 2, une méthode d'analyse descriptive originale portant sur l'extraction de motifs spatio-temporels, très pertinente pour des applications de veille épidémiologique. Nous avons appliqué cette méthode pour étudier l'évolution des épidémies de dengue.

Un autre aspect important lors de l'application de méthodes d'analyse aux données de santé est l'évaluation. Pour un seul problème, il y a très souvent un très grand nombre de techniques qu'il est possible d'appliquer. Il faut alors définir des critères pour comparer ces techniques et identifier la meilleure.

Dans la Section 1 du Chapitre 2, nous montrerons comment nous avons comparé des méthodes de classification classiques à la méthode que nous avons proposée pour la caractérisation du grade du cancer du sein. Une réflexion plus générale sur les méthodes d'évaluation est nécessaire notamment quand les méthodes nécessitent des données annotées (voir perspective du Chapitre 3 Section 3.6).

Le choix des algorithmes se base souvent sur la valeur de rappel car la santé est une question de vie et de mort et il ne faut pas rater de vrais positifs. Les questions de performances (temps de calcul, nombre maximum de données en entrée, etc.) sont également importantes dans un contexte temps réel. D'après notre expérience, l'utilité médicale du résultat de la méthode d'analyse et son interprétabilité sont tout

aussi essentielles. La méthode doit aider le professionnel de la santé dans sa pratique. Il ne peut utiliser l'outil d'analyse comme une boîte noire. Il doit être capable de comprendre les résultats de l'analyse.

Les méthodes à base de motifs développées et décrites dans la Section 2 du Chapitre 2 permettent un niveau d'interprétation important pour les professionnels de la santé. Toutefois, le très grand nombre de motifs générés rend cette interprétation difficile. Nous montrerons donc également quels sont les méthodes et outils (e.g. mesures d'intérêt, visualisations, etc.) que nous avons développés pour assurer l'interprétabilité par les professionnels de santé des sorties des processus d'analyse.

L'analyse peut être limitée par l'accessibilité aux données. En effet, les données utilisées en entrée des outils d'analyse existent souvent dans des contextes et systèmes différents (e.g. bases de données administratives, données cliniques, laboratoires, etc.) et sont possédées par des acteurs différents (hôpitaux, centres de recherche, assurance, etc.). Par conséquent, les données doivent être collectées et intégrées avant de pouvoir réaliser l'analyse. Des questions éthiques et juridiques peuvent alors se poser, telles que l'accès à la vie privée. *Comment assurer que la vie privée des patients est conservée suite au processus d'analyse ?* Diverses formes d'anonymisation sont possibles [Martinezemail et al., 2013, Domingo-Ferrer, 2008].

Comme nous l'évoquerons dans les perspectives du Chapitre 4, la question de la diversité des données est centrale à la collaboration que nous menons avec l'ICM sur la qualité de vie des patientes atteintes par un cancer du sein. Nous visons à comparer des données hétérogènes issues de questionnaires structurés remplis par les patients et de textes libres issus de leurs messages dans les forums de santé. Par ailleurs, nous posons également de nombreuses questions éthiques et juridiques et travaillons en étroite collaboration avec des juristes, dans le cadre d'un autre projet visant à identifier les personnes à risque suicidaire dans les données issues des réseaux sociaux. *Quelle forme de consentement est nécessaire ? Jusqu'où a-t-on le droit d'intervenir dans la vie des patients quand l'analyse découvre un risque potentiel ?*

Tous ces challenges sont au cœur des travaux réalisés et à venir.

FORMALISATION DE LA CONNAISSANCE MÉDICALE SOUS FORME DE MOTIFS

Sommaire

- A Introduction**
 - B Des méthodes descriptives efficaces basées sur les motifs**
 - 1 Décrire des données de plus en plus riches sémantiquement
 - 2 Choisir, représenter et interpréter les meilleurs motifs
 - C Vers des approches prédictives et prescriptives basées sur les motifs**
 - 1 Prédire le grade d'un cancer
 - 2 Prédire la catégorie d'un patient et le prochain évènement
 - 3 Prescrire en détectant une anomalie et en déclenchant une alerte
 - D Conclusions et perspectives**
 - 1 Des motifs de plus en plus riches sémantiquement
 - 2 Sélection et visualisation des motifs
 - 3 De nouvelles applications aux motifs, utiles aux professionnels de la santé
-

A Introduction

Parmi les méthodes d'analyse de données, la *découverte de motifs* (PD Pattern Discovery) a connu une croissance assez spectaculaire ces dernières années, sous l'impulsion des organisations propriétaires de grands volumes de données et soucieuses d'en extraire de la valeur ajoutée. La découverte de motifs cherche à induire des lois générales à partir des données stockées. Ces méthodes ont été beaucoup étudiées en bioinformatique pour l'analyse des données génomiques à large échelle et dans de nombreux domaines d'applications où des régularités peuvent être porteuses de valeur ajoutée (e.g. la découverte de règles d'associations dans des données transactionnelles [Agrawal et al., 1993], de motifs séquentiels dans des bases de séquences de comportements [Agrawal and Srikant, 1995, Mannila et al., 1997, Masegla et al., 1998] ou la découverte de motifs plus complexes tels que des sous-graphes [Inokuchi et al., 2000] ou des sous-arbres [Termier et al., 2002, Zaki, 2002], etc). Historiquement, l'équipe ADVANSE¹⁰ s'est focalisée sur l'extraction de motifs dans des grandes bases de données. Les premiers travaux ont porté sur l'extraction de motifs séquentiels (i.e. expression de régularités temporelles selon un ordre strict). Elles ont donné lieu à l'approche PSP [Masegla et al., 1998] citée plus de 313 fois sur Google Scholar, ce qui montre l'intérêt de la communauté pour ces thématiques.

À mon arrivée dans l'équipe en 2007, je me suis intéressée à l'apport des motifs séquentiels pour l'analyse des données de santé. Via diverses collaborations, nous nous sommes confrontés à des données médicales très différentes. Nous avons eu besoin d'étendre ces motifs pour prendre en compte des informations contextuelles et spatiales afin d'enrichir la lecture qu'un expert peut faire des motifs. Nous avons également eu le besoin d'étendre ces motifs pour intégrer un ordre temporel non strict afin de condenser l'information présente dans plusieurs motifs. Nous avons proposé de nouvelles formalisations pour décrire ces informations, des algorithmes efficaces pour l'extraction de ces motifs sur de gros volumes de données et des scénarios d'utilisation originaux dans le contexte médical.

Nous détaillons dans la suite ces différents travaux et les collaborations et encadrements associés. Dans la Section B, nous présentons les méthodes descriptives mises en places. Nous commençons dans la Section B.1 par la définition de nouveaux motifs : extraction de motifs séquentiels dans les puces à ADN (voir 1.1), de motifs contextuels dans les bases de type PMSI (voir 1.2), de motifs spatio-séquentiels dans les bases épidémiologiques (voir 1.3) et finalement de motifs partiellement ordonnés de nouveau dans les puces à ADN (voir 1.4). Nous décrivons ensuite, dans la Section B.2, différentes mesures de sélection de ces motifs et leur combinaison (voir 2.1) ainsi que des visualisations innovantes (voir 2.2) dédiées à l'exploration des experts pour faciliter l'appropriation de ces motifs. Dans la Section C, nous présentons ce qu'il est possible de faire à partir de ces motifs pour des applications prédictives et prescriptives : comment utiliser les motifs séquentiels pour prédire le grade du cancer (voir C.1), comment utiliser les motifs contextuels pour prédire la catégorie d'un patient et les événements à venir (voir C.2) et détecter des anomalies (voir C.3). Nous concluons dans la Section D. Les perspectives associées à ces différents travaux seront présentées dans le Chapitre 4 de Conclusions et Perspectives.

10. <http://www.lirmm.fr/recherche/equipes/advanse>

B Des méthodes descriptives efficaces basées sur les motifs

1 Décrire des données de plus en plus riches sémantiquement

Dans cette section, nous décrivons les différents motifs que nous avons définis pour prendre en compte la sémantique de plus en plus riche associée aux données séquentielles rencontrées lors de nos collaborations. Par soucis de concision, pour chacun de ces motifs, nous donnerons l'intuition de leur formalisation et de l'algorithme d'extraction en illustrant sur des exemples médicaux. Pour plus de détails, nous invitons le lecteur à se reporter aux articles cités.

1.1 Motifs séquentiels

TABLE 2.1 : Résumé des projets et des encadrements sur la thématique des motifs séquentiels.

Master/Thèse	Projets & Collaborations
P. Salle (2007-2010) – Thèse co-encadrement M. Teisseire	MMDN - ANR Pradnet (2007-2011)

Dans le cadre d'une collaboration avec le MMDN¹¹ puis via l'ANR PRADNET financée par la Fondation de Coopération Scientifique Maladie d'Alzheimer et Maladies Apparentées, nous avons cherché à obtenir des connaissances inattendues pour les biologistes. Pour cela, nous avons reconsidéré le problème de l'extraction des motifs séquentiels sur des données temporelles en organisant les données pour prendre en compte le degré d'expression des gènes (Thèse de P. Salle). Cette nouvelle approche a permis aux biologistes de faire apparaître de nouvelles corrélations entre gènes [Salle et al., 2009].

La table 2.2 décrit les données sur lesquelles nous avons appliqué cette méthode. Une ligne correspond à une puce et une colonne à un gène. À l'intersection d'une ligne et d'une colonne, se trouve la valeur d'expression du gène mesurée par la puce (e.g. le gène $G1$ a pour expression 6.76 dans la puce $P1$). Nous construisons des séquences pour chaque puce en ordonnant les gènes selon leur valeur d'expression. Dans le vocabulaire classique d'extraction de motifs, les gènes sont des *items*. Un *itemset* est un groupe non ordonné de gènes ayant une valeur d'expression similaire. Par exemple, $G2$ et $G3$ sont regroupés dans le même itemset car ils ont la même expression. En considérant un écart minimal égal à 0.1, $G5$ peut être regroupé dans deux itemsets : soit $it_1 = (G2\ G3\ G5)$ soit $it_2 = (G1\ G5)$. Une séquence $S = \langle it_a it_b \dots it_p \rangle$ est une liste non vide et ordonnée de p itemsets, i.e. de groupes de gènes ordonnés selon leur valeur d'expression. Pour chaque puce, nous générons autant de séquences qu'il y a de combinaisons d'itemsets possibles. Nous obtenons ainsi une *base de séquences*.

La table 2.3 donne un exemple de séquences associées à un écart minimal de 0.1, utilisées en entrée de l'algorithme d'extraction de motifs séquentiels. Un *motif séquentiel* est une sous-séquence fréquente de gènes. Un motif est supporté par une puce si ce motif est inclus dans une ou plusieurs séquences associées à la puce. Par exemple, le motif $M_1 = \langle (G2)(G5) \rangle$ est inclus dans l'une des deux séquences (ou les deux) associées à la puce $P1$. Donc M_1 est supporté par $P1$. Le *support* du motif M_1 correspond au pourcentage de puces qui supportent M_1 , $support(M_1) = 4/5$. Afin de ne conserver que les motifs les plus fréquents, un *support minimum* est fourni par le décideur. On n'extrait que les motifs ayant un support supérieur à ce seuil. Par exemple, si on fixe un support minimum à $2/5$, alors M_1 est fréquent mais $M_2 = \langle (G4)(G1) \rangle$ ne l'est pas car $support(M_2) = 1/5$.

11. <http://www.mmdn.univ-montp2.fr/>

Le motif séquentiel de gènes $\langle (G2\ G5)(G4) \rangle_{80\%}$ s'interprète ainsi : fréquemment (pour 80% des puces), les gènes $G2$ et $G5$ ont des niveaux d'expression similaires. Tous deux s'expriment moins fort que le gène $G4$.

TABLE 2.2 : Expression des gènes pour la puce à ADN $P1$ et deux séquences associées avec l'écart minimal de 0.1.

Gènes	G1	G2	G3	G4	G5
$P1$	6.76	6.65	6.65	9.65	6.75
Séquences associées	$\langle (G2\ G3\ G5)(G1)(G4) \rangle$				
	$\langle (G2\ G3)(G5\ G1)(G4) \rangle$				

TABLE 2.3 : Séquences obtenues à partir de 5 puces à ADN. Selon un écart minimum de 0.1 fixé par l'utilisateur, deux séquences sont associées à $P1$ et $P3$ et une séquence est associée à $P2$, $P4$ et $P5$. Les deux dernières colonnes illustrent le calcul du support pour deux motifs. Le premier sera fréquent mais pas le deuxième.

Puces	Séquences de gènes associés	Inclusion $\langle (G2)(G5) \rangle$	Inclusion $\langle (G4)(G1) \rangle$
$P1$	$\langle (G2\ G3\ G5)(G1)(G4) \rangle$		
	$\langle (G2\ G3)(G5\ G1)(G4) \rangle$	✓	
$P2$	$\langle (G1)(G4)(G2)(G3)(G5) \rangle$	✓	
$P3$	$\langle (G2\ G5)(G4)(G3)(G1) \rangle$		
	$\langle (G2)(G5\ G4)(G3)(G1) \rangle$	✓	✓
$P4$	$\langle (G1)(G4)(G5)(G2)(G3) \rangle$		
$P5$	$\langle (G1)(G2)(G3)(G4)(G5) \rangle$	✓	

En utilisant l'approche précédente, nous avons obtenu des motifs séquentiels qui ne sont pas toujours facilement compréhensibles et manipulables par les experts car ils ne permettent pas de quantifier la différence d'expression des gènes. Par exemple, si l'on considère le motif $\langle (G1\ G5)(G3) \rangle$, il n'est pas possible de savoir de combien l'expression de $G3$ est supérieure par rapport aux deux autres gènes $G1$ et $G5$. Si l'on considère la figure 2.1, dans le premier cas, l'expression de $G3$ est beaucoup plus grande que celle des gènes $G1$ et $G5$. Dans le deuxième cas, l'écart est réduit et dans le troisième l'écart est le plus petit par rapport au niveau de discrétisation envisagé. Or, un seul motif sera extrait avec la méthode traditionnelle pour ces trois cas. En biologie, les experts ont l'habitude d'interpréter des différences d'expressions entre gènes et ils ne retrouvent pas cette information dans les motifs obtenus avec l'approche classique. Nous avons donc proposé (Thèse de P. Salle) d'apporter cette information supplémentaire sur la lecture des motifs séquentiels en exploitant la logique floue en post-traitement comme [Fiot et al., 2007] pour quantifier les écarts entre itemsets [Bringay et al., 2009].

Le motif séquentiel de gènes à écarts flous $\langle (G1\ G5)(over\ expressed\ 0.8)(G3) \rangle_{80\%}$ s'interprète ainsi : fréquemment (pour 80% des puces), le gène $G3$ est exprimé **beaucoup plus fort** que les gènes $G1$ et $G5$ dont les expressions sont similaires. Le 0.8 décrit l'importance de l'appartenance de l'écart à la classe "over expressed"^a.

^a. Dans la suite de ce manuscrit, dans les exemples de motifs, nous ne reprendrons pas les éléments liés à la fréquence en sachant que ce critère de sélection sera toujours employé.

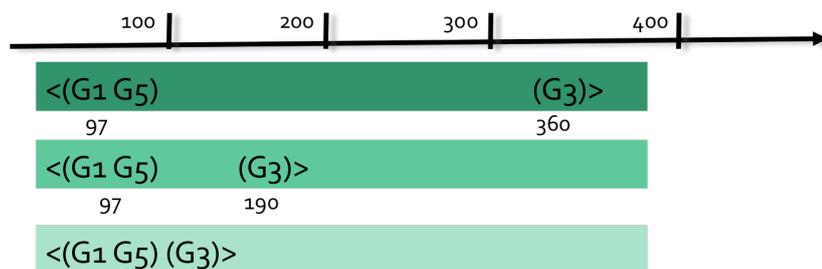


FIGURE 2.1 : Trois écarts pour un même motif : dans le premier cas, G_3 a une expression beaucoup plus forte que G_1 et G_5 qui ont une expression similaire, dans le deuxième cas, l'écart est réduit et dans le troisième l'écart est le plus petit vis à vis de la discrétisation envisagée.

Sélection de références

P. Salle, S. Bringay and M. Teisseire. Mining Discriminant Sequential Patterns for Aging Brain. Proceeding AIME'09. 18-22 July, Verona, (Italy), 2009 : 365-369 (**Rang A**)

S. Bringay, A. Laurent, B. Orsetti, P. Salle, M. Teisseire. Handling fuzzy gaps in sequential patterns : Application to health. FUZZ-IEEE, 2009, Jeju (Korea) : 1338-1345 (**Rang A**).

1.2 Motifs contextuels

TABLE 2.4 : Résumé des projets et des encadrements sur la thématique des motifs contextuels.

Master/Thèse	Projets & Collaborations
J. Rabatel (2008-2011) – Thèse co-encadrement P. Poncelet	Tecnalia (Espagne) (2008-2011)
J. Pinaire (Depuis 2014) – Thèse co-encadrement J. Azé, P. Landais	CHU Nîmes (Depuis 2014)

Les motifs séquentiels présentés précédemment ne tiennent généralement pas compte des informations contextuelles fréquemment associées aux données séquentielles. Par exemple, dans le cas des séquences d'actes réalisées pour des patients dans un hôpital, l'extraction classique de motifs séquentiels se focalise sur les séries d'actes sans considérer leur sexe, leur âge, leur poids, etc. Or, en considérant le fait qu'un motif séquentiel est spécifique à un contexte donné (e.g. les jeunes hommes), un professionnel de santé pourra adapter sa stratégie de soin au contexte du patient et prendre les décisions adéquates. Dans le cadre d'une collaboration avec la société Tecnalia (Thèse CIFRE de J. Rabatel)¹², nous avons redéfini des motifs dits *motifs contextuels* que nous décrivons dans la suite de cette section. Depuis 2014, dans le cadre d'une collaboration avec le CHU de Nîmes avec le Professeur P. Landais (Thèse de J. Pinaire), nous appliquons cette méthode sur les données du PMSI (Programme Médicalisé des Systèmes d'Information) dans le contexte de la cardiologie. L'objectif est de caractériser des trajectoires de patients pour la prise de décision médicale.

Considérons la base de données présentée dans le tableau 2.5 qui décrit différents actes médicaux (symbolisés par des lettres) effectués au cours du temps par des professionnels de santé entre 7h et 11h pour des patients dans un service. Les activités peuvent correspondre dans la vie réelle à : *petit-déjeuner*, *prise de température*, *bain*, *nouvelle perfusion*, *kinésithérapie*, etc. Ces données sont séquentielles car elles présentent des événements (les actes) disposés suivant un ordre (le temps). Par exemple, nous

12. Tecnalia <http://www.tecnalia.com/en/> Les données étudiées dans cette thèse sont des données séquentielles correspondant à des capteurs placés sur des trains. Pour la clarté de ce manuscrit d'HDR, les exemples ont été adaptés au contexte médical.

constatons que pour le patient P_1 , les actes a et d ont été réalisés ensembles entre 8h et 9h puis l'acte b entre 10h et 11h. En examinant le tableau 2.6, nous constatons également que le motif « a suivi plus tard par b » est vérifié par plus de 50% des patients (8 sur 14). En supposant que le professionnel de santé précise qu'il est intéressé par des actes qui apparaissent dans au moins 50% des cas (support minimum) de la base alors le motif $\langle(a)(b)\rangle$ est fréquent. Cet exemple considère la base comme un ensemble indivisible pour la recherche des motifs. Pourtant, les circonstances liées aux données impliquent l'existence de sous-ensembles de données rassemblant des propriétés similaires. Pour notre cas d'étude, par exemple, nous pouvons intégrer des informations supplémentaires comme dans le tableau 2.7 qui associe à chaque patient son âge (*jeune* ou *âgé*) et son sexe (*homme* ou *femme*). Ces informations contextuelles peuvent avoir une influence non négligeable sur ce qui se produit dans les données et l'extraction de motifs doit rendre cette influence perceptible pour l'utilisateur afin de lui offrir une vue contextualisée des données.

TABLE 2.5 : Exemple de base de données d'actes médicaux au cours du temps. Chaque ligne correspond à un patient et chaque colonne à un créneau horaire. Les lettres {a,b,c,d,e} sont des actes médicaux.

Patients	7h-8h	8h-9h	9h-10h	10h-11h
P_1		a,d		b
P_2	a,b		b	
P_3	a	a		b
P_4	c	a		b,c
P_5	d	a,b	b,c,d	
P_6		b		a
P_7		a	b	a
P_8	d	a		b,c
P_9		a,b	a	b,d
P_{10}			b,c,d	
P_{11}			b,d	a
P_{12}	e	b,c,d		a
P_{13}		b,d,e		
P_{14}	b		a	e

TABLE 2.6 : Mise en valeur du motif $\langle(a)(b)\rangle$ (en gras) soit l'acte a suivi de l'acte b . Ce motif est fréquent dans la base pour un support minimum de 50%.

Patients	7h-8h	8h-9h	9h-10h	10h-11h
P_1		a ,d		b
P_2	a ,b		b	
P_3	a	a		b
P_4	c	a		b ,c
P_5	d	a ,b	b ,c,d	
P_6		b		a
P_7		a	b	a
P_8	d	a		b ,c
P_9		a ,b	a	b ,d
P_{10}			b,c,d	
P_{11}			b,d	a
P_{12}	e	b,c,d		a
P_{13}		b,d,e		
P_{14}	b		a	e

TABLE 2.7 : Mise en valeur du motif $\langle(a)(b)\rangle$ (en gras) avec les informations contextuelles sur l'âge et le sexe. Ce motif est spécifique aux jeunes. Une seule personne âgée est concernée.

Patients	Age	Sexe	7h-8h	8h-9h	9h-10h	10h-11h
P_1	jeune	homme		a,d		b
P_2	jeune	homme	a,b		b	
P_3	jeune	homme	a	a		b
P_4	jeune	homme	c	a		b,c
P_5	jeune	homme	d	a,b	b,c,d	
P_6	jeune	femme		b		a
P_7	jeune	femme		a	b	a
P_8	jeune	femme	d	a		b,c
P_9	âgé	homme		a,b	a	b,d
P_{10}	âgé	homme			b,c,d	
P_{11}	âgé	homme			b,d	a
P_{12}	âgé	femme	e	b,c,d		a
P_{13}	âgé	femme		b,d,e		
P_{14}	âgé	femme	b		a	e

Considérons maintenant le motif $\langle(a)(b)\rangle$ dans le tableau 2.7, nous constatons que :

- ces actes sont fréquents dans la population jeune (7 jeunes sur 8) mais pas dans la population âgée (seulement 1 personne sur 6) ;
- ces actes demeurent fréquents chez les jeunes quel que soit leur sexe (5 jeunes hommes sur 5 et 2 jeunes femmes sur 3).

Savoir qu'un comportement est spécifique ou général à un contexte est alors utile pour une interprétation médicale. Dans cette approche, la propriété "d'être fréquent" dépend d'un contexte donné.

Le motif $\langle(a)(b)\rangle$ est un motif contextuel associé au contexte *jeune* qui s'interprète ainsi : la réalisation de l'acte médical *a* suivi de *b* est à la fois spécifique aux jeunes (car non fréquent dans la catégorie complémentaire âge) et représentatif (car fréquent pour tout type de jeunes patients qu'il soit un homme ou une femme).

De manière générale, l'extraction de motifs est un problème difficile qui nécessite de naviguer dans un très grand espace de recherche. La prise en compte des contextes étend encore cet espace de recherche. Via des propriétés théoriques intéressantes basées sur la fréquence et sur les propriétés formelles associées au treillis formé par les contextes, nous avons développé un algorithme très efficace pour l'extraction de ces motifs contextuels. Les expérimentations effectuées sur différents jeux de données réelles ont montré l'efficacité de l'approche proposée [Rabatel et al., 2010]. Cette approche a été généralisée à d'autres mesures d'intérêt que la fréquence, utiles pour la sélection des motifs tels que le gain d'information, le taux d'émergence, la confiance, etc. dont l'objectif est de s'appuyer sur les caractéristiques statistiques des motifs pour isoler les plus intéressants au sens de critères expert.

Sélection de références

- J. Rabatel, S. Bringay and P. Poncelet. *SO_MAD* : SensOr Mining for Anomaly Detection in Railway Data. Proceeding ICDM, July 20 - 22, Leipzig (Germany), 2009 : 191-205 (**Best paper selection**).
- J. Rabatel, S. Bringay and P. Poncelet. Anomaly Detection in Monitoring Sensor Data for Preventive Maintenance. *Journal Expert Systems with Applications*, 38, 2010 : 7003-7015 (**Impact Factor : 2,908**)
- J. Rabatel, S. Bringay and P. Poncelet. Mining Representative Frequent Patterns in a Hierarchy of Contexts. *IDA 2014* : 239-250.

1.3 Motifs spatio-temporels

TABLE 2.8 : Résumé des projets et des encadrements sur la thématique des motifs spatio-temporels.

Master/Thèse	Projets & Collaborations
H. Alatrística Salas (2010-2013)	Univ. Nouvelle Calédonie
Thèse co-encadrement F. Flouvat, N. Selmaoui et M. Teisseire	InVS, DASS, Institut Pasteur

Ces dernières années, l’explosion du nombre de sources d’informations spatiales et de systèmes d’information géographique (GIS), a fait émerger de nouveaux défis en matière d’analyse de données. En effet, les avancées technologiques en terme d’acquisition (e.g. images satellitaires, capteurs) permettent d’associer une information spatiale aux données séquentielles. Ces données sont difficiles à appréhender par leur seule lecture. La mise en place de moyens analytiques leur permettant de faire sens pour les experts est un enjeu critique. Les applications associées à ces analyses sont alors multiples comme la détection de changements abrupts (e.g. catastrophes naturelles), le suivi de phénomènes évolutifs (e.g. érosion côtière, désertification), le suivi de phénomènes qui se propagent (e.g. épidémies, pollution de rivières), etc.

Dans le cadre de la thèse de H. Alatrística Salas, nous nous sommes particulièrement intéressés à la dynamique d’évènements spatio-temporels [Yuan, 2009], en collaboration avec l’Université de la Nouvelle Calédonie, la DASS (Direction des Affaires Sanitaires et Sociales) et l’institut Pasteur. Ces phénomènes sont généralement associés à des entités spatiales (e.g. rivières, villes) correspondant aux zones où se déroulent les évènements, représentés par des caractéristiques statiques (e.g. le cours de la rivière, la surface d’une ville) qui ne changent pas dans une période de temps fixe [Asproth et al., 1995]. Des informations dynamiques sont associées aux informations géo-référencées (e.g. force du vent, température) et évoluent au cours du temps. Un phénomène spatio-temporel est un processus lié au changement, parfois récurrent ou périodique, pour lequel les caractéristiques d’un ensemble d’entités spatiales sont exprimées par des séries ou séquences d’évènements [Nadi and Delavar, 2003]. Par exemple, la dynamique d’une épidémie de dengue correspond aux interactions entre la pluie, le développement de points d’eau et des moustiques et tous autres facteurs contribuant à son évolution et que nous souhaitons découvrir.

Pour capter cette dynamique, nous avons défini de nouveaux motifs dits *motifs spatio-séquentiels*. La Figure 2.2 représente la propagation d’une épidémie pour trois zones Z_1 , Z_2 et Z_3 à trois dates consécutives. L’information relative au phénomène illustré dans la Figure 2.2, peut être représentée par une *base de données spatio-temporelles* comme celle décrite dans le Tableau 2.9. À partir d’une telle base, nous extrayons des motifs spatio-temporels qui prennent en compte le temps et l’espace. Ces motifs permettent, par exemple, de déterminer au cours du temps les occurrences des évènements et leurs localisations.

Un exemple de motif spatio-séquentiel est $\langle (- \bullet \textit{pluie forte})$
 $(- \bullet [\textit{pluie forte ET points d'eau beaucoup ET moustique moyen}]$
 $([\textit{moustique beaucoup ET dengue moyen}] \bullet \textit{moustique beaucoup}) \rangle$
 qui signifie qu’une forte pluie dans une zone proche suivie de nouveau d’une forte pluie, de la création de points d’eau et d’une augmentation du nombre de moustiques dans une zone proche sont suivis d’une augmentation du nombre de moustiques et de cas de dengue dans la zone d’étude. Le \bullet symbolise la relation spatiale de voisinage et le $-$ l’absence d’évènement.

Comme pour les motifs séquentiels contextuels, la prise en compte de la dimension spatiale génère un espace de recherche très grand qui nécessite la proposition de nouveaux algorithmes pour assurer l’efficacité de la méthode en présence de gros volumes de données [Alatrística Salas et al., 2012b]. Nous avons testé ces algorithmes sur des jeux de données synthétiques et réels. En particulier, nous avons

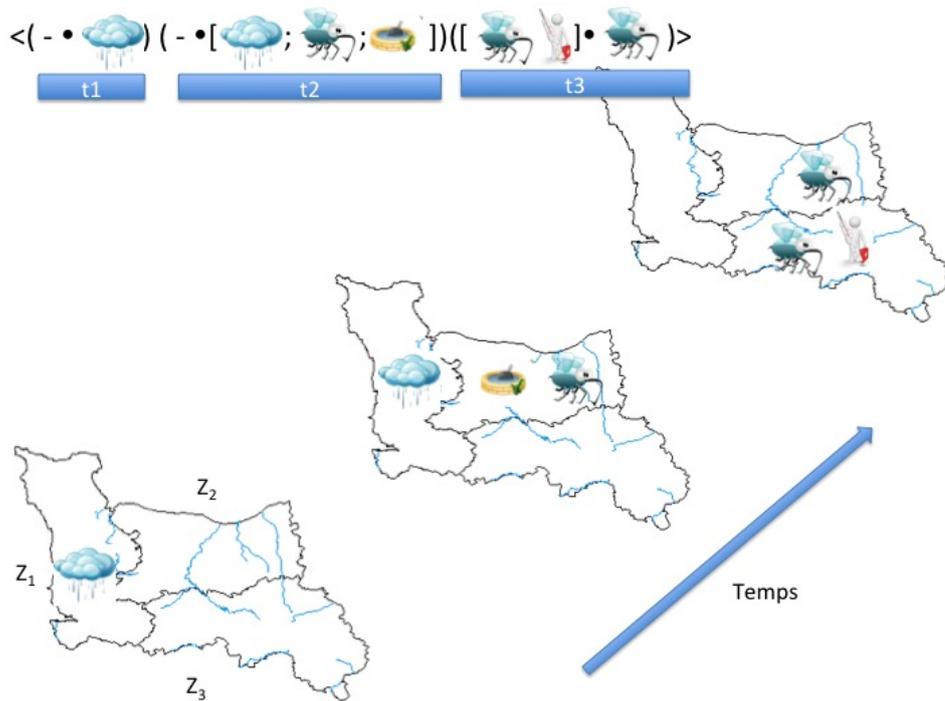


FIGURE 2.2 : Exemple d'un phénomène spatio-temporel : apparition des épidémies de dengue. L'espace est découpé en trois zones dans lesquelles on observe les évènements pluie, points d'eau, moustiques et cas de dengue pour 3 estampilles temporelles. Dans le coin supérieur droit, on présente un exemple de motif spatio-séquentiel extrait pour la zone Z_3 . Le – représente l'absence d'évènement et le • la relation spatiale.

TABLE 2.9 : Évolution de l'environnement pour les zones Z_1 , Z_2 et Z_3 pour 3 dates. Le – représente l'absence d'évènements.

Zone	Date	Pluie	Points d'eau	Moustiques	Cas de dengue
Z_1	t_1	<i>forte</i>	–	–	–
Z_1	t_2	<i>forte</i>	–	–	–
Z_1	t_3	–	–	–	–
Z_2	t_1	–	–	–	–
Z_2	t_2	–	<i>beaucoup</i>	<i>moyen</i>	–
Z_2	t_3	–	–	<i>beaucoup</i>	–
Z_3	t_1	–	–	–	–
Z_3	t_2	–	–	–	–
Z_3	t_3	–	–	<i>beaucoup</i>	<i>moyen</i>

travaillé sur les épidémies de dengue en Nouvelle Calédonie et en Guyane. Une épidémie désigne l'augmentation rapide de l'incidence d'une maladie contagieuse ou non¹³. Il n'existe ni vaccin ni médicament qui protègent contre la dengue qui est une maladie virale pouvant entraîner une fièvre hémorragique parfois mortelle. Mieux comprendre la propagation d'une telle épidémie est donc un enjeu crucial pour la haute autorité de santé afin de déclencher des alertes et mener des campagnes de prévention [Flamand et al., 2014, Alatrística Salas et al., 2012c].

13. Définition proposée pour l'Organisation Mondiale de la Santé (OMS).

Sélection de références

H. Alatrística-Salas, J. Azé, S. Bringay, F. Cernesson, N. Selmaoui-Folcher, M. Teisseire, A Knowledge Discovery Process for Spatiotemporal Data : Application to River Water Quality Monitoring. Ecological Informatics, Elsevier Ed., 2014 (**Impact Factor : 1,980**)

H. Alatrística-Salas, S. Bringay, F. Flouvat, N. Selmaoui-Folcher and M. Teisseire. The Pattern Next Door : Towards Spatio-sequential Pattern Discovery. Proceedings PAKDD 2012, Kuala Lumpur (Malaysia) (2) : 157-168 (**Rang A**)

C. Flamand, M. Fabrègue, S. Bringay, V. Ardillon, P. Quénel, JC. Desenclos, M. Teisseire. Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana. Journal of American Medical Informatics Association. 2014 Oct ;21(e2), 2014 : 232-240 (**Impact Factor : 3,932**)

1.4 Motifs partiellement ordonnés clos

TABLE 2.10 : Résumé des projets et des encadrements sur la thématique des motifs partiellement ordonnés

Master/Thèse	Projets & Collaborations
M. Fabrègue (2011-2014) – Thèse co-encadrement A. Braud, F. Le Ber et M. Teisseire	ANR FRESQUEAU (2011-2014)

Une limite très souvent formulée par les experts à propos des résultats des méthodes d'extraction de motifs est qu'ils sont difficiles à interpréter car les motifs générés sont parfois plus nombreux que les données initiales et très redondants dans la lecture. Nous nous sommes intéressés tout d'abord aux motifs clos pour réduire la redondance dans les motifs [Pasquier et al., 1999]. L'intuition est la suivante : un motif est clos s'il n'existe pas de motif plus grand dans lequel il serait inclus et qui aurait au moins le même support. Nous avons ensuite intégré la notion de motifs partiellement ordonnés clos, toujours dans l'objectif de réduire le nombre d'informations présentées aux experts (Thèse de M. Fabrègue). Contrairement aux motifs séquentiels qui représentent un ordre total entre les éléments d'un ensemble de séquences, les motifs partiellement ordonnés capturent une information partielle sur l'ordre entre des éléments. Le motif partiellement ordonné clos présenté dans la colonne de droite du Tableau 2.11 résume les deux motifs séquentiels de la colonne de gauche. Ce motif partiellement ordonné clos est supporté par au moins autant de séquences que les deux motifs séquentiels.

TABLE 2.11 : Des motifs séquentiels clos aux motifs partiellement ordonnés clos.

Motifs séquentiels clos	Motif partiellement ordonné clos
$\langle\langle(G1)(G2G3)(G4)\rangle\rangle$	
$\langle\langle(G1)(G5)(G4)\rangle\rangle$	

Le motif partiellement ordonné clos s'interprète ainsi : $G1$ a une expression moins forte que $G5$ et $G1$ a une expression moins forte que $G2$ et $G3$ qui ont une expression similaire. On ne peut ordonner $G5$ par rapport à $G2$ et $G3$. Tous ces gènes ont une expression moins forte que $G4$.

Ces motifs ont l'avantage de résumer des ensembles de motifs séquentiels, ce qui permet de condenser l'information, tout en restant interprétable et visualisable pour les humains grâce à la représentation sous forme de graphe. Cependant, l'extraction de tels motifs reste peu étudiée dans les bases de données de séquences. En effet, les méthodes existantes passent difficilement à l'échelle [Casas-Garriga, 2005, Pei et al., 2006]. De plus, elles n'extraient pas à la fois l'ensemble complet des motifs partiellement ordonnés clos et ne sont pas applicables sur n'importe quel type de bases de données de séquences. Dans le cadre de l'ANR FRESQUEAU¹⁴, nous avons développé un algorithme efficace pour l'extraction de motifs partiellement ordonnés clos en collaboration avec l'ENGEES. Comme pour les motifs séquentiels contextuels et les motifs spatio-temporels, l'extraction de tels motifs est plus complexe que l'extraction de motifs séquentiels classiques et nécessite la définition de nouveaux algorithmes pour prendre en compte le très grand espace de recherche généré. Du fait de l'explosion combinatoire qu'implique leur extraction, l'utilisation de certaines propriétés sur ces motifs, pour en réduire le nombre, est nécessaire. Nous nous sommes basés sur le paradigme *Pattern-Growth* et la projection de la base de données avec les préfixes fréquents ainsi que sur une optimisation qui a fait ses preuves en fouille de motifs séquentiels clos [Yan et al., 2003].

Sélection de références

- M. Fabrègue, A. Braud, S. Bringay, D. Le Ber, M. Teisseire : Mining closed partially ordered patterns, a new optimized algorithm. *Knowl.-Based Syst.* 79 : 68-79 (2015). (**Impact Factor : 3,058**)
- M. Fabrègue, A. Braud, S. Bringay, F. Le Ber, M. Teisseire : OrderSpan : Mining Closed Partially Ordered Patterns. *Proceedings IDA 2013, Londres (England)* : 186-197 (**Rang A**)
- M. Fabrègue, A. Braud, S. Bringay, C. Grac, F. Le Ber, D. Levet, M. Teisseire : Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. *Ecological Informatics* 24 (2014) : 210-221 (**Impact Factor : 1,980**)

14. <http://gls{ENGEES}-fresqueau.unistra.fr/>. Les données utilisées ont été des mesures associées à des prélèvements réalisés pour évaluer la qualité de l'eau des rivières.

2 Choisir, représenter et interpréter les meilleurs motifs

Dans les quatre sections précédentes, nous avons décrit de nouveaux types de motifs, donné l'intuition de leur formalisation et des méthodes d'extraction. Un point commun à toutes ces approches est que le nombre de motifs générés est très important même dans le cas des motifs partiellement ordonnés clos. Pour certaines applications, nous avons dû proposer de nouvelles mesures d'intérêt pour la sélection des "meilleurs" motifs au sens des experts avec qui nous avons été amené à collaborer. Par ailleurs, même après le filtrage de ces motifs, la lecture de listes plates de motifs ne permet pas toujours à l'expert de s'approprier les résultats de ces analyses descriptives. Nous travaillons donc sur des visualisations qui font émerger des régularités existant entre les motifs extraits. L'interaction que l'on peut ajouter à ces visualisations est également très importante pour la navigation de l'expert dans ces motifs et la découverte interactive de connaissances. Nous listons dans la suite les différentes mesures d'intérêt proposées et leur combinaison dans la Section 2.1 et décrivons des exemples de visualisations dans la Section 2.2.

2.1 Mesures d'intérêt et leur combinaison

TABLE 2.12 : Résumé des projets et des encadrements sur la thématique des mesures d'intérêt.

Master/Thèse	Projets & Collaborations
P. Salle (2007-2010) – Thèse co-encadrement M. Teisseire	MMDN - ANR Pradnet (2007-2011)
M. Fabrègue (2011-2014)	ANR FRESQUEAU (2011-2014)
Thèse co-encadrement A. Braud, F. Le Ber et M. Teisseire	
H. Alatrística Salas (2010-2013)	Univ. Nouvelle Calédonie
Thèse co-encadrement F. Flouvat, N. Selmaoui et M. Teisseire	INVS, DASS, Institut Pasteur

Nous nous demandons dans cette section *comment identifier les motifs les plus intéressants parmi les milliers, millions de motifs extraits ?* Cette notion d'intérêt varie selon le contexte applicatif. Une mesure d'intérêt permet de quantifier l'attractivité relative d'un motif pour un expert. Il ne faut pas confondre une mesure d'intérêt avec une mesure de similarité (ou de distance) qui permet d'estimer la différence entre deux motifs. De nombreuses études ont été réalisées en vue de comparer les mesures d'intérêt [Lenca et al., 2003, Blanchard, 2005].

Dans le cadre de l'extraction de motifs séquentiels dans les puces à ADN (voir Section 1.1), nous nous sommes intéressés à l'extraction des motifs correspondant à des connaissances déjà validées et à des connaissances plus inattendues. Pour cela, nous avons interrogé la base PUBMED (voir Section E) pour récupérer les articles traitant des gènes impliqués dans les motifs extraits. En appliquant différents ratios, nous avons distingué les *motifs populaires* composés de gènes déjà associés dans la littérature (connaissances validées) et les *motifs innovants* contenant des gènes non reconnus dans la littérature comme impliqués dans la pathologie étudiée (connaissances inattendues). Ces nouveaux gènes sont alors donnés à étudier aux biologistes [Bringay et al., 2010].

Dans le cadre de la description de données spatio-temporelles (Thèse de H. Alatrística Salas) (voir Section 1.3), nous avons choisi de nous focaliser sur la recherche de séquences fréquentes qui sont peu contredites par les données. Pour cela, nous avons étendu la mesure de la moindre contradiction définie par [Azé, 2003] pour les règles d'association, au contexte des séquences d'itemsets en prenant en compte l'ordre temporel d'apparition des événements puis aux motifs spatio-séquentiels en prenant en compte la dimension spatiale de ces motifs. Cette extension conserve l'esprit initial de la mesure qui vise à évaluer le nombre de fois où une règle est vérifiée vs. le nombre de fois où elle est invalidée dans les données. Une règle qui est plus fréquemment vérifiée qu'invalidée est *a priori* intéressante. Nous avons choisi

cette mesure car elle est simple à mettre en œuvre et à comprendre. Elle est donc relativement facile à appréhender par les experts.

Comme nous venons de le voir, il existe de nombreuses méthodes pour filtrer les motifs selon un critère d'intérêt spécifique aux besoins experts. Or, l'utilisateur peut vouloir les filtrer selon une combinaison de plusieurs critères sans en privilégier un en particulier. Nous avons proposé (Thèse de M. Fabrègue) pour cela une méthode itérative, qui filtre k motifs selon plusieurs dimensions d'intérêt. Nous l'avons appliquée pour filtrer des k motifs partiellement ordonnés clos (voir Section 1.4) mais elle peut être utilisée pour d'autres motifs. Les trois critères considérés sont : la fréquence (pourcentage de séquences dans une classe de données supportant un motif), la discriminance (rapport de la fréquence d'un motif dans une classe par rapport à une autre classe) et la non redondance (plusieurs motifs sont supportés par un nombre élevé de séquences communes). D'autres critères peuvent être utilisés. Des expérimentations sur des jeux de données réelles ont montré que la sélection basée sur les critères pris un par un ne permet pas de repérer les mêmes séquences. Avec notre méthode, la sélection obtenue identifie des séquences correspondant à un compromis entre les trois critères et a été jugée comme représentative du contenu de la base de données au sens des experts.

Sélection de références

P. Salle, S. Bringay, M. Teisseire, F. Chakkour, M. Roche, G. Devau, C. Lautier and JM. Verdier. GeneMining : Identification, Visualization, and Interpretation of Brain Ageing Signatures. Proceedings MIE'2009, Sarajevo (Bosnie-Herzégovine), 2009. Studies in Health Technology and Informatics : 767-771

H. Alatrística Salas, J. Azé, S. Bringay, F. Cernesson, F. Flouvat, N. Semaloui et M. Teisseire. Finding Relevant Sequences With The Least Temporal Contradiction Measure : Application to Hydrological Data. Proceedings AGILE 2012, Avignon (France) : 197-202

M. Fabrègue, A. Braud, S. Bringay, C. Grac, F. Le Ber, D. Levet, M. Teisseire : Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. Ecological Informatics 24 (2014) : 210-221 (**Impact Factor : 1,980**)

2.2 Visualisation de motifs

TABLE 2.13 : Résumé des projets et des encadrements sur la thématique de la visualisation de motifs.

Master/Thèse	Projets & Collaborations
A. Zine El Abidine - Master co-encadrement P. Poncelet et A. Sallaberry	IGMM (2010-2011)
H. Alatrística Salas (2010-2013)	PIKKO
Thèse co-encadrement F. Flouvat, N. Selmaoui et M. Teisseire	Univ. Nouvelle Calédonie
M. Fabrègue (2011-2014)	ANR FRESQUEAU (2011-2014)
Thèse co-encadrement A. Braud, F. Le Ber et M. Teisseire	

Nous nous demandons dans cette section *comment présenter les motifs les plus intéressants récupérés suite au filtrage pour faciliter l'interprétation des experts* ? De nouveau, la pertinence des visualisations est fortement liée au contexte applicatif.

Les techniques de visualisation ont été largement discutées dans la littérature [Bertin, 1983, Tufte, 1983, Peuquet, 1994, Ward et al., 2010]. Ces auteurs insistent, entre autre, sur l'importance de la représentation visuelle des informations de façon à rendre plus facile l'interprétation des résultats obtenus. Une représentation visuellement attrayante, montrant l'information à interpréter, est plus intéressante et souvent plus efficace en terme d'interprétation par les experts qu'un affichage immédiat de

quelques chiffres ou une représentation purement textuelle [Tufte, 1983]). Cela est d'autant plus vrai dans le cas des motifs qui, présentés sous la forme de listes plates, sont peu attractifs pour les experts. [Ware, 2004], spécialiste des études sur la perception, donne également des indications sur les spécifications techniques à prendre en compte lors de la représentation visuelle des données (e.g. sélection des couleurs, choix des formes, des polices) qui ont été suivies lors de la réalisation des visualisations que nous avons proposées.

Dans la littérature, on trouve peu d'approches pour la visualisation de motifs et ces approches sont quasiment toutes associées aux nouvelles techniques d'extraction générant les motifs. En effet, il est très difficile d'aborder ces problèmes de visualisation sans aborder la méthode utilisée pour générer les données à visualiser. Les méthodes d'analyse visuelle (VA Visual Analytics) sont alors très pertinentes [Keim et al., 2008] car elles se combinent étroitement avec les méthodes d'extraction de connaissances. Plusieurs approches ont été proposées pour les motifs séquentiels. Par exemple, [Wong et al., 2000] ont appliqué une technique d'extraction de motifs séquentiels aux données textuelles et l'ont accompagnée d'un prototype de visualisation pour l'analyse des motifs obtenus sur des grands corpus. [Subasic and Berendt, 2008] ont proposé une méthode et un outil de visualisation pour cartographier et interagir avec les publications scientifiques postées sur le Web en utilisant des méthodes de fouille de textes.

Dans la suite de cette section, nous décrivons trois systèmes de visualisation de motifs développés dans notre équipe et correspondant aux nouveaux motifs décrits dans les sections précédentes. Les couleurs étant souvent importantes pour la sémantique de ces motifs, nous avons ajouté une page web accessible à cette url (<http://www.univ-montp3.fr/miap/sbringay/hdr/images.html>) qui permet de visualiser les images en couleur (voir figures 2.3,2.4,2.5,2.6).

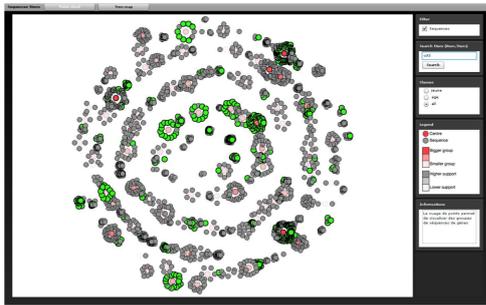
Système de visualisation des motifs de gènes

Dans le cadre d'une collaboration avec la société PIKKO et afin d'aider les biologistes dans l'analyse des motifs séquentiels extraits (voir Section 1.1), nous avons conçu un système de visualisation pour naviguer dans ces motifs et mettant en lien les articles de PUBMED qui traitent des différents gènes impliqués dans ces motifs [Sallaberry et al., 2011]. Trois vues ont été définies :

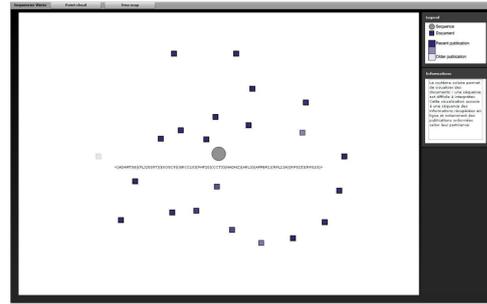
1. La figure 2.3 (a) représente des clusters de motifs. Les points gris sont des motifs. Les points rouges sont les centres des clusters. Plus un cluster contient de motifs, plus le rouge est intense. Il est possible de sélectionner des motifs (ici en vert) en indiquant le nom d'un gène. Le biologiste peut interagir en zoomant et dézoomant. Lorsqu'il sélectionne un motif, des informations contextuelles apparaissent. Cette vue ne prend pas en compte la hiérarchie intrinsèque qui existe entre les motifs (voir figure 2.3 (c)) ;
2. Nous avons donc utilisé une représentation appelée Treemap¹⁵ (voir figure 2.3 (d)). Chaque rectangle correspond à un nœud de la hiérarchie. La répartition des motifs dans les classes étudiées est représentée par la proportion des couleurs roses et vertes au niveau de chaque nœud. Le biologiste peut interagir en naviguant dans la hiérarchie par sélection (clic) d'un nœud ;
3. La dernière vue (voir figure 2.3 (b)) représente les documents (les carrés bleus) utiles à l'analyse d'un motif (le point gris placé au centre). Plus un document est proche, plus il contient d'informations sur les gènes impliqués dans le motif. Plus un article est de couleur foncé, plus il est récent.

Toujours sur la thématique des puces à ADN, nous avons travaillé dans le cadre d'un PEPS avec l'IGMM sur des données intégrant des aspects temporels. Il s'agit de puces à ADN réalisées sur des

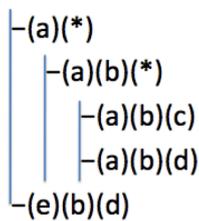
15. <http://www.cs.umd.edu/hcil/treemap-history/>



(a) Exemple de clusters de motifs



(b) Visualisation des meilleurs documents associés à un motif



(c) Hiérarchie de motifs



(d) Visualisation hiérarchique de motifs

FIGURE 2.3 : Trois vues de motifs de gènes : (a) cluster ; (b) document et (c) treemap.

cellules infectées par différentes souches de **HIV** à plusieurs estampilles temporelles. Nous avons implémenté un prototype [Abidine et al., 2013] (Master de Amal Zine El Abidine) décrit dans la figure 2.4. Chaque gène est associé à une ou plusieurs fonctions. Dans la visualisation, chaque ligne correspond à une fonction et chaque colonne à une estampille temporelle. Plus une fonction est représentée dans un ensemble de motifs, plus la ligne correspondante est épaisse. Cette vue permet au biologiste de visualiser à quel moment et jusqu'à quand une fonction est activée après infection par le virus.

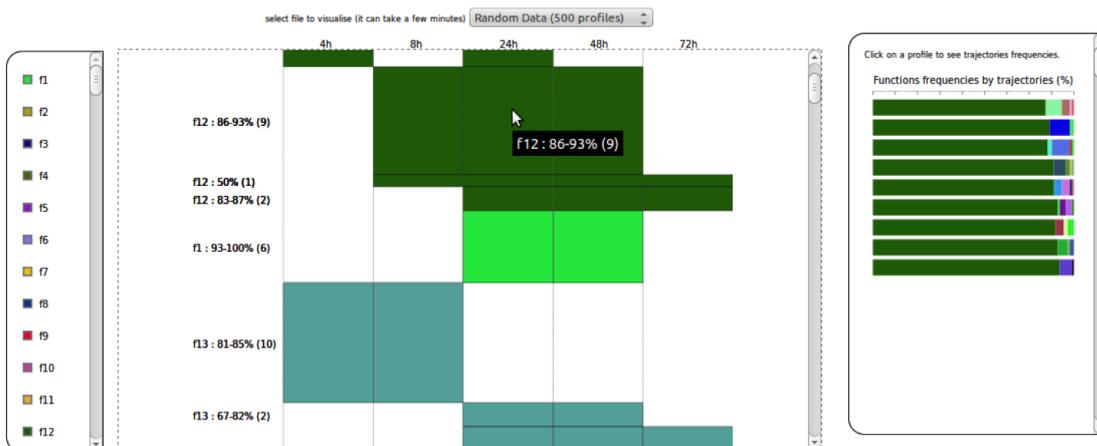


FIGURE 2.4 : Fonctions impliquées dans les motifs et leur évolution dans le temps.

Système de visualisation des motifs spatio-temporels

Nous avons proposé un système de visualisation aux experts pour mieux appréhender les interactions spatiales et temporelles entre les différents facteurs représentés via les motifs spatio-temporels (voir Section 1.3). À la différence des approches classiques, nous soulignons les dynamiques spatio-temporelles, tout en prenant en compte l'environnement proche. Nous ne nous limitons pas à la visualisation de motifs spatio-temporels mais nous proposons tout un environnement pour une analyse détaillée de ces motifs à différentes échelles (des motifs globaux aux objets spatiaux locaux). Plus précisément, notre environnement de visualisation offre les fonctionnalités suivantes :

1. une vue synthétique et schématique des motifs sous forme de graphes colorés (avec possibilité d'associer des icônes aux nœuds). Trois manières d'afficher un motif ont été proposées en fonction des besoins des experts (voir figure 2.5). Chaque cercle correspond à un évènement. Les cercles sont entourés de rouge quand ils correspondent à la dimension étudiée (cas de dengue dans cet exemple). Deux évènements sont reliés par une relation temporelle avec des arcs pleins et par une relation spatiale avec des arcs en pointillé ;
2. une vue détaillée des zones et des dates où sont apparus les événements (i.e. des occurrences des motifs). À partir d'un motif, il est possible d'identifier les zones impactées sur une carte (et inversement). Avec la frise chronologique, l'expert visualise les dates des événements représentés par les motifs (avec deux niveaux de détails) (voir figure 2.6 (a,b,d)) ;
3. des statistiques détaillées sur les zones (e.g. nombre d'habitants) et les caractéristiques temporelles des motifs (e.g. durée moyenne) (voir figure 2.6 (c)).

Notre système de visualisation peut être utilisé pour d'autres types de motifs, où les dimensions spatiales et/ou temporelles sont présentes, telles que les co-localisations [Shekhar and Huang, 2001] et les séquences temporelles [Tsoukatos and Gunopulos, 2001].

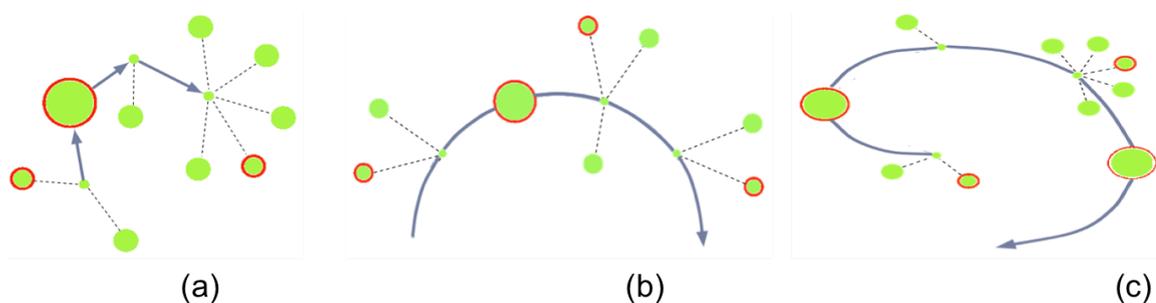


FIGURE 2.5 : Trois vues d'un motif spatio-temporel : (a) on optimise l'espace avec un algorithme de force. Le temps est représenté par un arc (b) ou une spirale (c).

Système de visualisation des motifs spatio-temporels

Nous présentons les trois vues développées dans le cadre du projet ANR FRESQUEAU [Accorsi et al., 2014] pour visualiser les motifs partiellement ordonnés clos et décrites dans la figure 2.7 :

1. la vue géographique (2.7 (b)) avec navigation sur une carte dans le cadre supérieur droit ;
2. la vue clustering (2.7 (a)) avec affichage des stations de prélèvements en fonction de leur comportement dans le cadre supérieur gauche. Les clusters sont représentés par des cercles positionnés au barycentre des stations qu'ils contiennent ;
3. la vue des motifs temporels (2.7 (c)) pour analyser les motifs sélectionnés à partir des deux vues précédentes dans la partie inférieure de l'écran. Cette vue est divisée en deux panneaux. Un premier

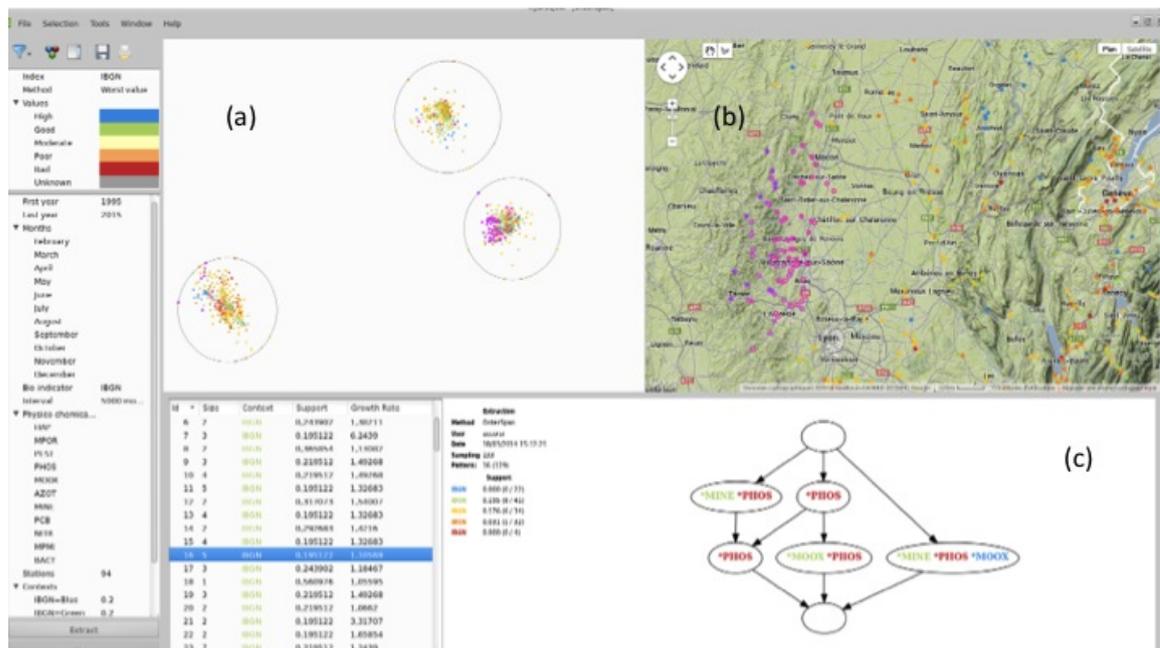


FIGURE 2.7 : Système de visualisation de motifs partiels ordonnés clos : application à l'évaluation de la qualité de l'eau des rivières : (a) vue cluster, (b) vue géographique et (c) vue d'un motif sélectionné.

ment orientée application, guidée par les besoins pratiques des experts, qui a connu un grand succès dans des domaines comme l'analyse financière ou les changements environnementaux [Keim et al., 2008]. Parmi les challenges identifiés par ces auteurs, dans le contexte de l'analyse de motifs en santé, on retrouve les problématiques liées au niveau de détail à présenter aux utilisateurs, au passage à l'échelle en présence de gros volumes de données, au besoin de définir de nouvelles métaphores comme nous avons été amenés à le faire pour visualiser les motifs spatio-séquentiels et les motifs partiellement ordonnés clos. Pour finir, la mise en place de protocoles d'évaluation rigoureux pour vérifier que les outils implémentés sont utiles et utilisables par les experts est essentielle.

Sélection de références

- P. Accorsi, N. Lalande, M. Fabrègue, A. Braud, P. Poncelet, A. Sallaberry, S. Bringay, M. Teisseire, F. Cernesson, F. Le Ber. HydroQual : Visual analysis of river water quality. IEEE VAST, 2014 : 123-132
- A. Sallaberry, N. Pecheur, S. Bringay, M. Roche, M. Teisseire. Sequential patterns mining and gene sequence visualization to discover novelty from microarray data. Journal of Biomedical Informatics 44(5), 2011 : 760-774 (**Impact Factor : 2,817**)
- A. Zine El Abidine, A. Sallaberry, S. Bringay, M. Fabrègue, C. Lecellier, N. H. Phan, P. Poncelet. Co2Vis : A visual analytics tool for mining co-expressed and co-regulated genes implied in HIV infections. Poster BioVis 2013, Atlanta (USA), 2013
- H. Alatrasta Salas, S. Bringay, F. Flouvat, N. Selmaoui-Folcher, M. Teisseire. S2PViewer : un prototype de visualisation de motifs spatio-temporels. In proceedings of SAGEO 2013, Brest, France, September 2013
- M. Fabrègue, S. Bringay, P. Poncelet, M. Teisseire and B. Orsetti. Mining microarray data to predict the histological grade of a Breast Cancer. Journal of Biomedical Informatics, 2011 Dec ;44 Suppl 1 :S12-6 (**Impact Factor : 2,817**)

C Vers des approches prédictives et prescriptives basées sur les motifs

TABLE 2.14 : Résumé des projets et des encadrements sur la thématique des méthodes prédictives et prescriptives.

Master/Thèse	Projets & Collaborations
M. Fabrègue (2010-2011) – Master co-encadrement P. Poncelet	INSERM Val d'Aurelle (2010)
J. Rabatel (2008-2011) – Thèse co-encadrement P. Poncelet	Tecnalia (Espagne) (2008-2011)

Dans les approches descriptives, le problème de la découverte de motifs a d'abord été introduit pour représenter le contenu d'une base de données dans le but de fournir des connaissances compréhensibles et interprétables par des experts pour les assister dans leur prise de décision. Plus récemment, leur champ d'application s'est élargi à d'autres types d'analyse. Plus particulièrement, des travaux ont cherché à exploiter les motifs extraits pour des applications prédictives et prescriptives et en particulier pour des problèmes de classification. Les approches basées sur les motifs ont en effet l'avantage d'être interprétables et compréhensibles par les décideurs, réduisant ainsi l'effet de « *boîte noire* » fréquemment rencontré dans les approches de classification. Par exemple, les professionnels de la santé à qui l'on propose un résultat de classification pour les aider dans une tâche de diagnostic, ont envie de comprendre le raisonnement ayant conduit à ce résultat. Dans la suite, nous présentons deux exemples d'applications prédictives et une application prescriptive. Nous donnons uniquement l'intuition de ces méthodes et renvoyons le lecteur aux articles correspondants pour plus de détails.

1 Prédire le grade d'un cancer

Nous avons travaillé en partenariat avec l'INSERM Val d'Aurelle (Master de Mickael Fabrègue), pour caractériser efficacement les différents grades de cancers du sein à partir de données issues de puces à ADN.

Pour chaque classe de données (grades de cancer), nous avons extrait des motifs séquentiels. La quantité de motifs pour chaque classe n'est pas la même. Nous avons donc pour chaque classe sélectionné k motifs en nous basant sur les écarts de supports des motifs dans chaque classe. Nous recherchons ceux ayant un support très élevé dans une classe et peu élevé dans les autres classes. Ces motifs sont alors utilisés comme descripteurs des classes. Lorsque l'on souhaite classer une nouvelle séquence de données non labellisée issue d'une puce, on compare cette séquence en terme d'inclusion aux motifs des différentes classes. À chaque fois qu'un motif est inclus dans les motifs choisis comme représentant d'une classe, on augmente le score de la classe pour ce motif. Un exemple de calcul est donné dans la table à tester 2.15.

TABLE 2.15 : Exemple de motifs associés aux deux classes C_1 et C_2 . Si l'on considère la séquence $S = \langle (a\ b)(e\ f\ g) \rangle$. Les 2 motifs de la classe C_1 sont inclus dans S contre 1 motif de la classe C_2 . S est attribuée à la classe C_1 .

C_1	C_2
$(a\ b)(e)$	$(c)(a)$
$(e\ f)$	$(e\ f)$

Cette méthode a été expérimentée sur différents jeux de la littérature disponibles en ligne depuis Gene Expression Omnibus¹⁶, en nous focalisant sur une liste de gènes connus pour leur implication

16. <http://www.ncbi.nlm.nih.gov/geo/> (mnibus2 KJX64-KJ125 (GSE2990), TAM (GSE6532), and TBG2

dans le cancer du sein [Sotiriou et al., 2003]. Nous avons considéré 3 grades [Elston and Ellis, 1991, Scarff and Torloni, 1968]. Nous avons comparé notre approche à tous les classificateurs implémentés dans le logiciel weka¹⁷. Les résultats obtenus améliorent significativement les résultats des travaux existants (F-mesure d'environ 0.96) [Fabrègue et al., 2011].

2 Prédire la catégorie d'un patient et le prochain évènement

Les travaux précédents ne tirent pas parti d'informations contextuelles lorsque celles-ci sont disponibles. Intégrer ces informations dans le processus de classification peut pourtant considérablement améliorer les résultats. Par exemple, supposons que nous cherchons à classer un patient dans une catégorie d'âge, i.e., à lui associer le label *jeune* ou *âgé* en fonction des actes médicaux reçus. La classification précédente basée sur les motifs s'appuiera uniquement sur les motifs extraits pour différencier les patients *jeunes* et *âgés*. Faisons maintenant l'hypothèse que cette séquence d'activités soit enregistrée en *été*. Les motifs contextuels nous permettent alors de tirer parti de cette information afin de considérer les différences spécifiques des actes réalisés en été qui existent entre les patients *jeunes* et *âgés* (e.g. consignes différentes pour l'hydratation quotidienne).

L'intuition de la méthode est la suivante. Lorsque l'on cherche à prédire la classe d'une séquence de données, plutôt que de se comparer à toute la base de données, on va utiliser le treillis des contextes pour se comparer uniquement aux motifs associés au contexte connu de la séquence. Si le patient est un *homme* et que les données ont été récoltées *en été*, on comparera sa séquence d'actes aux motifs associés aux deux contextes : "*homme+été+âgé*" et "*homme+été+jeune*" pour déterminer de quel contexte on se rapproche le plus et ainsi prédire une étiquette pour l'âge du patient.

Cette méthode a été testée sur différents jeux de la littérature (e.g. des données textuelles de Amazon, des puces à ADN, des mesures de consommation énergétique) et les résultats sont plutôt encourageants (F-mesure d'environ 0.8). L'apport de cette méthode est surtout lié à l'interprétation que l'expert peut faire lorsqu'on lui montre les motifs et les éléments de contextes qui ont permis la proposition de la classe.

Une extension assez simple de cette méthode consiste à prédire le ou les prochains évènements possibles à partir d'une séquence d'évènements. On utilise pour cela les informations contextuelles pour repérer les motifs associés au même contexte que la séquence. On cherche les motifs les plus similaires à la séquence et on en déduit une liste d'évènements possibles non déjà inclus dans la séquence étudiée. Par exemple, les deux motifs $M_1 = \langle (a)(b)(c)(d)(e) \rangle$ et $M_2 = \langle (a)(b)(c)(d)(f)(g) \rangle$ ont été extraits sur une base de séquences pour le contexte C . Si une nouvelle séquence se présente $S = (a)(b)(c)(d)$ dans le contexte C , on peut prédire que les prochains évènements seront (e) ou $(f)(g)$ et calculer une confiance dans cette prédiction qui prendra en compte le support des motifs M_1 et M_2 .

3 Prescrire en détectant une anomalie et en déclenchant une alerte

La détection d'anomalies dans les données séquentielles, se rapportant à la découverte de fragments de séquences qui ne correspondent pas aux comportements attendus, constitue un défi porteur d'enjeux considérables dans de multiples domaines. Par exemple, pour la surveillance d'équipements médicaux une telle anomalie peut annoncer un dysfonctionnement grave. De même, une anomalie relevée dans le battement cardiaque d'un patient peut mettre en évidence une maladie.

(GSE7390)

17. <http://www.cs.waikato.ac.nz/ml/weka/>

La détection d'anomalies dans les données séquentielles est une tâche difficile qui doit bien souvent faire face au manque de connaissances et de données utiles pour caractériser les différentes anomalies possibles. Beaucoup de données décrivent les comportements normaux et peu de données décrivent les comportements anormaux. Un comportement anormal peut également ne jamais avoir été rencontré et n'est donc pas décrit dans la base utilisée pour l'apprentissage. De nouveau, les informations contextuelles disponibles sont extrêmement importantes puisqu'un même comportement peut être considéré comme normal ou anormal en fonction des circonstances dans lesquelles il se manifeste. Par exemple, une approche pertinente de détection d'anomalies devra prendre en compte le fait que l'hydratation d'un patient attendue sera différente en été et en hiver.

En partant de l'idée générale qu'un motif fréquent représente un comportement attendu, nous avons utilisé les motifs contextuels pour détecter des anomalies. L'intuition est la suivante. On compare cette fois la séquence de donnée "écoutée" aux motifs associés au même contexte. Si la séquence s'avère moins en accord (en terme d'inclusion) avec ces motifs qu'avec les motifs d'autres contextes alors une alarme est déclenchée.

Comme précédemment, cette méthode a été testée sur les mêmes jeux de la littérature (e.g. Amazon, des puces à ADN, consommation énergétique, etc.) et les résultats sont plutôt encourageants (précision et rappel d'environ 0.85). De même, l'apport de cette méthode est surtout lié à l'interprétation que l'expert peut en faire.

Sélection de références

M. Fabrègue, S. Bringay, P. Poncelet, M. Teisseire and B. Orsetti. Mining microarray data to predict the histological grade of a Breast Cancer. *Journal of Biomedical Informatics*, 2011 Dec ;44 Suppl 1 :S12-6

(Impact Factor : 2,817)

J. Rabatel, S. Bringay and P. Poncelet. Anomaly Detection in Monitoring Sensor Data for Preventive Maintenance. *Journal Expert Systems with Applications*, 38, 2010 : 7003-7015 **(Impact Factor :**

2,908)

D Conclusions et perspectives

Lors de nos collaborations avec divers experts, entre autres du domaine de la santé, nous avons eu besoin de créer de nouveaux motifs pour coller à la réalité des données et des besoins d'analyse rencontrés et prendre en compte une sémantique de plus en plus riche. Nous avons ainsi intégré dans les motifs séquentiels qui considéraient déjà le temps, une composante spatiale pour les *motifs spatio-séquentiels* et une composante contextuelle pour les *motifs contextuels* (voir figure 2.8).

Afin de réduire la quantité de motifs présentés aux experts, nous avons travaillé sur la notion de *motifs partiellement ordonnés clos*, ainsi que sur la notion de *filtrage*. Afin de donner sens et vie à ces motifs, nous avons proposé des systèmes de *visualisation* innovants permettant à l'expert de s'approprier les résultats de ces méthodes descriptives. Nous avons également utilisé les motifs pour des approches prédictives et prescriptives, moins courantes dans la littérature comme montré dans le Chapitre 1 d'état de l'art. L'intérêt principal des approches à bases de motifs est qu'elles sont facilement interprétables par les experts et évitent le côté boîte noire de certaines méthodes traditionnelles. Les perspectives associées à ces approches sont multiples et détaillées ci-après.

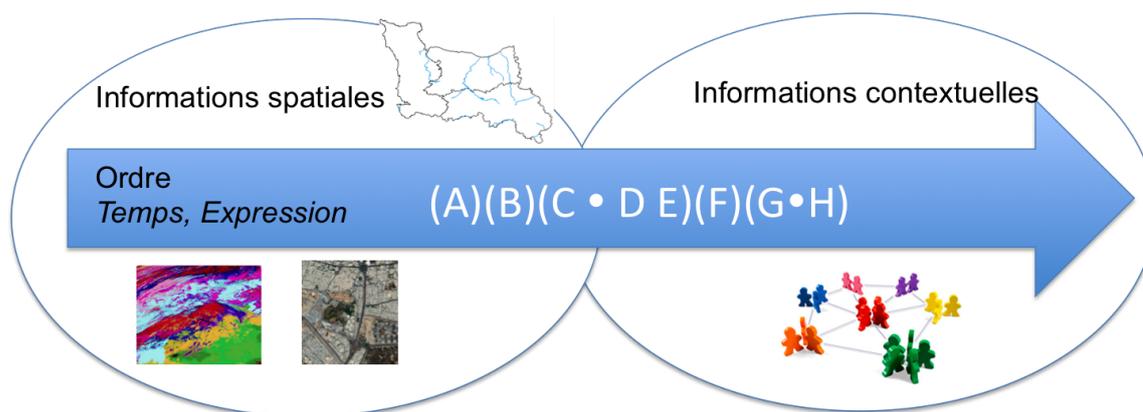


FIGURE 2.8 : Une information riche exploitée dans les motifs.

1 Des motifs de plus en plus riches sémantiquement

Tout d'abord, comme nous l'avons démontré dans ce chapitre, la sémantique des motifs extraits doit être considérée pour présenter aux experts des motifs réellement adaptés à leurs besoins applicatifs et en rapport avec tous les niveaux de description que l'on peut définir dans les données.

En particulier, nous n'avons pas exploité les structures qui existent naturellement dans les données "supplémentaires" que nous avons considérées comme des hiérarchies (e.g. hiérarchie sur les zones géographiques, les relations spatiales ou temporelles). Dans [Fabrègue et al., 2012], nous avons commencé à généraliser ce problème à des relations spatiales complexes entre objets géographiques mais l'étude n'a pas été exhaustive. De plus, les propriétés liées à l'inclusion de diverses granularités spatiales et temporelles doivent être étudiées car elles permettront très probablement d'améliorer le processus d'extraction.

Nous avons travaillé sur les données séquentielles et recherché des motifs en rapport avec cette séquentialité. Nous n'avons pas considéré encore des structures plus complexes que les séquences comme les graphes très étudiés actuellement (e.g. graphes dans les réseaux sociaux). Ces approches peuvent s'avérer réellement performantes comme le montrent les travaux prometteurs de [Pasquier et al., 2013, Sanhes et al., 2013] dans les bases de données spatiales.

2 Sélection et visualisation des motifs

Par ailleurs, si nous avons déjà travaillé sur les mesures d'intérêts qui permettent aux experts de sélectionner les motifs les plus pertinents, nous n'avons pas toujours été capables d'insérer la sélection des motifs dans le processus d'extraction ni d'utiliser des connaissances du domaine (e.g. méta-données, descriptions sémantiques, ontologies) pour améliorer à la fois le passage à l'échelle en réduisant l'espace de recherche mais également pour améliorer la qualité des motifs obtenus et leur interprétation en ajoutant des critères de sélection experts.

L'homme peut facilement interpréter des informations complexes, reconnaître rapidement des motifs via des jeux de placements, de couleurs, de formes, de textures, etc. Concevoir des visualisations facilitant ces processus cognitifs et lui permettant d'appréhender au mieux les notions abstraites portées par les motifs et qui représentent le monde qui l'entoure, est essentiel dans l'objectif de proposer des méthodes d'analyse réellement utilisables. Une généralisation des outils proposés est nécessaire pour s'adapter à tous les types de motifs.

3 De nouvelles applications aux motifs, utiles aux professionnels de la santé

Dans ces travaux, les applications des méthodes développées ont été guidées par les différentes collaborations locales et nationales que nous avons pu avoir avec des partenaires académiques et industriels. Il est possible d'envisager de nombreux autres domaines d'application pour lesquels la recherche des motifs serait pertinente. Beaucoup restent à explorer comme par exemple la fouille d'images (e.g. radiographie, scanner) [Kharat et al., 2014, Kambeitz et al., 2015] ou la fouille de données de capteurs [Pitarch et al., 2010] où beaucoup de données existent et beaucoup de méthodes efficaces et passant à l'échelle sont actuellement développées.

Finalement, il existe un réel besoin de collaboration entre experts de la fouille de données et experts du domaine de la santé, collaboration qui reste la clef du succès du processus d'extraction de connaissances.

Dans ce chapitre, nous avons présenté des méthodes de formalisation de la connaissance présente dans les bases médicales produites et reconnues par les professionnels de la santé. Dans le chapitre suivant, nous allons nous intéresser aux données produites par le patient dans les médias sociaux et nous montrerons que ces données peuvent apporter une réelle plus-value médicale.

ANALYSE DE LA SANTÉ VIA LES MÉDIAS SOCIAUX

Sommaire

- A Introduction**
 - B Motivations et challenges liés à l'analyse des productions des e-patients dans les médias sociaux**
 - C Enrichissement de messages**
 - 1 De quoi ?
 - 2 Comment ?
 - 3 Qui ?
 - 4 Quand ?
 - D Conclusions et perspectives**
 - 1 Des pré-traitements complexes
 - 2 Vocabulaire patient
 - 3 Ressources annotées
 - 4 Et les motifs ?
 - 5 Visualisations
-

A Introduction

Si depuis 2013, j'ai continué à travailler sur la thématique des motifs, décrite dans le Chapitre 2, la majorité de mes activités de recherche se focalise désormais sur l'analyse des productions des patients dans les médias sociaux. En effet, les données issues de ces médias sociaux adressent les défis décrits dans le Chapitre 1 Section F à savoir les caractéristiques des big data (volume, variété et vélocité) ainsi que des aspects temporels et spatiaux, tout en offrant l'avantage de leur disponibilité. Cette thématique étant plus récemment abordée, les publications associées sont moins nombreuses mais plusieurs sont en cours et destinées à des journaux de haut niveau scientifique. Il ne s'agit en rien d'un virage thématique car, comme je l'expliquerai dans la Section 2.1) puis dans la conclusion Section D, la recherche de motifs est très pertinente lorsque l'on travaille sur les textes et je l'utilise actuellement dans plusieurs chaînes de traitements, notamment pour détecter les messages subjectifs et en particulier les traces de sentiments.

Les médias sociaux dédiés à la santé correspondent à une nébuleuse de sites et de fonctionnalités sociales permettant aux internautes de publier du contenu (e.g. messages, photos, vidéos), d'exprimer des sentiments à propos de ces contenus (e.g. j'aime, j'aime pas avec "Likes", émotions via les smileys), de dialoguer de manière synchrone et asynchrone entre utilisateurs, éventuellement au sein de groupes (e.g. files de discussions, chat, mails), d'agréger les informations, etc. Dans le cadre du projet PATIENT'S MIND¹⁸ que je porte depuis 2003 et de l'ANR Jeune chercheur SFIR¹⁹ à laquelle je participe depuis 2012, nous avons commencé par étudier les forums de santé qui sont de plus en plus consultés par les patients [Huh et al., 2013]. Ces forums sont des espaces d'échanges où les patients partagent leurs sentiments à propos de leurs maladies, traitements, etc. Sous couvert d'anonymat, ils expriment très librement leurs expériences personnelles. Ces forums sont donc une source d'informations très utile pour les patients à la recherche d'information se rapportant à leur santé et de soutien humain. Ces forums sont également très utiles pour les professionnels de santé afin de mieux identifier et comprendre les problèmes, les comportements et les sentiments de leurs patients. Récemment, nous avons élargi cette étude aux réseaux sociaux qui font partie des médias sociaux et ont pour vocation plus particulière la mise en relation. On distingue les réseaux de contact comme *Facebook* qui favorisent le réseautage sous forme de cercles d'amis et les réseaux de contenu comme *Twitter* qui favorisent le partage au sein de larges communautés. Dans ces réseaux sociaux, les patients s'expriment également beaucoup à propos de leur santé mais différemment par rapport aux forums de santé, avec plus d'immédiateté, d'investissement émotionnel, notamment quand les messages sont destinés aux amis proches, etc.

Dans la suite de ce chapitre, nous reprendrons le terme de *e-patients*, utilisé par [Akerkar and Bichile, 2004], pour désigner les patients que l'on retrouve sur les médias sociaux. Ces auteurs décrivent la révolution de l'e-patient (*e-patient revolution*) qui se présente désormais chez le médecin avec des idées préconçues, issues de ses interactions dans le web social et qu'il est important de prendre en compte dans la relation entre le médecin et le patient.

Dans la Section B, nous présentons les motivations liées à l'analyse des productions des e-patients et les challenges rencontrés. Dans la Section C, nous décrivons les méthodes élaborées pour analyser les contenus produits par ces e-patients en vue de les enrichir selon 4 dimensions d'analyse. Les travaux présentés dans cette section ont été expérimentés sur des forums de santé mais peuvent facilement être transposés aux réseaux sociaux en général.

18. <https://www.lirmm.fr/patient-mind/>

19. <http://www.agence-nationale-recherche.fr/?Projet=ANR-12-JS02-0010>

B Motivations et challenges liés à l'analyse des productions des e-patients dans les médias sociaux

De nos jours, les internautes postent, bloguent, twittent, etc. sur presque tous les sujets, y compris des sujets liés à leur santé. Les médias sociaux dédiés à la santé offrant ces fonctionnalités (e.g. PatientWorld²⁰, PatientLikeMe²¹) ont prospéré sur le Web au cours de la dernière décennie et sont de plus en plus populaires [Hawn, 2009]. Via ces communautés, les e-patients discutent de manière très active. Par exemple, les patients souffrant de maladies graves, souvent chroniques, obtiennent des informations de la part d'experts médicaux mais également de non-experts sur les diagnostics, les résultats de leurs tests, les options possibles de traitements, etc. Si ces informations médicales leur sont très utiles, ils trouvent également dans ces médias une aide concernant beaucoup d'autres aspects essentiels à leur vie quotidienne, comme des détails techniques sur le port d'une perruque ou du support moral, de l'encouragement mutuel, allant bien au-delà du domaine clinique. Selon plusieurs études [Civan and Pratt, 2007, Meier et al., 2007, Jha and Elhadad, 2010], les e-patients se sentent mieux informés et donc plus forts (*patient empowerment* [Anderson and Funnell, 2010]) suite à l'échange d'expériences avec d'autres personnes vivant ou ayant vécu des expériences similaires.

Selon l'organisation HON²², Internet est la 2^{ème} source d'informations des patients après la visite chez le médecin. 24% des e-patients utilisent Internet pour rechercher des informations sur leur santé au moins une fois par jour (et jusqu'à 6 fois par jour) et 25% l'utilisent plusieurs fois par semaine. 90% utilisent des moteurs de recherche pour formuler leurs demandes. Comme la plupart des liens retournés les redirigent vers des forums, ceux-ci sont utilisés par plus de 50% des e-patients. Les réseaux sociaux occupent une place tout aussi importante. Depuis 2013, 2 milliards d'utilisateurs sont actifs dans les réseaux sociaux²³. Facebook est le champion avec 1,2 milliards, suivi par d'autres réseaux, dont Twitter avec 225 millions. Les e-patients utilisent ces réseaux pour partager leurs pensées, opinions et émotions avec leurs proches et cela inclut des informations sur leur santé. Les ressources issues du Web Social représentent donc une base volumineuse, riche, variée, unique et atypique de connaissances des perceptions qu'ont les e-patients de leur maladie et des soins qui leur sont éventuellement prodigués. Dans ce contexte éminemment subjectif, la caractérisation et la compréhension de ces perceptions sont difficiles, mais néanmoins particulièrement intéressantes dans la perspective de compléter et d'améliorer les programmes de santé.

Toutes ces informations disséminées dans le web social pour la santé peuvent être utilisées comme un vaste réseau de capteurs pour la modélisation de la santé publique à l'échelle de la population [Kautz, 2013]. En examinant manuellement un très grand nombre de tweets, [Kriek et al., 2011] ont montré que les symptômes auto-déclarés forment un signal fiable de l'apparition d'une maladie. Au cours des cinq dernières années, de nombreux travaux ont exploité les médias sociaux pour analyser la propagation des maladies [Sadilek et al., 2012a, Sadilek et al., 2012b, Collier et al., 2011, Culotta, 2010, Chunara et al., 2012]. Par exemple, [Kriek et al., 2011, Sadilek and Kautz, 2013] ont montré que Twitter peut être utilisé pour suivre et prévoir les épidémies de grippe. On trouve d'autres applications que l'épidémiologie. Par exemple, [Paul, 2011] utilise Twitter et des modèles thématiques pour capturer les liens entre symptômes et effets indésirables associés à des traitements. Ces travaux fournissent des preuves solides qu'il existe un réel "signal" dans les médias sociaux [Kautz, 2013], qui peut être exploité pour différentes applications de monitoring des populations. Ces applications peuvent aussi

20. www.entrepaticients.net/fr

21. www.patientslikeme.com/

22. HON (Health On the Net) How Do General Public Search Online Health Information? Avril 2011. http://www.hon.ch/Survey/khresmoi_general_public_survey_results.html

23. <http://b3b.info/2013/12/29/lexplosion-des-reseaux-sociaux-de-2008-a-2013/>

se placer au niveau de l'individu pour déclencher des alertes en cas de risque avéré. Par exemple, [Choudhury et al., 2013] proposent une méthode pour détecter les personnes dépressives en se basant sur leurs tweets. Ces études exploitent la quantité massive de données générées via les médias sociaux et en particulier la fréquence de termes spécifiques (e.g. H1N1, grippe, fièvre) [Chew and Eysenbach, 2010, Lampos and Christianini, 2010]. Ce type d'approches limite la portée de l'analyse à des listes de termes et peut manquer des indices retrouvés avec des méthodes basées sur des méthodes d'apprentissage telles que celles que nous décrirons dans la Section C de ce chapitre. Par ailleurs, les évaluations réalisées dans ces études sont généralement limitées à l'évaluation des algorithmes implémentés avec des mesures telles que le rappel, la précision, etc. mais leur impact en terme d'amélioration du processus de soins reste difficile à estimer, surtout quand les conclusions sont basées sur des échantillons restreints d'exemples bien choisis. Peu d'études cliniques valident ces travaux.

Le verrou principal lorsque l'on cherche à analyser des données issues des médias sociaux de santé est que la plupart des méthodes (semi-automatiques) existantes ont été créées et appliquées sur des textes relativement bien rédigés et structurés tels que des publications, des comptes rendus d'hospitalisation, etc. Concevoir des méthodes destinées aux textes produits par les e-patients est loin d'être trivial et présente un véritable défi technologique, et ceci pour différentes raisons : d'une part, les messages sont écrits suivant des normes peu contraintes (e.g. taille variable des messages, syntaxe et orthographe libres) ou relevant des genres textuels émergents et encore relativement peu étudiés, que ce soit au niveau sémiotique (e.g. émoticons, abréviations, sigles, marqueurs d'emphase, vocabulaire sociolectal) qu'au niveau sémantique (positionnements énonciatifs et dialectiques, modalités temporelles et aspectuelles, thématiques, etc.). Par ailleurs, si les éléments médicaux à rechercher sont généralement bien connus comme les facteurs de risque par exemple, leur expression symptomatique personnelle et intime est extrêmement variable. Comment relier la notion médicale de « vomissement » à « vomito done ! » dans un message de patiente anorexique. Pour finir, le volume des messages est généralement très important (140 millions de tweets sont produits chaque jour !²⁴). Pour ces différentes raisons, traiter les données des médias sociaux avec des méthodes semi-automatiques nécessite des adaptations importantes et complexes.

Dans la suite, nous allons décrire différentes contributions portant sur l'annotation sémantique des messages issus des médias sociaux. La plupart combine l'utilisation et la création de lexiques et des méthodes à base d'apprentissage quand l'élément à capturer est plus diffus. Pour chacune de ces méthodes, nous nous demanderons quels sont les meilleurs descripteurs en comparant ceux relatifs aux textes et ceux plus contextuels et/ou structurels. Ces annotations peuvent être utilisées pour différentes analyses *a posteriori*, pour visualiser le contenu des bases, recommander, classifier, prédire, etc.

La figure 3.1 représente un message enrichi. Une image en couleur est disponible à cette url (<http://www.univ-montp3.fr/miap/sbringay/hdr/images.html>). Un exemple d'informations que nous sommes capables d'extraire à partir de ces phrases enrichies est "*x% des patients femmes*[Qui] *qui postent des messages dans les forums de santé avant une première opération* [Quand] *sont stressées* [Émotion] *à propos des conséquences financières* [Thème]".

Les applications de ces annotations sont multiples. Par exemple, les chercheurs en santé, les administrateurs/modérateurs de médias sociaux et les e-patients peuvent tirer profit de ces informations. Nous travaillons actuellement sur la qualité de vie des patientes atteintes d'un cancer du sein dans le cadre d'une collaboration avec l'ICM²⁵ et l'I3M²⁶ et nous cherchons à montrer qu'il est possible de fournir des connaissances aux oncologues qui peuvent être comparées à celles issues d'essais cliniques plus clas-

24. <http://www.terrafemina.com/culture/culture-web/articles/9812-donnees-numeriques-140-millions-de-tweets-emis-chaque-jour.html>

25. <http://www.icm.unicancer.fr/fr>

26. <http://www.i3m.univ-montp2.fr/>

B. Motivations et challenges liés à l'analyse des productions des e-patients dans les médias sociaux⁴⁷

The image shows a forum post interface. On the left is a user profile card for '(Utilisateur)' with a pink ribbon icon, 'HORS LIGNE' status, 'Membre régulier' rank, and 'Message: 35'. The main post text is: 'Bonjour et bonne année tout le monde ! Ça va faire 3 mois que je suis sous Tamoxifene et depuis environ 2 semaines je ressens des douleurs vers ma cicatrice, comme si mon soutien gorges était trop serré mais ce n'est pas le cas... Ensuite quand je respire, à l'inspiration j'ai comme une douleur intercostal et c'est tout le temps. Avez-vous déjà connu ça? A bientôt'. A legend box on the right lists: 'Qui?' (purple line), 'Quand?' (blue line), 'Comment?' (red line), and 'A propos de quoi?' (teal line).

FIGURE 3.1 : Exemple de message enrichi. On retrouve dans ce message des informations temporelles (*ça va faire 3 mois et depuis environ 2 semaines*). On sait également, d'après le profil, que l'utilisateur est assez nouveau car son rang est *régulier* et il n'a posté que 35 messages. Il poste des questions *Avez vous déjà connu ça ?* et exprime son état affectif *je ressens des douleurs* à propos de thèmes précis (*tamoxifene, cicatrice, douleur intercostal*). Toutes ces informations sont capturées avec les méthodes automatiques que nous proposons.

siques. Par ailleurs, d'après une collaboration menée avec la responsable du site *Vivre Sans Thyroïde*²⁷ (forum de patients), les méta-informations que nous produisons sont également très intéressantes pour les modérateurs afin de repérer les très bons e-patients à valoriser ou à recommander et au contraire les mauvais e-patients à modérer ou à bannir. De plus, les conversations évoluent très souvent. Nous pouvons proposer à ces modérateurs des outils pour réorganiser les files de discussion semi-automatiquement et ainsi maintenir des files thématiquement cohérentes. Finalement, ces méta-informations sont également très pertinentes pour les membres des forums, souvent perdus dans la masse de messages, pour réaliser des recherches plus élaborées que des recherches plein texte, pour leur recommander des sujets, des e-patients ayant des profils similaires aux leurs ou des experts sur une de leur thématique d'intérêt.

27. <https://www.forum-thyroïde.net/>

C Enrichissement de messages

Dans cette section, nous décrivons les méthodes élaborées pour enrichir les contenus produits par les e-patients dans les médias sociaux à partir d'annotations sémantiques organisées selon 4 dimensions d'analyse illustrées par la Figure 3.1 : la thématique *de quoi parle le message* (voir Section C.1), le niveau d'expression de ce locuteur *comment s'exprime-t'il dans le message ?* (voir Section C.2), le locuteur *qui s'exprime dans le message* (voir Section C.3), la temporalité *à quel moment de l'histoire du patient se rapporte le message* (voir Section C.4) ?

1 De quoi ?

TABLE 3.1 : Résumé des projets et des encadrements sur la thématique de l'enrichissement des messages - de quoi ?

Master/Thèse	Projets & Collaborations
A. Abdaoui (Depuis 2013) – Thèse co-encadrement J. Azé	Patients Mind, forums
T. Opitz (2013-2014) – Post-doctorat co-encadrement C. Lavergne et C. Mollevi	ICM et I3M, QdV Cancer
M. Tapi NZali (Depuis 2014) – Thèse co-encadrement C. Lavergne et C. Mollevi	ICM et I3M, QdV Cancer

Pour identifier les thèmes discutés par les e-patients dans les messages, nous nous sommes intéressés à trois approches : 1) la prédiction supervisée de thèmes : on cherche à associer les messages à des catégories prédéfinies ; 2) la recherche d'informations : des thèmes sont décrits par des mots clés et on recherche les messages, parmi la multitude de messages, se rapportant à ces thèmes ; 3) la prédiction non supervisée : on cherche les thèmes d'intérêt décrits par les patients *sans a priori*, c'est-à-dire sans classe prédéfinie.

1.1 Prédiction supervisée de thèmes

Nous avons tout d'abord travaillé sur la détection supervisée de thématiques (Thèse de A. Abdaoui). Nous nous sommes intéressés aux forums de type *Ask the doctor service* via lesquels les utilisateurs posent des questions (rémunérées) à des professionnels de la santé. Pour cibler le bon médecin, les e-patients doivent choisir une catégorie pour chacune de leur question. Cette tâche prend du temps et est source d'erreurs et très souvent, les e-patients choisissent la catégorie "Autre", ce qui ne rend plus possible une recherche ultérieure par catégorie.

Dans le contexte déjà très étudié des techniques de catégorisation de textes [Yang and Liu, 1999], nous avons adapté les travaux de [Himmel et al., 2009] pour classifier automatiquement une demande à un expert médical. Dans notre travail, la tâche est différente car nous cherchons à recommander une liste courte des catégories les plus appropriées en fonction de la question du patient. Dans sa formulation la plus courante, le problème de recommandation est réduit au problème d'estimation (scores) d'items pouvant intéresser un utilisateur [Adomavicius and Tuzhilin, 2005]. Les items avec les scores les plus élevés sont recommandés. Ici, les items sont les catégories des questions et les scores sont estimés en combinant les prédictions de plusieurs modèles de classification entraînés sur des données d'apprentissage. Notre contribution principale a été de proposer un score robuste, adapté de [Kittler et al., 1998], donnant des recommandations précises. Ce score considère à la fois le nombre de modèles qui s'accordent pour assigner une question à une catégorie et les probabilités associées à ces prédictions. Nous avons utilisé différents descripteurs comme les Uni-grammes et Bi-grammes extraits du texte des questions et du titre,

le genre et l'âge du patient, etc. Nous avons évalué l'efficacité de notre approche sur les sites *AlloDocteur*²⁸ et *MaSanteNet*²⁹. Notre méthode s'est avérée efficace avec une F-mesure d'environ 0.90. Comme on s'y attendait, la combinaison des votes donne de meilleurs résultats que les classifieurs seuls car les ensembles de questions mal classées par les différents classifieurs ne se recouvrent pas forcément.

Ce travail peut être facilement amélioré en adaptant le nombre de recommandations selon les scores obtenus avec les différents classifieurs et la longueur de la question. En effet, nous pouvons recommander une seule catégorie si son score est le plus élevé et très différent des scores des autres catégories. On recommandera deux catégories si leurs scores sont élevés et éloignés des autres catégories, etc. De plus, nous avons remarqué que notre méthode a un meilleur comportement sur des questions longues. Le nombre de propositions peut donc également être adapté en fonction de la longueur de la question. Enfin, l'identité et l'histoire de l'utilisateur sont évidemment des éléments importants à inclure. Si un utilisateur pose beaucoup de questions dans la catégorie "foie", il a très probablement une maladie du foie et il continuera à poser des questions dans cette catégorie, même s'il s'est trompé de pathologie dès le début.

Finalement, ce travail n'est pas spécifique aux sites de type *Ask the doctor service* et peut être étendu à tous les types de forums de santé en ligne. Cette application peut s'avérer particulièrement pertinente pour les membres qui postent des messages dans les forums afin de les aider à choisir des files de discussions les plus adaptées à leurs thématiques. Cette application peut également être utile pour les modérateurs afin de repérer les messages mal classés. Une adaptation est possible pour mesurer les dérives au sein des files de discussion [Abadaoui et al., 2015].

1.2 Recherche d'informations

Nous nous sommes intéressés à une autre tâche classique en recherche d'informations consistant à retrouver des messages pertinents selon une liste de mots clés (Post-doctorat de T. Opitz). Nous avons défini une liste de thèmes d'intérêt à partir de questionnaires remplis par les patients dans leurs trajectoires de soin. Nous avons recherché des messages thématiquement liés dans des forums spécialisés. Une recherche plein texte ne retourne pas l'ensemble des messages ayant un rapport thématique avec les mots clés spécifiés par l'utilisateur, notamment à cause de la richesse morphologique de la langue française. Par exemple, une requête comme "*bouche sèche*", génère de nombreux faux négatifs car elle ne capture pas des occurrences comme "*bouche desséchée*" ou "*langue sèche*". Au contraire, des requêtes trop générales comme "*bouche*" ou "*sècheresse*" retournent de très nombreux faux positifs comme "*j'en ai l'eau à la bouche*".

Nous avons donc réalisé une expansion des requêtes basées sur les mots clés en définissant avec les professionnels de santé des termes morphologiquement proches [Jacquemin and Royauté, 1994], puis en utilisant automatiquement la synonymie. Pour ce faire, nous avons interrogé automatiquement plusieurs

28. *Allodocteurs.com* Le jeu de données comporte 16, 609 messages publiquement accessibles écrits par les 6, 256 membres entre Juin 2009 et Novembre 2013 et récoltés à la date du 19/11/2013. Ce forum contient 4, 615 files de discussions dont certaines contiennent plus de 150 réponses. Ce forum couvre une large gamme de sujets tels que le cancer, la nutrition, les médicaments potentiellement dangereux, la maternité, etc. Il contient deux catégories d'utilisateurs : les professionnels de santé et les non professionnels. En général, la première catégorie d'utilisateurs répond à des questions posées par la deuxième. Les professionnels de santé peuvent être des médecins généralistes ou spécialistes mais ils peuvent aussi être des étudiants en médecine. Bien que leur nombre soit restreint (16 utilisateurs), leur participation aux échanges sur le forum est considérable : ils ont postés 3, 050 messages sur l'ensemble des messages collectés.

29. Ce jeu de données est issu du site *MaSanteNet* qui est un *ask the doctor service*, soumis à rémunération qui permet à des internautes de poser des questions à des professionnels de santé. Toutes les questions posées sur le site ont une réponse. Plus de 12, 000 posts ont été récoltés à la date du 18/02/2014 équitablement répartis entre e-patient et médecin. Ce forum couvre une large gamme de sujets tels que la nutrition, la dermatologie, etc.

sites en ligne pour trouver des synonymes aux mots clés. Pour désambiguïser les synonymes et ne garder que les plus pertinents, nous avons interrogé un moteur de recherche pour construire des contextes (voir figure 3.2). Les contextes sont des sacs de mots associés à leurs fréquences, qui co-occurrent avec les termes de la requête initiale dans les extraits de pages (snippet) retournés par le moteur de recherche. Nous avons ensuite comparé les contextes associés aux requêtes initiales et les contextes associés aux requêtes basées sur les synonymes avec une formule adaptée du test statistique du χ^2 . Nous ne retenons que les requêtes thématiquement liés.

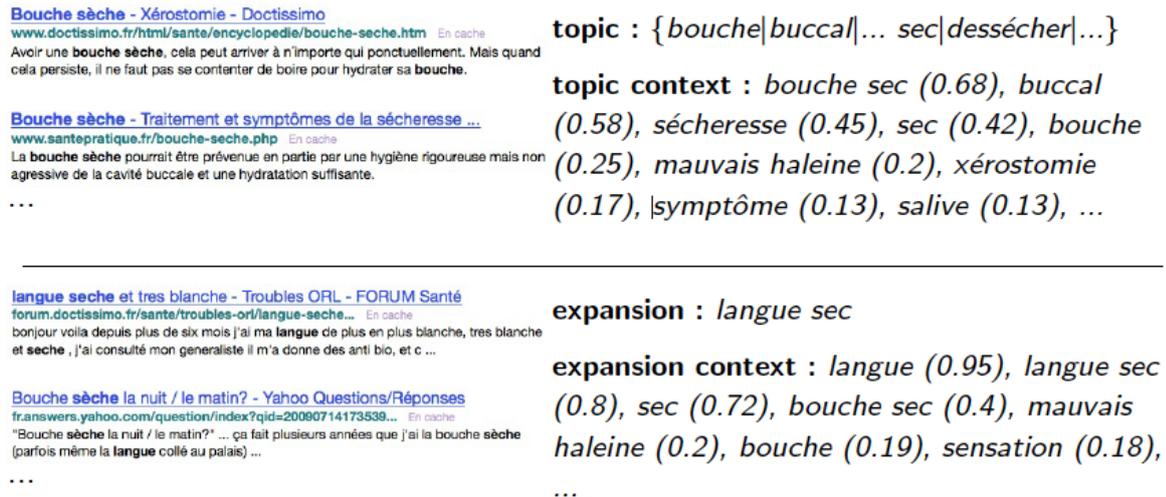


FIGURE 3.2 : Exemple de contextes construits. Sur la gauche de la figure, nous retrouvons les snippets de textes retournés par le moteur de recherche et sur la droite, la requête initiale (voir ligne *topic*) et une expansion de cette requête possible (voir ligne *expansion*) et les contextes correspondants, c'est à dire la suite de termes pondérés par leurs fréquences dans les snippets.

Nous avons expérimenté cette méthode sur le site *cancerdusein.org*³⁰ et obtenu des résultats très encourageants. Nous avons comparé les messages retournés uniquement avec les termes des requêtes sans expansion avec les messages retournés avec l'expansion morphologique manuelle uniquement, puis avec les messages retournés avec l'expansion automatique par synonyme. Nous avons évalué le nombre de vrais positifs pour différents sujets. Nous avons montré que nous retrouvons de nombreux messages pertinents qui auraient été oubliés avec une simple recherche par mots clé. Par exemple, pour la thématique "sein", on retrouve deux fois plus de messages (de 250 à 480). Une application possible consiste à intégrer ce type d'expansions pour améliorer les fonctionnalités de recherche dans les forums [Opitz et al., 2014].

1.3 Détection non supervisée de thèmes

Nous nous sommes également intéressés au problème dual de détection non supervisée de thèmes (Post-doctorat de T. Opitz et thèse de M. Tapi Nzali). Plutôt que de partir de listes de thèmes d'intérêt prédéfinis, nous recherchons *sans a priori* les thèmes dont les e-patients parlent.

Récemment, la détection de structures sémantiques latentes et de thèmes latents est devenue un domaine de recherche très actif dans la communauté de fouille de texte. Nous recherchons ici une caracté-

30. *cancerdusein.org*. Le jeu de données comporte 16, 961 messages publiquement accessibles écrits par les 675 membres du forum *CancerDuSein.org* entre octobre 2011 et octobre 2013, La récupération a été effectuée le 01 novembre 2013. Ce forum contient 1, 050 files de discussions dont certaines contiennent plus de 500 réponses. Ce forum couvre des sujets comme les traitements, les récurrences, les bonnes nouvelles, les nouvelles des anciennes, etc.

térisation fine des thèmes avec une information sur l'attribution d'un terme à un thème et l'attribution d'un thème à un message. Pour cela, nous appliquons la méthode très classique LDA [Blei et al., 2003] pour la modélisation des thèmes. C'est un modèle bayésien hiérarchique, devenu standard pour la détection non supervisée de thèmes dans les corpus textuels. Cette méthode est basée sur une représentation des documents sous la forme d'un sac de mots. Des distributions postérieures représentent chaque sujet comme une distribution de probabilités sur tous les termes et chaque message comme une distribution de probabilités sur tous les thèmes. De plus, chaque terme est attribué à un ou plusieurs thème(s) de chaque document.

Dans ce travail, nous avons considéré comme entrées, les termes ayant un intérêt médical tels que les termes de la ressource MeSH, les listes de médicaments et de traitements non conventionnels. Nous avons examiné différentes formes grammaticales après application de l'outil d'annotation de texte Tree-Tagger³¹ avec des informations grammaticales pour ne garder que les noms, les verbes, etc. Nous avons défini manuellement, en collaboration avec les professionnels de la santé, les paramètres du modèle, en particulier k le nombre de thèmes, α le nombre de descripteurs d'un thème (quelques termes vs. de nombreux termes) et δ pour l'attribution de thèmes à documents (quelques thèmes vs. de nombreux thèmes). Ces paramètres sont définis *a priori*. Nous avons appliqué cette méthode sur le site *cancerdusein.org* (voir Section 1.2) afin d'identifier et de quantifier des thèmes d'intérêt des e-patients utiles pour la compréhension de la qualité de vie par les oncologues. Nous avons comparé l'impact des catégories grammaticales sur les résultats et noté que si nous utilisions plutôt des noms, la plupart des thèmes étaient de nature factuelle alors que si nous introduisions des verbes, nous obtenions des informations sur les actions des utilisateurs (e.g. *attendre, consulter, rechercher*) et sur leurs sentiments (e.g. *ressentir, crier, avoir peur, accepter*). Si le niveau de description change selon ces catégories grammaticales, nous avons pu noter une relative stabilité des thèmes identifiés quelque soit le scénario retenu, due à la similarité des termes dominants (les plus fréquents). Lorsque l'on considère les adjectifs, on introduit de nombreuses références aux sentiments difficiles à interpréter médicalement. Nous pensons que l'analyse des sentiments doit être faite séparément avec des approches comme celle de [Lin et al., 2012] qui modélisent conjointement les thèmes et les sentiments avec une extension de la structure hiérarchique du modèle LDA. Dans le cadre de la thèse de M. Tapi Nzali, ayant débuté en octobre 2014, nous travaillons sur une adaptation de cette méthode pour suivre l'évolution temporelle des thèmes selon l'histoire du patient et le vocabulaire utilisé.

Sélection de références

A. Abdaoui, J. Azé, S. Bringay and P. Poncelet. Assisting e-patients in an Ask the Doctor Service. Proceedings MIE 2015, Madrid (Spain), 2015, Studies in Health Technology and Informatics (A paraitre)
 A. Abdaoui, J. Azé, S. Bringay, N. Grabar and P. Poncelet. Analyse des messages des patients et des médecins dans les forums de santé. *IC et Santé* (IC 2014). Clermont ferrand, France, 6 pages
 T. Opitz, J. Azé, S. Bringay, C. Joutard, C. Lavergne, C. Mollevi : Breast Cancer and Quality of Life : Medical Information Extraction from Health Forums. Proceedings MIE 2014, Istanbul (Turkish), 2014, Studies in Health Technology and Informatics : 1070-1074

2 Comment ?

Dans cette section, nous présentons uniquement les méthodes d'analyse de sentiments développées pour expliciter la manière dont les e-patients s'expriment. D'autres marqueurs comme l'incertitude [Thoumelin and Grabar, 2014] ou le risque³² [Kermisch, 2011] peuvent être étudiés de manière similaire.

31. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

32. <http://cediscor.revues.org/195>

TABLE 3.2 : Résumé des projets et des encadrements sur la thématique de l'enrichissement des messages - comment ?

Master/Thèse	Projets & Collaborations
S. Melzi (2013) – Master co-encadrement P. Poncelet	Patients Mind, forums
A. Abdaoui (Depuis 2013) – Thèse co-encadrement J. Azé	Patients Mind, forums
M. Tapi NZali (Depuis 2014) – Thèse co-encadrement C. Lavergne et C. Mollevi	ICM et I3M, QdV cancer

Dans le domaine du traitement du langage naturel, l'*analyse des sentiments* a beaucoup été étudiée depuis le début des années 2000. Elle implique l'extraction d'états émotionnels explicites ou implicites dans les textes [Liu, 2012]. La Figure 3.3 résume les différents éléments impliqués dans un sentiment. De nombreuses communautés ont donné des définitions et interprétations variées (e.g. en psychologie, linguistique, traitement du langage naturel, etc.). Plusieurs revues exhaustives présentent ces travaux [Pang and Lee, 2008, Jackiewicz et al., 2010, Roche and Prince, 2009]. L'analyse des sentiments comprend les 4 tâches principales suivantes :

1. l'analyse de la subjectivité [Wiebe et al., 2005] porte sur la détection de la présence de sentiments dans les textes ;
2. l'analyse de la polarité [Boiy et al., 2007] se concentre sur la détection de la polarité (*positive*, *négative* et *neutre*) des sentiments exprimés dans les textes ;
3. l'analyse de l'émotion [Lu et al., 2006] met l'accent sur la catégorie émotionnelle des textes (e.g. *colère*, *dégoût*, *peur*). Beaucoup de typologies d'émotions ont été définies. Celle de [Ekman, 1999] est souvent utilisée et décrit six émotions, mais beaucoup d'autres typologies existent [Plutchik, 1980, Pearl and Steyvers, 2010, Francisco and Gervás, 2006] ;
4. l'analyse de l'intensité du sentiment [Mulder et al., 2004] décrit différents niveaux d'intensité du sentiment (e.g. *très positif*, *très triste*). Ces approches offrent une granularité plus précise des opinions et des émotions exprimées.

Dans la suite, nous explorons les tâches 2 et 3 pour annoter sémantiquement les messages avec des informations sur la polarité et l'émotion dans la Section 2.1. Puis, nous expliquerons comment nous avons élaboré une ressource pour réaliser ces deux tâches en français (voir Section 2.2). Finalement, nous présenterons une méthode originale pour ajouter des informations à cet enrichissement en identifiant les sources (*qui ressent le sentiment ?*) et les cibles (*à propos de quoi ?*) de ces sentiments (voir Section 2.3).

2.1 Sentiment : polarités et émotions

La plupart des travaux de la littérature se concentrent soit sur la création de ressources pour décrire les sentiments, soit sur l'utilisation de ces ressources pour classer les textes selon ces sentiments (en *subjectif/objectif*, *positif/négatif/neutre*, *joie/colère*). Les méthodes utilisées pour analyser les sentiments sont généralement spécifiques au type de texte, par exemple aux tweets [Roberts et al., 2012], aux titres de presse [Strapparava and Mihalcea, 2008], etc. et aux domaines d'application, par exemple, l'impact de genre dans les négociations [Boneva et al., 2001], l'identification des emails suicidaires [Pestian et al., 2012], etc. Alors que beaucoup de ces méthodes sont efficaces sur de grands corpus de texte, leur efficacité est plus limitée dans le cas de textes courts tels que les tweets [Agarwal et al., 2011]. Il existe deux approches principales : l'une basée sur le comptage du nombre de termes d'une catégorie (e.g. nombre de termes positifs et négatifs) apparaissant dans un texte [Turney, 2002], et l'autre basée sur l'apprentissage à partir de données préalablement annotées [Pang, 2002, Esuli and Sebastiani, 2006]. Les approches hybrides se sont révélées être les meilleures [Wiegand and Klakow, 2010]. Dans le

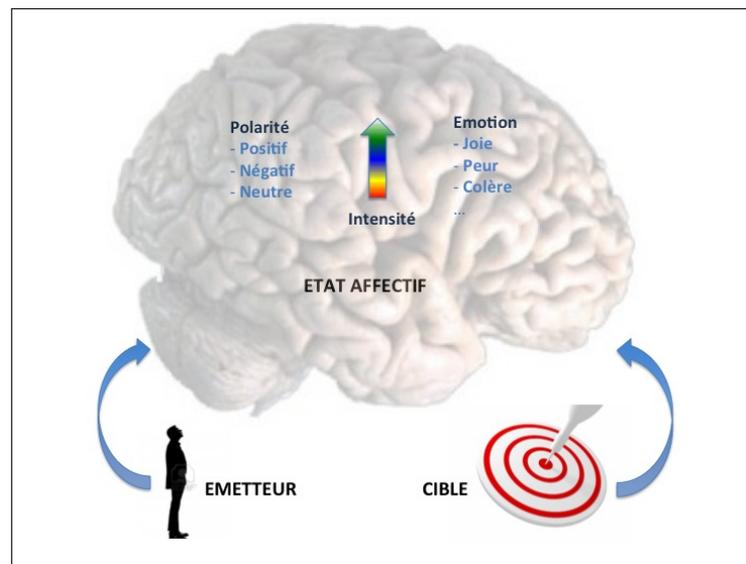


FIGURE 3.3 : Modèle des états affectifs. Un *état affectif* est ressenti par un *émetteur* (ou *source*). Il fait référence à une *polarité*, c'est-à-dire un jugement pouvant être *positif* s'il est lié à un effet bénéfique sur l'émetteur, *négatif* dans le cas contraire ou *neutre*. L'état affectif peut également faire référence à une *émotion* comme la *colère*, la *joie*, la *tristesse*, etc. Généralement, les émotions sont associées à une polarité. La joie est considérée par exemple comme positive, la colère comme négative et la surprise comme neutre. On peut associer différents niveaux d'*intensité* à l'état affectif (e.g., *très positif*, *un peu triste*, etc.). Pour finir, l'état affectif porte sur une *cible* qui est le réceptacle de l'opinion ou de l'émotion [Bringay et al., 2014].

contexte de blog et de micro-blog, les caractéristiques extraites des textes peuvent renforcer l'analyse de sentiments tels que des émoticônes [Alec et al., 2009, Davidov et al., 2010, Ruan, 2011], les hashtags [Barbosa and Feng, 2010], etc. En outre, plusieurs méthodes de classification peuvent être combinées avec des techniques de votes [Planté et al., 2008] ou en appliquant des méthodes de boosting et de bagging [Fan et al., 2011]. Des approches basées sur des systèmes incrémentaux sont également envisageables et pertinentes lorsque les données arrivent en flux [Wiebe and Riloff, 2011].

Dans nos travaux, nous nous sommes tout d'abord intéressés aux polarités et émotions exprimées par les e-patients dans les messages (master de S. Melzi puis thèse de A. Abdaoui). Par exemple, dans le message "*Je n'aime pas cette nouvelle perruque qui m'a coûté une fortune, elle me gratte ! !*", l'e-patient exprime une opinion (polarité négative) sur sa nouvelle perruque. Dans la phrase, "*Je me sens trop mal à cause de la douleur :-((*", l'e-patient exprime son mal être et sa tristesse via l'expression "*se sentir mal*" et le smiley. Nous avons défini une méthode pour identifier les phrases dans les messages contenant ces marques de sentiments. Pour cela, nous avons étudié des messages issus du site anglais SpineHealth³³ traitant de la douleur du dos, que nous avons annoté avec le lexique des mots d'émotions NRC [Mohammad, 2010], avec une liste d'expressions iconiques contenant entre autres les smileys (: -), LOL), et avec d'autres marqueurs graphiques comme la mise en majuscule pour l'emphase ou encore la présence de ponctuations répétées. Nous nous sommes également intéressés aux structures grammaticales fréquentes, spécifiques à l'opinion et à chaque émotion en appliquant des motifs séquentiels décrits dans le chapitre 2. Nous avons utilisé le contexte correspondant à la polarité de la phrase précédant ou suivant la phrase à classer dans le message.

Nous avons utilisé un sous ensemble annoté pour apprendre des modèles de classification et avons comparé différents algorithmes [Melzi et al., 2014]. SVM est celui qui donne les meilleurs résultats. Les Uni-grammes et les Bi-grammes se sont révélés être les meilleurs descripteurs, les smileys beaucoup moins à cause de leur utilisation ironique. Les motifs extraits ont sensiblement amélioré les résultats de la classification. Le contexte n'apporte pas d'information car deux phrases consécutives peuvent contenir des émotions décorréelées. La classification des polarités donnent de meilleurs résultats que la classification des émotions. Cela est cohérent avec l'accord entre annotateurs beaucoup moins bon pour les émotions que pour les polarités. La raison principale est que l'implicite est beaucoup plus fort lorsque l'on exprime des émotions que son opinion [Melzi et al., 2014]. Une étude des cas d'erreurs, nous a montré que lorsque le classifieur se trompe pour les émotions, il suggère généralement une étiquette associée à une classe proche (e.g. colère est plus proche de dégoût que de joie).

Cette annotation peut être très utile, combinée au marquage thématique présenté dans la section précédente car en s'inspirant des approches de [Liu et al., 2005] sur la comparaison de produits ou de [Carenini et al., 2006] sur le résumé d'opinions, nous pouvons étudier le ressenti des patients par rapport à des thématiques d'intérêt (e.g. les polarités et émotions associées à un traitement). Il est important de noter également que nous sommes encore en train d'améliorer cette chaîne de traitements en participant cette année au défi DEFT 2015³⁴ (Thèse de M. Tapi NZali et A. Abdaoui).

2.2 Constitution d'une ressource des émotions pour le français

Dans un deuxième temps, nous avons décidé de transposer ce travail au français. Pour utiliser la même méthodologie que décrite dans la section 2.1, nous avons besoin d'une ressource en français pour les sentiments et les émotions. Or, la plupart de ces ressources ont été créées pour des textes en anglais et l'analyse de polarité (*General Inquirer* [Stone and Hunt, 1963], *Linguistic Inquiry and*

33. <http://spinehealth.com/>

34. <https://deft.limsi.fr/2015/>

Word count [Tausczik and Pennebaker, 2010], MICROWNOP [Cerini et al., 2007], SENTIWORDNET [Baccianella et al., 2010], etc). Plusieurs ressources spécifiques (dictionnaire DAL [Whissell, 1989], WORDNET AFFECT [Strapparava and Valitutti, 2004], lexique NRC [Mohammad, 2010]), SENTICNET [Cambria et al., 2010] ont été créées pour les émotions. Il existe aussi des approches pour étendre ou spécialiser ces vocabulaires pour des domaines d'applications spécifiques en construisant des règles manuelles [Neviarouskaya et al., 2011], ou en identifiant des mots co-occurents avec des mots déjà identifiés comme désignant des émotions dans de grands corpus [Harb et al., 2008] ou le web [Kozareva et al., 2007].

En français, il n'existe pas de ressource équivalente au lexique NRC pour les émotions, seulement un lexique composé d'environ 1,300 termes [Augustyn et al., 2008]. Nous avons donc traduit semi-automatiquement la ressource NRC en interrogeant cinq traducteurs en ligne (thèse de A. Abdaoui). Nous avons utilisé la règle ci-contre : si trois traducteurs s'accordent, nous retenons la traduction. Un traducteur humain expert (Stage de Master de Traduction UM3 de C. Fournier) a validé entièrement ce lexique traduit et les polarités et émotions associées en deux mois. La règle automatique s'est avérée vraie pour plus de 90% des traductions validées ou invalidées. Nous avons procédé de même pour étendre cette ressource aux synonymes et finalement, nous avons obtenu plus de 14 000 termes polarisés et associés à des émotions. Cette ressource nous permet désormais de travailler sur des textes en français. Cette ressource est disponible pour la communauté [Abdaoui et al., 2014c]. Nous travaillons désormais sur la spécialisation de cette ressource pour le domaine médical. Par exemple, *crabe* n'est pas polarisé en général mais dans le domaine médical, *crabe* est *néгатif* car signifiant *cancer* dans les forums traitant de cette maladie.

2.3 Cibles et sources des sentiments

Après avoir identifié les phrases contenant des sentiments (opinions et émotions), nous nous sommes demandés qui exprimait ces sentiments et à propos de quels sujets. Notre avons proposé une méthode pour expliciter les traces d'émotions en identifiant dans les textes la source (qui ?) et la cible ou un contexte (quoi ?), si possible médical, facilitant l'interprétation.

Considérons uniquement le cas des cibles. Elles sont généralement présentes dans les textes sous la forme d'Entités Nommées (EN) comme des noms de personnes ou de médicaments, d'événements, de concepts abstraits, de caractéristiques associées à ces concepts abstraits ou de contextes généraux traités par l'état affectif [Popescu and Etzioni, 2005, Wilson et al., 2005]. Considérons les exemples décrits dans la Table 3.3. Il est parfois difficile de distinguer la cible des circonstances ou causes ayant suscité l'état affectif. C'est le cas par exemple dans la phrase P8. Ces exemples illustrent la complexité de la tâche visée, consistant à identifier des cibles pouvant s'exprimer de manières très variées. Contrairement à la plupart des approches de la littérature, nous avons pris le parti de ne pas limiter la cible à quelques mots mais au contraire de proposer le plus d'informations possibles comme dans les phrases P7 et P8.

Pour cela, nous avons proposé d'intégrer l'analyse sémantique de surface (*Shallow semantic parsing*) et d'utiliser la ressource lexicale FRAMENET³⁵. Basée sur la notion de rôle sémantique définie par [Baker et al., 1998], elle décrit schématiquement des situations grâce à un système relationnel de concepts (*quels éléments dans la phrase participent, subissent, causent, etc. une situation ?*). Une annotation basée sur cette ressource identifie dans les phrases des expressions de sentiments et explicite les constituants liés à ce sentiment. Nous avons proposé une typologie des annotations dédiées à notre contexte d'étude spécifique. À notre connaissance, il n'existe pas de méthode basée sur ce type d'annotations et personnalisée pour les forums de santé. Nous avons comparé cette approche à celle plus classique

35. <https://FrameNet.icsi.berkeley.edu/fndrupal/home>

P1 : J'ai peur de ce médicament.
P2 : J'ai peur de commencer la chimiothérapie.
P3 : J'ai peur de prendre de l'IVEMEND.
P4 : Le taux de tolérance pour la moyenne des patients pour ce médicament est excellent !
P5 : Son <u>taux de tolérance</u> est excellent .
P6 : J'ai peur .
P7 : J'ai peur de vivre dans la douleur encore une dizaine d'années.
P8 : On m'a donné une chance sur 10 de vivre pendant 10 ans, annoncé de réelles perspectives de <u>récurrence</u> et la peur est restée avec moi pendant tout ce temps.

TABLE 3.3 : Exemple d'expressions de la cible de l'émotion. Dans les phrases P1, P2 et P3, l'émotion *peur* porte sur une entité représentée respectivement par le concept général *médicament*, l'événement *début de la chimiothérapie* et l'**EN** *IVEMEND* (qui est un nom de médicament). Dans la phrase P4, la cible de la polarité est plus complexe et porte sur un *aspect*, une caractéristique : *le taux de tolérance* de l'entité *médicament*. Dans la phrase P5, seul l'aspect est présent. Dans la phrase P6, il n'y a pas de cible explicite. L'émotion fait référence au contexte général dans lequel la phrase est énoncée. Dans la phrase P7, la cible est détaillée dans le reste de la phrase et ne se limite pas à l'entité médicale *douleur*.

qui se base sur un calcul de distance (nombre de mots entre le mot de sentiment et une cible prédéfinie dans la phrase ou dans l'arbre syntaxique). Notre méthode [Bringay et al., 2014] a été expérimentée avec succès sur un jeu de données réelles et validée par 10 annotateurs humains, tous chercheurs en informatique. Notre méthode s'est avérée très efficace par rapport aux méthodes basées sur les distances dans la phrase mais incontestablement en raison de la robustesse de l'analyseur SEMAFOR³⁶ associé à FRAMENET utilisé sur les forums. La deuxième raison est qu'il est difficile d'établir une liste prédéfinie de sources et de cibles potentielles, contrairement aux revues de produits par exemple pour lesquels les cibles sont les caractéristiques des produits vendus identifiés par leur forte fréquence dans les revues.

Cette méthode pourra être généralisée à d'autres états affectifs comme l'incertitude et à d'autres domaines d'application que la santé. Nous n'avons pu appliquer directement ces travaux au français car il n'existe pas de ressource équivalente à FRAMENET. Nous travaillons actuellement sur une adaptation basée sur le réseau lexical JeuxDeMots³⁷, qui contient les rôles de *patient* et d'*agent* essentiels dans notre méthodologie pour simuler l'analyse en rôle sémantique. À partir du mot étiqueté par une émotion, on recherche dans le réseau lexical le chemin le plus court vers un autre mot de la phrase qui contient au moins un arc étiqueté par une relation *agent* ou *patient*. Un exemple de chemin entre un mot de sentiment et une source potentielle obtenu à partir de la phrase "j'aime le chocolat" est : *je* → *specialisation* → *personne* → *agent* → *aimer*.

Sélection de références

- A. Abdaoui, J. Azé, S. Bringay et P. Poncelet. FEEL : French Extended Emotional Lexicon. 2014. ISLRN : 041-639-484-224-2
- S. Melzi, A. Abdaoui, J. Azé, S. Bringay, P. Poncelet and F. Galtier. Patient's rationale : Patient Knowledge retrieval from health forums'. Proceedings 6th International Conference on eHealth, Telemedicine, and Social Medicine ETELEMED 2014, Barcelona (Spain), 2014 : 140-145
- S. Bringay, E. Kergosien, P. Pompidor, P. Poncelet : Identifying the Targets of the Emotions Expressed in Health Forums. Proceedings CICLing (2), Kathmandu (Nepal), 2014 : 85-97 (Rang B)

36. <https://code.google.com/p/semafor-semantic-parser/wiki/FrameNet>

37. JeuxDeMots est un jeu contributif dont le but est de construire un vaste réseau lexical-sémantique. Cette ressource, construite par les internautes, rassemble 112 types de relations dont 179 578 occurrences de la relation synonymie. <http://www.jeuxdemots.org/jdm-accueil.php>

3 Qui ?

TABLE 3.4 : Résumé des projets et des encadrements sur la thématique de l’enrichissement des messages - qui ?

Master/Thèse	Projets & Collaborations
A. Abdaoui (Depuis 2013) – Thèse co-encadrement J. Azé	Patients Mind, Vivre Sans Thyroïde

Dans cette section, nous nous focalisons sur les e-patients qui s’expriment dans les messages. Nous nous intéressons au rôle du e-patient dans la communication supportée par le média social (voir Section 3.1) puis à sa réputation dans la communauté (voir Section 3.2).

3.1 Détection de rôles médicaux

Les professionnels, les experts non professionnels, les e-patients et les proches des e-patients sont les trois principaux rôles sur lesquels nous avons travaillé. De nombreux travaux définissent la notion de rôle qui dépend du contexte d’application [Merton, 1957, White et al., 1976, Wolfe and Jensen, 2004]. Par exemple, le rôle est considéré comme une position professionnelle dans les mails des entreprises [McCallum et al., 2007]. Le rôle est lié au niveau de connaissances sur un sujet comme dans les discussions Web [Zhang et al., 2007], etc. Nous considérons dans ce travail la définition suivante adaptée de [Forestier et al., 2012], combinant ces deux aspects et adapté à notre cas d’étude. Nous ajoutons une notion portant sur le point de vue du lecteur, c’est-à-dire sur ses attentes par rapport aux productions textuelles de la personne selon son rôle : *Dans un forum de santé, un rôle est lié à une position (i.e. être ou non un professionnel de santé reconnu, être un proche du patient ou le patient) ainsi qu’à un ensemble de connaissances et compétences qui affectent le comportement d’un individu en fonction de sa position et qui sont attendues par les autres personnes qui sont reliées à lui au sein de la structure sociale.* Dans notre contexte, un *professionnel* sera un médecin, un interne, une infirmière. Les lecteurs attendent d’eux une certaine qualité de leurs messages : des réponses à leurs questions, des références, des avis, des descriptions claires des problèmes et des symptômes, une correction des fausses affirmations, etc. Un *expert* n’aura pas de qualification médicale mais s’exprimera comme un professionnel de la santé. Les lecteurs attendent également de ces experts une certaine qualité dans leurs réponses. Les *experts*, comme les professionnels, ne se focalisent pas sur leurs sentiments personnels et leurs expériences mais plutôt sur les demandes et soucis des personnes à qui ils répondent. L’*e-patient*, comme le *proche* n’aura pas de qualification médicale. On attend de lui qu’il pose les questions, exprime ses sentiments, etc.

Selon son besoin informationnel, le lecteur aura envie qu’on lui recommande soit un professionnel ou un expert quand il a, par exemple, une demande technique (*e.g. comment interpréter mes résultats de laboratoire ?*). L’e-patient en quête de réconfort recherchera plutôt un autre e-patient ayant vécu une expérience similaire à la sienne (*e.g. je me sens si triste depuis la récurrence de mon cancer*). Dans la suite, nous allons décrire une méthode qui distingue les messages des experts (professionnels ou non), des messages des non experts (e-patients ou proches) en se basant sur leurs productions textuelles. Nous avons simplifié le problème et ne parlerons plus dans cette section que d’experts et de non experts. Les figures 3.4 et 3.5 donnent des exemples de messages experts et non experts. Or, la plupart des sites ne donnent pas ces informations sur les rôles des utilisateurs.

Selon [Welser et al., 2007], chaque rôle social a une signature, qui peut être comprise comme l’ensemble des modèles comportementaux et structurels. De même, [Schwartz and Sprinzen, 1984] fait l’hypothèse que les acteurs dans une structure sociale avec des rôles similaires partagent des caractéristiques communes et des modèles communs de relations. Identifier ces rôles revient à identifier les personnes

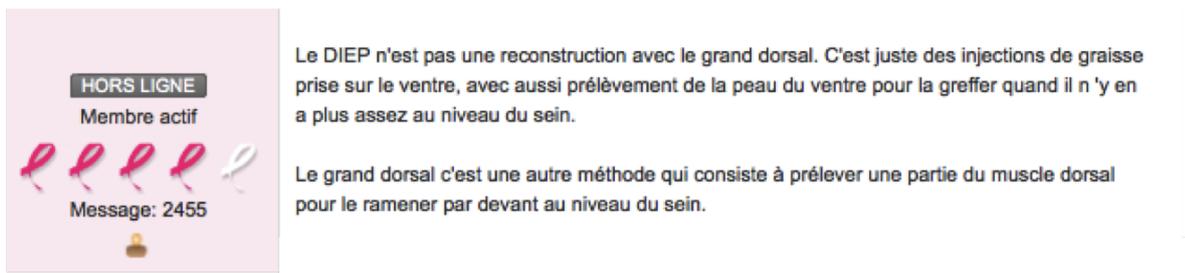


FIGURE 3.4 : Exemple de message expert. L'utilisateur a posté 2,455 messages. Son niveau de langue est élevé. Le message ne contient pas de faute d'orthographe. Sa description est factuelle.

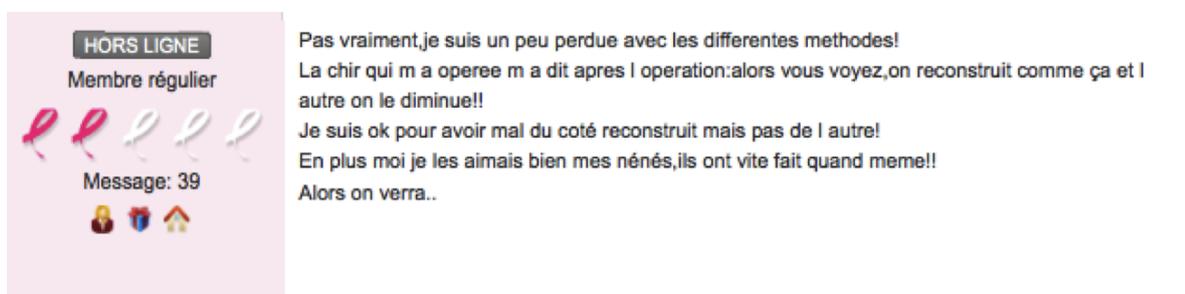


FIGURE 3.5 : Exemple de message non expert. L'utilisateur n'a posté que 39 messages. Le message contient des fautes d'orthographe, des abréviations, etc. L'auteur donne son ressenti (*je suis un peu perdue*).

qui partagent des caractéristiques et des modes de relations communes. Une étude statistique des textes nous a montré que les experts et les non experts, partagent certaines caractéristiques. Le sujet des messages est différent. En effet, les experts s'intéressent plus aux traitements alors que les messages des non experts portent plus sur les symptômes. Les experts utilisent un vocabulaire spécifique et codifié (e.g. *céphalée*), différent de celui utilisé par les non experts, relevant parfois plus de l'oral que de l'écrit dans sa forme (e.g. *mal de tête*). Les marques de politesse sont plus présentes dans les productions des experts (e.g. *Cordialement*) alors que les non experts utilisent plus le remerciement (e.g. *Merci à toutes*). Les experts font peu de fautes d'orthographe comparé aux non experts. Ces derniers utilisent beaucoup plus les formes interrogatives dans leurs productions. Les experts n'expriment pas leurs sentiments alors que les non experts le font beaucoup, par exemple pour exprimer leur tristesse liée à la maladie, la douleur, etc. Au contraire, les experts utilisent plus de mots d'incertitude, par exemple pour donner un avis médical incertain : *vous pouvez avoir une artérite*, etc.

Dans ce travail, nous avons proposé de considérer ces caractéristiques pour prédire efficacement le rôle médical. Nous avons proposé (thèse de A. Abdaoui) une approche de fouille de texte qui considère ces marqueurs associés aux classiques Uni-grammes et Bi-grammes afin pour distinguer les messages postés par les experts et ceux postés par les e-patients. Les meilleurs résultats ont été obtenus avec le classifieur SVM pour une F-mesure d'environ 0.95 [Abdaoui et al., 2014b]. Une perspective consiste à utiliser les approches à bases de motifs pour définir des nouveaux descripteurs comme nous l'avons fait pour les sentiments.

3.2 Confiance et réputation

Aujourd'hui, plus de 46% des e-patients issus de 12 pays différents utilisent Internet pour l'auto-diagnostic [Wald et al., 2007]. Les e-patients ont un fort désir d'apprendre, de comprendre leurs propres symptômes et d'avoir accès aux connaissances médicales. Bien que les informations médicales qu'ils trouvent sur le web soient facilement accessibles, la fiabilité de ces informations représente un risque majeur. Les conséquences d'un auto-diagnostic erroné sont difficiles à estimer si des mesures sont prises sans consulter un médecin. Or, seulement 21% des e-patients demandent une confirmation à leur médecin relative aux informations obtenues sur Internet [Wald et al., 2007]. S'il est difficile d'empêcher les e-patients de consulter des informations non fiables dans les forums de santé, il est possible de concevoir des outils pour mettre en évidence de l'information que l'on sait fiable et les auteurs de ces productions fiables.

Dans le cas des forums de santé, ce problème de fiabilité des informations est difficile à mesurer. Dans les travaux décrits dans la Section 3.1, nous nous sommes intéressés à la notion de rôle, en supposant que les informations données par une personne ayant le rôle d'expert n'étaient sans doute pas toutes fiables mais pouvaient majoritairement être recommandées. Ici l'approche est différente. Il ne s'agit plus d'associer un membre à une catégorie mais de lui donner un score de réputation qui représente l'opinion des autres membres à son sujet (thèse de A. Abdaoui). Pour calculer cette réputation, beaucoup de forums donnent un rang aux utilisateurs, qui est très simplement basé sur le nombre de messages postés depuis leur inscription. Un tel classement ne donne pas une bonne estimation de la réputation des utilisateurs. Une discussion avec les modérateurs du site *Vivre sans thyroïde* a confirmé cette intuition. En effet, ils savent que certains utilisateurs de confiance ne postent parfois que quelques messages alors des utilisateurs ayant été bannis peuvent avoir posté de grandes quantités de messages avant leur bannissement. De nombreuses définitions de la confiance et de la réputation existent [Deutsch, 1962, Golbeck, 2009]. Dans nos travaux, nous avons défini la confiance que l'utilisateur *A* donne à un utilisateur *B* comme : "la croyance de *A* dans l'exactitude de l'information donnée par *B*", et la réputation d'un utilisateur *A* comme "l'agrégation des valeurs de confiance données à *A*". La plupart des études sur la confiance dans les réseaux sociaux se basent sur des valeurs comptables (nombre de messages, nombre de réponses, date d'inscription, etc.) et des déclarations explicites de la part des autres membres (nombre de *likes*, nombre de citations, etc.) [Skopik et al., 2009, Wanas et al., 2008]. Or, les fonctionnalités de déclarations explicites sont rarement utilisées dans les forums de santé. Par exemple, dans le forum de santé français *CancerDuSein.org*, nous avons pu observer que seulement 2% des messages correspondent à une déclaration explicite de confiance avec la fonctionnalité "Like".

Nous avons donc décidé d'inférer ces déclarations explicites à partir des textes des messages. Généralement, les messages se présentent sous la forme d'une liste plate chronologique car les utilisateurs utilisent peu les fonctionnalités de type "répondre à" ou "citer une partie de message". Dans une première étape, nous avons élaboré une approche automatique à base de règles définies par observations des fils de discussions, qui reconstruit l'arbre d'une conversation en se basant sur différents indices structurels (e.g. utilisation de pseudos des autres membres de la conversation, @*Martine*, structuration de la conversation sous forme de questions et de réponses) et textuels (e.g. utilisation du tutoiement et du vouvoiement). Après cette étape, nous savons quel message répond à un autre message. Dans une deuxième étape, nous avons recherché des marqueurs de remerciement (e.g. *Merci à toutes !!*) et d'accord (*Je suis complètement d'accord avec toi*) que nous avons considéré comme des déclarations explicites de confiance au même titre que les "Likes" pour calculer un score de réputation au niveau du message. Pour un utilisateur, ce score est agrégé selon la moyenne obtenue pour tous ses messages et par topic car un utilisateur peut être spécialiste d'une thématique et incompetent dans une autre.

Nous avons validé manuellement *a priori* et *a posteriori* l'heuristique générant l'arbre des conversations avec une F-mesure d'environ 0.85 et les annotations des marques de confiance (remerciement et accord/désaccord) avec une F-mesure également d'environ 0.85. Pour évaluer notre score de réputation, nous l'avons comparé au rang du forum et effectivement il y a une corrélation, ce qui n'est pas étonnant car les anciens e-patients sont souvent les plus expérimentés. Nous allons maintenant utiliser ces résultats comme graines que nous allons propager dans le réseau des utilisateurs avec une approche de type *Trustrank* (Chen 2008). Nous travaillons actuellement avec des experts du site *Vivre Sans Thyroïde* pour valider les très mauvais et très bons utilisateurs identifiés avec notre méthode sur leur site. Une perspective consiste à intégrer l'évolution de l'expertise au cours du temps. À l'annonce de sa maladie, un patient peut être non expert mais devenir expert après quelques années.

Sélection de références

A. Abdaoui, J. Azé, S. Bringay, N. Grabar, P. Poncelet : Predicting Medical Roles in Online Health . Proceedings SLSP, Grenoble (France), 2014 : 247-258

4 Quand ?

TABLE 3.5 : Résumé des projets et des encadrements sur la thématique de l'enrichissement des messages - quand ?

Master/Thèse	Projets & Collaborations
T. Opitz (2013-2014) – Post-doctorat co-encadrement C. Lavergne et C. Mollevi	ICM et I3M, QdV cancer
M. Tapi NZali (Depuis 2014) – Thèse co-encadrement C. Lavergne et C. Mollevi	ICM et I3M, QdV cancer

Temporal Supervised Learning. Dernièrement, nous avons commencé à étudier (post-doctorat de T. Opitz et thèse de M. Tapi NZali), les aspects temporels présents dans les messages des patients. Par exemple, il peut être intéressant de replacer un message dans le contexte temporel de la trajectoire du patient. Nous avons exploré ces aspects dans le domaine du cancer du sein, où la trajectoire des patients est bien définie.

La trajectoire d'une patiente atteinte d'un cancer du sein commence par l'annonce. La chirurgie joue ensuite un rôle pivot. Lors de la chirurgie, on enlève la tumeur par une ablation totale ou partielle d'un des deux seins. Parfois d'autres opérations sont nécessaires assez rapidement après l'ablation. Une analyse des tumeurs est faite à l'occasion de la chirurgie. On distingue donc deux périodes : la pré-chirurgie dominée par les préoccupations, l'incertitude sur la gravité de la maladie et la post-chirurgie où les séquelles et l'issue du traitement sont plus concrets, prévisibles et certains. La chirurgie est généralement précédée ou suivie d'une chimiothérapie pour diminuer la taille de la tumeur ou éliminer les restes suite à l'ablation. Le patient est en rémission quand les opérations et les traitements sont réussis. Malheureusement, certains patients rechutent. Une nouvelle tumeur apparaît avec éventuellement des métastases dans d'autres parties du corps. En phase terminale, les patients ne reçoivent plus que des soins palliatifs.

Nous avons donc proposé une méthode qui identifie si un message posté par une femme atteinte d'un cancer du sein a été écrit en période pré-chirurgie, post-chirurgie et rechute. Nous avons pour cela appliqué différents classifieurs de la littérature et examiné le pouvoir prédictif d'un très grand nombre de descripteurs notamment ceux liés au temps et capturés avec l'outil HeidelbergTime : la date, la durée, les jours de la semaine, etc. Nous avons également considéré les N-grammes et des statistiques portant sur la structure lexicale telles que le nombre de mots, d'adjectifs, etc. Nous avons évidemment pris en compte des méta-données sur les messages telles que les catégories de forum, la date du post, etc. et des méta-données sur l'utilisateur, telles que son rang, son sexe, son âge, etc. Une vingtaine d'annotateurs ont

réalisé l'étiquetage des messages avec un accord inter-annotateur élevé (>60%). Finalement, nous avons obtenu une F-mesure d'environ 0.74 avec SVM qui peut facilement être améliorée en tenant compte de l'ordre chronologique entre les messages afin de corriger les mauvaises prédictions. Encore une fois les meilleurs descripteurs isolés sont les Uni-grammes et les Bi-grammes et les meilleurs résultats sont obtenus en combinant l'ensemble des descripteurs.

D Conclusions et perspectives

Dans ce chapitre, nous avons présenté différentes approches permettant de réaliser une analyse descriptive des productions des patients dans les médias sociaux et de les enrichir avec des annotations sémantiques. Nous avons considéré 4 niveaux d'analyse comme décrit sur la Figure 3.6 : le thème, le locuteur, les sentiments et la temporalité. Pour toutes ces analyses, nous avons exploré des approches supervisées et non supervisées et mis en place des chaînes de traitements incluant des étapes classiques en fouille de texte. Pour chaque approche, les contributions sont triples : 1) la formalisation de la tâche, 2) la recherche des meilleurs descripteurs qui sont le plus souvent des N-grammes combinés avec des lexiques spécialisés qu'il faut construire ; 3) la recherche du ou des ou de la meilleure combinaison de classifieur(s) et 5) les expérimentations à partir de données préalablement 4) annotées. Ces travaux ont des limites inhérentes listées ci-dessous.



FIGURE 3.6 : Enrichissement sémantique des messages.

1 Des pré-traitements complexes

Pour l'ensemble des méthodes présentées dans ce chapitre, nous avons dû mettre en place des chaînes de pré-traitements complexes liés à la nature des textes pour capturer la grammaire très imprécise, les fautes d'orthographe et tous les graphiques et représentations iconiques souvent différents d'un site à l'autre dans leur codage. Nous avons eu également besoin de ressources médicales comme des listes de médicaments, d'abréviations. Ce qui est difficile dans ces pré-traitements, c'est qu'ils doivent être adaptés d'un forum à un autre, d'un média à un autre et qu'ils nécessitent beaucoup de temps de développement. Par exemple, la "langue" des personnes âgées dans les forums traitant des maux de dos est très différente de celle des jeunes femmes dans les forums traitant de la maternité.

Une réflexion sur une chaîne de traitement uniformisée et spécifique aux médias sociaux et paramétrable selon les domaines et les types de locuteurs est nécessaire. Il est également important de noter que malgré les nombreuses études menées en traitement automatique de la langue naturelle en France, il y a un manque de ressources pour l'analyse automatique de textes spécifiques comme ceux que nous traitons. Nous avons du traduire la ressource NRC pour pouvoir travailler sur les émotions en français [Bringay et al., 2014]. Il n'existe pas de ressource équivalente à FRAMENET pour l'analyse des cibles et des sources (voir Section 2.3).

2 Vocabulaire patient

Les vocabulaires contrôlés jouent un rôle clé dans des applications biomédicales de fouille de textes. Des vocabulaires de santé normalisés tels que la *SNOMED* (Nomenclature systématisée de médecine), le *MeSH* (Medical Subject Headings), l'*UMLS* (Unified Medical Language System), etc. ont vu le jour. Ces vocabulaires contiennent seulement les termes utilisés par les professionnels de la santé. Depuis 10 ans, des vocabulaires dédiés aux consommateurs de soins (Consumer Health Vocabularies *CHV*), mettant l'accent sur le vocabulaire des patients, ont également été créés [Zeng et al., 2007]. Ces *CHV* lient des mots de tous les jours se rapportant au domaine de la santé à des mots d'argots techniques utilisés par les professionnels de santé (par exemple, le mot "onco" utilisé par les patients à "oncologue" utilisé par les professionnels de la santé). Seuls deux *CHV* sont actuellement disponibles : 1) *MedlinePlus*³⁸, librement disponible, est produit par la *National Library of Medicine* ; 2) *Open and collaborative Consumer Health Vocabulary* (CAO *CHV*)³⁹ est inclus dans l'*UMLS*. À notre connaissance, en français, il n'existe pas de *CHV*.

Nous travaillons actuellement (Thèse de Mike Tapi-Nzali) sur la constitution d'un tel vocabulaire dans le domaine du cancer du sein. Nous avons récupéré un très grand corpus de messages de patients dans les forums et dans le réseau Facebook que nous traitons pour extraire ce vocabulaire. Pour cela, nous avons adapté l'algorithme de [Porter, 1980] pour capturer les abréviations (e.g. *Onco* pour *Oncologue*). Nous utilisons une version du correcteur orthographique *Aspell*⁴⁰ pour identifier les mots médicaux sur lesquels les patients font des fautes d'orthographe fréquentes (e.g. *abcés* pour *abcès*). Nous utilisons aussi des mesures fréquentielles issues des travaux de [Ventura et al., 2014] pour identifier des termes fréquents chez les patients mais non fréquents chez les professionnels de santé. Si ces termes ne sont pas présents dans les ressources médicales comme le *MeSH*, nous les considérons comme des candidats. Nous utilisons alors l'*API WIKIPEDIA* pour les lier à des termes connus et présents dans les ressources médicales. Avec une telle approche, nous sommes par exemple capables de relier *crabe* et *cancer*. Pour valider et typer automatiquement ces nouvelles associations, nous utilisons la ressource lexicale *JeuxDeMots* [Lafourcade et al., 2014] introduite dans la section 2.3, pour identifier les termes reliés via le jeu par les utilisateurs. Les termes restant sont validés manuellement par un oncologue. Nous sommes actuellement en train d'exporter cette ressource au format SKOS (Simple Knowledge Organization System) pour publication dans le portail BioPortal (ANR SFIR⁴¹).

En réduisant l'écart entre les connaissances des patients et des professionnels de la santé, ce type de ressource se révèle crucial pour aider le patient à appréhender sa maladie et à participer au processus de décision médicale. Des relations docteur-patients réussies sont intrinsèquement liées à ce que le patient comprend et croit [Fiscella et al., 2004]. Certains chercheurs ont utilisé ces vocabulaires pour améliorer

38. <http://www.nlm.nih.gov/medlineplus/>

39. <http://www.consumerhealthvocab.org/>

40. <http://aspell.net/>

41. http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2[CODE]=ANR-12-JS02-0010

la lisibilité des documents [Wu et al., 2013] ou la compréhension du dossier patient [Balaji et al., 2013]. [Doing-Harris and Zeng-Treitler, 2011] ont utilisé ce vocabulaire pour coder des données dans une perspective de recherche et d'analyse des données. Dans notre contexte, ce vocabulaire sera essentiellement utilisé pour "corriger" les phrases des messages et ainsi améliorer les résultats de l'annotation sémantique.

3 Ressources annotées

Une limite très forte également à ces travaux est la création de ressources annotées pour l'apprentissage et la validation. Pour chaque expérimentation, nous avons besoin d'experts humains, parfois des professionnels de santé, rarement disponibles pour des tâches très fastidieuses. Certaines tâches d'annotation sont très difficiles, ce qui apparaît lorsque l'on considère les accords inter et intra-annotateurs.

Dans les dernières expérimentations réalisées, nous avons mis en place deux types de validation, *a priori* et *a posteriori*. La validation *a priori* consiste à réaliser le même travail que l'algorithme sans information sur les résultats de celui-ci. La validation *a posteriori* consiste à corriger les résultats de l'algorithme. L'annotation *a priori* donne des accords entre annotateurs moins forts car laissant plus de liberté à l'interprétation de l'annotateur. Elle donne une idée de la difficulté de la tâche que l'on donne au programme qui doit faire au moins aussi bien que les humains. D'un point de vue méthodologique, plusieurs types de calculs d'accord inter-annotateurs existent [Artstein and Poesio, 2008, Fort and Claveau, 2012]. Ces scores d'accord sont généralement utilisés comme mesure d'évaluation de la qualité de la ressource annotée produite [Dandapat et al., 2009]. Il peut être intéressant d'explorer les méthodes d'annotation agiles [Alex et al., 2010, Voormann and GUT, 2008] qui utilisent ces mesures pendant le processus d'annotation, pour assurer la cohérence des annotations et limiter les divergences.

4 Et les motifs ?

Les méthodes à base de motifs comme décrites dans le chapitre 2 sont très pertinentes pour identifier des relations dans les textes. En effet, les approches basées sur des techniques d'apprentissage comme celles que nous venons de décrire, ne sont souvent pas facilement interprétables par les experts contrairement aux approches de traitement automatique de la langue naturelle, comme les approches à base de patrons, qui sont par contre chrono-phages lorsque leur définition est manuelle. Des auteurs comme [Cellier et al., 2010] ont proposé de combiner des approches de fouille de motifs et de traitement automatique de la langue pour générer automatiquement des patrons utilisés pour identifier des relations dans les grands corpus textuels. Par exemple, [Béchet et al., 2014] ont appliqué cette méthode pour la découverte de relations entre gènes et maladies rares. Nous avons commencé à exploiter ce type d'approches pour nos travaux sur l'identification de messages subjectifs (contenant des sentiments voir Section 2.1). Nous nous intéresserons également à l'application de ces méthodes pour la construction automatique de vocabulaire patient en identifiant des structures syntaxiques récurrentes.

5 Visualisations

Comme expliqué dans le Chapitre 2 (Section 2.2), les visualisations en médecine jouent un rôle central et sont très demandées par les professionnels de la santé [Bui and Hsu, 2010, Preim and Bartz, 2007]. L'interactivité que l'on peut ajouter à ces visualisations est également très importante pour faciliter la découverte de connaissances. Les résultats de l'annotation des productions des patients avec les méthodes précédemment décrites ne dérogent pas à cette règle.

La difficulté est de choisir la métaphore visuelle appropriée pour ces nouvelles informations. Selon [Bui and Hsu, 2010], pour chaque élément de connaissance sélectionné, une représentation graphique doit être choisie, pour optimiser la compréhension de l'utilisateur basée sur la raison d'être de la donnée dans l'affichage et la capacité du lecteur à comprendre et utiliser la visualisation. Une piste intéressante consiste à explorer les visualisations dédiées aux textes et présentées par [Kostiantyn and Kerren, 2014]⁴², même si ces dernières ne sont pas spécifiques au domaine de la médecine. Nous devons identifier des visualisations adaptées à notre cas d'étude et qui prennent donc en compte les aspects temporels (e.g. sous forme de timeline) [Aigner et al., 2011], les réseaux d'interaction (e.g. qui poste à qui) [Brandes et al., 2013], les connexions thématiques entre chaque production des patients, etc. Une réflexion plus globale doit être menée en fonction des objectifs des applications et des utilisateurs envisagés pour intégrer ces notions si importantes de visualisation.

Les médias sociaux offrent donc un riche terrain d'études pour les applications médicales d'analyse de données sur lesquelles je vais continuer à travailler dans les années à venir.

42. <http://textvis.lnu.se/>

CHAPITRE

4

CONCLUSION, PERSPECTIVES

Sommaire

- A **Résumé des contributions**
 - B **Perspectives**
-

Dans ce chapitre, je résume les principales contributions et donne pour chacune les publications les plus significatives. Puis, j'introduis les deux principales perspectives au cœur de mes préoccupations actuelles.

A Résumé des contributions

Dans le Chapitre 2, j'ai présenté les différents types de motifs qui ont été définis au cours de mes travaux de recherche et d'encadrement de la recherche. Nous avons appliqué les algorithmes classiques d'extraction de motifs séquentiels sur des données de puces à ADN. L'originalité de ces travaux a été d'utiliser l'ordre entre les expressions des gènes à la place de l'ordre temporel classique [Salle et al., 2009]. Nous avons ensuite étendu ces motifs pour prendre en compte des informations contextuelles qui s'avèrent tout à fait pertinentes pour extraire des trajectoires de patients [Rabatel et al., 2014]. Nous avons également étendu ces motifs pour intégrer des informations spatiales qui permettent de suivre dans le temps et l'espace l'évolution des épidémies de dengue [Alatrística Salas et al., 2012c, Flamand et al., 2014]. Pour chacune de ces approches, nous avons défini des mesures d'intérêt pour la sélection des motifs et des visualisations qui permettent aux experts de s'approprier les résultats des méthodes d'analyse [Alatrística Salas et al., 2012a]. Toutefois, ces méthodes génèrent de très nombreux résultats. Nous avons ainsi travaillé sur une représentation plus condensée des informations contenues dans les motifs via la définition des motifs partiellement ordonnés clos qui ont permis d'évaluer la qualité de l'eau des rivières [Fabrègue et al., 2014]. Ces différents motifs sont efficaces pour des tâches de prédiction comme la prédiction du grade du cancer [Fabrègue et al., 2011] ou la détection d'anomalies [Rabatel et al., 2011].

Dans le Chapitre 3, j'ai présenté les travaux portant sur l'analyse des productions des patients dans les réseaux sociaux. Nous avons développé plusieurs méthodes permettant d'enrichir sémantiquement les messages des patients selon quatre dimensions : temporelle, thématique [Opitz et al., 2014], liée au locuteur et à son rôle dans l'organisation et à la confiance que l'on peut lui accorder [Abdaoui et al., 2014a, Abdaoui et al., 2014b] et à la manière dont ils s'expriment en nous focalisant notamment sur l'analyse de sentiments (polarité et émotion) [Melzi et al., 2014, Abdaoui et al., 2014c, Bringay et al., 2014]. La plupart des méthodes utilisées sont basées sur des lexiques existants ou créés par nos soins et des méthodes d'apprentissage pour capter les éléments diffus dans le discours et les signaux faibles. Ces différentes méthodes aboutissent sur des chaînes de traitements complexes que nous avons validées sur des données annotées. Nous pouvons désormais les combiner pour des applications comme l'étude de la qualité de vie [Garratt et al., 2002] des patients atteint d'un cancer du sein (collaboration avec l'ICM et l'I3M) ou la détection des personnes ayant déjà réalisé une tentative de suicide (collaboration avec le Service des Urgences Psychiatriques du CHU de Montpellier).

B Perspectives

Parmi les axes de recherche discutés dans ce mémoire, certains problèmes m'intéressent plus particulièrement et motivent mes travaux de recherche et d'encadrement de la recherche actuels et à venir. Sans revenir en détail sur les perspectives identifiées dans chaque section, j'aimerais conclure en généralisant ces perspectives selon les 2 axes qui ont guidé l'organisation de ce manuscrit :

– **Axe 1 : Extraction de motifs.** Afin d'initier et de maintenir des processus d'extraction de connaissances au sein des organisations, des plateformes accessibles aux non spécialistes de l'extraction de connaissances, comme les professionnels de santé, doivent être conçues. Ces plateformes peuvent être soit généralistes, soit spécialisées pour des communautés afin de répondre à des besoins spécifiques comme celles décrites dans le chapitre 2 (pour les épidémiologues ou les hydrobiologistes). Évidemment, plus une plateforme est spécialisée pour un métier et des tâches précises, plus elle a des chances d'être acceptée par les utilisateurs et moins elle sera réutilisable. Toutefois, parmi les fonctionnalités génériques à intégrer dans de tels environnements, on trouve :

1. des *méthodes de formalisation du contenu* des grandes bases de données, dont l'extraction de motifs étudiée dans ce manuscrit, pour résumer, condenser et permettre l'appropriation des informations et des connaissances par les non experts ;
2. des *méthodes de partage et de validation de ces connaissances* extraites au sein des communautés, qui sont soumises aux règles des organisations (e.g. droit d'accès, vie privée, responsabilités distribuées, etc). L'analyse réalisée sur les données brutes ne doit pas donner accès pour un utilisateur à des connaissances qu'il n'aurait pas sans cette analyse. Les règles et le fonctionnement de ces communautés étant très différents, ils appellent à de nouvelles méthodes et outils pour assister le cycle de vie des connaissances stockées, extraites, etc. ;
3. des *méthodes de visualisation interactive des connaissances*. Ces visualisations ont la difficile tâche de combler l'écart entre les représentations formelles des connaissances et les objectifs à réaliser par les utilisateurs. Ce problème reste grand ouvert surtout quand les représentations formelles sous-jacentes sont de plus en plus complexes à manipuler par les humains (e.g. séquences, graphes, multi-graphes, etc.). Ainsi, la combinaison étroite des méthodes de visualisations et d'extraction avec les activités sociales de la communauté a un potentiel très important pour améliorer le processus d'exploration collectif des connaissances ;
4. de l'*importante de l'accompagnement des utilisateurs non experts*, lors de l'appropriation de ces nouveaux outils, qui passe par des formations dans lesquelles nous avons déjà commencé à nous investir mais qui demandent encore beaucoup d'efforts de diffusion et de vulgarisation dans les années à venir.

Ces différentes contributions méthodologiques s'inscrivent dans le défi des *Sciences des données* qui combinent de multiples domaines (fouille de données, apprentissage, ingénierie des connaissances, visualisation analytique, statistiques, etc.) dans un contexte pluridisciplinaire.

– **Axe 2 : Analyse des productions des patients.** Ayant pris du recul par rapport aux nombreuses applications de la fouille des médias sociaux pour des applications médicales, comme présentées dans le chapitre 3, je prévois de travailler sur les questions suivantes :

1. *Mise en relation de connaissances avérées des professionnels de la santé et celles co-construites par les patients au travers de la nébuleuse du web social.* C'est un défi identifié dans le cadre de notre travail sur la qualité de vie. Les enquêtes classiques basées sur les questionnaires et les interviews en face-à-face souffrent de biais de réponse. Les patients donnent souvent des réponses évasives sur des sujets délicats comme certains effets secondaires (e.g. diarrhée, problèmes sexuels). Ils oublient des informations entre deux questionnaires, etc. Via le web social, au contraire, les malades parlent de ce qu'ils veulent, quand ils le veulent et d'où ils le veulent. Les données des médias sociaux ont-elles alors une plus-value médicale ? Oui ! Mais en considérant le fait que les messages étudiés ne représentent pas la population générale mais seulement les utilisateurs des médias sociaux. Les informations que nous extrayons de ces médias sont alors complémentaires à celles des études cliniques. Les sorties des méthodes décrites dans le Chapitre 3 seront les points d'entrée pour le calcul d'indicateurs statistiques, permettant de comparer la connaissance avérée et co-construite. Ces indicateurs restent à construire.
2. *Partage de corpus annotés pour une reproductibilité des expérimentations.* Si de nombreuses méthodes ont déjà montré l'intérêt de fouiller les données sociales pour des applications médicales (voir chapitre 3), la plupart ont été validées sur des données non accessibles, à partir d'exemples triés sur le volet, qui ne permettent pas de reproduire les expérimentations (à juste titre car les données sont généralement sensibles). Dans de futurs travaux, j'aimerais utiliser des méthodes de recherche d'informations pour collecter des textes hétérogènes issus du web social et librement accessibles (e.g. en licence Creative Commons⁴³). Ces données alimenteront un corpus de référence (benchmark) sur lequel des méthodes d'analyse pourront être comparées, une fois les données annotées selon la tâche visée, comme discuté dans le Chapitre 3, Section 3.
3. *Fiabilité des connaissances dans le Web Social.* Il est difficile d'empêcher les e-patients de consulter des informations non pertinentes ou non fiables dans les médias sociaux, en revanche il est possible de concevoir des outils pour mettre en évidence des informations de qualité dans ces forums. À partir de l'ensemble des caractéristiques obtenues avec les méthodes développées dans le chapitre 3, il est possible de classifier automatiquement les messages selon leur visée discursive. Par exemple, si un utilisateur recherche de l'information technique (e.g. *comment interpréter mes derniers résultats d'analyse ?*), il faut lui pointer des messages ne contenant pas/peu d'états affectifs, beaucoup de vocabulaire spécifique, postés par des personnes ayant un rang élevé dans le forum (considérés comme expert). Au contraire, si l'e-patient recherche du support moral (e.g. *je me sens triste depuis l'annonce de mon cancer*), il faut lui pointer des fils relatifs à sa thématique et contenant de nombreuses marques d'émotions. Ces approches sont à combiner avec celles sur le rôle et la confiance que l'on peut avoir dans les locuteurs (voir Chapitre 3, Section 3).
4. *Monitoring global de la santé des populations.* Les données issues du web social sont extrêmement variables du point de vue inter et intra individu mais elles peuvent toutefois être

43. Creative Commons (<http://creativecommons.fr/>) propose gratuitement six licences qui permettent aux titulaires de droits d'auteur de mettre leurs œuvres à disposition du public à des conditions prédéfinies. Les licences Creative Commons viennent en complément du droit applicable, elles ne se substituent pas au droit d'auteur. Simples à utiliser et intégrées dans les standards du web, ces autorisations non exclusives permettent aux titulaires de droits d'autoriser le public à effectuer certaines utilisations.

fouillées afin d'obtenir des modèles de prédiction permettant de déclencher efficacement des alertes et prescrire des réponses graduées selon le niveau de ces alertes (e.g. pour les personnes suicidaires). Les défis consistent ici à : 1) déterminer pour chaque application les caractéristiques pertinentes à repérer dans les textes des messages (e.g. capter l'humeur des personnes suicidaires, les facteurs de risque comme l'anorexie, la perte d'estime de soi, etc.) et dans les comportements sur le réseau social (e.g. dépôt de messages la nuit, augmentation frénétique du nombre de messages, etc.) ; 2) produire un modèle de détection robuste sur ces données particulières (hétérogènes, volumineuses, etc.), en combinant différentes méthodes pour capter les changements brusques ou plus insidieux (e.g. votes entre différents classificateurs, détection de concepts drift, active learning pour prendre en compte les réactions des experts en charge du monitoring, etc.). Ici encore, les questions éthiques et juridiques seront centrales (e.g. vie privée, consentement).

Mes futurs travaux de recherche et d'encadrement de la recherche se situent dans le domaine des **Sciences de données** (data science) et je vais continuer à m'intéresser à des méthodes et outils variés qui s'adaptent aux données massives.

Munie de ces nouvelles méthodes et outils et persuadée de leurs intérêts pour des applications dédiées aux professionnels de la santé et aux patients, je me situe désormais dans une perspective à très long terme de **médecine translationnelle**, visant à rapprocher la recherche fondamentale sur les méthodes des sciences de données, de la recherche clinique ayant des applications pratiques.

BIBLIOGRAPHIE

- [Abadaoui et al., 2015] Abadaoui, A., Azé, J., Bringay, S., and Poncelet, P. (2015). Assisting e-patients in an ask the doctor service. In *Studies in Health Technology and Informatics*, page to appear. 49
- [Abdaoui et al., 2014a] Abdaoui, A., Azé, J., Bringay, S., Grabar, N., and Poncelet, P. (2014a). Analysis of forum posts written by patients and health professionals. In *Studies in Health Technology and Informatics*, page 1185. 66
- [Abdaoui et al., 2014b] Abdaoui, A., Azé, J., Bringay, S., Grabar, N., and Poncelet, P. (2014b). Predicting medical roles in online health fora. In *2nd International Conference on Statistical Language and Speech Processing, SLSP 2014, Grenoble, France*, pages 247–258. 58, 66
- [Abdaoui et al., 2014c] Abdaoui, A., Azé, J., Bringay, S., and Poncelet, P. (2014c). Feel : French extended emotional lexicon. *ISLRN : 041-639-484-224-2*. 55, 66
- [Abidine et al., 2013] Abidine, A. Z. E., Sallaberry, A., Bringay, S., Fabrègue, M., Lecellier, C., Phan, N. H., and Poncelet, P. (2013). Co2vis : A visual analytics tool for mining co-expressed and co-regulated genes implied in hiv infections. In *Poster BioVis 2013, Atlanta (USA)*, pages 105–116. 34
- [Accorsi et al., 2014] Accorsi, P., Fabregue, M., Sallaberry, A., Cernesson, F., Lalande, N., Braud, A., Bringay, S., Ber, F. L., Poncelet, P., and Teisseire, M. (2014). Hydroqual : Visual analysis of river water quality. In *Visual Analytics Science and Technology, VAST 2014*, pages 123–132. 35
- [Adomavicius and Tuzhilin, 2005] Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering IEEE Transaction*, 17(6) :734–749. 48
- [Agarwal et al., 2011] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011). Sentiment analysis of twitter data. In *Workshop on Languages in Social Media, LSM '11*, pages 30–38, Stroudsburg, PA, USA. Association for Computational Linguistics. 52
- [Agrawal et al., 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *International Conference on Management of data, SIGMOD '93*, pages 207–216, New York, NY, USA. ACM. 21
- [Agrawal and Srikant, 1995] Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Eleventh International Conference on Data Engineering*, pages 3–14. IEEE Computer Society. 21
- [Aigner et al., 2011] Aigner, W., Miksch, S., Schuman, H., and Tominski, C. (2011). *Visualization of Time-Oriented Data*. Human-Computer Interaction. Springer Verlag, 1st edition. 64

- [Akerkar and Bichile, 2004] Akerkar, S. M. and Bichile, L. S. (2004). Doctor patient relationship : changing dynamics in the information age. *Journal of Postgraduate Medicine*, 50(2) :120–122. 44
- [Alatrística Salas et al., 2012a] Alatrística Salas, H., Azé, J., Bringay, S., Cernesson, F., Flouvat, F., Selmaoui-Folcher, N., and Teisseire, M. (2012a). Recherche de séquences spatio-temporelles peu contredites dans des données hydrologiques. *Mesurer et évaluer la Qualité des Données et des Connaissances*, pages 165–188. 66
- [Alatrística Salas et al., 2012b] Alatrística Salas, H., Bringay, S., Flouvat, F., Selmaoui-Folcher, N., and Teisseire, M. (2012b). The pattern next door : Towards spatio-sequential pattern discovery. In *Advances in Knowledge Discovery and Data Mining*, pages 157–168. Springer. 27
- [Alatrística Salas et al., 2012c] Alatrística Salas, H., Bringay, S., Flouvat, F., Selmaoui-Folcher, N., and Teisseire, M. (2012c). The Pattern Next Door : Towards Spatio-sequential Pattern Discovery. In *Advances in Knowledge Discovery and Data Mining*, volume 7302, pages 157–168. Springer. 28, 66
- [Alec et al., 2009] Alec, G., Lei, H., and Richa, B. (2009). Twitter sentiment classification using distant supervision. *Computer and Information Science*, 150(12) :1–6. 54
- [Alex et al., 2010] Alex, B., Grover, C., Shen, R., and Kabadjov, M. (2010). Agile corpus annotation in practice : An overview of manual and automatic annotation of cvs. In *4th Linguistic Annotation Workshop, LAW IV '10*, pages 29–37, Stroudsburg, PA, USA. Association for Computational Linguistics. 63
- [Anderson and Funnell, 2010] Anderson, R. and Funnell, M. (2010). Patient empowerment : Myths and misconceptions. *Patient Education and Counseling*, 79(3) :277–282. 45
- [Artstein and Poesio, 2008] Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4) :555–596. 63
- [Asproth et al., 1995] Asproth, V., Hakansson, A., and Rèvay, P. (1995). Dynamic information in GIS systems. *Computers, Environment and Urban Systems*, 19(2) :107–115. 27
- [Augustyn et al., 2008] Augustyn, M., Ben Hamou, S., Bloquet, G., Goossens, V., Loiseau, M., and Rynck, F. (2008). Constitution de ressources pédagogiques numériques : le lexique des affects. In *Autour des langues et du langage : perspective pluridisciplinaire*, pages 407–414. Presses Universitaires de Grenoble, Grenoble. 55
- [Azari et al., 2012] Azari, A., Janeja, V., and Mohseni, A. (2012). Predicting hospital length of stay (phlos) : A multi-tiered data mining approach. In *IEEE 13th International Conference on Data Mining Workshops 2012-2013*, pages 287–300. 9
- [Azé, 2003] Azé, J. (2003). Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. In *Extraction et gestion des connaissances, EGC'2003, Revue RIA-ECA numéro spécial EGC*, 17, pages 171–182. 31
- [Baccianella et al., 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. In *International Conference on Language Resources and Evaluation, LREC 2010*, volume 10, pages 2200–2204. 55
- [Baker et al., 1998] Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *International Conference on Computational Linguistics*, pages 86–90. 55
- [Balaji et al., 2013] Balaji, P. R., Houston, T. K., Brandt, C., Fang, H., and Yu, H. (2013). Improving patients' electronic health record comprehension with noteaid. In *MedInfo World Congress on Health and Biomedical Informatics*, volume 192 of *Studies in Health Technology and Informatics*, pages 714–718. IOS Press. 63

- [Barbosa and Feng, 2010] Barbosa, L. and Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In *International Conference on Computational Linguistics, COLING'2010*, pages 36–44. 54
- [Béchet et al., 2014] Béchet, N., Cellier, P., Charnois, T., and Crémilleux, B. (2014). Fouille de motifs séquentiels pour la découverte de relations entre gènes et maladies rares. *Revue d'Intelligence Artificielle*, 28(2-3) :245–270. 63
- [Beresniak et al., 2012] Beresniak, B., Sabatier, P., Achouh, P., Menasché, P., and Fabiani, J. (2012). Cost-effectiveness of mitral valve repair versus replacement by biologic or mechanical prosthesis. *The Annals of thoracic surgery*, 95(1) :98–104. 1
- [Bertin, 1983] Bertin, J. (1983). *Semiology of Graphics*. University of Wisconsin Press. 32
- [Blanchard, 2005] Blanchard, J. (2005). *Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association*. PhD thesis, Thèse de doctorat, Université de Nantes Atlantique. 31
- [Blei et al., 2003] Blei, D., , Ng, M., Andrew, Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3 :993–1022. 13, 51
- [Boiy et al., 2007] Boiy, E., Hens, P., Deschacht, K., and Moens, M. (2007). Automatic sentiment analysis in on-line text. In *11th International Conference on Electronic Publishing*, pages 349–360. Vienna, Austria. 52
- [Boneva et al., 2001] Boneva, B., Kraut, R., and Frohlich, D. (2001). Using email for personal relationships the difference gender makes. In *American Behavioral Scientist*, volume 45(3), pages 530–549. Atlanta, Georgia. 52
- [Brandes et al., 2013] Brandes, U., Freeman, L. C., and Wagner, D. (2013). Social networks. *Handbook of Graph Drawing and Visualization, Roberto Tamassia, Editor, CRC Press, June 24, 2013*. 64
- [Bringay et al., 2014] Bringay, S., Kergosien, E., Pompidor, P., and Poncelet, P. (2014). Identifying the targets of the emotions expressed in health forums. In *Computational Linguistics and Intelligent Text Processing - 15th International Conference, CICLing 2014, Nepal, Part II*, pages 85–97. 53, 56, 62, 66, 86
- [Bringay et al., 2009] Bringay, S., Laurent, A., Orsetti, B., Salle, P., and Teisseire, M. (2009). Handling fuzzy gaps in sequential patterns : Application to health. In *International Conference on Fuzzy Systems, FUZZ-IEEE 2009*, pages 1338–1345. 23
- [Bringay et al., 2010] Bringay, S., M, M. R., Teisseire, M., Poncelet, P., Rassoul, R., Verdier, J., and Devau, G. (2010). Discovering novelty in sequential patterns : application for analysis of microarray data on alzheimer disease. *Studies in Health Technology and Informatics*, 160(Pt 2) :1314–8. 31
- [Bui and Hsu, 2010] Bui, A. and Hsu, W. (2010). Medical data visualization : Toward integrated clinical workstations. *Medical Imaging Informatics*, pages 171–240. 63, 64
- [Cambria et al., 2010] Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010). Senticnet : A publicly available semantic resource for opinion mining. In *Commonsense Knowledge, Papers from the 2010 AAAI Fall Symposium, Arlington, Virginia, USA, November 11-13, 2010*, volume FS-10-02 of AAAI Technical Report. AAAI. 55
- [Carenini et al., 2006] Carenini, G., Ng, R., and Pauls, A. (2006). Multi-document summarization of evaluative text. In *Conference of the European Chapter of the Association for Computational Linguistics, EACL 2006*, pages 305–312. 54
- [Casas-Garriga, 2005] Casas-Garriga, G. (2005). Summarizing sequential data with closed partial orders. In *SIAM International Conference on Data Mining, SDM*. 30

- [Cellier et al., 2010] Cellier, P., Charnois, T., Plantevit, M., and Crémilleux, B. (2010). Recursive sequence mining to discover named entity relations. In Cohen, P. R., Adams, N. M., and Berthold, M. R., editors, *9th International Symposium Advances in Intelligent Data Analysis, IDA 2010, Tucson, AZ, USA*, volume 6065 of *Lecture Notes in Computer Science*, pages 30–41. Springer. 63
- [Cerini et al., 2007] Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. (2007). Language resources and linguistic theory : Typology, second language acquisition, english linguistics. A. Sansò, Ed. Milano, IT. 55
- [Charlet, 2002] Charlet, J. (2002). *L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales*. Habilitation à diriger des recherches. 7, 85
- [Charlet et al., 2014] Charlet, J., Mazuel, L., Declerck, G., Miroux, P., and Gayet, P. (2014). Describing localized diseases in medical ontology : an fma-based algorithm. *Studies in Health Technology and Informatics*, 205 :858–62. 15
- [Chew and Eysenbach, 2010] Chew, C. and Eysenbach, G. (2010). Pandemics in the age of twitter : Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS One*, 5(11). 46
- [Choudhury et al., 2013] Choudhury, M. D., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *7th International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA*. 46
- [Chunara et al., 2012] Chunara, R., Andrews, J., and Brownstein, J. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene*, 86(1) :39–45. 45
- [Civan and Pratt, 2007] Civan, A. and Pratt, W. (2007). Threading together patient expertise. In *AMIA American Medical Informatics Association Annual Symposium*, pages 140–144. New York, NY, USA. 45
- [Collier et al., 2011] Collier, N., Son, N., and Nguyen, N. (2011). Omg u got flu ? analysis of shared health messages for bio-surveillance. *Journal of Biomedical Semantics*, 2(5) :S9. 45
- [Culotta, 2010] Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. In *First Work-shop on Social Media Analytics*, pages 115–122. 45
- [Dandapat et al., 2009] Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009). Complex linguistic annotation - no easy way out ! a case from bengali and hindi pos labeling tasks. In *3rd Linguistic Annotation Workshop*, pages 10–18. Association for Computational Linguistics. 63
- [Davidov et al., 2010] Davidov, D., Tsur, O., and Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. In *International Conference on Computational Linguistics, COLING'2010*, pages 241–249. 54
- [Debruyne et al., 2015] Debruyne, P., Johnson, P., Pottel, L., Daniels, S., Greer, R., Hodgkinson, E., Kelly, S., Lycke, M., Samol, J., Mason, J., Kimber, D., Loucaides, E., Parmar, M., and Harvey, S. (2015). Optimisation of pharmacy content in clinical cancer research protocols : Experience of the united kingdom chemotherapy and pharmacy advisory service. *Clinical Trials*. 8
- [Deutsch, 1962] Deutsch, M. (1962). Cooperation and trust : Some theoretical notes. In *Nebraska Symposium on Motivation*, pages 275–320. Oxford, England : Univer. Nebraska Press. 59
- [Doing-Harris and Zeng-Treitler, 2011] Doing-Harris, M. K. and Zeng-Treitler, Q. (2011). Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of Medical Internet Research*, 13(2) :e37. 63
- [Domingo-Ferrer, 2008] Domingo-Ferrer, J. A. (2008). Survey of inference control methods for privacy-preserving data mining. Aggarwal CC, Yu PS, editors. *Privacy-preserving data mining*. Springer US, pages 53–80. 17

- [Dursun, 2014] Dursun, D. (2014). *Real-World Data Mining : Applied Business Analytics and Decision Making*. Published Dec 23, 2014 by Pearson FT Press. Part of the FT Press Analytics series. 6, 10
- [Ekman, 1999] Ekman, P. (1999). Basic emotions. *T. Dalgleish and T. Power (Eds.) The Handbook of Cognition and Emotion*, pages 45–60. 52
- [El-Sappagh et al., 2013] El-Sappagh, S. H., El-Masri, S., Riad, A. M., and Elmogy, M. (2013). Data mining and knowledge discovery : Applications, techniques, challenges and process models in healthcare. *International Journal of Engineering Research and Applications (IJERA)*, 3(3) :900–906. 6, 8, 15
- [Elston and Ellis, 1991] Elston, C. and Ellis, I. (1991). The value of histological grade in breast cancer : experience from a large study with long-term follow-up. *Histopathology*, 19 :403–10. 39
- [Esuli and Sebastiani, 2006] Esuli, A. and Sebastiani, F. (2006). Sentiwordnet : A publicly available lexical resource for opinion mining. In *5th Conference on Language Resources and Evaluation, LREC 2016*, pages 417–422. 52
- [Fabrègue et al., 2012] Fabrègue, M., Braud, A., Bringay, S., Le Ber, F., and Teisseire, M. (2012). Including spatial relations and scales within sequential pattern extraction. In *Discovery Science - 15th International Conference, DS 2012, Lyon, France*, pages 209–223. 41
- [Fabrègue et al., 2014] Fabrègue, M., Braud, S., Bringay, S., Grac, C., Ber, F. L., Levet, D., and Teisseire, M. (2014). Discriminant temporal patterns for linking physico-chemistry and biology in hydroecosystem assessment. *Ecological Informatics*, 24 :210–221. 66
- [Fabrègue et al., 2011] Fabrègue, M., Bringay, S., Poncelet, P., Teisseire, M., and Orsetti, B. (2011). Mining microarray data to predict the histological grade of a breast cancer. *Journal of Biomedical Informatics*, pages 12–16. 39, 66
- [Fan et al., 2011] Fan, W., Sun, S., and Song, G. (2011). Sentiment classification for chinese netnews comments based on multiple classifiers integration. In *International Joint Conference on Computer Sciences and Optimization*, pages 829–834. 54
- [Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). Data mining to knowledge discovery : an overview. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, *Advances in knowledge discovery and data mining*, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA. 10, 11, 85
- [Fiot et al., 2007] Fiot, C., Laurent, A., and Teisseire, M. (2007). Extended time constraints for sequence mining. *14th International Symposium on Temporal Representation and Reasoning, TIME 2007*, pages 105–116. 23
- [Fiscella et al., 2004] Fiscella, K., Meldrum, S., Franks, P., Shields, C., Duberstein, P., McDaniel, S., and al. (2004). Patient trust : is it related to patient-centered behavior of primary care physicians ? *Medical Care*, 42 :1048–1055. 62
- [Flamand et al., 2014] Flamand, C., Fabregue, M., Bringay, S., Ardillon, V., Quénel, P., Desenclos, J., and Teisseire, M. (2014). Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in french guiana. *Journal of American Medical Informatics Association*, 21(e2) :232–240. 28, 66
- [Forestier et al., 2012] Forestier, M., Stavrianou, A., Velcin, J., and Zighed, D. (2012). Roles in social networks : Methodologies and research issues. *Web Intelligence and Agent Systems*, 10 :117–133. 57
- [Fort and Claveau, 2012] Fort, K. and Claveau, V. (2012). Annotation manuelle de matchs de foot : Oh la la la ! l'accord inter-annotateurs ! et c'est le but. In *Traitement Automatique des Langues Naturelles, TALN 2012*, volume 2, pages 1–14. Grenoble, 4 au 8 juin 2012. 63

- [Francisco and Gervás, 2006] Francisco, V. and Gervás, P. (2006). Automated mark up of affective information in english texts. In *9th international conference on Text, Speech and Dialogue*, pages 375–382. Berlin, Heidelberg. 52
- [Garratt et al., 2002] Garratt, A., Schmidt, L., Mackintosh, A., and Fitzpatrick, R. (2002). Quality of life measurement : bibliographic study of patient assessed health outcome measures. *British Medical Journal*, 324(7351) :1417. 66
- [Golbeck, 2009] Golbeck, J. (2009). Trust and nuanced profile similarity in online social networks. *ACM Transactions on the Web, TWEB 2009*, 3(4) :12 :1–12 :33. 59
- [Harb et al., 2008] Harb, A., Plantié, M., Dray, G., Roche, M., Troussel, F., and Poncelet, P. (2008). Web opinion mining : How to extract opinions from blogs ? categories and subject descriptors. In *International conference on Soft Computing as Transdisciplinary Science and Technology*, pages 211–217. 55
- [Hawn, 2009] Hawn, C. (2009). Take two aspirin and tweet me in the morning : how twitter, facebook, and other social media are reshaping health care. *Health affairs*, 28(2) :361–368. 45
- [Himmel et al., 2009] Himmel, W., Reincke, U., and Michelman, H. W. (2009). Text mining and natural language processing approaches for automatic categorization of lay requests to web-based expert forums. *Journal Medical Internet Research*, 11(3) :1–1. 48
- [Hou and Zhang, 2008] Hou, S. and Zhang, X. (2008). Alarms association rules based on sequential pattern mining algorithm. In Ma, J., Yin, Y., Yu, J., and Zhou, S., editors, *5th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008*, pages 556–560. IEEE Computer Society. 9
- [Huh et al., 2013] Huh, J., Yetisgen-Yildiz, M., and Pratt, W. (2013). Text classification for assisting moderators in online health communities. *Journal of Biomedical Informatics*, 46(6) :998–1005. 44
- [Inokuchi et al., 2000] Inokuchi, A., Washio, T., and Motoda, H. (2000). An apriori-based algorithm for mining frequent substructures from graph data. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD'00*, pages 13–23. 21
- [Jackiewicz et al., 2010] Jackiewicz, A., Hunston, S., and El-Bèze, M. (2010). Introduction. *Traitement Automatique des Langues*, 51(3) :7–17. 52
- [Jackin et al., 2014] Jackin, B., Miyata, H., Ohkawa, T., and al (2014). Distributed calculation method for large-pixel-number holograms by decomposition of object and hologram planes. *Optics Letters*, 39(24) :6867–70. 15
- [Jacob and Ramani, 2012] Jacob, S. G. and Ramani, R. G. (2012). Article : Data mining in clinical data sets : A review. *International Journal of Applied Information Systems*, 4(6) :15–26. Published by Foundation of Computer Science, New York, USA. 9
- [Jacquemin and Royauté, 1994] Jacquemin, C. and Royauté, J. (1994). Retrieving terms and their variants in a lexicalised unification-based framework. In Croft, W. B. and van Rijsbergen, C. J., editors, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 132–141. ACM/Springer. 49
- [Jain et al., 1999] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering : A review. *ACM Computer Survey*, 31(3) :264–323. 9
- [Jha and Elhadad, 2010] Jha, M. and Elhadad, N. (2010). Cancer stage prediction based on patient online discourse. In *Workshop on Biomedical Natural Language Processing*, pages 64–71. Association for Computational Linguistics. 45

- [Joudaki et al., 2014] Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., and Arab, M. (2014). Using data mining to detect health care fraud and abuse : a review of literature. *Global Journal Health Science*, 7(1) :194–202. 8
- [Kambeitz et al., 2015] Kambeitz, J., Kambeitz-Illankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., Falkai, P., and Koutsouleris, N. (2015). Detecting neuroimaging biomarkers for schizophrenia : A meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology*. 42
- [Kautz, 2013] Kautz, H. (2013). Data mining social media for public health applications. In *Senior Member Presentation : 23rd International Joint Conference on Artificial Intelligence, IJCAI 2013*. Beijing, China. 45
- [Keim et al., 2008] Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G. (2008). Visual analytics : Definition, process, and challenges. In *Information Visualization*, pages 154–175. Springer-Verlag, Berlin, Heidelberg. 33, 37
- [Kermisch, 2011] Kermisch, C. (2011). *Le concept de risque. De l'épistémologie à l'éthique*. Paris, Lavoisier. 51
- [Khajehei and Etemady, 2010] Khajehei, M. and Etemady, F. (2010). Data mining and medical research studies. In *IEEE Second International Conference on Computational Intelligence, Modeling and Simulation, CIMSIM 2010*, pages 119–122. 6
- [Kharat et al., 2014] Kharat, A., Singh, A., Kulkarni, V., and Shah, D. (2014). Detecting neuroimaging biomarkers for schizophrenia : A meta-analysis of multivariate pattern recognition studies. *Indian Journal of Radiology Imaging*, 24(2) :97–102. 42
- [Kittler et al., 1998] Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3) :226–239. 48
- [Koh and Tan, 2005] Koh, H. C. and Tan, G. (2005). Data mining applications in healthcare. *Journal of healthcare information management*, 19(2) :64–72. 8, 15
- [Kostiantyn and Kerren, 2014] Kostiantyn, K. and Kerren, A. (2014). Text visualization browser : A visual survey of text visualization techniques. In *IEEE Information Visualization Conference (InfoVis 2014), Poster Abstracts of IEEE VisWeek 2014*. 64
- [Kozareva et al., 2007] Kozareva, Z., Navarro, B., Vazquez, S., and Montoyo, A. (2007). Uazbsa : a headline emotion classification through web information. In *4th International Workshop on Semantic Evaluations*, pages 334–337. Stroudsburg, PA, USA. 55
- [Kriek et al., 2011] Kriek, M., Dreesman, J., Otrusina, L., and Denecke, K. (2011). A new age of public health : Identifying disease outbreaks by analyzing tweets. In *Health WebScience Workshop, ACM Web Science Conference*. 45
- [Lafourcade et al., 2014] Lafourcade, M., Zarrouk, M., and Joubert, A. (2014). About inferences in a crowdsourced lexical-semantic network. In *14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 174–182. 62
- [Lamos and Christianini, 2010] Lamos, V. and Christianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In *Workshop on Cognitive Information Processing*, pages 411–416. 46
- [Lenca et al., 2003] Lenca, P., Meyer, P., Vaillant, B., Picouet, P., and Lallich, S. (2003). Evaluation et analyse multicritère des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information*, pages 220–246. 31
- [Lerman and Azé, 2007] Lerman, I. and Azé, J. (2007). *A New Probabilistic Measure of Interestingness for Association Rules, Based on the Likelihood of the Link*, pages 207–236. Springer. 9

- [Lin et al., 2012] Lin, C., He, Y., Everson, R., and Ruger, S. (2012). Weakly supervised joint sentiment-topic detection from text. *Knowledge and Data Engineering, IEEE Transactions*, 24(6) :1134–1145. 51
- [Liu, 2012] Liu, B. (2012). Sentiment analysis and opinion mining : Synthesis lectures on human language technologies. *Morgan and Claypool Publishers*, page 162. 52
- [Liu et al., 2005] Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer : Analyzing and comparing opinions on the web. In *International Conference on World Wide Web, WWW 2005*, pages 342–351. 54
- [Liu et al., 2013] Liu, N.-H., Chiang, C.-Y., and Hsu, H.-M. (2013). Improving driver alertness through music selection using a mobile eeg to detect brainwaves. *Sensors (Basel)*, 13(7) :8199–8221. 9
- [Lu et al., 2006] Lu, C., Hong, J., and Cruzlara, S. (2006). Emotion detection in textual information by semantic role labeling and web mining techniques. In *3rd Taiwanese French Conference on Information Technology, TFIT 2006*. Nancy, France. 52
- [Mannila et al., 1997] Mannila, H., Toivonen, H., and Verkamo, A. (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3) :259–289. 21
- [Martínezemail et al., 2013] Martínezemail, S., Sánchezemail, D., and Valls, A. (2013). A semantic framework to protect the privacy of electronic health records with non-numerical attributes. *Journal of Biomedical Informatics*, 46(2) :294–303. 17
- [Masseglia et al., 1998] Masseglia, F., Cathala, F., and Poncelet, P. (1998). The PSP approach for mining sequential patterns. *Principles of Data Mining and Knowledge Discovery*, pages 176–184. 21
- [McCallum et al., 2007] McCallum, A., Wang, X., and Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30 :249–272. 57
- [Meier et al., 2007] Meier, A., Lyons, E., Frydman, G., Forlenza, M., and K Rimer, B. (2007). How cancer survivors provide support on cancer-related internet mailing lists. *Journal of Medical Internet Research*, 9(2) :e12. 45
- [Melzi et al., 2014] Melzi, A., Abdaoui, A., Azé, J., Bringay, S., Poncelet, P., and Galtier, F. (2014). Patient's rationale : Patient knowledge retrieval from health forums. In *6th International Conference on eHealth, Telemedicine, and Social Medicine, ETELEMED 2014*. Barcelona, Spain. 54, 66
- [Merton, 1957] Merton, R. (1957). Social theory and social structure. *Free Press, New York, NY, US*, page 713. 57
- [Milley, 2000] Milley, A. (2000). Healthcare and data mining. *Health Management Technology*, 21(8) :44–47. 8
- [Mohammad, 2010] Mohammad, S. (2010). Emotions evoked by commonwords and phrases : Using mechanical turk to create an emotion lexicon. In *Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34. Stroudsburg, PA, USA. 54, 55
- [Moher et al., 2009] Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. (2009). Preferred reporting items for systematic reviews and meta-analyses : The prisma statement. *PLoS Med*, 6(7) :e1000097. 13
- [Mulder et al., 2004] Mulder, M., Nijholt, A., Uyl, M. D., and Terpstra, P. (2004). A lexical grammatical implementation of affect. In *7th International Conference on Text, Speech and Dialogue*, pages 171–177. Heidelberg. 52
- [Nadi and Delavar, 2003] Nadi, S. and Delavar, M. (2003). Spatio-temporal modeling of dynamic phenomena in GIS. In *9th Scandinavian Research Conference on Geographical Information Science, ScanGIS 2003*, pages 215–225. 27

- [Neviarouskaya et al., 2011] Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2011). Affect analysis model : Novel rule-based approach to affect sensing from text. *International Journal of Natural Language Engineering*, 17(1) :95–135. 55
- [Opitz et al., 2014] Opitz, T., Azé, J., Bringay, S., Joutard, C., Lavergne, C., and Mollevi, C. (2014). Breast cancer and quality of life : Medical information extraction from health forums. In *Studies in Health Technology and Informatics*, pages 1070–1074. 50, 66
- [Pang, 2002] Pang, B. (2002). Thumbs up ? : Sentiment classification using machine learning techniques. In *Conference on Empirical Methods on Natural Language Processing, EMNLP 2002*, volume 10, pages 79–86. 52
- [Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2) :1–135. 52
- [Parès et al., 2014] Parès, Y., Aimé, X., Charlet, J., and Jaulent, M. (2014). Towards an automatic harmonization of the representation of medical reports to assess their similarities. *Studies in Health Technology and Informatics*, 205 :858–62. 15
- [Pasquier et al., 2013] Pasquier, C., Sanhes, J., Flouvat, F., and Selmaoui-Folcher, N. (2013). Frequent pattern mining in attributed trees. In *Advances in Knowledge Discovery and Data Mining, PAKDD'13*, pages 26–37. Springer Berlin Heidelberg. 41
- [Pasquier et al., 1999] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999). Discovering frequent closed itemsets for association rules. In *7th International Conference on Database Theory, ICDT '99*, pages 398–416. Springer. 29
- [Paul, 2011] Paul, M. (2011). You are what you tweet : Analyzing twitter for public health. In *International Conference on Weblogs and Social Media, ICWSM 2011*. 45
- [Pearl and Steyvers, 2010] Pearl, M. and Steyvers, L. (2010). Identifying emotions, intentions, and attitudes in text using a game with a purpose. In *Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 71–79. USA. 52
- [Pei et al., 2006] Pei, J., Wang, H., Liu, J., Wang, K., Wang, J., and Yu, P. S. (2006). Discovering frequent closed partial orders from strings. *IEEE Transactions on Knowledge and Data Engineering*, 18(11) :2006. 30
- [Pestian et al., 2012] Pestian, J. P., Matykiewicz, P., LinnGust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K. B., Hurdle, J., Brew, C., et al. (2012). Sentiment analysis of suicide notes : A shared task. *Biomedical Informatics Insights*, 5(1) :3–16. 52
- [Peuquet, 1994] Peuquet, D. (1994). It's about time : A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 84(3) :441–461. 32
- [Pitarch et al., 2010] Pitarch, Y., Laurent, A., and Poncelet, P. (2010). Summarizing multidimensional data streams : A hierarchy-graph-based approach. In *14th Pacific-Asia Conference on Knowledge Discovery and Mining, PAKDD 2010*. Hyderabad, India. 16, 42
- [Plantié et al., 2008] Plantié, M., Roche, M., Dray, G., and Poncelet, P. (2008). Is a voting approach accurate for opinion mining ? In *Data Warehousing and Knowledge Discovery, 10th International Conference, DaWaK 2008, Turin, Italy, September 2-5*, pages 413–422. 54
- [Plutchik, 1980] Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Emotion : Theory, research, and experience*, 1 :3–33. 52
- [Popescu and Etzioni, 2005] Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346. Computational Linguistics Association. 55

- [Porter, 1980] Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3) :130–137. 62
- [Preim and Bartz, 2007] Preim, B. and Bartz, D. (2007). *Visualization in Medicine : Theory, Algorithms, and Applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 63
- [Rabatel et al., 2010] Rabatel, J., Bringay, S., and Poncelet, P. (2010). Aide à la décision pour la maintenance ferroviaire préventive. In *Extraction et Gestion des Connaissances, EGC'10*, Revue des Nouvelles Technologies de l'Information, pages 363–368. Cépaduès-Éditions. 26
- [Rabatel et al., 2011] Rabatel, J., Bringay, S., and Poncelet, P. (2011). Anomaly detection in monitoring sensor data for preventive maintenance. *Expert System Application*, 38(6) :7003–7015. 66
- [Rabatel et al., 2014] Rabatel, J., Bringay, S., and Poncelet, P. (2014). Mining representative frequent patterns in a hierarchy of contexts. In *Advances in Intelligent Data Analysis XIII - 13th International Symposium, IDA 2014, Leuven, Belgium*, pages 239–250. 66
- [Roberts et al., 2012] Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., and Harabagiu, S. M. (2012). Empatweet : Annotating and detecting emotions on twitter. In *8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 23–25. Istanbul, Turkey. 52
- [Roche and Prince, 2009] Roche, M. and Prince, V. (2009). A web-mining approach to disambiguate biomedical acronym expansions. *Informatica*, 34(2) :243–253. 52
- [Ruan, 2011] Ruan, L. (2011). Meaningful signs - emoticons. *Theory and Practice in Language Studies*, 1(1) :91–94. 54
- [Sadilek and Kautz, 2013] Sadilek, A. and Kautz, H. (2013). Modeling the impact of lifestyle on health at scale. In *Sixth ACM International Conference on Web Search and Data Mining*, pages 637–646. 45
- [Sadilek et al., 2012a] Sadilek, A., Kautz, H., and Silenzio, V. (2012a). Modeling spread of disease from social interactions. In *6th AAAI International Conference on Weblogs and Social Media, ICWSM 2012*. 45
- [Sadilek et al., 2012b] Sadilek, A., Kautz, H., and Silenzio, V. (2012b). Predicting disease transmission from geo-tagged micro-blog data. In *26th AAAI Conference on Artificial Intelligence*. 45
- [Sallaberry et al., 2011] Sallaberry, A., Pecheur, N., Bringay, S., Roche, M., and Teisseire, M. (2011). Sequential patterns mining and gene sequence visualization to discover novelty from microarray data. *Journal of Biomedical Informatics*, 44(5) :760–774. 33
- [Salle et al., 2009] Salle, P., Bringay, S., and Teisseire, M. (2009). Mining discriminant sequential patterns for aging brain. In *Artificial Intelligence in Medicine*, volume 5651 of *Lecture Notes in Computer Science*, pages 365–369. Springer Berlin / Heidelberg. 22, 66
- [Sanhes et al., 2013] Sanhes, J., Flouvat, F., Pasquier, C., Selmaoui-Folcher, N., and Boulicaut, J. (2013). Extraction de motifs condensés dans un unique graphe orienté acyclique attribué. In *Extraction et Gestion des Connaissances, RNTI E-24 EGC'13*, pages 205–216. Hermann-Éditions. 41
- [Scarff and Torloni, 1968] Scarff, R. and Torloni, H. (1968). Histological typing of breast tumors. *International histological classification of tumours*, 2(2) :13–20. 39
- [Schwartz and Sprinzen, 1984] Schwartz, J. and Sprinzen, M. (1984). Structures of connectivity. *Social Networks*, 6 :103–140. 57
- [Shekhar and Huang, 2001] Shekhar, S. and Huang, Y. (2001). Discovering spatial co-location patterns : A summary of results. In Jensen, C. S., Schneider, M., Seeger, B., and Tsotras, V. J., editors, *SSTD*, volume 2121 of *Lecture Notes in Computer Science*, pages 236–256. Springer. 35
- [Simoes et al., 2013] Simoes, P. D. A., Martins, P., Casagrande, R., and al (2013). Using a model of parallel distributed processing associated with data mining in the characterization of sexuality in a university population. *Studies in Health Technology and Informatics*, page 192 :1135. 15

- [Skopik et al., 2009] Skopik, F., Truong, H.-L., and Dustdar, S. (2009). Trust and reputation mining in professional virtual communities. *Web Engineering*, pages 76–90. 59
- [Song et al., 2013] Song, J., Ju-Hong, L., Joon-Hyuk, C., and Seok-Ju, C. (2013). Automatic differential diagnosis of pancreatic serous and mucinous cystadenomas based on morphological features. *Computers in biology and medicine*, 43 :1–15. 1
- [Sotiriou et al., 2003] Sotiriou, C., SY, N., McShane, L., Korn, E., Long, P., Jazaeri, A., Martiat, P., Fox, S., Harris, A., and ET, L. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *National Academy of Sciences*, 100(18) :10393–8. 39
- [Stam et al., 1995] Stam, C., Jelles, B., Achtereekte, H., Rombouts, S., Slaets, J., and Keunen, R. (1995). Investigation of eeg non-linearity in dementia and parkinson’s disease. *Electroencephalography and Clinical Neurophysiology*, 95(5) :309–17. 9
- [Stone and Hunt, 1963] Stone, P. J. and Hunt, E. B. (1963). A computer approach to content analysis : Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference, AFIPS ’63 (Spring)*, pages 241–256, New York, NY, USA. ACM. 54
- [Strapparava and Mihalcea, 2008] Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *Symposium on Applied Computing, SAC 2008*, pages 1556–1560. New York, NY, USA. 52
- [Strapparava and Valitutti, 2004] Strapparava, C. and Valitutti, A. (2004). Wordnet affect : an affective extension of wordnet. In *4th International Conference on Language Resources and Evaluation*, pages 1083–1086. Lisbon. 55
- [Subasic and Berendt, 2008] Subasic, I. and Berendt, B. (2008). Web mining for understanding stories through graph visualisation. In *8th IEEE International Conference on Data Mining, ICDM ’08*, pages 570–579, Washington, DC, USA. IEEE Computer Society. 33
- [Tausczik and Pennebaker, 2010] Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words : Liwc and computerized text analysis methods. *Journal of Language and Social Psychology* J LANG SOC PSYCHOL, 29(1) :24–54. 55
- [Termier et al., 2002] Termier, A., Rousset, M., and Sebag, M. (2002). Treefinder : a first step towards xml data mining. In *IEEE International Conference on Data Mining, ICDM 2002*, pages 450–457. 21
- [Thoumelin and Grabar, 2014] Thoumelin, P. C. and Grabar, N. (2014). La subjectivité dans le discours médical : sur les traces de l’incertitude et des émotions. In Reynaud, C., Martin, A., and Quiniou, R., editors, *14èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2014, Rennes, France, 28-32 Janvier, 2014*, volume E-26 of *Revue des Nouvelles Technologies de l’Information*, pages 455–466. Hermann-Éditions. 51
- [Tsoukatos and Gunopulos, 2001] Tsoukatos, I. and Gunopulos, D. (2001). Efficient Mining of Spatio-temporal Patterns. In *7th International Symposium on Advances in Spatial and Temporal Databases, SSTD ’01*, pages 425–442. Springer. 35
- [Tufté, 1983] Tufté, E. (1983). *The visual display of quantitative information*. Number 914 in *The Visual Display of Quantitative Information*. Graphics Press. 32, 33
- [Turney, 2002] Turney, P. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *40th Annual Meeting on Association for Computational Linguistics, ACL 2002*, pages 417–424. 52
- [Ventura et al., 2014] Ventura, J. A. L., Jonquet, C., Roche, M., and Teisseire, M. (2014). Towards a mixed approach to extract biomedical terms from text corpus. *International Journal of Knowledge Discovery in Bioinformatics IJKDB*, 4(1) :1–15. 62

- [Voormann and GUT, 2008] Voormann, H. and GUT, U. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, 4(2) :235–251. 63
- [Wald et al., 2007] Wald, H. S., Dube, C. E., and Anthony, D. C. (2007). Untangling the web—the impact of internet use on health care and the physician-patient relationship. *Patient Education and Counseling*, 68(3) :218–224. 59
- [Wanas et al., 2008] Wanas, N., El-Saban, M., Ashour, H., and Ammar, W. (2008). Automatic scoring of online discussion posts. In *2nd ACM Workshop on Information Credibility on the Web*, pages 19–26. New York, NY, USA. 59
- [Ward et al., 2010] Ward, M., Grinstein, G., and Keim, D. (2010). *Interactive Data Visualization : Foundations, Techniques, and Applications*. A. K. Peters, Ltd., Natick, MA, USA. 32
- [Ware, 2004] Ware, C. (2004). *Information Visualization : Perception for Design*. Interactive Technologies. Elsevier Science. 33
- [Wasan et al., 2006] Wasan, S. K., Bhatnagar, V., and Kaur, H. (2006). The impact of datamining techniques on medical diagnostics. *Data Science Journal*, 5 :119–126. 6
- [Welser et al., 2007] Welser, H., Gleave, E., Fisher, D., and Wiebe Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *The Journal of Social Structure*, 8 :564–586. 57
- [Whissell, 1989] Whissell, C. (1989). The dictionary of affect in language. *Academic Press*. 55
- [White et al., 1976] White, H., Boorman, S., and Breiger, R. (1976). Social structure from multiple networks. blockmodels of roles and positions. *American Journal of Sociology*, 81 :730–780. 57
- [Wiebe and Riloff, 2011] Wiebe, J. and Riloff, E. (2011). Finding mutual benefit between subjectivity analysis and information extraction. *IEEE Transactions on Affective Computing*, 2(4) :175–191. 54
- [Wiebe et al., 2005] Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3) :165–210. 52
- [Wiegand and Klakow, 2010] Wiegand, M. and Klakow, D. (2010). Bootstrapping supervised machine-learning polarity classifiers with rule-based classification. In *1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA 2010*, pages 59–66. 52
- [Wilson et al., 2005] Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, pages 347–354. 55
- [Wolfe and Jensen, 2004] Wolfe, A. and Jensen, D. (2004). Playing multiple roles : Discovering overlapping roles in social networks. In *Workshop on Statistical Relational Learning and Its Connections to Other Fields, ICML-04*. 57
- [Wong et al., 2000] Wong, P. C., Cowley, W., Foote, H., Jurrus, E., and Thomas, J. (2000). Visualizing sequential patterns for text mining. In *IEEE Symposium on Information Visualization, InfoVis'00*, pages 105–111. IEEE. 33
- [Wu et al., 2013] Wu, D., Hanauer, D., Mei, Q., Clark, P., An, L., Lei, J., Proulx, J., Zeng-Treitler, Q., and Zheng, K. (2013). Applying multiple methods to assess the readability of a large corpus of medical documents. *Studies in Health Technology and Informatics*, 192 :647–51. 63
- [Wu et al., 2014] Wu, H., Lin, S., and Liu, C. (2014). Analyzing patients' values by applying cluster analysis and Irfm model in a pediatric dental clinic in taiwan. *Scientific World Journal*. 8
- [Yan et al., 2003] Yan, X., Han, J., and Afshar, A. (2003). Clospan : Mining closed sequential patterns in large datasets. In *International Conference on Data Mining, SDM'03*, pages 166–177. 30

- [Yang and Parthasarathy, 2006] Yang, H. and Parthasarathy, S. (2006). Mining spatial and spatio-temporal patterns in scientific data. In *22nd International Conference on Data Engineering Workshops, ICDEW '06*, pages 146–, Washington, DC, USA. IEEE Computer Society. 15
- [Yang and Liu, 1999] Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *22Nd Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR 1999*, pages 42–49. New York, NY, USA. 48
- [Yuan, 2009] Yuan, M. (2009). Knowledge toward discovery about geographic dynamics in spatio-temporal databases. In *Geographic Data Mining and Knowledge Discovery, Second Edition*, pages 347–365. Edited by Harvey J. Miller and Jiawei Han. 27
- [Zaki, 2002] Zaki, M. (2002). Efficiently mining frequent trees in a forest. In *8th ACM Conference on Knowledge Discovery and Data Mining, SIGKDD 2002*, pages 71–80. ACM press. 21
- [Zeng et al., 2007] Zeng, Q., Tse, T., G, D., and al. (2007). Term identification methods for consumer health vocabulary development. *Journal Medical Internet Research*, 9(4). 62
- [Zhang et al., 2007] Zhang, J., Ackerman, M., and Adamic, L. (2007). Expertise networks in online communities : Structure and algorithms. In *16th International Conference on World Wide Web, WWW '07.*, pages 221–230. ACM, New York, NY, USA. 57

TABLE DES ILLUSTRATIONS

Liste des figures

1	Deux thématiques de recherche développées depuis 2007 avec les principaux encadrements et collaborations.	1
1.1	Flux de connaissances médicales dans le système de soins français, schéma adapté de [Charlet, 2002].	7
1.2	Analyse descriptive, prédictive et prescriptive.	9
1.3	Processus d'analyse de données, adapté de [Fayyad et al., 1996].	11
1.4	Nuages de mots des 10 thèmes associés à l'application des méthodes descriptives sur les données de santé.	14
2.1	Trois écarts pour un même motif : dans le premier cas, G3 a une expression beaucoup plus forte que G1 et G5 qui ont une expression similaire, dans le deuxième cas, l'écart est réduit et dans le troisième l'écart est le plus petit vis à vis de la discrétisation envisagée.	24
2.2	Exemple d'un phénomène spatio-temporel : apparition des épidémies de dengue. L'espace est découpé en trois zones dans lesquelles on observe les événements pluie, points d'eau, moustiques et cas de dengue pour 3 estampilles temporelles. Dans le coin supérieur droit, on présente un exemple de motif spatio-séquentiel extrait pour la zone Z_3 . Le – représente l'absence d'évènement et le • la relation spatiale.	28
2.3	Trois vues de motifs de gènes : (a) cluster ; (b) document et (c) treemap.	34
2.4	Fonctions impliquées dans les motifs et leur évolution dans le temps.	34
2.5	Trois vues d'un motif spatio-temporel : (a) on optimise l'espace avec un algorithme de force. Le temps est représenté par un arc (b) ou une spirale (c).	35
2.6	Système de visualisation de motifs spatio-temporels : application au suivi des épidémies de dengue : (a) vue géographique, (b) liste de motifs, statistiques associées à un motif dans les zones d'apparition (c) et les temps d'apparition (d).	36

2.7	Système de visualisation de motifs partiels ordonnés clos : application à l'évaluation de la qualité de l'eau des rivières : (a) vue cluster, (b) vue géographique et (c) vue d'un motif sélectionné.	37
2.8	Une information riche exploitée dans les motifs.	41
3.1	Exemple de message enrichi. On retrouve dans ce message des informations temporelles (<i>ça va faire 3 mois et depuis environ 2 semaines</i>). On sait également, d'après le profil, que l'utilisateur est assez nouveau car son rang est <i>régulier</i> et il n'a posté que 35 messages. Il poste des questions <i>Avez vous déjà connu ça ?</i> et exprime son état affectif <i>je ressens des douleurs</i> à propos de thèmes précis (<i>tamoxifene, cicatrice, douleur intercostal</i>). Toutes ces informations sont capturées avec les méthodes automatiques que nous proposons.	47
3.2	Exemple de contextes construits. Sur la gauche de la figure, nous retrouvons les snippets de textes retournés par le moteur de recherche et sur la droite, la requête initiale (voir ligne <i>topic</i>) et une expansion de cette requête possible (voir ligne <i>expansion</i>) et les contextes correspondants, c'est à dire la suite de termes pondérés par leurs fréquences dans les snippets.	50
3.3	Modèle des états affectifs. Un <i>état affectif</i> est ressenti par un <i>émetteur</i> (ou <i>source</i>). Il fait référence à une <i>polarité</i> , c'est-à-dire un jugement pouvant être <i>positif</i> s'il est lié à un effet bénéfique sur l'émetteur, <i>négatif</i> dans le cas contraire ou <i>neutre</i> . L'état affectif peut également faire référence à une <i>émotion</i> comme la <i>colère</i> , la <i>joie</i> , la <i>tristesse</i> , etc. Généralement, les émotions sont associées à une polarité. La joie est considérée par exemple comme positive, la colère comme négative et la surprise comme neutre. On peut associer différents niveaux d' <i>intensité</i> à l'état affectif (<i>e.g., très positif, un peu triste, etc.</i>). Pour finir, l'état affectif porte sur une <i>cible</i> qui est le réceptacle de l'opinion ou de l'émotion [Bringay et al., 2014].	53
3.4	Exemple de message expert. L'utilisateur a posté 2,455 messages. Son niveau de langue est élevé. Le message ne contient pas de faute d'orthographe. Sa description est factuelle.	58
3.5	Exemple de message non expert. L'utilisateur n'a posté que 39 messages. Le message contient des fautes d'orthographe, des abréviations, etc. L'auteur donne son ressenti (<i>je suis un peu perdue</i>).	58
3.6	Enrichissement sémantique des messages.	61

Liste des tables

1	Résumé des projets et des encadrements. Légende : M-Master, T-Thèse, PostD-Postdoctorant.	3
1.1	Requêtes utilisées pour réaliser la revue de la littérature.	12
1.2	Résultats quantitatifs des requêtes.	12
2.1	Résumé des projets et des encadrements sur la thématique des motifs séquentiels.	22

2.2	Expression des gènes pour la puce à ADN $P1$ et deux séquences associées avec l'écart minimal de 0.1.	23
2.3	Séquences obtenues à partir de 5 puces à ADN. Selon un écart minimum de 0.1 fixé par l'utilisateur, deux séquences sont associées à $P1$ et $P3$ et une séquence est associée à $P2$, $P4$ et $P5$. Les deux dernières colonnes illustrent le calcul du support pour deux motifs. Le premier sera fréquent mais pas le deuxième.	23
2.4	Résumé des projets et des encadrements sur la thématique des motifs contextuels.	24
2.5	Exemple de base de données d'actes médicaux au cours du temps. Chaque ligne correspond à un patient et chaque colonne à un créneau horaire. Les lettres {a,b,c,d,e} sont des actes médicaux.	25
2.6	Mise en valeur du motif $\langle\langle a \rangle\rangle\langle\langle b \rangle\rangle$ (en gras) soit l'acte a suivi de l'acte b . Ce motif est fréquent dans la base pour un support minimum de 50%.	25
2.7	Mise en valeur du motif $\langle\langle a \rangle\rangle\langle\langle b \rangle\rangle$ (en gras) avec les informations contextuelles sur l'âge et le sexe. Ce motif est spécifique aux jeunes. Une seule personne âgée est concernée.	26
2.8	Résumé des projets et des encadrements sur la thématique des motifs spatio-temporels.	27
2.9	Évolution de l'environnement pour les zones Z_1 , Z_2 et Z_3 pour 3 dates. Le – représente l'absence d'évènements.	28
2.10	Résumé des projets et des encadrements sur la thématique des motifs partiellement ordonnés	29
2.11	Des motifs séquentiels clos aux motifs partiellement ordonnés clos.	29
2.12	Résumé des projets et des encadrements sur la thématique des mesures d'intérêt.	31
2.13	Résumé des projets et des encadrements sur la thématique de la visualisation de motifs.	32
2.14	Résumé des projets et des encadrements sur la thématique des méthodes prédictives et prescriptives.	38
2.15	Exemple de motifs associés aux deux classes C_1 et C_2 . Si l'on considère la séquence $S = \langle\langle a \rangle\rangle\langle\langle b \rangle\rangle\langle\langle e \rangle\rangle\langle\langle f \rangle\rangle\langle\langle g \rangle\rangle$. Les 2 motifs de la classe C_1 sont inclus dans S contre 1 motif de la classe C_2 . S est attribuée à la classe C_1	38
3.1	Résumé des projets et des encadrements sur la thématique de l'enrichissement des messages - de quoi ?	48
3.2	Résumé des projets et des encadrements sur la thématique de l'enrichissement des messages - comment ?	52
3.3	Exemple d'expressions de la cible de l'émotion. Dans les phrases $P1$, $P2$ et $P3$, l'émotion <i>peur</i> porte sur une entité représentée respectivement par le concept général <i>médicament</i> , l'événement <i>début de la chimiothérapie</i> et l'EN <i>IVEMEND</i> (qui est un nom de médicament). Dans la phrase $P4$, la cible de la polarité est plus complexe et porte sur un <i>aspect</i> , une caractéristique : <i>le taux de tolérance</i> de l'entité <i>médicament</i> . Dans la phrase $P5$, seul l'aspect est présent. Dans la phrase $P6$, il n'y a pas de cible explicite. L'émotion fait référence au contexte général dans lequel la phrase est énoncée. Dans la phrase $P7$, la cible est détaillée dans le reste de la phrase et ne se limite pas à l'entité médicale <i>douleur</i>	56
3.4	Résumé des projets et des encadrements sur la thématique de l'enrichissement des messages - qui ?	57

3.5 Résumé des projets et des encadrements sur la thématique de l'enrichissement des messages - quand? 60

Liste des algorithmes

GLOSSARY

- ADN** Acide DésoxyriboNucléique. 3, 21, 23, 31, 33, 38, 66, 87
- ADVANCE** ADVanced Analytics for data ScienceE (<https://www.lirmm.fr/recherche/equipes/advance>). 2, 21
- ANR** Agence Nationale de la Recherche (<http://www.agence-nationale-recherche.fr/>). 3, 22, 29–32, 35, 44, 62
- API** Application Programming Interface. 11, 62
- CHU** Centre Hospitalier Universitaire. 2, 3, 24, 66
- CHV** Consumer Health Vocabulary. 62
- CIFRE** Conventions Industrielles de Formation par la REcherche. 3, 24
- CRCT** Congés pour recherches ou conversions thématiques. 2
- DA** Data Analytics. 6
- DASS** Direction des Affaires Sanitaires et Sociales (<http://www.dass.gouv.nc/portal/page/portal/dass/>). 3, 27
- DM** Data Mining. 6
- EBM** Evidence Base Medecine. 7
- EDA** Exploratory Data Analysis. 6
- EN** Entité Nommée. 55, 56, 87
- ENGEES** Ecole Nationale du Génie de l'Eau et de l'Environnement de Strasbourg (<https://engees.unistra.fr/>). 30
- GHM** Groupe Homogène de Malades. 7
- GIS** Système d'information géographique. 27
- HIV** Human Immunodeficiency Virus. 34
- HON** Fondation la santé sur Internet (https://www.hon.ch/home1_f.html). 45

- I3M** Institut de Mathématiques et de Modélisation de Montpellier (<http://www.i3m.univ-montp2.fr/>). 3, 46, 48, 52, 60, 66
- ICM** Institut du Cancer de Montpellier (<http://www.icm.unicancer.fr/fr>). 3, 17, 46, 48, 52, 60, 66
- IGMM** Institut de Génétique Moléculaire de Montpellier (<http://www.igmm.cnrs.fr/?lang=fr>). 3, 32, 33
- INSERM** Institut national de la santé et de la recherche médicale (<http://www.inserm.fr/>). 3, 38
- InVS** Institut de Veille Sanitaire (<http://www.invs.sante.fr/>). 3, 27
- KDD** Knowledge Discovery Process. 6, 10
- LDA** Latent Dirichlet Allocation. 13, 51
- LIRMM** Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (<http://www.lirmm.fr/>). 2
- MEDLINE** Medline (<http://www.ncbi.nlm.nih.gov/pubmed>). 12
- MeSH** Medical Subject Headings (<http://mesh.inserm.fr/mesh/>). 51, 62
- MMDN** Mécanismes Moléculaires dans les Démences Neurodégénératives. 3, 22, 31
- NCBI** National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>). 12, 13
- PD** Patern Discovery. 21
- PEPS** Projets Exploratoires Premier Soutien. 33
- PMSI** Programme Médicalisé des Systeme d'Information. 7, 15, 21, 24
- PRISMA** Preferred Reporting Items for Systematic Reviews and Meta-Analyses (<http://www.prisma-statement.org/>). 11
- PubMed** PubMed comprend plus de 24 millions de citations de littérature biomédicale (<http://www.ncbi.nlm.nih.gov/pubmed>). 11, 31, 33
- QdV** Qualité de vie. 48, 52, 60
- RSS** Résumé de Sortie Standardisé. 7
- SNIIRAM** Système national d'information inter-régimes de l'Assurance maladie (<http://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/sniiram/finalites-du-sniiram.php>). 15
- SNOMED** Systematized Nomenclature of Medicine (http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html). 62
- UMLS** Unified Medical Language System (<http://www.nlm.nih.gov/research/umls/>). 62
- VA** Visual Analytics. 33

PARTIE I

CURRICULUM VITAE

CURRICULUM VITAE

BRINGAY Sandra

Née le 20 Octobre 1979, à Montauban
Française, Célibataire, 1 enfant né en 2011
5 allée Frédéric Mistral,
34790 Grabels
(+33) [0]6 83 24 79 33
<http://www.lirmm.fr/~bringay/>

LIRMM, UMR 5506, Bâtiment 5
860 rue de St Priest,
34095 Montpellier cedex 5
(+33) [0]4 67 14 21 57
bringay@lirmm.fr

Dpt. MIAP, Université Paul Valéry Montpellier
Route de Mende, 34199 Montpellier
(+33) [0]4 67 41 86 36
sandra.bringay@univ-montp3.fr

CURSUS

SITUATIONS ADMINISTRATIVES

- Depuis 2007** **Maître de conférences** à l'Université Paul Valéry Montpellier, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), Equipe ADVANSE, Montpellier
- 2013-2014** **Délégation CNRS** (6 mois) et **Congé de recherche thématique** (6 mois) au LIRMM
- 2005-2007** **Attachée Temporaire d'Enseignement et de Recherche** à l'Université de Lille 2, Droit et Santé, Centre de Recherche en Informatique Médicale (CERIM), Lille
- 2002-2005** **Chargée de recherche** pour Alternattech, Financement Etat-Région dans le cadre du projet HTSC DocPatient, Amiens

PRIME : Depuis 2011, titulaire de la **Prime d'excellence scientifique**

FORMATIONS

- 2006** **Doctorat en informatique**, *mention Très Honorable*, Université Picardie Jules Verne, Laboratoire de Recherche en Informatique d'Amiens (LaRIA devenu MIS), Amiens
- 2002** **Diplôme d'Ingénieur** de l'Institut National des Sciences Appliquées (INSA) de Rouen, dans le département Génie Mathématique
Diplôme d'Etudes Approfondies, Sciences de l'Ingénieur, *mention Bien*, Université de Rouen

ACTIVITES D'ENSEIGNEMENT (DEPUIS 2007)

Enseignements à l'université Montpellier 3 :

L'université Paul Valéry Montpellier est une Université de lettres et sciences humaines. Depuis 8 ans en tant que maître de conférences, j'effectue mes enseignements au sein du département MIAp (Mathématiques et Informatique appliquée). Le volume d'enseignements que je réalise est approximativement de 230 heures TD chaque année. J'interviens dans les formations ci-dessous :

- **Compétences informatiques et C2I (enseignements transversaux pour tout niveau licence) :** la plus grosse part de mon activité pédagogique (130 heures TD cette année) se déroule dans le cadre de cours transversaux portant sur les bases des outils informatiques et des réseaux et sur la manipulation de logiciels de bureautique. Ces enseignements sont destinés à l'ensemble des étudiants de licence pour tous les cursus de notre université. L'objectif est de les préparer pour le Certificat Informatique et Internet (C2I) niveau 1. Ce diplôme spécifique vient s'ajouter au diplôme de licence. Il atteste de l'habileté des étudiants à manipuler des données numériques et de leur capacité à prendre en main les outils informatiques de base.
- **Base de données (licence et master) :** Si les étudiants de l'université Paul Valéry Montpellier ne sont pas des spécialistes de l'informatique, certains d'entre eux seront amenés à manipuler des gros volumes de données. Par exemple, dans les entreprises et collectivités, la mise en place d'outils de gestion informatiques nécessite très souvent la création, l'utilisation ou la consolidation de données structurées dans l'optique de les sauvegarder, d'effectuer des recherches, ou encore d'interagir avec d'autres systèmes. Pour cela, des bases de données sont mises en place à tous les niveaux : gestion du personnel, des fournisseurs, des achats... Nos étudiants, futurs décideurs, ont un rôle important à jouer lors de la conception de ces outils pour obtenir des bases performantes et réellement adaptées aux besoins des utilisateurs. Dans ce contexte, j'interviens dans différents modules. En licence, j'ai collaboré avec deux collègues pour mettre en place en janvier 2013 un nouveau niveau expert de 20h ouvert en option pour les enseignements transversaux (60 étudiant par an environ). Son objectif est de donner aux étudiants les compétences théoriques et pratiques nécessaires pour tenir ce rôle de décideur et interagir pertinemment avec les informaticiens. J'interviens également en master 1 IOD (Institution Organisation et Développement) pour une intervention plus théorique sur la modélisation des bases de données (20 heures TD) et en master 2 ESEEC (Expertise Socio-économique, Emplois et Compétences) sur la gestion de projet de bases de données (6 heures TD).
- **Prise de parole en public (formation doctorale) :** Quelle que soit la discipline du doctorant, celui-ci doit savoir s'exprimer clairement à l'oral lorsqu'il présente ses résultats à son équipe, exprime son point de vue pendant des réunions, fait part de ses objectifs et contraintes pendant la promotion de son projet, enseigne à des étudiants, valorise ses activités lors de conférences, séminaires... Et surtout, le jour de sa soutenance de thèse ! J'ai conçu cette formation avec l'objectif de rendre le doctorant autonome et performant lorsqu'il prend la parole en public et réalise des diapositives avec un logiciel de PAO (Présentation Assistée par Ordinateur). Dans ce contexte, j'effectue une intervention de 18h heures TD chaque année au sein des écoles doctorales 58 et 60 (les deux écoles doctorales de l'université Paul Valéry Montpellier). Les étudiants assistant à cette formation sont des doctorants en première année de thèse.
- **Rédaction de documents (formation doctorale) :** Les étudiants en thèse rencontrent de grandes difficultés au moment de la rédaction de leur mémoire. En effet, ils perdent beaucoup de temps en manipulant un logiciel de traitement de texte alors qu'ils devraient en gagner, notamment en automatisant certains traitements (table des matières, index, bibliographie...). Dans ce contexte, j'effectue une intervention de 24 heures TD chaque année au sein des écoles doctorales 58 et 60. Les étudiants assistant au cours sont des doctorants en deuxième année de thèse. Ce cours entre dans le cadre des modules de formations doctorales obligatoires.

Enseignement à l'université Montpellier 2 :

- **Fouille de données et santé :** Depuis 2007, je suis intervenue chaque année dans le Master 2 Recherche Informatique de Montpellier 2, dans le master STIC Santé et dans la dernière année de l'option Tic Santé de l'Institut Telecom (pour chaque formation pour une vingtaine d'étudiants environ 3 heures TD). J'ai dressé un panorama des méthodes et outils de fouille de données ayant été appliqués au domaine de la santé.
- **Encadrements :** Chaque année, j'encadre des Travaux Etudes et Recherche TER en master 1 et en

master 2 d'informatique. Je dirige également des stages d'analyse, des stages recherche et des stages professionnalisant dans les différents master 2 d'informatique de l'Université de Montpellier 2.

Administration

Chaque année, environ 3500 étudiants suivent nos cours d'informatique à l'Université Paul Valéry-Montpellier. Le travail de gestion impliqué par ce gros volume d'étudiants est assuré par cinq enseignants permanents. Pour pourvoir l'ensemble des TDs, nous engageons en moyenne 30 vacataires chaque semestre. Chaque semaine, nous gérons environ 80 groupes de TD de 5 niveaux différents dans lesquels tous les étudiants de la L1 ainsi que les L2 et L3 ayant pris l'option informatique sont répartis en fonction de leur compétences. Pour ces 5 niveaux, il faut disposer du même matériel, dispenser le même contenu pédagogique, organiser les mêmes examens... Une partie de ces enseignements est par ailleurs délocalisée dans une antenne à Béziers. Le travail de coordination prend donc une part très importante de notre temps et nous cherchons à rationaliser nos activités en mettant en place des procédures strictes et en créant des outils informatiques pour supporter les tâches répétitives.

Dans ce contexte, en plus du travail classique sur les supports pédagogiques, je coordonne l'intervention des vacataires dans les différents modules. Pour cela, chaque semaine, j'indique aux vacataires par mail ce qu'ils doivent réaliser en TD et je réponds à leurs questions. Pour faciliter ce travail, j'ai mise en place, avec A. Pinlou, un wiki sur lequel sont décrits pour chaque niveau de nos enseignements, les attentes pédagogiques lors des séances de TDs, les procédures d'inscriptions des étudiants, les procédures à appliquer pendant les examens... Je participe également à l'évolution et à la maintenance d'un outil de gestion permettant l'inscription des étudiants dans les groupes de TD, le suivi des présences, des notes...

Je suis également fortement impliquée dans l'organisation de la certification C2I (Certificat de l'Informatique et de l'Internet). En binôme avec A. Pinlou, nous assurons l'évolution et la maintenance d'une plate-forme d'examens qui permet aux étudiants d'accéder de manière sécurisée à des sujets différents les uns des autres et aux enseignants de récolter les copies. Nous avons par exemple développé une application permettant aux étudiants de répondre à des questions rédactionnelles qui sont ensuite réparties entre les correcteurs de telle sorte que ces derniers aient le moins possible de questions différentes à corriger. Cette application permet de gagner beaucoup de temps de correction et de garantir l'équité (tous les étudiants ayant eu la même question sont notés par un seul correcteur).

Chaque année, nous récoltons environ 9000 copies d'examens. J'ai mis en place un outil de gestion des grilles de correction qui permet de centraliser le détail des notes des chargés de TD. Depuis bientôt quatre ans, en collaboration avec A. Pinlou, nous tentons également de mettre en place des outils de correction semi-automatique de certains exercices de certification du C2I. À ce titre, le conseil d'administration de Montpellier 3 a supporté notre projet en 2011 en finançant un ingénieur à temps plein pendant 12 mois.

En collaboration avec G. Richomme, j'ai également développé une plateforme qui permet de suivre les corrections réalisées par les chargés de TDs, d'effectuer différents contrôles et d'éditer automatiquement les jurys pour les 5 niveaux que nous gérons sur les deux sites de Montpellier et Béziers.

Finalement, nous menons une réflexion globale sur l'ensemble de nos procédures et outils afin de les rendre plus robustes et garantir ainsi l'équité pour tous les étudiants.

Depuis 2014, je me suis investie dans l'administration de la licence MIASH, Mathématique Informatique Appliquées aux Sciences Humaines et Sociales, qui a ouvert en septembre 2014 à Montpellier (création de vidéos de présentations, poster, flyer, présentations aux journées portes ouvertes, au salon des étudiants...).

J'ai également pris la **responsabilité du projet de master MIASHS**, qui n'ouvrira qu'en 2017 (construction du programme, rédaction de la maquette et des différents descriptifs...). Ce master s'adresse à des étudiants ayant un bagage minimum en Informatique et en Mathématiques. Il formera des professionnels qui seront spécialistes en traitement de données SHS, avec une bonne connaissance des méthodes de traitement de gros volumes de données Big Data, de fouille de données, de données connectées Open Data, de visualisation, de Statistique et d'aide à la décision. Ce master sera ouvert pour les deux années en EAD et organisé pour permettre la formation continue et l'alternance.

PRIME : Depuis 2008, titulaire de la **Prime de responsabilités pédagogiques (PRP)**

RESPONSABILITE PEDAGOGIQUE D'ENSEIGNEMENT

- **Depuis 2014** Responsable du master MIASHS (Ouverture en 2017)
- **Depuis 2012** Responsable du niveau Expert BD des cours transversaux d'informatique (Licence L2, L3)
- **Depuis 2009** Responsable des modules « Prise de parole en public » (Doctorants)

COMITES DE SELECTION

- Depuis 2009 Elue dans le **pool d'experts** (section 27) pour les recrutements rattachés au LIRMM UMR 5506
- **Membre du comité de sélection** (section 27), Université de Toulouse III Paul Sabatier en 2015
- **Membre du comité de sélection** (section 27), Université Paul-Valéry en 2015
- **Membre du comité de sélection** (section 27), IUT Aix Marseille en 2014
- **Membre du comité de sélection** (section 27), Université de Montpellier 2 en 2010
- **Membre du comité de sélection** (section 27), CIRAD en 2008

ENCADREMENTS (DEPUIS 2007)

THESES : J'ai co-encadré 7 thèses dont 4 sont **soutenues**.

- **Depuis Octobre 2014** M. Tapi Nzali, Thèse de l'Université Montpellier 2 (spécialité Bio-Statistique), *Analyse des forums de santé pour mesurer la qualité de vie des patientes atteintes d'un cancer du sein*
Taux d'encadrement : 33%. Co-encadrement avec C. Lavergne et C. Mollevi
Soutenance prévue en 2017
Financement : Univ. Montpellier 2
- **Depuis mai 2014** J. Pinaire, Thèse de l'Université Montpellier 2, *Trajectoires de patients*
Taux d'encadrement : 33%. Co-encadrement avec J. Azé et P. Landais
Soutenance prévue en 2017
Financement : CHU Montpellier
- **Depuis septembre 2013** A. Abdaoui, Thèse de l'Université Montpellier 2, *Rôles dans les réseaux sociaux*
Taux d'encadrement : 50%. Co-encadrement avec J. Azé
Soutenance prévue en 2016
Financement : Averoès
- **2011-2014** M. Fabregue, Co-tutelle entre l'Université Montpellier 2 et l'Université de Strasbourg, *Extraction d'informations synthétiques à partir de données séquentielles : Application à l'évaluation de la qualité des rivières*
Taux d'encadrement : 40 %. Co-encadrement avec A. Braud, F. Leber et M. Teisseire
Soutenance : 26 Novembre 2014
Situation actuelle : Responsable recherche chez Aquabio
Financement : ANR Fresqueau
- **2010-2013** H. Alatriza Salas, Thèse de l'Université Montpellier 2 et l'Université de la Nouvelle Calédonie, *Extraction de relations spatio-temporelles à partir des données environnementales et de la santé*
Taux d'encadrement : 40 %. Co-encadrement avec F. Flouvat, N. Selmaoui et M. Teisseire
Soutenance : 04 Octobre 2013
Situation actuelle : Post-doctorant depuis septembre 2014 à l'Université de Lima au Pérou
Financement : ½ Région Languedoc Roussillon et ½ Université De la Nouvelle Calédonie
- **2008-2011** J. Rabatel, Thèse de l'Université Montpellier 2. *Fouille de données comportementales pour la maintenance ferroviaire*
Taux d'encadrement : 50%. Co-encadrement avec P. Poncelet
Soutenance : 21 septembre 2011
Situation actuelle : post-doctorant au LIRMM
Financement : CIFFRE avec la société Fatronick
- **2007-2010** P. Salle, Thèse de l'Université Montpellier 2. *Fouille de données bio-médicales*
Taux d'encadrement : 60%. Co-encadrement avec M. Teisseire
Soutenance : 01 Juin 2010
Situation actuelle : Responsable du département Recherche pour Expertise Radiologie
Financement : ANR Ubiquitus

STAGES POST-DOCTORAUX : J'ai collaboré avec 3 docteurs en stages post-doctoraux :

- **2014** T. Opitz (Chercheur INRA depuis sept. 2014): *Analyse des forums de santé pour une meilleure connaissance de la qualité de vie des patientes atteintes d'un cancer du sein*
Collaboration avec C. Joutard, C. Lavergne et C. Mollevi

- **2009** J. Nin Guerrero (Associate researcher at the Barcelona Supercomputing Center depuis 2013) : *Clustering hiérarchique et résumé de motifs séquentiels de gènes*
Collaboration avec M. Teisseire
- **2009** F. Chakkour (Associate professor at Aleppo University, Syrie) : *Recherche de documents pertinents associés aux motifs séquentiels de gènes*
Collaboration avec M. Roche et M. Teisseire

MASTER : J'ai co-encadré 10 masters 2 dont 2 sont en cours.

- **2015** C. Maigrot : *Détection des points de ruptures des personnes suicidaires dans le réseau social Facebook*
Co-encadrement avec J. Azé
- **2015** M. Vioulès : *Détection des changements de phases des personnes suicidaires dans le réseau social Twitter*
Co-encadrement avec J. Azé
- **2014** Y. Motie : *Elaboration d'un vocabulaire patient-médecin*
Co-encadrement avec J. Azé et T. Opitz
Situation suivante : Master Recherche à Berlin en sept. 2014
- **2014** O. NKaira : *Analyse de la confiance dans les forums de santé*
Co-encadrement avec A. Abdaoui et J. Azé
- **2013** A. Zine : *Fouille de données transcriptomiques dans le cadre de la lutte contre le HIV et le cancer*
Co-encadrement avec P. Poncelet et M. Teisseire
Situation suivante : Ingénieur LIRMM
- **2013** S. Melzi : *Analyse semi-automatique de fora de santé*
Co-encadrement avec P. Poncelet
Situation suivante : Master STIC Santé
- **2012** M. Nair : *Skyline et préférences*
Co-encadrement avec J. Munro (Optimal Medecine) et P. Poncelet
Situation suivante : Thèse Orange, Lannion
- **2011** M. Fabregue : *Fouille de données de santé*
Co-encadrement avec P. Poncelet et M. Teisseire
Situation suivante : Thèse ANR Fresqueau
- **2011** J. Dorado : *Classification automatique de Tweets*
Co-encadrement avec P. Poncelet et M. Roche en collaboration avec la société Web Report
Situation suivante : Master professionnel en 2012
- **2010** M. Touzani : *Entrepôt de données*
Co-encadrement avec A. Laurent, T. Libourel et J. Quinqueton
Situation suivante : Thèse LIRMM
- **2010** E. Capacho : *Skylines et environnement*
Co-encadrement avec M. Teisseire
Situation suivante : Ingénieur, Nantes
- **2009** T. Bouadi : *Skylines et préférences*
Co-encadrement avec M. Teisseire
Situation suivante : Thèse IRISA, Rennes
- **2008** H. Saneifar : *Ontologies floues et fouille de données*
Co-encadrement avec A. Laurent
Situation suivante : Thèse LIRMM, CIFFRE Satin IP
- **2008** J. Rabatel : *Fouille de données comportementales pour la maintenance ferroviaire*
Co-encadrement avec P. Poncelet et M. Teisseire
Situation suivante : Thèse LIRMM, CIFFRE Tecnalía

ACTIVITES DE RECHERCHE (DEPUIS 2007)

ADJOINTE DE L'ÉQUIPE : ADVANSE du LIRMM (Responsable d'équipe : P. Poncelet)

DOMAINES DE RECHERCHE : Ingénierie des connaissances, Extraction de connaissances, Fouille de données, Motifs, Informatique médicale, Santé

PRIME : Depuis 2011, titulaire de la **Prime d'excellence scientifique**

RESPONSABLE & MEMBRE DE PROJETS DE RECHERCHE

- **Porteur du projet « Parlons de nous »** (depuis 2013) *Fouille semi automatique de fora de santé*
Financement MSH-M Maison des Sciences de l'Homme de Montpellier en 2013 (10k€) puis réseau Inter-MSH en 2014 (10k€)
Il s'agit de mettre en place une collaboration nationale pour identifier des méthodes d'ingénierie des connaissances et de traitement automatique de la langue naturelle qui permettent d'extraire dans les fora de santé de la connaissance (marques de contexte, de structure et de sentiment), utile pour les professionnels de santé. <https://www.lirmm.fr/patient-mind/>
- **Porteur du PEPS HIV** (2012-2013) *Fouille de données transcriptomiques (HIV)*
Financement PEPS Projets Exploratoires Premier Soutien du CNRS
Co-responsable scientifique avec C. Lecellier (IGHM Montpellier) et P. Poncelet (10k€)
Il s'agit de mettre en œuvre des méthodes de fouille sur des données issues de puces à ADN qui permettent de distinguer différentes souches du HIV et ainsi de mieux comprendre les modulations de la maladie au niveau transcriptomique.
- **Membre du Groupe de travail Toxicovigilance** (depuis 2012) *Détection de Signal Automatisée*, dirigé par L. Faisandier de l'INVS
L'InVS possède une base qui recense tous les cas détaillés et répertoriés de signalement d'intoxications et cherche à définir des méthodes pour exploiter cette base (statistiques et fouille de données). Dans le cadre du groupe, nous travaillons sur l'identification des méthodes les plus pertinentes.
- **Membre de l'ANR SFIR** (2013-2017) *Indexation sémantique de ressources biomédicales francophones*, portée par C. Jonquet
Financement de l'Agence Nationale de la Recherche (270k€)
Cette ANR Jeune Chercheur vise à développer une plateforme dédiée à la communauté scientifique bio-médicale qui permettra de faciliter les croisements entre les jeux de données ouverts pour découvrir des nouvelles ressources ou données pertinentes.
- **Membre de l'ANR Fresqueau** (2011 – 2014) *Fouille de données pour l'évaluation et le suivi de la qualité hydrobiologique des cours d'eau*, portée par F. Leber et M. Teisseire
Financement ANR (852k€)
Cette ANR vise à développer des méthodes et outils dédiés à l'analyse des données de qualité des cours d'eau.
- **Membre du projet PRADNET** (2010-2013) *Primate Alzheimer's disease network*
Financement par la Fondation de Coopération scientifique Maladie d'Alzheimer et maladies apparentées (300 k€)
L'objectif de ce projet est de développer des outils permettant d'offrir à la communauté scientifique un modèle primate d'étude du vieillissement et de la maladie d'Alzheimer.
- **Membre du PEPS GeneMining** (2007-2008) *Fouille de puces transcriptomiques (Alzheimer)*
Financement PEPS (10k€)
Il s'agit de mettre en œuvre des méthodes de fouille sur des données issues de puces à ADN afin de mieux comprendre la maladie d'Alzheimer. La difficulté est ici le tout petit nombre de puces à notre disposition.
- **Membre du PEPS ST2I-SHS** (2007-2008) *Langage, Mémoire et Alzheimer*
Financement PEPS (10k€)
Il s'agit de mettre en œuvre des méthodes de fouille de données pour exploiter les données issues de retranscriptions d'interviews de patients atteints de la maladie d'Alzheimer. L'objectif est d'identifier des signes dans leur discours qui soient caractéristiques de la maladie.

RESPONSABLE ET MEMBRE DE COMITES D'ORGANISATION

- **2014 Atelier Forum de santé**, *Quand le patient prend le pouvoir sur sa santé*. Université de la e-Santé, Castres, 13 mai 2014. Co-organisatrice avec N. Souf
<https://www.lirmm.fr/patient-mind/pmwiki/pmwiki.php?n=Site.4juin2014>
- **2014 Atelier IC & Santé**, *Ingénierie des connaissances et santé*. Clermont Ferrand, 13 mai 2014. Co-organisatrice avec N. Souf et L. Tamine-Lechani
<https://www.lirmm.fr/ic-sante/>
- **2014 NLDB**, *19th International Conference on Application of Natural Language to Information Systems*. Montpellier, Juin 2014. Membre du comité d'organisation
<http://www2.lirmm.fr/~mroche/NLDB2014/Web/>
- **2013 Journée Thématique Thème C, GdR STIC Santé**, *Étude des fora de santé : A quoi pensent les patients ?* Montpellier, Co-organisatrice avec N. Souf
http://www.stic-sante.org/index.php?view=details&id=165:etude-de-fora-en-sante-a-quoi-pensent-les-patients&option=com_eventlist&Itemid=200
- **2011 Atelier ExCoco** – *Extraction de connaissances et contexte*, associé à la conférence IC 2011 : Chambéry, France, Mai 2011. Co-organisatrice avec N. Souf et A. Baynex
<http://www.excoco.org/>
- **2011 Atelier ECS** – *Extraction de connaissances et santé*, associé à la conférence EGC 2011 : Brest, France, Janvier 2011. Co-organisatrice avec N. Souf et A. Baynex
<http://www.ecs2011.org/>
- **2009 EDA** - *5ème journées francophones sur les entrepôts de données et l'analyse en ligne* : Montpellier, France, 4 et 5 Juin 2009. Co-organisatrice avec A. Laurent et M. Teisseire
<http://www.lirmm.fr/EDA09/>
- **2009 Decision and Health** - Session Invitée IS04 associée à la conférence First KES International Symposium on *Intelligent Decision Technologies* IDT'09, 23-24 avril 2009. Co-organisatrice avec M. Teisseire

RELECTURES D'ARTICLES

- International : JBI, MIE, MedInfo, IDA, ICDM, Fuzz-ieee, Discovery Science, Information Sciences, IJCNN, Health Informatic Journal...
- National : EGC, EDA, BDA, Eval'ECD, RECITAL, SIIM...

PROTOTYPAGE <http://www.lirmm.fr/recherche/equipes/avance>

- **2014 Visualisation d'indicateurs de qualité de l'eau des rivières**. Co-Responsable avec A. Braud, F. Leber et M. Teisseire
- **2013 Visualisation de motifs spatio-temporels**. Co-Responsable avec F. Flouvat, N. Selmaoui et M. Teisseire
- **2013 Visualisation de motifs séquentiels de gènes impliqués dans le HIV**. Co-Responsable avec P. Poncelet
- **2010 Typage de cancer du sein**. Co-Responsable avec P. Poncelet et M. Teisseire
- **2009 Outil de correction automatique de gros volumes de copies**. Co-Responsable avec A. Pinlou
- **2009 Visualisation de motifs séquentiels de gènes**. Co-Responsable avec M. Roche et M. Teisseire

JURYS

- **2015 Jury de Doctorat (hors doctorants encadrés)** : F. Soualah-Alila (Université de Dijon), examinatrice, *Conception d'une architecture sémantique basée sur un système de recommandation et des techniques d'ontologies évolutives. Application aux domaines du elearning/social learning/mobile learning*.
- **2012 Jury de Doctorat (hors doctorants encadrés)** : C. Lopez (Université de Montpellier 2), examinatrice, *Titration automatique de documents textuels*
- Rapporteur de stages de M2 Informatique (en moyenne deux rapports par an) pour Montpellier et Lyon

AUTRES RESPONSABILITES SCIENTIFIQUES

- **Depuis 2015** Membre de la commission Logiciel Recherche du LIRMM
- **Depuis 2014** Membre du Comité Scientifique de l'**Agora des Savoirs** <http://www.montpellier.fr/4152-agora-des-savoirs-2014-2015.htm>
- **Depuis 2014** Membre du bureau du collège **Ingénierie des connaissances de l'AFIA**
- **Présidente de session**, Construction d'ontologies, Conférence Ingénierie des Connaissances, Clermont Ferrand, 2014. Programme <http://www.irit.fr/IC2014/node/3>

RESPONSABLE & MEMBRE DE PROJETS INDUSTRIELS

- **Compilsoft** (2014-2015) Aide à la faisabilité technologique (30k€)
GIMMG : Mise au point d'un système de gestion de l'information produit multidimensionnel générique
Responsable scientifique
- **JVWEB** (2013-2014) Aide à la faisabilité technologique (30k€)
Aide à la classification automatique de flux de produits dans les places de marché
Co-responsable scientifique avec J. Azé et D. Ienco
- **Octipas** (2010-2011) Projet LRI d'Incubation Languedoc Roussillon (8k€)
Stratégie et solutions de e-commerce
Co-responsable scientifique avec P. Poncelet
- **WebReport** (2010-2011) Prestation de service (21k€)
Détection automatique de catastrophes via les tweets
Co-responsable scientifique avec M. Roche
- **PIKKO** (2009-2011)
Visualisation de motifs séquentiels de gènes
Membre de la collaboration
- **Fatronick** (2008-2011) Thèse CIFFRE et accompagnement (60k€)
Fouille de capteurs
Membre de la collaboration
- **EKIOO** (2009-2010) Projet LRI d'Incubation Languedoc Roussillon (10k€)
Conception d'un annuaire collaboratif pour les entreprises
Responsable scientifique
- **IBISKUS** (2009-2010) Prestation de service financée par OSEO (18k€), soutien à l'innovation et la croissance des PME
Logiciel d'analyse "intelligente" (datamining) de données décisionnelles
Membre de la collaboration

EXPERTISES

- **Depuis 2012** Rédaction d'une dizaine de **demandes d'agrément**s pour des PME du domaine de la santé (noms confidentiels) pour le Ministère de l'Enseignement Supérieur et de la Recherche
- **Depuis 2011** Rédaction d'une trentaine d'**expertises CIR Crédit Impôt Recherche** et **JEI Jeune entreprise Innovante** pour des PME du domaine de la santé (noms confidentiels) pour diverses délégations du Ministère de l'Enseignement Supérieur et de la Recherche

PUBLICATIONS

Je publie dans le domaine de l'**Extraction de connaissances** et dans le domaine de l'**Informatique médicale**.

La plupart des publications sont indexées par :

- DBLP Computer Science Bibliography (55 entrées le 27/12/2014) : <http://www.informatik.uni-trier.de/~ley/pers/hd/b/Bringay:Sandra.html>
- PUBMED (11 entrées le 27/12/2014) : <http://www.ncbi.nlm.nih.gov/pubmed?term=bringay>

Mes publications sont listées ci-dessous par année de parution. Elles sont classées selon les catégories Publication internationales avec comité de lecture et Publication nationales avec comité de lecture. Dans ces deux catégories, je sépare les publications liées à l'extraction de connaissances et celles liées à l'informatique médicale. Je distingue également les articles publiés dans les ouvrages numérotés avec un préfixe J et les articles dans les actes des conférences numérotés à l'aide du préfixe C.

Les impacts factor recensés dans cette liste sont issus des sites des journaux et mis à jour le 27/12/2014. Le rang des conférences est issu du site CORE (<http://103.1.187.206/core>) et mis à jour le 27/12/2014.

La plupart des articles ont été rédigés avec les doctorants que je co-encadre. La règle pour l'ordre des noms est la suivante : le doctorant en premier et les encadrants ou collaborateurs par ordre alphabétique ensuite. Parfois, les experts médicaux nous demandent également d'apparaître en dernier.

- [J1] Entrepôts de Données et Analyse en ligne. Rédactrices invitées : S. Bringay, A. Laurent, M. Teisseire. Numéro spécial de la Revue des nouvelles technologies de l'information (RNTI). Ed. Cépaduès. 2010

PUBLICATIONS INTERNATIONALES AVEC COMITÉ DE LECTURE

• **Extraction de connaissances**

- [J2] M. Fabrègue, A. Braud, S. Bringay, F. Le Ber, M. Teisseire: Mining closed partially ordered patterns, a new optimized algorithm. *Knowl.-Based Syst.* 79: 68-79 (2015) (**Impact Factor : 3.058**)
- [C1] P. Accorsi, M. Fabrègue, A. Sallaberry, F. Cernesson, N. Lalande, A. Braud, S. Bringay, F. Le Ber, P. Poncelet, M. Teisseire. *HydroQual: Visual Analysis of River Water Quality*. Proceedings VAST 2014: 123-132 (**Rang C**)
- [C2] S. Bringay, E. Kergosien, P. Pompidor, P. Poncelet: Identifying the Targets of the Emotions Expressed in Health Forums. Proceedings CICLing (2), Kathmandu (Nepal), 2014: 85-97 (**Rang B**)
- [C3] A. Abdaoui, J. Azé, S. Bringay, N. Grabar, P. Poncelet: Predicting Medical Roles in Online Health Fora. Proceedings SLSP, Grenoble (France), 2014: 247-258
- [C4] A. Abboute, Y. Boudjeriou, G. Entringer, J. Azé, S. Bringay, P. Poncelet: Mining Twitter for Suicide Prevention. Proceedings NLDB, Montpellier (France), 2014: 250-253 (**Rang C**)
- [J3] H. Alatrística-Salas, J. Azé, S. Bringay, F. Cernesson, N. Selmaoui-Folcher, M. Teisseire, A Knowledge Discovery Process for Spatiotemporal Data: Application to River Water Quality Monitoring. *Ecological Informatics*, Elsevier Ed., 2014 (**Impact Factor : 1,980**)
- [C5] J. Pasquet, S. Bringay, M. Chaumont: Steganalysis with cover-source mismatch and a small learning database. Proceedings EUSIPCO, Lisbon (Portugal), 2014: 2425-2429 (**Rang B**)
- [C6] J. Rabatel, S. Bringay, P. Poncelet: Mining Representative Frequent Patterns in a Hierarchy of Contexts. Proceedings IDA 2014, Lieuvén (Belgium): 239-250 (**Rang A**)
- [J4] M. Fabrègue, A. Braud, S. Bringay, C. Grac, F. Le Ber, D. Levet, M. Teisseire: Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. *Ecological Informatics* 24 (2014): 210-221 (**Impact Factor : 1,980**)
- [J5] H. Alatrística Salas, F. Flouvat, S. Bringay, N. Selmaoui-Folcher, M. Teisseire: A spatial-based KDD process to better manage the river water quality. *Revue Internationale de Géomatique* 23(3-4), 2013: 469-494
- [C7] A. Zine El Abidine, A. Sallaberry, S. Bringay, M. Fabrègue, C. Lecellier, N. H. Phan, P. Poncelet. *Co2Vis: A visual analytics tool for mining co-expressed and co-regulated genes implied in HIV infections*. Poster BioVis 2013, Atlanta (USA), 2013
- [C8] M. Fabrègue, Agnès Braud, Sandra Bringay, Florence Le Ber, Maguelonne Teisseire: OrderSpan: Mining Closed Partially Ordered Patterns. Proceedings IDA 2013, Londres (England): 186-197 (**Rang A**)
- [C9] S. Tahrat, E. Kergosien, S. Bringay, M. Roche, M. Teisseire: Text2Geo: from textual data to geospatial information. Proceedings WIMS, Madrid (Spain) 2013: 23
- [C10] M. Fabrègue, A. Braud, S. Bringay, F. Le Ber, M. Teisseire. Including Spatial Relations and Scales within Sequential Pattern Extraction. Proceedings Discovery Science, Lyon (France), 2012: 209-223 (**Rang C**)
- [C11] F. Bouillot, H. Phan Nhat, N. Béchet, S. Bringay, D. Ienco, S. Matwin, P. Poncelet, M. Roche, M. Teisseire: How to Extract Relevant Knowledge from Tweets? Springer CCSI (Communications in Computer and Information Science), post proceedings of International Workshop ISIP 2013, Bangkok, (Thailand): 111-120
- [C12] H. Alatrística-Salas, S. Bringay, F. Flouvat, N. Selmaoui-Folcher and M. Teisseire. The Pattern Next Door: Towards Spatio-sequential Pattern Discovery. Proceedings PAKDD 2012, Kuala Lumpur (Malaysia) (2): 157-168 (**Rang A**)
- [C13] H. Alatrística Salas, J. Azé, S. Bringay, F. Cernesson, F. Flouvat, N. Semaloui et M. Teisseire. Finding Relevant Sequences With The Least Temporal Contradiction Measure: Application to Hydrological Data. Proceedings AGILE 2012, Avignon (France): 197-202.
- [J6] J. Rabatel, S. Bringay and P. Poncelet. Mining Sequential Patterns: a Context-Aware Approach. In *Advances in Knowledge Discovery and Management*, Springer, Vol. 3 (AKDM-3) 2012: 23-41
- [C14] S. Bringay, N. Béchet, F. Bouillot, P. Poncelet, M. Roche, M. Teisseire. Towards an On-Line Analysis of Tweets Processing. DEXA, Toulouse (France) (2) 2011: 154-161 (**Rang B**)
- [J7] J. Rabatel, S. Bringay and P. Poncelet. Anomaly Detection in Monitoring Sensor Data for Preventive Maintenance. *Journal Expert Systems with Applications*, 38, 2010: 7003-7015 (**Impact Factor : 2,908**)

- [C15] J. Rabatel, S. Bringay and P. Poncelet. Contextual Sequential Pattern Mining. Proceeding of the Workshop DDDM2010 in conjunction with ICDM'2010, Sydney (Australia), 2010, 8 pages
- [C16] A. Sallaberry, N. Pecheur, S. Bringay, M. Roche and M. Teisseire. Discovering Novelty in Gene Data: From Sequential Patterns to Visualization. Proceedings ISVC10, Las Vegas (USA). G. Bebis et al. (Eds.), Part III, LNCS 6455, Springer-Verlag Berlin Heidelberg, 2010: 534-543 (**Rang C**)
- [C17] J. Rabatel, S. Bringay and P. Poncelet. Fuzzy Anomaly Detection in Monitoring Sensor Data. Proceedings FUZZ-IEEE 2010, Barcelona (Spain), July 18-23, 2010: 1-8 (**Rang A**)
- [C18] S. Bringay, A. Laurent, B. Orsetti, P. Salle and M. Teisseire. Handling Fuzzy Gaps in Sequential Patterns: Application to Health. Proceedings FUZZ-IEEE 2009, Jeju (Korea): 1338-1345 (**Rang A**)
- [C19] J. Rabatel, S. Bringay and P. Poncelet. SO_MAD: SensOr Mining for Anomaly Detection in Railway Data. Proceeding ICDM, July 20 - 22, Leipzig (Germany), 2009: 191-205 (**Best paper selection**)
- [C20] L. Di Jorio, S. Bringay, C. Fiot, A. Laurent and M. Teisseire. Sequential Patterns for maintaining ontologies over time. Proceedings Ontologies, DataBases, and Applications of Semantics OTM Conferences, Monterrey (Mexico), 2, 2008: 1385-1403
- [C21] H. Saneifar, S. Bringay and A. Laurent. S2MP: Similarity Measure for Sequential Patterns. Proceeding AusDM 2008, Adelaide (Australia), 2008: 95-104

- **Informatique médicale**

- [J8] L. Faisandier, A. Fouillet, DJ. Bicout, F. Golliot, I. Ahmed, S. Bringay, D. Eilstein: Surveillance et détection des événements inhabituels en toxicovigilance : revue des méthodes pertinentes. Rev Epidemiol Sante Publique. 2015 Apr;63(2):119-131 (**Impact Factor : 0,656**)
- [C22] T. Opitz, J. Azé, S. Bringay, C. Joutard, C. Lavergne, C. Mollevi: Breast Cancer and Quality of Life: Medical Information Extraction from Health Forums. Proceedings MIE 2014, Istanbul (Turkish), 2014, Studies in Health Technology and Informatics: 1070-1074
- [C23] A. Abdaoui, J. Azé, S. Bringay, N. Grabar, P. Poncelet: Analysis of Forum Posts Written by Patients and Health Professionals. Poster MIE 2014, Istanbul (Turkish), Studies in Health Technology and Informatics: 1185
- [C24] S. Melzi, A. Abdaoui, J. Azé, S. Bringay, P. Poncelet and F. Galtier. Patient's rationale: Patient Knowledge retrieval from health forums'. Proceedings ETELEMED 8th International Conference on eHealth, Telemedicine, and Social Medicine, 2014, Barcelona (Spain), 2014: 140-145
- [J9] C. Flamand, M. Fabregue, S. Bringay, V. Ardillon, P. Quénel, JC. Desenclos, M. Teisseire. Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana. Journal of American Medical Informatics Association. 2014 Oct;21(e2), 2014: 232-240 (**Impact Factor : 3,932**)
- [C25] D. Breton, S. Bringay, F. Marques, P. Poncelet and M. Roche. Epimining: Using Web News for Influenza Surveillance. Proceedings Workshop on Data Mining for Healthcare Management (DMHM 2012) in conjunction with PAKDD 2012, Kuala Lumpur (Malaysia), 2012: 11-21
- [J10] A. Sallaberry, N. Pecheur, S. Bringay, M. Roche, M. Teisseire. Sequential patterns mining and gene sequence visualization to discover novelty from microarray data. Journal of Biomedical Informatics 44(5), 2011: 760-774 (**Impact Factor : 2,817**)
- [J11] M. Fabregue, S. Bringay, P. Poncelet, M. Teisseire and B. Orsetti. Mining microarray data to predict the histological grade of a Breast Cancer. Journal of Biomedical Informatics, 2011 Dec;44 Suppl 1:S12-6 (**Impact Factor : 2,817**)
- [C26] C. Flamand, P. Quenel, V. Ardillon, L. Carvalho, S. Bringay and M. Teisseire. The Epidemiologic Surveillance of Dengue-Fever in French Guyana: When achievements trigger higher goals. Proceedings MIE 2011, Oslo (Norway), Studies in Health Technology and Informatics. 2011;169: 629-33
- [C27] S. Bringay, M. Roche, M. Teisseire, P. Poncelet, R. Abdel Rassoul, JM. Verdier and G. Devau. Discovering novelty in sequential patterns: application for analysis of microarray data on Alzheimer disease. Proceeding MedInfo'2010, Cape town (South Africa). Studies in Health Technology and Informatics. 2010;160(Pt 2), 2010:1314-8 (**Rang B**)
- [C28] A. Sallaberry, N. Pecheur, S. Bringay, M. Roche and M. Teisseire. SequencesViewer: Visualization of genes sequences. Demonstration MedInfo'2010, Cape town (South Africa), September 13-16 (**Rang B**)
- [C29] J. Nin, P. Salle, S. Bringay and M. Teisseire. Using OWA Operators for Gene Sequential Pattern Clustering. Proceeding CBMS'09., Albuquerque (USA), 2009 : 1-4
- [C30] P. Salle, S. Bringay and M. Teisseire. Mining Discriminant Sequential Patterns for Aging Brain. Proceeding AIME'09. 18-22 July, Verona, (Italy), 2009: 365-369 (**Rang A**)

- [C31] P. Salle, S. Bringay, M. Teisseire, F. Chakkour, M. Roche, G. Devau, C. Lautier and JM. Verdier. GeneMining: Identification, Visualization, and Interpretation of Brain Ageing Signatures. Proceedings MIE'2009, Sarajevo (Bosnie-Herzégovine), 2009. Studies in Health Technology and Informatics: 767-771
- [J12] N. Bricon-Souf, S. Bringay, F. Anceaux, S. Hamek, N. Degardin, C. Barry and J. Charlet. Informal Notes to support the Asynchronous Collaborative Activities. International Journal of Medical Information. 76 Suppl 3. Epub 2007 Apr 23, 2007: S342-8 (**Impact Factor : 3.214**)
- [C32] S. Bringay, C. Barry, J. Charlet and G. Krim. Proceedings MedInfo 2007, Brisbane (Australia): 2607-2609 (**Rang B**)
- [C33] S. Bringay, C. Barry and J. Charlet. Annotations for the collaboration of the health professionals. Proceedings AMIA 2006: 91-5 (**Rang A**)
- [C34] N. Bricon-Souf, S. Bringay, F. Anceaux, S. Hamek, N. Degardin, C. Barry and J. Charlet. A study of the communication notes for two asynchronous collaborative activities. Studies in Health Technology and Informatics 2006;124: 713-8
- [C35] S. Bringay, C. Barry and J. Charlet. Annotations: A Functionality to support Cooperation, Coordination and Awareness in the Electronic Medical Record. Proceedings COOP'2006, Carry-le-Rouet (France): 39-54
- [C36] S. Bringay, C. Barry, J. Charlet, G. Krim. Annotations for sharing and managing knowledge in the Electronic Health Record. Proceedings sort paper and Poster, MIE 2005, Genève (Suisse) (**Prix du meilleur poster**)
- [C37] S. Bringay, C. Barry, J. Charlet. A specific tool of Annotations for the Electronic Health Record. Proceedings International workshop IWAC 2005, Paris (France): 21-26
- [C38] S. Bringay, C. Barry, J. Charlet. Annotations for sharing and managing knowledge in the Electronic Health Record. Proceedings International Workshop on Knowledge Management and Organizational Memories in conjunction with IJCAI 2005, Edinburgh, (Ecosse): 11-23
- [C39] S. Bringay, C. Barry, J. Charlet. Annotations: A new type of document in the Electronic Health Record. Proceedings DOCAM Document Academy, 2004, San Francisco (USA), 15 pages
- [C40] S. Bringay, C. Barry, J. Charlet. The Health Record: Kernel of a Medical Memory. Proceedings International Workshop on Knowledge Management and Organizational Memories in conjunction with ECAI 2004, Valencia (Espagne): 18-32

PUBLICATIONS NATIONALES AVEC COMITE DE LECTURE

- **Extraction de connaissances**

- [C41] S. Bringay, E. Kergosien, P. Pompidor et P. Poncelet. Identifier la cible des émotions dans les forums de santé. Actes IC 2014, Clermont-Ferrand (France): 163-174
- [C42] S. Melzi, A. Abdaoui, J. Azé, S. Bringay, P. Poncelet, Florence Galtier: Que ressentent les patients ? Actes EGC 2014, Rennes (France): 449-454
- [J13] N. Selmaoui-Folcher, F. Flouvat, H. Alatrística Salas, S. Bringay: Motifs spatio-temporels. Enjeux et applications à l'environnement. Revue d'Intelligence Artificielle 27(4-5): 619-648 (2013)
- [C43] S. Tahrat, E. Kergosien, S. Bringay, M. Roche, Maguelonne Teisseire: Text2Geo : des données textuelles aux informations géospatiales. Actes EGC 2013, Toulouse (France): 407-412
- [C44] M. Fabrègue, A. Braud, S. Bringay, F. Le Ber, C. Lecellier, P. Poncelet and M. Teisseire. OrderGeneMiner : Logiciel pour l'extraction et la visualisation de motifs partiellement ordonnés à partir de puces à ADN. Démonstration EGC 2013, Toulouse (France)
- [C45] M. Fabrègue, A. Braud, S. Bringay, F. Le Ber, M. Teisseire. Extraction de motifs spatio-temporels à différentes échelles avec gestion de relations spatiales qualitatives. Inforsid 2012, Montpellier (France): 123-140
- [C46] H. Alatrística-Salas, S. Bringay, F. Flouvat, N. Selmaoui-Folcher and M. Teisseire. Vers une approche efficace d'extraction de motifs spatio-séquentiels. Extraction et gestion des connaissances. Actes EGC 2012, Bordeaux (France), Revue des Nouvelles Technologies de l'Information, Volume: RNTI-E-23, 2012: 201-212
- [C47] S. Bringay, N. Béchet, F. Bouillot, P. Poncelet, M. Roche and M. Teisseire. Analyse de gazouillis en ligne. Actes EDA 2011, Clermont Ferrand, (France): 87-102
- [J14] S. Bringay, A. Laurent, M. Teisseire. Éditorial. Technique et Science Informatiques 30(8), 2011: 931-932
- [J15] H. Alatrística Salas, J. Azé, S. Bringay, F. Cernesson, F. Flouvat, N. Semaloui et M. Teisseire. Recherche de séquences spatio-temporelles peu contredites dans des données hydrologiques. Revue des Nouvelles

Technologies de l'Information (RNTI), numéro spécial Qualité des Données et des connaissances/Evaluation des Méthodes d'Extraction des Connaissances dans les Données. Volume: RNTI-E-22, 2011: 165-188

- [C48] J. Rabatel, S. Bringay. Extraction de motifs séquentiels contextuels. Actes EGC 2011, Brest (France): 11-22 (nominé pour les meilleurs papiers)
- [C49] B. Rosoor, L. Sebag, S. Bringay, M. Roche. A la recherche des tweets porteurs d'informations journalistiques. Démonstration EGC 2011, Brest (France): 283-286
- [C50] B. Rosoor, L. Sebag, S. Bringay, P. Poncelet and M. Roche. Quand un tweet détecte une catastrophe naturelle... Actes VSST'2010, Toulouse (France), 2010: 15 pages
- [C51] S. Bringay, M. Teisseire, J. Gomila, D. Hoffschir and T. Vicaire. Les motifs séquentiels au service de la structuration des folksonomies. Actes IC'2010, Nimes (France), 2010: 133-144
- [C52] J. Rabatel, S. Bringay and P. Poncelet. Aide à la décision pour la maintenance ferroviaire préventive. Actes EGC 2010, Hammamet (Tunisie), 2010: 363-368
- [C53] A. Sallaberry, N. Pecheur, S. Bringay, M. Roche and M. Teisseire. SequencesViewer : Visualisation de séquences ordonnées de gènes ou comment rendre accessible des motifs séquentiels trop nombreux ? Actes EGC 2010, Hammamet (Tunisie), 2010: 387-392
- [C54] T. Bouadi, S. Bringay, P. Poncelet and M. Teisseire. Requêtes Skyline avec prise en compte des préférences utilisateurs pour des données volumineuses. Actes EGC 2010, Hammamet (Tunisie), 2010: 399-404
- [C55] S. Bringay, A. Laurent, P. Poncelet, M. Roche and M. Teisseire. Bien cube, les données textuelles peuvent s'agréger ! Actes EGC 2010, Hammamet (Tunisie), 2010 (nominé pour les meilleurs papiers): 585-596
- [C56] P. Salle, S. Bringay et M. Teisseire. DEMON : DEcouverte de MOTifs séquentiels pour les puces adN. Actes EGC 2009, Strasbourg (France): 459-460
- [C57] P. Salle, S. Bringay and M. Teisseire. Motifs Séquentiels Discriminants pour les puces ADN. Actes INFORSID 2009, Toulouse (France), 2009: 397-412
- [C58] P. Salle, S. Bringay, A. Laurent and M. Teisseire. Motifs séquentiels et écarts flous. Actes LFA'09, Annecy (France), 2009: 41-48
- [C59] W. Xing, P. Salle, S. Bringay and M. Teisseire. DEMON-Visualisation : un outil pour la visualisation des motifs séquentiels extraits à partir de données biologiques. Démonstration EGC 2009, Strasbourg (France): 491S
- [C60] J. Rabatel, S. Bringay, P. Poncelet and M. Teisseire. Aide au diagnostic de pannes guidée par l'extraction de motifs séquentiels. Revue des nouvelles technologies de l'information 2009, RNTI-E-18, 87-112
- [C61] H. Saneifar, S. Bringay, A. Laurent and M. Teisseire. S²MP : Une mesure de similarité pour les motifs séquentiels. Actes de l'atelier EvalECD'09 associé à EGC 2009 (2009), Strasbourg (France), 2009: 35-46
- [J16] S. Bringay, C. Barry and J. Charlet. Un modèle pour les annotations du dossier patient informatisé. Salembier et M. Zacklad (eds) : Annotations dans les documents pour l'action, Hermes Publishing, Londres-Paris 2007: 47-67
- [J17] M. Zacklad et al., 2006. Processus d'annotation dans les documents pour l'action : textualité et médiation de la coopération, in J.Charlet (Ed), Documents et Contenu : création, indexation, navigation, Hermès, Lavoisier, Paris, 25 pages
- [C62] S. Bringay, C. Barry, J. Charlet. Les annotations pour gérer les connaissances du dossier patient. Actes IC 2005, Nice (France): 73-84

• Informatique médicale

- [C63] T. Opitz, S. Bringay, J. Azé, C. Joutard, C. Lavergne and C. Mollevi. Paroles de patients dans les forums de santé: une perspective originale sur la qualité de la vie. Atelier Ingénierie des connaissances et santé, Clermont Ferrand (France): 6 pages
- [C64] A. Abdaoui, J. Azé, S. Bringay, P. Poncelet and N. Grabar. Analyse des messages des patients et des médecins dans les fora de santé. Atelier Ingénierie des connaissances et santé, Clermont Ferrand (France): 6 pages
- [J18] L. Di Jorio, S. Bringay, D. Brouillet, A. Laurent, S. Martin and M. Teisseire. Fouille de données issues d'études psychologiques liées au vieillissement : extraction de règles graduelles. Revue des sciences et technologies de l'information - TSI - 29/2010. Interface STIC-SHS: 939-957

- [C65] P. Salle, S. Bringay, M. Teisseire and G. Devau. Recherche de signatures pour la maladie d'Alzheimer : extraction de motifs séquentiels discriminants à partir de puces. Poster JFIM 2009, Nice (France), 2009
- [C66] G. Devau, P. Salle, R. Abdel Rassoul, S. Bringay, S. Alves, C. Lautier, N. Mestre-Francès, M. Teisseire and J.M. Verdier. Identification of gene expression changes in the brain of microcebus murinus during aging by using data mining. Poster 9ème colloque Société des Neurosciences, Bordeaux (France), 2009
- [C67] S. Bringay, N. Bricon-Souf, F. Anceaux, S. Hamek, C. Barry and J. Charlet. Un modèle des stratégies d'écriture supportant deux activités collaboratives asynchrones. Actes JFIM'2007, 2007
- [C68] S. Bringay, C. Barry, J. Charlet, G. Krim. Une fonctionnalité d'annotation pour le dossier patient informatisé. Actes JFIM'2005, Lille (France), 7 pages
- [C69] S. Bringay, C. Barry, J. Charlet. Annotations dans le cadre du Dossier Patient Hospitalier. Actes WSM 2004, Rouen (France)
- [J19] S. Bringay, C. Barry, J. Charlet. Les documents et les annotations dans le dossier patient hospitalier. Numéro Spécial de la revue I3 Information, Interaction, Intelligence. Vol. 4, Num. 1, 2004: 191-211

SYNTHESE QUANTITATIVE DES PUBLICATIONS

Dans le Tableau 1, je distingue les publications dans des conférences ou revues dont les thèmes portent sur l'extraction de connaissances (EC) ou sur l'informatique médicale (IM), mais bien souvent les thématiques se recouvrent. Montrer leur étroite imbrication est d'ailleurs l'objectif que je poursuis depuis mon arrivée à Montpellier. Je distingue également les publications avant et après mon recrutement à l'Université Paul-Valéry Montpellier. Dans le Tableau 2, j'ai reporté uniquement les publications dans les revues nationales et internationales et les conférences internationales de rang A*, A, B et C.

Tableau 1: Synthèse quantitative. Légende (EC Extraction de connaissance, IM Informatique médicale)

		Éditions ouvrage int.	Éditions ouvrage nat.	Revue int.	Revue nat.	Actes conf. int. avec comité de sélection.					Actes conf. nat. avec com	Workshop int.	Workshop nat.	Poster/Demo int	Poster/Demo Nat.	Autre	TOTAL
						A*	A	B	C	/							
2015	EC	1														0	
	IM	1														0	
2014	EC	2		1		2	1	1	1	2	2		2		2		11
	IM	1								2	2						5
2013	EC	1		1	1				2	1	1			1		8	
	IM															0	
2012	EC	1				1	1	1	2							6	
	IM															1	
2011	EC			2		1				3						6	
	IM	2						1								3	
2010	EC	1				1	1			6	1					10	
	IM			1	1								1				3
2009	EC	1		1	2				2	3	1		1				11
	IM															3	
2008	EC															2	
2007	IM	1		2	1				1								5
TOTAL		0	1	9	8	0	6	5	3	11	18	4	3	3	4	0	75
2006	IM	1														3	
2005	IM															5	
2004	IM	1														4	
TOTAL		0	1	9	9	0	7	5	3	11	22	8	4	4	4	0	87

SYNTHESE QUALITATIVES DES PUBLICATIONS

Tableau 2: Synthèse qualitative

	Éditions ouvrage nat.	Revue int/Chapitre de livre int.	Revue nat. /Chapitre de livre nat.	Actes conf. int. avec com. Sélec.			
				A *	A	B	C
2015		Knowl.-Based Syst Revue épidém. et santé pub.					
2014		2 Ecolo. Info. (IF 1,9) JAMIA (IF 3,932)		IDA	CICLING EUSIPCO	VAST NLDB	
2013		Revue Int. Geom.	Revue IA	IDA			
2012		AKDM	RNTI-E-23	PAKDD	Disco. sciences		
2011		2 J. Bio Info., (IF 2.817)	Ed. TSI RNTI-E-22		DEXA		
2010		J. Expert System (IF 2.908)		FUZZ-IEEE	MedInfo	ISVC	
2009	RNTI		RNTI-E-18	FUZZ-IEEE AIME			
2008							
2007		Int. J. of Med. Info. (IF 3.214)			MedInfo		
2006			Doc. & cont.	AMIA			
2005							
2004			RNTI-E-4				

PARTIE II

SÉLECTION DE PUBLICATIONS

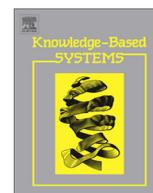
LISTE DES ARTICLES

■ Fouille de données

- Fabrègue M., Braud A., Bringay S., Le Ber F., Teisseire M. : Mining closed partially ordered patterns, a new optimized algorithm. *Knowl.-Based Syst.* 79 : 68-79 (2015) (**Impact Factor : 3.058**)

■ Santé

- Flamand C, Fabregue M, Bringay S, Ardillon V, Quénel P, Desenclos JC, Teisseire M. Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana. *J Am Med Inform Assoc.* 2014 Oct ;21(e2) :e232-40. (**Impact Factor : 3.932**)
- Fabregue M, Bringay S, Poncelet P, Teisseire M, Orsetti B. Mining microarray data to predict the histological grade of a breast cancer. *J Biomed Inform.* 2011. (**Impact Factor : 2.482**)



Mining closed partially ordered patterns, a new optimized algorithm



Mickaël Fabrègue^{a,b,*}, Agnès Braud^c, Sandra Bringay^d, Florence Le Ber^a, Maguelonne Teisseire^b

^a ICube, University of Strasbourg/ENGEEES, CNRS, Illkirch, France

^b TETIS, IRSTEA, Montpellier, France

^c ICube, University of Strasbourg, CNRS, Illkirch, France

^d LIRMM, Montpellier 3 University, CNRS, France

ARTICLE INFO

Article history:

Received 24 April 2014

Received in revised form 20 December 2014

Accepted 25 December 2014

Available online 17 January 2015

Keywords:

Data mining

Sequential patterns

Partially ordered patterns

ABSTRACT

Nowadays, sequence databases are available in several domains with increasing sizes. Exploring such databases with new pattern mining approaches involving new data structures is thus important. This paper investigates this data mining challenge by presenting *OrderSpan*, an algorithm that is able to extract a set of closed partially ordered patterns from a sequence database. It combines well-known properties of prefixes and suffixes. Furthermore, we extend *OrderSpan* by adapting efficient optimizations used in sequential pattern mining domain. Indeed, the proposed method is flexible and follows the sequential pattern paradigm. It is more efficient in the search space exploration, as it skips redundant branches. Experiments were performed on different real datasets to show (1) the effectiveness of the optimized approach and (2) the benefit of closed partially ordered patterns with respect to closed sequential patterns.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Due to the exponential growth of temporal and spatiotemporal databases, sequential pattern mining has become a very active research area. Many studies have demonstrated the usefulness of such patterns for analysis [1], classification [2,3] or prediction [4]. These patterns were introduced in [5] and are an extension of association rules [6]. Several algorithms to mine such patterns have been proposed and are presented in [7,8]. They are used when information is totally ordered according to a specific criterion, which is usually temporal. For instance, let us take the well-known “market basket” problem. We consider a customer database where the pattern $\langle(Bread)(Chocolate)\rangle$ is found. This means that the product *Bread* is frequently purchased before the product *Chocolate*. Mining such related items according to temporal aspects is very useful for specialists in various domains such as marketing [9], software engineering [10] or medicine [11]. Despite their advantages, sequential patterns often generate little information since they only provide totally ordered information about data. For example, let us consider a second pattern, $\langle(Bread)(Milk)\rangle$, discovered in the same database. If these two patterns describe the same

customers, their coexistence is not taken into account with sequential pattern approaches. However, they can be synthesized via partial ordering.

Fig. 1 presents a so-called partially ordered pattern that combines the two previous sequential patterns. This new pattern means that customers frequently purchase the product *Bread* before purchasing the two other products *Chocolate* and *Milk* which themselves are not ordered. Partially ordered patterns can be used in sequential databases and have many advantages: (1) they provide more information on order among elements; (2) they are represented as a directed acyclic graph, which facilitates the understanding; and (3) they summarize sequential pattern sets.

In a previous paper [12], we presented a method designed to directly extract closed partially ordered patterns in the general case of itemset sequences with item repetitions. In the present paper, we propose an improvement of our algorithm:

- Based on the property presented in [13], we present an optimized version of the so-called *OrderSpan* algorithm that explores the search space and outputs the complete set of closed partially ordered patterns.
- We use a new data structure to represent patterns in the algorithm. Properties of this new data structure lead to a different and generic way to remove the redundancy in patterns.
- We provide a complexity analysis of the approach and an upper bound on the number of extracted patterns given a minimum support.

* Corresponding author at: ICube, University of Strasbourg/ENGEEES, CNRS, Illkirch, France.

E-mail addresses: mickael.fabregue@teledetection.fr (M. Fabrègue), agnes.braud@unistra.fr (A. Braud), sandra.bringay@lirmm.fr (S. Bringay), florence.leber@engees.unistra.fr (F. Le Ber), maguelonne.teisseire@teledetection.fr (M. Teisseire).

The method proposed in [12] is close to the non-optimized algorithm presented in this paper. As we will see, the main difference is the use of an expanded data-structure that easily allows the addition of effective optimizations during the process. We thus integrated optimizations from [13].

Based on sequential pattern mining work, *OrderSpan* extracts partially ordered patterns based on the prefix and suffix properties of sequences. We opted to extract closed partially ordered patterns because they provide a compact representation of all partially ordered patterns. Thus, the output result set is smaller and it is possible to retrieve the complete set of all partially ordered patterns. There is no information loss. Our approach follows the *Pattern-Growth* paradigm on sequences, thus it is related to other approaches in sequential pattern mining. Some of these methods [13,14] are optimized to explore the search space of closed sequential patterns in a very efficient way. These optimizations are performed according to some properties that help to prune the search space to reduce its exploration. Thus, we analyzed closed sequential pattern properties that can be applied to the problem of mining closed partially ordered patterns. We adapted the optimization based on the equivalence databases proposed in *CloSpan* [13]. This property efficiently prunes the search space in the case of sequential pattern mining. We generalized it to the sub-search space that corresponds to a closed partially ordered pattern.

This paper is organized as follows. Section 2 gives some preliminary definitions on sequences and partially ordered patterns. Section 3 describes existing studies on partially ordered pattern mining. Section 4 introduces the *OrderSpan* algorithm including an optimization step and complexity analysis. Experimental results are presented in Section 5. Firstly, we compare the non-optimized and the optimized algorithm on a set of examples. Secondly, we compare the optimized version of *OrderSpan* with the algorithm proposed in [15]. Finally, we study the semantic aspects of closed partially ordered patterns.

2. Problem definition

Before presenting the partially ordered pattern concept, we provide some important definitions relative to closed sequential pattern mining. As we will see later, a partially ordered pattern is a more complex structure composed of closed sequential patterns. Let us first define a sequence (Definition 1), sub-sequence (Definition 2), a sequential pattern (Definition 3) and a closed sequential pattern (Definition 4).

Definition 1 (Sequence)

Let $\mathcal{I} = \{I_1, I_2, \dots, I_m\}$ be a set of **items**. An **itemset** IS is a non empty, unordered, set of **items** denoted $(I_{j_1} \dots I_{j_k})$ where $I_{j_i} \in \mathcal{I}$. Let \mathcal{IS} be the set of all **itemsets** built from \mathcal{I} . A **sequence** S is a non-empty ordered list of **itemsets** denoted $\langle IS_1 IS_2 \dots IS_p \rangle$ where $IS_j \in \mathcal{IS}$.

Definition 2 (Sub-sequence)

A **sequence** $S_\alpha = \langle IS_1 IS_2 \dots IS_p \rangle$ is a **sub-sequence** of another **sequence** $S_\beta = \langle IS'_1 IS'_2 \dots IS'_m \rangle$, denoted $S_\alpha \preceq_s S_\beta$, if $p \leq m$ and if there are integers $j_1 < j_2 < \dots < j_k < \dots < j_p$ such that $IS_1 \subseteq IS'_{j_1}$, $IS_2 \subseteq IS'_{j_2}, \dots, IS_p \subseteq IS'_{j_p}$.

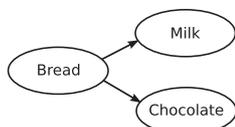


Fig. 1. Example of partially ordered pattern for the “market basket” problem.

Definition 3 (Sequential pattern)

Let S_p be a **sequence** and \mathcal{DB} a sequence database. Let $S' \subseteq \mathcal{DB}$ be the maximal set of **sequences** such that $\forall S_i \in S', S_p \preceq_s S_i$. $|S'|$ is called the **support** of S_p . S_p is called a **sequential pattern**, denoted **seq-pattern**, when $Support(S_p) \geq \theta$ where θ is a given value (minimum support).

Definition 4 (Closed sequential pattern)

Let S_p be a **sequential pattern**, S_p is a **closed sequential pattern** if there is no other **sequential pattern** S'_p such that $S_p \preceq_s S'_p$ and $Support(S_p) = Support(S'_p)$.

In the following, the database in Table 1 is used to illustrate these definitions. This database contains three sequences of itemsets based on the alphabet $\Sigma = \{a, c, d, e, f, g\}$. Given this database, the sub-sequences $\langle (g) \rangle$, $\langle (d) \rangle$ and $\langle (g)(d) \rangle$ are supported by sequences S_1, S_2 and S_3 , and their support is equal to 3. Thus with a minimum support $\theta = 2$, these sub-sequences are seq-patterns because $3 \geq \theta$. But sequences $\langle (g) \rangle$ and $\langle (d) \rangle$ are not closed seq-patterns since the sequence $\langle (g)(d) \rangle$ is such that $\langle (g) \rangle \preceq_s \langle (g)(d) \rangle$ and $\langle (d) \rangle \preceq_s \langle (g)(d) \rangle$ with an equivalent support. Finally, the sequences $\langle (g)(d) \rangle$, $\langle (cd)(a) \rangle$, $\langle (cd)(g) \rangle$, $\langle (g)(d)(e) \rangle$, $\langle (g)(d)(f) \rangle$ and $\langle (g)(e)(f) \rangle$ in Table 2, give the complete set of closed seq-patterns for $\theta = 2$.

These are given with the associated set of supporting sequences. Some sets of closed seq-patterns are supported by the same set of sequences. For instance $\langle (g)(d)(e) \rangle$, $\langle (g)(d)(f) \rangle$ and $\langle (g)(e)(f) \rangle$ are supported by S_2 and S_3 . A partial order can be used to obtain a synthetic representation of these closed seq-patterns relative to the sequence set $\{S_2, S_3\}$. We can define as many partial orders as there are corresponding sets of sequences. In our example, these sets are $\{S_1, S_2, S_3\}$, $\{S_1, S_2\}$ and $\{S_2, S_3\}$. Fig. 2a–c give the three partial orders corresponding to each set of sequences. We use two vertices labeled “(” and “)”, representing the beginning and end of the patterns.

These structures are used to represent partially ordered patterns which are defined in the following:

Definition 5 (Partially ordered sequence)

A **partially ordered sequence** is a set of itemsets with a partial order $(\mathcal{V}, <)$. It can be represented with a **labeled directed acyclic graph** $G = (\mathcal{V}, \mathcal{A}, \Sigma_{\mathcal{V}}, l_{\mathcal{V}})$ where:

- \mathcal{V} is the set of **vertices** and \mathcal{A} is the set of **arcs** where $\mathcal{A} = \{(u, v) \in <, \text{with } u, v \in \mathcal{V}\}$
- $\Sigma_{\mathcal{V}}$ is a finite alphabet representing possible vertex label values
- $l_{\mathcal{V}} : \mathcal{V} \rightarrow \Sigma_{\mathcal{V}}$ is a mapping giving the labeling on the **vertices**

In the graph, for all $u, v \in \mathcal{V}, u < v$ if there is a directed path from u to v . However, if there is no path from u to v or from v to u ,

Table 1
An example of a sequence database.

Seq id	Sequence
S_1	$\langle (cd)(a)(g)(d) \rangle$
S_2	$\langle (g)(cde)(f)(aeg) \rangle$
S_3	$\langle (g)(d)(e)(f) \rangle$

Table 2
Set of closed sequential patterns related to Table 1 with the minimum support $\theta = 2$.

Supporting seq set	Sequence
$\{S_1, S_2, S_3\}$	$\langle (g)(d) \rangle$
$\{S_1, S_2\}$	$\langle (cd)(a) \rangle$
$\{S_1, S_2\}$	$\langle (cd)(g) \rangle$
$\{S_2, S_3\}$	$\langle (g)(d)(e) \rangle$
$\{S_2, S_3\}$	$\langle (g)(d)(f) \rangle$
$\{S_2, S_3\}$	$\langle (g)(e)(f) \rangle$

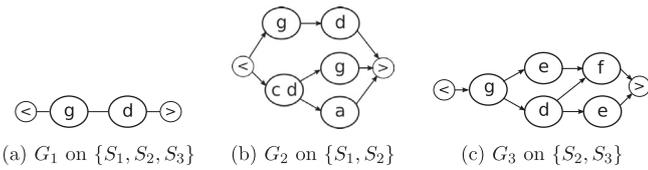


Fig. 2. Set of closed partially ordered patterns related to Table 1 with the minimum support $\theta = 2$.

these elements are not comparable. Each path in the graph is a **sequence**.

Definition 6 (Sub-partially ordered sequence)

Let G_α and G_β be two **partially ordered sequences**, G_α is a **sub-partially ordered sequence** of G_β , denoted $G_\alpha \preceq_g G_\beta$, if for all paths $path_{xi}$ in G_α there is a path $path_{\beta j}$ in G_β such that $path_{xi} \preceq_s path_{\beta j}$.

Definition 7 (Partially ordered pattern)

Let G_p be a **partially ordered sequence** and DB a sequence database. Let $S' \subseteq DB$ be the maximal set of **sequences** such that $\forall S_i \in S', G_p \preceq_g S_i$, i.e. all paths in G_p are supported by S_i . $|S'|$ is called the support of G_p . G_p is called a **partially ordered pattern**, denoted **po-pattern**, when $Support(G_p) \geq \theta$ where θ is a given value (minimum support).

Definition 8 (Closed partially ordered pattern)

Let G be a **partially ordered pattern**, G is a **closed partially ordered pattern** if there is no other **partially ordered pattern** G' such that $G \preceq_g G'$ and $Support(G) = Support(G')$.

To illustrate these definitions, let us consider the closed po-pattern G_3 given in Fig. 2c, which is supported by the sequences S_2 and S_3 . There are two paths between the itemset $\{g\}$ and the itemset $\{f\}$ given by the sequences $\langle (g)(e)(f) \rangle$ and $\langle (g)(d)(f) \rangle$, then $\langle (g) \rangle < \langle (f) \rangle$. Since this closed po-pattern is supported by the sequences S_2 and S_3 , $Support(G_3) = 2$. Given the po-pattern G_1 (Fig. 2a), we observe that $G_1 \preceq_g G_3$ since the path $\langle (g)(d) \rangle$ in G_1 is included in the paths $\langle (g)(d)(f) \rangle$ and $\langle (g)(d)(e) \rangle$. Then G_1 is a sub-partially ordered sequence of G_3 . Furthermore, each po-pattern in Fig. 2 is closed since it is not possible to add a new vertex or a new edge such that the support of the po-pattern do not decrease.

3. Related work

In the literature, po-pattern mining has been studied in two main contexts. The first involves mining po-patterns as frequent episodes occurring within a single sequence of events. In the second one, po-patterns are mined in a sequence database.

3.1. Mining po-pattern episodes in a single sequence

Fig. 3 represents a sequence composed of 21 timestamps with a sliding window covering five timestamps. The support of a po-pattern G is defined as the number of windows that support G within the considered sequence. Episodes are po-patterns representing a piece of repetitive information in a sequence according to an iterative temporal sliding window. Episode mining was first introduced by Mannila et al. [16]. The proposed algorithm, *Winepi*, mines episodes in an Apriori way by using a sliding window of fixed width. This method has drawbacks since mining huge

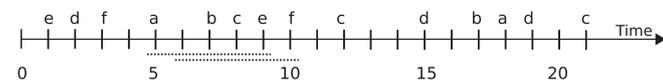


Fig. 3. Example of an event sequence with window of width 5.

databases leads to significant overhead. In addition, the algorithm does not merely extract closed po-patterns but rather the complete set of po-patterns. Therefore the number of patterns can be quite large. Other authors [17,18] proposed new algorithms to only mine closed episodes. Both use a *Pattern-Growth* paradigm to explore the search space. For this, the complete set of extracted patterns has to be stored in the computer memory in order to verify if each episode is closed or not during the process [18], or alternatively during a post-processing step [17].

3.2. Mining po-patterns in a sequence database

Mining episodes in a single sequence is very different from mining po-patterns in a sequence database. In [19], Pei et al. studied the problem of mining po-patterns in string databases. Despite the good performance of the *Frecpo* algorithm, they were able to only extract patterns on simple sequences that have non-repetitive items and no itemsets. This considerably reduces the potential applications of the algorithm as temporal databases are nowadays composed of multiple information for the same timestamp, i.e. multivariate, and the same piece of information can appear several times in a sequence. Since we aim at mining datasets composed of repetitive items and itemsets, we will not compare our approach with the *Frecpo* algorithm.

Alternatively, Garriga [15] presents another algorithm to extract closed po-patterns. Closed seq-patterns are thus first extracted using an algorithm such as *CloSpan* [13] or *BIDE* [14], and then a postprocessing operation is performed to convert a set of closed seq-patterns into po-patterns. Thus, the algorithm does not directly extract patterns, but uses an existing closed seq-pattern algorithm. Nevertheless, this approach only extracts a subset, not the complete set of closed po-patterns. The reason is that the algorithm proposed in [15] groups closed seq-patterns supported by the same set of sequences, noted S , in order to generate a closed po-pattern. Let \mathcal{CS} be the set of closed seq-patterns supported by S , the set of sequences supporting the generated closed po-pattern from \mathcal{CS} is necessarily equal to S . Then, for each set S of sequences supporting a generated closed po-pattern, there is a closed seq-pattern such that S is the maximal set of sequences supporting it. But some closed po-patterns are supported by a set of sequences S such that there is no closed seq-pattern maximally covered by S . We illustrate this given the set of sequences in Table 3 and its corresponding set of closed seq-patterns in Table 4 with a minimum support of 2. Based on this database, three closed po-patterns are shown in Fig. 4a–c. Nevertheless, the approach proposed in [15] is not able to extract the closed po-pattern G'_3 since in Table 4 there are no closed seq-pattern such that the list of sequences supporting it is equal to $\{S'_2, S'_3\}$. Conversely, the method proposed in this paper is able to extract such patterns.

The proposal in this paper is to directly mine closed po-patterns in a sequence database using a *Pattern-Growth* approach. Our method manages repetitive items and sequences composed of itemsets. We also adapt the optimization on equivalence databases used in closed seq-pattern mining [13]. In addition, closeness checking is directly performed during the process without having to consider already extracted po-patterns.

Table 3
A simple sequence database.

Seq id	Sequence
S_1	$\langle (a) \rangle$
S_2	$\langle (a)(b) \rangle$
S_3	$\langle (b)(a) \rangle$
S_4	$\langle (b) \rangle$

Table 4
Set of closed seq-patterns related to Table 3 with minimum support $\theta = 2$.

Supporting seq set	Sequence
$\{S'_1, S'_2, S'_3\}$	$\langle\langle a \rangle\rangle$
$\{S'_2, S'_3, S'_4\}$	$\langle\langle b \rangle\rangle$

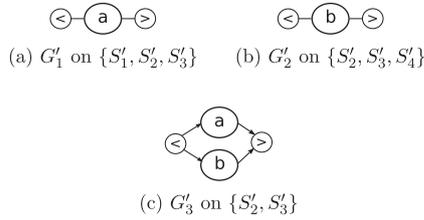


Fig. 4. Set of closed po-patterns related to Table 3 with a minimum support $\theta = 2$.

4. The OrderSpan algorithm

We now present the *OrderSpan* algorithm which is designed to meet the previously outlined challenges inherent to po-pattern mining: (1) mining po-patterns directly from a sequence database; (2) focusing extraction on closed po-patterns in order to reduce the result size; and (3) considering sequences of itemsets with repetitive items. This algorithm relies on a two-phase approach, based on the prefix and suffix properties of sequences. The following subsection presents the *Pattern-Growth* paradigm on sequences which is well-known in seq-pattern mining [20] and is used as a basis for our method.

4.1. The sequential pattern mining paradigm

Several useful methods have been proposed to tackle the sequence mining problem. We opted to base our algorithm on the *Pattern-Growth* paradigm which uses a divide-and-conquer strategy. This paradigm was first implemented in the *PrefixSpan* algorithm [20], which is currently one of the most efficient algorithms for extracting seq-patterns, in terms of both computation time and memory consumption.

To illustrate this paradigm, let us consider the tree in Fig. 5, which gives us the complete seq-pattern search space from the database of Table 1, with $\theta = 2$. Each node represents a sequence and, starting from the root, it is recursively possible to retrieve the complete set of seq-patterns. The number above each vertex is the support of the corresponding seq-pattern and dotted vertices represent closed seq-patterns. Thereby, given the sub-tree starting from the labeled vertex 'c' under the root, we obtain the following seq-patterns: $\langle\langle c \rangle\rangle$, $\langle\langle cd \rangle\rangle$, $\langle\langle cd(a) \rangle\rangle$, $\langle\langle cd(g) \rangle\rangle$, $\langle\langle c(a) \rangle\rangle$, $\langle\langle c(g) \rangle\rangle$, whose closed seq-patterns are: $\langle\langle cd(a) \rangle\rangle$ and $\langle\langle cd(g) \rangle\rangle$. Two operations are available in this tree, the *I-Extension* and the *S-Extension*. Given a sequence $S = \langle IS_1 IS_2 \dots IS_p \rangle$ and an item α , the operation $S \diamond \alpha$ concatenates the item α to S . The *I-Extension*, noted $S \diamond_i \alpha$, concatenates α in the last itemset IS_p of S , e.g. $\langle\langle c \rangle\rangle \diamond_i d$ gives the sequence $\langle\langle cd \rangle\rangle$. The *S-Extension*, noted $S \diamond_s \alpha$, concatenates α in a new itemset following IS_p , e.g. $\langle\langle cd \rangle\rangle \diamond_s g$ gives the sequence $\langle\langle cd \rangle g \rangle$. In Fig. 5,

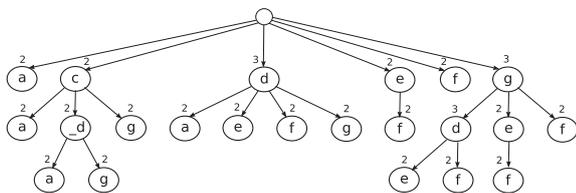


Fig. 5. The prefix tree representing the seq-pattern search space from the database in Table 1 with a minimum support $\theta = 2$.

a node starting with symbol ' _ ' represents an *I-Extension*, otherwise it represents an *S-Extension*.

Based on a depth-first search approach, the *Pattern-Growth* paradigm recursively divides the database by using database projections. Such projections are made according to a seq-pattern, called a prefix [20]. Given a prefix p , a projected sequence S with p is denoted $S|_p$, which is a suffix of S according to the first appearance of the prefix p in S . For instance, $\langle\langle ab \rangle\langle bc \rangle\langle ac \rangle\langle b \rangle\rangle|_{\langle\langle a \rangle\langle c \rangle\rangle} = \langle\langle ac \rangle\langle b \rangle\rangle$ and $\langle\langle ab \rangle\langle bc \rangle\langle ac \rangle\langle b \rangle\rangle|_{\langle\langle a \rangle\langle b \rangle\rangle} = \langle\langle _ \rangle\langle ac \rangle\langle b \rangle\rangle$. This last projected sequence starts with the symbol " _ " to take in consideration the possible *I-Extension* according to the item c . For each node in the tree, it is possible to build a projection $\mathcal{DB}|_p$ of the complete database according to the pattern p given by the node. Given \mathcal{DB} the sample database in Table 1, $\mathcal{DB}|_{\langle\langle g \rangle\langle d \rangle\rangle} = \{\langle\langle _ \rangle\langle e \rangle\langle f \rangle\langle aeg \rangle\rangle, \langle\langle e \rangle\langle f \rangle\rangle\}$. By scanning this projected database, we find the frequent items e and f with a support equal to 2. Thus, \mathcal{DB} will be recursively projected with prefixes $\langle\langle g \rangle\langle d \rangle\langle e \rangle\rangle$ and $\langle\langle g \rangle\langle d \rangle\langle f \rangle\rangle$.

4.2. From sequential pattern mining to partially ordered pattern mining

Previously, we highlighted the link between seq-patterns and po-patterns. This enables us to transform this sequence mining problem into a directed acyclic graph mining problem. In order to propose an efficient approach, we use an expanded partial order structure instead of the graph definition given by Definition 5. In this definition, vertex labels in a po-pattern represent an itemset, and two itemsets linked by a path are sequentially ordered. As presented in Section 4.1, sequential pattern mining has to consider *I-Extensions* and *S-Extensions*. Now, we consider these two extensions with two different arcs in the expanded partial order structure, defined as follows:

Definition 9. An **expanded partially ordered pattern** is a set of items with a partial order (\mathcal{V}, \prec) . It can be represented with a **labeled directed acyclic graph** $G^{exp} = (\mathcal{V}, \mathcal{A}, \Sigma_{\mathcal{V}}, \Sigma_{\mathcal{A}}, l_{\mathcal{V}}, l_{\mathcal{A}})$ where:

- \mathcal{V} is the set of **vertices** and \mathcal{A} is the set of **arcs**
- $\Sigma_{\mathcal{V}}$ is a set of items and $\Sigma_{\mathcal{A}} = \{I\text{-Extension}, S\text{-Extension}\}$
- $l_{\mathcal{V}} : \mathcal{V} \rightarrow \Sigma_{\mathcal{V}}$ and $l_{\mathcal{A}} : \mathcal{A} \rightarrow \Sigma_{\mathcal{A}}$ are mappings giving the labeling on **vertices** and **arcs**
- and given two vertices $u, v \in \mathcal{V}$, $l_{\mathcal{A}}(u, v) = "I\text{-Extension}"$ is equivalent to the sequence $\langle\langle uv \rangle\rangle$ and $l_{\mathcal{A}}(u, v) = "S\text{-Extension}"$ is equivalent to the sequence $\langle\langle u \rangle v \rangle$

We illustrate this definition by constructing an expanded version of the po-pattern G_2 , denoted G_2^{exp} , shown in Fig. 6. This transformation involves breaking each itemset in as many vertices as there are items in the itemset. This implies an *I-Extension* arc which separates each item of the itemset $\langle cd \rangle$. We use dotted arcs to distinguish *I-Extension* arcs from *S-Extension*. In itemsets, items are normally not ordered between them, so another possibility is to represent them by a clique or an hypergraph in the extended po-pattern. In order to follow the *Pattern-Growth* paradigm, items in itemsets are ordered according to their lexicographic value. For example, in Fig. 6 the broken itemset can be represented by $\langle cd \rangle$ or $\langle dc \rangle$, but $c < d$ so the arc $\langle c, d \rangle$ with $l_{\mathcal{A}}(c, d) = I\text{-Extension}$ is added to G_2^{exp} . This warrants generating po-patterns in a strict order to avoid unnecessary redundancy.

A closed po-pattern encompasses the complete set of closed seq-patterns covering a set of sequences S . In our case, the representation of a po-pattern is given by a sub-prefix-tree of the prefix-tree representing the whole seq-pattern search space. Let us consider the po-pattern G_2^{exp} (Fig. 6). Its sub-prefix-tree representation is given in Fig. 7a. This sub-prefix-tree represents the complete set of seq-patterns appearing in sequences S_1 and

S_2 of the sample database given in Table 1. For each seq-pattern S_p in the tree, there is at least one path p in G_2 such that $S_p \preceq_s p$. Thus, for each sequence subset in a sequence database, there is a sub-prefix-tree representing all seq-patterns covering this sequence subset.

4.3. Prefix-tree mining

The first algorithm step is based on this last property. Algorithm 1 extracts the complete set of sub-prefix-trees. Let us take a set of sequences \mathcal{S} , the algorithm first initializes a sub-prefix-tree which covers all sequences in \mathcal{S} . All frequent sub-prefix-trees on subsets in \mathcal{S} are then recursively extracted. This is based on the following assumption: for a sequence subset in the database, there is only one closed po-pattern, i.e. a sub-prefix-tree, describing its sequences [15]. To avoid the issue of mining the same pattern many times, we use a data structure called *ListSet*. It contains the list of sequence subsets covered by a previously extracted po-pattern. The algorithm computes a po-pattern that represents the covering sub-prefix-tree on \mathcal{S} . An example is given by the po-pattern in Fig. 7b, which is equivalent to the covering sub-prefix-tree (Fig. 7a).

Algorithm 1. ForwardTreeMining

```

input :  $\mathcal{S}$  a sequence set,  $\theta$  a minimum
        support, ListSet the set of
        sequence database subsets already
        explored
1 PartialOder  $\leftarrow$  new DAG;
2 PartialOder.begin.database = new
  ProjectedDatabase( $\mathcal{S}$ );
3 NodeQueue =  $\emptyset \cup$  PartialOder.begin;
4 while NodeQueue is not empty do
5   Node = NodeQueue.pop_back();
6   ListOcc  $\leftarrow$ 
  getListOccurrences(Node.database,  $\theta$ );
7   foreach Occ in ListOcc such that
  Occ.support =  $|\mathcal{S}|$  do
8     N' = Node.new();
9     N'.database =
  N.database.projectOn(Occ.item);
10    Extends the current node N with
  N';
11    NodeQueue = NodeQueue  $\cup$  N';
12  end
13  foreach Occ in ListOcc such that
  Occ.support <  $|\mathcal{S}|$  do
14    if ListSet does not contain the set
  of sequences covered by the
  occurrence Occ then
15       $\mathcal{S}'$  = set of sequences covered by
  the occurrence Occ;
16      ListSet = ListSet  $\cup$   $\mathcal{S}'$ ;
17      ForwardTreeMining( $\mathcal{S}'$ ,  $\theta$ , ListSet);
18    end
19  end
20 end
21 MergingSuffixTree(PartialOder);
22 return PartialOder;

```

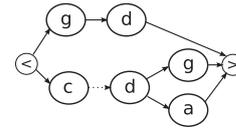


Fig. 6. Expanded version of the po-pattern G_2 .

Below, we detail each step of the algorithm. Lines [1–3]: a directed acyclic graph is initialized. According to the definitions given in Section 2, directed acyclic graphs contain two nodes representing the beginning and end of the po-pattern, labeled “ \langle ” and “ \rangle ”, respectively. Each node contains: (1) a label representing the extracted information, i.e. an item; (2) the projected database containing the suffixes of sequences in \mathcal{S} according to the prefix, i.e. seq-pattern, from the root node. A queue called *NodeQueue* is initialized with the “begin” node of the po-pattern. Lines [4–20]: *NodeQueue* is empty when there is no more node to be extended, i.e. the sub-prefix-tree is extracted. Line 6: the list of frequent occurrences is computed by scanning sequences in the projected database. Each occurrence is a frequent item in the projected database. Lines [7–12]: for each frequent occurrence (item) with a support equal to the cardinality of \mathcal{S} , it extends the current po-pattern, i.e. sub-prefix-tree. The aim is to fully cover sequences in \mathcal{S} . The I-Extension and the S-Extension are performed in line 10. Lines [13–19]: discovering of occurrences that have a support lower than the sequence set cardinality $|\mathcal{S}|$, which means that there is a more specific po-pattern covering a subset \mathcal{S}' such that $\mathcal{S}' \subset \mathcal{S}$. Therefore, if there is no po-pattern covering this subset \mathcal{S}' , a new one is extracted by recursively calling *ForwardTreeMining* on \mathcal{S}' . Line 21: when the extraction of the sub-prefix-tree is finished, the method *MergingSuffixTree* is applied on the po-pattern in order to prune and merge the redundant vertices. This operation is presented in the next section.

4.4. Prefix-tree pruning and merging

Since we use the prefix property to mine the complete set of sub-prefix-trees, a lot of redundancies are obtained. Let us consider the po-pattern given by Fig. 7b. It represents the complete set of seq-patterns on sequences \mathcal{S}_1 and \mathcal{S}_2 . Some of these seq-patterns are not closed and generate redundant information. For instance, the sub-sequence $\langle(c)(g)\rangle$ is supported by the sequence $\langle(cd)(g)\rangle$, and the sub-sequence $\langle(a)\rangle$ is supported by the sequence $\langle(cd)(a)\rangle$. By starting from the po-pattern ending vertex, vertices with the same label can be merged using the suffix property on sequences. Indeed, merging these vertices maintains the order among elements. For example, sequences $\langle(c)(a)\rangle$ and $\langle(cd)(a)\rangle$ are both suffixed by the suffix $\langle(a)\rangle$, and in the po-pattern, vertices representing this suffix with the label a can be merged into a single vertex labeled a . Then, it is possible to recursively merge all redundant vertices in a po-pattern. This process is called the *MergingSuffixTree* operation.

Fig. 8 provides an example of such an operation on the po-pattern in Fig. 7b. To illustrate the method, the operation is only performed on the parent nodes of the ending vertex. First, vertices representing the suffix $\langle(a)\rangle$ and labeled a are merged by retaining edges from their respective parent nodes. Next, the same operation is performed on vertices labeled g for the suffix $\langle(g)\rangle$ and on vertices labeled d for the suffix $\langle(d)\rangle$. This last suffix is a trivial case because there is only one vertex labeled d for the suffix $\langle(d)\rangle$. Once a set of vertices related to the same suffix are merged into one vertex v , this operation is recursively carried out on each parent node. Therefore, the operation executed on parent nodes of the merged

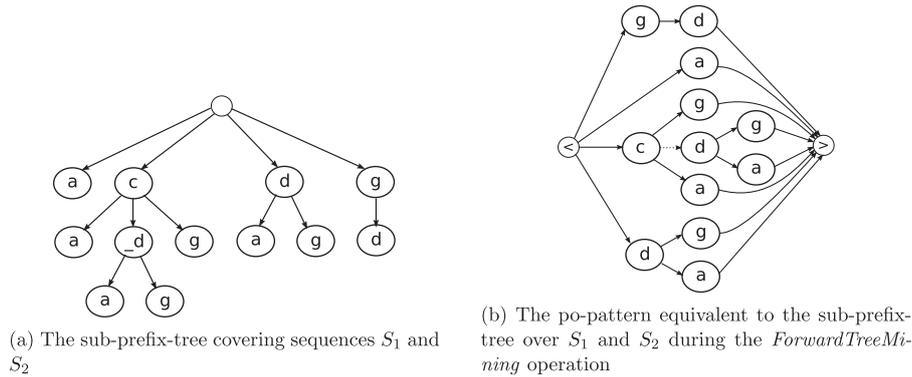


Fig. 7. From the sub-prefix-tree to the partially ordered pattern.

vertex a merges vertices labeled c and d which correspond to suffixes $\langle(c)(a)\rangle$ and $\langle(d)(a)\rangle$. In a po-pattern, a vertex v can have multiple suffixes represented by outgoing arcs of v . For example, the vertex labeled d has distinct suffixes $\langle(a)\rangle$ and $\langle(g)\rangle$. Note therefore that two vertices are merged only if they have exactly the same set of suffixes.

However, if we consider the suffix $\langle(c)(a)\rangle$, retaining edges from all parent nodes can generate transitive redundancy in the po-pattern. Indeed, in Fig. 8, the edge from vertex c to the merged vertex a is redundant since the order information is already given by the path represented by the sequence $\langle(cd)(a)\rangle$. Thus, after applying the *MergingSuffixTree* operation on a vertex, we have to check the transitive redundancy on the parent vertices. We can alternatively consider both *I-Extensions* and *S-Extensions* cases as described in Property 1.

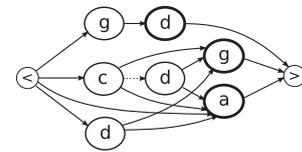


Fig. 8. Merging operation on the po-pattern covering S_1 and S_2 .

Property 1 (Redundant edge). Let $G^{exp} = (\mathcal{V}, \mathcal{A}, \Sigma_{\mathcal{V}}, \Sigma_{\mathcal{A}}, l_{\mathcal{V}}, l_{\mathcal{A}})$ be an extended partially ordered pattern with two vertices $\alpha, \gamma \in \mathcal{V}$, with one path from α to γ represented by a sequence S and with I_1 being the first item of S and I_n the last item of S , and three arcs (α, γ) , (α, I_1) , $(I_n, \gamma) \in \mathcal{A}$ such that $l_{\mathcal{A}}(\alpha, \gamma) = k$, $l_{\mathcal{A}}(\alpha, I_1) = l$ and $l_{\mathcal{A}}(I_n, \gamma) = m$. The arc (α, γ) is redundant if $\alpha \diamond_k \gamma \preceq_s \alpha \diamond_l I_1 \diamond_m \gamma$.

Fig. 9 illustrates this property.

Let us consider a merged vertex v , then the edge transitivity pruning involves in checking the redundancy for each pair of v parent vertices. We detail this process with the *EdgeTransitivityChecking* algorithm (see Algorithm 2) called after the end of the merging operation on a vertex.

Algorithm 2. EdgeTransitivityChecking.

```

input :  $v$  a merged vertex, PartialOrder
       the current po-pattern
1  $EdgeToRemove = \emptyset$ ;
2 foreach  $Edge$  in ParentEdges of  $v$  do
3   | if  $Edge$  is redundant according to
   | Property 1 then
4   | | Add  $Edge$  to  $EdgeToRemove$ ;
5   |
6 end
7 foreach  $Edge$  in  $EdgeToRemove$  do
8   | Remove  $Edge$  from PartialOrder;
9 end

```

Lines [1–6]: For each parent edge of a merged vertex v , we check if this edge is redundant according to Property 1. **Lines [7–9]:** Each redundant edge is removed from the po-pattern.

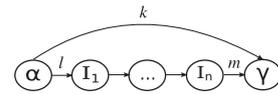


Fig. 9. Transitivity.

Thus, by removing transitive redundancy during the process, this operation ensures that all non-closed paths, i.e. non-closed seq-patterns, are removed. The *MergingSuffixTree* operation removes transitive redundancy and also recursively merges redundant vertices.

4.5. Optimization to reduce the ForwardTreeMining result

Reducing the po-pattern obtained from the *ForwardTreeMining* operation can be a costly operation due to substantial redundancy. A sub-prefix-tree contains the complete set of seq-pat- terns included in the associated closed po-pat- tern. Thus, a big-sized closed po-pattern (in the number of edges and vertices) implies a big-sized sub-prefix-tree. The worst case concerns the computation of a sub-prefix-tree covering a single sequence of the database, (feasible with a minimum support $\theta = 1$). In such a case, each sequence in the database has its equivalent closed po-pattern in the result. To illustrate this, let us take the sequence $S_3 = \langle(g)(d)(e)(f)\rangle$. The po-pattern corresponding to S_3 after the *ForwardTreeMining* operation is represented by Fig. 10.

Applying the *MergingSuffixTree* operation on it leads to the trivial closed po-pattern given in Fig. 11. We see that the sub-prefix-tree is huge compared to the corresponding merged closed po-pattern. In this example, the sequence is just composed of four items. Intuitively, the sub-prefix-tree gets exponentially bigger as the sequence becomes longer.

Although mining a sub-prefix-tree over a single sequence is the worst case, this is less of a problem with multiple sequences. In the above example, the closed po-pattern is represented by the deeper branch of the sub-prefix-tree, then intuitively it could be efficient to not explore the others branches since they are redundant.

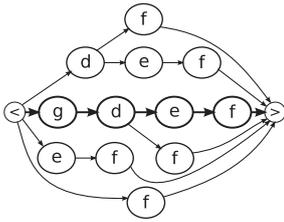


Fig. 10. The po-pattern over the sequence S_3 after the *ForwardTreeMining* operation.



Fig. 11. The closed po-pattern over the sequence S_3 .

This issue has been studied in the *CloSpan* [13] algorithm to skip redundant branches in the case of seq-pattern mining. In this section, we introduce a generalization of this optimization to extract a pre-merged po-pattern instead of a po-pattern representing the complete sub-prefix-tree. All redundant branches are merged together. In *CloSpan*, the authors use the sum of the projected sequence length to efficiently prune the prefix-tree during the extraction of closed seq-patterns. The notion of equivalent projected databases given different sequence prefixes underlies this concept. This method is clever, since considering two different frequent prefixes, if their projected databases are equal then they have the same set of *I-Extensions* and *S-Extensions* (same suffixes).

We illustrate this with the example given in Fig. 12 based on the example used in Section 4.2. During the mining process, the database is projected at each vertex in order to expand the po-pattern. We note that some of these projections are equivalent. In the example, the projected databases according to the sequence prefixes $\langle\langle d \rangle\rangle$ and $\langle\langle cd \rangle\rangle$ are identical because $DB|_{\langle\langle d \rangle\rangle} = DB|_{\langle\langle cd \rangle\rangle} = \{\langle\langle a \rangle\rangle(g)(d), \langle\langle _e \rangle\rangle(f)(a, e, g)\}$. They have exactly the same set of frequent suffixes, which are here $\langle\langle a \rangle\rangle$ and $\langle\langle g \rangle\rangle$.

Then it is not necessary to have two different branches in the po-pattern concerning prefixes $\langle\langle d \rangle\rangle$ and $\langle\langle cd \rangle\rangle$, as both contain exactly the same information. The optimization method we propose is based on this observation. Besides, checking the equivalence of two given databases leads to a potential overhead because of sequence scanning. To deal with this issue, the property proposed in *CloSpan* is based on the fact that two projections of the same sequence according to two prefixes are equivalent if they have the same length, i.e. the same length means the same suffix. Thus, we introduce Definition 10.

Definition 10 (*Projected sequences equivalence*)

Let S be a sequence and α, β be two sequence prefixes. The projected sequences $S|_{\alpha}$ and $S|_{\beta}$ are equivalent if $length(S|_{\alpha}) = length(S|_{\beta})$.

Indeed, when a sequence S is projected according to a sequence prefix, the projection is a suffix of S . When two projections of a

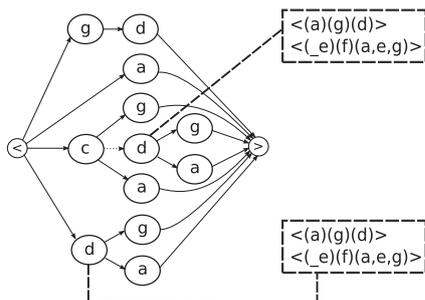


Fig. 12. Example of equivalent projected databases for the po-pattern in Fig. 10.

sequence have the same length, it means they contain the same suffix of the sequence, thus the two projections are equivalent. Based on this definition, we define now the equivalence of projected databases (see Definition 11).

Definition 11 (*Projected databases equivalence*)

Let DB be a sequence database and α, β two sequence prefixes. The projected databases $DB|_{\alpha}$ and $DB|_{\beta}$ are equivalent if $\forall S \in DB, S|_{\alpha}$ and $S|_{\beta}$ are equivalent.

Given this property, it is now possible to efficiently check whether two projections are equivalent during the po-pattern extension. This approach differs from the method proposed in [13]. In *CloSpan*, given a sequence database DB and two sequence prefixes α, β , the equivalence of their projected databases is verified only if $\alpha \preceq_s \beta$ or $\beta \preceq_s \alpha$. This is due to the way in which *CloSpan* explores the search space, it is then limited to a small part of the search space. In our definition, all projections of a po-pattern must be checked with others. The idea is to use this property during the extension process. Given an *I-Extension* or a *S-Extension* of a vertex v , two cases are possible: (1) there is no other vertex in the po-pattern having an equivalent projected database, then a new vertex v' is created and an edge is added between v and v' ; (2) there is another vertex v'' in the po-pattern having an equivalent projected database, then an edge is added between v and v'' . To illustrate this, let us consider the po-pattern in Fig. 12. With the optimization, the two vertices labeled d are now represented by only one vertex. This optimization applied during the *ForwardTreeMining* operation leads to the pre-merged po-pattern illustrated in Fig. 13. In this example, it significantly reduces the number of vertices at the end of the *ForwardTreeMining* operation. We observe that this optimization produces a transitive redundancy in the po-pattern. However, the *EdgeTransitivityChecking* function presented in Section 4.4 is well-adapted to such an issue. Removing the transitive redundancy in the case of vertex merging is equivalent to removing the transitive redundancy of a pre-merged po-pattern. The transitive redundancy produced by this operation is then automatically removed during the *MergingSuffixTree* process.

Optimizations from seq-pattern mining domain are then well-adapted to the case of mining closed po-patterns. This section studies the adaption of the property from the *CloSpan* algorithm, however the optimization proposed in the *BIDE* [14] algorithm may potentially be used or combined with the present optimization. Such a combined approach is not introduced in this paper and may be the subject of future work in closed po-pattern mining.

Fig. 14 synthesizes the *OrderSpan* overall process by illustrating interactions between the different algorithm parts. In the figure, operations that are specific to *OrderSpan* are dotted.

4.6. Complexity analysis

This section presents an analysis of the overall complexity of the *ForwardTreeMining* and the *MergingSuffixTree* operations. For clarity purpose, these two operations are denoted *FTM* and *MST*. The studied complexity concerns the worst case, thus we do not consider the optimization proposed in Section 4.5 because the worst case remains the same: if there are no equivalent projected databases during the process, then the optimized algorithm is equivalent to the non-optimized one.

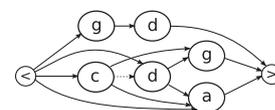


Fig. 13. Example of optimized po-pattern extension.

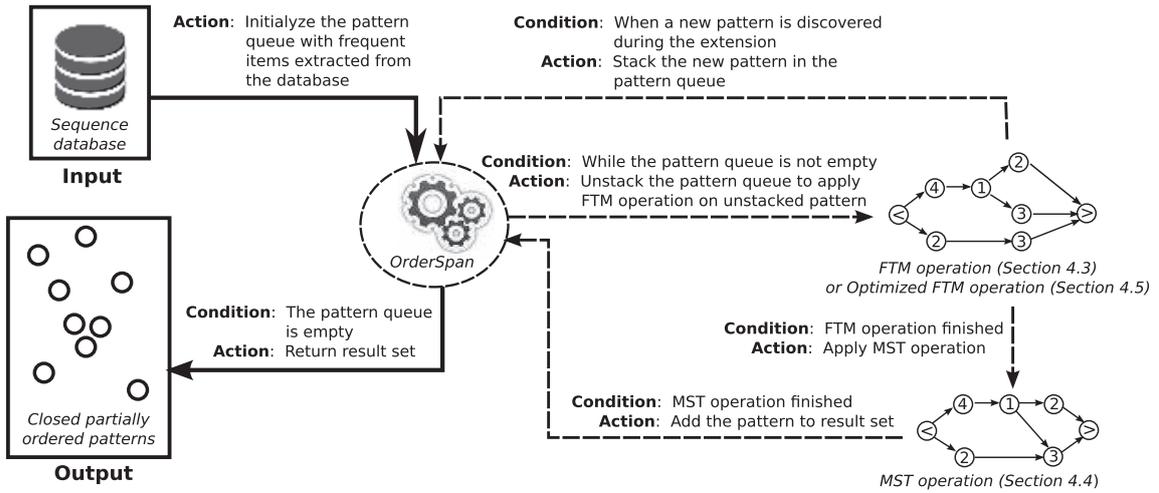


Fig. 14. Overall process.

Given a set of sequences \mathcal{S} , *FTM* gives the sub-prefix-tree in a po-pattern $G_T(\mathcal{V}_T, \mathcal{E}_T)$ covering \mathcal{S} , with \mathcal{V}_T being the set of vertices and \mathcal{E}_T the set of edges. *MST* merges redundancies from G_T to compute the associated closed po-pattern.

The *FTM* complexity is very close to the PrefixSpan complexity because it involves mining a complete sub-prefix-tree over a set of sequences. The worst-case complexity of PrefixSpan is given by the formula $\Theta((2 \cdot |\mathcal{I}|)^{S_{max}})$, with \mathcal{I} being the set of items in the complete database and S_{max} the longest sequence in the database. In our case, this formula represents the worst-case complexity for *FTM* over a single set of sequences \mathcal{S} . This gives the size of the biggest potential sub-prefix-tree (in number of vertices) that can be computed from a set of sequences. Then, we have $|\mathcal{V}_{T_{max}}| = (2 \cdot |\mathcal{I}|)^{S_{max}}$. The *FTM* complexity is given by Eq. (1).

$$\Theta((2 \cdot |\mathcal{I}|)^{S_{max}}) \quad (1)$$

The *MST* operation is applied after the extraction of the sub-prefix-tree G_T . This is equivalent to a *width-first* search in a tree because each vertex is explored only once. Such an algorithm has a well-known complexity linear with respect to the size of the tree, $\Theta(|\mathcal{V}_T|)$. Since in the worst case $|\mathcal{V}_{T_{max}}| = (2 \cdot |\mathcal{I}|)^{S_{max}}$, the *MST* complexity is the same as the complexity of *FTM* provided by Eq. (1).

Now, given the complexity of the *FTM* and the *MST* operations, we provide the overall worst-case complexity. Like existing pattern mining approaches, the complexity of our algorithm is related to the size of the result, i.e. the number of extracted patterns. Then, the overall worst-case complexity is given by Eq. (2) with respect to the size of \mathcal{G}_{max} , i.e. the complete set of extracted po-patterns.

$$\Theta(|\mathcal{G}_{max}|) \cdot (\Theta(FTM) + \Theta(MST)) = \Theta(|\mathcal{G}_{max}| \cdot 2 \cdot (2 \cdot |\mathcal{I}|)^{S_{max}}) \quad (2)$$

An upper bound for \mathcal{G}_{max} . Since there is only one possible closed po-pattern for a set of sequences, we can estimate the maximal number of closed po-patterns that can be extracted from a sequence database \mathcal{DB} , with $|\mathcal{DB}|$ being the number of sequences in \mathcal{DB} . According to a minimum support θ , the upper bound of the number of potential po-patterns in \mathcal{DB} , denoted $|\mathcal{G}_{max}|$, is the number of sequence subsets $\mathcal{S}' \subseteq \mathcal{DB}$ such that $|\mathcal{S}'| \geq \theta$. If $\theta = 1$, the case is trivial. The number of subsets is given by $2^{|\mathcal{DB}|}$. As we do not consider the empty set $\{\emptyset\}$, $|\mathcal{G}_{max}| = 2^{|\mathcal{DB}|} - 1$. If $1 < \theta \leq |\mathcal{DB}|$, the number of sequence subsets is given by Formula (3). It computes the sum of the binomial coefficients $\binom{|\mathcal{DB}|}{k}$ from $k = \theta$ to $|\mathcal{DB}|$, which is, for each iteration, the number of k -combinations of sequences in \mathcal{DB} .

$$|\mathcal{G}_{max}| = \sum_{k=\theta}^{|\mathcal{DB}|} \binom{|\mathcal{DB}|}{k} \quad (3)$$

5. Experiments

In this section, some tests were conducted on sequence databases with different characteristics. The experiments were performed on a laptop computer with an Intel Core i7 and 8 GB main memory, running on Debian sTable 7.0. We implemented the *OrderSpan* algorithm in C++. The experiments are divided into two subsections: (1) a performance study of the algorithm with respect to the computation time; (2) a qualitative study of closed po-patterns compared to closed seq-patterns.

5.1. Performance study

For the performance evaluation of *OrderSpan*, we selected three different datasets (*Gazelle*, *Kosarak* and *Sign*) that have been previously used in many sequential pattern mining papers [14,13,21] to evaluate algorithms. They are accessible online¹ and can be downloaded in SPMF format. Each dataset is briefly presented in the following.

Gazelle. This dataset was proposed during the KDD cup 2000.² It contains data on clickstreams and purchases from Gazelle.com. It is a very sparse dataset describing the behavior of thousand customers. A sequence describes each customer browsing by ordering his clickstreams.

Kosarak. This very large dataset contains sequences of anonymized clickstreams obtained from a news Hungarian website. Little information is available about this dataset.³ Due to the huge number of sequences in this dataset (almost a million), our experiments were conducted on a subset.

Sign. This latter dataset is from Boston University.⁴ It is a collection of utterances with an associated video segment for each utterance, recording a number of American Sign Language (ASL) gestural and grammatical fields over a time interval.

For this performance study, we intentionally choose very different and heterogeneous datasets. To characterize these various

¹ <http://www.philippe-fournier-viger.com/spmf/index.php?link=databases.php>.
² <http://www.kdd.org/kdd-cup-2000-online-retailer-website-clickstream-analysis>.
³ <http://fimi.ua.ac.be/data/>.
⁴ http://cs-people.bu.edu/panagpap/Research/asl_mining.htm.

Table 5
Datasets statistics.

Dataset	Sequences	Alphabet	Avg sequence size
Gazelle	59,601	497	2.51
Kosarak	70,000	21,144	7.98
Sign	730	267	51.99

datasets, we computed different statistics like the total number of items, the total number of sequences and the average sequence size (see Table 5).

Furthermore, we give an overall idea of the number of closed po-patterns according to the minimum support in each dataset (see Fig. 15a–c). In these experiments, we arbitrarily opted to represent the relative minimum support instead of the absolute minimum support. For example, the *Gazelle* dataset contains 59,601 sequences, and a relative minimum support equal to 0.001 means that we extracted all closed po-patterns that cover at least 60 sequences (absolute minimum support) because: $59,601 \times 0.001 = 59.601$. Like in all pattern mining problems, the number of closed po-patterns increases exponentially as the minimum support decreases. For instance, in the *Sign* dataset, 29,662 closed po-patterns were extracted with a relative minimum support of 0.38, and 147,004 were extracted with a relative minimum support of 0.32.

Fig. 16a–c compare the performance of the algorithm with and without optimization. Note that the optimization was more efficient when the minimum support was low. This behavior is due to the fact that when the minimum support decreases, the probability of having equivalent databases increases. For example, in the *Gazelle* dataset, with the relative minimum support 0.005, the optimized version is 1.14 times faster and, at a relative minimum support 0.00065, it is 2.83 times faster than the original algorithm. The performance gain obtained with the optimization is related to the size of the po-patterns after the *ForwardTreeMining* operation. Optimization reduces the number of vertices before the *MergingSuffixTree* operation. To illustrate this property, Fig. 17a–c show the average number of vertices per po-pattern before the *MergingSuffixTree* operation with and without optimization. The average number of vertices between the non-optimized and optimized method is strongly related to the observed difference in extraction time. For example, in the *Kosarak* dataset with a minimum support 0.001, the average number of vertices in the non-optimized version is 16.64 times bigger than the optimized one, while time extraction is 3.84 times faster in the optimized approach. In the *Sign* dataset, we observe that the difference between the average number of vertices in both versions is low, i.e. 1.40 times at minimum support 0.45 and 1.66 times at 0.32. The difference in extraction time is 1.08 times and 1.14 times, respectively. Fig. 17b illustrates an interesting case with the decrease in the average number of vertices for the non-optimized approach at a minimum support 0.001. This is due to

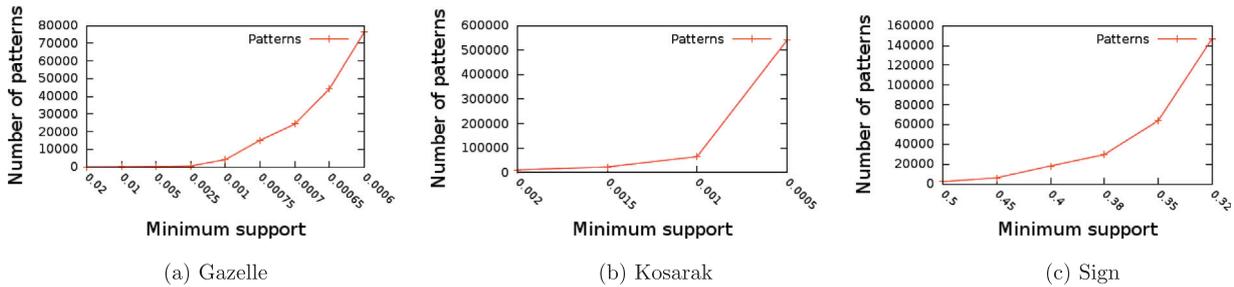


Fig. 15. Number of closed po-patterns.

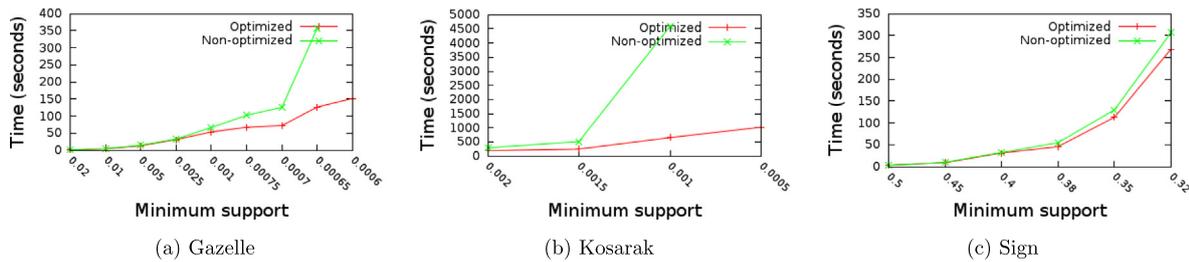


Fig. 16. Computation time.

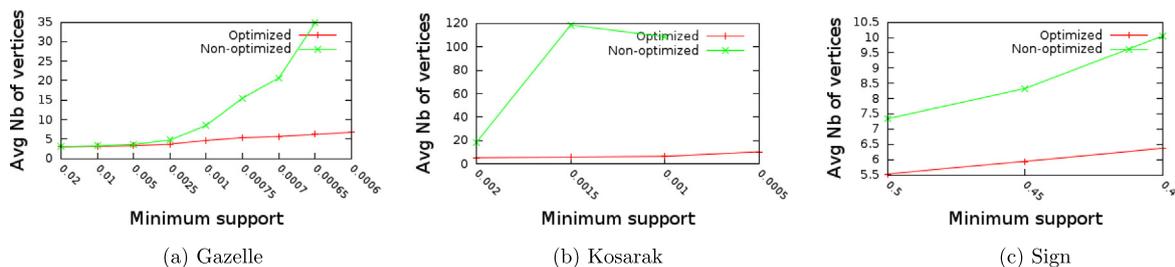


Fig. 17. Average number of vertices.

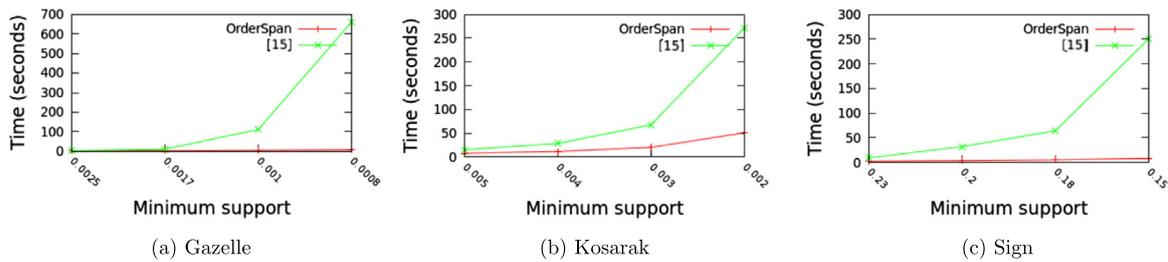


Fig. 18. Comparison between *OrderSpan* and the approach in [15].

the fact that closed po-patterns between the relative minimum supports 0.0015 and 0.001 are on average smaller than those extracted over the minimum support 0.0015.

For the non-optimized algorithm, we note that experiments on *Gazelle* at minimum support 0.0006 and *Kosarak* at minimum support 0.0005 are not presented due to the long computation time.

For each dataset, the optimization gain is linked to the average number of vertices in both approaches. The greater the difference, the faster the optimized approach is compared to the non-optimized approach. This relation depends on the dataset, for example the *Sign* dataset is less sensitive to optimization than the *Gazelle* and *Kosarak* datasets.

We also compare the efficiency of *OrderSpan* with the approach proposed in [15]. As explained in the related work section, the algorithm in [15] is a postprocessing applied on a set of previously extracted closed seq-patterns. Conversely to *OrderSpan*, it does not extract the complete set of closed po-patterns. Indeed, the author's aim is different and she searches for closed po-patterns such that, for each one of them, there is at least one closed seq-pattern on the same set of sequences (see Section 3). Before providing the comparison, we briefly describe the different steps of the algorithm, described in [15]:

1. First, the complete set of closed seq-patterns is extracted from the data given a minimum support threshold θ with the BIDE [14] algorithm.
2. Then, closed seq-patterns are grouped according to the sequence set that support them. Each group of closed seq-patterns represents the common information shared by a sequence set S of the database, if and only if there is a seq-pattern closed on S (see Section 4 in [15]).
3. Finally, for each group of closed seq-patterns, a closed po-pattern is generated. It consists in matching each itemset value and position of the closed seq-patterns to form the maximal paths in the associated closed po-pattern (see Section 5 in [15]).

Then, to perform a valid comparison, we modified *OrderSpan* to extract exactly the same set of closed po-patterns as the method presented in [15]. Such a comparison has the benefit of comparing an algorithm that directly extracts closed po-patterns (*OrderSpan*) with an algorithm that performs a postprocessing (approach in [15]). Fig. 18a–c give the results obtained on the three previously used datasets. In these experiments, we used the optimized version of *OrderSpan*.

Let us consider the *Kosarak* dataset with the relative minimum support 0.002, algorithm in [15] performs in 67.97 s while *OrderSpan* performs in 20.37 s. In the *Gazelle* dataset with the relative minimum support 0.001, algorithm in [15] performs in 111.09 s while *OrderSpan* performs in 5.43 s. In this last example, *OrderSpan* is almost 20 times faster. These results show that mining directly closed po-patterns in the data is more efficient than applying a postprocessing on a set of closed seq-patterns. The reason is as

follows, when the number of closed seq-patterns is important (many thousands), a postprocessing approach is costly because a comparison for many pairs of patterns has to be done. Furthermore, this comparison shows that our approach is also efficient when transposed to another context, e.g. by searching for a subset of closed po-patterns.

5.2. Semantic study

This last experimental section shows the usefulness of closed po-pattern mining compared to closed seq-pattern mining, based on a real dataset. As shown in the first definitions: (1) each path in a closed po-pattern is a closed seq-pattern and (2) there is only one closed po-pattern covering a given set of sequences. The po-patterns thus provide global detailed information about the order between items in a given set of sequences. This global information can be interesting for experts and easy to analyze. To illustrate this, we study closed po-patterns extracted from the hydro-ecological dataset *Fresqueau*.⁵ We provide this study since we are collaborating with hydrobiologists in the context of the *Fresqueau* ANR project. For a more complete and detailed application of closed po-patterns on hydro-ecology, the reader is encouraged to refer to [22], where a full qualitative study is provided. Several studies [23–27] have applied data mining approaches on hydro-ecological data, but none of them used an approach based on temporal patterns, like seq-patterns or po-patterns. The *Fresqueau* dataset provides temporal information about 2,505 French sampling sites on rivers. For each site, we have different samples obtained at different timestamps. These data are divided into two categories:

Biological data They concern the flora and the fauna species living in the river, such as macro-invertebrates and macrophytes. They also contain various biological indicators which give a global quality note about the hydro-ecosystem. In these experiments, we focus on the IBGN indicator. The IBGN is related to macro-invertebrate species sampled at a site. Some species are typical of a good water quality while some other species are not. These measures are obtained once a year at each site.

Physico-chemical data These represent a list of chemical and physical water measures. They are divided into two groups: (1) macro-pollutants like nitrates or suspended matter (2) micropollutants like pesticides. These measures are obtained every 2 months at each site.

⁵ <http://engees-fresqueau.unistra.fr/>.

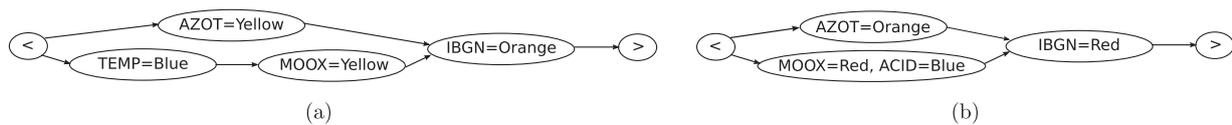


Fig. 19. Example of closed po-patterns in the *Fresqueau* dataset.

Specialists are concerned about the environmental impacts of macropollutants on water quality. For example, detecting the macropollutants and the order in which they vary in the river can explain biodiversity changes. More specifically, the objective is to link a biological quality index value (here IBGN) to one or more macropollutant quality values.

To process the dataset each variable is discretized. There are five quality values: *Very good*, *Good*, *Medium*, *Bad*, *Very Bad* respectively represented by colors *Blue*, *Green*, *Yellow*, *Orange* and *Red*. The five biological quality index values are given by the French DCE standard.⁶ The macropollutants are grouped and discretized in five quality values according to the SEQ-Eau standard.⁷ According to the expert's needs, we constrained the extraction of closed po-patterns within an interval ranging from 0 and 6 months before a biological sampling, by preprocessing the sampling site sequences.

Fig. 19a and b show us two examples of two closed po-patterns extracted from this database. These two patterns are related to the macropollutants frequently found before a biological sampling with a given quality. Recall that an orange IBGN value (Fig. 19a) is representative of bad water quality and a red IBGN value (Fig. 19b) is representative of very bad water quality. We provide some information about macropollutants appearing in both po-patterns:

- ACID** The water acidity, determined by the pH value of the water combined with the level of dissolved aluminum.
- AZOT** The organic nitrogenous matter, given by levels of NH_4^+ , NKJ and NO_2^- molecules.
- MOOX** The oxidizable organic matter, given by levels of NH_4^+ and NKJ molecules (like the *AZOT* characteristic) but also other elements like organic carbon and oxygen dissolved in water.
- TEMP** The water temperature.

The closed po-pattern in Fig. 19a shows three different characteristics frequently found from 0 to 6 months before sampling of an orange IBGN value. We can observe a yellow AZOT value and a blue TEMP value followed by a yellow MOOX value. The closed po-pattern in Fig. 19b (red IBGN value) shows an orange AZOT value and a red MOOX value frequently sampled at the same time as a blue ACID value. It is interesting to note that the two closed po-patterns have two characteristics in common: MOOX and AZOT. These two closed po-patterns highlight that, frequently, as the MOOX and ACID quality values increase, the IBGN quality value gets better. For example MOOX has a red quality value in the po-pattern corresponding to the red quality IBGN, and this same characteristic has a yellow value in the po-pattern corresponding to the orange quality IBGN. These first results are promising since they match expert's forecast. Indeed, MOOX represents all substances whose presence may cause the consumption of dissolved oxygen in streams. A red MOOX means that there are a lot of these substances in the water. This can lead to massive water deoxygenation, resulting in the death of macro-invertebrates. AZOT represents the organic

Table 6

Closed seq-patterns extracted from the closed po-pattern in Fig. 19b.

Sequential pattern
((AZOT = Orange)(IBGN = Red))
((MOOX = Red, ACID = Blue)(IBGN = Red))

nitrogenous matter derived from domestic, industrial and livestock waste discharges. The main impact is that these discharges contribute to the growth of algae and plants. A high presence of organic nitrogenous matter does not have a direct impact on macro-invertebrates, but it does on their habitat, including the flora.

We now compare these closed po-patterns with their related closed seq-patterns. For each closed po-pattern, it is possible to retrieve the set of corresponding closed seq-patterns by extracting each path in the po-pattern. For example, extracting each path of the closed po-pattern in Fig. 19b leads to the closed seq-patterns $\langle (AZOT = Orange)(IBGN = Red) \rangle$ and $\langle (MOOX = Red, ACID = Blue)(IBGN = Red) \rangle$ (see Table 6).

Two main results have to be noticed. (1) By extracting closed po-patterns, we get the information that closed seq-patterns included in a closed po-pattern occur in the same set of sampling sites. Conversely, with closed seq-pattern mining, we do not know which seq-patterns occur at the same set of sites. (2) Furthermore, closed seq-pattern mining provides a lot of seq-patterns that may be redundant and hard to analyze: the representation of closed po-patterns in a directed acyclic graph tackles this issue by automatically merging shared parts of closed seq-patterns.

6. Conclusion

Closed po-pattern mining requires the development of new techniques to extract such patterns in the general context of large temporal databases. Greater complexity and a much vaster search space require optimization techniques to efficiently discover po-patterns.

This paper presents OrderSpan, an algorithm which can be used to efficiently mine the complete set of closed po-patterns. Our method uses both the prefix and suffix properties of seq-patterns, based on *ForwardTreeMining* and *MergingSuffixTree* operations, respectively. It uses the equivalence database property to improve the *ForwardTreeMining* process by reducing redundant explorations. Subsequently, it is able to directly mine all kinds of sequential databases with sequences of itemsets. Compared to existing approaches, we directly mine po-patterns rather than generating them from seq-patterns or only considering sequences with no item repetitions.

We provide a performance evaluation on three classical sequence datasets to highlight the optimization gain. Furthermore, we perform a quality evaluation on a real dataset, to highlight the importance of mining closed po-patterns instead of mining closed seq-patterns. They prove to be more relevant for specialist analysis due to the synthetic information they provide.

As perspective, other sequential pattern mining optimization techniques [14] deserve to be studied, adapted and combined with our approach to even more improve its efficiency.

⁶ <http://www.eaufrance.fr/observer-et-evaluer/etat-des-milieux/regles-d-evaluation-de-l-etat-des/>.

⁷ <http://sierm.eaurmc.fr/eaux-superficielles/fichiers-telechargeables/grilles-seq-eau-v2.pdf>.

Acknowledgment

This work was funded by the French National Research Agency (ANR), as part of the ANR11_MONU14 Fresqueau project.

References

- [1] L. Geng, H.J. Hamilton, Interestingness measures for data mining: a survey, *ACM Comput. Survey* (2006) 38.
- [2] H. Cheng, X. Yan, J. Han, C. Hsu, Discriminative frequent pattern analysis for effective classification, in: *International Conference on Data Engineering, ICDE, 2007*, pp. 716–725.
- [3] H. Cheng, X. Yan, J. Han, P.S. Yu, Direct discriminative pattern mining for effective classification, in: *International Conference on Data Engineering, ICDE, 2008*, pp. 169–178.
- [4] M. Wang, X.-q. Shang, Z.-h. Li, Sequential pattern mining for protein function prediction, in: *Advanced Data Mining and Applications*, vol. 5139 of *ADMA*, 2008, pp. 652–658.
- [5] R. Agrawal, R. Srikant, Mining sequential patterns, in: *International Conference on Data Engineering, ICDE, 1995*, pp. 3–14.
- [6] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: *International Conference on Very Large Data Bases, VLDB, 1994*, pp. 487–499.
- [7] N.R. Mabroukeh, C.I. Ezeife, A taxonomy of sequential pattern mining algorithms, *ACM Comput. Surveys (CSUR)* 43 (2010) 3.
- [8] C.H. Mooney, J.F. Roddick, Sequential pattern mining – approaches and algorithms, *ACM Comput. Surveys (CSUR)* 45 (2) (2013) 19.
- [9] A. George, D. Binu, DRL-prefixspan: a novel pattern growth algorithm for discovering downturn, revision and launch (DRL) sequential patterns, *Central Eur. J. Comput. Sci.* 2 (2012) 426–439.
- [10] J. Ren, L. Wang, J. Dong, C. Hu, K. Wang, A novel sequential pattern mining algorithm for the feature discovery of software fault, in: *International Conference on Computational Intelligence and Software Engineering*, vol. 5854 of *CISE*, 2009, pp. 439–447.
- [11] A. Sallaberry, N. Pecheur, S. Bringay, M. Roche, M. Teisseire, Sequential patterns mining and gene sequence visualization to discover novelty from microarray data, *J. Biomed. Inform.* 44 (2011) 760–774.
- [12] M. Fabrègue, A. Braud, S. Bringay, F. Ber, M. Teisseire, OrderSpan: mining closed partially ordered patterns, in: *Advances in Intelligent Data Analysis XII*, vol. 8207 of *LNCS*, 2013, pp. 186–197.
- [13] X. Yan, J. Han, R. Afshar, CloSpan: mining closed sequential patterns in large datasets, in: *SIAM International Conference on Data Mining, 2003*, pp. 166–177.
- [14] J. Wang, J. Han, BIDE: efficient mining of frequent closed sequences, in: *International Conference on Data Engineering, ICDE, 2004*, pp. 79–90.
- [15] G. Casas-Garriga, Summarizing sequential data with closed partial orders, in: *SIAM International Conference on Data Mining, SDM, 2005*.
- [16] H. Mannila, H. Toivonen, A.I. Verkamo, Discovery of frequent episodes in event sequences, *Data Mining Knowl. Discovery* 1 (1997) 259–289.
- [17] N. Tatti, B. Cule, Mining closed strict episodes, *Data Mining Knowl. Discovery* 25 (2012) 34–66.
- [18] W. Zhou, H. Liu, H. Cheng, Mining closed episodes from event sequences efficiently, in: *Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, vol. 6118 of *PAKDD*, 2010, pp. 310–318.
- [19] J. Pei, H. Wang, J. Liu, K. Wang, J. Wang, P.S. Yu, Discovering frequent closed partial orders from strings, *IEEE Trans. Knowl. Data Eng.* 18 (2006) 1467–1481.
- [20] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, M.-C. Hsu, Mining sequential patterns by pattern-growth: the PrefixSpan approach, *IEEE Trans. Knowl. Data Eng.* 16 (2004) 1424–1440.
- [21] P. Papapetrou, G. Kollios, S. Sclaroff, D. Gunopulos, Discovering frequent arrangements of temporal intervals, in: *Proceedings of the Fifth IEEE International Conference on Data Mining, ICDM '05*, IEEE Computer Society, Washington, DC, USA, 2005, pp. 354–361.
- [22] M. Fabrègue, A. Braud, S. Bringay, F. Ber, M. Teisseire, Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment, *Ecol. Inf.* 24 (2014) 210–221.
- [23] M. Pugelj, S. Džeroski, Predicting structured outputs k-nearest neighbours method, in: *Discovery Science*, vol. 6926, 2011, pp. 262–276.
- [24] D. Kocev, A. Naumoski, K. Mitreski, S. Krstić, S. Džeroski, Learning habitat models for the diatom community in Lake Prespa, *Ecol. Model.* 221 (2) (2010) 330–337.
- [25] Y.-C.E. Yang, X. Cai, E.E. Herricks, Identification of hydrologic indicators related to fish diversity and abundance: a data mining approach for fish community analysis, *Water Resour. Res.* (2008) 44.
- [26] A. Bertaux, F. Le Ber, A. Braud, M. Trémolières, Identifying ecological traits: a concrete FCA-based approach, in: *Formal Concept Analysis*, vol. 5548 of *LNCS*, 2009, pp. 224–236.
- [27] E. Dakou, T. D'heygere, A. Dedecker, P. Goethals, M. Lazaridou-Dimitriadou, N. Pauw, Decision tree models for prediction of macroinvertebrate taxa in the River Axios (Northern Greece), *Aquatic Ecol.* 41 (2007) 399–411.

Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana

Claude Flamand,¹ Mickael Fabregue,² Sandra Bringay,^{2,3} Vanessa Ardillon,⁴ Philippe Quénel,¹ Jean-Claude Desenclos,⁵ Maguelonne Teisseire⁶

¹Epidemiology Unit, Institut Pasteur in French Guiana, Cayenne, French Guiana

²LIRMM, CNRS, UMR 5506, Montpellier, France

³MIAp Department, University Paul-Valéry, Montpellier, France

⁴Regional Epidemiology Unit of the French Institute for Public Health Surveillance, Institut de Veille Sanitaire, Cayenne, French Guiana

⁵French Institute for Public Health Surveillance (Institut de Veille Sanitaire), Saint-Maurice, France

⁶Laboratory Department of Information System, Irstea-TETIS, Montpellier, France

Correspondence to

Claude Flamand, Epidemiology Unit, Institut Pasteur in French Guiana, 23 Avenue Pasteur BP 6010, Cayenne, Cedex 97306, French Guiana; cflamand@pasteur-cayenne.fr

Received 11 September 2013

Revised 23 December 2013

Accepted 29 January 2014

Published Online First

18 February 2014

ABSTRACT

Objective To identify local meteorological drivers of dengue fever in French Guiana, we applied an original data mining method to the available epidemiological and climatic data. Through this work, we also assessed the contribution of the data mining method to the understanding of factors associated with the dissemination of infectious diseases and their spatiotemporal spread.

Methods We applied contextual sequential pattern extraction techniques to epidemiological and meteorological data to identify the most significant climatic factors for dengue fever, and we investigated the relevance of the extracted patterns for the early warning of dengue outbreaks in French Guiana.

Results The maximum temperature, minimum relative humidity, global brilliance, and cumulative rainfall were identified as determinants of dengue outbreaks, and the precise intervals of their values and variations were quantified according to the epidemiologic context. The strongest significant correlations were observed between dengue incidence and meteorological drivers after a 4–6-week lag.

Discussion We demonstrated the use of contextual sequential patterns to better understand the determinants of the spatiotemporal spread of dengue fever in French Guiana. Future work should integrate additional variables and explore the notion of neighborhood for extracting sequential patterns.

Conclusions Dengue fever remains a major public health issue in French Guiana. The development of new methods to identify such specific characteristics becomes crucial in order to better understand and control spatiotemporal transmission.

INTRODUCTION

Dengue virus, which is most commonly acquired through the bite of an *Aedes aegypti* mosquito, is the most important arthropod-borne viral disease affecting humans.¹ The increasing number of cases is associated with the expanding geographic range and the increasing intensity of transmission in affected areas.^{2,3} Recent estimates indicate 390 million infections per year worldwide, of which 96 million dengue infections per year are manifested.⁴ This virus has four serotypes—DENV-1, DENV-2, DENV-3, and DENV-4—although the existence of a fifth serotype has been discussed.⁵ The clinical forms of each serotype include asymptomatic infection, influenza-like illness, and severe forms—for example, fatal dengue hemorrhagic fever (DHF), dengue shock syndrome, encephalitis, and hepatitis. Even though several dengue vaccines are being developed,⁶ no vaccine or curative treatment is

currently available. Prevention strategies are limited to vector control, and treatment strategies are limited to supportive care to avoid shock syndrome.⁷

In Latin American and Caribbean countries, the reintroduction and dissemination of *A. aegypti* were observed in the 1970s after a reduction in vector control interventions that had been initiated in the 1960s. Since then, regular outbreaks have occurred on a 3–5-year cycle, and there has been an increase in severe forms of dengue, particularly DHF.⁸ In French Guiana, France's overseas territory in South America with 230 000 inhabitants, the epidemiology of dengue evolved from an endemo-epidemic to a hyper-endemic state.⁹ Five major epidemics linked to the circulation of one or two predominant serotypes have occurred over the last 10 years. These outbreaks usually last for 6–12 months and may affect nearly 10% of the population.

With the increasing frequency of epidemics and the resulting health, social, and economic impacts of dengue,¹⁰ the surveillance, control, and prevention of dengue have become social, political, and public health challenges that require specific preparedness activities.¹¹ One key element of an effective preparedness plan is the capacity to understand and predict the occurrence of dengue epidemics.

Epidemic dynamics are driven by complex interactions between intrinsic factors associated with human host demographics, vectors, and viruses, which drive multiannual dynamics, as well as extrinsic drivers, such as climate patterns, that potentially drive annual seasonality.

Previous investigators have created descriptive and predictive dengue models using various input variables,^{12–14} including climate data,^{15, 16} vector characteristics,^{17, 18} circulating viral serotypes, the immune status of the host population,¹⁵ or demographic data.^{19, 20} Even if the different studies in various affected areas do not always yield the same results, climatic variability is postulated to be one of the most important determinants of dengue epidemics; therefore, many studies have highlighted the influence of meteorological conditions on dengue incidence.²¹ The increase in temperature has been associated with dengue in Thailand,²² Indonesia,^{23, 24} Singapore,²⁵ Mexico,²⁶ Puerto Rico,²⁷ New Caledonia,²⁸ Guadeloupe,²⁹ and Sri Lanka.³⁰ An increase in humidity and high mosquito density increased the transmission rate of dengue fever in southern Taiwan.³¹ The abundance of predominant vectors is partly regulated by rainfall, which provides breeding sites and simulates egg hatching.^{32–36}

However, dengue patterns are dependent on the study area and are often characterized by non-linear



CrossMark

To cite: Flamand C, Fabregue M, Bringay S, et al. *J Am Med Inform Assoc* 2014;**21**:232–240.

dynamics, multi-annual oscillation, and irregular fluctuations in incidence; these factors complicate the understanding, detection, and prediction of both temporal and spatial transmission.

Data mining (ie, discovering useful, valid, unexpected, and understandable knowledge using databases) has been recognized as a promising new area for database research.³⁷ This area can be defined as efficiently discovering interesting information in large databases using statistical methods, database management techniques, and artificial intelligence.

Among the different data mining techniques, sequential pattern extraction³⁸ has received increased attention in recent years and has a wide range of applications in various areas, including finance, marketing, insurance, medical research, and sensor data. Traditional sequential pattern mining aims to extract sets of items that are commonly associated over time. However, this approach has rarely been applied to assess the spatiotemporal factors associated with infectious disease transmission.²⁰

The development of infectious disease surveillance in French Guiana in combination with technological advances in information systems offers new possibilities for applying data mining methods in future analyses.

We concentrated our efforts on applying sequential pattern mining to an epidemiological and meteorological dataset to identify potential drivers of dengue fever outbreaks. We used contextual sequential patterns, which extend the concept of traditional sequential patterns and were recently introduced by Rabatel *et al*.³⁹ to identify relationships. By considering the fact that a pattern is associated with one specific epidemiological or spatial context, the experts can then adapt their strategy depending on specific situations.

In this paper, we focus on the descriptive component, using different 'epidemiological contexts' to consider the impact of the interrelationships between dengue fever and climatic factors on specific epidemiologic figures. Our contribution is described in terms of methodology, epidemiological findings, and surveillance implications.

MATERIAL AND METHODS

Settings

French Guiana is located in South America between the Tropic of Cancer and the equator (4°00 north latitude and 53°00 west longitude); it is found between Brazil and Surinam. Its climate is typically tropical: hot and humid, with little variation in seasonal temperatures, heavy rainfall in the wet season from January to June, and low rainfall in the dry season from July to December. The relative humidity is high and varies between 80% and 90% according to the season. Primary health delivery differs according to location: in the coastal area, primary healthcare is delivered by 85 general practitioners (GPs), whereas further inland, care is provided by 17 public healthcare centers.⁴⁰

Epidemiological dataset

Epidemiologic data on dengue fever were obtained for the period from 2006 to 2011 from the multi-source surveillance system of the Regional Epidemiology Unit of the Institut de Veille Sanitaire (InVS).⁴⁰

Weekly numbers of biologically confirmed cases (BCCs), stratified according to the municipality of residence, were obtained from the laboratory surveillance system. This surveillance system, which collects individual information (including the patient's sex and age, area of residence, date of onset, date of blood sample, and results) from the seven laboratories

located in the coastal area, was authorized by the French Data Protection Agency (CNIL, N°1213498). In accordance with the CNIL, all of the data used in this study were aggregated so that they could not be associated with any specific individual.

The following criteria were used to define BCCs: virus isolation, viral RNA detection by reverse transcription-PCR (RT-PCR), detection of secreted NS1 protein, or a serological test based on an immunoglobulin M (IgM)-capture ELISA (MAC-ELISA).⁴¹ The dengue serotype data were identified for some of the BCCs (approximately 30% of the cases) by the National Reference Center (NRC) based at the Institut Pasteur in French Guiana (IPG).

Clinical case (CC) surveillance was set up from a sentinel network composed of 30 voluntary GPs located in the municipalities of the coastal area (representing approximately 35% of the GPs' total activity) and health centers located inland.⁴⁰ A CC was defined as a fever ($\geq 38^{\circ}\text{C}$) with no evidence of other etiology and associated with one or more non-specific symptoms, including headache, myalgia, arthralgia, and/or retro-orbital aches. The weekly number of CCs from 2006 to 2011 was included in the dataset.

For an outbreak in a given territory, we calculated the cumulative number of incident BCCs of dengue (BCC_i) and the clinical dengue incidence (CC_i) per week per 1000 residents. In the calculations, we assumed that the population of a territory was constant throughout a given year.

Weekly variation rates were calculated from the average of the four previous weeks for biological cases and CCs; 10th and 20th percentiles were used to classify the number of cases and the rates in 5 or 10 groups of similar size.

Meteorological dataset

Climatic records were obtained from Meteo France. Daily climate data, including cumulative rainfall (RR in mm), minimum and maximum temperatures (TN and TX in °C), sunstroke averages (INST in hours), wind strength at 10 meters (FXI in km/h), minimum and maximum relative humidity (UN and UX in %), and global brilliance (GLOT in KWh/m²/day), were collected from six meteorological stations (Cayenne, Kourou, Maripasoula, Matoury, Saint-Laurent, and Saint-Georges). From these daily data, weekly means were calculated throughout the study period. There were no missing values during this time period.

Weekly variation rates were calculated from the average of the four previous weeks for all of the meteorological indicators; 10th and 20th percentiles were used to classify the indicators and the rates in 5 or 10 groups of similar size.

Statistical analysis

The bivariate analyses were conducted using Stata V.12.⁴² The relationships between the epidemiological and meteorological data from 2006 to 2011 were studied at the national level of French Guiana and at different time scales using a Spearman rank correlation method. A p value <0.05 indicated statistical significance. On a weekly level, time-lagged correlation analyses (with a lag of 1–12 weeks) were performed on the time series of the weekly means of the meteorological variables and dengue incidence rates. Epidemic and non-epidemic years were compared to identify suitable meteorological patterns for dengue epidemics.

Contextual sequential pattern mining

The methodology involved three steps:

- ▶ Step 1: The spatiotemporal resolution and the epidemiological contexts were defined.
- ▶ Step 2: The sequence preprocessing module transformed the raw data into sequences of events.
- ▶ Step 3: The sequential patterns extraction module extracted frequent sequences of events for each context.

For the analyses performed after step 1, all the variables needed to fit the same spatiotemporal scale. A weekly temporal scale was used because weekly dengue surveillance data were available. The spatial distribution was based on homogeneous territories in terms of geographic distance and the movements of the population. Territories that consisted of several neighboring municipalities (figure 1) were established in collaboration with a local expert committee composed of epidemiologists, biologists, clinicians, entomologists, and specialists involved in the control and prevention of vector-borne diseases.⁴⁰

Five distinct epidemiological stages were defined by the expert committee⁴⁰:

- ▶ Stage 1: Sporadic transmission.
- ▶ Stage 2: Presence of dengue fever clusters in some areas.
- ▶ Stage 3: Pre-alert epidemic (when alert thresholds for CCs and BCCs are exceeded in the two following weeks).
- ▶ Stage 4: Confirmation of the epidemic (when thresholds are exceeded in the 2 weeks following the pre-alert epidemic).
- ▶ Stage 5: End of the epidemic.

For the subsequent analyses, five epidemiological phases were defined according to the different stages:

- ▶ Pre-epidemic (4 weeks preceding stage 3).
- ▶ Beginning of the epidemic (the first 4 weeks of stage 3).

- ▶ Ascending phase (from the 5th week following stage 3 to 4 weeks preceding the epidemic peak).
- ▶ Epidemic peak (from 3 weeks before to 3 weeks after the peak).

Descending phase (the end of the epidemic).

For each territory, the raw data included weekly epidemiological and meteorological data (table 1). For each week, the number of CCs and BCCs were known as well as the positivity rates of the blood samples, the values of local meteorological indicators, and the variation in the epidemiological and meteorological indicators. We defined contextual dimensions using either the epidemic or non-epidemic periods. We used 3-month periods (ie, quarter years) for the non-epidemic periods, and we used the epidemiological phases for the epidemic periods. We defined ‘epidemic’ or ‘non-epidemic’ weeks as *general contexts*, and the ‘pre-epidemic’ or ‘1st quarter of the year’ periods were denoted as *minimal contexts* in alignment with the hierarchies depicted in figure 2.

Each weekly value associated with a territory was called an *item* (eg, a maximum temperature variation of <-3% meant that the temperature decreased more than 3% compared to the previous 4 weeks). An *itemset* $it_1 = (i_1 \dots i_n)$ is a non-ordered set of items (eg, events that occurred during the same week). For example, a maximum temperature of 32.0–33.1°C, a maximum temperature variation of <-3%, and a cumulative rainfall of 85–158 mm is an itemset that indicates that for the designated week, the maximum temperature was between 32°C and 33.1°C, the maximum temperature decreased more than 3%, and the rainfall was between 85 and 158 mm.

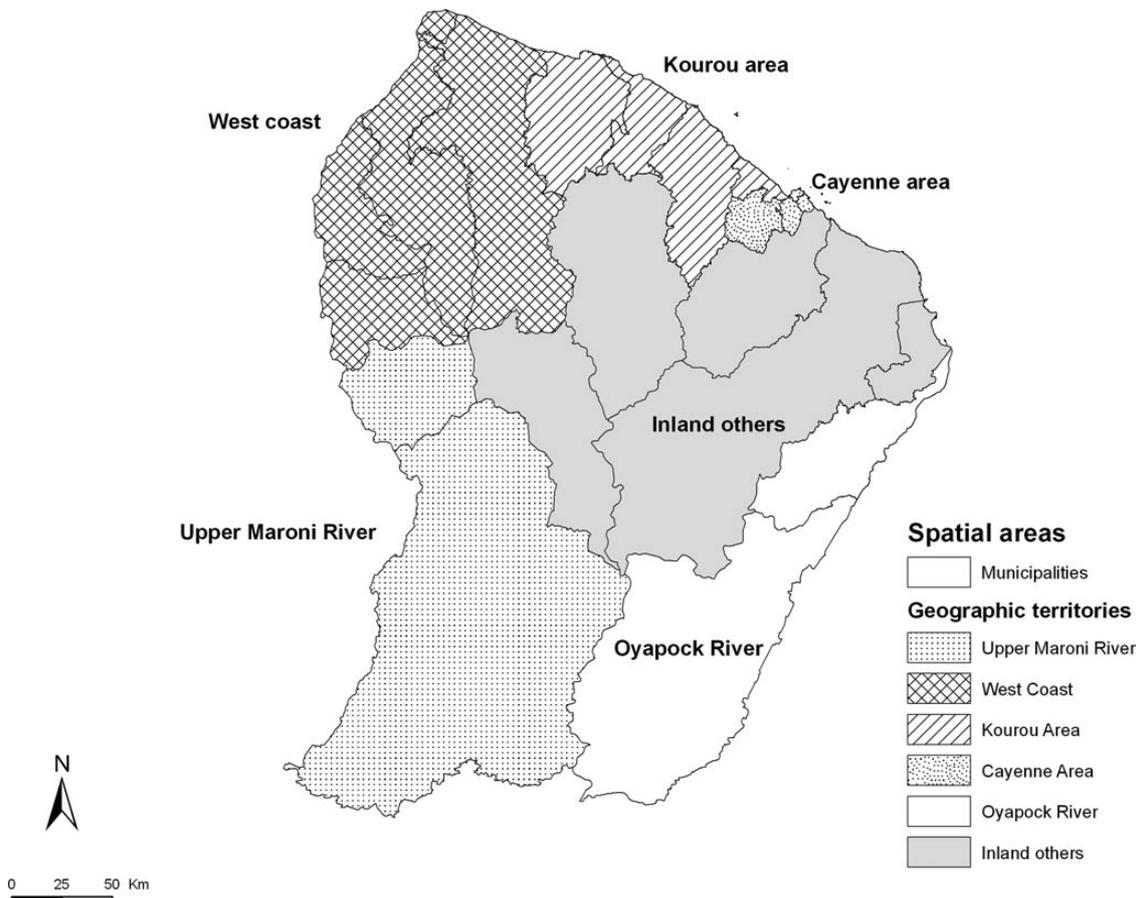


Figure 1 Spatial distribution of geographic territories for the dengue fever analysis, French Guiana, 2006–2011.

Table 1 Example of raw data, dengue fever, French Guiana, 2006–2011

Territory	Week	General context	Minimal context	BCC variation (%)	TX (°C)	TX. variation (%)	RR (mm)
T1	W2009 _{i-4}	Non-epidemic	2nd quarter	(-17; 0)	(32.0–33.1)	(-2; 0)	(85–158)
	W2009 _{i-3}	Non-epidemic	2nd quarter	(-17; 0)	(30.3–31.2)	<-3	(32–85)
	–	–	–	–	–	–	–
	W2009 _{i-1}	Non-epidemic	Pre-epidemic	(33; 80)	(30.3–31.2)	(-2; 0)	(158–327)
	W2009 _i	Epidemic	Beginning	>80	<30.3	(-2; -10)	(158–327)
	W2009 _{i+1}	Epidemic	Epidemic	>80	(30.3–31.2)	(-2; 0)	(85–158)
	–	–	Beginning	–	–	–	–
W2009 _{i+4}	Epidemic	Epidemic	(0; 33)	(31.2–32.0)	(0; 2)	(32–85)	
–	–	–	–	–	–	–	–
–	–	–	Epidemic peak	–	–	–	–
T2	W2010 _{i-4}	Non-epidemic	2nd quarter	(-17; 0)	>33.1	(-2; 0)	(32–85)
	W2010 _{i-3}	Non-epidemic	2nd quarter	(-17; 0)	(30.3–31.2)	<-3	(32–85)
	–	–	–	–	–	–	–
	W2010 _{i-1}	Non-epidemic	Pre-epidemic	(33; 80)	(31.2–32.0)	(-2; -10)	(158–327)
	W2010 _i	Epidemic	Beginning	>80	<30.3	(-2; -10)	(85–158)
	W2010 _{i+1}	Epidemic	Epidemic	(33; 80)	(30.3–31.2)	(-2; 0)	(85–158)
	–	–	Begin epidemic	–	–	–	–
W2010 _{i+4}	Epidemic	–	(-17; 0)	(30.3–31.2)	(0; 2)	<32	
–	–	–	Descending phase	–	–	–	

BCC, biologically confirmed case; RR, cumulative rainfall; TX, maximum temperature.

The second step consisted of transforming the raw data into sequences of events. The aim of this step was to build sequences by ordering the itemsets according to the week of their appearance during the periods of interest for the epidemiology of dengue.

The *sequence* $St_1 = \text{'(maximum temperature variation (< -3\%), rainfall variation (>157\%)) (number of clinical dengue cases variation >33\%)'}$ means that in territory T_1 , an increase in dengue cases >33% was preceded by maximum temperature decreases greater than 3% and associated with an increase in rainfall >157%. In step 2, we generated a sequence of events for territory 1 (see table 2).

We introduced constraints in this step to focus on more specific patterns that matched the specified domain constraints defined by the epidemiological and meteorological experts.

A constraint is a list of regular expressions, *exp*, separated by time intervals. An example of a constraint is $(exp1)[time1](exp2)[time2]:::[timek-1](expk)$, with *k* as the length of the constraint. For example, let P_c be a constraint and a time unit corresponding to a week, where $P_c = (UN) [1-3](CC)$. In other words, we extract all frequent patterns with a length of 2 (ie, the number of itemsets) where the characteristic humidity (UN) in the first itemset lasts for an interval of 1–3 weeks as well as

the number of CC. Table 2 provides some valid patterns according to this constraint.

The objective of step 3 was to build sequential patterns. Support for a pattern was obtained from the data sequences defined in step 1. For example (see table 2), the pattern $P \text{'(e2 e5)(e1)(e4)'}$ was included in two data sequences for zone T_1 . Thus, $support(P) = 2/4$.

To obtain the most frequent patterns, we used the PrefixSpan algorithm,⁴³ which extracts all the frequent *sequential* patterns according to the constraints defined. We only select patterns of size 1–3 with temporal intervals of 1–2 weeks between two itemsets. We also focus on patterns with at least one item related to the number of dengue cases in the given time interval. Support was calculated for all the minimal contexts of all the frequent patterns extracted. We considered that a pattern must have a support greater than 0.5 to be considered as a frequent pattern in a given minimal context.

The difference between the support of the pattern obtained in a context and the second highest support obtained in other contexts was calculated to provide a 'c-specificity' score to quantify the extent to which the pattern was specific to that context.³⁹ The sequential pattern extraction algorithms were applied using Weka Data Mining software.⁴⁴

Figure 2 Hierarchies of the epidemic and non-epidemic periods of dengue fever, French Guiana, 2006–2011.

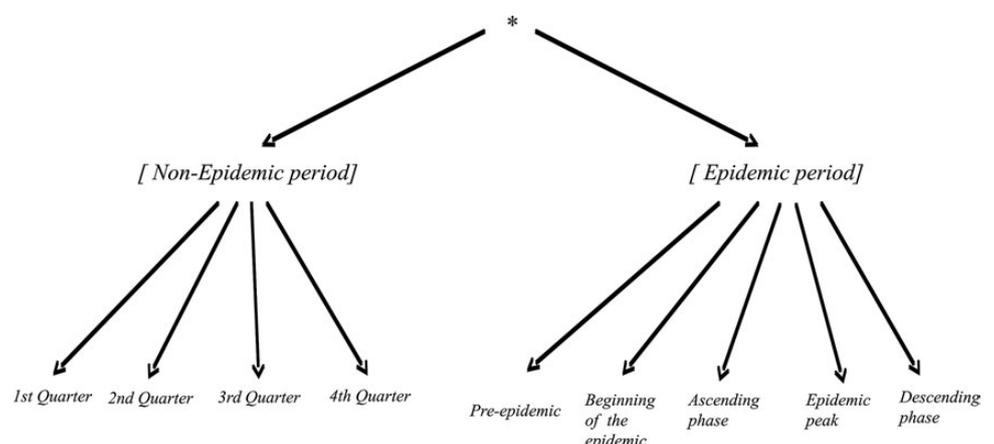


Table 2 Sequence of events for territory 1 for dengue fever, French Guiana, 2006–2011

Territory	Context	Associated event sequences
T1	Inter-epidemic	(e2 e3 e5)(e1)(e4)
	Inter-epidemic	(e5)(e2)(e4)
	Pre-epidemic	(e2 e5)(e1 e2)(e3 e4)
	Epidemic	(e3 e5)(e3)(e4)

Bold values are those selected in the extracted pattern cited in the example on the previous page.

RESULTS

Overall dengue incidence

From the beginning of 2006 to April 2011, 39 587 CCs and 11 133 BCCs were recorded in French Guiana. The national activity levels were strongly influenced by outbreak periods (figure 3). As shown in figure 3, three major outbreaks occurred during the study period. The average duration of these epidemics varied from 38 to 41 weeks.

Bivariate statistical analysis

During the study period, we found statistically significant positive correlations between dengue incidence and meteorological variables during the epidemic years for each family of variables (table 3, figure 4). The maximum correlation rates were obtained after a 4–6-week lag during the epidemic years.

Contextual sequential patterns extraction

The extracted sequential patterns showed temporal associations between local weather conditions, the evolution of dengue incidence, and time periods in the various territories of French Guiana.

Regardless of their position in the extracted sequential patterns, the meteorological variables were considered to have a relevant association; for example, an item included in an extracted pattern was considered to be associated with an epidemiological context whether it was in the 1st, 2nd, or 3rd itemset.

Outside epidemic periods, the 1st quarter of each year was characterized by minimum relative humidity greater than the median class (63–68%) (table 4). Low levels of incidence were frequently observed during this quarter, which was also marked by an increase in the number of clinical and BCCs without a high c-specificity score considering the evolution of the number of cases during outbreaks. This period was also marked by an increase in rainfall that was frequently associated with the

appearance of the 1st isolated clusters. The different epidemics in the study period all began during the 1st quarter of their respective years.

The 2nd and 3rd quarters were frequently associated with an increase in maximum temperatures, a decrease in the minimum relative humidity, and low levels of dengue incidence. The 4th quarter was marked by high maximum temperatures and low levels of rainfall. All of these results were compatible with the occurrence of the dry season. No specific evolution of dengue incidence was observed during this period.

Considering the fact that epidemic-period contexts were defined according to the epidemiological phases, items related to dengue incidence were frequently found in the epidemiological patterns (table 5). Nevertheless, our findings related to these items were compatible with the epidemiological phases defined by the local vector-borne disease expert committee.

The pre-epidemic periods were associated with a decrease in the maximum temperature (2–10% from the mean of the previous 4 weeks), a decrease in global brilliance (11–50%), and an increase in the minimum relative humidity (2–10%).

The beginning of an outbreak was frequently associated with a 4-week lag during which there was a strong increase in the minimum relative humidity (>40%), a decrease in the maximum temperature (–2 to 10%) (after a peak observed 1 or 2 months before the start of the epidemic), high levels of cumulative rainfall (158–327 mm), and a very slight increase in the maximum relative humidity. Similar to the pre-epidemic phase, a decrease in global brilliance was associated with the beginnings of the epidemics.

Importantly, epidemiological items included in the sequential patterns of the first two epidemic-period contexts suggested a premature evolution of the BCCs compared to the increase in CCs before the ascending phase of the epidemic. Dengue incidence-related items were frequently found in the sequential patterns extracted from the epidemic period contexts.

Except for the increase in global brilliance (between 62% and 67%) at the pre-epidemic peak, the evolution of specific weather conditions was not included in the sequential patterns that were associated with the phases surrounding the epidemic peak, where a predominance of the cumulative incidence occurred.

DISCUSSION

Sequential pattern mining is an important method that has been widely used by the data mining community in many different types of applications. In this paper, we have presented the critical steps of a data-mining project which will allow better understanding and prediction of temporal dynamics of dengue fever

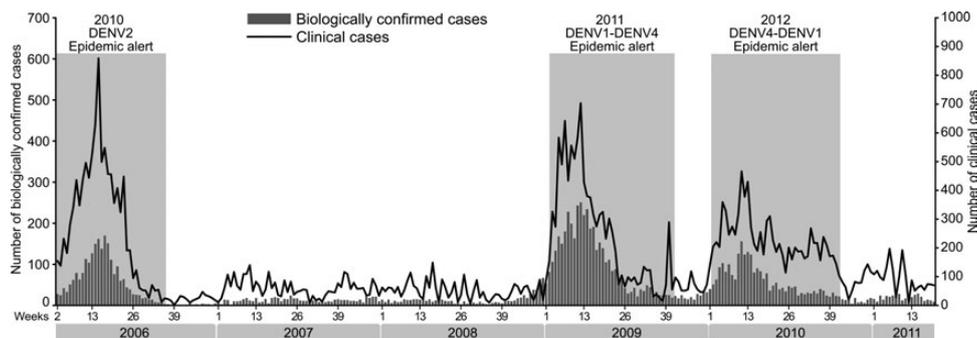


Figure 3 Weekly number of biologically confirmed and clinical cases of dengue fever and outbreak periods, French Guiana, January 2006–April 2011.

Table 3 Correlations between meteorological variables and dengue incidence

	Non-epidemic years		Epidemic years				
	Lag2wk	Lag3wk	Lag2wk	Lag3wk	Lag4wk	Lag6wk	Lag8wk
RR	0.06 -0.214*	0.01 0.275*	0.485*** 0.456***	0.498*** 0.465***	0.509*** 0.486***	0.498*** 0.474***	0.375*** 0.384***
TN	0.105 -0.041	0.171 0.04	0.501*** 0.521***	0.515*** 0.522***	0.516*** 0.528***	0.483*** 0.487***	0.436*** 0.428***
TX	-0.206* 0.144	-0.18* 0.191*	-0.678*** -0.693***	-0.702*** -0.703***	-0.716*** -0.721***	-0.646*** -0.670***	-0.502*** -0.549
INST	-0.167 0.152	-0.098 0.221*	-0.591*** -0.573***	-0.632*** -0.598***	-0.649*** -0.620***	-0.634*** -0.607***	-0.538*** -0.538***
FXI	0.284** 0.025	0.204* 0.009	0.338*** 0.397***	0.378*** 0.411***	0.405*** 0.441***	0.435*** 0.481***	0.431*** 0.475***
UN	0.191* -0.200*	0.114 -0.262*	0.563*** 0.519***	0.584*** 0.535**	0.611*** 0.556***	0.568* 0.514***	0.454*** 0.405***
UX	-0.252** -0.197*	-0.218* -0.226**	-0.269** -0.496***	-0.268** -0.490***	-0.260** -0.479***	-0.234** -0.482	-0.192* -0.457***
GLOT	-0.230*** 0.067	-0.166** 0.186*	-0.527*** -0.502***	-0.580*** -0.540***	-0.622*** -0.582***	-0.626*** -0.600***	-0.562*** -0.543***

Spearman's rank correlation test (*r*, significance score of *p* value).
 The first row represents correlation between the meteorological variable and clinical cases (CC) incidence. The second row represents correlation with biologically confirmed cases (BCC) incidence.
 Significance score: **p*<10⁻², ***p*<10⁻³, ****p*<10⁻⁴.
 FXI, wind strength; GLOT, global brilliance; INST, sunstroke average; Lag2wk, lag 2 weeks; RR, cumulative rainfall; TN/TX, minimum and maximum temperature; UN/UX, minimum and maximum relative humidity.

in French Guiana. In particular, we applied an algorithm for contextual sequential pattern extraction to identify the most important climatic factors related to dengue fever in French Guiana.

Our results suggest that the local climate has major effects on the occurrence of dengue epidemics in French Guiana and well known climatic factors were found as determinants in outbreak occurrence.

The correlation rates obtained from the clinical dengue incidence rates were compatible with the rates obtained from biologically confirmed dengue cases. The maximum correlation rates were obtained after a 4–6-week lag during the epidemic years. These findings are compatible with mosquito biology and the viral transmission cycle.

Maximum temperature, minimum relative humidity, global brilliance, and cumulative rainfall were identified as determinants of dengue epidemics, and the intervals of their values were quantified. For instance, the level of cumulative rainfall was frequently associated with the beginning of outbreaks (RR 158–327 mm), suggesting that dengue epidemics are associated with a rainfall level that was relatively high but not too extreme (which would destroy breeding sites via a ‘washing effect’).

The approach we developed helped us to explore the dataset by bringing various descriptive and analytical results together. The contextual analysis allowed us to make comparisons between temporal or spatial subgroups by identifying the most discriminating categories and anticipating possible classifications or typologies for situations. Compared with traditional models, such an approach is particularly useful for two main reasons: it can provide relevant insights that account for various temporal intervals and spatial units, and it is quite appropriate for comparing situations that can constrain analysts to multiply stratified analyses with traditional methods. The situations observed in French Guiana were particularly heterogeneous in space (ie, a small amount of the population lives in the Amazonian land area where the presence of vectors is low, and the urban coastal

area is home to 90% of the population) and time (ie, different seasons); thus, they were well suited to contextual approaches.

Another advantage is that the approach allows the simultaneous analysis of associations between many outcomes and various explanatory variables. For instance, we studied the associations between meteorological variables and CCs and BCCs while also exploring reactivity and the evolution of one indicator versus the others.

However, our study has several limitations. First, well-known factors that were not included in our dataset may have contributed to the epidemic dynamics in the different territories. We did not include any direct measurements of vector behavior as input variables; for example, mosquito prevalence, vector behavior, the presence of potential or confirmed breeding sites, or the prevalence of the dengue virus in the vector. Although this information is particularly important for estimating the transmission risk of vector-borne diseases, it requires intensive financial, laboratory, and technical resources that are usually not available in routine practice in the territory and for long time periods.

Other factors that can play a key role in transmission, such as environmental characteristics, social and demographic indicators, or the immune status of host populations, could not be explored in our study because the data were unavailable or not available in a temporal and spatial format. In the absence of seroprevalence data, future studies need to consider the population age distribution and human movement patterns to approximate the role of the immune status of the population. An older and thus more immune population reduces the probability of a vector feeding on a susceptible or infectious person, both of which drive transmission.

Second, the defined contexts were based on temporal periods and did not allow for the identification of possible spatial differences between the climatic drivers of dengue in the various territories.

Future work should integrate additional variables and create new contexts. Remote sensing data are currently being collected

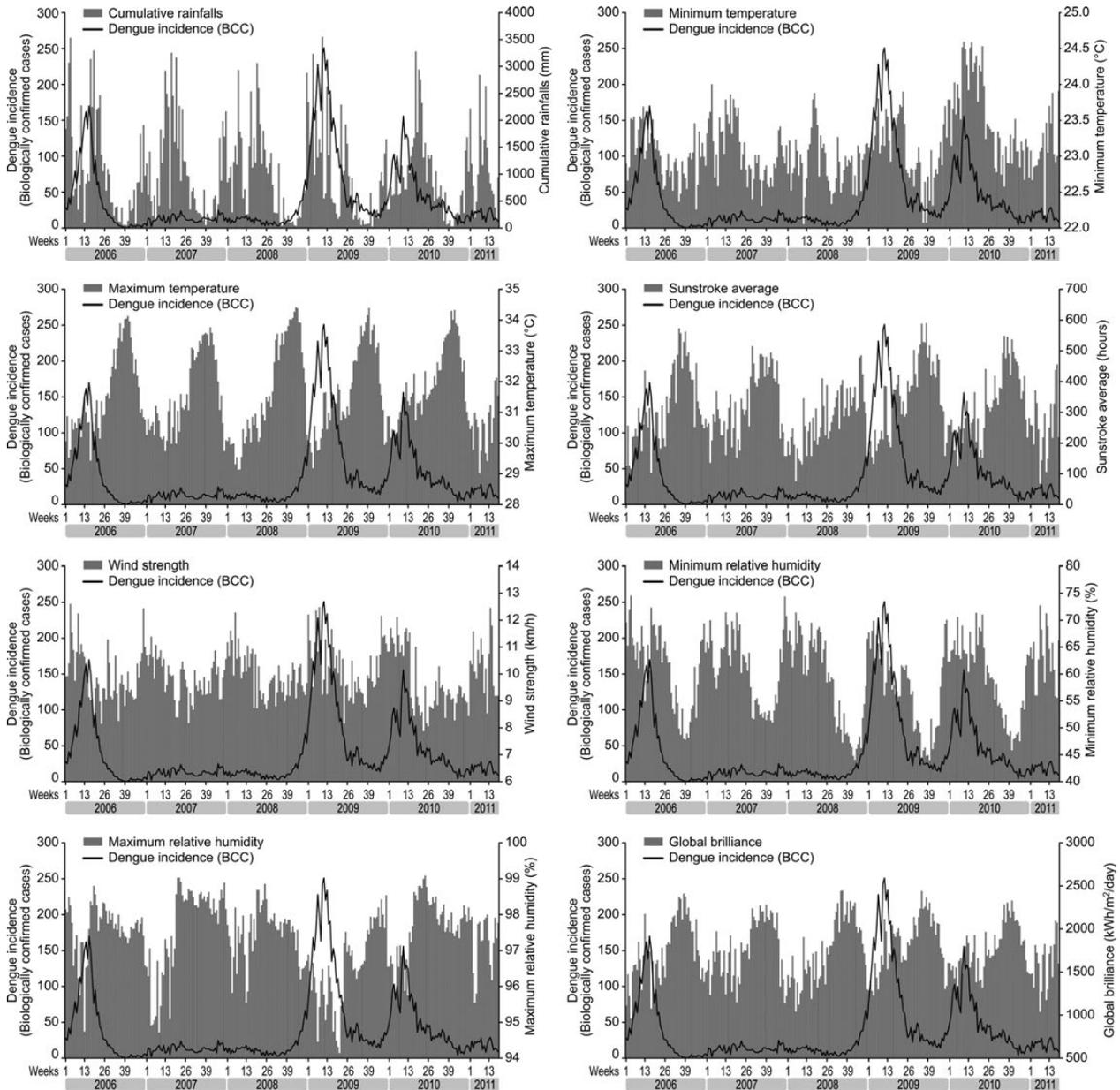


Figure 4 Weekly incidence of dengue fever (biologically confirmed cases, BCC) in French Guiana from April 2006 to April 2011 compared to crude meteorological variables for the same period: (A) cumulative rainfall; (B) minimum temperature; (C) maximum temperature; (D) sunstroke average; (E) wind strength; (F) minimum relative humidity; (G) maximum relative humidity; (H) global brilliance.

and may provide very useful information about environmental factors, such as types of habitats and types of areas (city centers, spontaneous settlements, road borders, collective buildings, individual houses with gardens, etc). Future studies should also include geographic areas as contexts to estimate the existing differences between the various regions of French Guiana. The results will help to target the territories in which the predictive models could be implemented to anticipate the risk of transmission of dengue fever. Creating hierarchies between the various contexts will enable researchers to estimate the contributions of the spatial or temporal units and consequently differentiate the most relevant contexts for developing predictive models. Among other possible future developments, we plan to take into account the notion of neighborhood in the extraction of the sequential patterns. A new method recently described by Alatavista *et al*⁴⁵ highlighted an extension of sequential patterns, called new spatio-sequential patterns, for analyzing the

evolution of areas considering their neighboring environment. Furthermore, the extraction could be used with the aim of determining the geographic clustering in French Guiana to identify the relevant spatial units for characterizing, monitoring, and predicting the local transmission of dengue. Accurate prediction of dengue outbreaks may lead to useful public health interventions. The final aim of our research project will be the development of predictive tools that allow for spatial identification of specific high risk areas whilst taking into account the temporal dynamics of dengue transmission.

CONCLUSION

Dengue remains a major public health issue in French Guiana. Our findings highlight the utility of the data mining approach to analyze disease surveillance data on a temporal and a spatial scale in relation to climatic, social, and environmental variables. Despite the heightened awareness among

Table 4 Non-epidemic contextual sequential patterns

Minimal context	Non-epidemic associated sequential patterns	Support	C-specificity
1st quarter (non-epidemic period)	(UN (63–68%)) (RR (85–158 mm))	0.72	0.22
	(UN (63–68%)) (TX<30°C)	0.67	0.57
	(UN (63–68%)) (UN>68%)	0.64	0.17
	(UN (63–68%)) (Var_CC>33%)	0.64	0.14
	(UN (63–68%)) (RR (158–327 mm))	0.57	0.09
	(UN (63–68%)) (Var_BCC>33%)	0.64	0.07
2nd quarter (non-epidemic period)	(Var_TX>2%) (UN>68%)	0.59	0.19
	(Var_TX>2%; Var_UN>7%)	0.74	0.19
	(Var_TX>2%; BCC (0;2))	0.74	0.09
	(TX (30.3–31.2°C))	0.89	0.08
	(Var_TX>2%) (TN (23.2–23.8°C))	0.63	0.08
	(Var_TX>2%) (Var_FXI<7%)	0.63	0.08
3rd quarter (non-epidemic period)	(Var_TX>2%) (RR<32 mm)	0.87	0.42
	(FXI<8.7 km/h)	0.65	0.14
	(Var_TX (0.7%; 1.9%))	0.94	0.13
	(Var_UN<–6%)	0.97	0.12
	(CC=0)	0.87	0.11
	(TN (22.4–22.8°C))	0.74	0.11
4th quarter (non-epidemic period)	(TX>33.1°C) (Var_TX (–2–0%))	0.85	0.59
	(TX>33.1°C) (TX (32.0–33.1°C))	0.76	0.56
	(TX>33.1°C) (Var_UN>7%)	0.82	0.53
	(TX>33.1°C) (Var_BCC (–17–0%))	0.79	0.27
	(TX>33.1°C)	0.94	0.16
	(TX>33.1°C) RR<32 mm)	0.82	0.11

BCC, biologically confirmed case; CC, clinical case; RR, cumulative rainfall; TN/TX, minimum and maximum temperature; UN/UX, minimum and maximum relative humidity.

Table 5 Epidemic contextual sequential patterns

Minimal context	Epidemic associated sequential patterns	Support	C-specificity
Pre-epidemic (4-week period)	(Var_TX ^o (–2% to –10%), CC _i <1%, BCC _i <0.3%) (CC _i <1%, BCC _i <0.3%)	0.57	0.57
	(Var_UN (2–7%), CC _i <1%, BCC _i <0.3%)	0.67	0.52
	(CC _i <1%, BCC _i <0.3%, Var_GLOT (–11% to –50%)) (CC _i <1%, BCC _i <0.3%)	0.57	0.57
	(Var_TX ^o (–2% to –10%), CC _i <1%, BCC _i <0.3%)	0.62	0.48
	(CC _i <1%, BCC _i <0.3%) (Var_BCC>40%, CC _i <1%, BCC _i <0.3%)	0.57	0.48
	(Var_UX (0.1–0.4%) CC _i <1%, BCC _i <0.3%)	0.57	0.43
	(Var_UN>7%, CC _i <1%, CC _i <1%)	0.57	0.38
	(Var_BCC>40%, CC _i <1%, BCC _i <0.3%) (CC _i <1%, BCC _i <0.3%)	0.57	0.38
	(Var_BCC>40%) (BCC _i (0.3–1.9%))	0.76	0.56
	(BCC _i (0.3–1.9%)) (BCC _i (0.3–1.9%))	0.86	0.51
Beginning of epidemic (4-week period)	(BCC _i (0.3–1.9%)) (Var_UX (0.1%; 0.4%))>	0.71	0.46
	(Var_BCC>40%) (RR (158–327 mm))	0.67	0.18
	(Var_BCC>40%) (Var_CC (0–33%))	0.81	0.16
	(Var_BCC>40%) (Var_UN (7–40%))	0.67	0.09
	(Var_BCC>40%) (UN (62–67%))	0.57	0.05
	(Var_GLOT>12%)	0.62	0.02
	(Var_BCC>40%) (Var_UX (0.1–0.4%))	0.62	0.01
	(UX<96%)	0.57	0.01
	(BCC>8) (TX (30.3°; 31.2°), BCC>8)	0.6	0.25
	(BCC _i (1.8%; 4.3%)) (BCC _i (1.8%; 4.3%))	0.6	0.17
Epidemic peak (7-week period)	(UN (62–67%)) (BCC>8)	0.65	0.10
	(Var_BCC (1–40%))	0.8	0.04
	(Var_GLOT>(3–12%))	0.65	0.04
	(Var_BCC<–33%) (Var_UN (2–7%))	0.85	0.30
	(Var_BCC<–33%) (Var_TX (0%))	0.85	0.30
Descendant phase	(Var_BCC<–33%) (Var_BCC<–33%)	0.90	0.23
	(Var_CC (–4–0%))	0.70	0.21
	(Var_BCC<–33%) (Var_TX>2%)>	0.85	0.20

BCC, biologically confirmed case; CC, clinical case; GLOT, global brilliance; RR, cumulative rainfall; TX, maximum temperature; UN/UX, minimum and maximum relative humidity.

health authorities of the importance of dengue prevention and vector control, various challenges still exist to better understand and accurately predict dengue epidemics. Better understanding of dengue epidemics is necessary for public

health interventions to mitigate the effect of these outbreaks, particularly in areas where resources are limited and where the medical infrastructure may become overwhelmed by significant epidemics.

Acknowledgements We are grateful to all of the collaborators involved in the surveillance system monitored by the Regional Epidemiology Unit. We wish to thank Dr Alain Bouix and Dr Stanley Carroll, coordinators of the GP's sentinel network; all the biological laboratories; Dr Dominique Rousset and Dr Séverine Matheus, virologists at the National Reference Center based at the Institut Pasteur in French Guiana; Dr Felix Djossou and Christelle Prince from the Infectious Tropical Disease Unit of the Hospital Center of Cayenne; and Dr Muriel Ville, who coordinates the Health Centers, for their help in the epidemiologic and virology data collection. We wish to thank Philippe Palany, Jean-Louis Maridet, and Christian Brevignon from Meteo France for their help with the meteorological data collection. We thank Laurel Zmolek-Smith from the University of Iowa for linguistic support.

Contributors CF conducted the study, performed data analysis, and drafted the manuscript. MF performed data mining analysis. VA contributed to the collection, classification, and interpretation of epidemiologic data. SB and MT contributed to the conception and design of the study and helped to draft the manuscript. PQ and J-CD contributed to the epidemiologic interpretation of the results of the study and helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

- Gubler DJ. The global emergency/resurgence of arboviral diseases as public health problems. *Arch Med Res* 2002;334:330–42.
- Guzman MG, Halstead SB, Artsob H, et al. Dengue: a continuing global threat. *Nat Rev Microbiol* 2010;8(Suppl):S7–16.
- WHO. *Dengue and severe dengue*. Fact Sheet No. 117, 2013. Geneva: World Health Organisation. <http://www.who.int/mediacentre/factsheets/fs117/en/>
- Bhatt S, Gething PW, Brady OJ, et al. The global distribution of dengue. *Nature* 2013;496:504–7.
- Normile D. Tropical medicine. Surprising new dengue virus throws a spanner in disease control efforts. *Science* 2013;342:415.
- Guy B, Saville M, Lang J. Development of Sanofi Pasteur tetravalent dengue vaccine. *Hum Vaccin* 2010;6:696–705.
- Beatty ME, Stone A, Fitzsimons D, et al. Best practices in dengue surveillance: a report from the Asia-Pacific and Americas Dengue Prevention Boards. *Plos Negl Trop Dis* 2010;4:e890.
- Halstead SB. Dengue in the Americas and Southeast Asia: do they differ? *Rev Panam Salud Publica* 2006;20:407–15.
- Quenel P, Dussart P, Marrama L, et al. Contributions de la recherche virologique, clinique, épidémiologique, socio comportementale et en modélisation mathématique au contrôle de la dengue dans les DFA. *Bulletin de Veille Sanitaire* 2009;3:1–16.
- Torres JR, Castro J. The health and economic impact of dengue in Latin America. *Cad Saude Publica* 2007;22(Suppl 1):S23–31.
- IMS Dengue. *Cire Antilles-Guyane: integrated management strategy for dengue prevention and control in the Caribbean subregion*. *Bull de Veille Sanit Antilles Guyane* 2009;8:2–15. http://www.invs.sante.fr/publications/bvs/antilles_guyane/2009/bvs_ag_2009_08.pdf
- Focks DA, Haile DG, Daniels E, et al. Dynamic life table model for *Aedes aegypti* (Diptera: Culicidae): analysis of the literature and model development. *J Med Entomol* 1993;30:1003–17.
- Focks DA, Daniels E, Haile DG, et al. A simulation model of the epidemiology of urban dengue fever: literature analysis, model development, preliminary validation, and samples of simulation results. *Am J Trop Med Hyg* 1995;53:489–506.
- Racloz V, Ramsey R, Tong S, et al. Surveillance of dengue fever virus: a review of epidemiological models and early warning systems. *PLoS Negl Trop Dis* 2012;6:e1648.
- Barbazan P, Yoksan S, Gonzalez JP. Dengue hemorrhagic fever epidemiology in Thailand: description and forecasting of epidemics. *Microbes Infect* 2002;4:699–705.
- Otero M, Solari HG. Stochastic eco-epidemiological model of dengue disease transmission by *Aedes aegypti* mosquito. *Math Biosci* 2010;223:32–46.
- Bartley LM, Donnelly CA, Garnett GP. The seasonal pattern of dengue in endemic areas: mathematical models of mechanisms. *Trans R Soc Trop Med Hyg* 2002;96:387–97.
- Wearing HJ, Rohani P. Ecological and immunological determinants of dengue epidemics. *Proc Natl Acad Sci USA* 2006;103:11802–7.
- Esteva L, Vargas C. A model for dengue disease with variable human population. *J Math Biol* 1999;38:220–40.
- Buczak AL, Koshute PT, Babin SM, et al. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. *BMC Med Inform Decis Mak* 2012;12:124.
- Halstead SB. Dengue virus-mosquito interactions. *Ann Rev Entomol* 2008;53:273–91.
- Focks D, Alexander N, Villegas E. Multicountry study of *Aedes aegypti* pupal productivity survey methodology: findings and recommendations. *Dengue Bull WHO* 2007;31:192–200.
- Arcari P, Tapper N, Pfueller S. Regional variability in relationships between climate and dengue/DHF in Indonesia. *Singap J Trop Geogr* 2007;28:251–72.
- Bangs M, Larasati R, Corwin A, et al. Climatic factors associated with epidemic dengue in Palembang, Indonesia: implications of short-term meteorological events on virus transmission. *Southeast Asian J Trop Med Public Health* 2006;37:1103–16.
- Burattini M, Chen M, Chow A, et al. Modelling the control strategies against dengue in Singapore. *Epidemiol Infect* 2007;136:309–19.
- Chowell G, Sanchez F. Climate-based descriptive models of dengue fever: the 2002 epidemic in Colima, Mexico. *J Environ Health* 2006;68:40–4.
- Keating J. An investigation into the cyclical incidence of dengue fever. *Soc Sci Med* 2001;53:1587–97.
- Descloux E, Mangeas M, Menkes CE, et al. Climate-based models for understanding and forecasting dengue epidemics. *PLoS Negl Trop Dis* 2012;6:e1470.
- Gharbi M, Quenel P, Gustave J, et al. Time series analysis of dengue incidence in Guadeloupe, French West Indies: forecasting models using climate variables as predictors. *BMC Infect Dis* 2011;11:166.
- Goto K, Kumarendran B, Mettananda S, et al. Analysis of effects of meteorological factors on dengue incidence in Sri Lanka using time series data. *PLoS ONE* 2013;8:e63717.
- Chen SC, Liao CM, Chio CP, et al. Lagged temperature effect with mosquito transmission potential explains dengue variability in southern Taiwan: insights from a statistical analysis. *Sci Total Environ* 2010;408:4069–75.
- Corwin A, Larasati R, Bangs M, et al. Epidemic dengue transmission in southern Sumatra, Indonesia. *Trans R Soc Trop Med Hyg* 2001;95:257–65.
- Chadee D, Shivnauth B, Rawlins S, et al. Climate, mosquito indices and the epidemiology of dengue fever in Trinidad (2002–2004). *Ann Trop Med Parasitol* 2007;101:69–77.
- Barrera R, Delgado N, Jiménez M, et al. Stratification of a city with hyperendemic dengue hemorrhagic fever. *Rev Panam Salud Publica* 2000;8:225–33.
- Depradine C, Lovell E. Climatological variables and the incidence of Dengue fever in Barbados. *Int J Environ Health Res* 2004;14:429–41.
- Nakhapakorn K, Tripathi NK. An information value based analysis of physical and climatic factors affecting dengue fever and dengue haemorrhagic fever incidence. *Int J Health Geogr* 2005;4:13.
- Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, eds. *Advances in knowledge discovery and data mining*. AAAI/MIT Press, 1996.
- Agrawal R, Srikant R. Mining sequential patterns. In: Yu PS, Chen ASP, eds. *Eleventh International Conference on Data Engineering*; IEEE Computer Society Press, 1995.
- Rabatel J, Bringay S, Poncelet P. Contextual Sequential Pattern Mining. In: *IEEE 2010 IEEE International Conference on Data Mining Workshops*. 2010:981–8.
- Flamand C, Quenel P, Ardillon V, et al. The epidemiologic surveillance of dengue fever in French Guiana: when achievements trigger higher goals. *Stud Health Technol Inform* 2011;169:629–33.
- Dussart P, Petit L, Labeau B, et al. Evaluation of two commercial tests for the diagnosis of acute dengue virus infection using NS1 antigen detection in human serum. *PLoS Negl Trop Dis* 2008;2:e280.
- StataCorp. *Stata statistical software: release 12*. College Station, TX: StataCorp LP, 2011.
- Pei J, Han J, Mortazavi-Asl B, et al. Mining sequential patterns by pattern-growth: the PrefixSpan approach. *IEEE Trans Knowledge Data Eng* 2004;16–11:1424–40.
- Hall M, Eibe F, Holmes G, et al. The WEKA data mining software: an update. *SIGKDD Explorations*. 2009;11:10–18.
- Salas HA, Bringay S, Flouvat F, et al. The pattern next door: Towards spatio-sequential pattern discovery. *16th PAKDD*, 2012.



Sequential patterns mining and gene sequence visualization to discover novelty from microarray data

A. Sallaberry^a, N. Pecheur^b, S. Bringay^{b,c}, M. Roche^b, M. Teisseire^{d,*}

^aLaBRI, INRIA Bordeaux Sud-Ouest, Pikko, 351, cours de la Libération, 33405 Talence Cedex, France

^bLIRMM, Univ. Montpellier 2 - CNRS, 161 rue Ada 34095 Montpellier Cedex 5 France

^cMIAp Department, Univ. Montpellier 3, route de Mende 34199 Montpellier cedex 5, France

^dCemagref, UMR TETIS, Maison de la teledetection, 500 rue Jean-François Breton, 34093 Montpellier, France

ARTICLE INFO

Article history:

Received 30 August 2010

Available online 16 April 2011

Keywords:

Visualization

Data mining

Bioinformatics

Sequential patterns

Microarray data

Gene data

ABSTRACT

Data mining allow users to discover novelty in huge amounts of data. Frequent pattern methods have proved to be efficient, but the extracted patterns are often too numerous and thus difficult to analyze by end users. In this paper, we focus on sequential pattern mining and propose a new visualization system to help end users analyze the extracted knowledge and to highlight novelty according to databases of referenced biological documents. Our system is based on three visualization techniques: clouds, solar systems, and treemaps. We show that these techniques are very helpful for identifying associations and hierarchical relationships between patterns among related documents. Sequential patterns extracted from gene data using our system were successfully evaluated by two biology laboratories working on Alzheimer's disease and cancer.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

DNA microarrays have been successfully used for many applications (e.g. diagnosis and characterization of physiological states). They allow researchers to compare gene expression in different tissues, cells or conditions [1] and provide some information on the relative expression levels of thousands of genes that are compared in a few samples, usually less than a hundred (e.g., Affymetrix U-133 plus 2.0 microarrays measure 54,675 values). Nevertheless, due to the huge amount of data available, how process and interpret them to make biomedical sense of them remains a challenge. Data mining techniques, such as [2–4], have played a key role in discovering previously unknown information and shown that they can be very useful to biologists in identifying relevant subsets of microarray data for further analysis [5].

However, the number of results is usually so huge that they cannot easily be analyzed by the experts concerned. In [6], we proposed a general process, called GeneMining, based on the DBSAP algorithm for extracting sequential patterns from DNA microarrays [7]. We obtained patterns of correlated genes ordered according to their level of expression. Although this method is useful, the way to

select relevant patterns is still not highly efficient. For instance, depending on the values of parameters, between 1000 and 100,000 patterns may be extracted, which are not easy to interpret. Thus, the main aim of this new work is to propose new visualization techniques to help biologists to navigate through the extracted patterns. Biologists are also faced with the problem of locating relevant publications about the genes involved in the patterns. Even if some tools are now available to automatically extract information from microarray data (e.g., [8] or [9]), there are still no user-friendly literature search tools available to analyze patterns.

In this paper, we describe an efficient tool to help biologists focus on new knowledge by navigating through large numbers of sequential patterns (i.e., sequences of ordered genes). Our contribution is twofold. First, we adapt three different techniques (i.e., point clouds, solar systems, and treemaps) to deal with data organized as a sequence and to produce an effective solution to the above problem. Second, using our system, the biologist can now be automatically provided with relevant documents extracted from the PubMed/MEDLINE database.¹ Although the methods described in this paper mainly focus on sequences extracted from DNA microarrays, they could easily be adapted to any other kind of sequential data.

The paper is organized as follows. In Section 2, we describe the data we are working with and give an overview of related

* Corresponding author. Fax: +33 467 548 700.

E-mail addresses: arnaud.sallaberry@labri.fr (A. Sallaberry), pecheur@lirmm.fr (N. Pecheur), bringay@lirmm.fr (S. Bringay), mroche@lirmm.fr (M. Roche), maguelonne.teisseire@cemagref.fr (M. Teisseire).

¹ www.ncbi.nlm.nih.gov/pubmed.

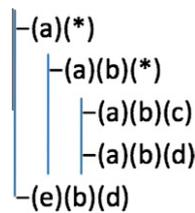


Fig. 1. Representation of the hierarchy.

work. In Section 3, we describe our proposal and the associated tool. In Section 4, we evaluate the new systems, and in Section 5 we present our conclusions and future work.

2. Preliminaries

In the framework of the Gene Mining² project (PEPS project funded by ST2I Institute of CNRS – France), we mined real data produced by the analysis of DNA microarrays (Affymetrix DNA U133 plus 2.0) to study Alzheimer’s disease (AD) using the DBSAP algorithm [7]. This dataset was used to discover classification tools to distinguish between two classes (AD and healthy individuals). In [7], we proposed to extract patterns of correlated genes ordered according to their level of expression. An example of pattern is $\langle\langle MRV1 \rangle\rangle(PGAP1, GSK3B)$ meaning that “the level of expression of gene *MRV1* is lower than that of genes *PGAP1* and *GSK3B*, whose levels are very similar”.

Although this method was useful since it proved that sequential patterns could be very useful for biologists, the way of selecting relevant patterns remained a challenge. Actually, depending on the values of parameters, 1000 to 100,000 patterns could be extracted and were consequently not easy to interpret. Biologists still needed a visualization tool to enable them to navigate through the huge amount of sequences, to select and order relevant novel sequences (e.g. sequences in which new gene correlations may exist), and to automatically query specific publications from Pubmed/MEDLINE (or other publication database) on the selected genes.

To summarize, an appropriate visualization tool needs to explore both kinds of data:

1. Gene sequences described by an ordered list of sets of genes and the class *supports* (i.e., the number of occurrences of this class in the database respecting this expression). As already mentioned, too many patterns are extracted. By using the k-means clustering algorithm with a sequence-oriented measure (S2MP [10]), we are able to identify groups of similar sequences and highlight a representative sequence called the centre. According to the centre, sequences can also be summarized [11] and organized according to a *hierarchy*. For example, the three sequences $\langle\langle a \rangle\rangle(b)(c)$, $\langle\langle a \rangle\rangle(b)(d)$ and $\langle\langle e \rangle\rangle(b)(d)$ can be presented as a tree (Fig. 1). The two first are summarized by $\langle\langle a \rangle\rangle(b)(*)$.
2. Documents in the literature dealing with genes from sequences. The documents are obtained from the bibliographical database Pubmed-MEDLINE (i.e. free digital archive of biomedical and life sciences literature) with or without gene synonyms [6]. We define a distance between a document and the gene sequence taking into account the publication date as well as the number of genes mentioned in the paper. The more recent the document and the more genes described in the paper, the closer the document will be to the sequence concerned.

The visualization tool, which is described in the following section, combines all these elements: support, class, groups, hierarchy, and sets of documents. To facilitate specific tasks, we propose three different visual representations [12]. The “Point Cloud” representation is mainly used to show the set of sequences while the “Solar System” is mainly used to focus on a specific sequence. Finally, the treemap is very useful when hierarchies and volumes have to be represented.

The combination of these representations allows the user to explore gene sequences efficiently and to identify relevant information. A typical use of the application consists in looking at the clusters and identifying those containing particular genes of interest. The user can visualize interesting clusters in more detail and select the sequences that appear to be the most relevant according to their support and the users previous conjectures. Users also need easy access to the bibliography related to a particular sequence to (in)validate their arguments. Indeed, they need to access the supports of higher levels of a sequence in the hierarchy to evaluate the potential role of each gene in this sequence and to access the groups containing sequences beginning with high levels elements of the hierarchy

In [13], a visualization tool based on point clouds representing groups of sequences is proposed. Sequences are placed according to an alignment in a 3-dimensional space. However, this approach is not able to account for the hierarchy of sequences. Indeed, most previous works concerning visualization of biological sequences focus on the representation of sequence alignments [14–16]. To the best of our knowledge, no method is currently available to visualize sequences and associated documents, as most previous works deal with visualization of parts of a document [17], information about documents [18] or a collection of documents [19]. None of these methods is suitable for our context.

3. Sequencesviewer

SequencesViewer [20] helps biomedical experts to browse and explore sequences of genes identified by knowledge discovery techniques (see Fig. 2). In the following we describe the main representations selected according to Shneiderman’s information visualization mantra [12]: “overview first (see Section 3.1), zoom and filter (group of sequences in Section 3.2), details-on-demand (sequences with documents in Section 3.2)”. A fourth view based on a treemap have been added to give another point of view of the input data (see Section 3.3).

3.1. Point cloud

The Point Cloud representation allows biologists to visualize groups of gene sequences (see Figs. 3, 4). It gives an overview of the centres of the groups, the distance from the centres, and associated sequences. Three steps are required to compute the relevant positions of centres to limit the number of occlusions. We combine three algorithms and adapt them to our problem. An efficient interaction mode is also added to help users find the information they require.

3.1.1. Main placement of the centres

The basic idea is to locate the centres in such a way that the Euclidean distances between them are proportional to the distances between the sequences given by a matrix of distances *D* containing S2MP measures [10].

Let d_{ij} be the matrix value for a centre *i* and a centre *j*. We want to find the coordinates $p_i = (x_i, y_i)$ for each centre *i* so that $\|p_i - p_j\| \approx d_{ij}$ where $\|p_i - p_j\|$ is the Euclidean distance between the centres *i* and *j*.

² This work was conducted in collaboration with the MMDN lab (‘Molecular mechanisms in neurodegenerative dementias’ laboratory, University of Montpellier 2).

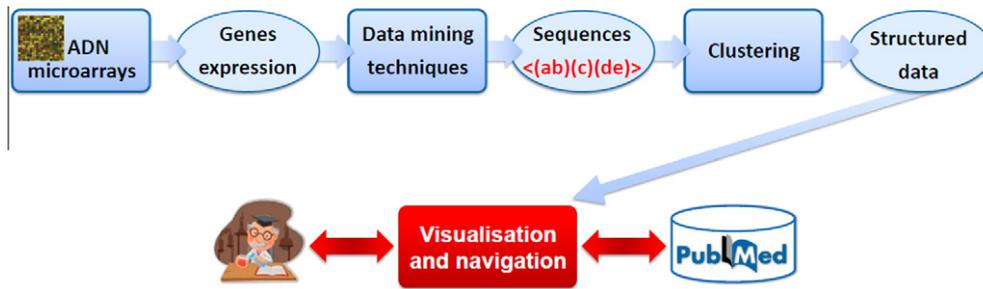


Fig. 2. SequencesViewer enables biologists to browse and explore sequences of genes and their related papers in Pubmed. These sequences are extracted and divided into groups using data mining and clustering techniques.

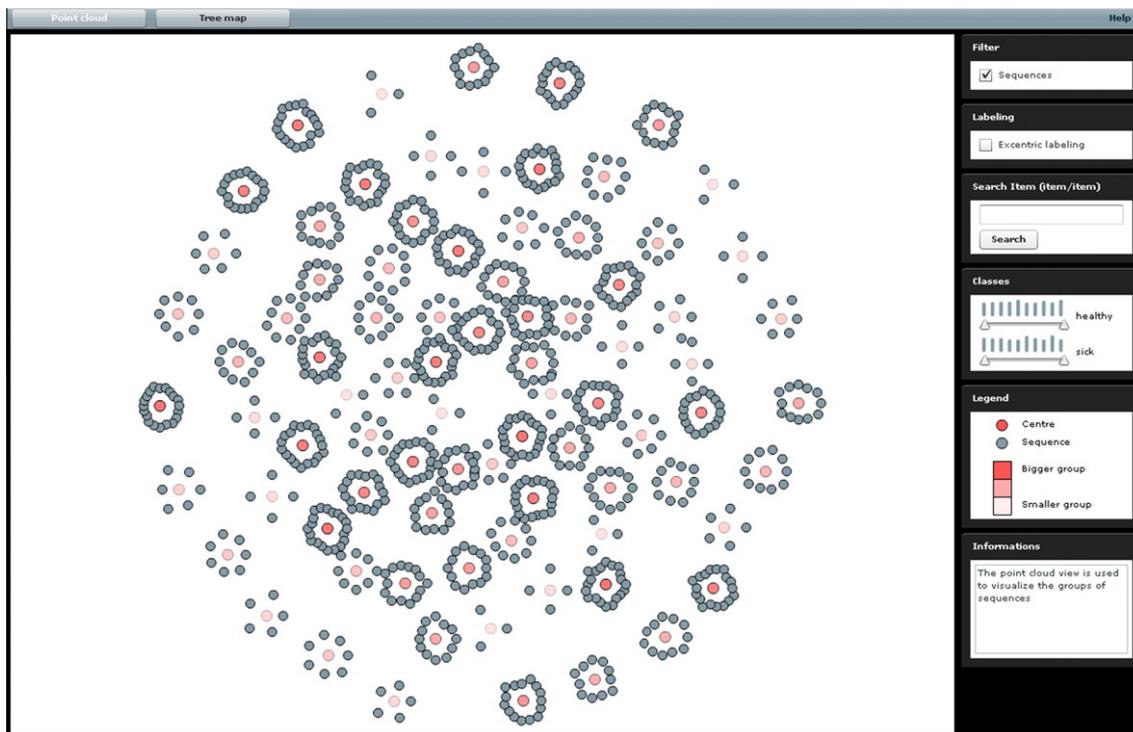


Fig. 3. Point cloud representation of sequences: the red nodes represent the centres of the groups and the grey one represents the related sequences.

Different techniques are described in the literature to assign a location to items in an N-dimensional space. Multidimensional Scaling (MDS) technique [21] is often used in information visualization and was first introduced by Togerson [22]. This technique produces representations that reveal similarities and dissimilarities in the dataset using a matrix of ideal distances. In our application, we want to find positions in a 2-dimensional space. We use a MDS optimization strategy called Stress Majorization [23], which consists of minimizing a cost function (i.e. stress function) that measures the square differences between ideal distances and Euclidean distances in 2-dimensional space:

$$\sigma(p) = \sum_{i < j \leq n} \omega_{ij} (d_{ij} - \|p_i - p_j\|)^2 \quad (1)$$

where $\omega_{ij} = d_{ij}^{-\alpha}$ and $p = p_1, p_2, \dots, p_n$ is the actual configuration. We use $\alpha = 2$, which appears to produce good results in most cases, as shown by [24].

Several techniques have been developed to minimize the stress function (see [21] for an overview). In our application, we chose a method introduced in [24] for its simplicity, fast convergence and for the quality of the results. It consists of successively computing a simple function that returns position p_i :

$$p_i^{[t+1]} \leftarrow \frac{\sum_{j:j \neq i} \omega_{ij} \left(p_j^{[t]} + s_{ij}^{[t]} \cdot (p_i^{[t]} - p_j^{[t]}) \right)}{\sum_{j:j \neq i} \omega_{ij}} \quad (2)$$

where $p_i^{[t]}$ is the position of the centre i at time t and

$$s_{ij}^{[t]} = \begin{cases} \frac{d_{ij}}{\|p_i^{[t]} - p_j^{[t]}\|} & \text{if } \|p_i^{[t]} - p_j^{[t]}\| \neq 0 \\ 0 & \text{else} \end{cases} \quad (3)$$

This iterative updating is performed for each node and repeated until a stable configuration is reached. At each step, $\sigma(p)^{[t]} \geq \sigma(p)^{[t+1]}$ and the stress function converge to a local minimum [25].

3.1.2. Initial placement of the centres

One important aspect of these methods is to find an initial placement of the centres before performing the iterative process. Random placement is not efficient because each time the algorithm is executed for the same data, the final layout changes. Moreover, the stress majorization converges slowly and it can fall into local minima. In our system, we use the fold-free embedding defined in [26]. The algorithm selects four centres c_1, c_2, c_3 and c_4 so that they are in the periphery of the point cloud. The pair (c_1, c_2) has

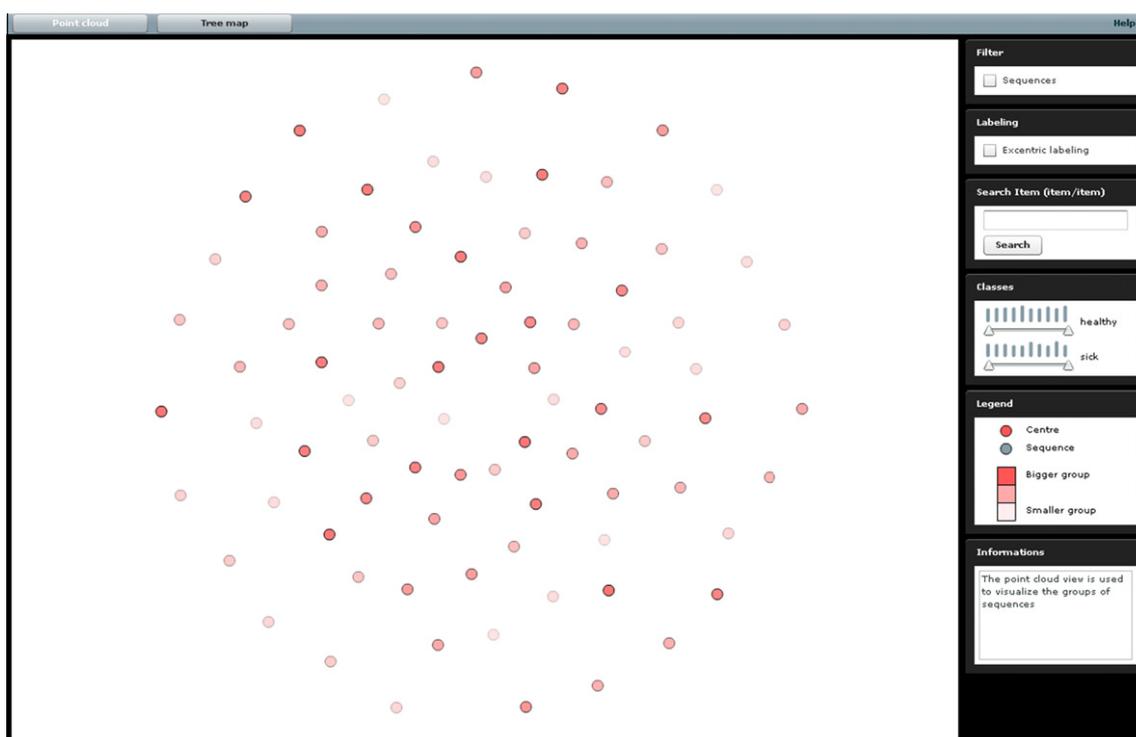


Fig. 4. Centres in the point cloud representation: checking a box enables the user to hide the sequences that are not the centre of their group.

to be roughly perpendicular to the pair (c_3, c_4) in the layout. A fifth centre c_5 is selected so that it lies in the middle of the point cloud. These centres are selected as follows:

1. Arbitrarily select a centre c_0 . c_1 is the centre so that the $d_{c_0c_1} \geq d_{c_0i}$ for each centre i with $d_{c_0c_1}$ the distance between c_0 and c_1 in the matrix D .
2. c_2 is the centre so that $d_{c_1c_2} \geq d_{c_0i}$ for each centre i . Thus, c_1 and c_2 are roughly opposite one another in the point cloud.
3. c_3 is the centre so that $|d_{c_1c_3} - d_{c_2c_3}| \leq |d_{c_1i} - d_{c_2i}|$ for each centre i . It is roughly equidistant from c_1 and c_2 .
4. As in the previous step, c_4 is one of the centres so that $|d_{c_1c_4} - d_{c_2c_4}| \leq |d_{c_1i} - d_{c_2i}|$ for each other centre i . Among this set of candidates, pick the one that maximizes $d_{c_3c_4}$. Thus, c_4 is roughly equidistant from c_1 and c_2 and roughly opposite c_3 .
5. As in the previous step, c_5 is one of the centres so that $|d_{c_1c_5} - d_{c_2c_5}| \leq |d_{c_1i} - d_{c_2i}|$ for each other centre i . Among these candidates, pick the one that minimizes $|d_{c_3c_5} - d_{c_4c_5}|$. Thus, c_5 is roughly in the middle of the graph.

x_i denotes $d_{c_3i} - d_{c_4i}$ and y_i denotes $d_{c_1i} - d_{c_2i}$. We can use (x_i, y_i) coordinates directly to locate each centre i . Unfortunately, this solution disregards the distance between i and c_5 . To overcome this problem, the method described in [26] computes the polar coordinate (ρ_i, θ_i) of a centre i so that $\rho_i = d_{c_5i} \times R$ and $\theta_i = \tan^{-1}\left(\frac{y_i}{x_i}\right)$. Actually, we compute θ_i more accurately according to trigonometry:

$$\theta_i = \begin{cases} \tan^{-1}\left(\frac{y_i}{x_i}\right) & \text{if } x_i > 0 \text{ and } y_i \geq 0 \\ \tan^{-1}\left(\frac{y_i}{x_i}\right) + 2\pi & \text{if } x_i > 0 \text{ and } y_i < 0 \\ \tan^{-1}\left(\frac{y_i}{x_i}\right) + \pi & \text{if } x_i < 0 \\ \frac{\pi}{2} & \text{if } x_i = 0 \text{ and } y_i \geq 0 \\ \frac{3\pi}{2} & \text{if } x_i = 0 \text{ and } y_i < 0 \end{cases}$$

3.1.3. Removing central overlap

The MDS method we implemented does not avoid overlapping of centres. Node occlusions can mislead the user by hiding information. We thus run a node overlap removal algorithm after the MDS placement step described above. Gansner and Hu [27] implemented a simple but effective solution based on a nice adaptation of the stress majorization process.

This solution is based on a Delaunay triangulation [28] computed for the set of centres and their current positions. A Delaunay triangulation is a triangulation that maximizes the minimum angle of all the angles of the triangles. We can represent the results of a triangulation on our centres as a planar graph $G(V, E)$ where V is the set of the centres and E is the set of the edges of triangles. The node overlap is removed iteratively:

1. First, we compute a Delaunay triangulation on the current layout. Let $G^{DT}(V, E^{DT})$ be the graph produced by the triangulation.
2. For each $\{i, j\} \in E^{DT}$ an overlap factor is computed:

$$t_{ij} = \max\left(\frac{a_i + a_j}{\|p_i - p_j\|}, 1\right) \quad (4)$$

where a_i is the radius of the centre i . $t_{ij} = 1$ if the centres i and j do not overlap. If $t_{ij} < 1$, we can remove the overlap by extending the length of the edge $\{i, j\}$ by this factor. A new ideal distance matrix is then computed: $d_{ij}^{DT} = s_{ij}^{DT} \|p_i - p_j\|$ where s_{ij}^{DT} is a factor computed from t_{ij} to damp it: $s_{ij}^{DT} = \min\{s_{max}, t_{ij}\}$, with $s_{max} > 1$ (1.5 in our implementation). s_{max} is the maximum amount of overlap we can remove at each step while keeping the same global configuration.

3. We now minimize the stress function using the process described above (see Eq. (2)) with d_{ij}^{DT} and s_{ij}^{DT} in spite of d_{ij} and s_{ij} .

$$\sigma^{DT}(p) = \sum_{i < j \leq n} \omega_{ij} \left(d_{ij}^{DT} - \|p_i - p_j\|\right)^2 \quad (5)$$

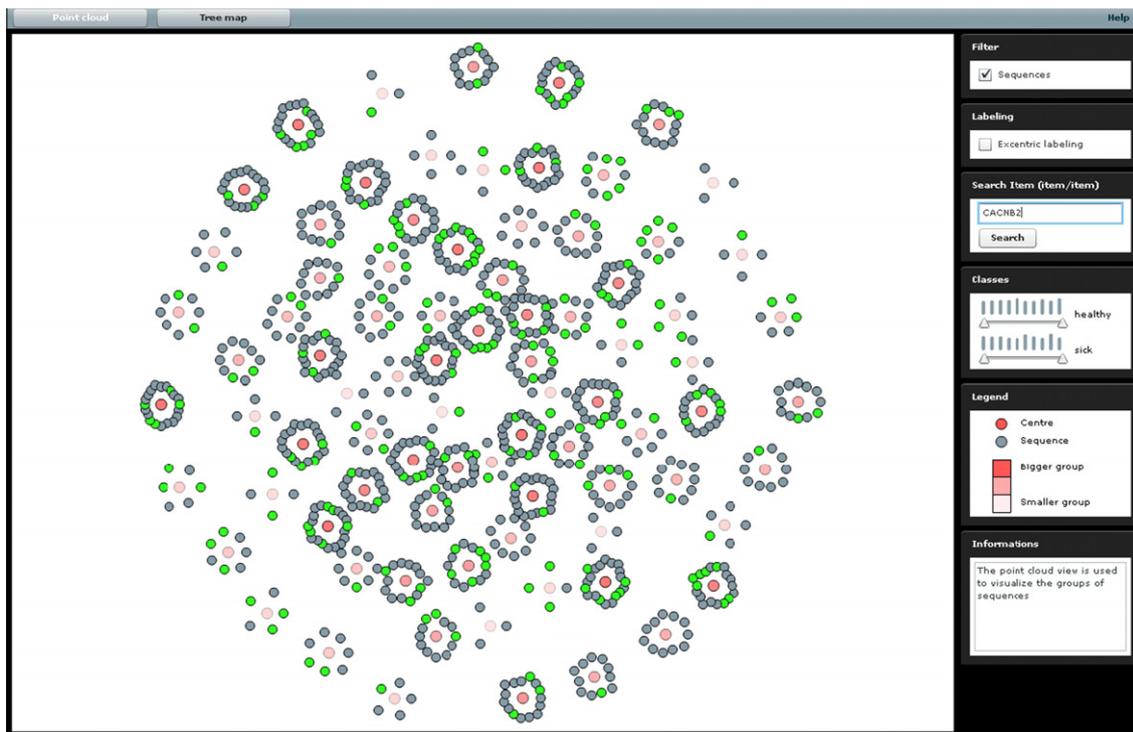


Fig. 5. Point cloud with sequences and highlighted searched items.

3.1.4. Interactions and navigation

The user can choose to visualize the centres (see Fig. 4) or the centres plus their associated sequences using the check box labelled *Sequences* (see Fig. 3). The colour of the centres is of different intensity, which is proportional to the number of sequences associated with the centre concerned. The legend on the right helps the user evaluate the size of the groups. An item can be searched and the sequences containing the searched term are highlighted. The screenshot in Fig. 5 represents a map with the highlighted sequences (in green³) resulting from a search operation.

Moreover, the user can move the whole map by dragging and dropping with a mouse. Zoom In/Out options are also available to the user by using the mouse wheel. A tooltip containing sequence informations is displayed when the user clicks on a sequence (see Fig. 6).

For each classes, a slider has been added to select a range of their corresponding support (see box *Classes* in Fig. 7). The sequences with a support out from this range are then filtered from the view. Inspired by Scented Widgets [29], a bar chart is displayed over each slider to represent the number of sequences sharing the corresponding support value.

Finally, we have implemented an excentric labelling technique [30]. When the user clicks on a free space in the map, labels of the sequences positioned inside a circle around the clicked point are displayed (see Fig. 8). To avoid overlaps, they are positioned far from their corresponding sequences: colours and lines are used to link the sequences with their own labels.

3.2. Solar system

When the user double-clicks on a sequence in the point cloud view, he/she accesses a second view (see Fig. 9) based on a solar

system metaphor [31]. This view allows only the group of the selected centre to be explored. The centre is positioned in the middle of the visualization area (position (0,0)). Then, each sequence i is placed at a coordinate (d_i, θ_i) where d_i is the S2MP measure between the sequence and its centre and $\theta_i = i \cdot \frac{2\pi}{n}$ where n is the number of sequences. Grey circles have been added to the visualization to help the user approximate the value of d_i .

Interactions techniques previously described (i.e. zoom, search, sliders, tooltip, moving the whole map, excentric labelling) are also available in this view except the removing of the sequences, i.e. it is useless to display the centre alone. The legend is also displayed.

A second view based on the solar system can be accessed from the first view by double-clicking on a sequence (see Fig. 10). This view represents the sequence and its associated set of text documents i.e. scientific papers dealing with the genes belonging to

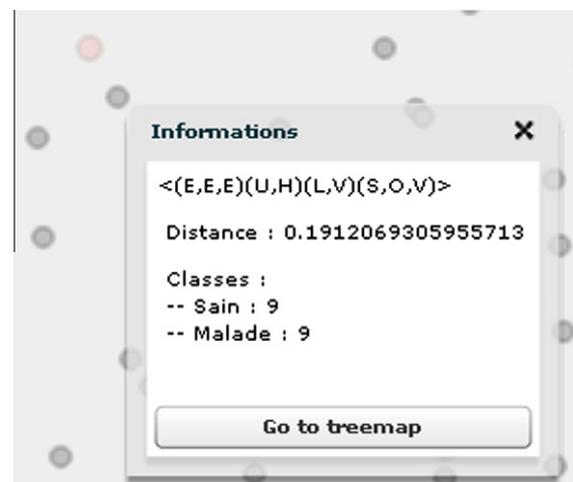


Fig. 6. Clicking on a sequence opens a tooltip containing its information.

³ For interpretation of colour in Figs. 1–19, the reader is referred to the web version of this article.

PARTIE III

ANNEXES

		DM1 Descriptive	DM2 Predictive	DM3 Prescriptive	TOTAL	
		98425	773575	226168	1098168	
AP1	Treatment effectiveness	2042	7	233	57	297
AP2	Healthcare management	2181	13	240	37	290
AP3	Customer relationships	45	1	4	1	6
AP4	Fraud and abuse	178	1	42	4	47
AP5	Genomics	90405	2035	7599	2181	11815
		94851	2057	8118	2280	12455

AP1	Treatment effectiveness	Année	Type	Thème	Méthode	Pays
	Dickerson	1	2014	Cas d'étude	Impact soin sur le fait de dormir	pourcentage usa
	Hirani	2	2014	Cas d'étude	Impact telecare on quality of life	clustering usa
	Cartwright	3	2013	Cas d'étude	Impact telecare on quality of life	clustering uk
	Spyridonos	4	2013	Cas d'étude	Diagnostic	Similarity Greece
	Co	5	2010	Cas d'étude	Impact EHR on quality of care	clustering usa
	Van Bystervel	6	2010	Cas d'étude	Impact soin sur enfant Down syndrome	pourcentage New Zealand
	Wu	7	2008	Cas d'étude	Impact du réseau sur prise en charge	clustering usa

Année	Nombre	Méthode	Nombre
>=2010	6	pourcentage	2
<2010	1	Clustering	5
Type	Nombre	Localisation	Nombre
Cas d'étude	7	Europe	3
Revue	0	USA	3
		New Zealand	1
Thème	Nombre		
Impact des si	5		
Impact envin	1		
Diagnostic	1		

AP2	Healthcare management	Année	Type	Thème	Méthode	Pays
	Khuba	1	2014	Cas d'étude	Health Information system	clustering Kenya
	Naghdi	2	2014	Cas d'étude	Impact of food on health	clustering Iran
	Allsop	3	2014	revue	Impact of ICT on pain	prisma Monde
	Hammer	4	2013	Cas d'étude	Impact of social capital on quality of care	regression ar Europe
	Sheedy	5	2014	Cas d'étude	Impact of care in case of stroke	regression new south w
	Andrew	6	2014	Cas d'étude	Impact of care in case of stroke	regression australie
	Kim	7	2014	Cas d'étude	Impact bad staffing on pain among c	regression Korea
	Jephcote	8	2013	Cas d'étude	Impact of road-transport emissions c	spatial patterneuk
	Trinh	9	2014	Cas d'étude	Impact of service duplication in case of hospital	clu USA
	Garcia-Olmc	10	2012	Cas d'étude	Detection comorbidity patterns	regression Espagne
	Stark	11	2011	Cas d'étude	Impact of care in case of type 2 diab	regression ar Allemagne
	Jung	12	/	/	/	/
	Ramirez	13	2003	Cas d'étude	Impact management of asthma	pattern recoUSA
	Bathikar	14	2002	Cas d'étude	Diagnostic medical	Artificial Neu USA

Année	Nombre	Méthode	Nombre
>=2010	11	Regression	6
<2010	2	Clustering	4
exclu	1	Prisma (revu)	1
		pattern	2
		Artificial Neu	1
Type	Nombre	Localisation	Nombre
Cas d'étude	12	Europe	5
Revue	1	USA	3
		Asie	1
		Afrique	2
		Monde	1
		Australie	1
			13

AP3	Customer relationships	Année	Type	Thème	Méthode	Pays
	Fu	1	2010	revue	age synthesis and estimation via face	monde

AP4	Fraud and abuse	Année	Type	Thème	Méthode	Pays
	Biafore	1	2003	Cas d'étude	détection fraude	pattern usa