Fouille de données : motifs minimaux, redescription d'espace et analyse du (e-)sport

François Rioult

Normandie Université

Habilitation à Diriger des Recherches

Pour obtenir le diplôme d'habilitation à diriger des recherches

Spécialité INFORMATIQUE

Préparée au sein de l'Université de Caen, Normandie

Fouille de données : motifs minimaux, redescription d'espace et analyse du (e-)sport.

Présentée et soutenue par
François RIOULT

HDR soutenue publiquement le 7 décembre 2017 devant le jury composé de		
Amedeo NAPOLI	DR CNRS, LORIA, Nancy	Rapporteur
Maguelonne TEISSEIRE	DR IRSTEA, TETIS, Montpellier	Rapporteure
Henry SOLDANO	MCF HDR, LIPN, Université de Paris-Nord	Rapporteur
Christine LARGERON	Professeure, LHC, Université Jean Monnet	Examinatrice
Christophe RIGOTTI	MCF HDR, LIRIS, INSA de Lyon	Examinateur
Bruno CRÉMILLEUX	Professeur, GREYC, Université de Caen Normandie	Examinateur
Sylvain PEYRONNET	Professeur, Ix-Labs, Rouen	Examinateur (Directeur de l'HDR)

UNICAEN
UNIVERSITÉ
CAEN
NORMANDIE

GREYC

Table des matières

1. Introduction

Ce document constitue une rapide synthèse des activités de recherches réalisées depuis ma thèse en 2005 [55], puis en tant que Maître de Conférences à l'Université de Caen Normandie depuis 2006, au laboratoire CNRS UMR6072 GREYC.

Le thème général de cette synthèse est la fouille de données symboliques ou *extraction de motifs* : selon une approche théorique – en lien avec les hypergraphes –, selon sa complexité en moyenne, et selon les usages des motifs extraits, en particulier dans le domaine émergent de l'analyse du sport et de sa déclinaison, le jeu vidéo compétitif. Nous nous attacherons à montrer que la découverte de motifs peut être considérée comme une opération permettant une *redescription des données*, dans un espace où elles sont prises en charge par des techniques éprouvées d'apprentissage automatique (arbres, neurones, *etc.*).

L'extraction de motifs, comme une étape de redescription des données, cadre bien avec la tendance actuelle de l'apprentissage automatique : un classifieur commence par quelques couches de filtrage ou *template matching* [66] à l'aide de SVM ou de couches de convolution, puis optimise la décision en se fondant sur les traits qui ont émergé. Désormais, le *soft computing* remplit une fonction de *recherche d'information symbolique* – typiquement des formes élémentaires de 5x5 pixels dans une image – sous optimisation. Dans une partie importante de ce rapport (partie II), c'est la fouille qui sera observée pour son apport en terme de redescription et d'enrichissement des données, et donc d'étape préliminaire pour des techniques d'apprentissage plus généralistes.

Illustrons sommairement ces transformations :

— au cours de la recherche d'un modèle de dialogue, nous avons besoin de prédire un événement dans une séquence de motifs. Par exemple nous pourrions utiliser des réseaux de neurones récurrents pour apprendre un modèle. Cependant, devant la faible taille de l'échantillon, nous préférons commencer par extraire des motifs séquentiels de nos données pour en décrire plus finement les aspects temporels. Outre leur intelligibilité, ces motifs procurent aux données un enrichissement qu'une méthode de décision prendra en compte, analysant la survenue des motifs dans les données plutôt que celle des seuls attributs ;

— pour remplacer (ou *imputer*) une valeur manquante par une valeur connue, on peut commencer par déterminer les explications du fait qu'elle est manquante, par exemple à l'aide de règles d'association non redondantes. Ces associations induisent un changement de représentation, dans laquelle les données sont matérialisées par un graphe d'associations. À l'aide d'un Pagerank sur ce graphe, une explication *raisonnable* émerge pour chaque valeur manquante, explication qui permet de contextualiser l'imputation.

Ce travail rapporte des recherches en collaboration étroite avec Arnaud Soulet (LI Tours), Loïck Lhote (GREYC), Gaël Lejeune et Leïla Ben Othman.

Plan du mémoire Trois parties structurent ce mémoire, depuis les fondements théoriques et algorithmiques de l'extraction de motifs (partie I) jusqu'à leur exploitation pour la redescription des données (partie II) ainsi que vers l'analyse du sport et de sa déclinaison électronique (partie III).

La première partie contient deux chapitres traitant des motifs minimaux. Le chapitre 3 expose un algorithme en profondeur pour extraire des motifs minimaux au sein d'une classe d'équivalence définie par une fonction et un langage donnés. Par exemple, pour la fonction « fréquence d'un motif », les classes d'équivalence sont définies par le support et les minimaux sont les motifs *libres*. Nous avons donc introduit la notion de système d'ensembles minimisables sur un système accessible [58] doté d'un opérateur cov aux propriétés de couverture (*i.e.* $cov(X \cup Y) = cov(X) \cap cov(Y)$), comme par exemple l'opérateur d'extension en analyse formelle des concepts.

Le chapitre 4 analyse le nombre moyen de traverses minimales dans un hypergraphe. Comme ensemble d'*hyperarêtes* (arêtes de taille quelconque), un hypergraphe est une structure bien adaptée pour représenter les collections de motifs manipulées par la fouille. De plus, le calcul des traverses minimales d'un hypergraphe, ces ensembles minimaux qui intersectent avec toute hyperarête, suscite un large intérêt en tant que problème séparateur de P et NP ; ce calcul occupe une position centrale en fouille puisqu'il cristallise la difficulté de passer d'une solution temporaire au problème à l'ensemble des candidats minimaux pour l'étape suivante. Le calcul de traverses minimales mérite donc un intérêt tout particulier.

La deuxième partie regroupe deux exemples de l'usage de la fouille pour la redescription des données. Au chapitre 5, un *modèle de l'interaction dialogique* est calculé à partir de dialogues annotés selon des modalités émotionnelles. Le cadre général est la réalisation d'un agent conversationnel animé, destiné à raconter une histoire à un enfant en l'impliquant émotionnellement. Pour cela, l'agent a besoin d'un modèle du dialogue et de connaissances sur la façon de susciter des émotions chez l'enfant, information disponible grâce à l'analyse des motifs dans les annotations. Le problème est envisagé sous l'angle de la prédiction d'un événement : l'intervention de l'enfant. Il s'agit là d'une application emblématique de la redescription de données par la fouille : une transformation de l'espace d'entrée, peu structuré conceptuellement ou temporellement, vers un espace des régularités sur lequel l'algorithme générique de décision peut s'appuyer.

Le chapitre 6 relate comment l'extraction de motifs participe à l'enrichissement des données incomplètes en expliquant les phénomènes de valeurs manquantes. Une typologie pour ces valeurs est proposée, qui peut être valorisée pour améliorer la performance de la phase d'imputation, car des valeurs manquantes ayant le même contexte d'apparition devraient naturellement être complétées de la même manière. Ce chapitre met en évidence la difficulté d'évaluer une méthode d'imputation des données, tant d'un point de vue de la génération de données de test que de l'indice de performance choisi.

La troisième et dernière partie rassemble quelques contributions méthodologiques au domaine émergent de l'analyse du sport et du sport électronique. Le chapitre 7 présente ainsi une analyse de l'abandon arbitral à l'aide de motifs émergents et une première étude de grande envergure sur

des données HawkEye, concernant les trajectoires de la balle de tennis. Ces résultats fournissent des préconisations à l'intention des joueurs et des entraîneurs, dans un but de développement personnel ou pour préparer une opposition.

Enfin, le dernier chapitre 8 rapporte des analyses de données de jeu vidéo pratiqué en compétition. Ces analyses valorisent des traces de jeu, par exemple des trajectoires. Des préconisations stratégiques peuvent alors émerger, de même que des indices sur l'*attention* [61, 132] à accorder aux traces, de façon à générer de l'engagement sur les médias associés.

2. Publications

Revues internationales avec comité de lecture

[1] J. David, L. Lhote, A. Mary, et F. Rioult. An average study of hypergraphs and their minimal transversals. *Theor. Comput. Sci.*, 596 :124–141, 2015.

[2] M. Laignelet et F. Rioult. Automatic tracking of obsolescent segments with linguistic cues. *TAL*, 51(1) :41–63, 2010.

[3] S. Mecheri, F. Rioult, B. Mantel, F. Kauffmann, et N. Benguigui. The serve impact in tennis : First large-scale study of big hawk-eye data. *Statistical Analysis and Data Mining : The ASA Data Science Journal*, pages 1–16, 2016.

[4] N. Scelles, C. Durand, S. T. Bah, et F. Rioult. Intra-match competitive intensity in french football ligue 1 and rugby top 14. *International Journal of Sport Management and Marketing*, 2011.

Conférences internationales avec comité de lecture

[5] L. Ben Othman, F. Rioult, S. Ben Yahia, et B. Crémilleux. Completing non-random missing values. Dans *4th International Conference on Intelligent Systems and Knowledge Engineering (ISKE'09)*, pages 227–232, Hasselt, Belgium, November 2009. World Scientific.

[6] L. Ben Othman, F. Rioult, S. Ben Yahia, et B. Crémilleux. Missing values : Proposition of a typology and characterization with an association rule-based model. Dans *11th International Conference on Data Warehousing and Knowledge Discovery (DaWak'09)*, volume 5691 of *Lecture Notes in Computer Science*, pages 441–452, Linz, Austria, September 2009. Springer.

[7] G. Lejeune, F. Rioult, et B. Crémilleux. Highlighting psychological features for predicting child interjections during story telling. Dans *Interspeech'16*, pages 2056–2059, 2016.

[8] L. Lhote, F. Rioult, et A. Soulet. Average number of frequent (closed) patterns in bernouilli and markovian databases. Dans *IEEE International Conference on Data Mining (ICDM'05)*, pages 713–716, Houston, USA, 2005.

[9] A. Pauchet, F. Rioult, E. Chanoni, Z. Ales, et O. Serban. Interactive narration requires interaction and emotion. Dans *5th International Conference on Agents and Artificial Intelligence*, pages 1–4, 2013.

[10] F. Rioult et B. Crémilleux. Condensed representations in presence of missing values. Dans *Symposium on Intelligent Data Analysis*, pages 578–588, Berlin, Germany, 2003.

[11] F. Rioult, S. Ferrandiz, M. Bastien, et M. Boullé. Information enhancement in a voluminous forum with automatic co-clustering. Dans *2nd International Symposium on Web AlGorithms, Deauville, France*, 2016.

[12] F. Rioult, J.-P. Métivier, B. Helleu, N. Scelles, et C. Durand. Mining tracks of competitive video games. Dans *{AASRI} Conference on Sports Engineering and Computer Science (SECS 2014)*, volume 8, pages 82–87, 2014.

[13] A. Soulet, B. Crémilleux, et F. Rioult. Condensed representation of emerging patterns. Dans *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04)*, pages 127–132, Sydney, Australia, 2004.

[14] A. Soulet et F. Rioult. Efficiently depth-first minimal pattern mining. Dans *Advances in Knowledge Discovery and Data Mining - 18th Pacific-Asia Conference, PAKDD 2014*, volume 8443 of *Lecture Notes in Computer Science*, pages 28–39. Springer, 2014.

[15] B. Ziani, F. Rioult, et Y. Ouinten. A constraint-based mining approach for multi-attribute index selection. Dans *(ICEIS 2012) - 14th International Conference on Enterprise Information Systems*, pages 93–98, 2012.

Ateliers internationaux avec comité de lecture

[16] B. Crémilleux, A. Soulet, et F. Rioult. Mining the strongest emerging patterns characterizing patients affected by diseases due to atherosclerosis. Dans *International Discovery Challenge Workshop co-located with ECML-PKDD 2003*, pages 59–70, Cavtat-Dubrovnik, Croatia, 2003.

[17] F. Dosseville, F. Rioult, et S. Laborde. Why do sports officials dropout ? Dans *Machine Learning and Data Mining for Sports Analytics, ECML/PKDD 2013 workshop*, pages 1–12, 2013.

[18] B. Jeudy et F. Rioult. *Post-proceedings of the International Workshop on Knowledge Discovery in Inductive Databases (KDID'04) co-located with the ECML-PKDD'04*, chapter Database Transposition for Constrained (Closed) Pattern Mining, pages 89–107. Springer, 2005.

[19] B. Jeudy et F. Rioult. Database transposition for constrained (closed) pattern mining. *CoRR*, abs/0902.1259, 2009.

[20] A. Pauchet, F. Rioult, E. Chanoni, Z. Ales, et O. Serban. Modélisation de dialogues narratifs pour la conception d'un aca narrateur. Dans *WACAI'12 workshop on Affects, Compagnons Artificiels et Interaction*, pages 1–8, 2012.

[21] F. Rioult. Mining strong emerging patterns in wide sage data. Dans *ECML/PKDD'04 Discovery Challenge*, pages 127–138, Pisa, Italy, 2004.

[22] F. Rioult, J.-F. Boulicaut, B. Crémilleux, et J. Besson. Using transposition for pattern discovery from microarray data. Dans *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'03)*, pages 73–79, 2003.

[23] F. Rioult et B. Crémilleux. *Post-proceedings of the International Workshop on Knowledge Discovery in Inductive Databases (KDID'06) co-located with the ECML-PKDD'06*,

chapter Mining Correct Properties in Incomplete Databases, pages 208–222. LNCS 4747. Springer, 2007.

[24] F. Rioult, S. Mecheri, B. Mantel, F. Kauffmann, et N. Benguigui. What can hawk-eye data reveal about serve performance in tennis ? Dans *Machine Learning and Data Mining for Sports Analytics, ECML/PKDD 2013 workshop*, pages 1–12, 2015.

[25] F. Rioult, C. Robardet, S. Blachon, B. Crémilleux, O. Gandrillon, et J.-F. Boulicaut. Mining concepts from large sage gene expression matrices. Dans *International Workshop on Knowledge Discovery in Inductive Databases KDID'03 co-located with ECML-PKDD'03*, Cavtat-Dubrovnik (Croatia), 2003.

[26] A. Soulet, B. Crémilleux, et F. Rioult. Mining and using an efficient condensed representation of emerging patterns. Dans *proceedings of the third International Workshop on Knowledge Discovery in Inductive Databases (KDID'04) co-located with the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases PKDD'04*, pages 97–108, Pisa, Italy, September 2004.

[27] T. Varin, F. Rioult, R. Bureau, et S. Rault. Determination of 2d pharmacophore with a new algorithm for determining emerging patterns. application to 5-ht7 ligands. Dans *Strasbourg Summer School on Chemoinformatics : CheminfoS3*, june 2008.

Chapitres de livres

[28] J. Besson, F. Rioult, B. Crémilleux, S. Rome, et J.-F. Boulicaut. *Informatique pour l'analyse du transcriptome*, chapter Solutions pour le calcul d'ensembles fréquents dans des données biopuces, pages 231–254. Hermès, 2004.

[29] F. Rioult, B. Zanuttini, et B. Crémilleux. *Advances in Intelligent Information Systems*, volume 265 of *Studies in Computational Intelligence*, chapter Nonredundant generalized rules and their impact in classification, pages 3–25. Springer, 2010.

[30] N. Scelles, F. Dosseville, C. Durand, et F. Rioult. *Comment concilier intensité compétitive et respect des arbitres ?*, pages 255–272. Publibook, 12/2011 2011.

[31] N. Scelles, F. Dosseville, C. Durand, et F. Rioult. *Les facettes de l'arbitrage : Recherches et problématiques actuelles*, chapter Comment concilier intensité compétitive et respect des arbitres ?, pages 262–278. Editions Publibook, Collection Université : Sport & Santé, 2011.

[32] A. Soulet, B. Crémilleux, et F. Rioult. *Condensed Representation of EPs and Patterns Quantified by Frequency-Based Measures*, pages 173–190. Springer, 2005.

[33] A. Soulet et F. Rioult. *Advances in Knowledge Discovery and Management : Volume 6*, chapter Exact and Approximate Minimal Pattern Mining, pages 61–81. Springer International Publishing, Cham, 2017.

Revues françaises avec comité de lecture

[34] L. Ben Othman, F. Rioult, S. Ben Yahia, et B. Crémilleux. Base de caractérisation des valeurs manquantes. *Technique et Science Informatiques*, 30(10) :1247–1270, 2011.

[35] S. Ferrari, C. Charnois, Y. Mathet, F. Rioult, et D. Legallois. Analyse de discours évaluatif, modéle linguistique et applications. *Revue des Nouvelles Technologies de l'Information*, E-17 :71–93, 2009. Numéro spécial Fouille de données d'opinions.

[36] F. Rioult. Découverte de motifs fréquents dans les bases de données, un cadre formel pour les méthodes. *Revue des sciences et technologies de l'information série Ingéniérie des systèmes d'information (RSTI-ISI)*, 9 :211–240, 2004.

[37] F. Rioult, J.-P. Métivier, B. Helleu, N. Scelles, et C. Durand. Fouille de traces de jeu vidéo en compétition. une approche stratégique et sportive. *Ingénierie des Systèmes d'Information*, 17(2) :99–120, 2012.

[38] O. Serban, A. Bersoult, Z. Ales, E. Lebertois, E. Chanoni, F. Rioult, et A. Pauchet. Modélisation de dialogues pour personnage virtuel narrateur. *Revue d'Intelligence Artificielle*, 28(1) :101–130, 2014.

[39] A. Soulet, B. Crémilleux, et F. Rioult. Représentation condensée de motifs émergents. Dans *4èmes journées d'Extraction et de Gestion des Connaissances (EGC'04)*, Revue des Nouvelles Technologies de l'Information, pages 265–276, Clermont-Ferrand, France, 2004. Cepaduès Editions.

Conférences françaises avec comité de lecture

[40] T. Charnois, A. Doucet, Y. Mathet, et F. Rioult. Trois approches du greyc pour la classification de textes. Dans *Défi fouille de texte 2008*, pages 171–180, 2008.

[41] F. Houben et F. Rioult. Généralisation d'étiquetage morpho-syntaxique par classification supervisée. Dans *Atelier Langues peu dotées, TALN-RECITAL'05*, pages 239–248, Dourdan, France, 2005.

[42] F. Houben et F. Rioult. Généralisation d'étiquetage morpho-syntaxique par classification supervisée. Dans *Journées internationales d'Analyse statistique des Données Textuelles (JADT 2006)*, 2006.

[43] B. Jeudy et F. Rioult. Extraction de concepts sous contraintes dans des données d'expression de gènes. Dans *Conférence d'Apprentissage (CAp'05)*, pages 265–280, Nice, France, 2005.

[44] M. Laignelet et F. Rioult. Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. Dans *Actes de la 16ème conférence Traitement Automatique des Langues Naturelles (TALN'09)*, pages 10 pages, actes électroniques, Senlis, France, juin 2009.

[45] L. Lhote, F. Rioult, et A. Soulet. Average number of frequent and closed patterns in random databases. Dans *Conférence d'Apprentissage (CAp'05)*, pages 345–360, Nice, France, 2005.

[46] F. Rioult. Fouille de traces pour la recommandation stratégique en sport électronique. Dans *Interactions, Contextes, Traces (ICT'09)*, Caen, France, mars 2009.

[47] F. Rioult. Interprétation graphique de la courbe ROC. Dans *Extraction et Gestion des Connaissances (EGC'11)*, pages 301–304, 2011.

[48] F. Rioult et B. Crémilleux. Optimisation d'extraction de motifs : une nouvelle méthode fondée sur la transposition de données. Dans *Conférence d'Apprentissage, (CAp'03)*, pages 299–313, Laval, France, 2003.

[49] F. Rioult et B. Crémilleux. Représentation condensée en présence de valeurs manquantes. Dans *XXIIè congrès Inforsid*, pages 301–317, Biarritz, France, 2004.

[50] F. Rioult et B. Crémilleux. Extraction de propriétés correctes dans des bases de données incomplètes. Dans *Conférence sur l'Apprentissage Automatique (CAp'06)*, pages 347–362, Tregastel, France, 2006.

[51] F. Rioult, B. Zanuttini, et C. Crémilleux. Apport de la négation pour la classification supervisée à l'aide d'associations. Dans F. d'Alché Buc, editeur, *Actes de la 10e Conférence d'Apprentissage (CAp 2008)*, pages 183–196. Cépaduès éditions, 2008.

[52] A. Soulet et F. Rioult. Extraire les motifs minimaux efficacement et en profondeur. Dans *14èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2014*, volume E-26, pages 383–394. Hermann-Éditions, 2014.

[53] M. Vernier, Y. Mathet, F. Rioult, T. Charnois, S. Ferrari, et D. Legallois. Classification de textes d'opinions : une approche mixte n-grammes et sémantique. Dans *3ème Défi Fouille de Texte 2007*, 2007.

[54] A. Widlöcher, F. Bilhaut, N. Hernandez, F. Rioult, T. Charnois, S. Ferrari, et P. Enjalbert. Une approche hybride de la segmentation thématique : collaboration du traitement automatique des langues et de la fouille de texte. Dans *2ème Défi Fouille de Texte 2006*, Friboug, Switzerland, 2006.

Autres publications

[55] F. Rioult. *Extraction de connaissances dans les bases de données comportant des valeurs manquantes ou un grand nombre d'attributs*. Thèse, Université de Caen Basse-Normandie, France, 2005.

Première partie

Motifs minimaux et traverses minimales d'hypergraphe

Notations

Cette partie est dédiée à la recherche de motifs minimaux d'*attributs*. Ces attributs sont utilisés pour la description des *objets* d'expérience. Ce cadre est très général et permet aussi bien de représenter des motifs (ou *itemsets*) dans des contextes booléens que des chaînes de caractères, ou toute structure que l'on peut représenter par des ensembles. Ce chapitre traite également d'hypergraphes, un formalisme utilisé pour représenter les contextes booléens ou les ensembles de motifs.

Nous utilisons les notations suivantes :

$\mathcal{A} = \{a_1, \ldots, a_m\}$	l'ensemble des *attributs*
$\mathcal{O} = \{o_1, \ldots, o_n\}$	l'ensemble des *objets*
$r = (\mathcal{A}, \mathcal{O}, R)$	un contexte relationnel mettant en relation objets et attributs à l'aide de R. On notera les objets comme des ensembles d'attributs et *vice-versa* : $o \equiv \{a \in \mathcal{A} \mid a\, R\, o\}$ et $a \equiv \{o \mid a\, R\, o\}$.
$X = \{x_1, \ldots\} \subset \mathcal{A}$	un *motif*.
$\mathcal{L} = 2^{\mathcal{A}}$	le *langage* des motifs
$supp(X) = supp(\wedge X)$	le *support* (dit conjonctif) d'un motif : $\{o \in \mathcal{O} \mid X \subseteq o\}$
$supp(\vee X)$	le support *disjonctif* : $\{o \in \mathcal{O} \mid X \cap o \neq \emptyset\}$
$H = (V, E)$	un hypergraphe de sommets $V = \{v_1, \ldots, v_m\}$ et d'hyperarêtes $E = \{E_1, \ldots, E_n\}$ avec $E_i \subseteq V$.

3. Extraction de minimaux en profondeur

Ce travail est le résultat d'une collaboration avec Arnaud Soulet *(LI Tours) [52, 14, 33].*

Notre première présentation concerne l'extraction en profondeur d'ensembles minimaux. On s'intéresse ici aux fonctions f sur les motifs, plus précisément aux fonctions *condensables*, c'est-à-dire qu'elles sont constantes sur des classes d'équivalence. Typiquement, en fouille de motifs, on s'intéresse au *support* d'un motif : $supp(X) = \{o \mid X \subseteq o\}$. On peut également s'intéresser à la *fréquence* du motif : $freq(X) = |supp(X)|$.

Les classes d'équivalence permettent de condenser le nombre de motifs à extraire. Par exemple, chaque classe possède un unique motif maximal, dit *fermé* : les motifs fermés fréquents fournissent donc une représentation condensée des motifs fréquents.

Si une classe n'a qu'un seul maximal, elle possède potentiellement plusieurs motifs minimaux. Ces motifs représentent une variation stricte dans la fonction, symbolisant le passage d'une classe à l'autre, du fermé du dessus vers le minimal du dessous [1] : $X \in \mathcal{L}$ est minimal si

$$\forall Y \subsetneq X, f(Y) \neq f(X).$$

Les motifs minimaux ont été relativement peu étudiés par rapport aux fermés : seulement 13% des publications sur les représentations condensées sont consacrés aux minimaux, contre 59% aux fermés [91]. Pourtant, leur connaissance est essentielle pour disséquer la sémantique implicative du fermé d'une classe, par séparation en une prémisse non redondante et une conclusion maximale, dans le but de former des règles non redondantes [137, 80, 29] ou maximisant les mesures d'intérêt [102]. Les règles séquentielles bénéficient également de la minimalité [114]. Enfin, les minimaux sont largement impliqués dans l'extraction des traverses minimales, auxquelles sont consacrées la prochaine section, elles-mêmes au cœur de la complexité algorithmique de la fouille de motifs.

Les minimaux ont été peu étudiés car leur extraction efficace est délicate. En effet, la vérification *apriori* de la minimalité d'un motif requiert celle de ses sous-ensembles. Par exemple, abc ne peut être minimal que si ab, ac et bc le sont, une information indisponible lors d'une exploration en profondeur a puis ab. Les approches par niveau sont donc naturelles mais font face à une explosion combinatoire en mémoire qu'on souhaiterait éviter grâce à un parcours en profondeur.

1. considérant la spécialisation (resp. la généralisation [117]) comme une *descente* (resp. montée) dans le treillis des motifs.

3.1. Ensemble d'objets critiques

Nous avons proposé [52] un algorithme à délai et espace polynomiaux, qui extrait en profondeur les motifs minimaux. Il est suffisamment générique pour adresser toute fonction cov : $2^{\mathcal{A}} \rightarrow 2^{\mathcal{O}}$ ayant la propriété de couverture :

$$cov(X \cup Y) = cov(X) \cap cov(Y).$$

Par exemple, pour les motifs, on peut choisir pour couverture l'*extension* (ou support) : $cov(X) = \{o \in \mathcal{O} \mid X \subseteq o\}$. Pour les chaînes, $S = abracadabra$ étant encodée par $\{(a,0),(b,1),(r,2)\ldots\}$, un opérateur de couverture est la liste des index : $cov(\{(a,0),(b,1)\}) = \{0,7\}$.

La notion clé est celle d'*objets critiques*, analogue à la notion d'arêtes critiques utilisée par [118], restreinte au calcul des traverses minimales. Pour l'élément e du motif X, l'ensemble des objets critiques $\widehat{cov}(X,a)$ représente la contribution de a à la couverture de X :

$$\forall a \in X, \widehat{cov}(X,a) = cov(X \setminus a) \setminus cov(a).$$

Par exemple, $\widehat{cov}(ab,a) = \{0,7\} \setminus \{0,3,5,7,10\} = \emptyset$: l'ajout de a à b n'a pas d'impact sur la couverture (et on verra que ab n'est donc pas minimal) ; en revanche $\widehat{cov}(ab,b) = \{0,3,5,7,10\} \setminus \{0,7\} = \{3,5,10\}$: à cause de b, ab ne couvre pas $\{3,5,10\}$.

La propriété de caractérisation suivante effectue le lien avec les minimaux :

Propriété 1 *X est minimal si*

$$\forall a \in X, \widehat{cov}(X,a) \neq \emptyset.$$

Preuve : voir [33].

Pour obtenir les minimaux, on élague les motifs qui contiennent un item non contributif pour le calcul de couverture.

En outre, si l'on souhaite explorer en profondeur, en ajoutant a' à X, il est intéressant de remarquer que les objets critiques de Xa' se calculent efficacement à partir de ceux de X, par intersection avec la couverture de a' :

Propriété 2

$$\forall (a,a') \in X^2, a \neq a' \implies \widehat{cov}(Xa',a) = \widehat{cov}(X,a) \cap cov(a').$$

Preuve : voir [33].

3.2. Algorithme en profondeur pour les minimaux

La figure 3.1 représente les opérations effectuées pour produire les versions incrémentales des ensembles critiques (colonne du milieu), par intersection avec la couverture de l'item ajouté au motif (colonne de gauche). La couverture de X est également calculée incrémentalement par intersection avec celle de l'item ; elle est nécessaire au calcul de $\widehat{cov}(Xa',a') = cov(X) \setminus cov(a')$.

L'algorithme DEFME [52, 14] (Algorithme 3.2) extrait les minimaux dit *acceptables* ([58] emploie le terme de *feasible*), par exemple fréquents. DEFME consiste à descendre le long d'une branche pour en calculer les ensembles critiques, tant qu'aucun n'est vide.

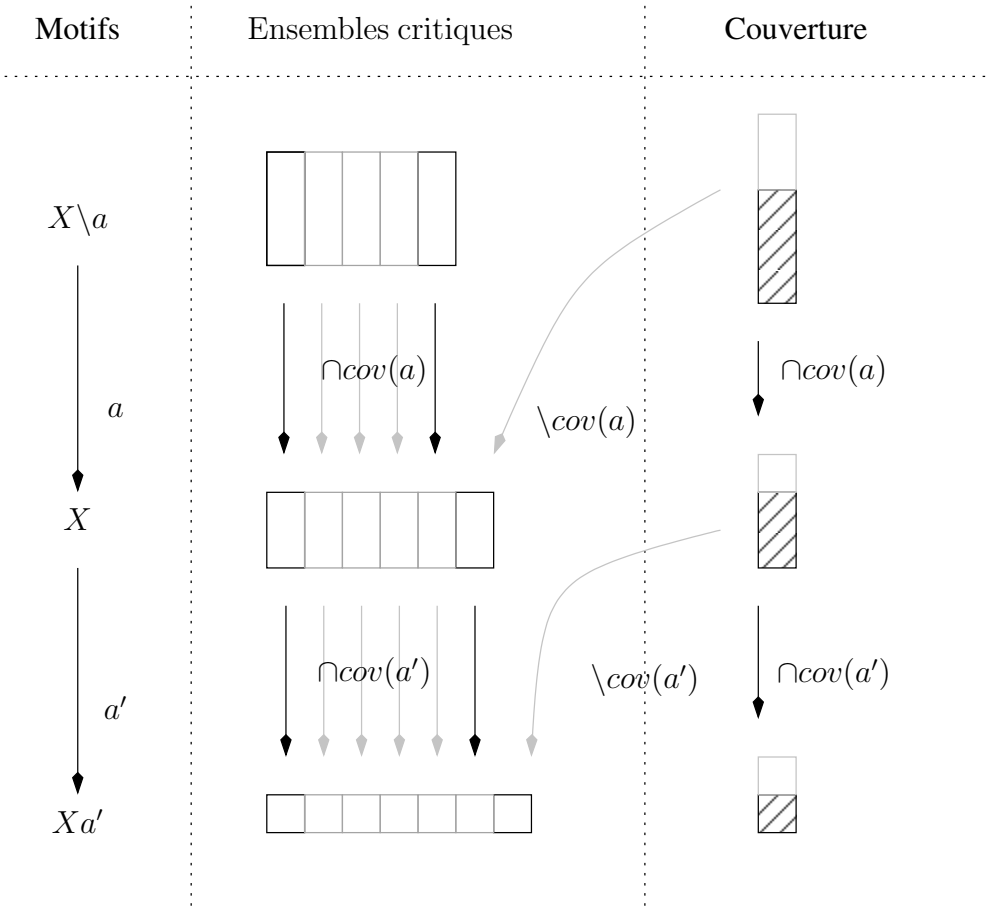

FIGURE 3.1. – Calcul en profondeur des ensembles critiques.

DEFME$(X, tail)$

Require: $tail$ est l'ensemble des items restant à parcourir.

Valeurs initiales : $X = \emptyset, tail = \mathcal{A}$.

1: **if** $\forall a \in X,\ \widehat{cov}(X, a) \neq \emptyset$ **then**
2: print X
3: **for all** $a \in tail$ **do**
4: **if** $feasible(Xa)$ **then**
5: $tail := tail \setminus \{a\}$
6: $\widehat{cov}(Xa, a) := cov(X) \setminus cov(a)$
7: **for all** $a' \in X (a' \neq a)$ **do**
8: $\widehat{cov}(Xa, a') := \widehat{cov}(X, a') \cap cov(a)$
9: **end for**
10: DEFME$(Xa, tail)$
11: **end if**
12: **end for**
13: **end if**

FIGURE 3.2. – Algorithme DEFME d'extraction en profondeur de minimaux acceptables.

3.3. Complexité

La figure 3.1 illustre trois étapes accomplies par l'algorithme au fil de la génération en profondeur :

— les trois motifs générés figurent dans la colonne de gauche, depuis un motif $X \setminus a$ jusqu'à Xa' ;
— on trouve les ensembles critiques dans la colonne du milieu, ils sont obtenus par intersection avec l'item ajouté d'une étape à l'autre, a puis a'. De plus, un nouvel ensemble critique est ajouté, par différence entre la couverture de l'item et celle du motif ;
— la couverture du motif est indiquée à la dernière colonne.

Dans le pire cas, on constate que les minimaux sont énumérables :

1. en espace $O(|\mathcal{O}| \cdot |\mathcal{A}|)$: c'est l'espace maximal pour stocher les ensembles critiques ;
2. en temps $O(|\mathcal{O}| \cdot |\mathcal{A}|^2)$ par motif minimal : les opérations de calcul des ensembles critiques (en $O(|\mathcal{A}|)$) sont effectuées $O(|\mathcal{A}|)$ fois par la boucle de la ligne 7 de l'algorithme.

L'espace nécessaire est donc *linéaire* en la « surface » de la base et l'algorithme énumère les minimaux avec un *délai polynomial*.

3.4. Expérimentations

Pour mettre à l'épreuve notre algorithme DEFME, nous l'avons comparé avec cinq algorithmes :

— deux d'entre eux effectuent des parcours en largeur : ACMINER [69] et FTMINER [100]. Ce dernier a la particularité d'être adapté aux contextes avec un grand nombre d'attributs

devant le nombre d'objets : FTMINER applique des critères d'élagage sur les supports des motifs ;

— les trois autres effectuent un parcours en profondeur en réordonnant les items : GRGROWTH [113], NDI [75] et l'algorithme TALKYG2 de la plateforme CORON [126].

Nous nous intéressons aux performances de chacun de ces prototypes lors de l'extraction de motifs minimaux pour la fonction de support, motifs dits *libres*. Les benchmarks proviennent des challenges FIMI [2] et Dicovery Challenge PKDD 2004 [3]. Les caractéristiques des benchmarks sont indiquées dans les trois premières colonnes de la table 3.1. Le seuil de fréquence est ensuite indiqué puis les temps d'exécution. La table 3.2 indique la consommation mémoire.

				temps d'extraction (s)					
benchmark	objets	items	minsup	ACMINER	FTMINER	GRGROWTH	NDI	TALKYG2	DEFME
74x822	74	822	88%	échec	158	122	échec	5 920	**45**
90x27679	90	27 679	91%	échec	270	206	échec	échec	**79**
chess	3 196	75	22%	6 623	345	**56**	187	1 291	192
connect	67 557	129	7%	34 943	échec	**37**	115	1 104	4 873
pumsb	49 046	2 113	51%	70 014	échec	**64**	212	6 897	548
pumsb*	49 046	2 088	5%	21 267	2971	**89**	202	échec	4 600

TABLE 3.1. – Caractéristiques des benchmarks, seuil de fréquence et temps d'extraction des motifs libres.

		mémoire (Mo)					
benchmark	minsup	ACMINER	FTMINER	GRGROWTH	NDI	TALKYG2	DEFME
74x822	88%	échec	12 467	1 990	échec	20 096	**3**
90x27679	91%	échec	6 763	2 929	échec	échec	**13**
chess	22%	3 914	7 990	914	1 684	12 243	**8**
connect	7%	2 087	échec	684	1 181	12 305	**174**
pumsb	51%	7 236	échec	916	1 818	30 941	**118**
pumsb*	5%	5 175	51 702	1 330	2 523	échec	**170**

TABLE 3.2. – Consommation de mémoire.

Les meilleures performances sont en gras dans la table 3.1. Les approches en profondeur GRGROWTH, NDI et DEFME dominent clairement les approches en largeur de ACMINER et FTMINER, grandes consommatrices de temps et de mémoire. GRGROWTH est le plus rapide, excepté sur les bases génomiques, qui comportent un grand nombre d'attributs. Dans ce cas, c'est notre approche DEFME qui est la plus efficace.

Comme on peut le lire à la table 3.2, DEFME est l'algorithme qui consomme le moins de mémoire. GRGROWTH et NDI ne sont pas adaptés aux données larges, même avec 200 Go de RAM et des seuils plutôt élevés. DEFME travaille avec une mémoire bornée, ce qui n'induit pas de limitation sur le seuil d'exploration.

2. http://fimi.ua.ac.be/data/
3. http://lisp.vse.cz/challenge/ecmlpkdd2004/

3.5. Extension à la minimalité approchée

Le cadre que nous proposons se prête bien à l'extraction de minimaux à δ près : étant donnée une fonction de calcul d'erreur ϵ, nous dirons que X est δ-minimal [33] si

$$\forall a \in X, \epsilon(\widehat{cov}(X, a)) > \delta.$$

Pour les motifs, si ϵ est la fonction cardinal et si l'opérateur de couverture la fonction de support, on retrouve les motifs δ-libres de [70]. Mais on peut également utiliser ϵ pour introduire une pondération sur les objets considérés.

Cette notion de minimalité approchée est précieuse, car elle ouvre la voie au calcul de motif émergents *minimaux*. Les motifs émergents [83] ont un rapport de support favorable à une classe plutôt qu'au reste et sont de bons motifs pour construire des classifieurs. La section suivante montre comment calculer ce rapport de support dans notre cadre formel de *système minimisable d'ensembles*.

Notons également le lien fort entre motifs émergents et traverses minimales, initialement pointé par [62] : les motifs émergents bondissant minimaux – *jumping emerging patterns*, de support nul dans la classe négative – sont les traverses minimales des complémentaires des objets de la classe négative, fréquentes dans les objets de la classe (voir une utilisation au chapitre 6 et la figure 7.1). Des *traverses minimales approchées* [29, 85, 104] sont donc pertinentes pour obtenir, entre autres, des émergents minimaux à δ-près qui ont le bon goût de maximiser les mesures d'intérêt [101, 102]. Notre cadre prend en charge cette minimalité approchée, l'algorithme 3.2 nécessitant simplement un relèvement de la borne inférieure de la ligne 5 (algorithme 3.2).

3.6. Système minimisable d'ensembles

Pour terminer cette présentation, précisons que notre algorithme s'applique dans tout *système minimisable d'ensembles* [33]. C'est généralement le cas, par exemple lorsqu'il dérive d'un *systèmes fortement accessible* [4] [68] qui a permis de formaliser la recherche d'une vaste gamme de motifs maximaux [58] : enveloppes convexes de points, bi-cliques (bi-clusters) maximales, sous-graphes clos dans des graphes de relation, motifs clos rigides avec jokers, séquences rigides dans des flux, patchs images clos par translation et rotation, *etc.* Pour peu que les objets à rechercher puissent être décrits par des ensembles et qu'on puisse définir un opérateur de couverture entre les attributs et les objets, notre algorithme de recherche de minimaux s'applique.

Donnons quelques exemples d'adaptation de l'opérateur de couverture :

— pour les motifs et l'opérateur d'extension (ou de support) comme couverture, les minimaux sont les motifs clé [64], libres [74] ou générateurs [98].

— en utilisant le complémentaire du support disjonctif, dit *couverture négative* :

$$\overline{cov_{\mathcal{A}}} : X \mapsto \{o \in \mathcal{O} \mid X \cap o = \emptyset\},$$

les motifs minimaux de ce système sont les *essentiels* [76] et les essentiels de support $\mathcal{O}$ (ou fréquence 100%) sont les traverses minimales ;

4. Dans l'espace de recherche, entre deux motifs *acceptables* ordonnés, il existe un chemin de motifs acceptables.

— pour le calcul de motifs émergents entre deux classes : si $\mathcal{O} = \mathcal{O}_1 \cup \mathcal{O}_2$, la confiance de la règle de classification $X \rightarrow c_1$ est $|\mathcal{O}_1 \cap cov(X)|/|cov(X)|$. En fait, toute mesure fondée sur une fréquence (lift, bond, *etc.*) peut être dérivée des couvertures positive et négative.
— en conjonction avec des fonctions d'agrégat comme min, max ou sum [124].

3.7. Conclusion

Nous avons présenté un cadre formel prenant en compte de nombreux langages de description – tant qu'il s'agit d'objets peuplés par des attributs et qu'un opérateur de couverture peut y être défini – pour concevoir un algorithme qui calcule efficacement les motifs minimaux, en espace et délai polynomial. En lien avec le chapitre suivant, les perspectives concernent l'analyse en moyenne de notre algorithme, en particulier lorsqu'il est utilisé pour extraire des traverses minimales.

4. Analyse en moyenne de l'extraction de traverses minimales

Ce travail [1] a été réalisé avec Loïck Lhote (GREYC), il est le prolongement de notre collaboration rapportée au chapitre 4 de [55] et publiée dans [8, 45].

Dans ce chapitre, nous analysons le nombre moyen de traverses minimales dans un hypergraphe. En tant qu'ensemble d'*hyperarêtes* (arêtes de taille quelconque), un hypergraphe est une structure bien adaptée pour représenter les collections de motifs manipulées par la fouille. De plus, le calcul des traverses minimales d'un hypergraphe (ou *Transversal Hypergraph Generation,* THG-*problem*), ces ensembles minimaux qui intersectent avec toute hyperarête, suscite un grand intérêt en tant que problème séparateur de P et NP ; les traverses minimales occupent une position centrale en fouille puisque leur calcul cristallise la difficulté à passer d'une solution temporaire au problème à l'ensemble des candidats minimaux pour l'étape suivante [115, 36].

L'étude de la complexité de ce problème mérite donc un intérêt tout particulier, d'autant qu'on le retrouve dans de nombreux domaines [97], au premier rang desquels figure la dualisation de formules booléennes monotones [90]. Une analyse dans le pire cas sur des hypergraphes variés [95] montre que les algorithmes phares [62, 84, 105] ne sont pas output–polynomial [1].

L'*analyse en moyenne* constitue une facette originale de la théorie de la complexité qui ne focalise pas sur le pire des cas. En effet, sur des hypergraphes pathologiques, on pourrait obtenir un nombre exponentiel de traverses minimales : par exemple, $\{\{v_1, v_2\}, \{v_3, v_4\}, \ldots \{v_{m-1}, v_m\}\}$, de taille m, possède $2^{\lfloor \frac{m}{2} \rfloor}$ traverses minimales [127]. L'étude du pire apporte ici peu d'information.

Pour mieux appréhender la réalité, nous étudions donc le nombre moyen de traverses minimale sur une distribution d'hypergraphes aléatoires. Ces travaux font suite à de premiers résultats sur le nombre moyen de motifs fréquents voire fermés, dans des bases de données suivant des modèles de Bernoulli ou de Markov [8, 45].

4.1. Les principes d'une analyse en moyenne

Dans cette section, nous examinons les différentes étapes d'une analyse en moyenne : définition de la taille des instances que nous manipulons, les paramètres d'intérêt, modèle des instances et calcul asymptotique.

1. Un algorithme est output-polynomial lorsque le délai entre deux sorties est polynomial en la taille de l'entrée ajoutée à celle de la sortie.

Étape 1 : définir la taille des instances

Cette étape est commune à tout type d'analyse de complexité, qu'elle soit dans le pire cas ou en moyenne.

Nous nous intéressons ici aux hypergraphes comme à des matrices booléennes $M = (m_{i,j})$ avec n lignes pour les hyperarêtes de l'hypergraphe et m colonnes pour les sommets. Nous nous restreignons au cas où les hypergraphes contiennent un nombre de sommets non négligeable devant le nombre d'arêtes et *vice-versa*, *i.e.* ces paramètres sont liés polynomialement : il existe α et β tels que $m = \beta \cdot n^{\alpha}$.

Cette restriction est nécessaire car nous étudions des asympotiques, simultanément sur le nombre de sommets et le nombre d'arêtes. La taille des instances est donc définie *à la fois* par m et n, mais nous ne retiendrons pas m pour la suite en vertu de la relation ci-dessus qui l'exprime en fonction de n.

Étape 2 : définir les paramètres d'intérêt

L'analyse se focalise sur une des spécificités de l'algorithme étudié : la complexité en temps, la taille de la sortie, le nombre d'opérations arithmétiques, *etc.* Par exemple, dans la suite, nous étudions :

— $T(\mathcal{H})$: le nombre de *traverses* de $\mathcal{H}$;
— $T_j(\mathcal{H})$: le nombre de *traverses* de $\mathcal{H}$ de taille j ;
— $M(\mathcal{H})$: le nombre de *traverses minimales* de $\mathcal{H}$;
— $M_j(\mathcal{H})$: le nombre de *traverses minimales* de $\mathcal{H}$ de taille j ;
— $C(\mathcal{H})$: la *complexité en temps* de l'algorithme MTMINER (voir section 4.3) ;
— $S(\mathcal{H})$: la *taille* de l'entrée ajoutée à celle de la sortie d'un algorithme d'extraction de traverses minimales.

Étape 3 : définir le modèle aléatoire

Le modèle aléatoire que nous considérons, noté $HG(n, m, p)$, généralise le modèle pour les graphes d'Erdös et Rényi [87, 88] : la matrice représentant l'hypergraphe est vue comme une famille de variables aléatoires suivant toutes la même loi de Bernoulli de paramètre p. En d'autres termes, un sommet appartient à une arête avec une probabilité p (ou $(1 - q)$), indépendamment des autres sommets et arêtes. Au total, ce sommet appartiendra en moyenne à $p \cdot n$ arêtes et il suffit de peu de sommets (de l'ordre de $O(\log n)$) pour obtenir une traverse.

Ce modèle ne correspond pas à la réalité, mais pour un n et un m donnés, il considère une infinité d'hypergraphes, tandis que l'analyse dans le pire cas n'en considère que quelques uns. De plus, il est possible d'y analyser le nombre moyen de traverses minimales.

Étape finale : calculer asymptotiquement

L'analyse en moyenne d'un paramètre d'intérêt X étudie le comportement asymptotique de $\max_{\mathcal{H}} X(\mathcal{H})$ lorsque la taille n de $\mathcal{H}$ tend vers l'infini. Notre choix de modèle définit une distribution de probabilité P_n sur les hypergraphes (t étant le nombre de 1 dans la matrice représentant

l'hypergraphe $\mathcal{H}$) :

$$\mathrm{P}_n[\mathcal{H}] = p^t(1-p)^{nm-t}.$$

Nous éviterons d'étudier les moments d'ordre supérieur à celui de l'espérance, car les calculs sont complexes ; nous nous efforcerons plutôt de montrer que la variable est *concentrée*, quand sa variance est négligeable devant le carré de son espérance :

$$\mathrm{V}_n[X] = \beta_n \cdot \mathrm{E}_n(X)^2 \qquad \text{avec} \qquad \beta_n \underset{n\to\infty}{\to} 0.$$

Dans ce cas, grâce à l'inégalité de Bienaymé-Tchébychef [129], X est proche de sa moyenne, avec grande probabilité, et sa variance n'apporte pas d'information.

4.2. Nombre moyen de traverses

Le nombre de traverses minimales est naturellement majoré par le nombre total de traverses. De plus, si nous considérons explorer l'espace de recherche en partant de l'ensemble vide, selon des motifs de taille j croissante, on peut remarquer que :

1. quand j est petit, il y a de fortes probabiltés pour qu'un motif qui est une traverse soit également une traverse minimale. Cela signifie que M_j, le nombre de traverses minimales de taille j, est proche du nombre T_j de traverses de longueur j.
2. à partir d'une certaine petite taille $j_0 = \log_{\frac{1}{q}} \frac{mp}{\ln\ln n}$, $T_{j\geq j_0}$ est concentrée [1].

Parmi les T_j traverses, M_j sont minimales ; mais $T_j - M_j$ ne le sont pas, elles sont des sur-ensembles d'éléments de T_{j-1}, ces derniers pouvent être complétés aux maximum de $m-j+1$ manières :

$$T_j - (m-j+1)T_{j-1} \leq M_j \leq T_j.$$

Sachant que $\mathrm{E}_n[T_j] = \binom{m}{j}(1-q^j)^n$, le nombre moyen de traverses minimales est :

$$\mathrm{E}_n[M] = \exp\left(\frac{(\log n)^2}{|\log q|} - \frac{\log n}{|\log q|}\log\log n + O(\log n)\right).$$

Preuve : voir [1].

De plus, pour tout (ϵ, ϵ') avec $\frac{2\log\log n}{\log n} < \epsilon < 1$ et $0 < \epsilon' < 1$,

$$\mathrm{P}_n\left[M > \mathrm{E}_n[M]^{1-\epsilon}\right] = 1 + O\left(\frac{\log^2 n}{n^{2-\epsilon'}}\right),$$

où le O ne dépend pas d'ϵ : il y a peu de chance d'obtenir plus de traverses minimales que prévu en moyenne.

Le modèle choisi produit donc des traverses courtes, de taille $O(\log n)$, ce qui n'empêche pas d'exhiber un nombre sous-exponentiel de traverses minimales.

4.3. Complexité moyenne de l'algorithme MTMINER

Après le nombre moyen de traverses minimales, examinons maintenant la complexité d'un algorithme d'extraction particulier : MTMINER [99]. MTMINER voit les traverses comme des motifs dont le complémentaire du support disjonctif est maximum – égal à n – et les extrait selon l'approche *générer & tester* de l'algorithme A-PRIORI :

1. *génération :* par niveau, ce qui permet de vérifier *apriori* la contrainte de minimalité ;
2. *tester :* ne conserver que les motifs qui sont minimaux [2] *i.e.* dont le support disjonctif croît *strictement* relativement à ses sous-ensembles.

MTMINER

Require: un hypergraphe $\mathcal{H}$
Ensure: l'ensemble MT des traverses minimales de $\mathcal{H}$
$MT := \{\{v\} |\ v \in \mathcal{V},\ supp(\vee\{v\}) = n\}$
$N_1 := \{\{v\} |\ v \in \mathcal{V},\ n > supp(\vee\{v\}) \neq 0\}$
$j = 1$
1: **while** $N_j \neq \emptyset$ **do**
2: **for all** $W = V \cup \{v_1, v_2\}$ avec $(V \cup \{v_1\}, V \cup \{v_2\}) \in N_j^2$ **do**
3: **if** W est minimal **then**
4: **if** $supp(\vee W) = n$ **then**
5: ajouter W à MT
6: **else**
7: ajouter W à N_{j+1}
8: **end if**
9: **end if**
10: **end for**
11: **end while**

FIGURE 4.1. – Algorithme $MTminer$ d'extraction en largeur de traverses minimales.

L'algorithme de la figure 4.1 détaille ces opérations, en constituant les listes N_j des minimaux de longueur j pour le support disjonctif. Cet algorithme suggère que la complexité $C(\mathcal{H})$ de MTMINER est fortement liée au nombre $N(\mathcal{H})$ de minimaux générés : si l est la taille maximale d'une traverse (*i.e.* le nombre d'itérations de la ligne 1) :,

$$N(\mathcal{H}) \leq C(\mathcal{H}) \leq l \cdot n^4 \cdot N(\mathcal{H}). \tag{4.1}$$

On peut expliquer la borne inférieure par le fait que l'algorithme génère *tous* les minimaux. Pour la borne supérieure :

— la génération des W par fusion à préfixe commun (ligne 2) est réalisée en $O(n^3 \cdot |N_j|)$;
— étant donné W, on teste sa minimalité (ligne 3) en vérifiant que les $|W|$ sous-ensembles de taille j sont minimaux, soit $O(n)$ opérations.

2. Notons qu'un motif minimal ne donne pas nécessairement naissance à une traverse. Mathias Hagen a ainsi défini des hypergraphes avec de nombreux minimaux mais peu de traverses [96].

Notre principal résultat [1] montre qu'asymptotiquement, l'espérance du nombre de minimaux équivaut à celui du nombre de traverses minimales : $\mathrm{E}_n[N] \sim_{n\to\infty} \mathrm{E}_n[M]$. La complexité étant liée au nombre de minimaux générés, soit $\mathrm{E}_n[C] = \mathrm{E}_n[N]$, on conclut, comme pour le nombre de traverses minimales, que la complexité est sous-exponentielle :

$$\mathrm{E}_n[C] = \exp\left(\frac{(\log n)^2}{|\log q|} - \frac{\log n}{|\log q|}\log\log n + O(\log n)\right).$$

Nous répondons également à la question de savoir si MTMINER est *output-polynomial*, en qualifiant le rapport entre la complexité et la taille totale de l'entrée plus la sortie : pour tout (ϵ, ϵ') vérifiant $\frac{4\log\log n}{\log n} < \epsilon < \frac{1}{2}$ et $0 < \epsilon' < 1$:

$$\mathrm{P}_n[C \leq S^{1+\epsilon}] = 1 + O\left(\frac{\log^2 n}{n^{2-\epsilon'}}\right).$$

En d'autres termes, la complexité du THG–problem sur des instances aléatoire est, selon toute probabilité, output-polynomial et même **output quasi-linéaire** car l'exposant de S dans l'équation précédente est proche de 1.

4.4. Conclusion

Nous avons montré que le nombre de traverses minimales est, en moyenne et avec une grande probabilité, *super-polynomial*. De plus, notre étude de MTMINER, algorithme qui n'est pas efficace dès qu'il s'agit de calculer de longues traverses, a montré qu'il était néanmoins output-quasi-linéaire.

Il semble raisonnable de penser que $\mathrm{E}_n[C/S^{1+\epsilon}]$ tend vers zéro, ce qui serait un résultat plus fort, mais nous n'avons pas pu le prouver. Des calculs intermédiaires montrent cependant qu'il existe un entier k tel que $\mathrm{E}_n[C]/\mathrm{E}_n[S] = O(n^k)$. Cela suggère que MTMINER serait à délai polynomial, en moyenne et avec une probabilité non négligeable. Il nous semble difficile d'aller au-delà, sachant que l'existence d'un algorithme à délai polynomial pour THG impliquerait que $\mathrm{P} = \mathrm{NP}$ [86].

Ayant une bonne connaissance de l'algorithme DEFME, décrit au chapitre précédent, qui extrait les minimaux en profondeur, il nous semble naturel d'envisager désormais son analyse en moyenne pour le calcul des traverses minimales.

Deuxième partie

Redescription d'espace par la fouille de données

Cette partie rassemble des contributions mettant en œuvre la fouille de motifs, sous l'angle de la transformation de l'espace de description des données.

Cette partie du mémoire illustre l'idée générale que, dans des données structurées (contexte booléens, chaînes, bases de séquences, graphes), l'extraction de motifs construit une redescription des données propice à une exploitation par une méthode de décision. Cette idée n'est pas nouvelle, comme en témoignent les démarches de compression [133], de redescription [71, 79], *etc.* Elle s'inscrit dans notre pratique actuelle de l'analyse de données où la fouille est une étape préliminaire pour les techniques d'apprentissage automatique.

Nous illustrons cette place de la fouille dans deux thématiques de recherche que nous avons dirigées :

— la prédiction d'événements dans les séquences ;

— la caractérisation et la complétion des valeurs manquantes.

Nous constaterons également que la fouille de données conserve sa pertinence : à ses débuts, elle adressait des volumes de données alors « importants », désormais considérés comme « petits » devant les promesses du *deep learning*, ce dernier s'accommodant mal d'une insuffisance de données. Nous montrons ici que la fouille reste un outil performant d'extraction de connaissances actionnables et interprétables.

5. Prédiction d'événement dans les séquences

Ce chapitre relate des travaux effectués dans le cadre de collaborations initiées par Alexandre Pauchet du LITIS de Rouen. Ces collaborarions ont élé formalisées par un projet ACAMODIA, PEPS CNRS INS2I-INSHS (2011 à 2013) que j'ai porté, puis par un projet ANR nommé NarECA, pour Narrative Embodied Conversational Agent, du programme CONTINT, de 2013 à 2018. NarECA est porté par Alexandre Pauchet ; j'en suis responsable pour le GREYC. NarECA et la région Basse-Normandie ont permis de financer 18 mois de post-doc pour Gaël Lejeune.

5.1. Introduction

Les projets ACAMODIA et NarECA proposent une approche originale pour la réalisation d'un agent conversationnel : le *modèle de dialogue* utilisé par l'agent est appris à partir d'expérimentations mettant en œuvre un parent narrateur et un jeune enfant. Pour la réalisation de l'agent, nous avons œuvré à l'apprentissage de ce modèle de dialogue.

La narration d'histoires est une situation classique participant du développement de l'enfant. Les contextes sociaux et langagiers apportés par l'adulte lui sont nécessaires dans son processus d'apprentissage des compétences socio-communicatives, cognitives et morales. Les enfants développent une *théorie de l'esprit* [59] durant leurs premières années et deviennent ainsi capables d'assimiler le fait qu'une personne est déterminée par ses propres intentions, émotions et états mentaux. Ce développement n'est possible qu'au travers des situations sociales de dialogue. Le discours des adultes concernant les états mentaux, en particulier, se révèle être un médiateur d'apprentissage du concept de cognition sociale – grâce à une participation active au dialogue et à des interactions dynamiques.

L'amélioration des interactions entre agents et utilisateurs est nécessaire [67] car leur contexte de déploiement est de plus en plus vaste : assistants en ligne ou *chatbot* [81], aide aux séniors [72], jeux conversationnels [77]. Cette amélioration demande une meilleure connaissance des états mentaux pour minorer l'effort cognitif de l'utilisateur et améliorer sa satisfaction. Un agent conversationnel se devrait de maintenir l'*engagement* dans l'interaction [111] en détectant quand il est nécessaire de réorienter la conversation. Éviter la rigidité d'un schéma d'alternance de tours de parole améliore l'impression de naturel [134].

La conception d'un modèle de dialogue est une tâche difficile et souvent pluridisciplinaire : traitement de signaux multimodaux (parole, gestes, regards, posture), reconnaissance et génération de langage naturel, gestion du dialogue, modélisation des émotions, prosodie et comportement non verbal. Notre standard d'interaction se situant dans le contexte de la narration adulte-

enfant, le modèle dialogique de l'agent devrait être adapté aux compétences socio-cognitives et langagières de l'enfant.

En ce qui concerne les systèmes et modèles du dialogue pouvant être intégrés dans les agents conversationnels, plusieurs approches existent.

L'approche *à états finis* (voir par exemple [116, 122]) qui représente la structure du dialogue par un automate à états finis dans lequel chaque énoncé conduit à un nouvel état. Cependant, leur linéarité est un désavantage et ce type de stratégie échoue dans la représentation des dimensions multiples impliquées à chaque tour de parole.

L'approche *par formulaire* représente le dialogue comme un processus de remplissage de formulaire contenant des entrées prédéfinies (voir par exemple [60]). Les contributions possibles sont fixées à l'avance.

L'approche *par planification* (exemple : [56]) combine la reconnaissance de plans et la théorie des Actes de Langage [123]. Cette approche est complexe du point de vue calculatoire et requiert des composants très avancés de TAL afin d'inférer les intentions du locuteur.

Le framework ISU (*Information State Update*) [107] utilise une représentation formelle du terrain commun, l'*état d'information*, ainsi qu'une structure gérant le raisonnement de l'agent.

L'approche *logique* représente le dialogue et son contexte par un formalisme logique et utilise des mécanismes tels que l'inférence et les jeux de dialogue (voir par exemple [103, 131]).

Les approches *par apprentissage* proposent des techniques telles que l'apprentissage par renforcement [89, 125], statistique [94], stochastique [93] ou profond [82], afin de modéliser le dialogue via des processus de Markov. Ces approches requièrent un important travail d'annotation.

La plupart de ces approches sont fondées sur des processus décisionnels de Markov mais nécessitent une représentation explicite des états mentaux qui induit une certaine rigidité dans l'alternance de la parole. Le projet ANR NarECA vise à la conception d'un modèle de dialogue combinant planification au service de la résolution de la tâche – prédiction et planification des interventions de l'enfant – et gestion réactive par jeux de dialogue pour les conventions/motifs dialogiques.

Notre modèle de dialogue émerge d'une démarche de prédiction d'événement dans des séquences – une thématique peu abordée [57], au contraire de la classification – est formalisée dans [135] et utilisée par exemple dans [138] pour prédire la survenue d'une panne. Cette démarche combine explicativité et actionnabilité.

5.2. Engagement de l'interlocuteur

Nous avons abordé la modélisation du dialogue narratif sous l'angle de l'*engagement de l'enfant*. L'agent peut guider le dialogue pour favoriser ou éviter les interventions, par exemple pour susciter l'émotion chez l'enfant ou conduire la narration à sa fin. En particulier, il est important que l'agent puisse savoir *combien de temps* il doit attendre une interaction. Si l'enfant est bavard, il n'a pas besoin d'attendre. À l'inverse, une pause trop importante serait maladroite et nuirait à la qualité de l'interaction.

Nous proposons d'aborder ce problème comme une tâche de *prédiction d'événement* dans une séquence d'itemsets : étant donné un historique d'énoncés, qualifiés par des psychologues selon

une grille d'annotation des actes de dialogues, nous cherchons à prédire sous quelles conditions l'enfant intervient. En effet, il ne répond pas systématiquement à un stimulus simple comme une sollicitation directe et il est parfois nécessaire de faire appel à ses émotions pour l'impliquer.

L'utilisation de méthodes de fouille de données séquentielles répond bien à la problématique, que ce soit en terme de type de données à traiter (symboliques plutôt que numériques), de tâche (une prédiction sur un symbole et non pas une régression). Cependant, nous verrons que la fouille montre également son avantage car elle permet une resdescription compensant *le faible volume de données annotées*.

5.3. Modélisation de dialogues

La méthode de modélisation du dialogue proposée est présentée Figure 5.1 :

1. *collecte et numérisation* d'un corpus de dialogues au format audio ou vidéo, composé d'histoires enfantines racontées par des parents à leur enfant [1] ;
2. l'étape *transcription et codage* consiste à produire des données brutes à divers niveaux de détails (tours de parole, énoncés, onomatopées, pauses, etc.) selon les caractéristiques que l'on souhaite exhiber ;
3. une phase d'*extraction de régularités* (modélisation) est appliquée aux annotations ;
4. le modèle peut alors être *valorisé* par l'agent, qui peut orienter sa stratégie.

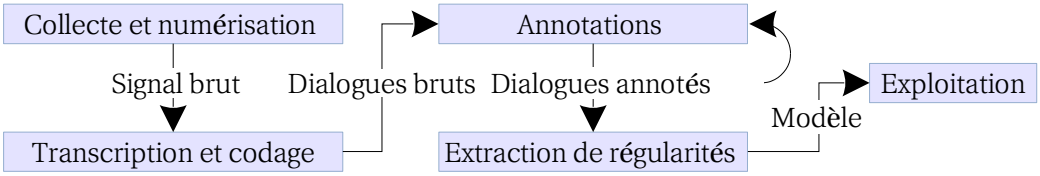

FIGURE 5.1. – Analyse du dialogue

Corpus de dialogues narratifs

Dans cette étude, nous utilisons d'abord un corpus de 90 dialogues entre parents et enfants âgés de 3, 4 et 5 ans, filmés en situation de récit d'histoires enfantines (10 enfants par tranche d'âge, 3 histoires différentes). Ces enregistrements sont retranscrits et annotés suivant une *grille mentaliste* [78] afin de faire ressortir les informations relatives aux états mentaux (croyances, volition [2], émotions, etc.) contenues dans les énoncés. La longueur moyenne des dialogues est de 90 énoncés.

Le Tableau 5.1 présente un exemple de dialogue provenant du corpus collecté. Chaque énoncé est caractérisé par un numéro de ligne, un locuteur (P : parent, E : enfant), une transcription et des annotations encodées suivant 5 dimensions :

1. Avant l'expérience, les parents prennent connaissance de l'histoire à raconter, type Petit Ours Brun, Babar, ... Pendant l'expérience, ils racontent l'histoire sans le support du livre.
2. Expression d'une volonté.

— la première colonne caractérise la nature de l'énoncé : une (A)ffirmation, une (Q)uestion, une demande d'attention - générale (G) ou concernant l'histoire (D) ;
— la seconde colonne définit la référence de l'énoncé. Il peut se référer à un personnage (P), à l'auditeur (H) ou au narrateur (R) ;
— la troisième colonne est dédiée aux états mentaux. Les interlocuteurs peuvent exprimer une (E)motion, une (V)olition, une cognition observable (B) ou non (N), une déclaration épistémique (K), une hypothèse (Y) ou une (S)urprise. La surprise se distingue des autres émotions de par son lien avec les croyances ;
— les deux dernières colonnes représentent les explications par (C)ause/conséquence, (O)pposition ou empathie (M), qui peuvent être utilisées, soit pour expliquer l'histoire (J), soit pour préciser une situation par l'évocation d'un contexte personnel (F).

Ligne	**Locuteur**	**Énoncé**	**Annotations**				
25	P	T'inquiète pas	A	P	E	-	-
26	P	Donc là ils se cachent	A	P	B	-	-
27	P	Ils cherchent	A	-	F	-	-
28	P	qui pourrait avoir pris la couronne.	Q	-	F	-	-
29	E	Elle est dedans, elle est dedans la couronne.	A	-	F	-	-
30	P	Donc là ils suspectent plein de monde, Cornélius, Céleste, la vieille dame...	A	P	Y	C	J
31	P	Qui a bien pu prendre la couronne ?	Q	-	F	-	-
32	E	La couronne elle est dedans.	A	-	F	-	-
33	P	Tu crois ? !	Q	H	K	-	-
34	E	Oui.	A	-	F	-	-
35	P	Mais Babar il ne sait pas qu'elle est dedans.	A	P	N	O	J
36	P	Donc il se dit que c'est une bombe, la couronne	A	P	N	C	J
37	P	ou je ne sais quoi.	A	R	N	-	-

TABLE 5.1. – Tableau d'annotation d'une narration parent-enfant

Par exemple, la ligne 35 (*Mais Babar il ne sait pas qu'elle est dedans.*) est encodée ainsi : l'énoncé est une affirmation (A) portant sur un état mental se référant à un personnage - "*Babar*" - (P) ; l'état mental correspondant - "*sait*" - se réfère à une cognition non observable (N) ; "*Mais*" dénote une justification par opposition (O) ; enfin, l'énoncé se réfère à l'histoire (J).

5.4. Extraction de régularités pour la modélisation du dialogue

Nous proposons de découper les retranscriptions de dialogue en *tours de parole*, caractérisés par un ensemble d'énoncés successifs provenant d'une seule personne (ici le parent ou l'enfant). Le problème revient à prévoir la fin du tour. Dans ce but, nous considérons des séquences de séries de vecteurs d'annotations (une série de vecteurs d'annotations par tour de parole) se terminant par une intervention de l'enfant. Les séquences du tableau 5.1 sont : $< (APE)(APB)(AF)(QF) >, < (APYCJ)(QF) >, < (QHK) >$, *etc.*

Pour extraire les régularités menant à la fin des séquences, les épisodes sont explorés par projections récursives grâce à un algorithme glouton. Dans l'exemple ci-dessus, l'algorithme débute avec l'épisode $< (Q) >$, commun à toutes les fins de séquences. L'algorithme est ensuite appelé une nouvelle fois sur les séquences projetées $< (APE)(APB)(AF) >, < (APYCJ) >$. (A) est ajouté à l'épisode qui devient $< (Q)(A) >$, qui est lui-même projeté une nouvelle fois : les séquences résultantes sont $< (APE)(APB) >$, *etc.* L'explosion combinatoire est limitée par deux contraintes anti-monotones : la fréquence d'apparition et la longueur moyenne des séquences et la distance moyenne en nombre d'énoncés à la fin de la séquence.

Au cours du traitement - dans lequel la séquence est parcourue de la fin vers le début - les épisodes obtenus ne sont pas nécessairement tous appropriés à la prédiction de la fin du tour de parole. Supposons, par exemple, que chaque séquence commence et termine par une (Q)uestion, l'algorithme décrit précédemment donnera comme prédicteur de fin $< (Q) >$, bien qu'il soit aussi un bon prédicteur de début. Pour éviter ce type de résultats défavorables, la distance moyenne de chaque épisode au début de la séquence doit être prise en compte. Si cette dernière est trop faible, l'épisode n'est pas conservé, ce qui garantit que la pertinence des régularités extraites sont pertinentes [20, 38].

Le processus d'extraction fournit un très grand nombre d'épisodes. Afin que l'expert puisse manuellement les évaluer, il est nécessaire d'en limiter le nombre. Dans cette optique, considérant que les épisodes sont des séquences de déplacements entre deux ensembles de vecteurs d'annotations, une approche par clustering de trajectoires [108] a été adoptée. Les déplacements sont classés et un représentant est obtenu pour chaque classe. Ce regroupement permet de passer de plusieurs centaines d'épisodes à seulement quelques dizaines de représentants que l'expert peut appréhender.

5.5. Expertise du modèle obtenu

L'évaluation par l'expert du modèle calculé souligne qu'un agent narrateur doit être interactif avec l'enfant (au travers de questions et de demandes d'attention) et ce d'autant plus avec les enfants en bas âge. Il apparaît essentiel de solliciter les enfants afin qu'ils interagissent. De plus, la compréhension des émotions et états mentaux des personnages peut être améliorée par une explication du comportement entraîné par l'état mental.

Selon l'âge des enfants, la table 5.2 résume les modèles des interactions de l'enfant, caractérisés par :

— leur *longueur moyenne*, qui correspond au nombre moyen d'énoncés entre le modèle et

l'interaction de l'enfant. Plus une séquence est courte, moins il y a d'énoncés entre elle et l'intervention de l'enfant ;
— le *modèle*, qui décrit une séquence d'annotations. Par exemple, la séquence E-Q symbolise une annotation E, suivie, plus ou moins tard, d'une annotation Q. Les annotations peuvent ne pas être dans la même dimension ;
— la *fréquence*, qui est le pourcentage de fois où le modèle apparaît.

3 ans			4 ans			5 ans		
longueur	**modèle**	**fréquence**	**longueur**	**modèle**	**fréquence**	**longueur**	**modèle**	**fréquence**
3,2	**E-Q**	**10,4%**	2,1	D-Q	14,9%	1,9	Q	35.4%
3,4	D-Q	16,8%	**2.2**	**E-Q**	**7,5%**	**2,2**	**E-E**	**9,1%**
3,5	J-Q	9,6%	2,2	Q-Q	12,7%	2,6	J-D	8,1%
3,5	D-Q-Q	9,6%	2,6	D-E	10,4%	2,7	E-D	6,1%
3,5	E-J	8,8%	2,8	D-D	11,2%	3,1	J-E	8,1%
4,3	D-E	12,8%	3,5	J	14,9%	3,4	V	13,1%
4,3	D-J	8,0%	3,8	B	7,5%	3,7	D-E	7,1%
5,4	B	10,4%	4,0	E-E	7,5%	3,8	J-J	6,1%
5,6	V	13,6%	4,1	E-D	6,7%	4,1	E-J	7,1%
			4,3	V	6,7%			

FIGURE 5.2. – Longueurs moyennes et fréquences des séquences en fonction de l'âge.

Les modèles fournis mettent en lumière les variations d'état mental améliorant l'interaction [9] :
— quel que soit l'âge, les séquences contenant des justifications sont fréquemment associées à divers indices (émotion, demande d'attention ou question). Dans ce contexte, l'interaction de l'enfant ne survient qu'entre 3,1 et 4,3 énoncés après le modèle ;
— la longueur des interactions décroît avec l'âge, de 3,2 à 1,9 énoncés ;
— le nombre d'énoncés auxquels sont associées des émotions est quasiment équivalent pour tous les âges. Néanmoins, plus l'enfant est âgé, plus les séquences d'émotions sont variées. Les séquences complexes (émotions et justifications : J-E ou E-J) n'apparaissent qu'avec les enfants les plus âgés ;
— à l'exception des demandes d'attention, les modèles les plus efficaces (en rouge et gras dans le tableau 5.2) contiennent toujours des émotions (E-Q ou E-E).

5.6. Prédiction de l'intervention de l'enfant

Les expériences relatées précédemment ont été poursuivies au cours du projet ANR NarECA, depuis 2013. Gaël Lejeune a proposé une approche combinant règles et extraction de motifs pour évaluer le modèle : étant donné l'historique d'un dialogue entre un narrateur (humain ou agent) et un enfant, la tâche consiste à prédire quel est le prochain locuteur. La motivation est double : premièrement, un tel modèle produit une probabilité sur le prochain locuteur, probabilité qui fournit une *mesure de l'engagement* que l'agent peut interpréter selon son plan. Deuxièmement, l'engagement permet à l'agent d'estimer le temps d'attente après une question directe ou une demande d'attention afin de ne pas dégrader la qualité de l'interaction.

Les dialogues DIT++

Dans le projet NarECA, les dialogues de narration sont analysés selon une grille DIT++ de codification des actes de dialogues. DIT++ (Dynamic Interpretation Theory [73]) permet d'annoter chaque groupe d'intonation (ou *breath-group* [110]) ou phrase à l'aide d'une paire *(fonction, dimension)*[3], voir Table 5.2 pour un exemple. Nous considérons donc des bases de séquences de motifs d'annotations sur des groupes d'intonnation.

Speaker	Breath Group	Function	Dimension
A	So, it's morning, chidren are coming to school	Inform	Task
C	**Yes**	Contact	Contact Management
A	Look at this child	Suggestion	Task
A	he does not seem happy to be there	Inform	Task
C	**Yeah, I saw**	Confirm	Task
A	And this boy, he has a ball...	Inform	Task
A	Look,	Suggestion	Task
A	it's Salim, he is calling his friends, [...]	Inform	Task
C	**Uh Uh**	Stalling	Time

TABLE 5.2. – Trois séquences qui se terminent par une intervention de l'enfant. Chaque ligne est un breath group, l'intervention de l'enfant est en gras.

Mesure de l'engagement

Nous avons comparé trois méthodes pour calculer la probabilité d'intervention de l'enfant, ainsi qu'une méthode hybride réunissant l'ensemble :

1. une *baseline* induisant une intervention en cas de sollicitation directe : question directe (DQ) et demande d'attention (CfA) ;
2. une méthode *directe*, exploitant les états mentaux signalés par les annotations pour prédire le prochain locuteur, à l'aide d'un classifieur standard (NaïveBayes, C4.5, ...) ;
3. une approche *motif séquentiel* : les dialogues sont recodés selon l'apparition des motifs séquentiels fermés (obtenus par table de suffixes selon [130]) ;
4. une approche *hybride* dans laquelle le classifieur dispose des attributs du motif à classer ainsi que des motifs séquentiels fermés apparus.

La table 5.3 rassemble les résultats.
L'approche *motif séquentiel* est très performante, pour deux raisons :

— le recodage des séquences de motifs en motifs de motifs séquentiels permet d'adresser des données séquentielles avec un classifieur standard ;
— elle projette les données initiales dans un espace de redescription, de dimension nettement supérieure à celui de la description initiale, dans laquelle la séparation des données est plus aisée.

3. La *fonction* qualifie si l'intention du locateur est d'apporter/obtenir de l'information ou s'il sollicite une action, la *dimension* qualifie la contribution du locuteur à la conversation : tâche, retour, animation des tours de parole, ...

Method	Variant	P	R	F_1	$F_{0.5}$
Baselines	DQ	65.5	44.2	52.8	59.8
	CfA	83.1	15.5	26.2	44.4
	DQ OR CfA	69.3	59.7	64.2	67.2
Direct Method	NaiveBayes	67.5	64.6	66.0	66.9
	SMO	69.2	70.1	69.7	69.4
	C4.5 Tree	69.0	69.5	69.2	69.6
	Random Forest	69.2	67.6	68.4	68.0
Pattern Mining	NaiveBayes	66.7	65.1	65.9	66.4
	SMO	71.3	62.0	66.3	69.2
	C4.5 Tree	**76.1**	61.2	67.9	72.6
	Random Forest	69.5	60.4	64.6	67.5
Hybridisation	NaiveBayes	68.8	70.2	69.5	69.1
	SMO	72.6	70.1	71.3	72.1
	C4.5 Tree	71.1	70.7	70.9	71.0
	Random Forest	74.1	**71.5**	**72.8**	**73.6**

TABLE 5.3. – Résultats ((P)recision, (R)appel, (F)-score) en classification pour les différentes méthodes de prédiction. Le F-score est calculé avec un classique β de 1 puis de 0,5 pour favoriser la précision.

Ce dernier argument illustre parfaitement le propos que nous souhaitons développer dans cette partie : au même titre que les noyaux, la fouille de données peut être considérée comme une méthode de redescription des données qui facilite leur séparation. Lorsque les données sont insuffisamment nombreuses pour rendre efficaces les approches à base de descente de gradient, la redescription par la fouille est prometteuse.

5.7. Conclusion

Au travers des projets PEPS Acamodia et ANR NarECA, et afin de piloter un agent conversationnel, nous avons été confronté à des problématiques de modélisation du dialogue à partir d'annotations. Pour cela, nous avons proposé une combinaison de fouille de motifs séquentiels et d'apprentissage automatique.

La fouille de motifs comme redescription des données, cela consiste à ne plus voir les données comme des séquences de motifs mais plutôt comme des motifs d'expression de séquence. Cette approche se montre pertinente dans notre configuration particulière où le modèle est obtenu à partir de relativement peu de données et nécessite d'être interprété.

6. Complétion de données manquantes par enrichissement

Ce chapitre reprend des travaux de Leïla Ben Othman pendant sa thèse [65] et de Saad Ouadrim lors de son stage de Master Recherche [120].

Nous traitons dans ce chapitre de la classique problématique des données contenant des *valeurs manquantes*. En fouille de motifs, ce phénomène désigne, pour un objet, les attributs dont on ne connaît pas la valeur.

En particulier, nous proposons une méthode de *complétion* (ou *imputation*), qui tire profit des conditions dans lesquelles les valeurs sont manquantes. Nous sommes par ailleurs en rupture avec les méthodes traditionnelles :

— nous ne considérons pas que les valeurs manquantes suivent un modèle aléatoire, mais ont en général une explication rationnelle qui sera valorisée pour aider à sa complétion ;
— nous ne remplacerons pas une valeur manquante par une valeur du domaine de définition, mais enrichirons les données avec de nouveaux symboles caractérisant les valeurs manquantes. Ces symboles dirigent ensuite la complétion vers un élément du domaine de définition ;
— les bases de données incomplètes utilisées lors de nos expérimentations ne seront pas produites par génération aléatoire de valeurs manquantes, mais utiliseront plutôt un modèle à base de règles qui permet de simuler de façon plus réaliste les correlations naturelles entre les valeurs manquantes ;
— pour évaluer la pertinence d'une méthode de complétion, il est spécieux de mesurer l'impact sur une performance en classification supervisée. En effet, le modèle calculé sur une complétion tend à être général, ce qui favorise la performance en classification.

Ce travail illustre le potentiel de la fouille de données en terme de redescription d'espace : les valeurs manquantes sont caractérisées par des associations induisant un graphe mis à profit pour la décision de complétion.

6.1. Caractérisation des valeurs manquantes

Prenons le cas de données d'un essai thérapeutique sur la maladie de Hodgkin, un cancer des ganglions lymphatiques. Les ganglions des patients de l'étude sont palpés par le médecin à la recherche d'un envahissement par le cancer qui rend le ganglion très dur au toucher, comme pourrait l'être un grain de riz. Un ganglion est donc caractérisé par un possible envahissement (oui/non) et la dimension de l'envahissement (sa longueur).
Le recensement des ganglions envahis pose déjà plusieurs problèmes de valeurs manquantes :

1. un ganglion non envahi n'est pas mesuré. Dans ce cas la dimension de l'envahissement est déclarée manquante ;
2. au cours des premières années de l'essai thérapeutique, les ganglions cervicaux gauche et droite n'ont pas été palpés, d'où deux valeurs manquantes pour chacun des ganglions : envahissement et dimension. Il s'agit là d'un classique problème de fusion de données ;
3. le protocole n'a pas été toujours respecté, certains ganglions n'ayant pas été palpés, donc non mesurés : lorsque l'envahissement est manquant, sa dimension l'est également.

Notre première contribution complète le travail séminal de modélisation de Little et Rubin [112], traditionnellement utilisé pour les valeurs manquantes. Il y est considéré qu'un *processus d'acquisition de données* transforme des entrées U, les *variables mesurées*, en sorties X, les *valeurs mesurées* [63], finalement les seules entrées à la disposition de l'expert. Cette transformation « efface » des valeurs en les rendant *manquantes*. Little et Rubin distinguent trois types :

— MCAR (missing completely at random) caractérise les valeurs manquantes sans explication rationnelle, *i.e.* sans corrélation avec les valeurs des autres entrées ;
— MAR (missing at random) qualifie les valeurs manquantes induites par des valeurs particulières des autres variables ;
— MNAR (missing not at random) caractérise les valeurs manquantes apparaissant sur une variable qui est effacée lorsqu'elle prend une valeur particulière.

Ce modèle induit de fortes contraintes qui limitent son utilisation dans la pratique. Le principal reproche tient à sa qualité de *modèle* : il propose une formulation pour l'opération d'acquisition mais n'indique pas comment procéder lors de données réelles, lorsque seule la sortie de l'acquisition est disponible. Dans la pratique, à moins de simuler l'acquisition et l'effacement, on ne dispose pas de la vérité sur les valeurs d'entrée ; en l'absence d'expertise, ce modèle est donc impossible à caractériser.

Ne prenant pas en compte la *localité* des situations, ce modèle affecte toutes les valeurs manquantes d'un même attribut à un *même type*. Enfin, il ne prend pas en compte les phénomènes de cascade de valeurs manquantes, lorsque la présence d'une valeur manquante sur un attribut déclenche une valeur manquante sur un autre attribut.

Notre typologie différencie les valeurs manquantes selon les relations qu'elles entretiennent avec les valeurs d'autres attributs (mesurées ou manquantes). Ces relations sont mises en évidence par des règles d'association dont les valeurs manquantes sont les conclusions.

Plus précisément, nous extrayons des règles $X \rightarrow miss(attr)$ concluant sur la présence d'une valeur manquante. Selon le type de la prémisse, une valeur manquante sera :

— *directe* : quand il existe une relation entre la valeur manquante, qu'on cherche à caractériser, et des données mesurées. Par exemple, la règle $essaiH7 \rightarrow miss(chd)$ [1] indique que l'essai H7 n'a pas pris en compte les envahissements du *chd* (ganglion cervical haut droit), dont les valeurs sont manquantes ;
— *indirecte* : quand il existe une relation entre la valeur manquante qu'on cherche à caractériser, et d'autres valeurs manquantes. Par exemple, la règle $miss(chd) \rightarrow miss(chddim)$ indique qu'on ne connaît pas la dimension de l'envahissement du *chd* lorsqu'on ne sait pas s'il est envahi ;

1. $miss(attribut)$ indique que $attribut$ est manquant.

— *hybride* : quand il existe à la fois une relation avec les données mesurées et d'autres valeurs manquantes. Par exemple, $plaq <= 600 \wedge miss(chd) \rightarrow miss(chg)$ signifie que, lorsque le taux de plaquettes est inférieur à 600 et l'envahissement du *chd* est inconnu, l'envahissement du *chg* est aussi manquant ;
— *aléatoire* : quand il n'existe aucune relation avec des données mesurées ou d'autres valeurs manquantes. Les relations étant obtenues par l'extraction de règles d'association, une valeur est aléatoire si elle n'est caractérisée par aucune règle en-dessus du seuil de fréquence minimale.

Cherchant à modéliser des situations claires relevant de la constitution des données, nous nous limitons à des règles de confiance 100%, à prémisse minimale pour une conclusion donnée, appelées *implications propres* dans [128]. Les prémisses de ces règles sont des *jumping emerging patterns* – présents dans la partie $\mathcal{D}_{miss(attr)}$ des données contenant les attributs manquants, absents de l'autre. Ces motifs sont extraits efficacement en poussant la contrainte de support dans $\mathcal{D}_{miss(attr)}$ dans un algorithme d'extraction de traverses minimales des complémentaires de $\mathcal{D}_{\overline{miss}(attr)}$ [65]. La contrainte de support est facile à pousser, que ce soit en utilisant l'algorithme en profondeur DEFME (*cf.* algorithme 3.2) ou MTMINER (algo. 4.1). Grâce à la minimalité approchée (*cf.* section 3.5), on peut obtenir des règles plus souples, de confiance moindre, généralement à un δ absolu près.

Les tables 6.1 à 6.3 illustrent la méthode sur un exemple jouet :
— les données initiales de la table 6.1 présentent huit objets décrits par quatre attributs, certaines valeurs manquent. En particulier, pour l'attribut A_4, dont le domaine de valeur est $\{h, i\}$, toutes les occurrences de i ont été transformées en valeur manquante ;
— les règles, de fréquence minimale de 2, sont à la table 6.2 ;
— les caractérisations de chaque valeur manquante sont reportées à la table 6.3.

À la différence de la modélisation classique, notre typologie autorise une certaine variété selon les objets. Sur l'exemple de la table 6.3, les valeurs manquantes sur A_4 (originellement que des i), sont de quatre types différents. Cette variété offre un point de vue réaliste sur les situations d'acquisition de données.

	A_1	A_2	A_3	A_4
o_1	a	c	**?**	h
o_2	**?**	c	e	**?**
o_3	a	c	**?**	h
o_4	a	d	f	**?**
o_5	**?**	c	f	**?**
o_6	b	**?**	f	h
o_7	a	**?**	g	**?**
o_8	**?**	d	g	**?**

TABLE 6.1. – Exemple d'un contexte incomplet.

	règle	fréquence	support
R_1	$a \wedge c \rightarrow miss(A_3)$	2	$\{o_1, o_3\}$
R_2	$miss(A_1) \rightarrow miss(A_4)$	3	$\{o_2, o_5, o_8\}$
R_3	$a \wedge h \rightarrow miss(A_3)$	2	$\{o_1, o_3\}$
R_4	$c \wedge miss(A_4) \rightarrow miss(A_1)$	2	$\{o_2, o_5\}$
R_5	$c \wedge h \rightarrow miss(A_3)$	2	$\{o_1, o_3\}$
R_6	$d \rightarrow miss(A_4)$	2	$\{o_4, o_8\}$
R_7	$g \rightarrow miss(A_4)$	2	$\{o_7, o_8\}$

TABLE 6.2. – Règles de caractérisation des valeurs manquantes

	A_1	A_2	A_3	A_4
o_1	-		{direct}	-
o_2	{hybride}	-	-	{indirect}
o_3	-	-	{direct}	-
o_4	-	-	-	{direct}
o_5	{hybride}	-	-	{indirect}
o_6	-	{random}	-	-
o_7	-	{random}	-	{direct}
o_8	{random}	-	-	{direct, indirect}

TABLE 6.3. – Typologie des valeurs manquantes.

6.2. Enrichissement de données incomplètes

Pour attribuer cette valeur de remplacement à une valeur manquante, l'ensemble des règles de caractérisation est assimilé à un ensemble de graphes orientés, dits *graphes de caractérisation* (cf. Figure 6.1 pour nos données d'exemple). Les nœuds sont des valeurs mesurées ou manquantes, les arcs symbolisent les associations. Du fait de l'orientation des arcs, le graphe, suggère un raisonnement abductif : on cherche une explication aux causes (les valeurs manquantes) à partir des faits (les valeurs connues).

Sur les conseils d'Arnaud Soulet [92], nous avons encadré un stage de master visant à utiliser l'algorithme du PageRank [106] (ou marche aléatoire avec téléportation) pour faire émerger l'explication à chaque valeur manquante dans chaque graphe de caractérisation. Cet algorithme consiste à simuler une marche aléatoire dans un graphe orienté et à calculer la probabilité de présence d'un surfeur aléatoire en chaque nœud du graphe. Le nœud ayant la plus forte probabilité de présence est retenu comme explication.

Sur nos données d'exemple, les symboles expliquant la présence de valeurs manquantes permettent de les enrichir comme illustré à la table 6.4.

Bien que ce ne soit pas le cas dans notre exemple, certaines valeurs manquantes peuvent ne pas être enrichies, lorsque le graphe ne contient pas de source. La situation extrême concerne les valeurs manquantes déclarées aléatoires, faute d'un seuil suffisant. Il est alors d'usage de calculer un *modèle des données non manquantes* afin de prédire les valeurs manquantes (voir

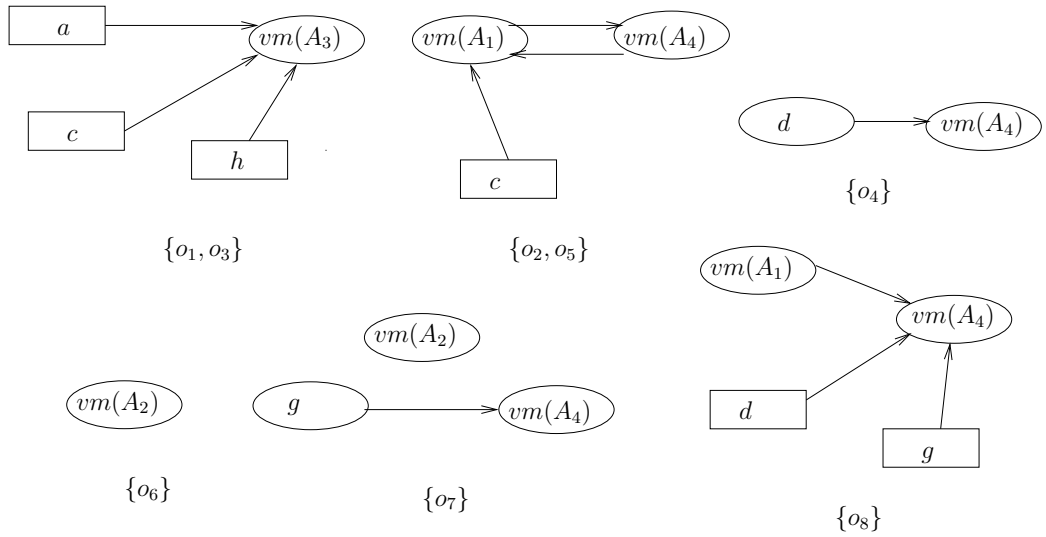

FIGURE 6.1. – Graphes de caractérisation pour les données de la table 6.2, soulignés par leur support.

par exemple [119]).

L'enrichissement est donc une étape préliminaire à la prédiction de la valeur de complétion. Cette étape revient à un *un clustering de couples* $(miss(attribute), object)$ fournissant pour chaque valeur manquante un identifiant de groupe, à valoriser par le modèle obtenu sur les données complètes.

6.3. Évaluation de la méthode de complétion

Traditionnellement, les méthodes de complétion sont évaluées selon le protocole suivant :

1. insérer des valeurs manquantes aléatoires dans une base complète, en effaçant une partie des *données initiales* ;
2. appliquer la méthode de complétion ;
3. comparer les données complétées avec les données initiales ;
4. ou mesurer la performance d'une méthode de classification supervisée.

La première de ces étapes est tout-à-fait critiquable. Comme nous l'avons précédemment évoqué, des valeurs manquantes apparaissent dans des conditions particulières que l'aléatoire seul ne peut retranscrire. Ainsi, il nous paraît important que des expérimentations *réalistes* insèrent des valeurs manquantes aléatoires autant que des valeurs manquantes directes, voire indirectes. Dans sa thèse [65], Leïla Ben Othman échantillonne une petite quantité d'associations exactes d'une base complète pour en effacer les conclusions.

Le point 3 est d'une évaluation classique est peu discutable : on souhaite que la méthode de complétion fournisse des données conformes aux originales. En revanche, une bonne per-

	A_1	A_2	A_3	A_4
o_1	a	c	ach	h
o_2	c	c	e	c
o_3	a	c	ach	h
o_4	a	d	f	d
o_5	c	c	f	c
o_6	b	?	f	h
o_7	a	?	g	g
o_8	?	d	g	dg

TABLE 6.4. – Enrichissement des données de la table 6.1.

formance en classification supervisée sur les données complétées ne représente pas un critère pertinent. En effet, nous avons constaté qu'un certain taux de valeurs manquantes aléatoires orientait le calcul vers un modèle plus général donc plus performant. L'effacement de données agit comme une technique de *régularisation* pénalisant la complexité d'un modèle afin d'éviter le phénomène de surapprentissage. Les architectures actuelles en Deep Learning utilisent pour cela des couches de *dropout* qui annulent un signal selon une probabilité uniforme.

Dans sa thèse, Leïla Ben Othman a donc préféré examiner la stabilité des données pour mettre à l'épreuve notre méthode d'enrichissement. Cette stabilité mesure le taux d'accord (indice de Rand [121]) entre un clustering sur les données initiales et un clustering sur les données complétées. Hélas, les résultats n'offrent pas de pertinence statistique, en premier lieu car les méthodes de clustering utilisées, comme *K-means*, n'ont pas la réputation d'être stables.

Depuis 2011, nous avons poursuivi ces recherches en améliorant la méthode d'enrichissement de façon à ce qu'elle valorise les résultats d'une marche aléatoire (*cf.* section précédente). Bien que procédant d'une démarche raisonnablement argumentée, l'apport de l'enrichissement reste très difficile à évaluer d'un point de vue quantitatif : d'une part, les données générées sont par nature incomparables avec les données initiales ; d'autre part, l'augmentation importante du domaine de définition que l'enrichissement génère pertube profondément la taille des espaces mis en œuvre, rendant hypothétique l'obtention d'une mesure de qualité adéquate à tous les espaces.

6.4. Conclusion

Comme nous l'avons montré, l'évaluation d'une méthode de complétion est un challenge, au moins aussi important que la conception de la méthode. Cette thématique nous aura également permis de rebondir sur les fondamentaux de la fouille, rendus nécessaires pour résoudre le problème de l'extraction de caractérisations minimales et les valeurs manquantes fournissent une nouvelle occasion d'illustrer notre thème de la redescription par la fouille : les données ne sont plus seulement des couples attribut/objet mais deviennent des motifs, des règles puis des graphes de dépendance.

Les perspectives de ce travail concernent donc la complétion par injection du Pagerank des données incomplètes dans les graphes représentant les données complètes.

Troisième partie

Fouille de données (e-)sportives

Cette partie recense des applications de la fouille pour l'analyse de données relatives au sport ou au *e-sport*, le jeu vidéo pratiqué en compétition. Ces contributions sont plus méthodologiques que théoriques et sont donc relatées sommairement. Elles trahissent un intérêt personnel voire précoce pour ces domaines émergents, et une volonté de fédérer des experts (sportifs, psychologues, sociologues, économistes, *etc.*) autour de projets de valorisation de leurs données.

Au-delà de la légèreté du contexte sportif devant des enjeux sociétaux plus *sérieux*, comme le cancer ou Alzheimer, notre intérêt se situe également dans une meilleure compréhension des pratiques sportives au travers de l'analyse de données. Le sport, comme tout autre sujet, est traversé par la déferlante *Big Data*, que ce soit pour accompagner la pratique sportive, prévenir les blessures, analyser et optimiser la performance, ou améliorer l'expérience utilisateur : la donnée y est l'un des premiers vecteurs d'action.

Quant à la pratique du sport électronique, autrefois confidentielle, elle captive désormais plusieurs millions de spectateurs et ses dotations se chiffrent en millions de dollars. Grande pourvoyeuse de traces, cette pratique suscite un intérêt croissant dans la communauté Fouille de données, comme en témoigne son inscription depuis 2015 aux thèmes de l'atelier Machine Learning and Data Mining for Sports Analytics d'ECML/PKDD.

7. Fouille de données sportives

En terme d'analyse de données de sport réel, j'ai été un élément moteur dans les collaborations avec des chercheurs multi-disciplinaires évoluant en Sciences et Techniques des Activités Physiques et Sportives du laboratoire CesamS - EA4260 de l'Université de Caen Normandie, auquel je suis officiellement associé depuis 2017.

Ces collaborations sont difficiles à valoriser dans la communauté "fouille de données" car elles se limitent souvent à proposer un cadre applicatif original à des méthodes maîtrisées. Considérant également que la science des données doit opérer dès la collecte, notre rôle aura souvent été déterminant pour des chercheurs en sciences humaines égarés devant la masse de données [4, 31, 30].

Plus précisément, deux collaborations majeures sont rapportées ci-après, l'une sur la mise en évidence des raisons de l'abandon arbitral et l'autre sur l'analyse d'un gros volume de trajectoires en Tennis.

7.1. Analyse de l'abandon arbitral

Un sondage de 135 questions a été effectué auprès de 1 718 arbitres (35% de femmes) d'une moyenne de 38 ans, officiant dans 35 disciplines sportives différentes. Un dixième de la population évolue au niveau international, le reste se répartit pour moitié entre les niveaux national et régional. Le sondage posait en particulier la question suivante : « Avez-vous eu un jour l'envie d'abandonner l'arbitrage ? ».

Pour caractériser cette population « abandonniste », nous avons proposé le calcul de motifs émergents minimaux avec l'enchaînement complémentaires/traverses minimales/contrainte de support (*cf.* sections 3.5, 6.1 et figure 7.1). Un étude statistique sur les motifs produits met en évidence, entre autres, l'isolement des arbitres et leur obligation à officier pour assurer la survie de leur club d'affiliation [17].

Cet exemple de collaboration montre qu'en sport comme ailleurs, où la tradition des statistiques fondées sur les tests est bien ancrée culturellement, les experts sont plutôt friands de méthodes exploratoires et, dans ce cas particulier, d'émergents minimaux.

7.2. Big Tennis Data

Ce projet a été accompagné par un financement CPER sur le sujet des humanités numériques pour quatre mois de postdoc.

Nous avons pu réaliser une première analyse de grande envergure de données HawkEye[1],

1. Cette société équipe de dix caméras un court de tennis et propose des reconstructions de la trajectoire à des fins d'arbitrage. `https://www.hawkeyeinnovations.com/`

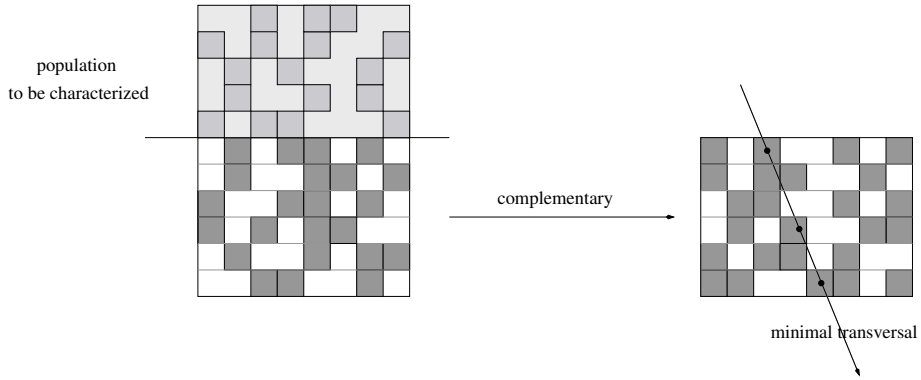

FIGURE 7.1. – Extraction de motifs émergents minimaux.

concernant 260 000 trajectoires de la balle de tennis au cours de 1729 matches disputés entre 2003 et 2008 au cours des tournois majeurs du circuit professionnel.

Sous l'angle de l'impact du service, nous avons reconstitué l'effet Magnus : un solide en rotation $\overrightarrow{\omega}$ subit une force dite de Magnus, orientée selon $\overrightarrow{v} \wedge \overrightarrow{\omega}$, qui courbe la trajectoire. Précisons que les trajectoires vendues par HawkEye en sont une approximation polynomiale de degré 4, ce qui n'est pas physiquement cohérent, mais informatiquement avantageux. Malgré cela, l'approximation linéaire du coefficient de Magnus fonctionne bien. Ces résultats fournissent des préconisations aux joueurs et entraîneurs, enrichissent le développement personnel ou préparent une opposition [3, 24]. Des analyses préliminaires indiquent par exemple les distributions des angles et coefficients d'effet appliqué à la balle lors du service (figure 7.2).

Un travail important a été mené pour nettoyer ces données très volumineuses et reconstruire l'effet Magnus. Les perspectives concernent la mesure de l'impact du service sur le déroulement de l'échange.

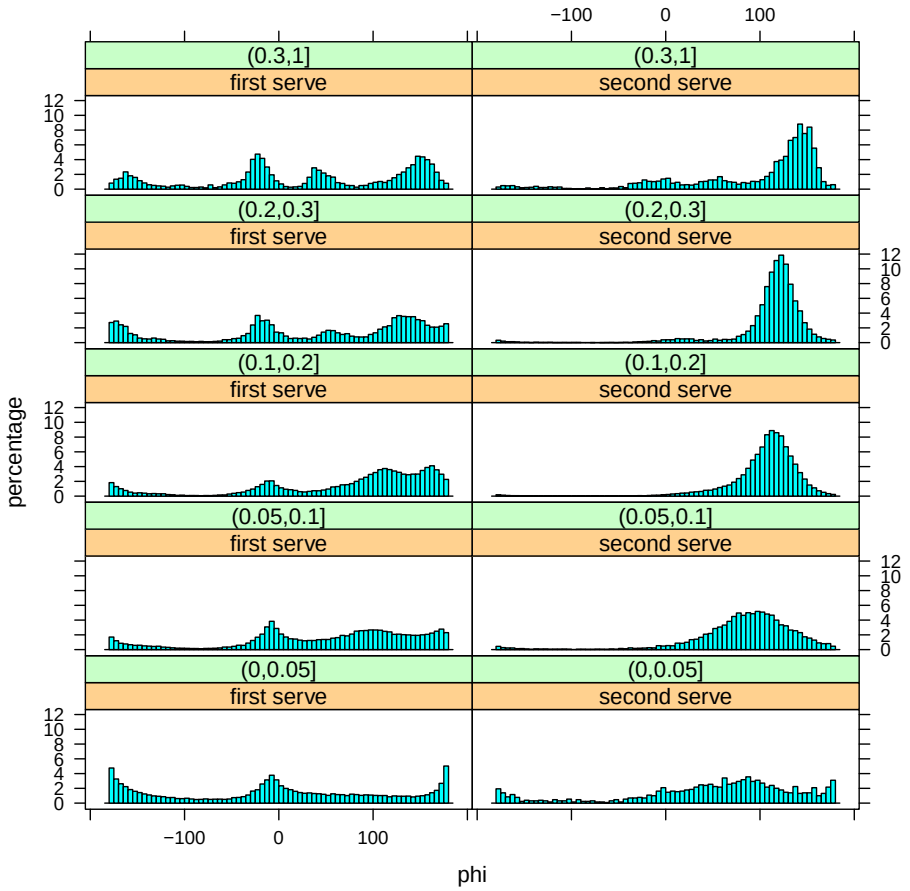

FIGURE 7.2. – Distribution de l'angle ϕ pour les droitiers selon l'intensité du coefficient de Magnus, croissant de bas en haut de la figure. ($\phi = 180^o$ pour un lift pur, $\phi = 90^o$ pour un slice pur).

8. Fouille de données e-sportives

Bien que l'intelligence artificielle ait fait de nombreux progrès, l'opposition à une entité mécanique n'a ni l'inventivité ni la réactivité propre à l'humain. Les amateurs de jeu vidéo apprécient de se mesurer entre eux, par défi et esprit compétitif. Lorsque qu'il permet la constitution d'équipes, le jeu devient un lieu collectif et convivial : Internet et les équipements connectés abolissent les distances. Des communautés d'intérêt se développent autour d'un jeu et le font vivre sur la durée, en alimentant des forums, en rédigeant des guides ou en organisant des événements.

Cette pratique amène une professionnalisation des acteurs autour d'une discipline : le e-sport. Les audiences massives des événements incitent leurs organisateurs à adopter une démarche de sport-spectacle, afin de garantir une expérience complète au fan.

8.1. Motivations

L'industrie du jeu vidéo génère toujours plus de traces d'interaction avec les utilisateurs. Elle peut désormais capter ces traces et les analyser, d'autant plus facilement que les deux mutations suivantes opèrent :

1. la dématérialisation en *computer as a service* ou *cloud gaming* [1] : le jeu tourne sur un serveur externe, le joueur interagit avec son flux vidéo et les traces sont captées par le serveur ;
2. *Game as A Service* : une majorité de jeux est désormais gratuite, sur le modèle *free-to-play* et le retour sur investissement s'effectue sur les transactions effectuées à l'intérieur du jeu (*in-game*). En effet, les joueurs passent de plus en plus de temps sur un jeu, maintenant une activité régulière au cours de plusieurs années. L'industrie du jeu est donc très demandeuse de CRM pour satisfaire ses utilisateurs car la donnée peut être valorisée pour générer de l'engagement : le middle-ware B2B ou B2C sur le *game analytics* s'emballe [2]. En particulier, la méthode d'apprentissage fait appel à l'*attention*[61] qui consiste à nourrir un RNN avec un flux plutôt qu'un vecteur de taille fixe, tout en lui permettant d'accéder à sa mémoire [136, 132].

Par ailleurs, il existe plusieurs raisons de considérer le e-sport davantage comme un objet sportif qu'un simple jeu :

— il implique des humains soumis à une performance ;

1. voir `http://www.blade-group.com/` pour une pure approche orientée ressource de calcul ou `https://www.blacknut.com/`, qui se veut le Netflix du jeu vidéo.

2. *cf.* les sites des startups `http://haste.net` à propos de l'infrastructure réseau, `http://microcoaching.net` une place de marché pour le coaching de joueur, `http://http://esports.one/` pour un commentaire automatique de retransmission à l'aide d'infographie statistique, `https://www.showdown.cc/` pour la génération de résumé automatique de partie.

— les performances peuvent être analysées, expliquées et améliorées au cours d'un suivi longitudinal ;
— selon l'axe stratégique du collectif, la plupart des types de confrontation sportive sont disponibles dans des jeux réalistes (football, basket, ...) ou fantaisistes (par exemple le rugby avec les MOBA, étudiés ci-dessous).

8.2. Analyse de trajectoires pour DotA

Une partie de ce travail a été réalisé par Alexandre Letois au cours de son stage de Master Recherche [109].

Nous avons étudié le jeu Defense Of The Ancients (DotA), un jeu de bataille en ligne (ou MOBA pour Multiplayer Online Battle Arena), prenant la forme d'un gagne-terrain analogue au rugby, opposant deux équipes d'avatars pilotés par autant d'humains.

Notre volonté était de valoriser des traces de jeu, par exemple certaines trajectoires, et d'en déduire des préconisations stratégiques.

Les deux camps de base de chaque équipe sont situés aux coins opposés de la carte (*cf.* figure 8.1). Trois chemins principaux (un central, deux latéraux) relient les deux bases. Toutes les 30 secondes, trois petites factions quittent chaque base par les trois chemins et se dirigent vers la base adverse. Sur leur chemin, elles rencontrent une faction adverse. Les trois affrontements résultant définissent une *ligne de front*, initialement sur la diagonale descendante.

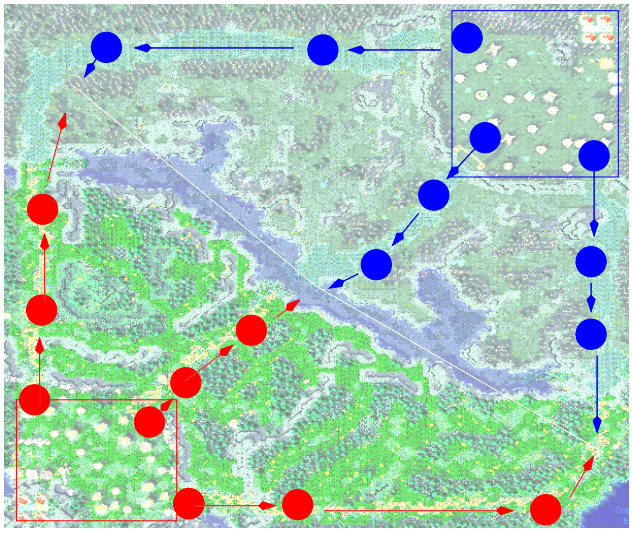

FIGURE 8.1. – La carte d'un MOBA (ici DotA).

Comme tous les jeux de gagne-terrain, le concept de ligne de front est primordial dans les MOBA. Sans intervention des joueurs, cette ligne est en équilibre instable et oscille lentement. Des tours de défense (trois par chemin, représentées par des cercles sur la figure), viennent stopper et saturer ces oscillations naturelles.

En incarnant un héros doté de pouvoirs, le joueur humain accompagne et favorise la progression de ses factions autonomes, par exemple en neutralisant un ou plusieurs joueurs adverses. Un MOBA est finalement une sorte de rugby à cinq sur un très grand terrain, avec trois ballons dans trois lignes différentes !

Obtenir des préconisations stratégiques est un objectif ambitieux, sur le chemin duquel nous avons repéré des facteurs topologiques déterminants pour le gain d'un match. Nous avons étudié la répartition spatiale des cinq joueurs d'une équipe, représentée par un pentagone (*cf.* figure 8.2) dont les caractéristiques géométriques sont mesurées (diamètre, aire et périmètre) ainsi que des caractéristiques stratégiques, sous l'angle du jeu collectif :

— capacité de regroupement (moyenne des distances au barycentre) ;
— inertie de regroupement (moyenne des carrés des distances) ;
— distances à la base et au but ;
— distance au barycentre de l'ennemi.

8.3. Conclusion

Nous avons montré que les moyennes de ces mesures au cours du deuxième quart-temps de la partie permettent à un algorithme d'apprentissage classique (arbre, SVM, *etc.*) d'obtenir 95% d'aire sous la courbe ROC pour la prédiction de l'équipe victorieuse [12]. Nous exploitons actuellement la chronologie précise de ces mesures pour nourrir un réseau de neurones récurrent apprenant les événements marquants d'une rencontre de haut niveau, en espérant aller ainsi vers la mise en évidence de comportements caractéristiques des équipes impliquées.

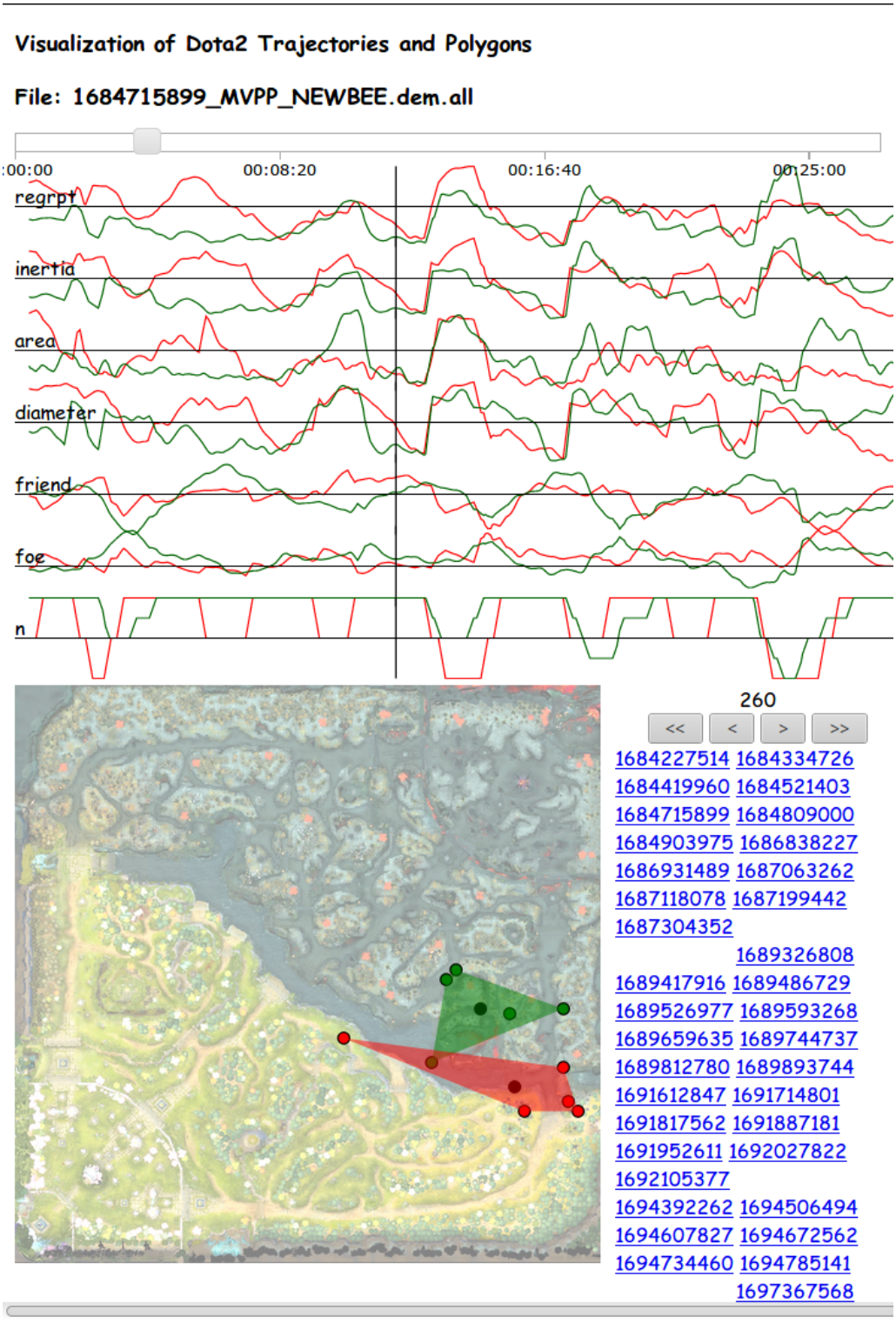

FIGURE 8.2. – Représentation des équipes sous forme d'un pentagone (*cf.* `https://rioultf.users.greyc.fr/polygon/`).

Conclusion

Comme le montre ce travail, notre spectre de recherche est large, du fondamental à l'applicatif. La première partie a rapporté des recherches théoriques sur un domaine *délaissé* de la fouille de données : l'extraction de minimaux et l'analyse de complexité en moyenne. La deuxième partie a décrit les contributions de jeunes chercheurs que nous avons encadrés, à propos de techniques de transformation d'espace par la fouille, afin de modéliser un dialogue ou de compléter des valeurs manquantes. La dernière partie a illustré des applications de la fouille sur des données sportives, dans un domaine scientifique multi-disciplinaire, ou émergent : le e-sport.
Nous souhaitons désormais orienter notre projet de recherche vers les axes suivants :

1. contribuer à l'amélioration des connaissances en matière d'emboîtement d'hypergraphe ;
2. discrétiser des trajectoires dans un espace porteur de sémantique et rendant possible l'application d'algorithmes de fouille.

Contribution à la théorie des hypergraphes

En présence de valeurs manquantes, nous avions montré qu'il est possible d'extraire des connaissances valides, sous forme de motifs ensemblistes [23, 50]. Par exemple, si un motif est fréquent, bien que des valeurs soient manquantes, il sera *a fortiori* fréquent dans toute complétion de la base de données. Cependant, le motif fréquent maximal sera plus court dans la base incomplète que dans toute base complète. L'étude de l'impact des valeurs manquantes met donc en œuvre deux ensembles de motifs (*i.e.* des hypergraphes), les hyperarêtes de l'un étant incluses dans les hyperarêtes de l'autre : ils constituent un emboîtement d'hypergraphes.

L'étude de cet emboîtement, en particulier la *mesure de son épaisseur*, est essentielle pour évaluer l'impact des valeurs manquantes. Cette mesure se voulant objective, relative à l'écart entre les bordures induit par la présence de valeurs manquantes, elle permet de qualifier proprement l'impact d'une méthode d'imputation, tâche centrale en gestion des valeurs manquantes.

D'autre part, nous disposons désormais d'un algorithme DEFME d'extraction en profondeur de motifs minimaux, à espace et délai polynomial (*cf.* chapitre 3). Outre la bonne maîtrise de la complexité algorithme que nous en avons, il se comporte en pratique nettement mieux que MTMINER dont nous avons réalisé l'étude en moyenne (*cf.* chapitre 4). De là, il conviendrait d'affiner nos analyses de complexité de l'extraction des traverses en prenant appui sur notre algorithme DEFME.

Enfin, le cadre formel d'ensembles minimisables que nous avons proposé permet d'obtenir des solutions *approchées* à ce problème fondamental de traverses minimales. Les applications de cette approximation concernent par conséquent le calcul de traverses minimales approchées ou de motifs émergents dans leur généralité.

Analyse de configuration spatiale dans le sport

Nous pensons que l'analyse des pratiques collectives sportives fait face à deux défis :

1. obtenir à faible coût des traces d'interaction, par exemple des trajectoires ;
2. discrétiser l'espace des trajectoires pour effectuer une analyse symbolique des comportements.

Concernant le premier défi, nous souhaiterions impliquer des industriels compétents en réalisation de terrains de sport ou de gymnases afin d'étudier l'opportunité d'insérer une matrice de capteurs RFID dans le sol. Nous pensons que les procédés de production de terrain synthétiques, intérieurs comme extérieurs, peuvent intégrer ces dispositifs, qui ne requièrent des cobayes que l'insertion d'un patch RFID dans une de leurs chaussures.

Il s'agit là d'une étape importante vers la démocratisation de l'accès à une technologie de captation de trajectoires, aujourd'hui plutôt réservée à l'élite des clubs de football professionnels et des équipes nationales. Pourtant, les organisateurs de manifestations sportives voient dans Big Data l'opportunité d'améliorer l'expérience du fan : entrer dans les détails de l'analyse du match implique fortement le spectateur. On peut supposer que les données de rencontres sportives seront toujours plus accessibles ; l'industrie a besoin de moyens génériques pour fournir ces données et donc de *terrains connectés*.

Nos tentatives dans ce sens ont donné lieu à quelques projets assez stériles. Nous avons notamment réalisé l'équipement vidéo d'un gymnase à l'aide de quatre caméras IP dont le flux était enregistré sur ordinateur. Faute d'investissement humain de la part de l'équipe pédagogique et par manque d'une solution acessible pour convertir les vidéos en trajectoires, cet équipement est actuellement sous-utilisé. Même si le matériel se démocratise pour le grand public [3], un projet de recherche ambitieux, faisant appel à des partenaires équipementiers du sport, en électronique et en *spatial analytics*, pourrait être pertinent.

Enfin, notre *graal* en matière d'analyse de configuration spatiale concernerait l'obtention de marqueurs « épigénétiques » du comportement d'une équipe particulière. Dans le sport collectif, on peut effectivement remarquer que les joueurs, au cours de l'entraînement, vont acquérir des routines collectives liées au *jouer ensemble*. Dans de nombreux sports comme le handball, des combinaisons orchestrées sont répétées à l'entraînement.

Cette traçabilité d'une équipe est-elle possible à qualifier ? Des réseaux récurrents de neurones répondraient par l'affirmative au vu des trajectoires, alors que nous souhaitons explorer une autre voie, vers la discrétisation de l'espace des trajectoires pour une sémantique des interactions (attaque, défense, bloc, diversion, *etc.*), en lien avec la *géomatique*. Une fois l'espace discrétisé, des techniques classiques d'apprentissage automatique sur des données symboliques pourraient obtenir un succès raisonnable.

3. voir par exemple `http://inout.tennis` pour un hawkeye personnel à 200€, localisant uniquement les rebonds, ou `https://www.stats.com/sportvu-basketball-media/` pour un système de six caméras qui équipe les arènes de NBA.

Bibliographie

[56] J. Allen et C. Perrault. Analyzing intention in utterances. *Artificial Intelligence*, 15(3) :143–178, 1980.

[57] C. M. Antunes et A. L. Oliveira. Kdd 2001 workshop on temporal data mining. Dans *Temporal data mining : An overview*, 2001.

[58] H. Arimura et T. Uno. Polynomial-delay and polynomial-space algorithms for mining closed sequences, graphs, and pictures in accessible set systems. Dans *SDM*, pages 1087–1098, 2009.

[59] J. W. Astington et J. Baird. *Why language matters for theory of mind*. Oxford University Press, New York, 2005.

[60] H. Aust, M. Oerder, F. Seide, et V. Steinbiss. The philips automatic train timetable information system. *Speech Communication*, 17(3-4) :249–262, 1995.

[61] D. Bahdanau, K. Cho, et Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[62] J. Bailey, T. Manoukian, et K. Ramamohanarao. A fast algorithm for computing hypergraph transversals and its application in mining emerging patterns. Dans *International Conference on Data Mining (ICDM'03)*, pages 485–489, 2003.

[63] F. A. A. Bashir et H.-L. Wei. Using nonlinear models to enhance prediction performance with incomplete data. Dans *Proceedings of the International Conference on Pattern Recognition Applications and Methods - Volume 1 : ICPRAM,*, pages 141–148. INSTICC, SciTePress, 2015.

[64] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, et L. Lakhal. Mining frequent patterns with counting inference. *SIGKDD Explor. Newsl.*, 2(2) :66–75, 2000.

[65] L. Ben Othman. *Conception et validation d'une méthode de complétion des valeurs manquantes fondée sur leurs modèles d'apparition*. Thèse, Université de Caen Basse-Normandie, 2011.

[66] Y. Bengio et Y. LeCun. Scaling learning algorithms towards ai. Dans L. Bottou, O. Chapelle, D. DeCoste, et J. Weston, editeurs, *Large-Scale Kernel Machines*. MIT Press, 2007.

[67] W. Boisseleau, O. Serban, et A. Pauchet. Building a narrative conversational agent using a component-based architecture. Dans *Proc. of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pages 1653–1654, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems.

[68] T. Boley, M. Horváth, A. Poigné, et S. Wrobel. *Efficient Closed Pattern Mining in Strongly Accessible Set Systems (Extended Abstract)*, pages 382–389. Springer Berlin Heidelberg, 2007.

[69] J.-F. Boulicaut, A. Bykowski, et C. Rigotti. Approximation of frequency queries by means of free-sets. Dans *Principles of Data Mining and Knowledge Discovery (PKDD'00), Lyon, France*, pages 75–85, 2000.

[70] J.-F. Boulicaut, A. Bykowski, et C. Rigotti. Free-sets : a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery journal*, pages 5–22, 2003. Kluwer Academics Publishers.

[71] B. Bringmann et A. Zimmermann. *Tree 2 – Decision Trees for Tree Structured Data*, pages 46–58. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[72] J. Broekens, M. Heerink, et H. Rosendal. Assistive social robots in elderly care : a review. *Gerontechnology*, 8(2), 2009.

[73] H. Bunt. The dit++ taxonomy for functional dialogue markup. Dans *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24, 2009.

[74] T. Calders et B. Goethals. Minimal k-free representations of frequent sets. Dans *Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'03), Cavtat-Dubrovnik, Croatia*, pages 71–82, 2003.

[75] T. Calders et B. Goethals. Quick inclusion-exclusion. Dans *Proceedings ECML-PKDD 2005 Workshop Knowledge Discovery in Inductive Databases*, volume 3933 of *LNCS*. Springer, 2005.

[76] A. Casali, R. Cicchetti, et L. Lakhal. Essential patterns : A perfect cover of frequent patterns. Dans *DaWaK*, pages 428–437, 2005.

[77] J. Cassell. Embodied conversational interface agents. *Commun. ACM*, 43(4) :70–78, Apr. 2000.

[78] E. Chanoni. Comment les mères racontent une histoire de fausses croyances à leur enfant de 3 à 5 ans ? *Enfance*, pages 181–189, 2009.

[79] V. Chaoji, A. Hoonlor, et B. Szymansk. Recursive data mining for role identification. Dans *CSTST*, pages 218–225, 2008.

[80] B. Crémilleux et J. Boulicaut. Simplest rules characterizing classes generated by delta-free sets. Dans Springer, editeur, *International Conference on Knowledge Based Systems and Applied Artificial Intelligence (Expert System), Cambridge, UK*, pages 33–46, 2002.

[81] P. B. de Byl. An online assistant for remote, distributed critiquing of electronically submitted assessment. *Educational Technology & Society*, 7 :29–41, 2004.

[82] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, et A. Acero. Recent advances in deep learning for speech research at microsoft. Dans *Acoustics, Speech and Signal Processing*, pages 8604–8608, 2013.

[83] G. Dong et J. Li. Efficient mining of emerging patterns : discovering trends and differences. Dans *Knowledge Discovery and Data Mining (KDD'99), San Diego, USA*, pages 43–52, 1999.

[84] G. Dong et J. Li. Mining border descriptions of emerging patterns from dataset pairs. *Knowledge and Information Systems*, 8 :178–202, 2005.

[85] N. Durand et M. Quafafou. Approximation of frequent itemset border by computing approximate minimal hypergraph transversals. Dans *Data Warehousing and Knowledge Discovery - 16th International Conference, DaWaK 2014, Munich, Germany, September 2-4, 2014. Proceedings*, pages 357–368, 2014.

[86] T. Eiter. Exact transversal hypergraphs and application to boolean μ-functions. *J. Symb. Comput.*, 17 :215–225, March 1994.

[87] P. Erdös et A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6 :290–297, 1959.

[88] P. Erdös et A. Rényi. On the evolution of random graphs. Dans *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, pages 17–61, 1960.

[89] M. Frampton et O. Lemon. Recent research advances in reinforcement learning in spoken dialogue systems. *Knowledge Engineering Review*, 24(04) :375–408, 2009.

[90] M. Fredman et L. Kachiyan. On the complexity of dualization of monotone disjunctive normal forms. *Journal of Algorithms*, 21(2) :618–628, 1996.

[91] A. Giacometti, D. H. Li, P. Marcel, et A. Soulet. 20 years of pattern mining : a bibliometric survey. *SIGKDD Explorations*, 15(1) :41–50, 2013.

[92] A. Giacometti, D. H. Li, et A. Soulet. *Balancing the Analysis of Frequent Patterns*, pages 53–64. Springer International Publishing, Cham, 2014.

[93] D. Griol, L. F. Hurtado, E. Segarra, et E. Sanchis. A statistical approach to spoken dialog systems design and evaluation. *Speech Commun.*, 50(8-9) :666–682, 2008.

[94] D. Griol, F. Torres, L. F. Hurtado, E. Sanchis, et E. Segarra. Different approaches to the dialogue management in the dihana project. Dans *10th Speech and Computer Conference(SPECOM)*, pages 203–206, Amsterdam, The Netherlands, 2005.

[95] M. Hagen. Lower bounds for three algorithms for the transversal hypergraph generation. Dans *Graph-Theoretic Concepts in Computer Science*, volume 4769 of *LNCS*, pages 316–327. Springer, 2007.

[96] M. Hagen. Remarks about the HBC-algorithms. Technical report, Private Conversation, 2007.

[97] M. Hagen. *Algorithmic and computational complexity issues of MONET*. Thèse, Firedrich-Schiller-Universität Jena, Germany, 2008.

[98] T. Hamrouni. Key roles of closed sets and minimal generators in concise representations of frequent patterns. *Intell. Data Anal.*, 16(4) :581–631, 2012.

[99] C. Hébert, A. Bretto, et B. Crémilleux. A data mining formalization to improve hypergraph transversal computation. *Fundamenta Informaticae*, 80(4) :415–433, 2007.

[100] C. Hébert et B. Crémilleux. Mining frequent *delta*-free patterns in large databases. Dans *Discovery Science*, pages 124–136, 2005.

[101] C. Hébert et B. Crémilleux. Optimized rule mining through a unified framework for interestingness measures. Dans *Data Warehousing and Knowledge Discovery, 8th International Conference, DaWaK 2006, Krakow, Poland, September 4-8, 2006, Proceedings.*, pages 238–247, 2006.

[102] C. Hébert et B. Crémilleux. A unified view of objective interestingness measures. Dans *5th International conference on Machine Learning and Data Mining (MLDM'07)*, pages 533–547, 2007.

[103] J. Hulstijn. Dialogue games are recipes for joint action. Dans *Proc. of Gotalog'00*, 2000.

[104] M. N. Jelassi. *Etude, représentation et applications des traverses minimales d'un hypergraphe. (Representation and applications of hypergraph minimal transversals)*. Thèse, Jean Monnet University, Saint-Étienne, France, 2014.

[105] D. J. Kavvadias et E. C. Stavropoulos. An efficient algorithm for the transversal hypergraph generation. *J. Graph Algorithms Appl.*, 9(2) :239–264, 2005.

[106] T. Largillier et S. Peyronnet. *Informatique Mathématique, une photographie en 2014*, chapter Algorithmique du web : autour du pagerank, pages 49–80. Le comptoir des presses d'universités, 2014.

[107] S. Larsson et D. Traum. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural language engineering*, 6(3&4) :323–340, 2000.

[108] J.-G. Lee, J. Han, et K.-Y. Whang. Trajectory clustering : a partition-and-group framework. Dans *Int. conf. on Management of data*, pages 593–604. ACM, 2007.

[109] A. Letois. Analyse stratégique de trajectoires dans du jeu vidéo compétitif. Master's thesis, Université de Caen Normandie, 2015.

[110] J. Liscombe, J. Bell Hirschberg, et J. J. Venditti. Detecting certainness in spoken tutorial dialogues. Dans *Proc. of Interspeech 2005*, 2005.

[111] D. Litman et K. Forbes-Riley. Spoken tutorial dialogue and the feeling of another's knowing. Dans *Proc. of the SIGDIAL 2009 Conference : The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, pages 286–289, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[112] R. Little et D. Rubin. *Statistical Analysis with Missing Data*. John Wiley, New York, 1987.

[113] G. Liu, J. Li, et L. Wong. A new concise representation of frequent itemsets using generators and a positive border. *Knowl. Inf. Syst.*, 17(1) :35–56, Oct. 2008.

[114] D. Lo, S.-C. Khoo, et L. Wong. Non-redundant sequential rules - theory and algorithm. *Inf. Syst.*, 34(4-5) :438–453, 2009.

[115] H. Mannila et H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3) :241–258, 1997.

[116] M. McTear. *Spoken dialogue technology : toward the conversational user interface*. Springer-Verlag New York Inc, 2004.

[117] T. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2) :203–226, 1980.

[118] K. Murakami et T. Uno. Efficient algorithms for dualizing large-scale hypergraphs. Dans *ALENEX*, pages 1–13, 2013.

[119] L. B. Othman et S. B. Yahia. Yet another approach for completing missing values. Dans *CLA*, pages 155–169, 2006.

[120] S. Ouaadrim. Complétion de valeurs manquantes par surfeur aléatoire. Master's thesis, Université de Caen Normandie, 2014.

[121] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336) :846–850, 1971.

[122] M. Schröder. The semaine api : Towards a standards-based framework for building emotion-oriented systems. *Adv. in Hum.-Comp. Int.*, 2010, Jan. 2010.

[123] J. Searle. *Speech Acts : An Essay in the Philosophy of Language*. Cambridge University, 1969.

[124] A. Soulet et B. Crémilleux. Adequate condensed representations of patterns. *Data Min. Knowl. Discov.*, 17(1) :94–110, 2008.

[125] P. Su, D. Vandyke, M. Gasic, D. Kim, N. Mrksic, T. Wen, et S. J. Young. Learning from real users : Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems. *CoRR*, 2015.

[126] L. Szathmary et A. Napoli. CORON : A Framework for Levelwise Itemset Mining Algorithms. Dans B. Ganter, R. Godin, et E. Mephu Nguifo, editeurs, *Third International Conference on Formal Concept Analysis - ICFCA '05*, pages 110–113, Lens/France, Feb. 2005.

[127] K. Takata. A worst-case analysis of the sequential method to list the minimal hitting sets of a hypergraph. *SIAM J. Discret. Math.*, 21 :936–946, December 2007.

[128] R. Taouil et Y. Bastide. Computing proper implications. Dans *9th International Conference on Conceptual Structures : Broadening the Base - ICCS'2001*, 2001.

[129] P. TCHÉBYCHEF. Des valeurs moyennes (traduction du russe, n. de khanikof.). *Journal de mathématiques pures et appliquées 2*, (2) :177–184, 1867.

[130] E. Ukkonen. Maximal and minimal representations of gapped and non-gapped motifs of a string. *Theoretical Computer Science*, 410(43) :4341–4349, 2009.

[131] S. Varges, S. Quarteroni, G. Riccardi, A. V. Ivanov, et P. Roberti. Leveraging pomdps trained with user simulations and rule-based dialogue management in a spoken dialogue system. Dans *Proc. of the 10th SIGDIAL Conference*, pages 156–159. The Association for Computer Linguistics, 2009.

[132] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, et G. E. Hinton. Grammar as a foreign language. *CoRR*, abs/1412.7449, 2014.

[133] J. Vreeken, M. van Leeuwen, et A. Siebes. Krimp : mining itemsets that compress. *Data Min. Knowl. Discov.*, 23(1) :169–214, 2011.

[134] N. Ward et D. Devault. Ten challenges in highly-interactive dialog system. *AAAI Spring Symposium Series*, 2015.

[135] G. M. Weiss. Timeweaver : a genetic algorithm for identifying predictive patterns in sequences of events. Dans *In Proc. of the Genetic and Evolutionary Computation Conference*, pages 718–725. Morgan Kaufmann, 1999.

[136] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, et Y. Bengio. Show, attend and tell : Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015.

[137] M. Zaki. Generating non-redundant association rules. Dans *ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, USA*, pages 34–43, 2000.

[138] M. J. Zaki, N. Lesh, et M. Ogihara. Planmine : Predicting plan failures using sequence mining. *Artif. Intell. Rev.*, 14(6) :421–446, 2000.