Notice des activités de recherche

Antoine Amarilli

Ce document présente une synthèse de mes travaux de recherche. Il couvre mes études de master, l'année de stages à l'étranger que j'ai effectué ensuite, ma thèse de doctorat [Ama16], et les travaux qui ont suivi.

Introduction

Le principal thème de ma recherche est la théorie de la gestion de données *incertaines*. J'ai travaillé sur ce thème pendant ma thèse, avec mon directeur Pierre Senellart, et au sein de collaborations internationales, commencées pendant mon année de stage à l'étranger (avant ma thèse) ou par la suite, et que je continue aujourd'hui : avec Tova Milo (Tel Aviv University) et Yael Amsterdamer (Bar Ilan University, Israel), sur l'interrogation de la foule [1] ; avec Michael Benedikt (University of Oxford), sur la réponse aux requêtes en monde ouvert ; avec Daniel Deutch (Tel Aviv University) sur l'incertitude sur des données ordonnées.

La recherche sur les données incertaines vise à étendre les techniques classiques de gestion de données développées pour les bases de données relationnelles [AHV95], en y ajoutant la possibilité de représenter différentes formes d'incertitude sur les données :

— les *valeurs manquantes*, par exemple les `NULL`s en SQL, qui servent à compléter des champs dont la valeur est inconnue ou non renseignée ;
— l'*incomplétude* dans les données, pour gérer des sources et des informations non exhaustives ;
— les *erreurs* dans les données, par exemple pour prendre en compte la possibilité que certains faits soient erronés ou incorrects.

La gestion de données incertaines est motivée par les nombreuses sources de données utilisées aujourd'hui, notamment sur le Web, qui sont toujours plus hétérogènes et moins fiables. On peut penser en particulier aux données extraites automatiquement à partir de pages Web par le traitement des langues naturelles [CBK+10] ; aux bases de connaissances structurées comme Wikidata réalisées collaborativement par de nombreux utilisateurs [VK14] ; aux données extraites en sondant des foules d'utilisateurs[1] [AGM+13 ; PGP+12] ; aux techniques incertaines d'intégration de données qui créent des appariements de schéma [2] probabilistes [DHY07] ; et plus généralement aux connaissances produites par la fouille de données et par l'apprentissage. Dans ces contextes, il est impératif de prendre en compte les erreurs et les données manquantes : on ne peut pas espérer corriger tous les problèmes manuellement, comme pour une base de données traditionnelle.

Je me suis intéressé à plusieurs problématiques théoriques autour de ce thème général, qui se sont intégrées à plusieurs communautés voisines : théorie des bases de données (ICDT), logique informatique (LICS), automates et algorithmes (ICALP), et intelligence artificielle et représentation des connaissances (IJCAI). Ma recherche trouve une unité autour des notions d'*incertitude* et de

1. En anglais, *crowdsourcing*
2. En anglais, *schema mappings*

structure : j'ai généralement cherché à garantir la décidabilité ou la tractabilité des tâches de gestion de données incertaines en imposant des conditions structurelles sur les données et les règles.

Ma recherche s'est notamment focalisée autour de quatre directions principales, présentées dans les quatre premières sections de ce document. Ces directions représentent autant de collaborations indépendantes, dont trois sont internationales :

1. J'ai travaillé sur la gestion de données relationnelles probabilistes, en imposant des *hypothèses de structure sur les instances* (formellement, des bornes sur la largeur d'arbre [3]), qui permettent d'assurer la tractabilité de l'évaluation probabiliste de requêtes et d'établir des liens avec les semi-anneaux de provenance.
2. J'ai étudié le *raisonnement en monde ouvert* sous des contraintes logiques, dans les contextes des logiques de description et des règles existentielles, et dans le contexte des bases de données sous l'hypothèse de finitude, pour démontrer la décidabilité de nouveaux langages de règles.
3. J'ai travaillé sur la *fouille de données à partir de la foule*[1], notamment sur l'identification des ensembles d'objets fréquents [4] et sur l'extrapolation de valeurs numériques manquantes suivant un ordre partiel.
4. J'ai proposé de nouvelles représentations pour la gestion de bases de données relationnelles avec un *ordre sur les faits*, notamment pour représenter l'incertitude sur cet ordre, et pour étudier le problème des réponses possibles et des réponses certaines.

Une cinquième section résume enfin d'autres travaux que j'ai entrepris sur des thèmes différents.

1 Provenance et probabilités sur les instances quasi-arborescentes [5]

Collaborateurs : Pierre Bourhis (CNRS CRIStAL), Pierre Senellart (Télécom ParisTech)

Publications :
— Un article [**ABS16**] à la conférence **PODS'16**
— Un article [**ABS15**] à la conférence **ICALP'15**
— Un article [**AS13**] à la conférence **BNCOD'13**

Les formalismes de bases de données relationnelles *probabilistes* [SOR+11] permettent de représenter de l'incertitude sur les faits d'une base de données, pour indiquer qu'ils ne sont pas nécessairement corrects. Le formalisme probabiliste le plus simple et le mieux compris est le modèle TID [6] : chaque fait de la base de données est annoté avec une probabilité d'apparition entre 0 et 1, et l'on suppose que les faits sont présents ou absents avec la probabilité indiquée et que ces événements sont tous indépendants.

Lorsqu'on évalue des requêtes sur une base de données TID, on souhaite obtenir les *réponses possibles* pour la requête, annotées par leur *probabilité* d'être vraie d'après la distribution de probabilité décrite par l'instance. La recherche sur les bases de données probabilistes a vite déterminé que cette tâche d'évaluation de requêtes était difficile, même lorsque la requête est fixée, et même dans le langage simple des *requêtes conjonctives*, dont l'évaluation est pourtant facile et hautement parallélisable sur les bases de données classiques. Cette direction de recherche a abouti à un résultat de dichotomie [DS12] pour caractériser les requêtes dont l'évaluation est facile, et a montré que, pour toute autre requête, l'évaluation probabiliste sur des instances TID arbitraires est infaisable en fonction de l'instance d'entrée (à savoir, #P-difficile).

3. En anglais, *bounded treewidth*
4. En anglais, *frequent itemsets*
5. En anglais, *treelike*
6. Pour *tuple-independent databases*

Plutôt que d'étudier quelles requêtes sont évaluables sur n'importe quelle instance d'entrée, nous avons voulu déterminer les classes d'*instances* probabilistes pour lesquelles l'évaluation de requête est toujours faisable efficacement. L'objectif de cette recherche est d'exploiter la *structure* des bases de données utilisées en pratique, qui ne sont pas arbitraires, pour montrer que l'évaluation probabiliste de requêtes peut parfois être faisable pour des requêtes riches si les instances d'entrée ont une certaine forme. Mon étude est également motivée par notre travail antérieur [**AS13**], qui avait construit des liens entre bases de données probabilistes et documents XML probabilistes [KS13], en conjonction avec des résultats préexistants [CKS09] qui montraient la faisabilité de l'évaluation probabiliste de requêtes pour XML. Pour généraliser cela, nous avons ainsi cherché à *déterminer les familles d'instances relationnelles pour lesquelles l'évaluation probabiliste de requêtes est toujours faisable.*

Notre premier travail [**ABS15**] a identifié une classe d'instances qui satisfaisait cette condition : celles dont la *largeur d'arbre*[7] est bornée par une constante, ce qui revient à imposer, en un sens précis, qu'il s'agit presque d'arbres. Cette condition avait déjà été étudiée par Courcelle [Cou90], hors du cadre probabiliste, pour montrer que l'évaluation de requêtes devenait faisable sur de telles instances, même pour le langage expressif de la *logique monadique du second ordre*. Notre travail étend ce résultat à l'évaluation *probabiliste* de telles requêtes, et montre que cette tâche est toujours tractable sur de telles instances : elle peut même être effectuée en complexité *linéaire* en les données, au coût des opérations arithmétiques près. En fait, on peut même montrer, au-delà du formalisme TID et de son hypothèse d'indépendance, que ce résultat s'étend à de nombreux formalismes probabilistes existants, même avec des corrélations, du moment que celles-ci sont également de largeur d'arbre bornée en un certain sens.

Au-delà de l'évaluation probabiliste, ces résultats sont obtenus à travers le formalisme plus général de la *provenance* [BKT01 ; CCT09] pour les bases de données, une construction abstraite qui permet d'indiquer comment les résultats d'une requête dépendent de l'instance sur laquelle on l'évalue. Notre résultat consiste en fait à démontrer que l'on peut calculer efficacement, sur les instances de largeur d'arbre bornée, des représentations de la provenance des requêtes, sous la forme de *circuits de provenance* [DMR+14], que l'on peut ensuite utiliser pour le calcul probabiliste. Nos constructions étendent ainsi les techniques de provenance expressive à base de semi-anneaux [GKT07] au-delà des langages habituels de requête sur les bases de données, en les appliquant par exemple aux automates d'arbre.

Dans le contexte des instances d'arité 2 (c'est-à-dire des graphes étiquetés), nous avons ensuite démontré [**ABS16**] que la largeur d'arbre est *la* bonne mesure de la complexité des instances pour l'évaluation probabiliste de requêtes. Plus spécifiquement, nous avons démontré que cette tâche est infaisable sur *toute* famille d'instances qui n'est pas de largeur d'arbre bornée, ou inconstructible en un certain sens technique, quelles que soient les autres restrictions que l'on impose sur les instances d'entrée. En d'autres termes, toute condition sur les instances qui garantit la faisabilité de l'évaluation probabiliste revient à borner leur largeur d'arbre, ou entraîne leur inconstructibilité. Notre résultat utilise des techniques récentes pour l'extraction de mineurs dans des graphes [CC14] suivant le résultat de Robertson et Seymour [RS86], et nous permet d'améliorer des bornes inférieures préexistantes [KT10 ; GHL+14] sur les formulations non-probabilistes de ces problèmes. Notre résultat démontre l'infaisabilité de l'évaluation de requêtes dans le langage de la logique du premier ordre, mais nous étudions également des requêtes plus simples (des unions de requêtes conjonctives avec inégalités), en montrant qu'elles ne peuvent pas être efficacement évaluées par les méthodes probabilistes à base de diagrammes de décision binaires ordonnés[8].

Nos résultats concluent ainsi à une dichotomie sur l'évaluation probabiliste de requêtes du point de vue de l'instance, à la manière de la dichotomie déjà connue pour les requêtes [DS12] : l'évaluation

7. En anglais, *treewidth*
8. En anglais, *ordered binary decision diagram* (OBDD)

probabiliste de requêtes sur une famille d'instances est faisable si leur largeur d'arbre est bornée, même dans des formalismes probabilistes riches et pour des langages de requêtes expressifs ; mais elle est infaisable sur *toute* famille d'instances de largeur d'arbre non bornée, sous des conditions naturelles sur la constructibilité et la signature.

2 Raisonnement en monde ouvert

Collaborateurs : Michael Benedikt, Michael Vanden Boom (University of Oxford), Pierre Bourhis

Publications :
— Un article [**ABB+16**] à la conférence **IJCAI'16**
— Un article [**AB15a**] à la conférence **IJCAI'15**
— Un article [**AB15b**] à la conférence **LICS'15**

Le problème de *réponse aux requêtes en monde ouvert*[9] (QA) considère une instance I, des contraintes logiques Σ, et une requête q : il consiste à calculer les *réponses certaines* à q sous I et Σ, c'est-à-dire, les réponses qui sont vraies sur *toutes* les complétions possibles des faits de l'instance I qui satisfont les contraintes Σ. Le problème QA est fondamental pour l'étude des bases de données *incomplètes* : l'instance I est un ensemble *incomplet* de faits, et Σ décrit des règles logiques que l'on veut imposer sur les faits inconnus. Le problème QA permet ainsi de compléter l'instance I par ses conséquences certaines d'après les contraintes logiques, et d'évaluer des requêtes sur le résultat.

Le problème QA a donc été étudié par des communautés multiples pour trouver des langages de contraintes aussi expressifs que possible pour Σ tels que QA soit décidable (et si possible faisable) :

Bases de données relationnelles. Le problème QA a été étudié à l'origine pour les *bases de données relationnelles* sous une formulation équivalente en termes d'*inclusion de requêtes*[10] [JK84] sous les contraintes d'intégrité classiques en bases de données [AHV95].

Logiques de description. La communauté des *logiques de description* (DL) s'est spécifiquement intéressée au développement de langages expressifs pour Σ, et au compromis entre expressivité et complexité. Malgré leur expressivité, les langages DL imposent toutefois que l'*arité* des faits de l'instance soit au plus 2, c'est-à-dire que les relations entre objets ne connectent que deux objets au plus, au contraire des bases de données relationnelles classiques.

Règles existentielles. Les *règles existentielles* permettent d'exprimer des contraintes logiques décidables sans imposer que l'arité soit 2 [BLM10], mais en contrepartie elles ne peuvent pas exprimer d'autres opérateurs logiques comme la disjonction ou la négation.

En étudiant le problème QA, j'ai d'abord cherché à établir des connexions entre les approches étudiées par ces communautés, et plus spécifiquement celle des logiques de description et des règles existentielles, afin d'obtenir de nouveaux résultats de décidabilité pour des langages expressifs. Notre travail [**AB15a**] étudie ainsi la décidabilité de règles logiques qui combinent deux formalismes : les règles existentielles, en arité arbitraire, et des DL capturées par un fragment logique expressif [Pra09] en arité 2. L'objectif de ce travail est d'obtenir le meilleur des deux mondes : pouvoir exprimer à la fois des contraintes expressives sur les données en arité 2, et de contraintes moins expressives sur les données relationnelles en général.

Nous montrons qu'une combinaison naïve de ces deux formalismes aboutit à un langage pour lequel le problème QA n'est plus décidable. L'indécidabilité provient surtout des *assertions de fonctionnalité* que les DL peuvent exprimer, par exemple, « une personne n'est née qu'à un seul endroit ». Nous avons montré comment rétablir la décidabilité en imposant des conditions sur les règles existentielles : restreindre au fragment *frontier-1* [BLM+09], et interdire certains motifs cycliques en un certain

9. En anglais, *open-world query answering*
10. En anglais, *query containment*

sens technique. Nous montrons que QA est décidable pour ce langage restreint combiné avec des contraintes expressives en arité 2, et que l'on peut même l'enrichir avec des *dépendances fonctionnelles* (qui généralisent les assertions de fonctionnalité à l'arité arbitraire), si on impose une condition d'*absence de conflits*[11] [CGP12]. Nos résultats sont ainsi une première étape pour raisonner sur des données incomplètes avec des contraintes logiques combinant des opérateurs logiques arbitraires, des dépendances fonctionnelles, et des contraintes en haute arité.

Ma seconde direction de recherche autour du problème QA se place dans le contexte des bases de données, pour des contraintes incluant des dépendances fonctionnelles ainsi que des *dépendances d'inclusion unaires*, un cas particulier de règles du fragment frontier-1. Notre travail [**AB15b**] étudie plus spécifiquement l'impact de l'hypothèse de *finitude* qui est couramment faite en bases de données : on suppose que l'ensemble des faits manquants est nécessairement fini. Cette hypothèse modifie la définition du problème QA (une réponse à la requête peut être certaine sur les complétions finies mais ne pas l'être sur les complétions infinies), mais elle limite surtout les outils auxquels on peut avoir recours. En effet, la plupart des techniques connues pour QA, notamment la construction de la *poursuite*[12], font appel à des ensembles de faits infinis. La finitude est pourtant une hypothèse naturelle que l'on souhaiterait pouvoir imposer sur les faits manquants lors du raisonnement.

Mon travail a ainsi montré comment réduire le problème QA sous l'hypothèse de finitude au problème QA sans cette hypothèse, pour le langage de contraintes que nous étudions, en utilisant une construction de clôture finie [CKV90] pour compléter les contraintes par leurs conséquences selon la finitude. La preuve de ce résultat s'obtient par une construction complexe de modèles finis universels pour ces contraintes, en utilisant plusieurs outils développés récemment pour l'étude du problème QA sous l'hypothèse de finitude [BGO10 ; Ros11 ; ILS14]. Mes résultats fournissent ainsi une première compréhension de l'impact de la finitude pour le problème QA, dans le cas de langages en arité arbitraire intégrant des dépendances fonctionnelles.

J'ai travaillé dans un troisième temps sur le problème QA dans le contexte des règles existentielles et plus généralement des logiques gardées [ANB98 ; BCS11], en étudiant l'impact sur la décidabilité que pouvait avoir l'ajout de relations interprétées de manière spéciale, à savoir, des relations transitives, des clôtures transitives, et des relations d'ordre total. Notre travail [**ABB+16**] démontre que le problème QA peut rester décidable dans ce contexte sous certaines hypothèses : les relations spéciales ne sont jamais utilisées comme gardes et (pour les relations d'ordre) sont *couvertes* (dans les règles et dans la requête) par des relations normales.

3 Foules d'utilisateurs et fouille de données

Collaborateurs : Yael Amsterdamer (Bar Ilan University, Israel), Tova Milo (Tel Aviv University), Pierre Senellart

Publications :
- Un article [**AAM14a**] à la conférence **ICDT'14**
- Un article [**AAM14b**] au workshop **UnCrowd'14**
- Un article [**AAM+16**] en soumission

Cet axe de ma recherche étudie l'interrogation d'une foule d'utilisateurs[1], et plus précisément à l'*extraction de données* à partir de la foule. Il s'intègre au projet ERC MODAS de Tova Milo [DM12], et fait suite à leur travail [AGM+13] qui applique des techniques de *fouille de données* à la foule, pour apprendre des *règles d'association*[13]. La vision est celle d'un système qui pose des questions

11. En anglais, *non-conflicting condition*
12. En anglais, *chase*
13. En anglais, *association rules*

à la foule pour constituer ou compléter des bases de connaissances structurées, en regroupant des informations connues des humains mais auparavant inaccessibles aux machines.

L'étude de la foule pose bien entendu des questions de gestion de l'incertitude. En effet, les réponses fournies par la foule ne sont pas toujours *correctes* [WP10], et sont souvent entachées d'erreurs, qui peuvent être involontaires ou délibérées. Par ailleurs, les données que l'on peut se permettre d'extraire sont nécessairement *incomplètes* par rapport à toutes celles qu'on aurait voulu obtenir, car chaque question posée à la foule entraîne de la latence et un coût financier (pour rétribuer les utilisateurs). Un algorithme d'acquisition de données sur la foule se mesure donc suivant deux dimensions : le *nombre de requêtes* qu'il pose aux utilisateurs, et la *complexité computationnelle* au sens classique pour le choix des questions à poser.

Notre travail s'est d'abord intéressé à l'extraction des ensembles d'objets fréquents[4], une tâche classique de la fouille de données, que nous réexaminons en supposant l'existence d'une taxonomie sur les objets à la manière de [SA95], et que nous généralisons pour l'extraction à partir de la foule. La taxonomie induit alors une structure de treillis distributif sur les ensembles d'objets, pour laquelle la recherche des ensembles fréquents revient à déterminer un prédicat booléen monotone par des requêtes sur la foule. Notre premier article [**AAM14a**] étudie la complexité de ce problème.

Nous avons ensuite cherché à généraliser ces questions pour apprendre des valeurs *numériques* à partir de la foule, par exemple la valeur de support pour chaque ensemble d'objets, en supposant que les valeurs sont monotones par rapport à une structure sous-jacente (par exemple la taxonomie). Nous avons esquissé des approches pour ce problème dans un cadre général [**AAM14b**] pour nous concentrer ensuite sur le problème plus précis des valeurs numériques inconnues contraintes par un ordre partiel [**AAM+16**]. Ce problème se pose par exemple pour catégoriser un produit dans une taxonomie de catégories : l'objet a un score de compatibilité inconnu avec chaque catégorie, et on sait que l'objet est toujours plus compatible avec une catégorie qu'avec une sous-catégorie plus spécifique. Une fois que la foule nous a fourni le score de compatibilité d'un produit avec certaines catégories, notre travail vise à compléter les valeurs qui nous manquent, sans effectuer davantage de requêtes : on cherche notamment à identifier les meilleures catégories où ranger le produit.

Notre travail généralise ainsi l'interpolation linéaire sur un ordre total et l'étend aux ordres partiels, un problème général qui n'avait pas encore été étudié. Nous formalisons cela comme le calcul de l'espérance de variables aléatoires correspondant aux valeurs inconnues, dans le polytope des contraintes linéaires imposées par l'ordre partiel, suivant la distribution uniforme. Nous concevons un algorithme pour calculer l'espérance en complexité $\mathrm{FP}^{\#\mathrm{P}}$, et montrons que ce problème est #P-difficile, à l'aide de résultats de théorie des ordres partiels [BW91]. Nous présentons ensuite deux manières d'assurer la faisabilité : le calcul approché, avec l'échantillonnage de polytopes [KLS97], et un algorithme dynamique exact en temps polynomial pour les ordres partiels arborescents.

4 Gestion de données ordonnées incertaines

Collaborateurs : M. Lamine Ba (QCRI), Daniel Deutch (Tel Aviv University), Pierre Senellart

Publication : — Un article [**ABD+16**] en soumission

Ce travail étudie la gestion de l'incertitude sur des données relationnelles *ordonnées* avec un ordre sur les tuples, pour le langage de requêtes classique de l'algèbre relationnelle.

Ce contexte est intéressant car la gestion de données relationnelles ordonnées, même sans incertitude, peut faire *apparaître* de l'incertitude sur le résultat des opérations. Par exemple, si l'on considère deux listes ordonnées de résultats issus de la même source et qu'on souhaite les combiner en calculant l'union de ces deux listes, plusieurs ordres sont possibles sur le résultat, car on ne sait pas comment les tuples de la première liste doivent être ordonnés par rapport à ceux de la seconde. Nous avons

généralisé cette observation pour développer une sémantique multiensembliste [14] pour les opérateurs de l'algèbre relationnelle positive, ainsi qu'un système de représentation des relations ordonnées incertaines s'appuyant sur les ordres partiels, à la façon de [GM99]. Nous avons ensuite étendu ce formalisme avec un opérateur d'accumulation dépendant de l'ordre, pour évaluer des requêtes qui examinent l'ordre sur les données. Les applications que nous envisageons pour ce travail incluent par exemple la recherche de fondations formelles pour les méthodes d'agrégation de classements [15] [DKN+01] utilisées pour réconcilier des listes de résultats, ainsi que pour les méthodes de gestion des préférences [JKS14].

Notre travail étudie la complexité, dans ce contexte, de la recherche des résultats possibles et des résultats certains [AKO07], deux problèmes classiques sur les données incertaines. Nous avons montré que ces problèmes sont infaisables dans le cas général, mais qu'on peut efficacement calculer les résultats certains si on n'utilise pas l'opérateur d'accumulation, ou si on effectue l'accumulation dans un *monoïde cancellatif*. Nous avons ensuite montré comment rendre ces tâches plus faciles en limitant la *structure* des relations ordonnées fournies en entrée. En effet, on peut s'attendre que les données à intégrer soient souvent totalement ordonnées, ou totalement dépourvues d'ordre. Si l'on interdit certains opérateurs, nous avons ainsi montré que le calcul des résultats possibles et certains était faisable, à requête fixée, sur des relations ainsi restreintes, ou plus généralement quand l'ordre partiel sur les données d'entrée est bornée en termes de *largeur* [16] [BLS87], ou en termes d'*ia-largeur*, un paramètre que nous introduisons.

5 Autres travaux

Complétude dans les bases de connaissances. J'ai collaboré avec Luis Galárraga et Fabian M. Suchanek (Télécom ParisTech) sur le problème d'estimer la complétude des informations répertoriées dans les bases de connaissances (Wikidata et YAGO). Cette tâche est importante pour déterminer, par exemple, si les résultats d'une requête sont exhaustifs ou non, mais elle est difficile à cause de l'hypothèse du monde ouvert : les faits manquants peuvent être incorrects ou simplement omis. Nous avons travaillé à appliquer l'approche AMIE [GTH+13] d'extraction automatique de règles, pour apprendre comment estimer la complétude. Ce travail est actuellement en soumission [**GRA+16**].

Possibilité pour le XML probabiliste. J'ai travaillé sur les représentations probabilistes pour les documents XML [KS13], qui décrivent des documents arborescents structurés au contenu incertain. J'ai étudié le problème de déterminer si un document donné est un monde possible d'un document probabiliste, en calculant sa masse dans la distribution de probabilité. J'ai identifié quelles variantes de ce problème peuvent être résolues efficacement, et démontré l'infaisabilité des autres.

Ce travail a été publié [**Ama14**] au workshop **AMW'2014** puis dans la revue *ISI* [**Ama15**].

Extraction d'entités sur le Web à partir d'identifiants uniques. J'ai travaillé avec Fabian M. Suchanek (Télécom ParisTech) sur un travail par Aliaksandr Talaika et Joanna Biega (Max Planck Institute for Informatics) sur l'extraction de données structurées à partir du Web, en recherchant des identifiants avec une structure fixe (ISBN, GTIN, DOI, etc.). Cette technique simple mais efficace peut extraire des millions d'entités uniques à partir d'un corpus de pages Web.

Ce travail a été présenté [**TBA+15**] au workshop **WebDB'15** de la conférence SIGMOD.

Alignement holistique de bases de connaissances. J'ai travaillé pendant mon stage de master sur les techniques d'alignement de bases de connaissances [ES07], qui identifient des liens entre des

14. En anglais, *bag semantics*
15. En anglais, *rank aggregation*
16. En anglais, *width*

sources de données de manière automatique, pour étendre un travail préexistant [SAS11].

Mes travaux ont été résumés dans un papier invité sur YAGO [**AGP+14**] et leur application pour l'exploration du Web caché a été présentée au workshop **VLDS'12** de VLDB [**OAS12**].

Politiques de tarification [17] pour documents XML. J'ai travaillé avec Tang Ruiming et Stéphane Bressan (National University of Singapore) et Pierre Senellart (Télécom ParisTech) sur la tarification des documents XML. Nous avons étudié la question de fixer un prix qui permette à l'utilisateur d'acheter un échantillon d'un document XML qui l'intéresse.

Ce travail [**TAS+14**] a été présenté à **DEXA'14** et publié dans *TLKDS* [**TAS+16**].

Transducteurs pondérés à états finis pour la reconnaissance vocale. Mon travail de master à Google New York avec Cyril Allauzen et Mehryar Mohri a donné lieu [**AAM15**] à un **brevet** sur le décodage pour la reconnaissance vocale à base de méthodes de minimisation du risque de Bayes [18].

Auto-Références

[AAM+16] Antoine Amarilli, Yael Amsterdamer, Tova Milo et Pierre Senellart. "Top-k Queries on Unknown Values under Order Constraints". Preprint : `https://a3nm.net/publications/amarilli2016top.pdf`. 2016.

[AAM14a] Antoine Amarilli, Yael Amsterdamer et Tova Milo. "On the Complexity of Mining Itemsets from the Crowd Using Taxonomies". In : *Proc. ICDT*. 2014, p. 15–25. URL : `https://arxiv.org/abs/1312.3248`.

[AAM14b] Antoine Amarilli, Yael Amsterdamer et Tova Milo. "Uncertainty in Crowd Data Sourcing Under Structural Constraints". In : *Proc. UnCrowd*. T. 8505. LNCS. Springer Berlin Heidelberg, 2014, p. 351–359. URL : `https://arxiv.org/abs/1403.0783`.

[AAM15] Antoine Amarilli, Cyril Allauzen et Mehryar Mohri. *Minimum Bayesian Risk Methods for Automatic Speech Recognition*. United States Patent 9123333. 2015. URL : `https://a3nm.net/publications/amarilli2014minimum.pdf`.

[AB15a] Antoine Amarilli et Michael Benedikt. "Combining Existential Rules and Description Logics". In : *Proc. IJCAI*. AAAI Press, 2015, p. 2691–2697. URL : `https://arxiv.org/abs/1505.00326`.

[AB15b] Antoine Amarilli et Michael Benedikt. "Finite Open-World Query Answering with Number Restrictions". In : *Proc. LICS*. 2015, p. 305–316. URL : `https://arxiv.org/abs/1505.04216`.

[ABB+16] Antoine Amarilli, Michael Benedikt, Pierre Bourhis et Michael Vanden Boom. "Query Answering with Transitive and Linear-Ordered Data". In : *Proc. IJCAI*. To appear. 2016.

[ABD+16] Antoine Amarilli, Lamine M. Ba, Daniel Deutch et Pierre Senellart. "Possible and Certain Answers for Queries over Order-Incomplete Data". Preprint : `https://a3nm.net/publications/amarilli2016possible.pdf`. 2016.

[ABS15] Antoine Amarilli, Pierre Bourhis et Pierre Senellart. "Provenance Circuits for Trees and Treelike Instances". In : *Proc. ICALP*. T. 9135. LNCS. Springer Berlin Heidelberg, 2015, p. 56–68. URL : `https://arxiv.org/abs/1511.08723`.

[ABS16] Antoine Amarilli, Pierre Bourhis et Pierre Senellart. "Tractable Lineages on Treelike Instances : Limits and Extensions". In : *Proc. PODS*. To appear. 2016. URL : `https://a3nm.net/publications/amarilli2016tractable.pdf`.

[AGP+14] Antoine Amarilli, Luis Galárraga, Nicoleta Preda et Fabian M. Suchanek. "Recent Topics of Research around the YAGO Knowledge Base". In : *Proc. APWEB*. T. 8709. LNCS. Springer International Publishing, 2014, p. 1–12. URL : `https://zenodo.org/record/34912`.

[Ama14] Antoine Amarilli. "The Possibility Problem for Probabilistic XML". In : *Proc. AMW*. 2014. URL : `http://ceur-ws.org/Vol-1189/paper_2.pdf`.

[Ama15] Antoine Amarilli. "Possibility for Probabilistic XML". In : *Ingénierie des Systèmes d'Information* 20.5 (2015), p. 53–75. URL : `https://arxiv.org/abs/1404.3131`.

17. En anglais, *data pricing*
18. En anglais, *Minimum Bayes Risk decoding*

[Ama16] Antoine Amarilli. “Leveraging the Structure of Uncertain Data”. 2016-ENST-0021. Thèse de doct. Télécom ParisTech, 2016.

[AS13] Antoine Amarilli et Pierre Senellart. “On the Connections between Relational and XML Probabilistic Data Models”. In : *Proc. BNCOD*. 2013, p. 121–134. url : http://pierre.senellart.com/publications/amarilli2013connections.pdf.

[GRA+16] Luis Galárraga, Simon Razniewski, Antoine Amarilli et Fabian M. Suchanek. “Predicting Completeness in Knowledge Bases”. Preprint : https://a3nm.net/publications/galarraga2016predicting.pdf. 2016.

[OAS12] Marilena Oita, Antoine Amarilli et Pierre Senellart. “Cross-Fertilizing Deep Web Analysis and Ontology Enrichment”. In : *Proc. VLDS*. 2012, p. 5–8. url : http://pierre.senellart.com/publications/oita2012crossfertilizing.pdf.

[TAS+14] Ruiming Tang, Antoine Amarilli, Pierre Senellart et Stéphane Bressan. “Get a Sample for a Discount”. In : *Proc. DEXA*. T. 8644. LNCS. Springer International Publishing, 2014, p. 20–34. url : https://a3nm.net/publications/tang2014get.pdf.

[TAS+16] Ruiming Tang, Antoine Amarilli, Pierre Senellart et Stéphane Bressan. “A Framework for Sampling-Based XML Data Pricing”. In : *Transactions on Large-Scale Data and Knowledge-Centered Systems* 24 (2016), p. 116–138. url : https://a3nm.net/publications/tang2014framework.pdf.

[TBA+15] Aliaksandr Talaika, Joanna Biega, Antoine Amarilli et Fabian M. Suchanek. “IBEX : Harvesting Entities from the Web Using Unique Identifiers”. In : *Proc. WebDB*. ACM, 2015, p. 13–19. url : https://arxiv.org/abs/1505.00841.

Références

[AGM+13] Yael Amsterdamer, Yael Grossman, Tova Milo et Pierre Senellart. “Crowd Mining”. In : *Proc. SIGMOD*. 2013.

[AHV95] Serge Abiteboul, Richard Hull et Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[AKO07] Lyublena Antova, Christoph Koch et Dan Olteanu. “World-Set Decompositions : Expressiveness and Efficient Algorithms”. In : *Proc. ICDT*. 2007.

[ANB98] Hajnal Andréka, István Németi et Johan van Benthem. “Modal Languages and Bounded Fragments of Predicate Logic”. In : *J. Philosophical Logic* 27.3 (1998).

[BCS11] Vince Bárány, Balder ten Cate et Luc Segoufin. “Guarded Negation”. In : *ICALP*. 2011.

[BGO10] Vince Bárány, Georg Gottlob et Martin Otto. “Querying the Guarded Fragment”. In : *Proc. LICS*. 2010.

[BKT01] Peter Buneman, Sanjeev Khanna et Wang-Chiew Tan. “Why and Where : A Characterization of Data Provenance”. In : *Proc. ICDT*. 2001.

[BLM+09] Jean-François Baget, Michel Leclère, Marie-Laure Mugnier et Eric Salvat. “Extending Decidable Cases for Rules with Existential Variables”. In : *Proc. IJCAI*. 2009.

[BLM10] Jean-François Baget, Michel Leclère et Marie-Laure Mugnier. “Walking the Decidability Line for Rules with Existential Variables”. In : *Proc. KR*. 2010.

[BLS87] Andeas Brandstädt, Van Bang Le et Jeremy P. Spinrad. “Posets”. In : *Graph Classes. A Survey*. 1987. Chap. 6.

[BW91] Graham Brightwell et Peter Winkler. “Counting Linear Extensions”. In : *Order* 8.3 (1991).

[CBK+10] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr. et Tom M. Mitchell. “Toward an Architecture for Never-Ending Language Learning”. In : *Proc. AAAI*. 2010.

[CC14] Chandra Chekuri et Julia Chuzhoy. “Polynomial Bounds for the Grid-Minor Theorem”. In : *Proc. STOC*. 2014.

[CCT09] James Cheney, Laura Chiticariu et Wang-Chiew Tan. “Provenance in Databases : Why, How, and Where”. In : *Foundations and Trends in Databases* 1.4 (2009).

[CGP12] Andrea Calì, Georg Gottlob et Andreas Pieris. “Towards More Expressive Ontology Languages : The Query Answering Problem”. In : *Artif. Intel.* 193 (2012).

[CKS09] Sara Cohen, Benny Kimelfeld et Yehoshua Sagiv. "Running Tree Automata on Probabilistic XML". In : *Proc. PODS*. 2009.

[CKV90] Stavros S. Cosmadakis, Paris C. Kanellakis et Moshe Y. Vardi. "Polynomial-Time Implication Problems for Unary Inclusion Dependencies". In : *J. ACM* 37.1 (1990).

[Cou90] Bruno Courcelle. "The Monadic Second-Order Logic of Graphs. I. Recognizable Sets of Finite Graphs". In : *Inf. Comput.* 85.1 (1990).

[DHY07] Xin Dong, Alon Y. Halevy et Cong Yu. "Data Integration with Uncertainty". In : *Proc. VLDB*. 2007.

[DKN+01] Cynthia Dwork, Ravi Kumar, Moni Naor et Dandapani Sivakumar. "Rank Aggregation Methods for the Web". In : *Proc. WWW*. 2001.

[DM12] Daniel Deutch et Tova Milo. "Mob data sourcing". In : *Proc. SIGMOD*. 2012.

[DMR+14] Daniel Deutch, Tova Milo, Sudeepa Roy et Val Tannen. "Circuits for Datalog Provenance." In : *Proc. ICDT*. 2014.

[DS12] Nilesh Dalvi et Dan Suciu. "The dichotomy of probabilistic inference for unions of conjunctive queries". In : *J. ACM* 59.6 (2012).

[ES07] Jérôme Euzenat et Pavel Shvaiko. *Ontology Matching*. Springer, 2007.

[GHL+14] Robert Ganian, Petr Hliněný, Alexander Langer, Jan Obdržálek, Peter Rossmanith et Somnath Sikdar. "Lower Bounds on the Complexity of MSO_1 Model-Checking". In : *JCSS* 1.80 (2014).

[GKT07] Todd J. Green, Grigoris Karvounarakis et Val Tannen. "Provenance Semirings". In : *Proc. PODS*. 2007.

[GM99] Stéphane Grumbach et Tova Milo. "An Algebra for Pomsets". In : *Inf. Comput.* 150.2 (1999).

[GTH+13] Luis Antonio Galárraga, Christina Teflioudi, Katja Hose et Fabian Suchanek. "AMIE : Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases". In : *Proc. WWW*. 2013.

[ILS14] Yazmín Ibáñez-García, Carsten Lutz et Thomas Schneider. "Finite Model Reasoning in Horn Description Logics". In : *Proc. KR*. 2014.

[JK84] David S. Johnson et Anthony C. Klug. "Testing Containment of Conjunctive Queries under Functional and Inclusion Dependencies". In : *JCSS* 28.1 (1984).

[JKS14] Marie Jacob, Benny Kimelfeld et Julia Stoyanovich. "A System for Management and Analysis of Preference Data". In : *PVLDB* 7.12 (2014).

[KLS97] Ravi Kannan, László Lovász et Miklós Simonovits. "Random Walks and an $O^*(n^5)$ Volume Algorithm for Convex Bodies". In : *Random Struct. Algorithms* 11.1 (1997).

[KS13] Benny Kimelfeld et Pierre Senellart. "Probabilistic XML: Models and Complexity". In : *Advances in Probabilistic Databases for Uncertain Information Management*. 2013.

[KT10] Stephan Kreutzer et Siamak Tazari. "Lower Bounds for the Complexity of Monadic Second-Order Logic". In : *Proc. LICS*. 2010.

[PGP+12] Aditya Parameswaran, Hector Garcia-Molina, Hyunjung Park, Neoklis Polyzotis, Aditya Ramesh et Jennifer Widom. "Crowdscreen : Algorithms for Filtering Data with Humans". In : *Proc. SIGMOD*. 2012.

[Pra09] Ian Pratt-Hartmann. "Data-Complexity of the Two-Variable Fragment with Counting Quantifiers". In : *Inf. Comput.* 207.8 (2009).

[Ros11] Riccardo Rosati. "On the Finite Controllability of Conjunctive Query Answering in Databases under Open-World Assumption". In : *JCSS* 77.3 (2011).

[RS86] Neil Robertson et Paul D. Seymour. "Graph minors. V. Excluding a Planar Graph". In : *J. Comb. Theory, Ser. B* 41.1 (1986).

[SA95] Ramakrishnan Srikant et Rakesh Agrawal. "Mining Generalized Association Rules". In : *VLDB*. 1995.

[SAS11] Fabian M Suchanek, Serge Abiteboul et Pierre Senellart. "PARIS : Probabilistic Alignment of Relations, Instances, and Schema". In : *PVLDB* 5.3 (2011).

[SOR+11] Dan Suciu, Dan Olteanu, Christopher Ré et Christoph Koch. *Probabilistic Databases*. Morgan & Claypool, 2011.

[VK14] Denny Vrandečić et Markus Krötzsch. “Wikidata : a Free Collaborative Knowledgebase”. In : *CACM* 57.10 (2014).

[WP10] Peter Welinder et Pietro Perona. “Online crowdsourcing : rating annotators and obtaining cost-effective labels”. In : *Proc. CVPRW*. 2010.