Haut Conseil de l'Évaluation de la Recherche et de l'Enseignement Supérieur

DOCUMENT D'AUTOÉVALUATION

Équipe BD

Campagne d'évaluation 2023-2024 — Vague D

Table des matières

1 INFORMATIONS GÉNÉRALES SUR L'ÉQUIPE BD **3**
1.1 Les thématiques, scientifiques et leurs enjeux . . . 3
Thématiques scientifiques et enjeux . . . 3
Avancées scientifiques majeures . . . 4
Animation et vie scientifique de l'équipe . . . 7
Recommandations concernant l'organisation et la vie de l'équipe . . . 7
Recommandations concernant les perspectives scientifiques à cinq ans et la faisabilité du projet . . . 8

2 INTRODUCTION DU PORTFOLIO **9**

3 AUTOÉVALUATION DU BILAN **10**
3.1 Autoévaluation de l'équipe . . . 10
Domaine 2. Attractivité . . . 10
Domaine 3. Production scientifique . . . 11
Domaine 4. Inscription des activités de recherche dans la société . . . 13

4 RÉFÉRENCES BIBLIOGRAPHIQUES SIGNIFICATIVES DE BD **15**

A ANNEXE — MEMBRES PERMANENTS AU 31/12/2022 **17**

1 INFORMATIONS GÉNÉRALES SUR L'ÉQUIPE BD

Nom de l'équipe : Bases de Données (BD)

Responsable de l'équipe : Bernd AMANN

	2017	2018	2019	2020	2021	2022
PR	2	2	2	2	1	1
MCF HDR	1	1	1	1	1	2
MCF	3	3	3	3	3	2
DR	0	0	0	0	0	0
CR HDR	0	0	0	0	0	0
CR	1	1	1	0	0	0
Total permanents	7	7	7	6	5	5
Émérites	0	0	0	0	0	0
Doctorants	4	4	2	1	0	2
Ingénieurs CDD ou hors tutelles	2	0	0	0	0	0
Post-doc, ATER, etc.	1	0	0	0	0	0
Stagiaires	4	2	3	1	3	6
Total non permanents	11	6	5	2	3	8
Total avec émérites	**18**	**13**	**12**	**8**	**8**	**13**
Equivalent temps plein recherche	4.0	4.0	4.0	3.0	2.5	2.5

TABLE 1 – Personnels BD sur la période 2017-2022 (au 1er juillet de chaque année)

Pendant la période, l'équipe BD a perdu deux membres permanents. Benjamin Piwowarski (CR) a changé d'équipe en 2019 et un Anne Doucet (PR) est partie à la retraite en 2020. Nous avons pu recruter un Maître de Conférences (attaché pour l'enseignement à l'Université de Nanterre) début 2023 et une demande de recrutement d'un MCF est en cours dans le cadre de la révision des effectifs 2023. L'équipe compte actuellement un seul professeur et une demande de PR est prévue pour la révision des effectifs 2024. Les années 2020 et 2021 sont également marquées par une baisse importante de membres non-permanents (stagiaires, doctorants), en partie liée aux difficultés de recrutement et d'encadrement pendant la période de crise COVID-19.

1.1 Les thématiques, scientifiques et leurs enjeux

Thématiques scientifiques et enjeux

L'objectif principal de l'équipe BD est de proposer de nouvelles solutions pour la mise en œuvre pratique de tâches complexes de gestion, transformation et d'analyse de données. Les enjeux scientifiques consistent à proposer des modèles, des algorithmes et des architectures pour réaliser des tâches d'exploration et d'analyse de données complexes en identifiant des défis liées à l'intégration d'infrastructures de gestion de données (big data, noSQL, cloud, cluster) avec des méthodes récentes en intelligence artificielle (apprentissage supervisé et non-supervisée, fouille de données) pour leur valorisation.

Nous pouvons identifier six thématiques de recherche majeurs étudiées dans l'équipe. Malgré la diversité de ces thématiques et de leurs enjeux, elles partagent l'objectif de traiter des problèmes complexes sur des données réelles en partant d'une modélisation formelle (logique, statistique) des données et des opérations (transformation, interrogation, agrégation) jusqu'au déploiement de solutions avec des technologies récentes pour le traitement parallélisé de grands volumes de données (big data, cloud), et en particulier la plateforme Apache Spark. Suivant cette démarche "de la modélisation jusqu'au déploiement", nos recherches s'inscrivent dans les défis en Science des Données et les axes AID (Intelligence Artificielle et Science des Données) et et ASN (Architecture, Systèmes et Réseaux) du laboratoire LIP6.

Thématiques et enjeux

Thème 1 – Qualité de données analytiques. Les sciences des données sont fondées sur l'apparition de nouveaux outils d'analyse et de visualisation de données complexes. Une première étape dans l'utilisation de ces outils est l'extraction et la transformation de données par des enchaînements de requêtes analytiques (filtrage, projection, partitionnement, agrégation) pour créer des vues pour une analyse plus approfondie. Un enjeu important dans cette première étape est l'absence d'outils pour vérifier la cohérence des traitements effectuées et des données utilisées, ce qui introduit souvent des erreurs importants dans les résultats d'analyse obtenus. L'équipe

travaille sur ces enjeux depuis plusieurs années et encadrées deux thèses pendant la période d'évaluation, dont une en collaboration avec la société SAP (Thème 1).

Thème 2 – Recommandation et analyse de réseaux sociaux. Les réseaux sociaux sont devenus des sources d'informations cruciales avec des nombreuses applications comme la recommandation de contenus et de produits. L'analyse des données produites par l'activité humaine dans ces réseaux (diffusion de messages, déplacements dans les réseaux de mobilité) est souvent complexe à cause de la quantité, la complexité et la qualité des données à traiter. Un enjeu important est alors de proposer des méthodes qui produisent des résultats d'analyse (recommandations) de bonne qualité et qui passent également à l'échelle. Dans le cadre de diverses collaborations, l'équipe a co-encadré une thèse sur la recommandation des points d'intérêts dans les réseaux de mobilité et deux thèses sur l'analyse des données du réseau social Twitter. (Thème 2).

Thème 3 – Inférence et satisfiabilité de schémas JSON. Ces dernières années ont vu l'utilisation généralisée de JSON comme format de données ouvert pour stocker, échanger et interroger des collections de données massives. Un enjeu important est l'absence d'informations sur la structure de ces données (schémas) qui réduit fortement leur utilisation. Pour répondre à cet enjeu, l'équipe collabore et co-encadré des thèses avec plusieurs partenaires académiques pour développer des solutions pour l'inférence de schémas JSON à partir de données massives et l'analyse formelle (satisfiabilité, inclusion, équivalence) de ces schémas (Thème 3).

Thème 4 – Évolution des sciences. Les archives bibliographiques telles que le Web of Science, ISTEX, arXiv ou PubMed représentent des ressources importantes pour l'étude de l'évolution des sciences. Révéler des motifs d'évolution significatifs à partir d'archives documentaires, a de nombreuses applications pour positionner des travaux de recherche dans leur contexte scientifique et historique ou évaluer le potentiel d'innovation et de transfert technologique d'un résultat scientifique. Bien qu'il existe de nombreux modèles performants pour la représentation compacte du contenu thématique d'un corpus documentaire, la représentation de l'évolution de ce contenu pose encore des défis importants de modélisation et de passage à l'échelle sur des grands corpus. L'équipe a abordé ces sujets dans le cadre d'un projet ANR et de deux thèses, dont une est en cours (Thème 4).

Thème 5 – Représentation d'informations et génération de résumés. La représentation statistique d'informations est un enjeu central pour l'utilisation des techniques récentes d'apprentissage supervisé et non-supervisé. La qualité des représentations dépend fortement des données utilisées, et un enjeu important est de développer des modèles de représentation capables d'exploiter des données divers (multi-modèles) et incertaines. L'équipe a étudié ces enjeux en collaboration étroite avec l'équipe MLIA du LIP6 (maintenant ISIR).

Thème 6 – Interrogation et raisonnement RDF. Le format RDF (Resource Description Framework) est une norme permettant de décrire de manière souple des données (entités et relations) et des connaissances (hiérarchies de classes et de propriétés, ontologies) sous la forme de *graphes de connaissances*. Ces graphes peuvent devenir très volumineux en termes de nombre de nœuds et de liens. En particulier, après avoir matérialisé toutes les relations sémantiques définies dans les ontologies associées aux données, ces graphes peuvent contenir des milliards de nœuds reliés. L'interrogation de ces grands graphes RDF avec le langage SPARQL pose des problèmes de performance et de passage à l'échelle, et de nombreuses solutions utilisant des technologies noSQL ont été présentées dans la littérature. L'équipe a contribué à cet effort avec le Laboratoire d'Informatique Gaspard Monge (LIGM) en proposant des nouvelles techniques d'optimisation des requêtes distribuées pour les requêtes SPARQL dans l'environnement Apache Spark (Thème 5).

Avancées scientifiques majeures

Les différents thèmes de recherche de l'équipe se distinguent par les types de données traitées et les problèmes abordés. Ils partagent souvent la même approche scientifique qui consiste à analyser les problèmes du point de vue de l'exploitation des données complexes (données multi-dimensionnelles, réseaux sociaux, JSON, RDF) et à trouver des solutions qui passent à l'échelle en utilisant les récentes technologies de traitement parallèle des données.

Thème 1 – Qualité de données analytiques

De plus en plus de jeux de données sont créés par des par des experts d'une manière ad-hoc en utilisant des outils d'analyse et de visualisation de données (PowerBI, SAP BusinessIntelligence, Tableau). Ces outils permettent de générer et composer des requêtes analytiques (SQL) pour construire des nouvelles vues. Les opérations sont souvent complexes (agrégation, projection, filtrage) et appliquées à des données imparfaites (incomplètes, erronées). Dans ce contexte, un objectif de l'équipe est de fournir aux experts des outils théoriques et pratiques pour évaluer et améliorer la qualité des données générées.

(1) Données manquantes et imputation. Une première contribution est un modèle de représentation compacte (minimale) de données manquantes sous forme de tables de motifs [Hannou et al., 2019c] (prix du meilleur article). La nouveauté de ce modèle est qu'il permet de générer automatiquement des tables de motifs décrivant les fragments des données manquantes ou erronées produites par des requêtes analytiques [Hannou et al., 2019b]. Une deuxième contribution est un nouvel algorithme à base de règles d'imputation pour réparer les erreurs introduites par des données manquantes.

(2) Augmentation de schémas. Une deuxième contribution sur la qualité des données étend les travaux existants sur l'augmentation de tables relationnelles

au cas de tables de données analytiques. La solution présentée dans [Liu et al., 2020] définit des critères de qualité formels pour les schémas obtenus par la jointure (augmentation) de tables analytiques.

Ces critères sont ensuite utilisés pour définir des opérations de réparation automatique pour notifier la génération d'attributs ambigus, déduire les fonctions d'agrégation applicables sur les nouveaux attributs, et compléter les résultats de fusion obtenus par une augmentation de schéma incomplète. La solution proposée inclut une analyse théorique minutieuse et une implémentation dans le système SAP HANA. Dans la continuité de ce travail, [Simon et al., 2023b] introduit une nouvelle condition de *résumabilité généralisée* (generalized summarizability) et des règles de propagation pour détecter et annoter des séquences de requêtes d'agrégation qui ne respectent pas cette condition.

Thème 2 – Recommandation et analyse de réseaux sociaux

L'explosion des données circulant dans les réseaux sociaux rend possible et nécessaire l'extraction des préférences des utilisateurs et la recommandation des données publiées. Les modèles de recommandation actuels souffrent de plusieurs problèmes liés à la qualité des données et à la taille du réseau, qui sont abordés par l'équipe.

(1) Recommandation de Points d'Interêt Géographiques (POI). La rareté des données et le comportement de mobilité hétérogène (urbain, régional, mondial) des utilisateurs dans les réseaux sociaux géo-localisés rend le problème de recommandation de points d'intérêt (POI, Point-Of-Interest) difficile, en particulier pour les grandes zones géographiques et les ensembles de données mondiaux. Dans une thèse

co-encadré avec Télécom-Paris [Griesner et al., 2018a] nous explorons l'impact du clustering spatial sur la qualité de la recommandation. L'approche proposée combine le clustering spatial avec les influences des utilisateurs. Elle est basée sur un modèle de factorisation de Poisson construit sur un réseau social implicite, déduit des modèles de mobilité géographique. Les expériences montrent que notre approche permet d'obtenir une qualité de recommandation nettement supérieure à celle des autres techniques de recommandation de pointe. Dans ce contexte, en particulier pour valider des solutions de recommandation de points d'intérêt géographiques (POI) à large échelle, nous avons également proposé une solution efficace pour enrichir des très grands jeux de données contenant des traces de déplacement d'utilisateurs, en les croisant avec une base décrivant des points d'intérêt et leurs catégories. La méthode conçue est robuste au déséquilibre des données lorsque certains POI sont beaucoup plus populaires que les autres [Gueye et al., 2020b]. Pour le cas de la recommandation de notes à des utilisateurs basée sur l'historique des notes précédemment attribuées (approche collaborative), nous avons proposé une nouvelle méthode qui détermine automatiquement le nombre de voisins similaires à prendre en compte dans le choix d'un article à recommander [Fopa et al., 2022b].

(2) Analyse de réseaux sociaux. Dans le cadre d'une thèse co-encadrée avec le laboratoire CEDRIC du CNAM, nous proposons un modèle de recommandation de contenu inspiré des méthodes de filtrage collaboratif qui repose sur l'homophilie présente dans le réseau social Twitter [Grossetti et al., 2018]. Nous comparons différentes méthodes de détection de communautés sur Twitter et observons finement le type des communautés produites par ces méthodes. Cette analyse de l'aspect communautaire de Twitter permet de quantifier l'effet de bulle produit par les algorithmes de recommandation. En utilisant les différents liens (topologiques, sémantiques, influence) entre les communautés, nous montrons qu'il est possible de limiter cet effet de cloisonnement des informations [Grossetti et al., 2021]. Une deuxième thèse, également co-encadrée avec le CEDRIC, analyse le comportement des utilisateurs dans les réseaux sociaux afin de détecter des comptes "anormaux" [Debure et al., 2020a]. Nous proposons une approche de détection d'utilisateurs populaires qui s'appuie sur une modélisation de l'évolution de la popularité sous la forme de motifs fréquents et un modèle de matching des motifs continu qui passe à l'échelle. Dans [Debure et al., 2021], nous proposons également une méthode de clustering basé sur le PageRank qui permet d'identifier des groupes d'utilisateurs partageant le même rôle, en utilisant les graphes d'interactions qu'ils génèrent. Enfin, nous nous sommes intéressés au passage à l'échelle d'algorithmes de calculs de plus courts chemins et de détection de communautés utilisés dans les systèmes de recommandation. Les méthodes de partitionnement des graphes proposées permettent l'exécution de ces algorithmes sur des architectures massivement

parallèles, en minimisant les interactions entre nœuds de calcul par une nouvelle technique de partitionnement d'arêtes qui prend en compte les spécificités des algorithmes de calcul et les propriétés topologiques des grands graphes du monde réel. L'approche est conçue pour améliorer la performance des algorithmes basés sur les marches aléatoires, utilisés notamment pour la recommandation [Li et al., 2017]. Nous avons également proposé une stratégie de partitionnement de multi-graphes, où l'affectation des arêtes à des blocs utilise un score de proximité des arêtes avec les profils des blocs basé sur les étiquettes.

Thème 3 – Inférence et satisfiabilité de schémas JSON

Le format JSON n'impose pas la définition de schémas pour la génération de données, mais l'absence de schémas a des conséquences négatives importantes pour les utilisateurs qui ne disposent pas de description des propriétés structurelles pour faciliter la formulation de requêtes et pour les systèmes qui ne peuvent pas exploiter de nombreuses optimisations basées sur les schémas. Un deuxième aspect plus formel concerne le raisonnement sur l'implication et l'équivalence de schémas JSON qui ont des applications directes dans l'analyse de l'évolution des schémas et l'intégration de données.

(1) Inférence de schémas JSON. Une solution est de générer des schémas JSON a posteriori à partir des données. Ceci pose plusieurs défis liés à la taille et la complexité des données JSON. Une première approche proposée dans [Baazizi et al., 2017a], puis étendue avec la notion de *parametricité*, offre le choix du niveau d'abstraction des schémas inférés [Baazizi et al., 2019b]. De plus, une inférence interactive guidée par l'utilisateur a été décrite formellement dans [Baazizi et al., 2019a] puis démontrée expérimentalement dans [Baazizi et al., 2020a]

(2) Satisfiabilité de schémas JSON. JSON Schema [1] utilise des assertions structurelles combinées à l'aide d'opérateurs logiques et son expressivité entraîne une complexité rédhibitoire pour la plupart des problèmes de décision, en particulier, la satisfiabilité. La satisfiabilité, l'inclusion et l'équivalence de schémas JSON peuvent être réduits au problème de la génération de témoins (satisfiabilité = existence de témoin). La satisfiabilité, l'inclusion et l'équivalence des schémas sont décidables, cependant, aucun algorithme de génération de témoins n'a encore été formellement décrit. Dans un article publié dans une conférence majeure du domaine [Attouche et al., 2022a], nous proposons un premier algorithme direct pour la génération de témoins de schémas JSON, et nous étudions son efficacité dans des expériences sur plusieurs collections de schémas, y compris des milliers de schémas du monde réel. D'autres travaux liés à ce travail, analysent les motifs fréquemment utilisés dans les schémas réels en se focalisant sur l'usage de la négation, opérateur peu fréquent dans les langages de schémas [Baazizi et al., 2021].

Thème 4 – Évolution des sciences

Dans le cadre du projet ANR EPIQUE, nous avons adopté une représentation de l'évolution de la science sous la forme de réseaux reliants des thèmes de recherche (topics) présents pendant différentes périodes dans un corpus scientifique. Les outils existants de génération et d'analyse de ces "réseaux d'évolution thématiques" étaient limités au traitement de corpus de taille moyenne et une exploration visuelle simple des réseaux construits. Notre objectif était de développer des solutions efficaces pour générer et interagir avec les réseaux d'évolution sur des très grands corpus. Nous avons défini un modèle de graphe avec un langage d'interrogation qui, combiné à des méthodes de visualisation, permet d'explorer des grands réseaux thématiques complexes en formulant des contraintes sur le contenu et la structure du réseau [Li et al., 2021]. Le prototype résultant passe à l'échelle et traite efficacement des réseaux avec des milliers de topics extraits de corpus contenant plusieurs millions de documents. Dans ce contexte, nous avons également proposé une nouvelle méthode de calcul de similarité parallèle pour aligner des topics similaires situés dans des périodes de temps différentes.

Thème 5 – Représentation d'informations et génération de résumés

(1) Représentation de mots. Représenter la sémantique d'un mot est un problème de longue date qui conditionne plusieurs applications majeures telles que la traduction automatique, l'analyse de sentiments et le résumé de texte. Des études psychologiques ont montré que le sens des mots est ancré dans la perception humaine, et un enjeu pour améliorer la qualité des représentations de mot est le développement de modèles de représentation qui tirent parti des connaissances implicites présentes dans ces sources d'information multi-modales (texte, image, son).

1. `https://json-schema.org`

Dans la cadre de nos travaux sur la représentation de mots multi-modale, nous avons d'abord proposé un modèle multimodal d'apprentissage de représentations de mots [Zablocki et al., 2018b] et ensuite travaillé sur des modèles permettant d'apprendre des représentations de phrases ancrées dans les images.

(2) Génération et résumé automatique de texte. Les modèles de génération de textes les plus performants actuellement sont basés sur des architectures de type Transformer qui permettent de prendre en compte un contexte bien plus large que les réseaux de neurones récurrents. Un enjeu important dans ce contexte est aujourd'hui le développement de modèles qui sont plus résistants au biais d'exposition, tout en nécessitant moins de données en apprentissage. Nous avons travaillé sur des modèles plus résistants au biais d'exposition tout en nécessitant moins de données en apprentissage. Dans [Scialom et al., 2019a], nous avons proposé de découpler le problème de générer un texte cohérent de celui de générer un texte qui contienne les éléments informatifs demandés et nous avons montré qu'il était possible de définir des indicateurs fiables sans toutefois découpler la syntaxe et le contenu sémantique.

3) Représentation d'informations incertaines. Les graphes de données sont de plus en plus utilisés pour la représentation d'informations complexes. L'information représentée dans ces graphes de données peut être déséquilibrée en termes de nombre de relations entre les nœuds et de données connues pour chaque nœud. Un enjeu important est alors de définir des modèles qui prennent en compte l'incertitude d'informations liée à cette absence de relations et de données. Les modèles probabilistes sont particulièrement adaptés pour la représentation d'informations incertaines dans des graphes de données incertaines. Nous avons développé des modèles qui prennent en compte l'incertitude des données où les nœuds sont représentés avec une distribution à variance variable. Ces techniques ont été appliquées à la recommandation [dos Santos et al., 2017] et à la classification dans les graphes [dos Santos et al., 2018].

Thème 6 – Interrogation et raisonnement RDF

Nous avons proposé des solutions qui passent à l'échelle pour exécuter efficacement des requêtes SPARQL sur des très grands graphes dans l'environnement de calcul parallèle Apache Spark. Nous avons défini un modèle de coût permettant de choisir parmi les deux opérateurs standards de jointure distribuée (jointure parallèle par hachage ou jointure par boucles imbriquées avec diffusion) qui tient compte des différents modèles de stockage distribué de la plateforme Apache Spark pour minimiser les transferts de données (shuffling). Nos résultats expérimentaux montrent qu'une utilisation combinée des deux opérateurs de jointure permet d'accélérer significativement les requêtes [Naacke et al., 2017a]. Ce travail a été ensuite adapté pour des flux de données RDF et étendu à du raisonnement à base de règles logiques dont le formalisme s'inspire du langage *Answer Set Programming* [Ren et al., 2017, Ren et al., 2018a].

Animation et vie scientifique de l'équipe

Les membres permanents de l'équipe se réunissent régulièrement (une fois par mois en moyenne) pour discuter des actions à mener au niveau scientifique et pédagogique. Les encadrants rencontrent leurs doctorants et stagiaires sur une base hebdomadaire pour les guider et suivre l'avancement de leurs travaux. L'équipe organise des séminaires conjoints (deux trois fois par an) où les doctorants et les stagiaires peuvent présenter leurs travaux et échanger leur expérience. Nous encourageons nos doctorants à participer à des séminaires et à des écoles d'été.

En dehors des réunions formelles, les membres de l'équipe se rencontrent régulièrement lors des pauses-café et déjeuner pour discuter des problèmes quotidiens et s'organiser. Pour le partage d'informations, l'équipe a mis en place plusieurs espaces de partage (nuage LIP6, site web). Quand les conditions financières le permettent, le budget de l'équipe est mutualisé de manière collégiale pour financer des missions ou acheter des équipements.

Recommandations concernant l'organisation et la vie de l'équipe

Rapport 2018. Il faudrait essayer d'anticiper les effets déstabilisants qui vont mécaniquement être provoqués par des départs en retraite et de possibles mutations. Par exemple, en renforçant les collaborations intra-équipe, ce qui peut passer par le fait que toute nouvelle thèse soit co-encadrée.

Réponse. Nous avons suivi la recommandation d'impliquer au moins deux encadrants de l'équipe dans chaque thèse pour renforcer les collaborations intra-équipe. Pendant la période, l'équipe a perdu deux membres permanents (un changement d'équipe et un départ à la retraite). Nous avons pu recruter un maître de conférence (attaché pour l'enseignement à l'Université de Nanterre) début 2023 et un recrutement d'un MCF est en cours

dans le cadre de la révision des effectifs 2023. L'équipe compte actuellement un seul professeur et une demande de PR est prévue pour la révision des effectifs 2024.

Rapport 2018. Étant donné la taille de l'équipe (moyenne, comparée à d'autres du laboratoire, bien qu'au final d'une bonne taille, c'est-à-dire gérable), le nombre de sujets de recherche est élevé, sans priorité affichée. Une réflexion pourrait être menée au sein de l'équipe pour définir des priorités entre ces différents sujets.

Réponse. Le nombre de thématiques de recherche décrit dans le rapport reste assez élevée. Néanmoins on a commencé à recentrer plus les thématiques. Ainsi le Thème 5 n'est plus couvert avec le départ d'un membre de l'équipe vers l'équipe MLIA en 2019, le Thème 6 a été abandonnée depuis 2019. Actuellement, l'effort est surtout concentré sur les quatre thèmes restants impliquant plusieurs chercheurs permanents par thème.

Recommandations concernant les perspectives scientifiques à cinq ans et la faisabilité du projet

Rapport 2018. L'application des compétences de l'équipe à la représentation et la gestion des données textuelles, des réseaux sociaux, mais aussi des données plus classiques du web ainsi que de l'IoT donne un spectre trop large pour pouvoir être maîtrisé avec les forces en présence, d'autant plus qu'il est très pertinent de poursuivre les réelles collaborations avec les collègues qui apportent leurs compétences en apprentissage automatique et fouille de données. Une réflexion stratégique devrait être menée sur la pertinence de resserrer l'équipe autour d'un noyau restreint de problématiques. Et de faire converger les forces en présence, pour porter ou participer à des soumissions de projets typiquement au niveau H2020 mais aussi développer des collaborations industrielles directes (telles celles évoquées à propos de SAP et ATOS)

Réponse. Le nombre de thématiques de recherche présentés dans l'introduction reste important. Néanmoins, une analyse plus fine de l'évolution des activités permet de voir que la plupart des thématiques sont animées par au moins deux permanents et que les différentes thématiques couvrent des problématiques de traitement et d'analyse de données (modélisation de données structurées et complexes, optimisation et parallélisation de traitements de données, qualité de données) qui sont au cœur des compétences de l'équipe. Plus généralement, le recouvrement des différentes thématiques reflète la tendance noSQL (not only SQL) d'adapter des solutions de modélisation et de gestion de données à des besoins particuliers de traitements et d'analyse de données. En ce qui concerne le montage de projet, l'équipe s'est concentrée sur le renforcement des collaborations industrielles dans le cadre de thèses CIFRE et de contrats de collaborations directes (voir Domaine 4).

2 INTRODUCTION DU PORTFOLIO

Cette section identifie les éléments de portfolio présentés par l'équipe BD. Chaque élément disposant de sa propre fiche explicative, nous nous contentons ici d'en donner une liste simple en indiquant le thème de recherche concerné à la fin :

- *Élément 1 (publication) :* L'article *"Discovering and Merging Related Analytic Datasets"* [Liu et al., 2020] présente un travail effectué dans le cadre d'une thèse CIFRE avec la société SAP France sur la découverte et l'augmentation semi-automatique de tables de données analytiques (Thème 1).
- *Élément 2 (publication) :* L'article *"An homophily-based approach for fast post recommendation in microblogging systems"* [Grossetti et al., 2018] présente un résultat d'une collaboration à long terme avec le laboratoire CEDRIC du CNAM Paris sur l'analyse du comportement des utilisateurs dans les réseaux sociaux et la mise à l'échelle d'algorithmes utilisés dans les systèmes de recommandation (Thème 2).
- *Élément 3 (vidéo) :* La vidéo *"Inférence Interactive de Schema JSON"* présente à travers d'un cas d'usage un prototype qui offre la possibilité de visualiser des schémas JSON inférés automatiquement à partir de collections de documents JSON et de choisir, de manière interactive, le niveau de précision souhaité pour chaque fragment du schéma sans impacter les autres fragments.
- *Élément 4 (vidéo) :* la vidéo *"EPIQUE : Extracting Meaningful Science Evolution Patterns from Large Document Archives"* présente un prototype développé dans le cadre du projet ANR EPIQUE pour générer et explorer l'évolution des thèmes de recherche dans les archives scientifiques. L'exemple choisi est un corpus sur les recherches sur glyphosate extrait de l'archive Web of Science (Thème 4).

3 AUTOÉVALUATION DU BILAN

3.1 Autoévaluation de l'équipe

Domaine 2. Attractivité

Référence 1. L'unité s'est assigné des objectifs scientifiques pertinents.

Invitations dans des institutions académiques ou des congrès internationaux. En tant que coordonnateur du projet ANR EPIQUE, l'équipe a participé à la table ronde "Digital Technology and Heritage – Challenges and Issues", Numérique et Patrimoine, ANR, 11-12 Mars 2021 (programme). Dans le même contexte, l'équipe a été coorganisateur d'une session à la conférence IHSPSBB à Oslo en 2019 sur le thème "Digital history and philosophy of science : The reconstruction of scientific phylomemies as a tool for the study of the life sciences" (programme)

Conférences où des membres de l'équipe sont membres du PC. Les membres de l'équipe ont participé aux comités de programme de nombreuses conférences internationales : ADBIS (2017-2022), APVP (2018), AWD (2020), DATA (2023), (2023), *EDBT* (2017 et 2020), ICCS (2020-2021), *VLDB* (2018), WISE (2019-2022), ICWE (2021-2022), MEDES (2022)

Ils ont également contribué en tant que relecteur dans des journaux internationaux : Knowledge and Information Systems (KAIS), The VLDB Journal, Information Retrieval (IR), International Journal of Data Science and Analytics (JDSA).

Organisation de conférences. L'équipe a organisé la 26e Conférence en "Gestion de Données - Principes Technologies et Applications" (BDA 2020). A cause de la crise COVID-19, cette conférence a été organisée en mode virtuel sur une durée de trois jours avec environ 120 participants.

Responsabilités éditoriales dans des revues et des collections. un membre de l'équipe est co-éditeur d'un numéro spécial du journal Transactions on LargeScale Data- and Knowledge-Centered Systems XLIX, volume 12920 of Lecture Notes in Computer Science. Springer et des proceedings de la Conérence BDA'2020.

Participation à des instances de pilotage de la recherche et d'expertise scientifique. Un membre de l'équipe est membre du comité de direction du GDR Madics depuis sa création. Un membre de l'équipe a également été membre des Comités d'évaluation HCERES du laboratoire LIMOS (décembre 2020) et du laboratoire iCube (janvier 2017). Les membres de l'équipe ont été sollicités comme experts scientifiques par des agences nationales (CNRS, Ministère de l'Industrie, Universités) et internationales (Austrian Science Fund) pour l'évaluation de soumissions de projets (JEI-CIR, Initiative d'Excellence Paris Seine CYU Cergy, STIC-AMSUD, ECOS Nord).

Prix, distinctions. L'équipe a reçu deux Best Paper Award. Le premier pour un article [Baazizi et al., 2020b] à la conférence nationale BDA 2020 et le second pour unenpublication [Hannou et al., 2019c] à la conférence internationale DBKDA 2019.

Invitation. Un membre de l'équipe a été invité à participer au séminaire Dagstuhl 17262 "Federated Semantic Data Management" (juin 2017).

Référence 2. L'unité est attractive par la qualité de sa politique d'accompagnement des personnels.

Thèses et stages. Tous les membres de l'équipe sont des enseignants-chercheurs et notre activité de recherche repose principalement sur l'encadrement de doctorants et de stagiaires sur des sujets de recherche spécifiques. Pendant la période, quatre doctorantes et doctorants de l'équipe ont soutenu leur thèse et deux thèses coencadrées par un membre de l'équipe ont été soutenues au laboratoire Cedric. Trois thèses sont actuellement en cours dans l'équipe et une thèse est co-encadrée avec une équipe du LAMSADE. Chaque thèse est encadrée par un ou deux membres de l'équipe. Les doctorantes et doctorants sont systématiquement premiers auteurs de leurs publications. L'équipe incite les doctorants à participer à des écoles d'été et à soumettre des publications dans des sessions doctorants. Pour chaque doctorant, les encadrants organisent une réunion hebdomadaire régulière pour le suivi de la thèse. Nous organisons également des réunions régulières pour des présentations en interne. L'équipe aide les doctorantes et doctorants à la recherche d'un post-doc ou d'un travail après leur thèse.

Membres permanents. Pendant la période d'évaluation, l'équipe a perdu un CR (changement d'équipe en 2019) et un professeur (départ à la retraite en 2020). L'équipe n'a pas recruté depuis 2013 avant le dernier recrutement en janvier 2023 d'un MCF, qui effectue ses enseignements à l'Université de Nanterre et ses recherches au LIP6.

Pendnant la période, un membre de l'équipe a passé son HDR. L'équipe est indiquée comme équipe d'accueil sur un poste de MCF ouvert en 2023.

Chercheurs invités. Pendant la période d'évaluation, l'équipe a accueilli six chercheurs permanents (Université Cheikh Anta Diop de Dakar, Université de Caracas, Université de British Columbia) pour des durées entre un et 10 mois.

Référence 3. L'unité est attractive par la reconnaissance de ses succès à des appels à projets compétitifs.

Contrats de recherche et de collaborations. L'équipe a été *coordonnateur* du projet ANR EPIQUE (AAP ANR La Révolution numérique : rapports aux savoirs et à la culture) sur l'analyse de l'évolution des sciences dans les archives scientifiques (Thème 4, site web). Ce projet a permis de financer une thèse et d'obtenir un financement de thèse DIM avec ISC-PIF (le candidat a malheureusement abandonné sa thèse après).

Dans le Thème 2 L'équipe a participé à un projet *FUI ABCD* (Advertising Big Collaborative Data) sur des nouvelles méthodes de profilage utilisateur à partir de l'analyse des données transactionnelles et de navigation, d'une part, ainsi que celle des contenus pertinents provenant des réseaux sociaux, d'autre part.

Le projet *CNRS Senagro* est financé par le dispositif de soutien aux collaborations avec l'Afrique Subsaharienne de la Direction Europe de la Recherche et Coopération Internationale (DERCI). L'objectif est de développer avec des collègues de l'Université Cheikh Anta Diop de Dakar (UCAD) un kit de formation à la data science pour l'agriculture intelligente au Sénégal (page web).

Dans le cadre d'un projet *Sorbonne Université Emergence* (DNA Storage) l'équipe BD a collaboré avec le Laboratoire de Biologie Moléculaire et Cellulaire des Eucaryotes (LBMCE) et l'équipe DELYS du LIP6 sur l'archivage de données massives dans des séquences ADN (page web).

L'équipe BD a également participé à plusieurs *projets LIP6* avec les équipes MOVE sur l'extraction des motifs dans les lignes de produits logiciels , avec l'équipe ComplexNetworks sur l'analyse topologique et requêtes interactives dans des grands graphes sémantiques et l'équipe LFI sur la cartographie automatique d'un datalake par l'apprentissage par renforcement avec budget contraint.

Autres collaborations académiques. Dans le Thème 1, l'équipe a collaboré avec l'*Institut Fraunhofer* (Karlsruhe) sur la représentation compacte de données manquantes pour l'analyse et la réparation de requêtes analytiques. Un deuxième sujet dans ce thème, en collaboration avec la société *SAP France*, concerne le problème de requêtes analytiques incohérentes dans le contexte de processus d'analyses interactives de données multidimensionnelles.

Dans le Thème 2, l'équipe a collaboré avec *Télécom Paris* sur la recommandation de points d'intérêt (POI) et avec l'Université Cheikh Anta Diop de Dakar (*UCAD*) sur la modélisation de propagation de maladies infectieuses à partir de données de mobilités. Dans le même thème, une deuxième collaboration importante avec le laboratoire *CEDRIC* du Conservatoire National des Arts et Métiers (CNAM) a porté sur la diversité des recommandations (effet bulle), recommandation de Tweet et le partitionnement de grands graphes hétérogènes dans le contexte des données Twitter.

Dans le Thème 3, l'équipe a collaboré avec le laboratoire *LAMSADE* de l'Université Paris-Dauphine, l'*Université de Pise*, l'*Université de Basilicate* et l'*Université de Passau*, sur l'inférence des schémas JSON à partir de données massives et sur l'analyse formelle de propriétés de schémas JSON (satisfiabilité, inclusion, équivalence).

Dans le Thème 4, l'équipe a collaboré avec l'Institut d'Histoire et de Philosophie des Sciences et Techniques (*IHPST*), l'Institut des Systèmes Complexes de Paris Ile de France (*ISC-PIF*) et l'*IRISA Rennes*.

Les travaux dans le Thème 5 ont été effectués en collaboration avec l'équipe *MLIA*. Les thématiques autour de la recherche d'information et l'apprentissage automatique se sont arrêtés en 2019 avec le départ du membre porteur de la thématique vers l'équipe MLIA (ISIR).

Dans le Thème 6, l'équipe a collaboré avec l'Institut d'électronique et d'informatique Gaspard-Monge (*IGM*) de l'Université Paris-Est Marne-la-Vallée (UPEM) sur l'optimisation de requêtes SPARQL et le raisonnement sur les flux RDF.

Référence 4. L'unité est attractive par la qualité de ses équipements et de ses compétences techniques.

L'équipe n'est pas concernée par ce point.

Domaine 3. Production scientifique

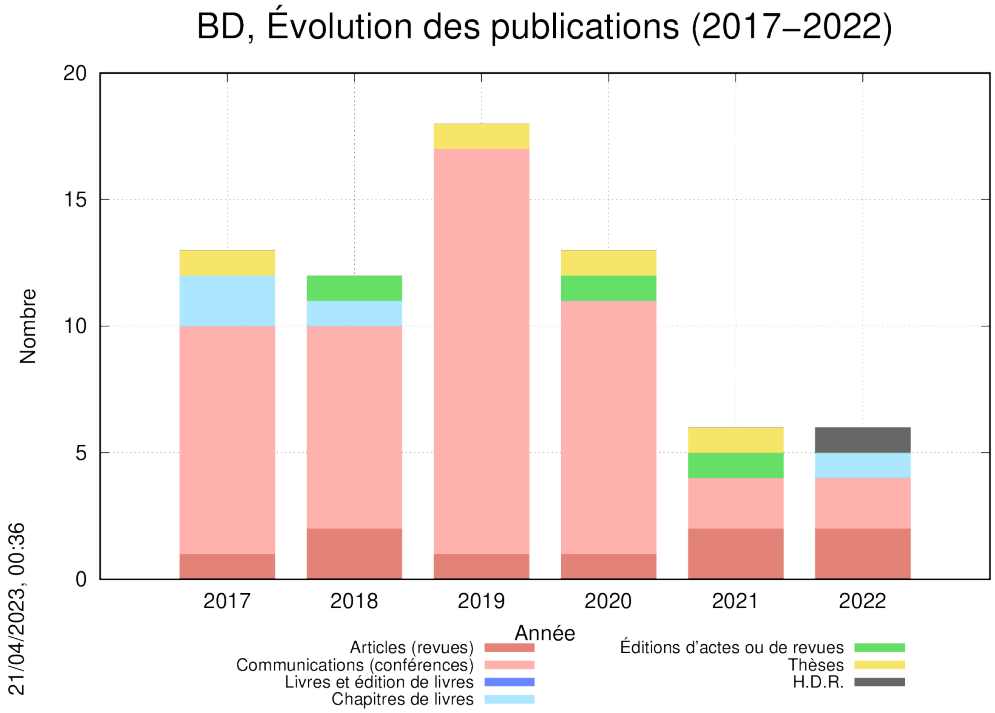

FIGURE 1 – Évolution des publications entre 2017 et 2022

	2017	2018	2019	2020	2021	2022
Articles (revues)	0.25	0.50	0.25	0.33	0.80	0.80
Communications (conférences)	2.25	2.00	4.00	3.33	0.80	0.80

TABLE 2 – Publications par ETPR par an entre 2017 et 2022

Référence 1. La production scientifique de l'unité satisfait à des critères de qualité.

L'équipe continue à faire un effort pour améliorer la qualité des supports de publications visés. Ainsi, pendant la période, l'équipe a réussi à publier dans des conférences et des journaux (Data and Knowledge internationaux de très bon niveau :

Thème 1. Information Systems (2020), International Conference on Database and Expert Systems Applications (DEXA 2019)

Thème 2. International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2018), Data and Knowledge Engineering (DKE 2022), International Conference on Extending Database Technology (EDBT 2018), International Conference on Scientific and Statistical Database Management (SSDBM 2017), International Journal of Web Information Systems (2021)

Thème 3. Proceedings of the VLDB Endowment (PVLDB 2022), International Colloquium on Automata, Languages, and Programming (ICALP 2019), The VLDB Journal (2019), International Conference on Extending Database Technology (EDBT 2020 et 2017)

Thème 4. Big Data Research (2021)

Thème 5. ACM Transactions on Knowledge Discovery from Data (TKDD 2018), ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), AAAI Conference on Artificial Intelligence (AAAI, 2018)

Thème 6. IEEE International Conference on Big Data (2017 et 2018)

Référence 2. La production scientifique de l'unité est proportionnée à son potentiel de recherche et correctement répartie entre ses personnels.

Le nombre de publications scientifiques était très bon jusqu'en 2020, suivi d'un baisse de publications pour certains membres en 2021 et 2022. Cette variation importante s'explique par l'absence de nouveaux recrutements en thèses à partir de 2018 (pas de thèse en cours entre juin 2021 et octobre 2021). L'année 2022 est marquée par une bonne reprise avec l'encadrement de 7 stagiaires Master et le recrutement de 4 nouveaux doctorants à partir de 2022 (un 5e recrutement est actuellement en cours). Cet effort de recrutement commence à porter ses fruits avec des nouvelles soumissions en 2023.

La plupart des publications sont des résultats de travaux de thèses encadrées par un ou deux membres permanents de l'équipe. Les doctorants sont généralement indiqués comme premiers auteurs et invités à présenter leurs travaux aux conférences. Tous les membres de l'équipe sont publiants, avec au moins deux publications par an en moyenne.

Référence 3. La production scientifique de l'unité respecte les principes de l'intégrité scientifique, de l'éthique et de la science ouverte. Elle est conforme aux directives applicables dans ce domaine.

L'équipe favorise clairement les supports de publications de qualité et validé par la communauté. Elle vérifie toutes les sollicitations par des revues "prédatrices" (page wikipedia) avant d'accepter de contribuer. Quand c'est possible, l'équipe publie ses résultats de recherche et les données utilisées pour les expériences sous forme d'archives sur son site web (page projets).

Domaine 4. Inscription des activités de recherche dans la société

Référence 1. L'unité se distingue par la qualité et la quantité de ses interactions avec le monde non-académique.

L'équipe fait un gros effort pour établir et pérenniser des collaborations industrielles. Elle a depuis plusieurs années une collaboration avec la société SAP France autour de la qualité des données analytiques (une thèse CIFRE) et l'optimisation de pipelines de données (trois stages Master en 2022 et 2023). Cette collaboration industrielle nous a permis également de produire deux publications majeures dans des journaux internationaux importants (Information Systems, ACM Data and Information Quality. Nous sommes également en discussion avec la société SAP France pour l'établissement d'un contrat cadre de collaboration. Une deuxième collaboration importante avec la société Zeenea autour de la construction et l'enrichissement de catalogues de données est également en cours (deux stages de Master et un contrat CIFRE en cours de soumission).

Référence 2. L'unité développe des produits à destination du monde culturel, économique et social.

Le résultat de la collaboration avec SAP France a été intégré dans un prototype du produit SAP Data Intelligence et les premiers résultats de la collaboration avec Zeenea ont été intégrés dans le produit Zeenea Studio. L'équipe interagit également avec le comité de standardisation d'un langage de schémas JSON pour les choix adoptés dans les nouvelles versions, notamment pour lever des ambiguïtés relevées par l'étude du Draft de la dernière version.

Référence 3. L'unité partage ses connaissances avec le grand public et intervient dans des débats de société.

Formation initiale. L'équipe est très investie dans la formation initiale en Informatique à Sorbonne Université. Elle est responsable et assure majoritairement les enseignements des six cours en Licence et Master Informatique de Sorbonne Université : un cours en L2 (BD - Bases de Données) avec 380 inscrits et un cours en L3 (SGBD - Systèmes de Gestion de Bases de Données) avec 90 inscrits, deux cours en M1 (MLBDA - Modèles et Bases Avancées, SAM - Stockage et Accès aux Mégadonnées) et deux cours en M2 (BDLE - Bases de Données Large Échelle, LODAS - Linked Open Data, Apprentisssage Symbolique) dans le parcours DAC du Master Informatique (page web). Un membre de l'équipe est également *coresponsable du parcours DAC* depuis sa création en 2014.

Formation continue. Les Sciences de Données sont de plus en plus dans le cœur de nombreux secteurs informatiques avec un besoin important de former les professionnels sur les dernières technologies Big Data. L'équipe

participe depuis plusieurs années à la formation continue *Machine Learning et Intelligence Artificielle* (MLIA, deux jours par an lien) et, en 2023, à une formation continue pour le *projet Sèmè City* au Bénin (lien). Deux membres de l'équipe ont également participé à la formation de *professeurs du Secondaire en Informatique* et à la formation de préparation à l'*Agrégation en Informatique* créé en 2021 à Sorbonne Université.

Médiation scientifique. Deux membres de l'équipe ont contribué au livre "Les Big Data à Découvert" édité par le CNRS.

4 RÉFÉRENCES BIBLIOGRAPHIQUES SIGNIFICATIVES DE BD

[Attouche et al., 2022a] Attouche, L., Baazizi, M.-A., Colazzo, D., Ghelli, G., Sartiani, C., and Scherzinger, S. (2022a). Witness Generation for JSON Schema. *Proceedings of the VLDB Endowment (PVLDB)*, 15(13) :4002–4014.

[Baazizi et al., 2017a] Baazizi, M.-A., Ben Lahmar, H., Colazzo, D., Ghelli, G., and Sartiani, C. (2017a). Schema Inference for Massive JSON Datasets. In *Extending Database Technology (EDBT)*, Venise, Italy.

[Baazizi et al., 2020a] Baazizi, M.-A., Berti, C., Colazzo, D., Ghelli, G., and Sartiani, C. (2020a). Human-in-the-Loop Schema Inference for Massive JSON Datasets. In *EDBT 2020 - 23nd International Conference on Extending Database Technology*, pages 635–638, Copenhagen, Denmark. OpenProceedings.org.

[Baazizi et al., 2019a] Baazizi, M.-A., Colazzo, D., Ghelli, G., and Sartiani, C. (2019a). A Type System for Interactive JSON Schema Inference (Extended Abstract). In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 101 :1–101 :13, Patras, Greece.

[Baazizi et al., 2019b] Baazizi, M.-A., Colazzo, D., Ghelli, G., and Sartiani, C. (2019b). Parametric schema inference for massive JSON datasets. *The VLDB Journal*, 28(4) :497–521.

[Baazizi et al., 2020b] Baazizi, M.-A., Colazzo, D., Ghelli, G., Sartiani, C., and Scherzinger, S. (2020b). Not Elimination and Witness Generation for JSON Schema (short version). In *36ème Conférence sur la Gestion de Données – Principes, Technologies et Applications.*, Paris, France.

[Baazizi et al., 2021] Baazizi, M.-A., Colazzo, D., Ghelli, G., Sartiani, C., and Scherzinger, S. (2021). An Empirical Study on the "Usage of Not" in Real-World JSON Schema Documents. In *40th International Conference on Conceptual Modeling ER 2021*, Virtual conference, Canada.

[Debure et al., 2020a] Debure, J., Brunessaux, S., Constantin, C., and Du Mouza, C. (2020a). A pattern-based Approach for an Early Detection of Popular Twitter Accounts. In *International Database Engineering & Applications Symposium (IDEAS)*, pages 1 – 9, Séoul, France.

[Debure et al., 2021] Debure, J., Brunesseaux, S., Constantin, C., and Du Mouza, C. (2021). An Interaction Profile-based Classification for Twitter Users. In *International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA) 2021*, pages 21–25, Valence, Spain.

[dos Santos et al., 2018] dos Santos, L., Piwowarski, B., Denoyer, L., and Gallinari, P. (2018). Representation Learning for Classification in Heterogeneous Graphs with Application to Social Networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(5) :1 – 33.

[dos Santos et al., 2017] dos Santos, L., Piwowarski, B., and Gallinari, P. (2017). Gaussian Embeddings for Collaborative Filtering. In *40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1065–1068, Tokyo, Japan. ACM.

[Fopa et al., 2022b] Fopa, M., Gueye, M., Ndiaye, S., and Naacke, H. (2022b). A parameter-free KNN for rating prediction. *Data and Knowledge Engineering*.

[Griesner et al., 2018a] Griesner, J.-B., Abdessalem, T., Naacke, H., and Dosne, P. (2018a). ALGeoSPF : A Hierarchical Factorization Model for POI Recommendation. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 87–90, Barcelona, Spain.

[Grossetti et al., 2018] Grossetti, Q., Constantin, C., Du Mouza, C., and Travers, N. (2018). An Homophily-based Approach for Fast Post Recommendation in Microblogging Systems. In *21st International Conference on Extending Database Technology (EDBT 2018)*, pages 229–240, Vienne, Austria.

[Grossetti et al., 2021] Grossetti, Q., Du Mouza, C., Travers, N., and Constantin, C. (2021). Reducing the filter bubble effect on Twitter by considering communities for recommendations. *International Journal of Web Information Systems*, 17(6) :728–752.

[Gueye et al., 2020b] Gueye, I., Naacke, H., and Gançarski, S. (2020b). Enriching Geolocalized Dataset with POIs Descriptions at Large Scale. In *Innovations and Interdisciplinary Solutions for Underserved Areas*, volume 321 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 264–273. Springer International Publishing.

[Hannou et al., 2019b] Hannou, F.-Z., Amann, B., and Baazizi, M.-A. (2019b). Explaining Query Answer Completeness and Correctness with Partition Patterns. In *30th International Conference on Database and Expert Systems Applications - DEXA 2019*, volume 11707 of *Lecture Notes in Computer Science*, pages 47–62, Linz, Austria.

[Hannou et al., 2019c] Hannou, F.-Z., Amann, B., and Baazizi, M.-A. (2019c). Exploring and Comparing Table Fragments With Fragment Summaries. In *Eleventh International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA)*, pages 31–38, Athènes, Greece. IARIA.

[Li et al., 2021] Li, K., Naacke, H., and Amann, B. (2021). An Analytic Graph Data Model and Query Language for Exploring the Evolution of Science. *Big Data Research*, 26 :100247.

[Li et al., 2017] Li, Y., Constantin, C., and Du Mouza, C. (2017). SGVCut : A Vertex-Cut Partitioning Tool for RandomWalks-based Computations over Social Network graphs. In *International Conference on Scientific and Statistical Database Management, SSDBM*, pages 39 :1–39 :4, Chicago, United States.

[Liu et al., 2020] Liu, R., Simon, E., Amann, B., and Gançarski, S. (2020). Discovering and merging related analytic datasets. *Information Systems*, 91 :101495.

[Naacke et al., 2017a] Naacke, H., Amann, B., and Curé, O. (2017a). SPARQL Graph Pattern Processing with Apache Spark. In *GRADES (Graph Data-management Experiences & Systems), Workshop, SIGMOD 2017*, pages 1–7, Chicago, United States.

[Ren et al., 2017] Ren, X., Curé, O., Naacke, H., Lhez, J., and Li, K. (2017). Strider R : Massive and Distributed RDF Graph Stream Reasoning. In *IEEE International Conference on Big Data, Big Data 2017*, pages 3358–3367, Boston, United States. IEEE.

[Ren et al., 2018a] Ren, X., Curé, O., Naacke, H., and Xiao, G. (2018a). BigSR : real-time expressive RDF stream reasoning on modern Big Data platforms. In *IEEE International Conference on Big Data*, pages 811–820, Seatle, WA, United States.

[Scialom et al., 2019a] Scialom, T., Lamprier, S., Piwowarski, B., and Staiano, J. (2019a). Answers Unite ! Unsupervised Metrics for Reinforced Summarization Models. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, ACL Anthology, pages 3237–3247, Hong Kong, China. Association for Computational Linguistics.

[Simon et al., 2023b] Simon, E., Amann, B., Liu, R., and Gançarski, S. (2023b). Controlling the correctness of aggregation operations during sessions of interactive analytic queries. *J. Data and Information Quality*.

[Zablocki et al., 2018b] Zablocki, É., Piwowarski, B., Soulier, L., and Gallinari, P. (2018b). Learning Multi-Modal Word Representation Grounded in Visual Context. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, United States.

A ANNEXE — MEMBRES PERMANENTS AU 31/12/2022

La table ci dessous liste les membres permanents de l'équipe BD.

NOM	Prénom	Corps	Employeur
AMANN	Bernd	PR	Sorbonne Université
BAAZIZI	Mohamed-Amine	MCF	Sorbonne Université
CONSTANTIN	Camélia	MCF	Sorbonne Université
GANÇARSKI	Stéphane	MCF (HDR)	Sorbonne Université
NAACKE	Hubert	MCF (HDR)	Sorbonne Université

ÉLÉMENT DE PORTFOLIO 01

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément :

Discovering and merging related analytic datasets

URL de l'élément : `https://hal.sorbonne-universite.fr/hal-02459098v1`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Cet article est le résultat d'une première collaboration importante avec la société SAP France dans le cadre de la thèse CIFRE de Rutian Liu [3] et une publication dans le journal Information Systems [4]. Les résultats pratiques de ce travail ont été développés avec le logiciel SAP HANA pour ensuite être intégré dans le produit SAP Data Intelligence. Cette collaboration a permis de continuer une collaboration sur la qualité de requêtes analytiques [5] et sur la parallélisation de data pipelines dans Data Intelligence (un stage M2 en 2022 et deux stages M2 en cours en 2023).

3 PRÉSENTATION DE CET ÉLÉMENT

La croissance importante des sources de données générées rend leur utilisation de plus en plus complexe. Par exemple, la société SAP fournit à ces clients des milliers de jeux de données analytiques prédéfinis et personnalisables pour divers domaines d'applications industrielles et commerciales (gestion de la chaîne logistique, gestion de la relation client, planification de ressources d'entreprises). Ces jeux de données sont définis comme des vues analytiques (multidimensionnelles) sur les données transactionnelles stockées et gérées par la suite SAP S4/HANA et contiennent des informations avec des mesures sophistiquées prêtes à être utilisées par les utilisateurs.

Des outils de préparation des données permettent aux utilisateurs de créer leurs propres ensembles de données analytiques de haute qualité à partir des données transactionnelles et des données analytiques existants, et de créer des visualisations de données personnalisées interactives et perspicaces. Un besoin typique des utilisateurs de ces outils est d'augmenter le schéma d'une table de données analytiques existant avec de nouveaux attributs, provenant d'un ou plusieurs jeux de données sémantiquement liés, qui peuvent représenter des détails supplémentaires sur les dimensions ou de nouvelles mesures. Il s'agit d'un besoin essentiel dans de nombreux scénarios tels que la création de tableaux de bord ou la préparation de données pour apprendre des modèles prédictifs. Malgré l'importance de cette tâche, les outils actuels offrent peu d'aide aux utilisateurs pour créer de nouvelles tables analytiques. Ce manque de soutien conduit les utilisateurs à redéfinir à plusieurs reprises des ensembles de données analytiques similaires d'une manière parfois incohérente ou à faire appel à leur service informatique pour créer leurs ensembles de données personnalisés.

Cet article étend les solutions existantes sur l'augmentation de tables relationnelles [1, 2, 6] au cas de tables de données analytiques. Le défi principal dans l'augmentation du schéma d'une table analytique est de respecter des contraintes imposées par les dimensions et les mesures présentes dans les tables utilisées et les propriétés des fonctions d'agrégation. Ainsi, une requête de jointure simple peut introduire des incohérences dans la table augmentée en produisant plusieurs lignes par clé (row multiplication) et des mesures agrégées incorrectes ou ambigües (incorrect / ambiguous reduction). L'article présente un modèle d'augmentation de données analytiques avec des critères de qualité formels pour les tables augmentées par des jointures avec d'autres tables analytiques. Ces critères sont e utilisés (1) pour identifier automatiquement des opérations de réparation, (b) pour notifier aux utilisateurs la génération d'attributs ambigus, (c) pour déduire des fonctions d'agrégation applicables sur les nouveaux attributs, et (d) pour compléter les résultats de fusion obtenus par une augmentation de schéma incomplète. La solution proposée inclut une analyse théorique minutieuse et une implémentation dans le système SAP HANA.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Halevy, Hongrae Lee, Fei Wu, Reynold Xin, and Cong Yu. Finding Related Tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 817–828, New York, NY, USA, 2012. ACM.

[2] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, and Nan Tang. The Data Civilizer System. In *CIDR*, 2017.

[3] Rutian Liu. *Semantic Services for Assisting Users to Augment Data in the Context of Analytic Data Sources*. Theses, Sorbonne Université, June 2020.

[4] Rutian Liu, Eric Simon, Bernd Amann, and Stéphane Gançarski. Discovering and merging related analytic datasets. *Information Systems*, 91 :101495, July 2020.

[5] Eric Simon, Bernd Amann, Rutian Liu, and Stéphane Gançarski. Controlling the Correctness of Aggregation Operations During Sessions of Interactive Analytic Queries. *ACM Journal of Data and Information Quality*, January 2023.

[6] Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. InfoGather : Entity augmentation and attribute discovery by holistic matching with web tables. In *Proceedings of the 2012 International Conference on Management of Data - SIGMOD '12*, pages 97–108. ACM Press, 2012.

ÉLÉMENT DE PORTFOLIO 02

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément :

An Homophily-based Approach for Fast Post Recommendation in Microblogging Systems

Fichier de élément :

`https://hal.sorbonne-universite.fr/hal-01679120v1`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Cet travail est le résultat d'une collaboration avec le laboratoire CEDRIC du CNAM Paris dans le cadre de la thèse de Quentin Grossetti [2]. Il a donné lieu à une publication dans la conférence EDBT 2018 [3]. Ce travail a été poursuivi avec la thèse de Jonathan Debure [1], et celle de Yifan Li [4], également co-encadrées avec le CEDRIC, pour analyser le comportement des utilisateurs dans les réseaux sociaux et réaliser le passage à l'échelle d'algorithmes de calculs utilisés dans les systèmes de recommandation.

3 PRÉSENTATION DE CET ÉLÉMENT

Les services de micro-blogging sont devenus récemment une source d'information importante. Cependant, victimes de leur succès, ils doivent actuellement gérer une quantité sans précédent d'informations générées par les utilisateurs. Il devient par conséquent difficile pour les utilisateurs de trouver dans ces services des contenus proches de leurs intérêts. Nous nous sommes intéressés dans ce travail au problème de la recommandation des utilisateurs à suivre sur un sujet donné. Pour cela nous avons défini des scores basés sur la topologie du graphe social et sur le contenu textuel des micro-blogs. Après une étude approfondie d'un large jeu de donnée issu de Twitter, nous nous appuyons sur ces conclusions concernant le fonctionnement et les usages de Twitter pour définir un modèle de propagation qui repose sur l'homophilie présente dans le réseau pour recommander des messages aux utilisateurs. Nous avons montré que la durée de vie des messages était extrêmement courte et qu'il était par conséquent nécessaire de recommander des messages rapidement après leur publication. Notre approche s'appuie sur la construction d'un graphe de similarités lié aux interactions des utilisateurs. Ces similarités sont cependant assez faibles compte tenu du faible nombre d'interactions en commun entre utilisateurs, il est donc crucial d'être en mesure de surmonter ce manque d'information. Nous construisons un réseau de similarités permettant de mettre à profit facilement la transitivité entre utilisateurs. Cette transitivité est la clé pour recommander des messages sur lesquels on ne dispose que de peu d'interactions. Notre méthode de recommandation est capable de recommander rapidement des messages après leur publication et de passer à l'échelle. Pour ce faire, nous nous appuyons sur le pré-calcul de recommandations pour certains comptes choisis dans le graphe.

Nous présentons plusieurs expérimentations pour démontrer la qualité de prédiction de notre modèle et sa capacité à passer à l'échelle. En comparant les résultats avec d'autres modèles de recommandation nous pouvons analyser le fonctionnement standard de ces méthodes avec leurs forces et leurs faiblesses. Certaines méthodes ne demandent aucun travail d'initialisation et peuvent facilement se mettre à jour mais ont tendance à recommander des messages populaires du fait des marches aléatoires. Certaines méthodes sont capables de fournir un nombre très élevé de recommandations aux utilisateurs mais ne disposent pas d'un excellent ratio précision/rappel, signe que les messages recommandés ne sont pas toujours pertinents. En revanche, le coût de calcul d'initialisation est trop élevé et ne permet pas d'envisager ce modèle de façon réaliste pour une plateforme comme Twitter. D'autres méthodes ont des coûts d'initialisation sont négligeables tandis que le coût de propagation est très élevé, ce qui limite également l'utilisation de ce type de modèle avec de larges plateformes de micro-blogging. Nous montrons également que les scores calculés par notre méthode peuvent être mis à jour à faible coût tout en proposant des recommandations pertinentes.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Jonathan Debure. *Détection de comportements et identification de rôles dans les réseaux sociaux*. Theses, Conservatoire national des arts et metiers - CNAM, July 2021.

[2] Quentin Grossetti. *Système de recommandation sur les plateformes de micro-blogging et bulles filtrantes*. Theses, Sorbonne Université, November 2018.

[3] Quentin Grossetti, Camelia Constantin, Cédric Du Mouza, and Nicolas Travers. An Homophily-based Approach for Fast Post Recommendation in Microblogging Systems. In *21st International Conference on Extending Database Technology (EDBT 2018)*, pages 229–240, Vienne, Austria, March 2018.

[4] Yifan Li. *Edge partitioning of large graphs*. Theses, Université Pierre et Marie Curie - Paris VI, December 2017.

ÉLÉMENT DE PORTFOLIO 03

Vidéo

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : Inférence Interactive de Schema JSON

URL de l'élément : `https://dropsu.sorbonne-universite.fr/s/7nf6RQz7FTRR8zK`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Il s'agit d'une vidéo accompagnant la démonstration d'un outil présenté dans une conférence internationale [1]. Cette démonstration est l'implantation de l'approche d'inférence interactive de schéma dont les fondements théoriques ont été présentés dans [2]. Elle s'appuie également sur un effort d'implantation ayant permis la réalisation des expérimentations des publications décrivant l'inférence de schémas paramétriques [3] qui a suscité beaucoup d'intérêt auprès de la communauté puisque de nombreuses extensions ont récemment été publiées, les plus importantes étant [5] et [4] publiés dans les conférences majeurs du domaine (SIGMOD et EDBT).

3 PRÉSENTATION DE CET ÉLÉMENT

JSON est le format de données semi-structurées le plus utilisé. Il tient son succès de son expressivité et de sa flexibilité puisqu'il permet de décrire des données complexes sans définir de schéma préalablement. Cette liberté s'accompagne de nombreux défis lorsqu'il s'agit d'exploiter de manière correcte et efficace les données représentées dans ce format qui sont souvent fortement imbriquées et volumineuses. L'inférence de schéma *a posteriori* est une solution attractive pour exploiter de manière efficace ces données puisqu'elle permet d'extraire une description succincte de la structure sous-jacente des données et ainsi de formuler des requêtes et des programmes d'extraction de données sensés.

Or, une même collection JSON peut être décrite suivant différents niveaux d'abstraction et différents usages font appels à différents niveaux de précision. Il devient alors plus judicieux d'offrir la possibilité de choisir de manière fine le niveau de précision souhaité pour chaque fragment de données, ce qui, d'un point de vue théorique, revient à définir une technique d'inférence non déterministe, guidée par l'interaction avec l'utilisateur et efficace en présence de grands volumes de données.

Pour répondre à ces contraintes, nous avons proposé une approche interactive s'appuyant sur des techniques de réécriture d'arbres [2] et garantissant, de manière formelle, la correction des schémas obtenus par simple réécriture de schémas plus précis. L'avantage est de pouvoir utiliser seulement les schémas, pendant l'interaction, sans avoir à accéder aux données qui sont souvent volumineuses et empêcheraient toute interaction fluide avec l'utilisateur.

La réalisation logicielle présentée dans [1] exploite cette technique de réécriture et démontre la faisabilité de l'approche sur des schémas inférés de données réelles de taille importante et présentant de nombreuses variations structurelles qui tendent à confirmer l'utilité de l'approche dans le cas pratique.

Une extension récente visant à générer automatiquement des requêtes à partir de l'interaction de l'utilisateur a été mise en œuvre et a permis de démontrer l'utilité d'une approche de découverte des données, guidée par les schémas.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Mohamed-Amine Baazizi, Clément Berti, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Human-in-the-Loop Schema Inference for Massive JSON Datasets. In *EDBT 2020 - 23nd International Conference on Extending Database Technology*, pages 635–638, Copenhagen, Denmark, March 2020. OpenProceedings.org.

[2] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. A Type System for Interactive JSON Schema Inference (Extended Abstract). In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 101 :1–101 :13, Patras, Greece, July 2019.

[3] Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Parametric schema inference for massive JSON datasets. *The VLDB Journal*, 28(4) :497–521, August 2019.

[4] Hanâ Lbath, Angela Bonifati, and Russ Harmer. Schema inference for property graphs. In *EDBT 2021-24th International Conference on Extending Database Technology*, pages 499–504, 2021.

[5] William Spoth, Oliver Kennedy, Ying Lu, Beda Hammerschmidt, and Zhen Hua Liu. Reducing ambiguity in json schema discovery. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1732–1744, 2021.

ÉLÉMENT DE PORTFOLIO 04

Vidéo

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : EPIQUE : Extracting Meaningful Science Evolution Patterns from Large Document Archives

URL de l'élément : `http://nextcloud.iscpif.fr/index.php/s/KBJk5MEyyoweQcp`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Cette video a été produite dans le cadre du projet ANR interdisciplinaire EPIQUE [1] (2017-2021) porté par l'équipe BD et avec trois partenaires académiques : l'Institut des Systèmes Complexes (ISC-PIF), l'IRISA de Rennes et une UMR en philosophie des sciences de Paris 1 (IHPST). Elle est un des résultats importants de la thèse de Ke LI (2018-2021) [1]. Ce travail donné lieu à trois publications internationales dont une revue.

Le projet Epique a permis à l'équipe BD d'acquérir de nouvelles compétences en fouille de données textuelles et ainsi de s'ouvrir à de nouveaux axes de recherche mais en les abordant, de manière différente, à travers un prisme "BD" qui cible en priorité la gestion efficace des données accompagnée de langages hautement expressifs pour les manipuler. Il a ouvert des perspectives intéressantes qui sont actuellement explorées dans la thèse de Hamed Rahimi (démarrée en 2021).

3 PRÉSENTATION DE CET ÉLÉMENT

Le projet Epique part du constat qu'il existe une demande croissante d'outils pratiques pour explorer l'évolution de la recherche scientifique publiée dans des archives bibliographiques telles que le Web of Science (WoS), arXiv, PubMed ou ISTEX. Mettre en évidence des modèles d'évolution à partir de ces archives documentaires est un besoin exprimé dans de nombreux cas d'usage, par exemple en histoire des sciences ou dans la gouvernance de l'activité scientifique.

Le projet Epique vise à reconstruire l'évolution des sciences dans le temps, en utilisant des méthodes quantitatives pour analyser des grands corpus textuels de publications scientifiques. Il s'agit de produire une *cartographie* des sciences au fil du temps. Les domaines, représentés par des mots-clés, sont positionnés par période croissante : des plus anciennes en haut de la carte aux plus récentes en bas de la carte. Pour représenter le changement progressif du contenu d'un domaine, des liens connectent les domaines similaires appartenant à deux périodes consécutives. La structure de la carte est un graphe complexe : il contient des motifs d'évolution qu'il s'agit de mettre en évidence. Une des nouveautés du projet est de rendre les cartes **interactives** afin d'augmenter les possibilités d'exploration et à terme découvrir des nouvelles formes d'évolution des sciences. Un modèle pour caractériser des motifs d'évolution a été défini ainsi qu'un langage déclaratif permettant d'interroger des graphes d'évolution. Le langage a l'avantage d'être composable afin de pouvoir rechercher des motifs en exprimant successivement plusieurs requêtes.

Ce travail a fait l'objet d'une publication dans la revue Big Data Research [3] et une autre dans le workshop BigVis@EDBT [2].

De plus, nous avons abordé un problème de performance soulevé lors de l'alignement temporel des domaines scientifiques représentés par des vecteurs. Un algorithme de calcul parallèle et distribué de similarité entre les domaines a été conçu et publié dans la conférence IEEE BigData [4].

Actuellement nous poursuivons les travaux sur l'analyse de l'évolution des sciences en nous appuyant sur les avancées récentes dans les domaines des modèles de langues et des embeddings de graphe. Les données textuelles des documents scientifiques sont combinées avec des graphes de citation pour représenter l'influence des communautés scientifiques sur les évolutions thématiques.

1. http ://www-bd.lip6.fr/wiki/site/recherche/projets/epique/start

Enfin, notre expertise en gestion efficace de grands volumes de données est mise à profit pour concevoir des solutions d'optimisation de chaînes de traitement manipulant à la fois des masses de données textuelles et des grands graphes.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Ke Li. *Exploring Topic Evolution in Large Scientific Archives with Pivot Graphs*. Theses, Sorbonne Université, June 2021.

[2] Ke Li, Hubert Naacke, and Bernd Amann. Exploring the Evolution of Science with Pivot Topic Graphs. In *International Workshop on Big Data Visual Exploration and Analytics BigVis at EDBT 2020*, Copenhague, Denmark, March 2020.

[3] Ke Li, Hubert Naacke, and Bernd Amann. An Analytic Graph Data Model and Query Language for Exploring the Evolution of Science. *Big Data Research*, 26 :100247, November 2021.

[4] Hubert Naacke, Ke Li, Bernd Amann, and Olivier Curé. Efficient similarity-based alignment of temporally-situated graph nodes with Apache Spark. In *High Performance Big Graph Data Management, Analysis, and Mining*, IEEE Int'l Conference on Big Data, Los Angeles, United States, December 2019.