جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THÈSE DE DOCTORAT

Monsieur Saâd BELEFQIH

Soutiendra publiquement sa thèse de Doctorat en Informatique

Le Samedi 18 Juillet 2026 à 10h00 au Grand Amphi à l'ENSIAS de Rabat

Intitulé de la thèse

A Contribution to Transformer-Based Semantic Schema Extraction from Schemaless Data Sources

Président :

Pr. Mohamed LAZAAR, PES, École Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Université Mohammed V de Rabat, Rabat.

Directeur de thèse :

Pr. Ahmed ZELLOU, PES, École Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Université Mohammed V de Rabat, Rabat.

Rapporteurs :

Pr. Hassan SILKAN, PES, Faculté des Sciences, Université Chouaib Doukkali, El Jadida.

Pr Khalid HOUSNI, PES, Faculté des Sciences, Université Ibn Tofail, Kénitra.

Pr. Soumia ZITI, PES, Faculté des Sciences, Université Mohammed V de Rabat, Rabat.

Examinateur(s) :

Pr. Noureddine FALIH, MCH, Faculté Polydisciplinaire, Université Sultan My Slimane, Béni Mellal.

Pr. Asmae RETBI, MCH, Ecole Mohammadia d'Ingénieurs, Université Mohammed V de Rabat, Rabat.

جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

Résumé:

Les sources de données sans schéma sont devenues centrales dans les systèmes d'information modernes, car elles permettent le stockage et la gestion de données hétérogènes, incomplètes et évolutives. Les bases de données NoSQL et les représentations semi-structurées offrent ainsi une flexibilité importante lorsque la structure des données est variable ou non entièrement connue à l'avance. Cependant, cette flexibilité rend la structure moins visible, puisque le schéma n'est pas explicitement déclaré avant le stockage, mais demeure réparti entre les instances stockées. Par conséquent, les entités, les propriétés, les types de données, les structures imbriquées, les relations, les contraintes et les variations structurelles doivent être inférés avant que les données puissent être efficacement comprises, interrogées, validées, intégrées et maintenues.

Cette thèse développe trois contributions complémentaires. La première contribution propose une approche fondée sur l'Intelligence Artificielle (IA) pour l'extraction sémantique de schémas dans les bases de données NoSQL. Elle transforme les données JSON en représentations sous forme de triplets et applique des modèles d'embeddings contextuels afin de soutenir l'analyse de similarité sémantique et la consolidation du schéma. Cette contribution évolue d'une approche basée sur BERT vers un cadre orienté RDF utilisant Sentence-BERT, des contraintes sémantiques, l'optimisation du schéma et la visualisation. L'objectif est de produire des schémas à la fois informatifs sur le plan structurel, significatifs sur le plan sémantique et réutilisables.

La deuxième contribution introduit le Schema Validation and Evaluation Framework pour les schémas extraits des bases de données JSON. Ce cadre répond au besoin d'évaluer les schémas inférés à l'aide de critères explicites et reproductibles. Il évalue la qualité du schéma selon plusieurs dimensions, notamment l'exactitude des types de données, les champs obligatoires et optionnels, la prise en charge des types multiples, la cohérence de la structure des collections, la récupération des relations entre entités et la détection de l'évolution temporelle. Ces dimensions fournissent une base structurée pour déterminer si un schéma extrait reflète les données sous-jacentes de manière exacte et cohérente.

La troisième contribution propose une approche fondée sur l'apprentissage par renforcement pour l'évolution dynamique des schémas RDF dans les bases de données NoSQL. L'évolution du schéma est modélisée comme un processus décisionnel séquentiel dans lequel les triplets compatibles avec RDF sont traités au moyen des actions Add, Merge, Modify et Ignore. L'apprentissage par renforcement guide ces décisions selon la cohérence sémantique, le contrôle de la redondance et la préservation des contraintes, permettant ainsi au schéma d'évoluer progressivement lorsque la source de données change.

Les contributions proposées sont évaluées à travers des scénarios de données sans schéma. L'approche initiale d'extraction sémantique a atteint une précision de 1,0, une exactitude de 0,833 et un score F1 de 0,909, tandis que l'approche améliorée orientée RDF a atteint une cohérence de schéma de 100 % et a réduit la redondance de 13,33

%. Pour l'évolution dynamique des schémas, l'évaluation sur 100 000 documents a montré des performances stables, avec des scores macro-F1 allant de 0,903 à 0,917 et des valeurs d'exactitude équilibrée allant de 0,907 à 0,919. Ces résultats indiquent que les contributions proposées soutiennent l'extraction de schémas, la cohérence sémantique, la cohérence structurelle, l'évaluation de la qualité des schémas et leur évolution adaptative dans les conditions évaluées.

Mots-clés: Extraction de schémas; Sources de données sans schéma; Schéma RDF; Évaluation de schémas; Évolution de schémas; Transformer; BERT; SBERT; Apprentissage par renforcement

Abstract:

Schema-less data sources have become central to modern information systems because they support the storage and management of heterogeneous, incomplete, and evolving data. NoSQL databases and semi-structured representations provide important flexibility when data structures are variable, partially known, or subject to change over time. However, this flexibility also reduces the visibility of the data structure, since the schema is not formally specified before storage but remains distributed across the stored instances. As a result, entities, properties, data types, nested structures, relationships, constraints, and structural variations must be recovered before the data can be effectively understood, queried, validated, integrated, and maintained.

The thesis develops three complementary contributions. The first contribution proposes an AI-based approach for semantic schema extraction in NoSQL databases. It transforms JSON-based data into triplet-based representations and applies contextual embedding models to support semantic similarity analysis and schema consolidation. This contribution evolves from a BERT-based extraction approach toward an RDF-oriented framework using Sentence-BERT, semantic constraints, schema optimization, and visualization. The objective is to produce schemas that are structurally informative, semantically meaningful, and reusable.

The second contribution introduces the Schema Validation and Evaluation Framework for extracted schemas in JSON databases. This framework addresses the need to assess inferred schemas through explicit and reproducible criteria. It evaluates schema quality according to several dimensions, including data type accuracy, required and optional fields, multiple type support, collection structure consistency, entity relationship recovery, and temporal evolution detection. Through these dimensions, the framework provides a structured basis for determining whether an extracted schema accurately and consistently reflects the underlying data.

The third contribution proposes a reinforcement learning-based approach for dynamic RDF schema evolution in NoSQL databases. In this approach, schema evolution is modeled as a sequential decision-making process in which incoming RDF-compatible triplets are handled through Add, Merge, Modify, and Ignore actions. Reinforcement learning guides these decisions according to semantic coherence, redundancy control, and constraint preservation, allowing the schema to evolve progressively as the data source changes.

The proposed contributions were assessed through schema-less data scenarios. The initial semantic extraction approach achieved a precision of 1.0, an accuracy of 0.833, and an F1-score of 0.909, while the enhanced RDF-oriented approach reached 100% schema coherence and reduced redundancy by 13.33%. For dynamic schema evolution, experiments conducted at 100,000 documents showed stable performance, with macro-F1 scores ranging from 0.903 to 0.917 and balanced accuracy values ranging from 0.907 to 0.919 across the evaluated datasets. These results indicate that the proposed contributions support schema recovery, semantic coherence, structural consistency, schema quality assessment, and adaptive schema evolution under the evaluated conditions.

Keywords: Schema Extraction; Schema-Less Data Sources; RDF Schema; Schema Evaluation; Schema Evolution; Transformer; BERT; SBERT; Reinforcement Learning.