UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR

AVIS DE SOUTENANCE DE THÈSE

Irvin DONGO ESCALANTE

CANDIDAT(E) au DOCTORAT INFORMATIQUE,
à **L'UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR**
SOUTIENDRA PUBLIQUEMENT sa THÈSE

le **20 décembre 2017** à **14h30**
à L'UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR
IUT de BAYONNE à ANGLET

SUR LE SUJET SUIVANT :

"Anonymisation des documents RDF"

JURY :

José AGUILAR, Professeur, UNIVERSITÉ DES ANDES (VENEZUELA)
Béchara AL BOUNA, Directeur de Recherche, UNIVERSITÉ ANTONINE (LIBAN)
Firas AL KHALIL, Directeur de Recherche, UNIVERSITY COLLEGE DE CORK (IRLANDE, ou EIRE)
Yudith CARDINALE, Professeur, UNIVERSITÉ SIMON BOLIVAR DE CARACAS (VENEZUELA)
Richard CHBEIR, Professeur des Universités, UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR
Alban GABILLON, Professeur des Universités, UNIVERSITÉ POLYNÉSIE FRANÇAISE
Sebastien LABORIE, Maître de Conférences, UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR
Said TAZI, Maître de Conférences, UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR

Pau, le 06 décembre 2017

Le Président et,
Par délégation, la Vice-Présidente de la Commission de la Recherche

Isabelle BARAILLE

DONGO ESCALANTE Irvin Franco Benito

Titre français de la thèse : Anonymisation de documents RDF - Towards RDF Anonymization

Résumé en français (chaque résumé ne doit pas dépasser 4000 caractères (espace et ponctuation compris) :

Avec l'avancée du Web Sémantique et des initiatives Open Linked Data, une grande quantité de documents RDF sont disponibles sur Internet. L'objectif est de rendre ces données lisibles pour les humains et les machines, en adoptant des formats spéciaux et en les connectant à l'aide des IRIs (International Resource Identifier), qui sont des abstractions de ressources réelles du monde. L'augmentation du nombre de données publiées et partagées augmente également le nombre d'informations sensibles diffusées. En conséquence, la confidentialité des entités d'intérêts (personnes, entreprises, etc.) est un véritable défi, nécessitant des techniques spéciales pour assurer la confidentialité et la sécurité adéquate des données disponibles dans un environnement où chaque utilisateur a accès à l'information sans aucune restriction (Web).

Ensuite, trois aspects principaux sont considérés pour assurer la protection de l'entité: (i) Préserver la confidentialité, en identifiant les données qui peuvent compromettre la confidentialité des entités (par exemple, les identifiants, les quasi-identifiants); (ii) Identifier l'utilité des données publiques pour diverses applications (par exemple, statistiques, tests, recherche); et (iii) Les connaissances antérieures du modèle qui peuvent être utilisées par les pirates informatiques (par exemple, le nombre de relations, une relation spécifique, l'information d'un nœud).

L'anonymisation est une technique de protection de la confidentialité qui a été appliquée avec succès dans les bases de données et les graphes. Cependant, les études sur l'anonymisation dans le contexte des documents RDF sont très limitées. Ces études sont les travaux initiaux de protection des individus sur des documents RDF, puisqu'ils montrent les approches pratiques d'anonymisation pour des scénarios simples comme l'utilisation d'opérations de généralisation et d'opérations de suppression basées sur des hiérarchies. Cependant, pour des scénarios complexes, où une diversité de données est présentée, les approches d'anonymisations existantes n'assurent pas une confidentialité suffisante.

Ainsi, dans ce contexte, nous proposons une approche d'anonymisation, qui analyse les voisins en fonction des connaissances antérieures, centrée sur la confidentialité des entités représentées comme des nœuds dans les documents RDF. Notre approche de l'anonymisation est capable de fournir une meilleure confidentialité, car elle prend en compte la condition de la diversité de l'environnement ainsi que les voisins (nœuds et arêtes) des entités d'intérêts. En outre, un processus d'anonymisation automatique est assuré par l'utilisation d'opérations d'anonymisations associées aux types de données.

Résumé en anglais :

With the advance of the Semantic Web and the Open Linked Data initiatives, a huge quantity of RDF data is available on Internet. The goal is to make this data readable for humans and machines, adopting special formats and connecting them by using International Resource Identifiers (IRIs), which are abstractions of real resources of the world. As more data is published and shared, sensitive information is also provided. In consequence, the privacy of entities of interest (e.g., people, companies) is a real challenge, requiring special techniques to ensure privacy and adequate security over data available in an environment in which every user has access to the information without any restriction (Web).

Then, three main aspects are considered to ensure entity protection: (i) Preserve privacy, by identifying and treating the data that can compromise the privacy of the entities (e.g., identifiers, quasi-identifiers); (ii) Identify utility of the public data for diverse applications (e.g., statistics, testing, research); and (iii) Model background knowledge that can be used for adversaries (e.g., number of relationships, a specific relationship, information of a node).

Anonymization is one technique for privacy protection that has been successfully applied in practice for databases and graph structures. However, studies about anonymization in the context of RDF data, are really limited. These studies are initial works for protecting individuals on RDF data, since they show a practical anonymization approach for simple scenarios as the use of generalization and suppression operations based on hierarchies. However, for complex scenarios, where a diversity of data is presented, the existing anonymization approaches does not ensure an enough privacy.

Thus, in this context, we propose an anonymization framework, which analyzes the neighbors according to the background knowledge, focused on the privacy of entities represented as nodes in the RDF data. Our anonymization approach is able to provide better privacy, since it takes into account the l-diversity condition as well as the neighbors (nodes and edges) of entities of interest. Also, an automatic anonymization process is provided by the use of anonymization operations associated to the datatypes.

Résumé autre langue (si le cas se présente) :

Mots clés en français :

RDF, Web Sémantique, Anonymisation, Datatypes

Mots clés en anglais :

RDF, Semantic Web, Anonymization, Datatypes

Mots clés autre langue (si le cas se présente) :