MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ ABDELHAMID IBN BADIS - MOSTAGANEM

Faculté des Sciences Exactes et d'Informatique

Département de Mathématiques et informatique

Filière : Informatique

RAPPORT DE MINI-PROJET

Option : **Ingénierie des Systèmes d'Information**

THÈME :

Edition Collaborative des données liées

(Linked Data)

Étudiants : « **Zaidi Houcine** et **Ouali Youcef**»

Encadrant : « Mechaoui Moulay Driss »

Année Universitaire 2021-2022

Résumé

Mots-clés :

Abstract

Keywords :

Table des figures

1.1 Un schéma de description défini en RDFS 5

Liste des tableaux

1.1 Par exemple : Soit la phrase simple suivante : la tour Eiffel est créée en 1887 a paris . Cette phrase est composée des parties suivantes 7
1.2 Résumé de comparaison entre les algorithmes 9

Liste des abréviations

CRDT Conflict-free Replicated Data Type

LOD Linked Open Data

RDF Resource Description Framework

RDFS Resource Description Framework Schema

Table des matières

Introduction Générale **2**

1 Web et données sémantique **3**
1.1 RDF . . . 3
1.2 RDFS . . . 4
1.3 LINKED DATA (DONNÉES LIÉES) . . . 5
1.3.1 LINKED OPEN DATA . . . 5
1.4 Edition des données sémantiques . . . 6
1.5 CRDT (type de donnée répliqué commutatif) . . . 6
1.6 Les algorithmes d'édition de données sémantique . . . 7
1.6.1 OR-Set . . . 7
1.6.2 C-SET [7] . . . 7
1.6.3 SU-Set . . . 8
1.6.4 B-SET [8] . . . 8
1.6.5 srCE . . . 8
1.6.6 LD-SET . . . 8
1.6.7 Le modèle CCI : . . . 8
1.7 Etude comparative entre ces algorithmes . . . 8

Conclusion **9**

Introduction Générale

Dans le Web sémantique et dans le Web en général, un problème fondamental est la comparaison et l'appariement des données et la capacité de résoudre la multiplicité des références de données aux mêmes objets du monde réel, en définissant les correspondances entre les données sous forme de liens de données. La tâche de couplage de données devient de plus en plus importante à mesure que le nombre de données structurées et semi-structurées disponibles sur le Web augmente. La transformation du Web d'un « Web de documents » en un « Web de données », ainsi que la disponibilité d'importantes collections de données générées par des capteurs (Internet des objets), mènent à une nouvelle génération d'applications Web fondées sur l'intégration des données et des services. Parallèlement, de nouvelles données sont publiées chaque jour à partir de contenus générés par les utilisateurs et de sites Web publics. mène à une nouvelle génération d'applications Web En termes généraux, le couplage de données est la tâche de déterminer si deux descriptions d'objets peuvent être liées l'une à l'autre pour représenter le fait qu'elles se réfèrent au même objet du monde réel dans un cas décrivant les objets du monde réel à travers des sources de données hétérogènes, en supposant que la similarité entre deux descriptions de données est plus élevée, plus grande est la probabilité que la liaison inclut également la tâche de définir des méthodes, des techniques et des outils (semi-)automatisés pour Dans ce contexte, l'une des initiatives les plus importantes dans le domaine du Web sémantique est le lien entre les grandes collections de données déjà disponibles sur le Web [1]. exemple de données liées montre comment la tâche de liaison de données est cruciale sur le Web de nos jours. les méthodes et les techniques de liaison des données sur le Web sémantique. En outre, il existe d'importants travaux décrivant des domaines de recherche très proches des caractéristiques de liaison de données sémantiques qui nécessitent des solutions spécifiques tant en termes de nouvelles techniques que dans Par exemple, d'une part, le couplage de données nécessite de traiter la complexité sémantique typique du couplage ontologique, mais d'autre part, la grande quantité de données disponibles du champ de couplage de données sur le Web sémantique. Dans ce document, nous donnons une définition générale de ce domaine, afin de souligner les problèmes et de décrire les solutions. Nous allons mieux définir le problème de liaison de données, en discutant également de nombreux algorithmes. L'objectif de ce projet est d'étudier les approches existantes en matière d'édition collaborative des données sémantiques, afin de faire une comparaison collaborative entre ces approches, et de proposer un éditeur collaboratif distribué pour les données sémantiques capables de supporter des groupes dynamiques où les utilisateurs peuvent se joindre et partir à tout moment.[2]

Chapitre 1

Web et données sémantique

Le Web sémantique Pour bien comprendre , nous commençons par une définition. Le grand dictionnaire terminologique définit le terme sémantique comme « l'ensemble des relations entre les caractères, ou groupes de caractères, et leur significations, indépendamment de la façon de les employer ou de les interpréter. » Il précise par la suite que « si, en linguistique, la sémantique porte sur l'étude du sens à partir de la combinaison des mots, en intelligence artificielle, elle porte sur la capacité d'un réseau [le Web] à représenter de la manière la plus humaine possible des relations entre des objets, des idées ou des situations. » Le terme sémantique implique donc que la machine ne se contentera plus de présenter visuellement les données du Web, mais, en les reliant, elle pourra conserver les significations qui leur sont attribuables. Or, en transformant le contenu du Web pour qu'il soit « compréhensible » par la machine et non seulement présentable, nous permettons à cette même machine d'être plus efficace dans le traitement de l'information. Ainsi, le dialogue avec les moteurs de recherche devient possible. Nous sommes alors en mesure de nous exprimer dans des termes que nos ordinateurs peuvent aussi interpréter et échanger. Il est également possible d'automatiser, d'intégrer et de réutiliser l'information entre diverses applicationsLe Web sémantique est décrit généralement comme un Web destiné aux machines. Disposer d'un Web dont le contenu est abordable par les machines peut apporter de grands bénéfices : L'automatisation de nombreuses tâches fondées sur le contenu comme la recherche de ressources ayant un contenu particulier, la comparaison du contenu de ressources (pages, bases de données, ontologies, etc...). Le Web sémantique permettrait de résoudre la relative difficulté de trouver de l'information sur le web.[3] Le Web sémantique a pour objectif de transformer le World Wide Web actuel, entièrement tourné vers la présentation des documents, vers un Web dont le contenu serait compréhensible par les machines. La vision s'appuie sur l'utilisation d'ontologies. Le Web sémantique n'est pas un Web séparé mais une extension du Web actuel, dans lequel l'information est bien définie, permettant ainsi aux ordinateurs et aux personnes de travailler en coopération. Le Web sémantique permettra aux machines de comprendre les documents et les données sémantiques, et non la parole humaine et les écrits.[3]

1.1 RDF

RDF est un acronyme de Resource Description Framework. RDF est une norme W3C des technologies Web sémantiques et la base de l'architecture standardisée des technologies Web Sémantique. Le RDF est un langage simple pour exprimer des modèles de données sous forme d'objets « ressources » et de leurs relations. Le RDF est un langage simple pour

exprimer des modèles de données sous forme d'objets « ressources » et de leurs relations. Il sera utilisé pour annoter des documents écrits dans des langages non structurés, ou comme une interface pour des documents écrits dans des langages ayant une sémantique équivalente (des bases de données, par exemple). Un document RDF est un ensemble de triplets de la forme <sujet, prédicat, objet>. Les éléments de ces triplets peuvent être des URIs (Universal Resource Identifiers), des littéraux ou des variables. Cet ensemble de triplets peut être représenté de façon naturelle par un graphe (plus précisément un multi-graphe orienté étiqueté), où les éléments apparaissant comme sujet ou objet sont les sommets, et chaque triplet est représenté par un arc dont l'origine est son sujet et la destination son objet .[3] RDF (Resource Description Framework), est un standard décrivant :

- des ressources, une ressource pouvant être n'importe quoi (personnes, lieux, animaux, documents, concepts, etc.) ;
- la description de ces ressources par des attributs et des relations ;
- le framework contenant un modèle de données, des langages et des syntaxes.

1.2 RDFS

Le schéma RDFS donne véritablement sa sémantique à la description RDF[4] Il permet aussi de définir un vocabulaire pouvant être utilisé pour décrire des ressources. On peut imaginer à loisir de nombreux vocabulaires différents, adaptés chacun à un domaine ou à une application spécifique. Notons que les vocabulaires, appelés aussi schémas de description, sont eux-mêmes écrits en RDF, en utilisant des balises de l'espace de nom RDFS(e.g, rdfs :Class, rdfs :subClassOf,rdfs :domain, rdfs :range,...). la figure 1 présente un schéma de description écrits en RDF. Deux de vocabulaire sont définis pour ce schéma. #**Personne**, #**Chercheur**, #**Doctorant** sont des classes de ressources d'un annuaire universitaire, #**Chercheur** et #**Doctorant** sont les sous classes de #**Personne**. #**nom** et #email sont des propriétés applicables aux ressources de la classe #**Personne** pour donner le nom et l'adresse email de chacune. #**sousDirection** est une propriété d'association entre #**Doctorant** et #**Chercheur**. La figure 1.1 illustre un example d'un shema RDFS.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<rdfs:Class rdf:ID="Personne"/>
<rdf:Property rdf:ID="nom">
   <rdfs:domain rdf:resource="#Personne"/>
   <rdfs:range
        rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Property>
<rdf:Property rdf:ID="email">
   <rdfs:domain rdf:resource="#Personne"/>
   <rdfs:range
        rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Property>
<rdfs:Class rdf:ID="Doctorant">
   <rdfs:subClassOf rdf:resource="#Personne"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Chercheur">
   <rdfs:subClassOf rdf:resource="#Personne"/>
</rdfs:Class>
<rdf:Property rdf:ID="sousDirection">
   <rdfs:domain rdf:resource="#Doctorant"/>
   <rdfs:range rdf:resource="#Chercheur"/>
</rdf:Property>
</rdf:RDF>
```

FIGURE 1.1 – Un schéma de description défini en RDFS

1.3 LINKED DATA (DONNÉES LIÉES)

Le Web est de plus en plus considéré comme un espace d'information global constitué non seulement de documents liés, mais aussi de données liées. Plus qu'une simple vision, le Web of Data qui en résulte est né de la maturation de la pile de technologies du Web sémantique et de la publication d'un nombre croissant d'ensembles de données selon les principes des Données Liées [1].

1.3.1 LINKED OPEN DATA

Linekd Open Data définit une vision des données liées et accessibles globalement sur l'internet, basée sur les normes RDF du web sémantique. LOD est souvent considéré comme un nuage de données virtuel où chacun peut accéder à toutes les données qu'il est autorisé à voir et peut également ajouter à toutes les données sans perturber la source de données originale. Cela fournit un environnement ouvert où les données peuvent être créées, connectées et consommées à l'échelle de l'internet. Une théorie de base de LOD est que les données ont plus de valeur si elles peuvent être connectées à d'autres données. Dans ce contexte, les données sont toute information structurée basée sur le web. Le LOD a été proposé comme base pour un gouvernement ouvert et pour résoudre de nombreux

problèmes d'intégration de données

Avantages LOD :

En outre, la mise en relation des ensembles de données ouvertes favorise la créativité et l'innovation, car tous les développeurs, les citoyens et les entreprises peuvent utiliser tous ces ensembles de données pour mettre les choses en contexte et créer des connaissances et des applications. Certains des avantages des données ouvertes liées sont :

- Utilisation efficace des ressources : Les données ouvertes liées réduisent la redondance en s'appuyant sur le travail des autres plutôt que de reproduire les systèmes existants.
- Amélioration de la qualité des informations : Les données ouvertes liées encouragent la normalisation des métadonnées et des formats de données, ce qui rend les données plus fiables et crédibles.
- Création de valeur ajoutée : En se connectant directement à d'autres données, les données ouvertes liées permettent aux utilisateurs de découvrir, d'utiliser et de réutiliser les informations de manière nouvelle et inattendue.
- Identification des lacunes dans les informations : Les données ouvertes liées permettent de mettre en évidence et de corriger les erreurs de données.
- Amélioration de la transparence : Les données ouvertes liées créent les moyens pour les citoyens et les groupes de défense de demander des comptes au secteur privé et aux gouvernements.

1.4 Edition des données sémantiques

Le développement fulgurant du Web 2.0 est à l'origine d'une nouvelle génération des éditeurs appelés 'éditeurs collaboratifs' passant de la centralisation à la décentralisation et de l'individu à la communauté. Depuis son introduction, le type de données répliqué commutatif (CRDT) a été largement étudié et continue d'être l'objet de nombreux travaux de recherche. Cet ouvrage présente la conception d'une nouvelle approche originale destinée à la réplication optimiste pour l'édition collaborative des stores sémantiques sur réseaux P2P. L'idée principale de ce travail est de concevoir un nouveau type de données commun et réplicatif pour les entrepôts sémantiques qui dépasse les limites d'un éditeur centré sur une architecture client/serveur à une architecture dynamique P2P, et cela afin de supporter la construction des connaissances de façon collaborative, de supporter le passage à l'échelle en termes d'utilisateurs et ressources, de supporter la dynamicité des pairs et d'assurer la disponibilité des triples-stores [5]. edition des données semantique.txt Displaying edition des données semantique.txt.

1.5 CRDT (type de donnée répliqué commutatif)

C'est un type de données répliqué pour lequel certaines propriétés mathématiques simples garantissent une cohérence éventuelle. Dans l'État style, les états successifs d'un objet devraient former un semi-réseau monotone, avec fusionner le calcul d'une limite inférieure. Dans le style op-based, concurrent En supposant seulement que le sous-système de communication assure la livraison éventuelle (en ordre causal pour les objets basés sur les opérations), les CRDT sont garantie de converger vers un état commun et correct, sans

nécessiter synchronisation. L'idée principale est de trouver le type de données pour les déplacements naturels Partant de ce principe, tout type de CRDT a été démontré dans Convergence si toutes les opérations sont référencées par des identifiants uniques différent Le défi se résume donc à concevoir des types de données et des techniques approprié pour assurer la commutativité des opérations. [6] L'idée consiste à associer un identifiant unique à chaque emplacement partagez des symboles , des lignes ou des atomes des documents . Si l'operation est généré, l'identifiant unique est également associé au paramètre positionelle Cependant, la gestion des identifiants est un problème très sérieux car La précision de cette méthode repose sur l'unicité et la préservation de l'identifiant Un changement dans l'ordre général des opérations. Par conséquent, la plage de valeurs de l'identifiant doit être choisie pour être compacte. Ainsi, entre les deux identifiants Etant donné, il doit toujours être possible de créer un nouvel identifiant. Parmi les algorithmes de CRDT existants, nous distinguons : OR-Set , C-SET ,SU-Set , B-SET, sr-CE et LD-SET Nous présentons ci-après chacun de ces algorithmes.

TABLE 1.1 – Par exemple : Soit la phrase simple suivante : la tour Eiffel est créée en 1887 a paris . Cette phrase est composée des parties suivantes

Sujet (ressource)	Tour Eiffel
Prédicat (propriété)	Créée/située
Objet (littéral)	1887/ Paris

1.6 Les algorithmes d'édition de données sémantique

1.6.1 OR-Set

OR-Set est conçu pour la mise à jour basée sur les éléments, il n'est pas adapté pour les opérations basées sur les modèles comme les opérations de mise à jour SPARQ Afin de réduire la complexité de la communication, au lieu d'étiqueter chaque élément séparément Remarquez que dans le cas d'OR-Set, le nombre de tours serait égal à le nombre de triples exploités, car nous devons envoyer un message pour chaque un. Les triples en double n'affectent pas la complexité des tours, ils sont traités comme un insertion normale.

1.6.2 C-SET [7]

est un CRDT conçu pour le type de données de l'ensemble. compteur associé à chaque élément pour garder une trace du nombre de fois qu'il a été ajouté et supprimé. Un élément est considéré comme membre du C-Set si son compteur est supérieur à zéro. C-Set est convergent, laisse la causalité au sous-jacent réseau, mais malheureusement, quand une suppression suivie d'un insert sont exécutés l'intention n'est pas préservée. La figure 4 présente un exemple de violation de l'intention dans l'ensemble C. À partir d'un point où un élément x a 3 dans son compteur et donc, pas un membre de l'ensemble. Si deux nœuds décident simultanément pour insérer x et le supprimer immédiatement, ils convergeront vers un état où x est un membre de l'ensemble, ce qui n'était l'intention d'aucun d'eux.

1.6.3 SU-Set

SU-Set a été développée pour servir comme une base pour les entrepôts RDF qui pourrait s'implémenter dans un moteur d'inférence RDF. SU-Set modifie le comportement des opérations à diffuser de telles sortes que l'ensemble des triplets seront affectés un par un.

1.6.4 B-SET [8]

repose sur l'entreposage de « pierres tombales », c.-à-d. les éléments supprimés sont simplement cachés à l'utilisateur via l'opération de recherche. L'utilisation de pierres tombales n'est pas appropriée pour les grands ensembles comme ceux que nous pouvons trouver dans le réseau de données, car leur complexité spatiale est élevée.

1.6.5 srCE

Une importante propriété du modèle srCE est que ses opérations ne s'exécutent pas directement sur l'ensemble final mais elles doivent d'abord s'exécuter sur l'un des multi-sets supplémentaires, selon le type de l'opération. Par la suite, une opération interne se génère en vue de produire un ensemble cohérent.

1.6.6 LD-SET

LD-Set est conforme à CRDT, il garantit la cohérence éventuelle dans tout les cas LD-Set concentre sur un mécanisme de réplication optimiste. Il vise à soutenir l'édition collaborative sociale de Données liées au sein d'une communauté virtuelle d'utilisateurs de différents sites, tout en maintenant une cohérence éventuelle afin de mise à jour simultanée

1.6.7 Le modèle CCI :

Dans le contexte de l'édition collaborative, un système d'édition collaborative partageant des données répliquées est considéré comme correct s'il assure un modèle de cohérence, plus précisément le modèle de CCI qui garantit à la fois la convergence des répliques et préserve la causalité et l'intention de toute opération générée dans le système.

1.7 Etude comparative entre ces algorithmes

SrCE est une approche destinée à la réplication optimiste pour l'édition collaborative des stores sémantiques sur un réseau P2P. Un réseau P2P est une collection de machines, appelées appepaires , qui échangentdes données via un système de communication. La spécificité est que ces machines peuvent à la fois être client et/ou serveur. C-Set aussi assure la convergence mais ne mentionne pas la garantir des critères de causalité, de cohérence et de préservation de l'intention du modèle CCI. B-Set est conçu non seulement pour assurer la convergence des répliques triples mais aussi pour préserver les intentions de l'utilisateur intégrées dans une architecture distribuée. Les ensembles d'opérations sont également définis afin de permettre l'édition simultanée des mêmes mémoires triples partagées. SU-set permet d'assurer la causalité, la convergence et les intentions dans une vision d'un "Live" Linked Data. LD-Set se concentre sur un mécanisme de réplication optimiste. Elle vise à

TABLE 1.2 – Résumé de comparaison entre les algorithmes

CCI	B-set	C-Set	SU-Set	LD-Set	Or-Set	SrCE
causalité		-				
cohérence		-	-	X		X
l'intention	X	-	X			X
convergence	X	X	X	X	X	X

l'édition collaborative sociale de données liées distribuées au sein d'une communauté virtuelle d'utilisateurs de différents sites, tout en maintenant une cohérence éventuelle afin de permettre mise à jour simultanée. apres avoir certaints algorithmes de CRDT nous terminons notre étude comparative par 1.2 qui résume une comparaison entre eux

Conclusion

Dans ce chapitre, nous avons présenté le web semantique, linked data et quelques modèles d'édition collaboratifs basés sur l'approche des CRDT. nous avons évalué les algorithmes CRDT étudiés dans ce chapitre selon le modèle CCI. Malgré leur robustesse elles ont leurs faiblesses. Par conséquent, dans le chapitre suivant, nous travaillerons sur l'amélioration de l'un des algorithmes susmentionnés.

Bibliographie

[1] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, "Linked data on the web (ldow2008)," in *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, (New York, NY, USA), p. 1265–1266, Association for Computing Machinery, 2008.

[2] T. Berners-Lee, J. Hollenbach, K. Lu, and J. Presbrey, "Tabulator redux : Browsing and writing linked data," *CEUR Workshop Proceedings*, vol. 369, 01 2008.

[3] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, vol. 284, pp. 34–43, May 2001.

[4] T. A. Ta, *Web sémantique et réseaux sociaux - Construction d'une mémoire collective par recommandations mutuelles et représentations.* Theses, Télécom ParisTech, July 2005.

[5] ZARZOUR-H, *Idition collaborative de Donnies Simantiques.* OMNISCRIPTUM, 2018.

[6] M. Shapiro and N. Preguiça, "Designing a commutative replicated data type," 2007.

[7] K. Aslan, P. Molli, H. Skaf-Molli, and S. Weiss, "C-Set : a Commutative Replicated Data Type for Semantic Stores," in *RED : Fourth International Workshop on REsource Discovery*, (Heraklion, Greece), May 2011.

[8] H. Zarzour and M. Sellami, "B-set : A synchronization method for distributed semantic stores," in *2012 IEEE International Conference on Complex Systems (ICCS)*, pp. 1–6, 2012.