État de l'art

Apperçu du domaine

Notre étude a la particulatrité de s'étendre sur le domaine de la fouille de donnée et du Web sémantique. En effet, on utilise les techniques de la fouille pour l'extraction des données depuis les différentes sources d'informations, et le Web sémantique afin de donner au métadonnées une structure plus lisible par la machine.

Dans ce chapitre on présente les téchnologies du Web sémantique que nous avons utilisé, puis nous effectuons une étude préliminère autour des travaux de recherche qui précède notre étude toute en introduisant les concepts et la problèmatique de notre sujet.

Technologies du Web sémantique

Intérêt du Web sémantique

Le Web sémantique est un domaine de recherche né des travaux de Tim Berners-Lee l'un des pionniers dans ce secteur [BLHL01] dont le but était d'ajouter du sens aux contenus du Web. Ce n'est pas une question d'ajouter une autre alternative du Web mais c'est plustôt d'étendre le Web actuel dans le but d'utiliser et manipuler le maximum de son contenu informatiquement. En clair, c'est de permettre à des programmes informatiques de traiter un ensemble étendu de données issues du Web.

Modèle RDF

Au centre du Web sémantique, comme la brique d'argile qui permet d'ériger les plus grands édifices, se trouve le modèle Resource Description Framework (RDF) . RDF [1] est un standard de World Wide Web Consortium (W3C), il se base sur un modèle de graphe sous forme de triplets (sujet, prédicat, objet) qui permettent d'exprimer tous types d'assertions. Il s'agit d'un cadre de description de ressources, d'une façon formelle sur le Web. C'est la première brique de standard du Web sémantique qui recouvre à la fois un modèle et plusieurs syntaxe pour publier des données variées sur le Web.

1. http ://www.w3.org/RDF/

Dans RDF :

— Les ressources sont un concept de base du Web sémantique, tout ce qui peut être référencé est une ressource. Dans un contexte plus technique : on déduit que tout ce qui peut être identifié par un Uniform Resource Identifier (URI) / Internationalized Resource Identifier (IRI) peut être considéré comme ressource.
— Un ensemble d'attributs décrivent la ressource, qui possèdent des caractéristiques et des relations avec d'autres ressources.
— Le cadre standardise la syntaxe de ces descriptions, les modèles et les langages.

La plus petite structure de description en RDF est le triplet.

FIGURE 1 – triplet RDF

Un triplet décrit une ressource, l'associe à une propriété et une valeur de cette propriété qui peut être une nouvelle ressource liée.
Par exemple "Moncef a écrit une page QuadsRDF.html à propos des quadruplets RDF" peut être décomposée en deux triplets ayant comme sujet "QuadsRDF.html" : <QuadsRDF.html, auteur, Moncef> et <QuadsRDF.html, thème, quadruplets RDF>.
On peut schématiser cela de la manière suivante :

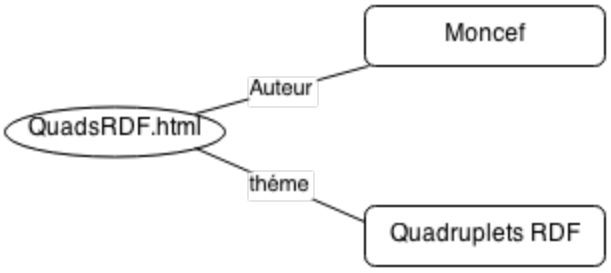

FIGURE 2 – Deux Triplets liés au même sujet

SPARQL

Si RDF fournit un modèle universel de représentation de métadonnées. D'autres niveaux de traitements ont été standariser au dessus de lui et notament l'interrogation de ces métadonnées. Protocol and RDF Query Language (SPARQL) fournit le langage d'interrogation du Web sémantique, et en cela il est à RDF ce que Structured Query Language (SQL) est au bases de données relationnelles.

SPARQL est un langage d'interrogation de graphe RDF dont l'énoncé de base est lui aussi un triplet (ressource, propriété, valeur).Il est une recommandation du W3C depuis juillet 2008. Poser une question en SPARQL consiste à écrire un graphe requête pour lequel on cherche des occurences dans le graphe cible.

N-Quads

N-Triples est une simple syntaxe ligne délimitée(line-delimited) pour les graphes RDF. N-Quads [2], un format qui s'étend N-Triples avec le contexte. Chaque triplet dans un document N-Quads peut avoir une valeur de contexte en option :
<subjet> <prédicat> <objet> <contexte>.

La notion de provenance est essentiel lors de l'intégration des données provenant de différentes sources ou sur le Web . Par conséquent, l'état de l'art des référentiels RDF (sujet, prédicat, objet, contexte) quadruplet, lorsque le contexte indique généralement la provenance d'une déclaration donnée.

2. http ://sw.deri.org/2008/07/n-quads/

Base de Connaissances

Introduction

Une base de connaissance regroupe des informations spécifiques à un domaine donné, sous un format exploitable par un ordinateur. Elle peut contenir des régles, des faits ou d'autres représentations. Les base de connaissances regroupent des informations structurées et nous particulièrement on cherche à exploiter ces informations pour les mettres dans une nouvelle structure plus facilement exploitable par la machine.

DBpedia

C'est un projet universitaire et communautaire d'extraction et d'exploitation automatique des données de wikipedia. C'est également un ensemble de données structurées et normalisées au format du Web sémantique. DBpedia 3.9 est la dernière version de DBpedia datant de Juin 2013.

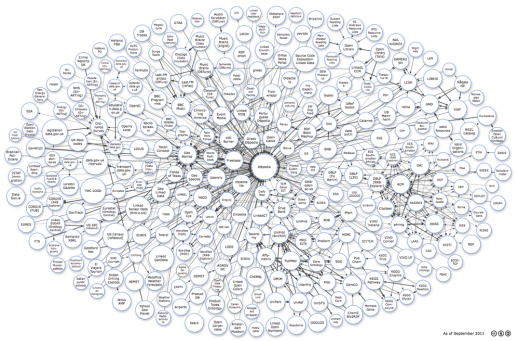

FIGURE 3 – DBpedia

Cette base de connaissance est écrite en Scala et Java. Elle adopte les normes du Web sémantique et du réseau Linked Open Data. Pour chaque document encyclopédique, il existe une page de ressources contenant toutes les données sous forme de triplets RDF.

YAGO

YAGO [3] est une large base de connaissance sémantique, délivré de Wikipedia, WordNet et GeoNames. Actuellement elle contient plus de 10 million entités (personnes, organisations, villes, etc..) et plus de 120 million faits de ces entités.
Les caractéristiques principales de YAGO :

— Une précision de 95%, chaque relation est annoté avec sa valeur de confiance.
— YAGO combine la taxonomie propre de WordNet avec la richesse du système de catégorie Wikipedia, l'attribution des entités à plus de 350 000 catégories.
— YAGO est une une ontologie qui attache une dimension temporelle et spacial pour plusieurs de ces faits et entités.

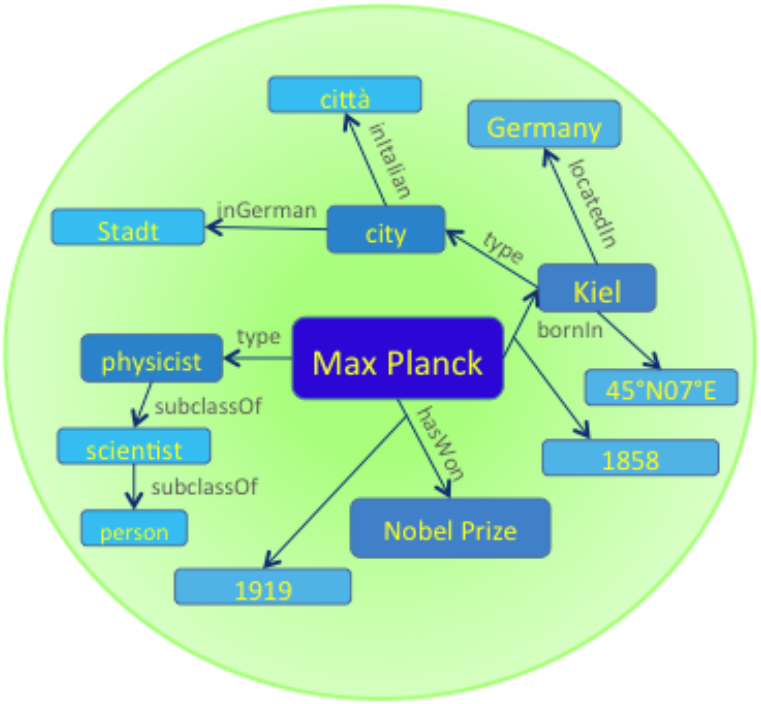

FIGURE 4 – YAGO

3. http ://www.mpi-inf.mpg.de/yago-naga/yago/

Différentes approches d'annotation temporelle

Introduction

Dans le domaine du Web sémantique, il y a plusieurs extensions de RDF qui ont été proposé pour : la vérité, la confiance, la certitude, le temps ect... Par exemple : Pour la verité de certaines triplets où le degré de la vérité est entre 0 et 1, l'instance "Rome is a big city to degree 0.8" peut être représentée par $(Rome, type, big\_city) : 0.8$.

De même pour la certitude un autre forme a été proposé : $(Max, hasSupervisor : (0.9, 2003), William)$ à la forme générale suivante $(s, p : (x, t), o)$
La certitude x est représentée sous forme d'un pourcentage, 90% pour cet exemple.

L'annotation temporelle

La nécessité de l'annotation temporelle sur les documents Web a été évoqué dans des nombreux travaux de recherche. La première approche formelle au problème de modélisation et d'interrogation temporelle en RDF a été introduite par Gutierrez et al [GHV05].

Ensuite, Udrea et al [URS06] ont remis en question la notion d'annoter temporellement les graphes RDF et depuis plusieurs travaux de recherche ont évoqué cette problèmatique. Ces derniers définissent le triplet annoté de la forme suivante $(s, p : t, o)$, t est une étiquette temporelle. De plus ils ont donné des algorithmes pour interroger les données RDF annotées.

RDF Temporel ou tRDF

Pour introduire RDF temporel "tRDF" on commence par les exemples suivants :
Il y a des triplets comme par exemple : " Mary est toujours la mère de John " qui n'ont pas une caractéristique temporelle explicite parce qu'il est toujours valide. Mais il y a aussi des triplets ayant une valeur vrai que dans une plage temporelle bien précise, par exemple : " Bill Clinton est le président de Etats Unis ", n'est valide que dans l'intervalle $[1993 - 2001]$.

Donc il y a des triplets qui ne peuvent être reconnus que dans des périodes temporelles précises.

D'après Andrea et al [PUS08] l'annotation tRDF peut être exprimé de la manière suivante (n est un nombre entier, T appartient à un interval de temps, s le sujet, p le prédicat, v l'objet) :

1. (s, $p : T$, v), ce type de triplet représente une relation entre le sujet et le prédicat et l'objet dure un temps T (dans n'importe quel point de temps dans T).
2. (s,p:<n:T>,v), ce triplet présente une relation entre s, p et v qui dure au moins n point de temps différents dans T.
3. (s, $p : [n : T]$, v), ce triplet présente une relation entre s, p et v qui dure au plus n points de temps différents dans T.

L'importance de l'annotation temporelle dans LOD

Présentation du LOD

Linked Open Data, Web de données (LOD) [4], c'est un moyen de publier des données structurées sur le Web où les données contenues dans des bases de données sont exposées sur le Web avec leur sémantique, ce qui donne la possibilité au métadonnées d'être connectées et enrichies d'une manière solide, permet d'avoir plusieurs représentations d'un même contenu et fait des rapprochements entre des ressources connexes.

Au cours des dernières années, le Web de donnée a développé dans une grande fusion de divers ensemble de données provenant de plusieurs domaines. Ce dernier décrit les ressources identifiées par des URI en représentant leurs propriétés et des liens vers d'autres ressources. L'ensemble des données fournit des connaissances du monde réel.

Relation entre l'annotation temporelle et LOD

Les informations sur un intervalle temporel de validité pour les évènements décrits par des triplets RDF jouent un rôle important dans un grand nombre d'applications. Un grand nombre de triplets dans LOD ne sont valides que

4. http ://linkeddata.org/

dans un certain intervalle de temps qu'ils l'appellent la protée de leurs temps. Par exemple dans DBpedia ils indiquent que "Mario Balotelli joue pour les équipes AC Lumezzane et le Milan AC". Lorsqu'on modélise des connaissances du monde réel, Mario Balotelli ne peut pas jouer au même temps avec AC Lumezzane et le Milan AC.

Les logiques temporelles d'informations ont besoin d'avoir de la protée temporelle des faits tels que "Mario Balotelli joue pour l'équipe AC Milan". Une approche a été proposée pour détecter la portée des évènements visés par des triplets RDF par Rule et al [Ani], elle se compose de quatre étapes principales :

- Les données du document Web sont normalisées pour tenir compte de l'importance des dates figurants dans les documents.
- La sortie de la phrase est comparée avec un ensemble d'intervalles de temps pertinents pour obtenir des notes de significations pour chaque intervalle.
- Un ensemble d'intervalles plus importants est sélectionné.
- Les intervalles sélectionnés sont fusionnés lorsque c'est possible.

La plateforme DeFacto(Deep Fact Validation) [LGMN12] a été utilisée pour la validation des états on cherchant des sources qu'elle confirme sur le Web.

Les triplets sont représentés par des faits et peuvent être associé à un contexte temporelle. Par exemple, $< Balotelli, team, ACMilan >$ se réfère à un événement de $2003 - 2009$, une annotation temporelle est rattachée au fait comme suit $< f, [ti, tj] >$.

Cette approche combine deux types d'informations : les informations temporelles recueillies dans des documents Web et les informations temporelles contenues dans les bases de connaissances, pour associer des intervalles de temps au triplets RDF.

Temps valide des triplets dans les données géospatiales liées

Bereta et al [BSK13] introduisent la composante temporelle des données du modèle stRDF et le langage de requêtes stSPARQL, récemment proposés pour la présentation et l'interrogation des données géospatiales liées qui changement dans le temps.

L'introduction du temps dans les modèles de données et les langages de requêtes, a été l'objet de recherches approfondies dans le champs des bases de données relationnelles.

Les trois types distincts de temps qui ont été étudiées :

— L'action temporelle indépendante, par exemple (01/12/1954 c'est l'anniversaire de John).
— Le temps d'évènement ou un fait vrai dans l'application (entre $2001 - 2012$ John a été professeur).
— Le délail de transaction qui est le moment où un fait est en cours dans la base de données (l'heure système qui présente l'heure exact quand John est un professeur "$2001 - 2012$" est en cours dans la base de données).

Bereta et al [BSK13] présentent également le concept de horodatages anonymes dans les graphes RDF, par exemple le quadruplet(quad) de la forme $(s, p, o)[t]$, où t est une horloge ou un timestamp x anonyme déclarant que le triplet est valable dans un certain point de temps inconnue.

L'idée principale est d'intégrer les informations géospatiales pour le modèle de graphe RDF temporel. Le langage d'interrogation spatial and temporal Protocol and RDF Query Language (stSPARQL) [5], ajoute deux nouveaux types de variables spatiales et temporelles, aux variables SPARQL standards.

Base de données temporelles

Une base de données temporelle est une base de données avec des acpects de temps intégrés(temps-valide, temps-transaction), c'est à dire un modèle

5. http ://www.strabon.di.uoa.gr/stSPARQL

de données temporelles et une version temporelle du langage structuré de requête(SPARQL, SQL).

En effet, le *temps valide* dénote la période du temps durant laquelle un fait est vrai par rapport à la réalité. Le *temps-transaction* est la période de temps pendant laquelle un fait est stocké dans une base de données.

Dans le contexte de l'annotation temporelle des graphes RDF, les besoins se résume comme suit :

- L'accès à des différents versions d'une ontologie.
- Récupération des informations passées sur les sites Web.
- La distribution des mise à jour des journaux.

En vertu des travaux de Antoniou et al [AvH04] présentent une ontologie du service Web, pour monter qu'une ontologie peut passer par plusieurs états, ainsi que d'autres recherches dont l'objectif d'analyser et justifier les besoins cités auparavant.

Une base de données temporelle peut être exprimé comme un répertoire d'informations temporelles. Gutiérrez et al [GHV07], montrent qu'il y aura deux manières pour ajouter des dimensions temporelles dans un graphe RDF intemporel :

- Étiqueter les éléments soumis à des changements, les triplets par exemple, à chaque changement un nouveau graphe créer et l'ancien état sera stocké quelque part.
- Versionner : capture de temps de transaction, l'étiquetage est mieux que les versions pour les raisons suivantes :
  - Il conserve le principe de la nature distribuée et extensible de RDF.
  - Si la nouvelle version n'affecte que quelques éléments cela implique la création d'un nouveau graphe, de ce fait on aura des contraites de mémoire.

Gutiérrez et al [GHV07], ont travaillé sur le domaine temporel à base de points et ils ont aussi codé les points du temps en intervalle.

Ces derniers ont proposé un vocabulaire pour affirmer les moments où les triplets sont valables dans un graphe RDF.

Graphe Temporel

Un graphe temporel c'est des triplets(s, p, o) avec des étiquettes temporelles qui représentent la période dans laquelle il est valable dans le monde réel. Exemple le triplet (s, p, o) est valable dans un temps t, $(a, b, c)[t]$, ou autrement dans un intervalle de temps $[t1, t2]$, $(a, b, c)[t1, t2]$.

L'idée générale de Pugliese et al [PUS08] est d'annoter RDF avec un interval de temps. Ces derniers ont proposé un graphe temporel d'indexation "tGRIN". C'est une structure d'indexation qui construit un index spécialisé pour RDF temporels qui sont stockés dans une base de données relationnelle "RDBMS".

D'autres efforts pour stocker RDF dans une base relationnelles :
— Jena2 de Apache
— Sesame de openRDF.org
— 3store ou triplestore de University of Southampton

D'autes index temporels sont implémentés (R+ trees, SR-trees, ST-index, and MAP21) mais l'index tGRIN présentent des performances supérieures selon les expérimentations faites dans [PUS08].

Synthèse

Plusieurs travaux de recherches on été mis au point pour résoudre le problème des données qui présentent un sémantique temporel dans les graphes RDF. Nous avons étudié ces travaux afin d'avoir une vision globale sur la problèmatique et pour voir ce qui est déjà fait dans ce domaine. On s'inspire de ces travaux pour proposer une nouvelle approche qui soit satisfaisante pour annoter temporellement les métadonnées de Wikipedia.

Extraction des données

Introduction

L'extraction, la fouille de données, ou encore la fouille des connaissances à partir de données, ont pour objet l'extraction d'un savoir, d'une connais-

saince, ou dans notre cas une connaissance mise en relation temporelle à partir de grandes quantité de données par des méthodes automatiques.

Différentes approches de l'extraction

Une approche proposée par Zweigenbaum et Tannier [ZT13] consiste à détecter les relations temporelles entre les évènements et les expressions temporelles à partir des comptes rendu hospitaliers.

La détection des relations temporelles entre les évènements dans un texte, fournit de bonnes informations pour l'extraction.
C'était les défis de TempEval Verhagen et al [VSCP10] qui ont abordé en "domaine ouvert", en cherchant à détecter en TempEval2 cinq types de relations temporelles :
$(Before, After, Overlap, Before\_or\_Overlap, Overlap\_or\_Before)$ et Identifier les relations temporelle décrivant la chronologie du séjour hospitalier.
Les relation à trouver dans des différentes situations :

- Entre un événement et une date ou autre événement qui domine.
- Entre un événement et la date de création de cette élément.
- Entre deux événement principaux de deux phrases consécutives.

Identifier les informations temporelles décrivant la chronologie entre ces événements.

Ces derniers utilisent des différents classifieurs (table de décision, arbre de décision, JRip, classifieurs bayésien naif) et le classifieur à arbre de décision J48 implémenté dans weka.

La question c'est d'identifier les situations les plus importantes à traiter et les méthodes à utiliser. Zweigenbaum et Tannier [ZT13] utilisent une méthode d'apprentissage supervisé avec un ensemble de données et des classifieurs entrainés pour chaque situation. L'évaluation a été appliquée sur un corpus d'apprentissage qui contient 190 échantillons, dont 120 échantillons de test.

On peut utiliser cette méthode pour les propriétés de DBpedia à la place des ces comptes rendus hospitaliers et chercher à chaque fois d'apprendre à partir d'un motif qui peut être temporel, spacial, ect... Au lieu

d'une procédure de décision gloutonne ou aléatoire, une relation de décision globale pourrait être implémentée pour étudier toutes les relations temporelles prédites.

Le but de Kessler et al [KTH+13] est d'extraire les dates saillantes (importantes) qui méritent de figurer dans une chronologie événementielle. Ces derniers ont utilisé une approche d'apprentissage pour extraire les dates saillantes pour un thème donné.

La méthode consiste d'annoter automatiquement les informations événementielles. C'est à dire repérer et baliser (Event) les occurrences d'événements au sens TimeML [6] et de les classifier selon l'ontologie définie par le schéma d'annotation.

Synthèse

L'extraction des informations temporelles est une étape primordial. On pourrait s'inspirer des méthodes présentées précédemment pour répondre aux objectifs fixés au démarrage de cette étude.

6. http ://timeml.org/site/index.html

Bibliographie

[Ani] Anisa Rula and Matteo Palmonari and Axel-Cyrille Ngonga and Daniel Gerber and Jens Lehmann and Lorenz Buhmann. Hybrid Acquisition of Temporal Scopes for RDF Data.

[AvH04] G. Antoniou and F. van Harme. A semantic web prime. *Semantic Web Primer. MIT Press*, 2004.

[BLHL01] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, pages 284–289, 2001.

[BSK13] Konstantina Bereta, Panayiotis Smeros, and Manolis Koubarakis. Representation and Querying of Valid Time of Triples in Linked Geospatial Data. *ESWC*, pages 1,15, 2013.

[GHV05] C. Gutierrez, C. Hurtado, and A. Vaisman. Temporal rdf. *Second European Semantic Web Conf. (ESWC' 05)*, pages 93, 107, 2005.

[GHV07] Claudio Gutierrez, Carlos Hurtado, and Alejandro Vaisman. Introducing Time into RDF. *IEEE Transactions on Knowledge and Data Engineering*, pages 207,218, 2007.

[KTH+13] Remy Kessler, Xavier Tannier, Caroline Hagege, Veronique Moriceau, and Andre Bittar. Extraction de dates saillantes. *Traitement Automatique des Langues, numéro spécial sur le traitement automatique des informations temporelles et spatiales*, 2013.

[LGMN12] J. Lehmann, D. Gerber, M. Morsey, and A.-C. Ngonga. Defacto - deep fact validation. *ISWC*, 2012.

[PUS08] Andrea Pugliese, Octavian Udrea, and V.S Subrahmanian. Scaling RDF with time. *Proc. of the 17th International Conference on World Wide Web (WWW 2008)*, pages 605,614, 2008.

[URS06] Octavian Udrea, Diego Reforgiato Recupero, and V. S. Subrahmanian. Annotated RDF. In York Sure and John Domingue, edi-

tors, *ESWC*, volume 4011 of *Lecture Notes in Computer Science*, pages 487–501. Springer, 2006.

[VSCP10] Verhagen, Sauri, Caselli, and Pustejovsky. Semeval-2010 task 13 : Tempeval-2. *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57,62, 2010.

[ZT13] Pierre Zweigenbaum and Xavier TANNIER. Extraction des relations temporelles. *TALN-RECITAL*, 2013.

Glossaire

IRI Internationalized Resource Identifier. 2

LOD Linked Open Data, Web de données. 7

RDF Resource Description Framework. 1

SPARQL Protocol and RDF Query Language. 3

SQL Structured Query Language. 3

stSPARQL spatial and temporal Protocol and RDF Query Language. 9

URI Uniform Resource Identifier. 2, 7

W3C World Wide Web Consortium. 1