Introduction générale

Depuis la création du web il y a vingt-cinq ans déjà, ce monde virtuel a vécu une évolution constante pour nous rendre une multitude de services et pour s'adapter à notre quotidien. Au fil des années, plusieurs versions du web sont apparues : le web documentaire, le web applicatif, le web social, le web mobile, etc...

Dans le contexte de l'évolution du web une nouvelle version dite le web 3.0 ou encore le web sémantique et sociale qui vise à propager nos modèles et leurs logiques; s'apprête à avoir le jour. Il y a plusieurs facettes du web, et la web sémantique offre un élément de réponse à l'intégration de chacune de ces facettes : il propose d'utiliser des métadonnées pour annoter les ressources du web, et d'exploiter la sémantique des schémas de ces annotations pour les traiter avec intelligence.

Le domaine du web sémantique compte plusieurs travaux de recherches sur les métadonnées du web afin d'avoir une nouvelle version bien structurée qui soit capable d'en assurer le contrôle efficace. Dans ce contexte, DBpedia est une base de donnée structurée qui contient des informations extraites de wikipedia et rend ces informations disponibles sur le web.

Aussi, Resource Description Framework (RDF) est le premier des standards de la web sémantique et se trouve être un modèle à plusieurs syntaxes, dans une est "Turtle" pour publier des données à thèmes variés sur le web.

Ce langage de modélisation permet à quiconque de décrire des ressources sur le web et aussi des ressources du web. Dans ce modèle connu comme étant la "lingua franca" du web, tout est exprimé sous forme de triplets (subject, predicate, object) où chaque triplet contribue à une description du monde.

Néanmoins, des faits tels que ceux donnés dans DBpedia sont en mesure d'être adaptés au changement constant du monde. RDF n'est pas bien équipé pour exprimer d'une manière cohérente la validité temporelle des états, tels que "Obama est le président des états-Unis depuis 2008".

Pour surmonter ce gap avec une modélisation RDF adéquate, plusieurs anciens travaux de recherches ont proposé d'attacher à ces triplets des annotations temporelles, ceci revient à une formalisation de ces états avec

des contraintes temporelles comme des quadruplets de la manière suivante $(subject, perdicate, object, time)$ à la place du formalisme de triplet habituel.

La théorie derrière un modèle de données basé sur des quadruplets évolue autour des termes de représentation, de connaissance, de raisonnement mais aussi d'interrogation. Or, le problème c'est qu'ils ne donnent aucune indication sur la façon dont les annotations temporelles sont créées.

Qui ou comment génère-t-on ces annotations ?
Dans le cadre de DBpedia, on veut extraire les informations temporelles automatiquement. On veut aussi utiliser les informations temporelles explicites de wikipedia et les informations temporelles déduites à partir de l'historique de modification des pages web.

De plus wikidata de "wikimedia foundation" donne des informations sur les articles wikipedia, dont une partie contient des indicateurs temporels qu'on veut extraire par la même occasion.

L'objectif de ce stage est d'une part, l'extraction des informations temporelles depuis les différents documents en utilisant les techniques de fouille de données. D'autre part, d'annoter ces documents selon leurs contextes. Enfin mettre ces informations sous forme de quadruplets structurés dans BDpedia.

État de l'art

Apperçu du domaine

Notre étude a la particulatrité de s'étandre sur le domaine de la fouille de donnée et du web sémantique.

En effet, on va utiliser les téchniques de la fouille pour l'extraction des données dans les différents documents, et le web sémantique pour avoir une structure plus lisible par la machine.

Différentes approches d'annotation temporelle

La nécessité de l'annotation temporelle sur les documents web ou l'adressage des changements d'une ontologie au meilleur de notre connaissance 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].

Udrea et al [URS06] étaient les premiers qui ont mis en question la notion d'annoter temporellement les graphes RDF. Ces derniers définissent le triplet annoté par un token de la forme suivante $(s, p : t, o)$, t est un token, où la propriété est annotée plutôt que le triplet. De plus ils ont donné des algorithmes pour interroger les données RDF annotées.

Par ailleurs la modélisation du temps est présente presque dans toutes les applications web, Abiteboul et al [Abi97], cela implique l'apparition de plusieurs extensions de RDF qui soutiennent le raisonnement temporel, le raisonnement de l'incertitude.
$(Max, hasSupervisor : (0.9, 2003), William)$ à la forme générale suivante $(s, p : (x, t), o)$
La certitude x est représentée sous forme d'un pourcentage, 90% pour cet exemple.

D'après Zimmermann et al [ZLPS12], dans le domaine du web sémantique, il y a plusieurs extensions de RDF qui ont été proposé pour : la vérité, la confiance, le temps ect...
Par exemple pour la verité de certaines triplets où le degré de la vérité est entre 0 et 1 pour l'instance "Rome is a big city to degree 0.8" peut être représentée par $(Rome, type, big\_city) : 0.8$.

Puissance de RDF

Les graphes RDF sont représentés comme une base de connaissance à partir de laquelle on peut générer d'autres nouvelles connaissances, d'autres graphes.

Cela peut poser des problèmes parce qu'une implication dans le cadre temporel est un peu complexe dans RDF que dans le cas des bases de données standards. Par ailleurs il présente un véritable défi dans ce domaine.

Contexte des DB 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 de données temporelles et une version temporelle du langage structuré de requête(Structured Query Language - 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.

*Objectifs de l'annotation temporelle des graphes RDF :

- 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.

Antoniou et al [AvH04]présentent une ontologie du service web, pour monter qu'une ontologie peut passer par plusieurs états.

Une base de données temporelle peut être exprimer comme un répertoire d'informations temporelles. Gutiérrez et al [GHV07], expliquent 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 temporeaux connus comme (R+ trees, SR-trees, ST-index, and MAP21), l'index tGRIN présentent des performances supérieures selon les expérimentations faites dans [PUS08].

RDF Temporel

Pour introduire le 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 importance temporelle parce qu'ils sont toujours valable. Mais il y a aussi des triplets valable que dans une plage temporelle bien précise, par exemple " Bill Clinton est le président de Etats Unis", est valable dans l'intervalle $[1993 - 2001]$.

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

L'annotation tRDF peut être exprimé (n est un nombre entier, T appartient à un interval de temps)

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 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.

Pugliese et al [PUS08] ont développé tGRIN et ils ont fait une comparaison pour montrer ça performance (tGRIN, JENA2, sesame, 3store).

Une autre Approche Annotation avec des faits

Linked Open Data (LOD), c'est un moyen de publier des données structurées sur le web, ce qui donne la possibilité au métadonnées d'être connectés et enrichis 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 LOD a développé dans une grande fusion de divers ensemble de données provenant de plusieurs domaines.

Linked Open Data 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.

Les informations sur un interval 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 et valables que 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 ". On veut modéliser les connaissances du monde réel donc cela n'est pas possible.

C'est à dire 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 "Article non publié".
L'algorithme 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.

Un ensemble d'intervalles déconnectés sont retournées par l'algorithme.

L'évaluation était faite à partir des données extraites de DBpedia et la base de connaissances de YAGO. L'algorithme DeFacto est utilisé pour l'extraction des données.

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 $2003 - 2009$. Ils définissent une annotation temporelle avec des faits comme suit $< f, [ti, tj] >$.

Cette approche combine deux types d'informations : les informations temporelles recuiellies 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 de base 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 un 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] introduisent également le concept de horodatages anonymes dans les graphes RDF, par exemple quads 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 x.

L'idée principale est d'intégrer les informations géospatiale pour le modèle de graphe RDF temporelle. Le langage d'interrogation stSPARQL, ajoute deux nouveaux types de variables spaciales et temporelles, aux variables SPARQL standards.

Bereta et al [BSK13] décrivent les motifs des triplets temporelles qui est une expression de la forme suivante (s, p, o, t), une forme qu'on souhaite utiliser dans notre cas d'étude, où (s, p, o) est le triplet motif et t c'est une période temporelle ou une variable.

Synthèse

Plusieurs travaux de recherches on été mis au point pour résoudre ce problème des données qui présentent un sémantique temporel dans les graphes RDF. On s'inspire de ces travaux pour proposer une nouvelle approche qui peut être satisfaisante pour résoudre ce gap.

Extraction des données

L'extraction, la fouille de données, ou encore la fouille des connaissance à partir de données, à pour objet l'extraction d'un savoir, d'une connaissaince, ou dans notre cas une connaissance mise en relation temporelle à partir de grande quantité de données par des méthodes automatiques.

Une approche proposée par Zweigenbaum et al[?] 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 des bonnes informations pour l'extraction.
C'était les défis de TempEval qui ont abordé cette problématique en "domaine ouvert", ont 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 al [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 nos documents DBpedia à la place des ces comptes rendus hospitaliers. 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 et de les classifier selon l'ontologie définie par le schéma d'annotation.
"Event" c'est un tag pour les éléments dans un texte indique que c'est un événement sémantique.

Synthèse

Dans cette étude l'extraction des informations temporelle est une étape essentielle. On pourrait s'inspirer des méthodes présentées ci-dessous pour répondre aux objectifs fixés au démarrage du projet.

Bibliography

[Abi97] S. Abiteboul. Querying semi-structured data. *Sixth Intel Conf. Database Theory (ICDT' 97)*, 1997.

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

[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.

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

[ZLPS12] Antoine Zimmermann, Nuno Lopes, Axel Polleres, and Umberto Straccia. A general framework for representing, reasoning and

querying with annotated semantic web data. *Journal of Web Semantics, Elsevier*, 2012.

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