

Annotation Temporelle Des Données DBpedia

Moncef BEN RAJEB

Mars 2014

Introduction générale

Depuis la création du web il y a de cela vingt-cinq ans déjà, ce monde virtuel a vécu une évolution constante. Il offre une multitude de services aux utilisateurs individuels, aux entreprises, mais aussi à la société. Au fil des années, plusieurs versions du web ont vu le jour : 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 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 le 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 est un objet de recherche sur les métadonnées du web. L'objectif principale de la naissance du web sémantique c'est d'en avoir une nouvelle version du web bien structurée qui soit capable d'en assurer le contrôle efficace des métadonnées. Dans ce contexte, DBpedia¹ est une base de donnée structurée qui contient des informations extraites de Wikipedia² 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"³ 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.

1. <http://dbpedia.org/About>

2. <http://wikipedia.org>

3. <http://www.w3.org/TeamSubmission/turtle/>

Néanmoins, des faits tels que ceux donnés dans DBpedia sont en mesure d’être adaptés au changement perpétuel 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 problème 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, predicate, 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.

De ce fait, 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 et d’autre part, d’annoter ces ressources selon leurs contextes, tout ceci mettre ces informations sous forme de quadruplets structurés dans BDpedia.

Portée de ce document

Ce mémoire de master résume les recherches, réflexions, modélisations, propositions et développements réalisés durant ce stage. Il conclut la seconde année de master Web Intelligence. Le contenu est organisé de la manière suivante :

- Une première partie présente l'état de l'art réalisé sur l'annotation temporelle des triplets RDF dans les bases de connaissances.
- La deuxième partie englobe les diverses propositions pour répondre aux besoins identifiés.
- La troisième partie développe l'aspect technique de la mise en oeuvre.
- La quatrième partie ouvre des perspectives.

État de l’art

Apperçu du domaine

Notre étude a la particularité de s’étendre sur le domaine de la fouille de donnée et du web sémantique.

En effet, on va utiliser les techniques 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

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 pour 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 [?].

Ensuite, Udrea et al [?] ont remis en question la notion d’annoter temporellement les graphes RDF et depuis plusieurs travaux de recherche ont évoqué cette problématique. 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.

Base de données temporelles

Une base de données temporelle est une base de données avec des aspects 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 (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 résument comme suit :

- L'accès à des différentes versions d'une ontologie.
- Récupération des informations passées sur les sites web.
- La distribution des mises à jour des journaux.

En vertu des travaux de Antoniou et al [?] présentent une ontologie du service web, pour montrer 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 exprimée comme un répertoire d'informations temporelles. Gutiérrez et al [?], 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 contraintes de mémoire.

Gutiérrez et al [?], 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 [?] 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’autres 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 [?].

tRDF ou RDF Temporel

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 [?] 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. $(s, p : \langle n : T \rangle, v)$, 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 .

Une autre Approche Annotation avec des faits

Linked Open Data (LOD)⁴, c'est un moyen de publier des données structurées sur le web, 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 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.

4. <http://linkeddata.org/>

Les informations sur un intervalle temporel de validité pour les événements 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 dans un certain intervalle de temps qu'ils l'appellent la portée 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, 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 portée temporelle des faits tels que "Mario Balotelli joue pour l'équipe AC Milan". Une approche a été proposée pour détecter la portée des événements visés par des triplets RDF par Rule et al [?], 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) [?] a été utilisée pour la validation des états en cherchant des sources qu'elle confirme sur le web.

Les triplets sont représentés par des faits et peuvent être associés à un contexte temporelle. Par exemple, $\langle \textit{Balotelli}, \textit{team}, \textit{ACMilan} \rangle$ se réfère à un événement de 2003 – 2009. Ils définissent une annotation temporelle avec des faits comme suit $\langle f, [t_i, t_j] \rangle$.

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 [?] 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 changent 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 délai 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 [?] 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 géospatiales pour le modèle de graphe RDF temporel. Le langage d'interrogation stSPARQL, ajoute deux nouveaux types de variables spatiales et temporelles, aux variables SPARQL standards.

Synthèse

Plusieurs travaux de recherches ont été mis au point pour résoudre le problème des données qui présentent une sémantique temporelle dans les graphes RDF. On s'inspire de ces travaux pour proposer une nouvelle approche qui

soit satisfaisante pour annoter temporellement les métadonnées de Wikipedia.

Extraction des données

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

Une approche proposée par Zweigenbaum et Tannier [?] consiste à détecter les relations temporelles entre les événements et les expressions temporelles à partir des comptes rendus hospitaliers.

La détection des relations temporelles entre les événements dans un texte, fournit de bonnes informations pour l'extraction.

C'étaient les défis de TempEval Verhagen et al [?] 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 temporelles décrivant la chronologie du séjour hospitalier.

Les relations 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 cet élément.
- Entre deux événements 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 naïf) 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 [?] utilisent une méthode d'apprentissage supervisé avec un ensemble de données et des classifieurs entraînés 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 [?] 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 TimeML⁵ 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.

5. <http://timeml.org/site/index.html>

Points de départ

RDF

RDF est l'abréviation de "Resource Description Framework" qui se base sur un modèle de graphe. 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.

Dans RDF :

- Les ressources sont un concept de base du web sémantique. D'où : "tout ce qui peut être référencé est une ressource". Et dans un contexte plus technique : On déduit que tout ce qui peut être identifié par un URI/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.

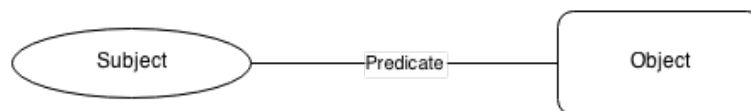


FIGURE 1 – triplet RDF

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> (figure2) et <QuadsRDF.html, thème, quadruplets RDF> (figure3). On peut schématiser cela de la manière suivante :

Analyse des besoins

Introduction

Le large succès de Wikipedia (qui est le 2ème site le plus visité sur internet) et le progrès des techniques d'extraction des données ont abouti à la naissance de la construction automatique de larges bases de connaissances comme DBpedia, YAGO, etc...

Beaucoup de connaissances sont construites en se basant sur l'extraction automatique des faits relationnels dans un texte. Malheureusement, les bases de connaissances convergent sur les faits statiques et ne donnent pas une grande importance à la dimension temporelle. Malgré le fait que la majorité des faits évoluent avec le temps, ou n'est valide que dans une période temporelle précise. Ainsi, nous remarquons que le temps a une dimension significatif dans ces bases de connaissances.

Dans cette étude on veut extraire des faits temporels et des événements depuis des informations semi-structurées de Wikipedia et Wikidata ; et textuelles de Wikipedia.

La dimension temporelle est particulièrement importante dans les relations binaires comme `isPresidentOf`, `isCEOof`, `isMarriedTo`, on peut être mariée à plusieurs épouses mais dans des différents intervalles de temps mais "On ne prend pas compte des exceptions de mariage polygames".

Une base de connaissances contenant plusieurs présidents des États-Unis ne peut être consistante que lorsqu'on ajoute une dimension temporelle à ces faits. De plus l'annotation temporelle aide à faire la distinction entre les faits courants et les faits dépassés. Par exemple le fait "Kennedy est le président des États-Unis" est correct, mais n'est plus valide. Lorsqu'on attache une annotation temporelle à un fait comme celui là, il devient universellement valide.

Problématique

Lorsqu'on parcourt un document "Wikipedia" on trouve beaucoup d'informations temporelles qui sont généralement liées à un contexte précis. Il

est plus difficile d'exploiter ces informations si elles ne possèdent pas une structure claire et lisible par la machine.

Il se trouve que des informations liées au même contexte temporel dans DBpedia sont exprimées de la manière suivante :

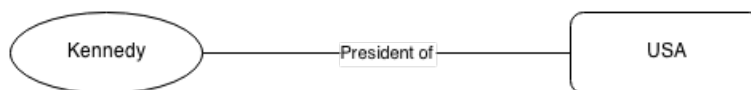


FIGURE 5 – triplet "Kennedy"

Le premier triplet n'a pas une sémantique valide que en tenant compte du triplet suivant :



FIGURE 6 – triplet presidential term "Kennedy"

On vise plutôt à annoter les triplets (s,p,o) avec une étiquette temporelle qui précise la validité de ce terme dans un cadre logique qui appartient au monde réel où en dehors de ce cadre, on peut dire que ce triplet RDF n'est pas valide et qu'on ne peut pas l'utiliser.

Sources de données

Les faits temporels extraits de Wikipedia consiste en deux phases principales : extractions des données semi-structurées (tableaux, infoboxes) et l'extraction depuis le texte wikipédia.

Texte Wikipedia

On cherche à extraire les informations temporelles qui ont un contexte de validité lié à un fait à partir des données textuelles des pages de Wikipedia. Par exemple sur la page de John Fitzgerald Kennedy, on retrouve : Kennedy a visité Berlin Ouest le 23 Juin 1963.

On veut extraire les ressources en donnant une nouvelle structure pour mieux présenter leur contexte de validité temporelle.

Infobox Wikipedia

Les infoboxes : Contiennent généralement les informations les plus importantes des entités décrites dans l'article. Par exemple dans l'infobox de "Kennedy" l'ancien président des États-Unis sur wikipédia on trouve la date d'élection, le prédécesseur du président, la durée du mandat, etc. . .
Chaque infoBox a un type particulier comme : évènement historique, élection, compétition ect. . .

John Fitzgerald Kennedy	
	
Fonctions	
35 ^e président des États-Unis	
20 janvier 1961 – 22 novembre 1963 (2 ans, 10 mois et 2 jours)	
Élection	8 novembre 1960
Vice-président	Lyndon B. Johnson
Prédécesseur	Dwight D. Eisenhower
Successeur	Lyndon B. Johnson
Sénateur du Massachusetts	
3 janvier 1953 – 22 décembre 1960	
Prédécesseur	Henry Cabot Lodge, Jr.
Successeur	Benjamin A. Smith II
Représentant du 11 ^e district du Massachusetts	
3 janvier 1947 – 3 janvier 1953	
Prédécesseur	James Michael Curley
Successeur	Tip O'Neill

FIGURE 7 – Info Box Wikipedia "Kennedy"

Tableau

Dans les tableaux "Wikipedia" figurent des informations temporelles plus structurées et plus faciles à extraire, qu'on cherche à récupérer.

Vous trouverez ci-dessous des informations sur les anciens présidents des États-Unis.








	Nom		Dates du mandat	Durée (j.)	Parti	Notes
1	George Washington		1789 1797	2 865 ^{note 1}	Indépendant (Fédéraliste)	Premier président des États-Unis, c'est le général vainqueur des Britanniques et héros de l'Indépendance . Très populaire, il est le seul président ayant été élu deux fois à l'unanimité (il n'avait pas d'opposant), en 1789 et 1792. Sous son mandat, un <i>Bill of rights</i> comprenant 10 amendements à la Constitution est ratifié, garantissant un certain nombre de droits individuels comme la liberté du culte, d'expression, de la presse, de port d'armes (amendements interprétables de différentes façons) ou encore un certain nombre de protections judiciaires. Cependant, l'esclavage reste permis (il possédait lui-même des esclaves dans sa propriété de <i>Mount Vernon</i> en Virginie). La banque des États-Unis est créée sous l'impulsion de son secrétaire au Trésor, Alexander Hamilton, afin de faire jouer un rôle central à l'État fédéral en matière économique. Alors qu'il ne souhaitait pas se représenter, il effectue un second mandat pour éviter que les tensions vives entre Alexander Hamilton et Thomas Jefferson ne débouchent sur un éclatement de la nation. Il refuse de se présenter à nouveau pour un troisième mandat.
2	John Adams		1797 1801	1 460	Fédéraliste	Vice-président de George Washington, il est élu par 72 voix contre 68 face à Thomas Jefferson qui devient son vice-président ^{note 2} . Sa présidence est notamment marquée par l'opposition entre les fédéralistes (John Adams) et les républicains-démocrates (Thomas Jefferson). Premier président à résider, en 1800, à la Maison-Blanche à Washington, devenue capitale fédérale. Candidat à un second mandat, il arrive derrière Thomas Jefferson et Aaron Burr.
3	Thomas Jefferson		1801 1809	2 922	Républicain-démocrate	C'est l'un des Pères de l'Indépendance et le principal rédacteur de la <i>Déclaration d'Indépendance</i> . Il fut pendant plusieurs années ambassadeur des États-Unis auprès de la France. Membre du Parti républicain-démocrate, vice-président de John Adams, il est élu contre Aaron Burr au 36 ^e tour de scrutin à la <i>Chambre des représentants</i> , réalisant la première alternance de l'histoire des États-Unis. <i>Achat de la Louisiane</i> à la France en 1803. Il entretient une relation ambiguë avec l'esclavage (il possède des esclaves dans sa propriété de <i>Monticello</i> en Virginie).
4	James Madison		1809 1817	2 922	Républicain-démocrate	Un des auteurs de la Constitution, Secrétaire d'État sous Jefferson, il est élu en 1808 en partie sur son habileté diplomatique à une période où la France et le Royaume-Uni sont prêts à faire la guerre aux États-Unis. Il fut réélu en 1812 contre le candidat républicain démocrate dissident DeWitt Clinton alors que le pays est en guerre avec le Royaume-Uni, guerre qui dura de 1812 à 1814, soldée par un <i>statu quo ante bellum</i> .
5	James Monroe		1817 1825	2 922	Républicain-démocrate	Triomphe définitif du Parti républicain-démocrate : sans opposant, il est réélu en 1820 à l'unanimité moins une voix du <i>collège des grands électeurs</i> . Il est le concepteur de la doctrine Monroe, à l'origine de l'isolationnisme américain et de la réduction de l'influence politique des puissances européennes (Grande-Bretagne, Espagne, France...) sur le continent américain.
6	John Q. Adams		1825 1829	1 461	Républicain national	Fils du 2 ^e président, John Adams, il est élu par le Congrès (aucun candidat n'avait obtenu la majorité absolue) face à Andrew Jackson, puis échouera face à celui-ci en 1828. Il était un adversaire acharné de l'esclavage des Noirs.
7	Andrew Jackson		1829 1837	2 922	Démocrate	Major général, vainqueur de la bataille de la Nouvelle-Orléans (Chalmette) en 1815. Il obtient la majorité des voix en 1824, mais ne parvient pas à obtenir la majorité absolue et échoue au Congrès face à John Q. Adams. En 1828, il affronte de nouveau celui-ci et le bat à une large majorité. Premier président démocrate, il a renforcé la démocratie aux États-Unis. Sa mémoire est cependant ternie par son soutien très actif à la <i>déportation des Amérindiens</i> à l'ouest du Mississippi et à l'esclavage des Noirs.

FIGURE 8 – Exemple de tableau Wikipedia

Wikidata

Sur la page wikidata, liée au contenu de la page wikipedia de Kennedy, on a intérêt à extraire les informations temporelles (fig).


NDL identifier	00445469 	[edit]
	1 source	
		[add]
date of birth	May 29 1917 <i>Gregorian</i>	[edit]
	1 source	
		[add]
date of death	November 22 1963	[edit]
	1 source	
		[add]

FIGURE 9 – Exemple Wikidata

Historique des pages wikipedia

On souhaite, si c'est possible, extraire des informations temporelles relatives à l'historique de modifications des pages de wikipedia. Car, il se trouve qu'il y a beaucoup d'informations liées à cet historique.

Résumé

Il s'avère que beaucoup de points du temps sont liés à plusieurs sources d'information événementielle ; on cherche à extraire ces informations afin de les mettre sous une forme plus adéquate.

Notre modélisation

Modèle quaternaire : un modèle qui capte la base du fait avec un indice temporel.

<politician> served as <politician office> from <date> to <date>

f1 : Kennedy holdsPoliticalPosition PresidentOfUSA

f2 :f1 startedOnDate

f3 :f1 endedOnDate

HappenedDate est utilisée pour dire que le fait est valide que dans ce point du temps.

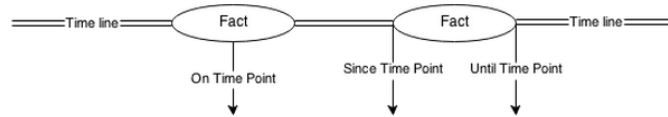


FIGURE 10 – Event Time Line

Pour surmonter ce problème et exprimer la validité temporelle d'un triplet RDF d'une manière à la fois intelligente et lisible par la machine ; on souhaite rattacher au triplets valides que dans une plage temporelle bien précise une étiquette temporelle adéquate.

Exemple :



FIGURE 11 – Modélisation quadruplet

On s'intéresse au format N-Quads qui est un standard w3c basé sur la forme N-Triples. L'avantage est qu'il se distingue par la possibilité d'encoder des graphes multiples. Les quadruplets vont être formalisés de la manière suivante :

$\langle s, p, o, [t1, t2] \rangle$, un sujet, prédicat, objet avec une intervalle de temps.

$\langle s, p, o, t \rangle$, de même avec un point de temps t .

Analyse temporelle

Représentation

Les informations temporelles peuvent avoir des repèsentations différentes :

- Un évènement “ Je vous propose un rendez-vous *demain* pour parler de mon projet 'PiSharing' “.
- Une connaissance “ Jacques Chirac est le président de la république Française “ **mais quand ?**.

Ambiguïtés temporelles

Le présent par exemple peut avoir plusieurs sens ou contextes : présent de narration, présent de généralité, présent qui réfère au futur proche, etc...

Les signaux temporelles sont ambigus : réunion de 14h à 16h, il court pour rattraper le temps, tu tournes après la rivière, etc...

La plupart des expressions sont floues : il y a deux ans, chaque deux semaines, j'arrive dans deux secondes, etc...

L'analyse du temps s'inscrit dans la compréhension globale des textes, et des événements auxquels on fait référence dans ce texte.

Modalité : l'équipe de France voulait gagner la coupe du monde en 2006.

Anaphore : cela pourrait avoir lieu dans les éditions suivantes.

Les événements décrits (et que l'on souhaite fixer temporellement) peuvent être : duratifs ou ponctuels/accomplis ou inaccomplis.

De même pour les dates qui peuvent être : Date absolue "le 18 mars, c'est mon anniversaire" ; Date relative par rapport au moment de l'énonciation : " il y a deux ans ". Pour la durée : Durée absolue " durant 2 ans " ; Durée relative : " depuis un an ".

On trouve aussi : Expression de fréquence " tous les ans, le vendredi 13 ", Expression plus complexe " après la Révolution Tunisienne ".

Résumé

Dans l'encyclopédia libre Wikipedia, il se trouve qu'il y a beaucoup d'informations temporelles. Dans cette étude on va plutôt essayer d'extraire des dates saillantes en fonction de leurs sémantiques événementiels.

Proposition

Introduction

L'annotation d'un triplet RDF est une façon d'ajouter des metadonnées à un triplet RDF pour décrire la validité temporelle, une restriction spatiale, ect...

Comment on utilise les annotations temporelles ? Sur le site sig.ma créer par the digital enterprise research institute in Ireland, la plateforme fournit un moteur de recherche par mot clé qui permet de récupérer des images et des textes accessibles par des annotations RDF, ainsi que d'une liste d'URI synonymes correspondant à la clé de recherche et des liens vers des sources Web contenant des données RDF pertinentes.

Schéma de modélisation

Tout d'abord, et à partir des différents sources d'informations on veut récupérer les informations temporelles dans leurs contextes sémantiques à l'aide de nos différents extracteurs implémentés. Ensuite, les mettre dans un ensemble de fichiers comme le montre le schéma ci-dessous.

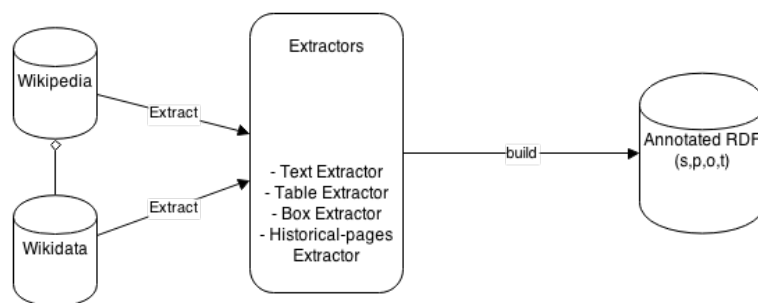


FIGURE 12 – Modélisation générale

Certes, on peut créer des “patterns” partons ou motifs à travers les faits temporels dans la base de connaissances mais on cherche ici à extraire les informations des textes et des données semi-structurée afin de leur trouver une autre structure plus adéquate (tel un triplet annoter ou bien un quadrupet).

Démarche

Au début, nous avons commencé par une procédure d'extraction classique avec un parseur avec l'API SAX pour extraire des données temporelles des dumps xml de wikipédia. Nous avons réussi à extraire plusieurs informations temporelles mais nous avons rencontré des problèmes relatives à la taille du dumps wikipédia/wikidata et au fait de relier ces informations temporelles au bon contexte du triplet qu'on veut annoté.

Ce fait nous a amené à chercher une autre solution que celle choisi au départ. Nous avons trouvé une solution plus intelligente pour former les quads ou des triplets annotés à partir des dumps DBpedia.

Sur le site de DBpedia, nous avons étudié les deux formats du dumps JSON et CSV afin de voir la structure des informations. Suite à nos observations nous avons décidé de manipuler les fichiers CSV pour voir la logique et la structure de ces informations.

Nous cherchons à former les quadruplets à partir des faits existants dans DBpedia. Dans cette dernière, on trouve des faits temporels et d'autres faits relié à leurs contexte.

Une liste des faits sera proposer en output et un expert sera placer pour juger la validité des données trouvées par notre algorithme sous forme de résultats labellisés en premier temps puis d'enregistrer les ressources dans des fichiers de quadruplets RDF classés selon des catégories bien spécifiques. L'expert peut sélectionner la liste des indicateurs *tokens* temporels misent par défaut dans le code de l'application, ajouter des indicateurs à la liste existante ou bien introduire une nouvelle liste d'indicateurs. Sachant que plus on introduit des indicateurs temporels, plus on aura des couples.

Nous avons testé notre algorithme avec les deux indicateurs *YEAR* et *DATE*. L'algrithme cherche à trouver une liste de couple de faits temporels et faits reliés possible qui seront les entrées d'une requête *SPARQL*.

L'expert intervient pour juger la valider des résultats trouvés. Un seul output valide peut donnée un ou plusieurs résultats "des triplets annotés".

Nous avons défini une base SPOTbase (*Subject, Predicate, Object, Time*) qui contient les quadruplets extraits par catégorie.

Nos algorithmes manipulent des ressources DBpedia afin de les mettre sous une autre forme valide dans le temps et tout en gardant une sémantique correcte.

Procédure

Notre proposition consiste à extraire des faits temporels de la forme suivante :

****Entity*with*TemporalToken**** puis de chercher d'autres propriétés reliées à ces faits temporels ****Entity*with*OtherWord****. On donne la main à un expert pour choisir un couple de faits puis de valider les résultats générés automatiquement par notre algorithme afin de mettre en place les quadruplets dans la base SPOTbase.

Étapes

La première étape consiste à trouver tout les propriétés et les stocker dans un fichier pour les exploiter comme une base de faits.

```
Algorithme : getProperties(propFile)
writer<=bufferWriter(propFile)
resultSet<=queryExecution(query)
writeProperties(resultSet,writer)
```

L'étape suivante consiste à trouver la liste des faits temporels à partir de la liste d'indicateurs mise par défaut ou introduite par l'expert.

```
Algorithme : getTemporalFacts(tToken, propFile)
buff <= ReadFile(propFile)
tf<=ListFacts(buff,tToken)
removeDuplication(tf)
```

La procédure d'extraction continue pour extraire les faits reliés, à partir de la liste de faits temporels trouvés. On cherche à extraire un nombre maximal de faits reliés pour avoir plus de résultats.

```
Algorithme : findCouple
getPairListAtt(file,tf)
```



```
listToken<=findTokens(tf,tp)
hashSet<=getHashSetList(listToken)
printList(hashSet)
```

Le choix de l'expert consiste à sélectionner le couple de faits pour lancer automatiquement l'algorithme qu'on a conçu permettant de récupérer les informations nécessaires pour former des triplets annotés.

Notre besoin se résume comme suit :

```
if (x propTemp y) and (x prop*** z) then
(x prop*** z) y with (y is the temporal annotation)
```

Les résultats de la requête labelisés seront affichés dans un *TextArea* à l'expert en premier lieu puis ils peuvent être stocker dans un fichier sous forme de quadruplets RDF. Les quadruplets annotés seront sauvegardés dans le fichier nommé par l'expert dans le dossier de la base de donnée SPOTbase.

Pour obtenir des quadruplets on a essayé de trouver des propriétés qui se rencontrent avec une procédure automatique mais le choix de l'expert et la validation des résultats à toujours important. Cela peu avoir un impacte déterminant sur le reste du travail et la nature des résultats.

Récapitulatif

Nous avons construit à partir de la base de connaissance DBpedia une autre base de triplets temporairement annotés qu'on a nommé SPOTbase. SPOTbase contient une liste de fichiers dans lesquels on a des quadruplets (*subject, predicat, object, time*) avec une sémantique correcte.

Par la suite, nous avons choisi d'extraire des propriétés temporelles liées à une liste d'indicateurs pour tourner nos algorithmes.

Une première étape consiste à extraire tous les faits DBpedia et les stockés dans un fichier puis la seconde consiste à proposer une liste de couple de faits à l'expert.

Dans l'étape suivante on va interroger DBpedia en utilisant une requête SPARQL qui retourne trois variables (x,y,z) le choix de l'expert est primordiale pour valider les résultats trouvés. Enfin les ressources formées en quadruplets seront stockés dans SPOTbase.
(X prop_liée Y) annoté avec Z .

Notre nouvelle approche à la particularité de manipule directement les ressources de DBpedia. Des anciens travaux de recherches ont évoqué la même problématique en essayant de rattacher des annotations aux données de wikipedia alors que dans notre étude nous cherchons à regrouper des propriétés des triplets existants afin de former des triplets annotés.

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

- [VMS⁺10] Verhagen, M., Sauri, Caselli, T., Pustejovsky, and J. 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.