

La Web Semántica como plataforma para sistemas de recomendación

Guido Zuccarelli

21 de octubre de 2014

Índice general

1. Introducción	3
1.1. Evaluaciones entrecruzadas	3
1.2. La web semántica	5
1.3. Reviews en la web semántica	5
1.4. Organización	5
2. Caso de Estudio	7
3. Enfoque General	8
3.1. Aplicaciones en la Web Semantica	8
3.2. Principios	8
3.3. Tecnologías	8
3.4. Estrategia propuesta	8
3.5. Selección de vocabularios	9
3.5.1. Review Ontology	10
3.5.2. hReview	11
3.5.3. dataVocabulary	12
3.5.4. Schema.org	13
3.6. Recolección y extracción de los datos	15
3.6.1. Búsqueda	15
3.7. Evaluación de Calidad de los Datos	16
3.8. Curado de los Datos	17
3.9. Integración de los datos	17
3.10. Publicación del Dataset Curado	17
3.11. Explotación del Dataset	17
4. Selección de Vocabularios	18
5. Recolección de Datos	19
6. Evaluación de Calidad de los Datos	20
7. Curado de los Datos	21
8. Integración de los Datos	22

9. Publicación del Dataset Curado	23
10.Explotación del Dataset	24
11.Conclusiones y trabajo futuro	25
12.Bibliografía	26
A. Publicaciones	28

Capítulo 1

Introducción

1.1. Evaluaciones entrecruzadas

Llamaremos ítem, a un elemento del mundo real que es aprovechado por las personas, dicho ítem puede ser, un objeto, un servicio, una idea, un programa, etc. Además, debería poder ser abstraído a un modelo representado por una computadora, de manera tal, que esa representación describa con precisamente a qué se refiere ante la interpretación del lector.

Para ello, el modelo del ítem contendrá un conjunto de atributos que lo describen. Algunos más importantes que otros.

Por ejemplo, imaginemos que se debe representar una bicicleta mediante un conjunto de pares atributo;valor, tendríamos algo como esto.

Tipo: Bicicleta

Sexo: Unisex

Talle: 51

Cuadro-Material: Aluminio

Cuadro-color: Rojo

Cuadro-Tipo: Ruta

Horquilla-Material: Fibra de carbono

Horquilla-Color: Rojo

Asiento-tipo: Adamo

Asiento-Color: Blanco

La cantidad de atributos que pueden utilizarse para describir el ítem puede ampliarse prácticamente hasta el cansancio.

Se pueden optar también por una representación como ésta

Tipo: Bicicleta

Talle: 51

Marca: Merida

Modelo: Reacto 500

Puede aprovecharse el hecho de que es muy probable que ya existan definiciones del ítem que intentamos describir, por lo que si se utilizan sólo los atrib-



utos que identifican al ítem, será suficiente para que el lector tenga la interpretación correcta del mismo.

Cada ítem es de alguna manera utilizado por uno o más usuarios. Y cada usuario puede de la misma manera ser representado en la computadora mediante pares de atributo;valor. Y al igual que los ítems algunos atributos servirán para identificar al usuario.

Esta relación de uso entre usuario e ítem, contiene un grado de conformidad entre el primero y el segundo, si el ítem cumplió o no con las expectativas del usuario debería poder modelarse también para ser representado computacionalmente.

Definiremos a una reseña (de aquí en adelante review) representación mediante la computadora, de la medida de conformidad con respecto a dicha relación entre usuario e ítem.

El objetivo de un review, es que un usuario pueda reflejar el sentimiento que le generó utilizar el ítem e informarlo a otros usuarios.

También los reviews dispondrán de un conjunto de atributos de los cuales algunos serán indispensables y otros que enriquecerán no sólo su valor intrínseco, sino también su utilidad para el contexto que será planteado.

Imaginemos un usuario que realiza un review sobre la bicicleta, podría generar el siguiente review:

Puntuación: 4

Título: “Buena opción”

Texto: “Liviana y cómoda, pero un poco rígida para doblar.”

Fecha: 15/10/2012

Al terminar de leer esta sección, el lector debe entender a que te referís con reviews (en términos generales), que muchas de ellas son publicadas en la web en redes sociales, en sitios de productos etc, y que sirven para ???. Podés agregar algunos ejemplos e incluso imágenes. No queda claro por que en el titulo de la sección dice “entrecruzadas”, ¿es importante o fue solo una elección al azar?. Para que este se conecte con el que sigue, podés dejar picando algo como “muchos han intentado procesar automáticamente las opiniones de los usuarios para ... pero eso es muy difícil porque...”

1.2. La web semántica

En sus comienzos en los 90, la web podía verse como un conjunto de sitios web que ofrecían una colección de documentos web, con el objetivo de comunicar información a los usuarios.

Con el correr de los años, múltiples tecnologías se fueron implementando y permitieron el desarrollo de una web mucho más grande y aprovechable... y esto es una referencia a un libro sobre web semántica [1]

Al terminar de leer esta sección, el lector debe entender que es la web semántica y a que apunta. Si dejaste picando el tema de automatización en el párrafo anterior, acá se va a imaginar porque hablas de ws. Habla muy brevemente de triplets y menciona RDF. También mencioná linked open data para darle una idea de que la web semántica es una web de datos interconectada. Podés tomar lo que ya escribiste en la propuesta. Ejemplifica con dbpedia o algo así... En un capítulo mas adelante vas a entrar en detalle en web semántica, rdf, etc. Acá contás solo lo suficiente para que se entienda lo que vas a proponer en la próxima sección

1.3. Reviews en la web semántica

Acá es donde presentas el problema a resolver. Ya contaste que son los reviews y por que alguien querría integrarlos. Ya contaste que es la web semántica y linked open data. Ahora tenés que contar que ha habido iniciativas para darle semántica a los reviews y que existen datos; ahora hay una posibilidad de aprovechar esa información. Y ahí decís algo como lo que dijiste al final de la propuesta: *El objetivo principal de esta tesis es evaluar la viabilidad, y entender los desafíos de la utilización de la información contenida en la web semántica en la construcción de sistemas de recomendación. Para eso, y con foco en el caso particular de opiniones de usuarios sobre distintos tipos de recursos se buscará: capturar, extraer, validar calidad, curar, integrar, publicar y explotar los datos disponibles.*

1.4. Organización

ya saben lo que vas a contar; acá les das una idea de como te vas a organizar para contarlos - puede ser algo como lo que está a continuación. Para el caso de estudio podés tomar el texto que escribimos en el artículo

El capítulo 2 presenta una aplicación de ejemplo que muestra claramente el problema que se quiere resolver y que servirá como referencia a lo largo de esta tesis. El capítulo 3 introduce los conceptos de Web Semantica, sus principios y tecnologías y presenta la estrategia de solución en términos generales. Los capítulos 4 a 10 discuten en detalle cada uno de pasos de la estrategia elegida. Finalmente, el capítulo 11 presenta los resultados observados, saca conclusiones al respecto, y plantea trabajo futuro. El anexo 12 presenta una publicación que fue resultado del trabajo efectuado en este trabajo de tesis.

Capítulo 2

Caso de Estudio

Capítulo 3

Enfoque General

Lo que vamos a construir es un caso de aplicación de la web semantica....

3.1. Aplicaciones en la Web Semantica

3.2. Principios

Dar historia (muy breve), definiciones (RDF, tripletas, owl, inferencia, ..)
Escribir a medida que lo necesitas-.

3.3. Tecnologías

Triplestores, frameworks, extractores, crawlers, motores de búsqueda semanticos...

3.4. Estrategia propuesta

Con el fin de crear una aplicación que satisfaga los requerimientos mencionados anteriormente, se debe encontrar un procedimiento que incluya desde obtener los datos relevantes de la web hasta llevarlos a un estado que permita una explotación satisfactoria.

El procedimiento debería incluir los siguientes pasos

Selección de vocabularios Tanto la recolección como la publicación de datos en la web semántica involucra la elección del o de los vocabulario/s que mejor modelen el dominio de problema, con la excepción que para su publicación existe la posibilidad de desarrollar uno propio que se ajuste correctamente en caso de que no se encuentre uno existente.

Recolección y extracción de los datos ?

Evaluación de Calidad de los Datos ???

Curado de los Datos ???

Integración de los datos ??

Publicación del Dataset Curado ??

Explotación del Dataset ??

acá sería bueno incluir un diagrama que muestre el pipeline. En la lista de arriba con un párrafo se explica cada paso. Luego, en las secciones que sigue se analiza mas cada paso, pero se lo plantea como un problema.. se analizan los retos; entonces los capítulos 4 a 9, explican como los resolviste. Alternativamente, se puede hablar menos acá, solo lo suficiente para que se entienda la estrategia general, y luego en los capítulos 4 a 9 se analiza el problema en detalle y se propone la solución. Con esta ultima forma, todo lo que se refiere a los vocabularios quedaría en el capitulo 4 y no acá. Puede ser mejor.

3.5. Selección de vocabularios

Seleccionar un vocabulario implica analizar varios aspectos del mismo, no sólo su definición e implementación, sino también el uso práctico dado por sus usuarios. En primer lugar se debe comprobar que los nombres de las propiedades que posee sean correctamente auto-explicativas. Supongamos por ejemplo que existe un ítem con un rating agregado modelado por una ontología que posee las siguientes propiedades:

1. minrating
2. ratingValue
3. countRating

Las dos últimas propiedades resultan fácilmente identificables, ratingValue se trata del promedio de puntaje, y ratingCount la cantidad de puntajes que le fueron otorgados, pero la propiedad minRating podría generar distintas formas de interpretación, alguien podría suponer que se trata del valor mínimo que fue adquirido por un usuario, o el valor mínimo que un usuario puede otorgar. Y muchas veces la documentación de la ontología (si es que existe) no es suficiente.

Luego se deberá analizar si existen las propiedades para cubrir las necesidades mínimas de los casos de uso.

Y por último se debería intentar buscar ejemplos reales que muestren el uso que le dieron los usuarios a la ontología, para determinar qué propiedades están

incorrectamente interpretadas o también para los casos donde las propiedades que se encuentren en desuso.

Con estas precauciones en mente se puede emprender la búsqueda, que podría tener como comienzo búsquedas en search engines. Existen dos buscadores específicos para esta tarea.

En la actualidad existen cuatro vocabularios que cumplen los requerimientos mínimos para modelar el dominio de problema planteado.

3.5.1. Review Ontology

RDF Review Vocabulary

También conocido como Review Ontology, es una de las ontologías más antiguas de review, que fue pensada para uso del lenguaje

RDF y está definido bajo el namespace <http://purl.org/stuff/rev#>.

Fue utilizado para la construcción Revyu, y sirvió como guía para otras ontologías.

Consta de tres clases y trece propiedades.

Clases

Comment: Un comentario sobre el review.

Feedback: Expresa la utilidad del review.

Review: El review mismo.

Propiedades

commenter: Especifica el usuario que realizó el comentario del review. Tiene dominio Feedback o Comment y rango foaf:Agent.

Actualmente se encuentra en desuso.

hasReview: Enlaza el ítem evaluado con el review. Su dominio es rdfs:Resource (siendo ésta la clase del ítem evaluado) y su rango es Review. Es una de las propiedades principales.

hasComment: Idem anterior pero con el comentario en lugar del review.

Se encuentra actualmente en desuso.

hasFeedback: Idem anterior pero con feedback en lugar de comentario.

Se encuentra actualmente en desuso.

maxRating: Establece el puntaje máximo que es posible otorgar por un usuario a través de la propiedad rating. Tiene como dominio Review y como rango Literal siendo este último un número positivo. Su ausencia en un review asume su valor por defecto (5). Por lo que si bien no es indispensable que esté, se deberá respetar la convención al momento de generar el rating.

minRating: De la misma forma que el anterior, sólo que establece el puntaje mínimo y su valor por defecto es (1).

positiveVotes: Se refiere a la cantidad de votos positivos que tuvo el review, otorgados por usuarios que lo leyeron y lo encontraron útil. Su dominio es Review y su rango Literal siendo este último un número positivo.

Se encuentra actualmente en desuso.

rating: Una de las propiedades principales, indica el valor numérico otorgado por el creador del review sobre el ítem evaluado. Su dominio es Review y su

rango es Literal siendo este último un número entre los valores de minRating y maxRating.

reviewer: Especifica el usuario que realizó el review. Tiene dominio Review y rango foaf:Person.

text: Otra de las propiedades principales, define el texto que describe el sentimiento del usuario hacia el ítem. Tiene como dominio Review y como rango Literal.

totalVotes: Exactamente igual a positiveVotes.

title: El título del review . Tiene dominio Review y rango Literal. Subclase de dc:title

No tiene demasiada utilidad para el caso de estudio y además se encuentra en desuso.

type: Enuncia el tipo de ítem que clasifica taxonómicamente al ítem evaluado. Su dominio es rdfs:Resource (siendo ésta la clase del ítem evaluado) y su rango no se encuentra especificado.

Es una propiedad muy útil pero actualmente se encuentra en desuso.

date: Si bien no está definida dentro del vocabulario, es correcto utilizar <http://purl.org/dc/terms/date>, implica la fecha en la que se realizó el review.

3.5.2. hReview

Como se mencionó anteriormente, está establecido por convención, que los microformatos son una forma de publicar información en la web semántica, pero al no disponer de namespaces, no pueden ser representados por ninguna ontología o ningún otro lenguaje de la misma.

Al no poder utilizar ontologías que modelen reviews surgió la necesidad de crear un estándar específico para embeberlos dentro de HTML utilizando microformatos.

Dicho estándar se encuentra actualmente en la versión 0.4 y propone el uso de 10 propiedades, que se supone, deberían ser suficientes para cubrir todas las necesidades a la hora de generar un review.

summary (opcional): Puede ser el título o nombre del review, o es posible también hacer una pequeña sinopsis del mismo. Se encuentra en desuso.

type (opcional): Representa el tipo de ítem evaluado, pero se encuentra acotado a alguno de estos product — business — event — person — place — website — url .

También está en desuso.

item: Esta propiedad enuncia toda la información que se crea necesaria para identificar al ítem, mínimamente los atributos nombre, url y foto.

Para lograr bajo una propiedad cubrir todos los atributos, el rango de la misma debería ser un hCard, que luego contendrá las propiedades mínimas necesarias: fn, url, photo y cualquier otra que se quiera adicional.

reviewer (opcional): Al igual que ítem, indica todo lo necesario para identificar a la persona autora del review, para lo cual también deberá representarse con un hCard.

dtreviewed (opcional): Se refiere a la fecha en la que fue creado el review.

rating: El valor numérico con el cual el usuario expresa su satisfacción con el ítem, y está formado por un entero con un solo decimal de precisión, que se encuentra dentro del rango 1.0 a 5.0. Dicho rango puede ser alterado con la presencia de las propiedades worst y best que restringen el valor mínimo y máximo respectivamente.

descripción (opcional): Establece el valor textual con el cual el usuario expresa su satisfacción con el ítem, creando una sinopsis detallada del mismo.

tags (opcional): Una etiqueta intenta establecer una idea acerca de qué se trata el contenido en una sola palabra para una rápida identificación o para mejorar las búsquedas.

permalink (opcional): Genera una URI que identificará al review creándole una especie de ID, que será útil para los casos donde se podría repetir la publicación del mismo.

license (opcional): Expresa la licencia del review.

El indicador “(opcional)” se refiere a si es indispensable para conformar un hReview o no, de manera tal que si no se encuentra una propiedad que no está marcada como opcional no podrá ser considerado un hReview.

Cabe destacar que muchas de las propiedades opcionales, podrían ser necesarias para el caso de estudio.

El problema con este vocabulario surge a la hora de trabajar con la información obtenida, que no puede ser representada por ningún otro lenguaje de la web semántica, por lo que se vió la necesidad de mapear las propiedades de hReview, a otro vocabulario que sí pueda.

Esto llevó a que en la web donde definen hReview lo consideren compatible con RDF Review Vocabulary, por lo que en Noviembre de 2007 se creó una herramienta que transforma de uno al otro hreview2rdfxml.xml, pero existe un problema con el rango de algunas propiedades, por ejemplo reviewer (propiedad homónima en ambos vocabularios, pero una con rango hCard y otra con rango foaf:Person).

En general si bien este vocabulario bien utilizado puede ser efectivo, resulta poco flexible y complicado de manejar por parte de quien quiera explotarlos, el motivo por el cual se ha vuelto muy popular es su facilidad para generarlos, dado que microformats es un lenguaje muy sencillo y cualquier persona con un relativamente mínimo conocimiento de HTML puede generar sin problemas un hReview, teniendo además herramientas online como opción, que generan el código a travez de un formulario. Como es el caso de <http://microformats.org/code/hreview/creator>.

3.5.3. dataVocabulary

RDF Data Vocabulary

En Mayo de 2009 Google anuncia la introducción de los llamados “Google Rich Snippets”, estos fragmentos enriquecidos son una convención de etiquetas (con soporte para RDFa-lite y microdata) que permitían agregar información útil a los SERP del buscador de Google. De manera tal que los datos que contenían estos fragmentos, recibían un tratamiento especial.

Sin embargo en el anuncio, Google revela que el soporte se limitó al uso de las clases y propiedades del vocabulario definido en una página notoriamente improvisada llamada <http://rdf.datavocabulary.org/> .

En ella se establecían modelos de clases para varios tipos de ítems, tales como Persona, Organización o Producto y también para los Reviews.

La clase Review, quedó definida bajo el namespace <http://data-vocabulary.org/Review> e incluía las siguientes propiedades:

itemreviewed: Enlace al ítem que está siendo evaluado.

rating: El valor numérico con el cual el usuario expresa su satisfacción con el ítem, tiene como rango un valor numérico bajo la clase `xsd:string` o `Rating`, y los valores posibles se encuentran en escala de 1 a 5, pudiendo la misma ser alterada con la presencia de las propiedades `worst` y `best` que restringen el valor mínimo y máximo respectivamente.

reviewer: El autor del review, su rango es `dvocab:Person` o `xsd:string`.

dtreviewed: La fecha en la que se realizó el review, no contiene un rango específico pero aclara que debe respetar el formato ISO para las fechas.

description: El cuerpo del review que representa el valor textual de satisfacción del usuario con el ítem.

summary: Un resumen corto del review.

Se puede notar la excesiva similitud de este vocabulario con `hReview`, queda claro que no hubo una intención de innovar algo, sino de representar el `hReview` en microdatos, probablemente por el apuro en la que `data vocabulary` fue creado.

Vale aclarar que limitar las clases y propiedades posibles en RDFa es básicamente hacerle perder el sentido al lenguaje (la descentralización de los vocabularios sobre los términos) haciendo que el lenguaje se utilice como si fuese microformatos, pero perdiendo su valor más importante (la simplicidad) de manera tal que tomó la inflexibilidad de microformatos y la complejidad de RDFa.

Más adelante los Google Rich Snippets incluyeron también soporte para microformatos (lo que incluía `hReview`).

3.5.4. Schema.org

Schema.org

Luego de los fragmentos enriquecidos y el improvisado vocabulario `data-vocabulary` creado para soportar los fragmentos, Google en conjunto con Bing y Yahoo crearon en el 2011 `Schema.org`, con el objetivo de obtener un vocabulario más completo y organizado para la implementación de los snippets con RDFa y microdata.

Su lanzamiento provocó la inmediata obsoletización de `data-vocabulary` cuyas clases fueron todas reemplazadas por equivalentes dentro de la nueva ontología:

<http://www.data-vocabulary.org/Address> -> <http://schema.org/PostalAddress>

<http://www.data-vocabulary.org/Geo> -> <http://schema.org/GeoCoordinates>

<http://www.data-vocabulary.org/Organization> -> <http://schema.org/Organization>

<http://www.data-vocabulary.org/Person> -> <http://schema.org/Person>

<http://www.data-vocabulary.org/Event> -> <http://schema.org/Event>

<http://www.data-vocabulary.org/Product> -> <http://schema.org/Product>

<http://www.data-vocabulary.org/Review> -¿<http://schema.org/Review>

<http://www.data-vocabulary.org/Offer> -¿<http://schema.org/Offer>

Este nuevo vocabulario, recibe constantes actualizaciones, al punto que, al día de la fecha, la ontología cuenta con 946 clases.

En particular, la clase review, definida bajo el namespace <http://schema.org/Review> y es subclase de `schema:CreativeWork` que a su vez es subclase de `schema:Thing`

·

Propiedades de Review

- `itemReviewed` El ítem que está siendo evaluado, tiene rango `schema:Thing`
- `reviewBody` El valor textual del review de la evaluación, que tiene rango `schema:Text`
- `reviewRating` El valor numérico del review de la evaluación, que tiene rango `schema:Rating`

Propiedades de CreativeWork

- `about` El tema del contenido. Rango `schema:Thing`.
- `aggregateRating` El promedio de la acumulación de uno o más ratings dentro de un review. Tiene rango `schema:AggregateRating`
- `author` El autor del contenido. Tiene rango `schema:Person` o `schema:Organization`
- `comment` Comentarios sobre el contenido. De rango `schema:Comment` o `schema:UserComments`.
- `commentCount` Cantidad de comentarios. Rango `schema:Integer` .
- `creator` El autor del contenido. Tiene rango `schema:Person` o `schema:Organization`
- `dateCreated` Fecha en la que fue creado. Rango `schema:Date` .
- `dateModified` Fecha en la que fue modificado por última vez. Rango `schema:Date`

·

- `datePublished` Fecha en de la primer publicación. Rango `schema:Date` .
- `publisher` El publicador. Tiene rango `schema:Organization`
- `review` El review sobre el contenido. Tiene rango `schema:Review` .
- `text` Contenido textual. Tiene rango `schema:Text` .

Properties de Thing

- `additionalType` Tipos adicionales para el ítem que en general son utilizados para especificar tipos externos, como podría ser por ejemplo una película de clase <http://schema.org/Movie> con el tipo adicional <http://dbpedia.org/ontology/> . Tiene rango `schema:URL` .
- `alternateName` Especifica un alias. Rango `schema:Text` .
- `description` Una descripción corta del ítem. Rango `schema:Text`
- `name` Establece el nombre. Rango `schema:Text` .
- `sameAs` Una referencia a una url que desambiguadamente represente al ítem, como podría ser una URL de wikipedia, dbpedia, freebase, etc . Rango `schema:URL`.
- `url` URL del ítem. Rango `schema:URL`

Unas 60 propiedades de CreativeWork fueron omitidas por no iban al caso ya que sólo tienen razón de ser para otras clases que también heredan de CreativeWork. Lo importante aquí es remarcar un par de situaciones interesantes:

=Las propiedades Review:reviewBody, CreativeWork:text y Thing:description podrían contener el valor textual de la evaluación del review semánticamente correcta por la definición de cada una.

=Las propiedades CreativeWork:author y CreativeWork:creator son exactamente iguales.

=Los valores CreativeWork:dateCreated CreativeWork:dateModified y CreativeWork:datePublished podrían generar confusión.

=Sería sintácticamente correcto que un Review utilice la propiedad Review

Más adelante se mostrarán y enumerarán los problemas que estas cuestiones (causadas por el afán de hacer uso de la herencia lo más posible) generan.

3.6. Recolección y extracción de los datos

Como se mencionó antes, la web contiene grandes cantidades de documentos publicados con información semántica. Pero la tarea de encontrarlos, con el agregado de que sólo una pequeña porción de ellos será relevante para los requerimientos no es trivial en lo absoluto debido a la inmensidad del universo en el que se encuentran. La forma de llevar a cabo este objetivo está atada al hardware disponible, tanto para almacenar los datos, como para el tiempo que va a emplear la ejecución de esta tarea.

Dado que las bases de datos semánticas sólo almacenan información en forma de tripletas o cuádrupletas, los documentos encontrados deberán someterse a un proceso de extracción que seleccione las sentencias HTML y las convierta a alguno de los lenguajes que soportan tripletas o cuádrupletas. Para esto existen múltiples herramientas.

Una vez encontrados, descargados y transformados los documentos HTML a documentos semánticos puede construirse la base de datos semántica con la información recolectada.

Estos son los cuatro pasos necesarios para lograr tener el dataset semántico con el cual se puede empezar a trabajar. Cada uno de ellos posee distintas alternativas para su realización, algunas se describirán a continuación

3.6.1. Búsqueda

OJO: ESTO ESTABA EN EL ARCHIVO `busqueda.tex` no en el `opciones-Busqueda`

La forma de ejecución de esta tarea dependerá de algunos aspectos:

Recursos de hardware disponibles

Cantidad y calidad de la información requerida

El grado de atemporalidad mínima tolerable en los datos

El primer paso para realizar la recolección es intentar responder la siguiente pregunta: ¿Dónde encuentro la información?

Una vez seleccionados los vocabularios, se necesitará obtener fuentes de datos que contengan sus datos publicados en esos vocabularios.

Para lograrlo se podrá utilizar como punto de partida:

Sitios indexadores: Son algunos sitios que disponen de un dataset muy grande procesado con documentos indexados, que ofrecen consultar dicho dataset mediante servicios web. Generalmente proveen una API donde se pueden consultar los datos mediante distintos grados de flexibilidad.

Síndice, LOD cloud cache y UriBurner son algunos ejemplos de estos sitios. Se puede utilizar entonces, los servicios web a fin de obtener una lista de URL que se cree que tendrán la información necesaria.

Sitios autoritativos: Son sitios conocidos que generan información relevante y la publican en las ontologías y vocabularios seleccionados. Ejemplos aplicados al caso de estudio podrían ser IMDB (que publica reviews de películas bajo el vocabulario schema), o Rottentomatoes (que hace lo mismo pero no solo con películas). Se podrán utilizar entonces estos sitios como punto base para un posible crawling.

Catálogos de endpoints: Son catálogos provistos por algunos sitios que mantienen una lista actualizada de SPARQL endpoints junto con el estado de disponibilidad en el que se encuentran. Por ejemplo los sitios <http://labs.mondeca.com/sparqlendpointsstatus/> y <http://www.w3.org/wiki/SparqlEndpoints> que este último además provee detalles sobre la información de los mismos.

Consultar estos sitios entonces generará una lista de endpoints SPARQL que se podrá utilizar para consultar la información necesaria.

Volcados de datos: Son datasets muy grandes que fueron el resultado de un web crawling, los cuales están disponibles para su descarga y se pueden utilizar para procesarlos y obtener los datos requeridos. El más importante es <http://challenge.semanticweb.org/2014/>.

3.7. Evaluación de Calidad de los Datos

Como sabemos los reviews son creados por usuarios que en su mayoría no están familiarizados con el desarrollo de una aplicación, esto causa que una parte muy significativa del contenido publicado no esté de la forma adecuada para ser procesado. La falta de calidad en el contenido publicado puede deberse tanto a errores por parte del usuario como por parte del publicador.

El publicador deberá asegurarse que los datos respeten rigurosamente las ontologías en las cuales se publican.

El sitio web en el cual el usuario se encuentre realizando el review deberá guiarlo en todo lo posible para lograr que la evaluación quede en un formato adecuado.

Aunque también existen muchos problemas que no dependen del sitio web, generalmente errores semánticos de calidad, donde lo redactado esta hecho de forma inconsistente o insuficiente.

Este paso entonces tiene por objetivo encontrar todos los problemas de calidad que pueda haber en el dataset que acaba de ser descargado y extraído y que generen inconvenientes para una posterior integración/explotación.

3.8. Curado de los Datos

En el paso anterior se describieron los problemas en la calidad del dataset que tendrán un impacto en la aplicación resultante, o que, podrían limitar o imposibilitar la realización de la misma. Dichos problemas pueden ser muchos, y algunos muy difíciles de solucionar, será parte entonces del proceso, reconocer aquellos que su resolución sea viable y además encontrar la forma de implementarla.

Cabe destacar que el proceso de evaluar el dataset en búsqueda de problemas y implementar soluciones es iterativo, debido a que una mejora puede conllevar a nuevos problemas.

3.9. Integración de los datos

Esto va acá o es parte de Explotación?

Los datos recolectados fueron generados por distintos usuarios, en múltiples sitios y bajo distintas ontologías y estándares. Estos aún curados, necesitan un último paso para poder realizar una explotación, y es la integración.

Este gran y dificultoso proceso abarca cualquier operación que intente dar una visión más unificada de la información. Este proceso puede tener distintos niveles y aspectos a integrar, como podría ser por ejemplo, unificar las ontologías de los reviews, o bien en una nueva ontología de review, o bien en una ya existente, para lograr tener información semánticamente más parecida. Ya que este proceso puede ser muy dificultoso, habrá que ver en base a los requerimientos, qué aspectos de los datos se integrarán.

Una vez realizada la integración mínima necesaria par also requerimientos, el dataset ya estará listo para su explotación.

3.10. Publicación del Dataset Curado

???

3.11. Explotación del Dataset

Acá es donde efectivamente se discute que implica cubrir los requerimientos de la aplicación . Puede ser que acá tambien te refieras a como la forma en la que querés explotar los datos impacta en las fases anteriores. Incluso, habría que pensarlo, en este capitulo tal vez conviene primero hablar de explotación y luego de las otras fases - porque explotación determina mucho que es lo que vas a mirar en las fases anteriores, ¿no?

Capítulo 4

Selección de Vocabularios

Capítulo 5

Recolección de Datos

Objetivo

Como se indicó anteriormente la información semántica que va a ser necesaria para construir se encuentra en la web en forma de documentos HTML que tienen la particularidad de ser muy efímeros, de manera tal que un documento que poseía datos relevantes a la fecha, puede al día siguiente, o dejar de estar disponible on-line, o haber cambiado de forma tal que la información de éste ya no es relevante, o ya no la posee.

En [] sección 4.2 Challenges for the selection of data sources se generó una estadística de este caso, donde se estableció que en promedio 62 % de los documentos encontrados, continuaban on-line luego de un año, y de estos, sólo un 56 % aún poseían datos relevantes.

Si bien armar un dataset con sólo información extraída de los documentos sin descargar estos últimos es posible, la situación anterior genera la necesidad de mantenerlos en una copia local para evitar una posible pérdida de los mismos.

El objetivo entonces será armar un repositorio local con los documentos on-line descargados que se cree que tienen la información necesaria.

Estrategia

Recolección y extracción de los datos Como se mencionó antes, la web contiene grandes cantidades de documentos publicados con información semántica. Pero la tarea de encontrarlos, con el agregado de que sólo una pequeña porción de ellos será relevante para los requerimientos no es trivial en lo absoluto debido a la inmensidad del universo en el que se encuentran. La forma de llevar a cabo este objetivo está atada al hardware disponible, tanto para almacenar los datos, como para el tiempo que va a emplear la ejecución de esta tarea. Dado que las bases de datos semánticas sólo almacenan información en forma de tripletas o cuádrupletas, los documentos encontrados deberán someterse a un proceso de extracción que seleccione las sentencias HTML y las convierta a alguno de los lenguajes que soportan tripletas o cuádrupletas. Para esto existen múltiples herramientas. Una vez transformados los documentos HTML a documentos semánticos puede construirse la base de datos semántica con la información recolectada.

Capítulo 6

Evaluación de Calidad de los Datos

Capítulo 7

Curado de los Datos

Capítulo 8

Integración de los Datos

Capítulo 9

Publicación del Dataset Curado

Capítulo 10

Explotación del Dataset

Capítulo 11

Conclusiones y trabajo futuro

Capítulo 12

Bibliografía

Bibliografía

- [1] Grigoris Antoniou and Frank Van Harmelen. *A Semantic Web Primer*.

Apéndice A

Publicaciones