La Web Semántica como plataforma para sistemas de recomendación

Guido Zuccarelli

October 1, 2014

Abstract

1

1 Introducción

1.1 Evaluaciones entrecruzadas

Definiremos a una reseña (de aquí en adelante review) como la forma que tiene un usuario de transmitir mediante el uso de la computadora, su grado de satisfacción con respecto a un ítem.
Estará conformado mediante una serie de atributos m'inimos necesarios y otros que ser'an que si bien no son indispensables, enriquecen no sólo su valor intrínseco, sino también su utilidad para el contexto que será planteado.
Llamaremos ítem, a un elemento que es aprovechado por las personas, dicho ítem puede ser, un objeto, un servicio, una idea, un programa, etc. Por ejemplo, una pelicula.
Este último debería poder ser identificable, para ello, será necesario como m'inimo un atributo (o un conjunto) que cumpla esa función, si tomamos como ejemplo un libro, sólo disponer del atributo ISBN, será suficiente apra identificarlo.
El review es una expresión de la persona que utilizó el ítem que refleja su grado de conformidad con el mismo, con el objetivo de informarlo a otras personas. Por lo que entonces, el atributo mínimo será aquel que refleje este sentimiento del usuario hacia el ítem, el mismo puede ser un texto explicativo (por ejemplo Ün pel'icula hermosa, pero me parecieron flojos lso actores¨) o un valor num'erico dentro de un rango determinado (7 en escala de 1 a 10).
A partir del atributo base, existen muchos otros atributos que, bien construidos, aportan mucha riqueza al aprovechamiento del review. Por ejemplo, el autor, que será un atributo que identifica a la persona creadora del review. Otro caso es el de la fecha.

1.2 La web sem'antica

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

1.3 RDF y OWL

1.4 Reviews en la web sem'antica

1.5 Linked data cloud

Con el fin de crear una aplicación que satisfaga los requerimientos mensionados 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
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 cuadrupletas, 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 cuadrupletas. 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. 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.
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.
Integración de los datos
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.