Mecanismos de Anotación semántica de Contenidos en Plataformas de Redes Sociales

Edwin F. Caldón[1]

Gustavo Uribe[1]

Diego M. López[1]

José Palazzo Moreira de Oliveira[2]

Leandro Krug Wives[2]

Resumo: La mayoría de la información que se encuentra en Internet es textual y sin ninguna estructura formal que permita a las máquinas entender y aprovechar dicha información de manera automática. En una red social todos los contenidos que están compartiendo son valiosos, por lo tanto es necesario definir mecanismos para dar estructura y enriquecer semánticamente la información existente, para que esta sea aprovechada por otras aplicaciones software. En este artículo se propone un mecanismo basado en ontologías de dominio para la anotación semántica de contenidos en plataformas de redes sociales que pueda ser implementado en plataformas existentes de código abierto para redes sociales. El mecanismo está basado en técnicas de Procesamiento de Lenguaje Natural y Anotaciones semánticas automatizadas de la Web semántica.

[1] Grupo de Ingeniería Telemática, Universidad del Cauca, Popayán, Colombia
{ecaldon, guribe, dmlopez @unicauca.edu.co}
[2] Instituto de Informática, UFRGS, Porto Alegre, RS, Brasil
{palazzo, wives @inf.ufrgs.br}

Abstract: Most of the information published in Internet is text, without any formal structure that allows machines to automatically analyze and understand this information. In a social network, all the content that is shared is valuable, therefore it is necessary to find mechanisms to add structure and meaning to the information, so it can be used by other software applications. In this paper, we propose a mechanism based on domain ontologies for semantic annotation of content published in social network sites that can be implemented into existing open source social network platforms. The mechanism is based on techniques of Natural Language Processing and Automated Semantic Annotation.

1 Introducción

Debido especialmente a la complejidad de la información en salud, el uso de tecnologías semánticas se constituye como una herramienta muy importante para soportar a los usuarios y gestores de información existente en redes sociales en salud, en los procesos de búsqueda y selección de contenidos. Desde el punto de vista semántico, el principal problema es que la mayoría de la información que se comparte en plataformas de redes sociales es entregada en documentos en texto plano, lo cual implica que no hayan meta-datos o información adicional que ayude a los buscadores web a procesar la información y sugerir a los usuarios el contenido más relevante.

La "web semántica" es una tecnología que entre otras cosas permite anotar formalmente los datos (dar más descripción acerca de los datos), mediante lenguajes especializados como XML, RDF u OWL para representar los conceptos de un dominio de conocimiento. Esto es, para cada área del conocimiento se agrupan conceptos y se relacionan de tal forma que describan el conocimiento existente de forma general o específica evitando ambigüedades conceptuales [1]. El uso de la Web semántica permite mejorar la calidad de los contenidos, especialmente su relevancia.

El objetivo de este artículo es proponer un mecanismo basado en ontologías de dominio para la anotación semántica de contenidos en plataformas de redes sociales que pueda ser integrado en plataformas existentes de código abierto para redes sociales.

2 Plataformas para Redes Sociales

Con el objetivo inicial de analizar los mecanismos existentes para la anotación semántica de contenidos en redes sociales, se hace un análisis de la arquitectura (componentes y sus interrelaciones) de las plataformas para redes sociales más importantes. Las plataformas para redes sociales a analizar incluyen OpenSocial de Google [2], el Framework de Desarrollado Web Django [3] y la plataforma de redes sociales Elgg [4].

2.1 La plataforma Elgg

Elgg es una plataforma abierta para redes sociales personales, desarrollado con tecnología LAMP (Linux, Apache, MySQL y Php) [4]. En la Figura 1 se muestra un esquema conceptual de los componentes de Elgg (Core de Elgg), donde la clase ElggEntity es la entidad que representa cualquier elemento de la plataforma, además de controlar permisos, propiedad, entre otros. El elemento ElggObject es un objeto que representa el tipo entradas como blogs, archivos y favoritos (bookmarks). El elemento ElggUser representa cada usuario del sistema y ElggSite representa cada sitio web creado en una instalación Elgg. ElggGroup es un sistema colaborativo multi-usuario, también llamado comunidades. Todos estos elementos (clases) al estar heredando de ElggEntity poseen propiedades y comportamientos comunes como el Identificador único global (GUID), permisos de acceso, información sobre el propietario o entidad a la que pertenece.

A cada una de las entidades se les puede agregar más información mediante metadatos como etiquetas (tags), número ISBN, ubicación del archivo o información del idioma. También es posible agregar anotaciones, que es la información adicionada por terceros como comentarios y puntuaciones [4].

Si bien Elgg captura metadatos de los datos compartidos (Metadata), esto es un proceso manual y subjetivo, además, tiene componentes bien definidos y la integración de aplicaciones se hace a través de APIs, por lo que la interacción con la plataforma es rígida y hay que adaptarse a las interfaces proporcionadas, creando una dependencia con la plataforma de desarrollo.

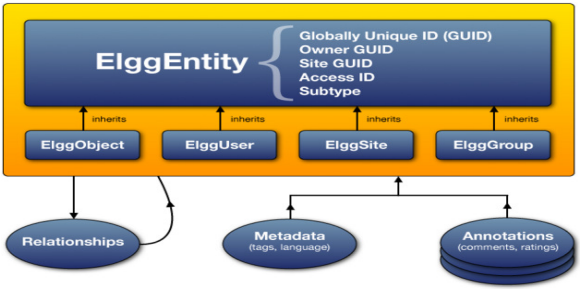

Figura 1: Modelo de datos de Elgg (http://docs.elgg.org/wiki/Engine/DataModel)

2.2 Modelo del Framework Django

En la Figura 2 se presenta la interacción de los componentes internos del framework Django al responder una petición web de un usuario o una aplicación. En este proceso se puede apreciar que Django proporciona herramientas básicas con las cuales se puede desarrollar cualquier aplicación web. Estas herramientas son: un servidor web de desarrollo; un Middleware; programas especializados en responder peticiones de los elementos internos del framework, por ejemplo mapear una petición URL (o dirección URL) a una funcionalidad del sitio como calcular la fecha (view middleware); y la herramienta de acceso a datos (Mapeador de objetos a tablas u ORM) [3].

Al tener bien definidos estos elementos básicos, Django provee flexibilidad para crear aplicaciones web complejas como redes sociales. La filosofía tras Django es desarrollar pequeñas aplicaciones con una funcionalidad bien definida, como autenticación, envío y recepción de correos, seguridad, entre otras. Esta ventaja ha permitido desarrollar iniciativas para web semántica, tal es el caso de django-rdf y django-foaf [5] que se pueden integrar a cualquier proyecto o sitio web. Del mismo modo existe Ruby on Rails (RoR), un framework de desarrollado web basado en el lenguaje de programación Ruby. El framework RoR tiene el plugin SWORD [6] para manipular ontologías en RDF. Aunque los enfoques de ambos frameworks son distintos, convergen en la idea de proveer plugins o módulos que permitan dar significado a los datos de las aplicaciones web, sin recurrir a terceros. Estos enfoques se desarrollan actualmente de manera exploratoria y no hay esfuerzos por parte de estas comunidades para desarrollar proyectos relacionados con la salud.

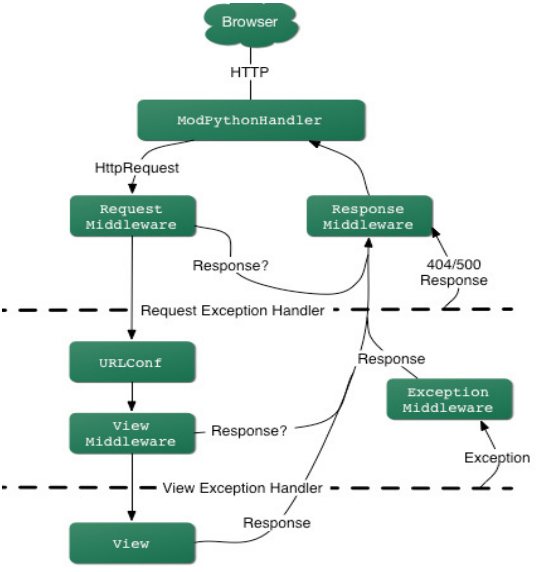

Figura 2: Proceso de peticiones en Django. Tomado de http://www.djangobook.com/en/1.0/chapter03/

2.3 Modelo del API Open Social

Open Social es un servicio abierto proporcionado por Google y sus socios a través de APIs para crear aplicaciones web [2]. En la Figura 3 se muestra que mediante las APIs se puede acceder a información del perfil de los usuarios (User data), información de los amigos y sus actividades (social graph) y gadgets (Templates) que se pueden embeber en cualquier sitio web. En otras palabras, las fuentes de datos y el grafo social pueden estar en Orkut, Yahoo, MySpace, LinkedIn, Hi5 o Ning, dotadas de gadgets (pequeñas aplicaciones o plugins) que tienen el mismo medio de comunicación: Open Social.

La gran ventaja de Open Social es la posibilidad de usar varios lenguajes de programación del lado del cliente para crear aplicaciones. A pesar de esta ventaja, aún no se provee un servicio que permita hacer anotaciones semánticas, que por otra parte, y aprovechando la estructura de esta API, podría cubrir diferentes niveles de datos como: perfiles, actividades y datos compartidos, de una manera efectiva. Otra desventaja del modelo Open Social es que no permite compatibilidad con otras aplicaciones que usen un API diferente como el API privativa de Facebook, por ejemplo.

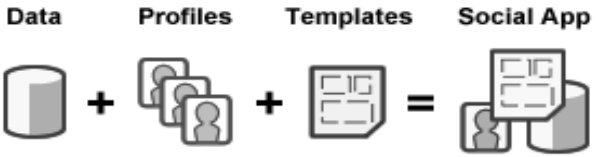

Figura 3: Estructura de Open Social. Tomado de http://wiki.opensocial.org/index.php?title=Articles_%26_Tutorials

3 Ontologías en el Dominio de la Salud

De acuerdo con [7], una ontología es una representación de los universales o las clases de la realidad y las relaciones existentes entre ellos, donde los universales son "los invariantes de la realidad o patrones en el mundo aprehendido por las ciencias específicas". Esta es sin embargo una definición desde el punto de vista de la filosofía que va más allá de la definición comúnmente aceptada en los círculos de la informática. Las ontologías en informática se relacionan principalmente con la representación del conocimiento. La apropiación del conocimiento de un dominio específico (por ejemplo la medicina) está basada en las terminologías, vocabularios y sistemas de clasificación existentes, así como en las especificaciones del significado de los términos y conceptos en un dominio (ontologías de dominio) y las relaciones entre ellos (a través de ontologías de nivel superior o de referencia), lo que permite la representación del conocimiento y la comunicación, el procesamiento de la máquina y la inferencia. Existe un sistema de ontologías que explica los diferentes tipos de ontologías, con la ontología filosófica (también llamada ontología

Universal) en la parte superior para explicar la naturaleza del mundo, seguido de ontologías de nivel superior u ontologías de referencia (Reference Ontology) que establecen las relaciones entre las ontologías de dominio (Domain Ontology) y finalmente ontologías de aplicación (Application Ontology) que representan los conceptos de un sub-dominio o área de aplicación en concreto [8].

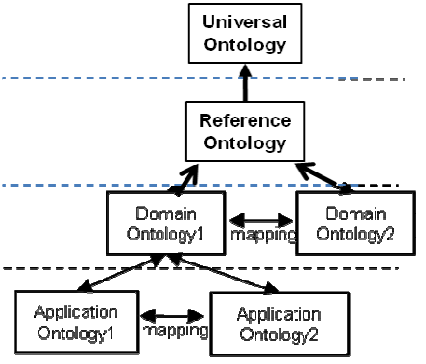

Figura 4: Una Jerarquía de Ontologías, adaptado de [8]

A continuación se presentan los principales enfoques ontológicos propuestos en el ámbito de la salud.

UMLS (Unified Medical Language System) es un meta-tesauro, es decir un amplísimo vocabulario, multipropósito y multi-lenguaje, que contiene información sobre conceptos médicos y biomédicos, incluyendo diferentes términos y sus relaciones. UMLS fue creado en 1986 por la Biblioteca Nacional de EE.UU. de Medicina (NLM) con el propósito de integrar la información de una variedad de diferentes fuentes bibliográficas. UMLS es actualmente la fuente más rica de la terminología biomédica, tesauros y sistemas de clasificación. Sin embargo, el uso de UMLS para aplicaciones más allá de la indexación de documentación médica por parte de la NLM se ve obstaculizado por el hecho de que muchos de sus fuentes están sujetos a licencias individuales [9][10]. Otra de sus desventajas es que no tiene una representación formal única.

La Nomenclatura Sistematizada de Términos Médicos-Terminología Clínica (SNOMED-CT) es una terminología completa y multiaxial, que intenta cubrir todo el campo de la medicina. Comprende las estructuras del cuerpo humano, los procedimientos médicos y otros aspectos pertinentes relacionados con la salud, incluyendo también el contexto social de las personas. Desde el punto de vista estructural, SNOMED CT proporciona jerarquías múltiples del tipo "es-un" que incluyen cerca de 310.000 nodos. Los nodos, que representan cada uno de los conceptos, se refieren sobre todo a las clases de entidades individuales (tales como las enfermedades, procedimientos, resultados de laboratorio, medicamentos, entidades geográficas, etc.), aunque hay todavía controversia sobre si SNOMET-CT se refiere a los

objetos en sí, por ejemplo, si el concepto de "dolor en el pecho" se refiere a la ocurrencia (evento) de un dolor en el pecho de un paciente determinado, o simplemente a su mención (registro) en la historia clínica del paciente. Desde abril de 2007, SNOMED CT es propiedad, mantenido, y distribuido por el Organismo Internacional de Desarrollo de Estándares para Terminologías en Salud (IHTSDO), una organización sin ánimo de lucro con sede en Dinamarca [11].

El sistema LOINC (Logical Observation Identifiers Names and Codes) es una terminología que se utiliza especialmente en el área de órdenes y resultados de Laboratorios clínicos. LOINC es un sistema multi-axial que se creó en 1994 como respuesta a la demanda de sistemas de intercambio electrónico de datos clínicos desde los laboratorios hacia los hospitales, consultorios médicos, etc. LOINC ha sido identificado por la Organización de Desarrollo de Estándares HL7 como el código preferente para los nombres de las pruebas de laboratorio en las transacciones entre los centros de salud, laboratorios, equipos de laboratorio de pruebas, y las autoridades de salud pública. A diferencia de SNOMED, los códigos LOINC no están organizados de forma simétrica o jerárquica, con lo que los códigos son asignados de manera arbitraria [12] [13].

El proyecto OBO (Open Biomedical Ontologies) es una iniciativa destinada a la construcción y mantenimiento de una colección evolutiva de ontologías interoperables que de forma inequívoca representen los tipos de entidades en la realidad biológica y biomédica. Esto incluye también un proceso de aseguramiento de calidad para todas las ontologías existentes y creadas por parte de la Fundación (OBO Foundry) que soporta el desarrollo de OBO. Mientras que las ontologías en OBO tienen la intención de representar a las entidades la realidad, los sistemas tradicionales de la terminología como los mencionados anteriormente, están diseñados para reflejar las declaraciones de los profesionales de la salud realizan cerca de esa realidad.

OBO es una especialización de la ontología formal básica (BFO). BFO es una ontología formal de alto nivel basada en principios probados para la construcción de ontologías, que se subdivide la realidad en dos categorías ortogonales [14]. BFO se utiliza como base para definir un subconjunto de ontologías de dominio construido para fines específicos y permite, por ejemplo, para superar la redundancia SNOMED-CT. Dependiendo del propósito y la granularidad del sistema biomédico que se considere, OBO define por ontologías para anatomía, genética, bioquímica, fenotipos, secuencias y técnicas de investigación. Además de OWL-DL, OBO utiliza un lenguaje propietario denominado OBO-EDIT [15]. Las ontologías OBO son de dominio público por lo tanto pueden ser usadas libremente por cualquier persona o institución.

4 Descripción del Servicio

La plataforma Elgg sirve como interfaz para interactuar con los recursos que ofrece una red social. La Figura 5 muestra de forma general el servicio de anotación, en términos de Elgg, un plugin. Este plugin consta de una herramienta para el Procesamiento de Lenguaje

Natural (Natural Language Processing) como Morphosaurus [16], la cual permite obtener los componentes léxicos y sintácticos del recurso, cabe aclarar que por el momento sólo se esta analizando recursos textuales como documentos de texto, entrada de blogs y/o páginas web.

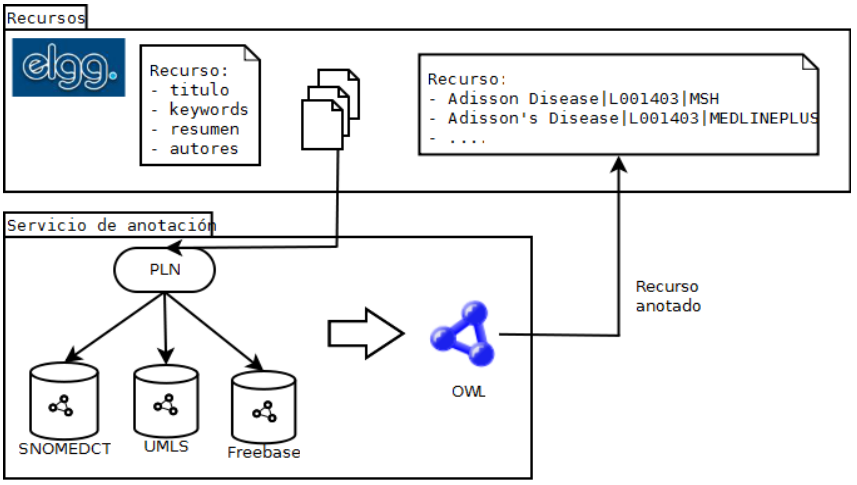

Figura 5: Descripción general del servicio de anotación semántico

Morphosaurus halla aquellas palabras y frases representativas en el documento mediante técnicas de Aprendizaje de Máquina eliminando las "stop words" (palabras comunes) y realizando steamming (obtener la raíz de la palabra) con sus respectivos pesos (importancia en el documento), este listado de palabras y frases no tienen ninguna relación entre si, no tienen semántica.

Con el listado de palabras y frases, se analizan frente a los repositorios de términos y conceptos médicos (SNOMED-CT, UMLS, OBO) y también frente a repositorios de conceptos y términos genéricos (Freebase). Por ejemplo el concepto "dolor de cabeza" puede tener varios términos asociados como "migraña", "dolor cabeza" o "dolor craneal" así que las palabras o frases se analizan con los términos de los repositorios antes mencionados. Este análisis arroja una estructura formal (Ontología) en formato OWL, donde las palabras y frases que inicialmente estaban aisladas y sin sentido ahora pueden estar relacionadas; pero esta relación no sería directa; puesto que esto implicaría crear una ontología donde se incluyan dichas palabras o frases.

Los términos y conceptos definidos a partir de las palabras y las frases obtenidas del texto, serán ahora los encargados de describir el documento; pero dichos descriptores son más ricos en información ya que están inmersos en unas ontologías de dominio médico y una ontología genérica.

Además de los nuevos descriptores del texto analizado, se cuenta con los metadatos tradicionales como resumen, palabras claves, tags (las palabras y frases obtenidas de la herramienta de PLN) y autores; este último elemento usa una ontología de descripción de perfiles como FOAF (Friend Of a Friend) y aprovechando que se está trabajando en el contexto de una red social, dicho perfil alimenta la información del cualquier recurso ya que informaría de los autores que están trabajando en la misma temática que quien escribió el recurso analizado.

5 Conclusiones

Actualmente las redes sociales ofrecen la posibilidad a los usuarios de enriquecer semánticamente sus contenidos mediante el uso de tags (etiquetas). Las etiquetas se adicionan al texto de forma manual por los usuarios que publican el texto, siendo sin embargo este proceso sujeto al conocimiento que los usuarios (editor y lector) tengan del dominio (por lo tanto el proceso de etiquetado es una tarea subjetiva) además del lenguaje que usan. Además, comúnmente este etiquetado (enriquecimiento semántico del contenido) no tiene una estructura formal que pueda ser procesada y entendida por aplicaciones y/o servicios como los buscadores.

Para soportar de manera automática e inteligente el proceso de anotación semántica de la información en sitios de redes sociales es posible el uso de la Web Semántica (WS), un área de estudio que entre otras cosas permite anotar formalmente los datos para dar significado al contenido de los sitios web, mediante lenguajes especializados como XML y RDF, además del uso de ontologías o conceptos relacionados de acuerdo a un dominio concreto de conocimiento. Este artículo propone un servicio basado en ontologías de dominio para la anotación semántica de contenidos en plataformas de redes sociales que pueda ser integrado en plataformas existentes de código abierto para redes sociales. El servicio está basado en técnicas de Procesamiento de Lenguaje Natural y Anotaciones semánticas automatizadas de la Web semántica.

6 Agradecimientos

Este trabajo ha sido soportado por el proyecto CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico/Cyted SALUS y CNPq Pro-Sul AVAL-SAÚDE, así como el proyecto QUIPU– UPCH, un programa auspiciado por el Fogarty International Center/National Institutes of Health (FIC/ NIH), proyecto: D43TW008438.

7 Referencias

[1] Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The semantic web revisited. IEEE Intelligent Systems.

[2] Good, R. (2007). Open Social: Google's New Social Networking Platform - What Is It And Why It Matters. Robin Good - MasterNewMedia. Recuperado a partir de http://www.masternewmedia.org/social_networking/social-networkingplatforms/Open-Social-Google-social-networking-platform-what-is-it-20071102.htm.

[3] Holovaty, A., & Kaplan-Moss, J. (2008). The Definitive Guide to Django: Web Development Done Right (Paperback) (First edition., pág. 447). Apress print book. Recuperado a partir de http://www.djangobook.com/en/1.0/chapter03/.

[4] Elgg site. (2010, Febrero 1). Elgg Data Model. Developer documentation. Recuperado a partir de http://docs.elgg.org/wiki/Engine/DataModel.

[5] Larlet, D. (2007). SemanticDjango Tools for semantic stuff in Django. Semantic Django. Recuperado Febrero 11, 2010, a partir de http://semanticdjango.org/.

[6] Mesnage, C., & Oren, E. (2007). Extending Ruby on Rails for Semantic Web Applications. Springer-Verlag Berlin Heidelberg, (Lecture Notes in Computer Science 4607), pp. 506–510.

[7] Smith B, & Brochhausen M. (2008). Establishing and Harmonizing Ontologies in an Interdisciplinary Health Care and Clinical Research Environment, eHealth: Combining Health Telematics, Telemedicine, Biomedical Engineering and Bioinformatics to the Edge IOS Press; 219 -33.

[8] Blobel B. Ontology driven health information systems architectures enable pHealth for empowered patients. Int J Med Inform. 2010 (In Press)

[9] Unified Medical Language System. (2010, Octubre 1). Disponible en http://www.nlm.nih.gov/research/umls/ .

[10] Freitas F., & Schulz S. (2009). Survey of current terminologies and ontologies in biology and medicine, Electronic Journal of Communication Information & Innovation in Health; 3 (1): 7- 18.

[11] International Health Terminology – Standards Development Organization. (2010, Octubre 1). Disponible en http://www.ihtsdo.org/snomed-ct/ .

[12] Sitio web LOINC Website. (2010, Octubre 2). Disponible en http://loinc.org/background .

[13] Jayaratna P., & Sartipi K. (2009). Tool-assisted HealthCare Knowledge to HL7 Message Translation, International Conference on Complex Medical Engineering, p. 1 – 7.

[14] Open Biomedical Ontologies http://www.obofoundry.org/. Last accessed November 2009.

[15] Smith B., Ashburner M., Rosse C., Bard J., Bug W., Ceusters W., Goldberg L. J., Eilbeck K., Ireland A., Mungall C. J., OBI Consortium, Leontis N., Roca-Serra P., Ruttenberg A., Sansone S. A., Scheuermann R. H., Shah N., Whetzel P. L., & Lewis S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration; 27 (11): 1251 – 5.

[16] P. Daumke; S. Schulz; M. L. Müller; W. Dzeyk; L. Prinzen; E. J. Pacheco; P. Secco Cancian; P. Nohama; K. Markó. (2010). Subword-based Semantic Retrieval of Clinical and Bibliographic Documents. Disponible en: http://morphwww.medinf.uni-freiburg.de/index.html.