Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 153-156 recibido 03-04-2017 revisado 12-05-2017 aceptado 18-05-2017

REDES: Reconocimiento de Entidades Digitales: Enriquecimiento y Seguimiento mediante Tecnologías del Lenguaje

REDES: Digital Entities Recognition: Enrichment and Tracking by Language Technologies

L. Alfonso Ureña López[1], Andrés Montoyo Guijarro[2], Mª Teresa Martín Valdivia[1], Patricio Martínez Barco[2]

[1] SINAI - Universidad de Jaén
Campus Las Lagunillas s/n, 23071, Jaén
{laurena,maite}@ujaen.es

[2] GPLSI - Universidad de Alicante
San Vicente del Raspeig, s/n, 03690, Alicante
{montoyo,patricio}@dlsi.ua.es

Resumen: El principal objetivo de este proyecto es el desarrollo de un modelo de integración capaz de definir y crear perfiles de entidades digitales. Estas entidades digitales incluirán no sólo las características básicas sino también sus rasgos lingüísticos y sociales, utilizando e integrando todas las fuentes de información disponibles. Concretamente se hará uso de tres tipos de fuentes en la Web: datos no estructurados, datos estructurados y datos abiertos enlazados. A partir de esta gran cantidad de información heterogénea, y mediante el diseño y desarrollo de herramientas, recursos y técnicas basadas en Tecnologías del Lenguaje Humano (TLH), se definirán y generarán entidades digitales entendidas como una estructura de información semántica donde encajar estos datos, con especial atención a las dimensiones espacial (ubicación geográfica) y temporal (variación de los datos que conforman la entidad a lo largo del tiempo).
Palabras clave: Procesamiento de lenguaje natural, PLN, análisis de sentimientos y opiniones, entidad digital, enriquecimiento semántico

Abstract: The main objective of this project is to develop an integration model able to define and create digital entities profiles. Such digital entities will include not only the basic, but also their linguistic and social features by means of using and integrating different information sources available. More specifically, three will be the Web sources: unstructured and structured data, but and also linked open data. Starting from this huge and heterogeneous amount of information, digital entities will be generated by means of the design and development of tools, resources and techniques based on NLP. Such entities will consist in a structure of semantic information where to place such data (with special attention to the spatial dimensions (geographical location) and temporal (variation of data that compose the entity during time).
Keywords: Natural language processing, NLP, sentiment analysis, opinion mining, digital entity, sentiment enrichment

1 Introducción

Actualmente, la Web 2.0 está cambiando la sociedad en la que vivimos haciendo necesario hablar de identidad digital para referirnos a cualquier objeto que deja un rastro en Internet a través de la generación de contenidos en la Red. Cada vez es más común que no solo las personas sino cualquier entidad (ya sea una empresa, un partido político o una ciudad)

ISSN 1135-5948

tengan un perfil digital asociado a redes sociales, blogs, portales administrativos o gubernamentales. Además, la información asociada a estas entidades digitales empieza a enlazarse y entremezclarse entre los distintos tipos de información (estructurada o no, multimodal y multilingüe, abierta o privada).

Durante los últimos años han aparecido sistemas que tratan de gestionar y analizar los documentos de la web social. Sin embargo, tales sistemas se centran en analizar la propia información más de una manera genérica y aislada que como datos asociados a una entidad digital, entendiendo ésta como un conjunto de características y relaciones en el mundo digital. Precisamente, consideramos que el concepto de entidad digital y su explotación en distintas aplicaciones es lo que generará un valor añadido a nuestros sistemas, aportando un avance significativo en la integración de conocimiento ya no solo de la web social sino de cualquier otra fuente de información disponible. Así, este proyecto identificará en primer lugar las entidades digitales y posteriormente las completará con toda la información extraída de los distintos medios. De esta manera, estas entidades serán enriquecidas semánticamente con el fin generar extensas pero depuradas bases de conocimiento que estarán a disposición de la comunidad científica para continuar explorando todo el potencial de la propuesta.

Así pues, el objetivo principal de este proyecto consiste en desarrollar un modelo de integración capaz de definir y crear perfiles de entidades digitales. Estas entidades digitales incluirán no solo las características básicas sino también sus rasgos lingüísticos y sociales, utilizando e integrando todas las fuentes de información disponibles. Concretamente, haremos uso de tres tipos de fuentes en la web: datos no estructurados, datos estructurados y datos abiertos enlazados. A partir de esta gran cantidad de información heterogénea, y mediante el diseño y desarrollo de herramientas, recursos y técnicas basadas en TLH, se definirán y generarán entidades digitales entendidas como una estructura de información semántica donde encajamos todos estos datos, con especial atención a las dimensiones espacial (ubicación geográfica de la entidad) y temporal (variación de los datos que conforman la entidad a lo largo del tiempo).

Desde el punto de vista científico-técnico, el proyecto plantea la combinación de modelos cognitivos del lenguaje, grandes bases de conocimiento públicas y enlazadas, y modelos multidimensionales de análisis para desarrollar métodos, recursos y herramientas eficientes y eficaces de extracción y análisis de cualquier información digital. El carácter abierto de la arquitectura diseñada contribuirá al desarrollo e integración en cualquier campo de la sociedad. Asimismo, este proyecto plantea un cambio de paradigma en el procesamiento de la información, apostando por una estrategia aglutinante de información con alto contenido semántico y su integración en la red de datos enlazados. Partiendo de una ontología núcleo y del lenguaje como rasgo principal para la definición de una entidad en el mundo digital, todo el proyecto es una semilla ambiciosa en la adquisición de conocimiento integrado a nivel global. Atributos adicionales, relaciones con otras entidades, su enriquecimiento con datos procedentes de distintas fuentes, el desarrollo de sistemas inteligentes con estas entidades como fuentes de conocimiento y otras posibilidades se abren ante este nuevo paradigma.

Los resultados esperados del proyecto REDES tendrán un impacto directo, ya no solo en empresas dedicadas expresamente al seguimiento y análisis de productos y servicios, sino en cualquier organización pública o privada que desee generar conocimiento a partir de las entidades digitales identificadas y procesadas.

2 Objetivos

El presente proyecto implica una serie de retos y objetivos específicos del proyecto global en el ámbito de la investigación de las TLH que se detallan a continuación:

O1: Definir entidades digitales. La definición de entidades digitales supone la determinación de un constructo que represente de una manera genérica a una entidad del mundo real. La entidad digital no sólo estará compuesta por datos presentes en Internet, sino también por información elaborada a partir de los datos que se identifiquen en la Red sobre dicha entidad.

O2: Procesar información heterogénea procedente de la web y web social. La web social, surgida de la transformación que supuso la Web 2.0, ha generado nuevos tipos de datos relacionados con la interacción entre personas y entes en la Red. El objetivo se centra en mejorar

la adquisición y producción de información a partir de datos no estructurados de la web, en general, así como su combinación con la información procedente de las relaciones de las entidades presentes en los datos no estructurados de la web social.

O3: Procesar información heterogénea procedente de la web de datos: La reutilización de información procedente de fuentes de datos abiertos y fuentes de datos abiertos enlazados supone un nuevo reto que proporcionará un salto cualitativo en cuanto a la generación de información y conocimiento. Para ello es necesario el desarrollo de nuevas metodologías, técnicas y recursos que permitan la correcta extracción de los datos procedentes desde las diferentes fuentes de la web de datos (web 3.0) para su posterior integración con el resto de datos disponibles.

O4: Enriquecer semánticamente las entidades digitales: La combinación de la información y conocimiento derivados de los objetivos 2 y 3 procedentes de la web, la web social y la web de datos debe formalizarse en la entidad digital mediante diferentes técnicas de homogeneización de dicha información y conocimiento.

O5: Monitorizar en el tiempo y en el espacio las entidades digitales: La información que caracteriza a una entidad digital es susceptible de ser modificada por la acción del contexto temporal y espacial en el que se desarrolla. La recuperación, extracción y normalización de la información temporal y espacial que acompaña a las propiedades de la entidad permitirá contextualizar el conocimiento de manera dinámica mediante su evolución a lo largo del tiempo o situándose en áreas geográficas diferentes.

O6: Integrar la información generada en el modelo de entidad digital: La definición, implantación y evaluación del modelo de integración del conocimiento junto con la plataforma que recoge todas las herramientas, técnicas y recursos enumerados anteriormente será otro de los grandes retos a abordar por el proyecto.

Para la consecución del objetivo global y los objetivos específicos del proyecto global anteriores, se propone la coordinación de dos subproyectos complementarios cuyos objetivos específicos particulares abarcarán los objetivos globales planteados, y cuya reunificación aportará el valor añadido que se busca con la coordinación.

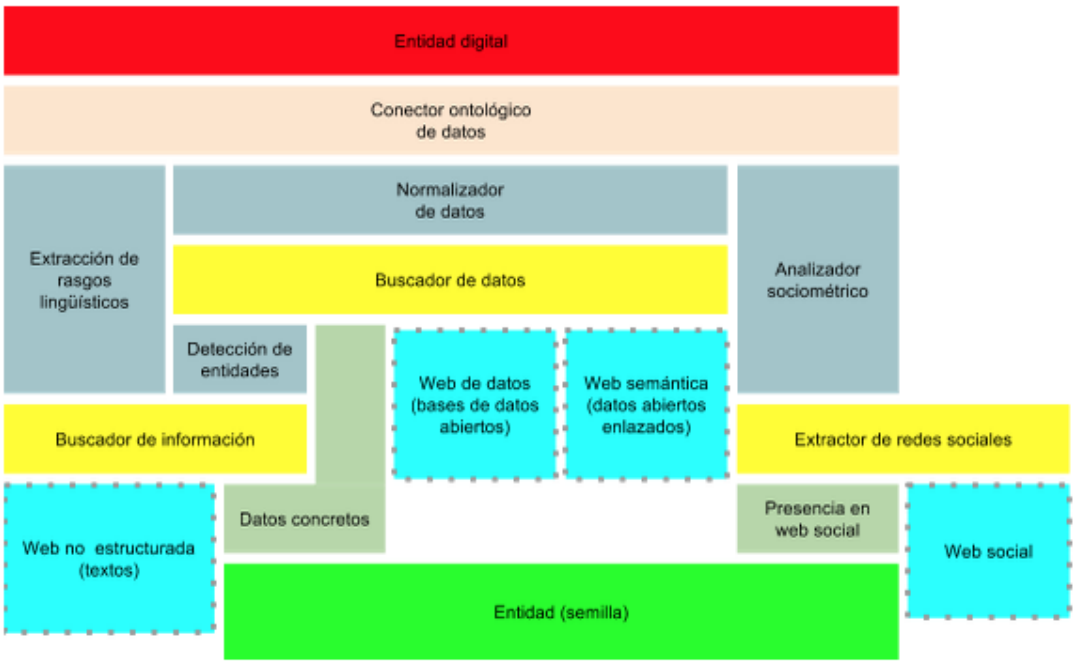

Figura 1. Modelo de integración de entidades digitales

3 Propuesta

El objetivo principal de este proyecto consiste en desarrollar una plataforma en la que se integren las distintas técnicas, recursos y herramientas de TLH con el objetivo de implementar sistemas capaces de definir y crear perfiles de entidades digitales. Estas entidades digitales incluirán no solo las características básicas sino también sus rasgos lingüísticos y sociales, utilizando e integrando todas las fuentes de información disponibles. Concretamente haremos uso de tres tipos de fuentes disponibles en la Web:

1. Fuentes de datos no estructuradas: principalmente las relativas a la Web Social (blogs, microblogs, comentarios, foros y redes sociales), aunque también desde fuentes formales como periódicos y portales de noticias. Se produce aquí un intenso proceso de análisis de texto para la extracción de la información.
2. Fuentes de datos estructuradas: en formato digital, pero sin estructura semántica (ontológica), como pueden ser bases de datos públicas y portales de transparencia con datos abiertos.
3. Fuentes de datos abiertos enlazados: para la extracción de información de fuentes semánticas, con ontologías definidas y sobre las que hemos llegado a un acuerdo ontológico en el mapeado de sus datos (aserciones) sobre el esquema ontológico definido en nuestro sistema.

A partir de este magma de información, y mediante el diseño y desarrollo de herramientas y técnicas basadas en TLH, se definirán y generarán entidades digitales entendidas como una estructura de información semántica donde se integran todos estos datos, con especial atención a las dimensiones espacial (ubicación geográfica de la entidad) y temporal (variación de los datos que conforman la entidad a lo largo del tiempo).

La figura 1 muestra la manera en la que se pueden integrar distintos componentes para construir un sistema capaz de integrar entidades digitales, con el objeto que permita la gestión y seguimiento de entidades digitales.

El diseño de los módulos del plan de trabajo propuesto se corresponde con las líneas de actuación marcadas en los objetivos del proyecto.

En el módulo 1 se gestiona el proyecto y se diseñan mecanismos de coordinación que permitan una comunicación fluida y una colaboración eficiente entre los distintos miembros del proyecto. El módulo 2 se centra en la identificación y especificación de entidades digitales. En el módulo 3 se desarrollan sistemas de recuperación de información de la Web heterogénea. El módulo 4 contempla el tratamiento inteligente de la información heterogénea en la web. Finalmente, mediante el módulo 5, se implementará la arquitectura que se describe a continuación y que permitirá la gestión y seguimiento de entidades digitales

En el tiempo en el que el proyecto lleva en ejecución, los trabajos realizados se han materializado en diferentes contribuciones como publicaciones en revistas, congresos, organización de eventos o participación en evaluaciones competitivas (Jiménez-Zafra et al., 2016) (Plaza del Arco et al., 2016) (Fernández et al., 2017) (Gutiérrez et al., 2016).

Agradecimientos

El proyecto REDES está financiado por el Ministerio de Economía y Competitividad con número de referencia TIN2015-65136-C2-1-R y TIN2015-65136-C2-2-R.

Bibliografía

Fernández, J., F. Llopis, P. Martínez-Barco, Y. Gutiérrez, y A. Díez. 2017. Analizando opiniones en las redes sociales. *Procesamiento del Lenguaje Natural*, 58: 141-148.

Gutiérrez, Y., S. Vázquez, y A. Montoyo. 2016. A semantic framework for textual data enrichment. *Expert Systems with Applications*, 57: 248-269.

Jiménez-Zafra S.M., M.T. Martín-Valdivia, E. Martínez, y L.A. Ureña. 2016. Combining resources to improve unsupervised sentiment analysis at aspect-level. *Journal of Information Science*, 42 (2): 213-229.

Plaza del Arco, F.M., M.T. Martín-Valdivia, S.M. Jiménez-Zafra, M.D. Molina González, y E. Martínez-Cámara. 2016. COPOS: Corpus Of Patient Opinions in Spanish. Application of Sentiment Analysis Techniques. *Procesamiento del Lenguaje Natural*, 57: 83-90.