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.