AORESCU: análisis de opinión en redes sociales y contenidos generados por usuarios

AORESCU: Opinion Analysis in Social Networks and User-Generated Contents

José A. Troyano Jiménez
ITALICA - Universidad de Sevilla
Av. Reina Mercedes, s/n. 41012 - Sevilla
troyano@us.es

L. Alfonso Ureña López
SINAI - Universidad de Jaén
Campus Las Lagunillas s/n, 23071 - Jaén
laurena@ujaen.es

Manuel J. Maña López
LABERINTO - Universidad de Huelva
Carretera de Palos s/n, 21819 - Huelva
manuel.mana@dti.uhu.es

Fermín Cruz Mata
ITALICA - Universidad de Sevilla
Av. Reina Mercedes, s/n. 41012 - Sevilla
fcruz@us.es

Fernando Enríquez de Salamanca Ros
ITALICA - Universidad de Sevilla
Av. Reina Mercedes, s/n. 41012 - Sevilla
fenros@us.es

Resumen: El proyecto AORESCU tiene como objetivos la recopilación y el procesamiento de la información generada por los usuarios sobre una entidad con idea de obtener a partir de ella una serie de indicadores que permitan evaluar la imagen que los usuarios tienen de la misma. La información recuperada puede ser estructurada (p.e. valoraciones numéricas) y no estructurada (fundamentalmente en forma de textos en lenguaje natural). Las técnicas y herramientas utilizadas en el proyecto son adaptables a cualquier dominio. No obstante, se ha elegido el ámbito turístico como dominio de aplicación al tratarse de un sector con una importante actividad económica y para el que es fácil encontrar contenidos para analizar. El proyecto tiene cuatro partes fundamentales: la recuperación de información de distintas fuentes sobre las entidades que pertenecen al dominio de aplicación (hoteles, restaurantes, espacios naturales, monumentos,…), la definición de un modelo de datos para representar esta información, el desarrollo de herramientas de análisis de textos para procesar los comentarios de los usuarios y el desarrollo de una aplicación web que permita analizar los datos procesados.
Palabras clave: Análisis de Opiniones, Procesamiento de Lenguaje Natural, Recuperación de Información, Extracción de Opiniones

Abstract: AORESCU project main goals are focused on the retrieval and processing of information generated by users about an entity. The idea is to get insights from this information that help us to understand the perception of users about an entity. We can retrieve two types of information from web 2.0 sources: structured information (e.g. numerical rating) and unstructured (mainly in the form of texts in natural language). The techniques and tools used in the project are adaptable to any domain. We chose the tourism sector as application domain since it is a sector with an important economic activity and because it is easy to find user generated content about touristic resources. The project has four main phases: the retrieval of information from different sources about the entities (for the tourism sector, these entities are hotels, restaurants, natural spaces, monuments,...), the definition of a data model to represent this information, the development of text analysis tools to process user comments and the development of a web application to query and analyze the processed data.
Keywords: Opinion Analysis, Natural Language Processing, Information Retrieval, Opinion Extraction

1 Introducción

La necesidad de conocer qué se dice de una persona, una empresa o cualquier organización no es algo nuevo. Ya en 1852 un agente de prensa polaco llamado Romeike fundó en Londres la primera empresa de *press clipping*. Este nuevo modelo de negocio, en aquellos tiempos, consistía en elaborar informes basados en recortes de prensa para personajes públicos, que estaban interesados en saber qué se decía de ellos. Desde entonces, el valor de la imagen no ha dejado de crecer, hasta el punto de que en algunos casos ya no está claro qué es más importante para una empresa: invertir en mejorar su producto o invertir en imagen.

En la actualidad ya no sólo se trata de analizar medios tradicionales como los periódicos. La irrupción del concepto web 2.0, que posibilita a cualquier usuario publicar contenidos mediante distintos canales (foros, blogs, microblogs, redes sociales…), multiplica el número de fuentes de información y plantea nuevos problemas de análisis de las mismas.

Los retos que plantea la extracción de información desde estas fuentes son complicados y desde hace unos años están siendo abordados por investigadores en varios campos. En concreto, las áreas de trabajo denominadas análisis de sentimientos y minería de opiniones están centradas en la resolución de este tipo de problemas. En ambos casos se aplican técnicas propias del procesamiento del lenguaje natural y de la minería de textos para extraer conocimiento desde textos subjetivos. El análisis de sentimientos se centra en determinar la actitud del autor de un texto con respecto a un determinado tema. La minería de opiniones, por su parte, analiza los textos a un nivel de granularidad más fino y se plantea identificar qué opina el autor del texto sobre aspectos concretos del tema sobre el que escribe (un producto, una institución, una persona, un partido político…).

2 Objetivos

El objetivo principal del proyecto es el desarrollo de un sistema para el análisis de la opinión expresada en contenidos generados por usuarios sobre una determinada entidad (empresa, producto, institución, personaje...). La idea es obtener de forma automática una serie de indicadores que resuman la imagen que los usuarios tienen sobre la entidad en función de la información generada por ellos mismos.

Los objetivos específicos del proyecto AORESCU son:

- Procesar tanto información no estructurada (por ejemplo comentarios escritos en lenguaje natural) como estructurada (como por ejemplo vínculos entre usuarios, etiquetas o información temporal).
- Aplicar técnicas de minería de textos y de procesamiento del lenguaje natural para analizar las opiniones expresadas en textos sin ningún tipo de formato.
- Desarrollar una herramienta de análisis de textos adaptable a diferentes dominios con facilidad. Para ello se separan los aspectos genéricos del análisis de contenidos, de los propios de cada dominio, quedando recogidos estos últimos en una serie de recursos lingüísticos específicos del dominio.
- Definir una taxonomía de características para cada dominio que recoja los aspectos sobre los cuales el sistema será capaz de extraer las opiniones desde los textos.
- Desarrollar una aplicación concreta en el contexto del sector turístico, que permita identificar la opinión de un colectivo de usuarios sobre productos y servicios turísticos.

3 Propuesta

Los objetivos planteados en el proyecto presentan una serie de retos relacionados con distintas líneas de investigación del área del Procesamiento del Lenguaje Natural. En concreto, el proyecto requiere fundamentalmente de la aplicación de técnicas de recuperación de información y extracción de información. Las técnicas de recuperación de información son necesarias para acceder a los contenidos publicados por usuarios sobre las entidades objeto de análisis. Las dificultades que presentan las particularidades de los textos escritos por usuarios (como la baja calidad o la concisión) se unen a las típicas dificultades de

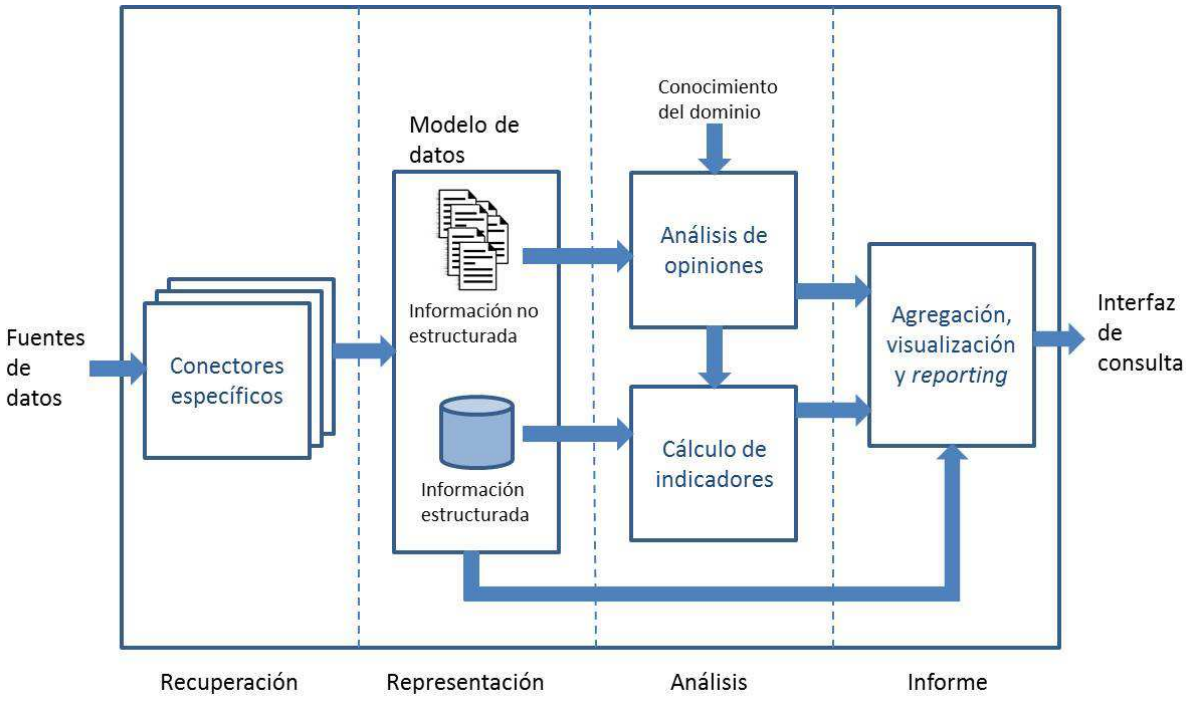

Figura 1. Arquitectura general del sistema

la recuperación de información propias de las distintas ambigüedades encontradas en cualquier texto en lenguaje natural. Las técnicas de extracción de información, por su parte, constituyen un elemento clave para el proyecto. Con ellas somos capaces de procesar los comentarios de los usuarios, que son las contribuciones más valiosas a la hora de obtener una imagen clara de la percepción que los usuarios tienen de una entidad. Si con las informaciones estructuradas es factible obtener una valoración cuantitativa que responda a la pregunta ¿cómo se valora? (por ejemplo agregando los *ratings* de distintos usuarios que han opinado), con la información no estructurada se pueden responder a preguntas más complejas, del tipo ¿por qué se valora? o ¿cuáles son los aspectos mejor y peor valorados? Evidentemente estas preguntas son mucho más interesantes pero también requieren de métodos más sofisticados para obtener la información que permita responderlas.

El proyecto AORESCU se articula en torno al desarrollo de un sistema que permita integrar las soluciones a los retos investigadores anteriormente mencionados en un entorno real. La Figura 1 muestra la arquitectura general del sistema, en ella se observan las cuatro fases principales en las que se ha dividido el proyecto: recuperación, representación, análisis e informe. La fase de recuperación contempla la implementación de conectores específicos para cada una de las fuentes analizadas (como por ejemplo redes sociales o webs de opinión). Dependiendo de las fuentes, esta recuperación se realiza con la ayuda de una API si ésta es ofrecida por la fuente o mediante *crawlers* específicos si no se dispone de dicha interfaz. La fase de representación está dedicada al desarrollo de modelos de datos que permitan almacenar la información extraída para su posterior procesamiento. La fase de análisis constituye el núcleo del proyecto y en ella se incluyen los algoritmos que permiten obtener indicadores de la percepción de los usuarios a partir del procesamiento de los contenidos publicados por ellos. Por último, la fase de informe tiene como objeto el desarrollo de una interfaz de consulta que permita explorar de forma dinámica el conocimiento generado a partir del análisis de las fuentes de información procesadas.

4 Resultados

AORESCU tiene un plazo de ejecución de 3 años de los cuales ya se han cubierto 18 meses. Nos encontramos, por tanto, en el ecuador del proyecto. En este período de tiempo se han

conseguido resultados tanto en el plano investigador como en el plano aplicado.

Aunque las técnicas aplicadas en el proyecto son aplicables a cualquier dominio, para poder llevar a la práctica estas ideas se hace necesario decidir un dominio de aplicación. Una vez que este dominio está concretado, se pueden validar experimentalmente las técnicas de adaptación a dominios específicos y también se puede concretar el desarrollo de un sistema siguiendo la arquitectura presentada en la figura 1.

Se ha elegido el sector turístico al tratarse de un sector con una importante actividad económica y para el que es fácil encontrar contenidos para analizar. A partir de ahí, se han elegido fuentes de datos para analizar distintas categorías relacionadas con el turismo (alojamiento, gastronomía, naturaleza, cultura, espectáculos y servicios). En este momento el sistema se encuentra desarrollado a un cincuenta por ciento. Se han desarrollado los conectores para las fuentes de datos, la capa de representación también está finalizada y se ha iniciado el desarrollo de los indicadores de análisis e interfaces de consulta para las categorías de alojamiento y gastronomía.

Se está siguiendo una metodología de desarrollo por fases en el desarrollo del sistema, aunque en paralelo se trabaja en el plano investigador experimentando con técnicas que permitan extraer información que pueda ser integrada en forma de indicadores en el sistema.

En el ámbito investigador, son varias las contribuciones que se han publicado en el período que lleva el proyecto. Los trabajos están relacionados fundamentalmente con las etapas de recuperación y análisis que son las que plantean los retos más interesantes desde el punto de vista investigador.

En el ámbito de la recuperación, el trabajo (Cotelo et al., 2014) presenta un método para obtener de forma automática consultas adaptativas a partir de un conjunto de *hashtags* semilla.

En el contexto del análisis de información no estructurada, se han publicado trabajos relacionados con la clasificación de documentos (Montejo-Ráez et al., 2013), con la extracción de información (Cruz et al., 2013) y con el análisis de estructuras y fenómenos lingüísticos en textos de opinión como son la negación y la especulación (Cruz Díaz et al., 2015).

También se han publicado trabajos que tienen como objeto la generación de recursos léxicos que sirvan de apoyo a tareas de análisis de opinión. En esta línea están (Molina-González et al., 2013) y (Cruz et al., 2014) en los que se presentan sendos métodos para la construcción de lexicones de palabras de opinión.

Agradecimientos

El proyecto AORESCU (P11-TIC-7684 MO) está financiado por la Consejería de Innovación, Ciencia y Empresas de la Junta de Andalucía.

Bibliografía

Cotelo, J.M., Cruz, F.L. Troyano, J.A. 2014. Dynamic topic-related tweet retrieval. *JASIST*. 65(3): 513-523

Cruz Díaz, N.P., Taboada, Mitkov, R. 2015. A Machine Learning Approach to Negation and Speculation Detection for Sentiment Analysis. *JASIST*. Pendiente de publicación.

Cruz, F.L., Troyano, J.A., Enríquez, F., Ortega, F.J., Vallejo, C.G. 2013. 'Long autonomy or long delay?' The importance of domain in opinion mining. *Expert Systems with Applications*. 40(8): 3174-3184.

Cruz, F.L., Troyano, B., Pontes, F., Ortega, F.J. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications.* 41(13): 5984–5994.

Molina-González, M. Dolores, Martínez-Cámara, Eugenio, Martín-Valdivia, M. Teresa, Perea-Ortega, Jose M. 2013. Semantic Orientation for Polarity Classification in Spanish Reviews. *Expert Systems with Applications*. 40(18):7250-7257.

Montejo-Ráez, Arturo, Martínez-Cámara, Eugenio, Martín-Valdivia, M. Teresa, Ureña-López, L. Alfonso. 2014. A Knowledge-Based Approach for Polarity Classification in Twitter. *JASIST*. 65(2):414-425.