

# Trabajo de doctorado: Recuperación de información orientada a la minería de opiniones

Javi Fernández, José M. Gómez, and Patricio Martínez-Barco

Universidad de Alicante

## 1. Introducción

El objetivo principal del trabajo de doctorado es el desarrollo de un sistema de recuperación de información orientado a la minería de opiniones, en el que se investiguen las influencias positivas entre las tres áreas que abarca: *recuperación de información* (RI), *rastreo web* (RW) y *minería de opiniones* (MO).

## 2. Descripción

Hasta la fecha se han realizado diversos trabajos relacionados en las diferentes áreas mencionadas, con el fin de realizar un acercamiento inicial a cada una de ellas y preparar las técnicas y herramientas que serán necesarias para alcanzar el objetivo final.

- Respecto a la MO, los estudios realizados se centran en la categorización de textos según categorías relacionadas con la subjetividad, como pueden ser la polaridad o la emoción. En ellos se han realizado diferentes experimentaciones utilizando recursos semánticos como *SentiWordNet* además de otras técnicas conocidas en PLN.
- Respecto a la RI, también se han estudiado las influencias de recursos semánticos en el cálculo de la relevancia de los documentos recuperados, tales como *clases semánticas* y *WordNet*. También se han realizado diferentes experimentaciones y posteriores evaluaciones de un sistema de RI con usuarios reales dentro de un ámbito concreto, oportunidades de negocio y exportación inteligente.
- Respecto al RW, se ha desarrollado un sistema de crawling muy adaptable y escalable, con el que se estudiarán las influencias de las relaciones entre documentos en la Web dentro de la RI y la MO. Debido a la falta de recursos que evalúen este tipo de sistemas desde el punto de vista científico, no se ha podido evaluar convenientemente. Este será uno de los objetos de estudio en posteriores trabajos.

En el último trabajo se ha realizado un estudio intensivo sobre la influencia de varias de las técnicas existentes en PLN en la tarea de clasificación de opiniones. Esta clasificación incluye, además de la tradicional categorización de objetividad y polaridad, categorías que reflejan otros aspectos relevantes de las opiniones:

la intensidad de la opinión y el tipo de emoción expresada. Evaluar este tipo de categorías ha sido posible gracias a la utilización del corpus *EmotiBlog*, un corpus de granularidad fina y una gran variedad de información anotada.

### **3. Trabajo futuro**

En la actualidad se está desarrollando un sistema de recuperación de opiniones en el marco de la telefonía móvil. El objetivo de este sistema es el de obtener, para un producto dado, todas las opiniones y su valoración general dentro de un conjunto de documentos pertenecientes a dominios web seleccionados. Para ello se han utilizado tanto herramientas existentes como desarrolladas específicamente para esta tarea, utilizando *InTime* como sistema de integración. Entre las herramientas existentes podemos destacar las siguientes: *Minipar* para el análisis léxico y búsqueda de relaciones entre palabras; *Weka* para el aprendizaje y clasificación de documentos según polaridad; y *Lucene* para la recuperación final de documentos. Las herramientas desarrolladas se enfocan en técnicas de rastreo y obtención de documentos en la Web, modelos de selección de términos, aproximaciones para la detección semiautomática de atributos y medidas para el cálculo de las valoraciones finales. Este sistema posteriormente será trasladado al ámbito socio-económico.