Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 159-162 recibido 15-04-14 revisado 30-06-14 aceptado 02-07-14

Análisis Semántico de la Opinión de los Ciudadanos en Redes Sociales en la Ciudad del Futuro

Opinion Mining in Social Networks using Semantic Analytics in the City of the Future

Julio Villena-Román
Adrián Luna-Cobos
Daedalus, S.A.
Av. de la Albufera 321
28031 Madrid, España
{jvillena, aluna}@daedalus.es

José Carlos González-Cristóbal
Universidad Politécnica de Madrid
E.T.S.I. Telecomunicación
Ciudad Universitaria s/n
28040 Madrid, España
jgonzalez@dit.upm.es

Resumen: En este artículo se presenta un sistema automático de almacenamiento, análisis y visualización de información semántica extraída de mensajes de Twitter, diseñado para proporcionar a las administraciones públicas una herramienta para analizar de una manera sencilla y rápida los patrones de comportamiento de los ciudadanos, su opinión acerca de los servicios públicos, la percepción de la ciudad, los eventos de interés, etc. Además, puede usarse como sistema de alerta temprana, mejorando la rapidez de actuación de los servicios de emergencia.
Palabras clave: Análisis semántico, redes sociales, ciudadano, opinión, temática, clasificación, ontología, eventos, alertas, big data, consola de la ciudad.

Abstract: In this paper, a real-time analysis system to automatically record, analyze and visualize high level aggregated information of Twitter messages is described, designed to provide public administrations with a powerful tool to easily understand what the citizen behaviour trends are, their opinion about city services, their perception of the city, events of interest, etc. Moreover, it can used as a primary alert system to improve emergency services.
Keywords: Semantic analytics, social networks, citizen, opinion, topics, classification, ontology, events, alerts, big data, city console.

1 Introducción

El objetivo final de las decisiones de las administraciones públicas es el bienestar ciudadano. Sin embargo, no siempre es fácil para los gestores identificar rápidamente los asuntos más importantes que afrontan sus ciudadanos y priorizarlos según la importancia real que los propios ciudadanos les asignan. El ciudadano se trata desde un punto de vista dual: como el principal usuario de los servicios que presta la ciudad, pero también, como un sensor proactivo, capaz de generar grandes cantidades de datos, por ejemplo en redes sociales, con información útil de su grado de satisfacción sobre su entorno. Por ello, el análisis de la opinión ciudadana es un factor clave dentro de la ciudad del futuro para identificar los problemas de los ciudadanos. Sin embargo, toda esta información no es realmente útil a no ser que sea automáticamente procesada y anotada semánticamente para distinguir la información relevante y lograr un mayor nivel de abstracción, y es aquí donde las tecnologías lingüísticas juegan un papel clave.

Este artículo presenta un sistema para el análisis en tiempo real de información en Twitter (no se aborda expresamente ninguna otra fuente de información). El sistema permite recopilar y almacenar los mensajes, analizarlos semánticamente y visualizar información

[1] Este trabajo ha sido financiado parcialmente por el proyecto Ciudad2020: *Hacia un nuevo modelo de ciudad inteligente sostenible* (INNPRONTA IPT-20111006), cuyo objetivo es el diseño de la Ciudad del Futuro, persiguiendo mejoras en áreas como la eficiencia energética, sostenibilidad medioambiental, movilidad y transporte, comportamiento humano e Internet de las cosas.

agregada de alto nivel. Aunque existen diversos trabajos que tratan el análisis semántico en redes sociales (TwitterSentiment, Twendz, SocialMention, etc.), no se conoce la existencia de un sistema que integre un análisis semántico completo y con capacidades de tiempo real, almacenamiento y capacidad de agregación estadística orientado a las ciudades inteligentes. Destaca un trabajo en esta línea (C. Musto et. al., 2014) pero centrado en análisis de cohesión social y sentido de pertenencia a la comunidad.

El objetivo último es proporcionar a los administradores públicos una herramienta potente para entender de una manera rápida y eficiente las tendencias de comportamiento, la opinión acerca de los servicios que ofrecen, eventos que tengan lugar en su ciudad, etc. y, además, proveer de un sistema de alerta temprana que consiga mejorar la eficiencia de los servicios de emergencia.

2 *Arquitectura del Sistema*

El sistema está formado por cuatro bloques principales, mostrados en la Figura 1.

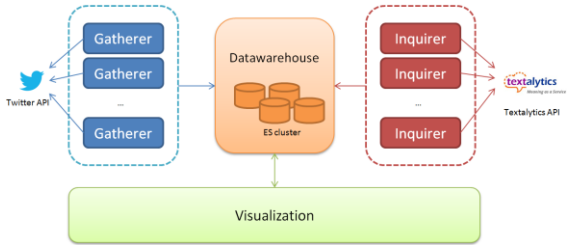

Figura 1: Arquitectura del sistema

El componente central es el ***datawarehouse***, el repositorio de información principal, capaz de almacenar el gran volumen de datos a los que hace frente el sistema además de proporcionar funcionalidad avanzada de búsqueda. Este componente se basa en Elasticsearch (Elasticsearch, 2014), motor de búsqueda en tiempo real, flexible y potente, de código abierto y distribuido. Su buena escalabilidad en escenarios con gran cantidad de datos fue el factor decisivo en la selección de esta tecnología.

El segundo componente lo forman un conjunto de procesos ***recolectores*** que implementan el acceso a los documentos vía consultas a las API de Twitter. Estos recolectores pueden ser configurados para filtrar tweets según una lista de identificadores de usuario, listas de palabras clave a seguir cómo términos o hashtags y localizaciones geográficas a las que restringir la búsqueda.

Un tercer componente, formado por un conjunto de procesos ***consumidores***, tiene como tarea anotar los mensajes de Twitter utilizando las APIs de Textalytics[2].

Se han diseñado dos modelos de clasificación temática (usando la API de clasificación de textos) específicos para este proyecto: *SocialMedia* y *CitizenSensor*, descritos más adelante. También se utiliza la API de extracción de *topics* para anotar entidades nombradas, conceptos, expresiones monetarias, URI, etc. Con la API de análisis de sentimiento se extrae la polaridad del mensaje, así como indicaciones acerca de su subjetividad o si expresa ironía. Por último, se utiliza *user demographics* para obtener información del tipo, sexo y edad del autor del tweet.

El proceso más pesado computacionalmente es la anotación semántica del texto y por lo tanto constituye el cuello de botella del sistema. Sin embargo, los procesos consumidores anotan los mensajes que aún no han sido procesados en orden descendente respecto a la fecha de indexación, de tal manera que la información más reciente siempre es la que está disponible primero. Esta característica es clave para poder reaccionar de forma temprana a alertas. Si el ratio de entrada de mensajes que los recolectores indexan en el sistema es mayor que lo que los consumidores son capaces de anotar, no será posible acceder a toda la información semántica de los mensajes en tiempo real, pero una vez que esta situación se revierta y el sistema consiga anotar a una velocidad mayor que los nuevos documentos, seguirá anotando los que quedaron sin procesar.

Por último, se ha definido un ***sistema de visualización*** para explotar los datos generados.

3 *Etiquetado semántico*

Se ha invertido un gran esfuerzo en la tarea de etiquetado semántico para este escenario particular: fragmentos cortos de texto, con capitalización inadecuada, faltas de ortografía, emoticonos, abreviaturas, etc. Los procesos consumidores proporcionan múltiples niveles de análisis según se describe a continuación.

En este primer despliegue se analizan exclusivamente tweets en español. Como las herramientas de procesamiento lingüístico

[2] http://textalytics.com

utilizadas en el sistema están ya disponibles para otros idiomas (inglés, francés, italiano, portugués y catalán), el sistema podría ser fácilmente extendido en estos aspectos. Sin embargo, sí habría que llevar un trabajo específico para migrar las ontologías de clasificación automática desarrolladas específicamente para este proyecto.

3.1 Clasificación automática

El algoritmo de clasificación de texto utilizado combina una clasificación estadística con un filtrado basado en reglas, que permite obtener un nivel de precisión bastante altos. Por ejemplo, evaluando el corpus Reuters-21578, se obtienen precisiones de más del 80% (Villena-Román et al., 2011). En concreto, se han diseñado dos ontologías específicas para este caso de uso.

El modelo de *SocialMedia* define los temas generales de clasificación, que ha sido desarrollado favoreciendo su precisión cuando se evalúan textos que proceden de redes sociales, respecto a los modelos generales ya disponibles y que se han usado satisfactoriamente en otros ámbitos.

Por otro lado, *CitizenSensor* se orienta a características propias del ciudadano como sensor de eventos de la ciudad, tratando de clasificar aspectos tales como su ubicación, eventos que ocurren en la ciudad o posibles catástrofes o alertas. Estos modelos se han desarrollado en base a reglas dónde se definen términos (o patrones) obligatorios, prohibidos, relevantes e irrelevantes para cada categoría.

3.2 Extracción de entidades

Este proceso se lleva a cabo combinando varias técnicas de procesamiento de lenguaje natural para obtener análisis morfosintáctico y semántico del texto y a través de estas características, identifican distintos tipos de elementos significativos.

Actualmente el sistema identifica (con flexión, variantes y sinónimos) distintos tipos de elementos: entidades nombradas (personas, organizaciones, lugares, etc.), conceptos (palabras clave relevantes para el texto tratado), expresiones temporales, expresiones monetarias y URIs.

Este análisis se apoya en recursos lingüísticos propios y reglas heurísticas. Evaluaciones internas sitúan al sistema en niveles de precisión y cobertura similares a otros sistemas de extracción de entidades.

3.3 Análisis de sentimiento

El análisis de sentimiento se realiza en otro nivel de análisis semántico, para determinar si el texto expresa un sentimiento positivo, neutral o negativo.

Este análisis se compone de varios procesos (Villena-Román et al., 2012): primero se evalúa el sentimiento local de cada frase y posteriormente se identifica la relación entre las distintas frases dando lugar a un sentimiento global. Además, empleando el análisis morfosintáctico, se detecta también la polaridad a nivel de a entidades y conceptos. El sistema ha sido evaluado en diversos foros obteniendo valores de medida-F superiores a 40%.

3.4 Características demográficas

Este módulo de análisis extrae características demográficas relativas al usuario que ha generado el texto analizado. Utilizando técnicas de extracción de información y algoritmos de clasificación, se estiman parámetros tales como el tipo de usuario (persona u organización), el sexo del usuario (hombre, mujer o desconocido) y su rango de edad (<15, 15-25, 25-35, 35-65 y >65 años).

Para realizar esta estimación, se utiliza la información del usuario en Twitter, el nombre asociado a su cuenta y la descripción de su perfil. El modelo se basa en n-gramas y ha sido desarrollado utilizando Weka.

3.5 Ejemplo de etiquetado

La Figura 2 muestra un mensaje anotado por el sistema, con salida JSON. Se pueden observar las distintas categorías asignadas para el modelo *CitizenSensor* (etiqueta "sensor"). El sistema identifica la ubicación del usuario (una vía pública), además de dos posibles alertas: aviso meteorológico por viento e incidencia por congestión de tráfico. Además, según el modelo *SocialMedia* (etiqueta "topic") se clasifica el mensaje dentro de la categoría de "medio ambiente". Posteriormente se muestran las entidades y conceptos detectados en el texto ("Gran Vía" y "viento", respectivamente), el análisis de sentimiento (se trata de un mensaje objetivo, no irónico y con polaridad negativa) y el análisis del usuario (el autor es una mujer de edad entre los 25 y los 35 años).

```
{
 "text":"el viento ha roto una rama y hay un
atascazo increible en toda la gran vía...",
 "tag_list":[
   {"type":"sensor", "value":"011002
Ubicación - Exteriores - Vías públicas"},
   {"type":"sensor", "value":"070700 Alertas
meteorológicas - Viento"},
   {"type":"sensor", "value":"080100
Incidencia - Congestión de tráfico"},
   {"type":"topic", "value":"06 medio
ambiente, meteorología y energía"},
   {"type":"entity", "value":"Gran Vía"},
   {"type":"concept", "value":"viento"},
   {"type":"sentiment", "value":"N"},
   {"type":"subjectivity", "value":"OBJ"},
   {"type":"irony", "value":"NONIRONIC"},
   {"type":"user_type", "value":"PERSON"},
   {"type":"user_gender", "value":"FEMALE"},
   {"type":"user_age", "value":"25-35"}
  ]
}
```


Figura 2: Ejemplo de anotación del sistema

4 *Módulo de visualización*

El módulo de visualización ofrece una interfaz web, que permite ejecutar consultas complejas de manera estructurada y presenta información de alto nivel, agregada y resumida.

La consola se define mediante elementos denominados *widgets*, configurados en una plantilla específica para los diferentes casos de uso del sistema y adaptada a cada necesidad. Para el desarrollo de los diferentes elementos se han utilizado librerías JavaScript existentes para la creación de gráficos[3], para la representación de mapas[4], y componentes propios.

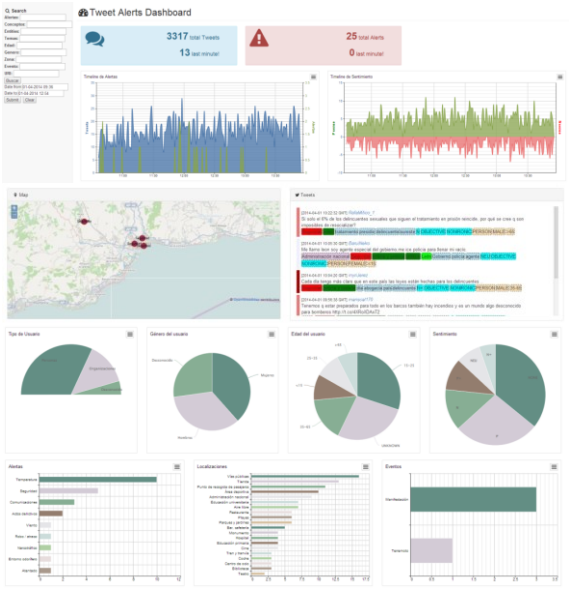

Figura 3: Consola de visualización

[3] http://www.highcharts.com
[4] http://openlayers.org

5 *Conclusiones y trabajos futuros*

Actualmente el sistema está en fase beta, acabando la puesta a punto de los diferentes módulos, y estará listo para ser desplegado en distintos escenarios a corto plazo. Las evaluaciones (informales) preliminares de la precisión de los diferentes módulos muestran que los resultados son totalmente válidos para cumplir con los objetivos de diseño del sistema.

Analizando el aspecto tecnológico, las capacidades de almacenamiento del sistema permiten, no sólo analizar los datos en tiempo real, sino también permiten aplicar algoritmos de minería de datos sobre los datos almacenados para, de esta manera, entender mejor las particularidades de la población, mediante técnicas de perfilado y *clustering* para identificar distintos grupos de ciudadanos que se encuentran en la ciudad, comparar singularidades entre los grupos detectados, etc.

Además se está investigando para explorar en el análisis de movilidad en la ciudad (cómo, cuándo y por qué los ciudadanos se mueven de un lugar a otro), la detección de los temas más relevantes a nivel de barrio o zona, y realizar un análisis de reputación o personalidad de marca.

Bibliografía

Musto, C., G. Semeraro, P. Lops, M. Gemmis, F. Narducci, L. Bordoni, M. Annunziato, C. Meloni, F.F. Orsucci, G. Paoloni. 2014. Developing a Semantic Content Analyzer for L'Aquila Social Urban Network. *In Proceedings of the 5th Italian Information Retrieval Workshop (IIR), Rome, Italy*.

Elasticsearch.org. Open Source Distributed Real Time Search & Analytics. 2014. *http://www.elasticsearch.org*

Villena-Román, J., S. Collada-Pérez, S. Lana-Serrano, and J.C. González-Cristóbal. 2011. Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization. *In Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-11), May 18-20, 2011, Palm Beach, Florida, USA. AAAI Press.*

Villena-Román, J., S. Lana-Serrano, C. Moreno-García, J. García-Morera, and J.C. González-Cristóbal. 2012. DAEDALUS at *RepLab 2012: Polarity Classification and Filtering on Twitter Data. CLEF 2012 Labs and Workshop Notebook Papers.*