Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 197-200 recibido 15-04-14 revisado 30-06-14 aceptado 02-07-14

track-It! Sistema de Análisis de Reputación en Tiempo Real

track-It! Real-Time Reputation Analysis System

Julio Villena-Román
Janine García-Morera
Daedalus, S.A.
Av. de la Albufera 321
28031 Madrid, España
{jvillena, jgarcia}@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: Este artículo presenta un sistema automático para recoger, almacenar, analizar y visualizar de manera agregada información publicada en medios de comunicación sobre ciertas organizaciones junto con las opiniones expresadas sobre ellas por usuarios en redes sociales. Este sistema permite automatizar la elaboración de un análisis de reputación completo y detallado, según diferentes dimensiones y en tiempo real, permitiendo que una organización pueda conocer su posición en el mercado, medir su evolución, compararse con sus competidores, y detectar lo más rápidamente posible situaciones problemáticas para ser capaces de tomar medidas correctoras.
Palabras clave: Reputación, extracción de información, análisis semántico, análisis de sentimiento, clasificación, opinión, redes sociales, RSS.

Abstract: This paper presents an automatic system to collect, store, analyze and display aggregated information published in mass media related to certain organizations together with user opinions about them expressed in social networks. This system automates the production of a complete, detailed reputation analysis, in real time and according to different dimensions, allowing organizations to know their position in the market, measure their evolution, benchmark against their competitors, and detect trouble situations to be able to take early corrective actions.
Keywords: Reputation, information extraction, semantic analytics, sentiment analysis, classification, opinion, topics, social networks, RSS.

1 Introducción

La reputación corporativa es el conjunto de percepciones que tienen sobre una organización todos los grupos de interés implicados: clientes, empleados, accionistas, proveedores, etc. Se ve afectada por todas las noticias sobre la organización en medios de comunicación y las opiniones, recomendaciones, etc. (beneficiosas o perjudiciales) de usuarios en redes sociales.

La gestión de esta información se convierte en algo cada vez más valioso, ofreciendo la oportunidad de responder a las expectativas sobre la organización y estar más protegidas frente a crisis eventuales. Pero el volumen de contenido es tan grande que las tecnologías de análisis automático se hacen indispensables para poder procesar toda esta información.

Este artículo describe track-It!, un sistema automático para recoger, almacenar, analizar y visualizar de manera agregada opiniones recogidas en Internet. Existen sistemas similares en el mercado, enfocados a vigilancia de marca (SproutSocial, comScore, Engagor, etc.) o al análisis de la voz del cliente (Vocus, Customerville, etc.).

El objetivo final es conocer "qué, cómo y cuánto" se dice de la organización en tiempo real, y medir cómo afecta esta información a su reputación y a la de sus competidores, actuando en consecuencia.

[1] Este trabajo ha sido financiado por los proyectos Ciudad2020: Hacia un nuevo modelo de ciudad inteligente sostenible (INNPRONTA IPT-20111006) y MA2VICMR: Mejorando el Acceso, el Análisis y la Visibilidad de la Información y los Contenidos Multilingüe y Multimedia en Red para la Comunidad de Madrid (S2009/TIC-1542).

2 *Arquitectura del sistema*

El sistema se compone de los cinco componentes mostrados en la Figura 1. El **recolector** de información es el responsable de capturar de manera continua y en tiempo real la información proveniente de diferentes fuentes de Internet. El **analizador** utiliza técnicas de procesamiento del lenguaje natural para procesar semánticamente la información. La información obtenida se almacena en el **datawarehouse**. El módulo de **agregación** selecciona, filtra, ordena y consolida toda la información referida a una misma organización y la analiza de forma global, obteniendo los valores de reputación agregados. Finalmente el componente de **visualización** presenta la información de forma gráfica e interactiva.

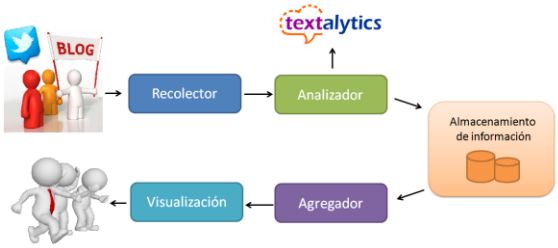

Figura 1: Arquitectura del sistema

3 *Recolector*

El recolector comprueba con una cierta periodicidad si en una serie de fuentes configuradas se ha recibido nuevo contenido. Se utilizan peticiones a la API de Twitter[2] para recolectar tweets con el nombre de la organización o alguno de sus alias. Además se utiliza un lector RSS para recoger las noticias publicadas en medios de comunicación relevantes.

4 *Analizador*

El analizador procesa y extrae información relevante de todo el contenido recogido de la red, empleando las funcionalidades lingüísticas de procesamiento y análisis de texto que ofrece Textalytics[3], nuestro portal de servicios lingüísticos en la nube:

1) Detección de idioma para filtrar el idioma de interés, si la fuente no lo incluye.
2) Clasificación automática, para determinar de qué aspectos del modelo de clasificación reputacional se habla (ética empresarial, organización de la empresa, trato a clientes).
3) Análisis de sentimientos para determinar si la polaridad del mensaje es positiva o negativa para la organización.
4) Detección de entidades para establecer la entidad a la que se refiere el texto y evitar problemas de desambiguación.

El objetivo es almacenar información clasificada sobre todos los aspectos que pueden afectar a la reputación de la empresa.

4.1 Detección de idioma

Se utiliza la detección automática de idioma cuando no se conoce el idioma a priori o no se puede extraer de la fuente. El detector se basa en técnicas estadísticas basada en la representación de N-gramas y es capaz de detectar hasta 60 idiomas distintos.

4.2 Modelo de clasificación

El módulo de clasificación automática de textos utiliza un modelo previamente entrenado para determinar la temática de la información. El algoritmo de clasificación utilizado consiste en un modelo híbrido que combina un procedimiento de clasificación estadístico con un filtrado basado en reglas (Villena-Román et al., 2011). El resultado es una lista de las categorías más representativas, ordenadas de mayor a menor relevancia. El clasificador ha sido evaluado con diferentes modelos de ámbito general, por ejemplo, Reuters-21578, en el que se obtienen precisiones superiores al 80%.

Categoría	Subcategorías
10 Oferta	11 Satisfacción necesidades
	12 Gestión de reclamaciones
	13 Trato a clientes
	14 Relación calidad/precio
	15 Calidad de productos y servicios
	16 Garantía de productos y servicios
20 Trabajo	21 Capacidad de los empleados
	22 Bienestar de los empleados
	23 Igualdad de oportunidades
	24 Remuneración a empleados
30 Integridad	31 Ética
	32 Asuntos corporativos
	32 Transparencia
40 Estrategia y liderazgo	41 Organización y estructura
	42 Liderazgo
	43 Visión y estrategia
50 Innovación y flexibilidad	51 Innovación
	52 Adaptabilidad al cambio
60 Responsabilidad social	61 Contribución a la sociedad
	62 Apoyo de causas sociales
	63 Protección del medio ambiente
70 Situación financiera	71 Potencial de crecimiento futuro
	72 Gestión financiera
	73 Resultados financieros

Figura 2: Ontología del modelo de reputación

El sistema emplea el modelo de reputación mostrado en la Figura 2. Para mejorar la

[2] https://dev.twitter.com/docs/api/1.1
[3] http://textalytics.com

precisión de la clasificación, se han desarrollado dos variantes de este modelo, una para textos largos (para noticias), entrenado estadísticamente con textos de entrenamiento e incluyendo en las reglas los términos empresariales más utilizados en cada categoría, y otro para fragmentos de texto cortos, en el que la clasificación se basa fundamentalmente en el filtrado mediante reglas muy específicas.

4.3 Análisis de sentimientos

El módulo de análisis de sentimientos permite determinar si la opinión o el hecho expresado en el texto es positivo, negativo o neutro, o bien si no expresa sentimiento. El análisis se basa en la información incluida en un modelo semántico que incluye reglas y recursos etiquetados (unidades con polaridad, modificadores).

Además, se utiliza un análisis morfosintáctico del texto para dividir el texto en segmentos, controlar mejor el alcance de las unidades semánticas del modelo, detectar la negación e identificar entidades. Se utiliza un algoritmo de agregación (basado en el uso de medias y desviaciones típicas que incluye la detección de valores *outliers*) para calcular el valor de la polaridad global del texto a partir de la polaridad de los diferentes segmentos y el de la polaridad final de las entidades y conceptos, a partir del valor de cada una de sus menciones.

El sistema ha sido evaluado en foros competitivos obteniendo valores de medida-F superiores al 40% y siendo el mejor sistema de los presentados (Villena-Román et al., 2012).

4.4 Extracción de entidades

La extracción de entidades se lleva a cabo mediante procedimientos de extracción de información basados en análisis morfosintáctico y semántico del texto, apoyado en recursos lingüísticos y reglas heurísticas, permitiendo identificar los elementos significativos. La salida será el listado de entidades encontradas junto a su información asociada (como el tipo de entidad). Evaluaciones internas sitúan al sistema en niveles de precisión y cobertura similares a otros sistemas existentes.

5 Almacenamiento de información

La información se almacena en un repositorio centralizado. Actualmente se utiliza una base de datos relacional (MySQL) pero el *roadmap* de producto plantea la migración del sistema a Elasticsearch (Elasticsearch, 2014), por sus características de tratamiento de grandes volúmenes de datos, eficiencia de búsquedas, escalabilidad y soporte a fallos.

6 Agregación de la información

La reputación de una organización en cada una de las categorías del modelo se calcula aplicando un algoritmo de agregación similar al del módulo de análisis de sentimientos, sobre todo el conjunto de textos comprendidos en un rango de fechas y provenientes de una o varias fuentes. Además, a partir de la reputación por categorías se obtiene la reputación general de la organización, teniendo en cuenta que las diferentes categorías no están interrelacionadas.

7 Visualización

Finalmente se muestra toda la información mediante tablas y gráficos.

Se ha desarrollado un escenario piloto de seguimiento de empresas del IBEX-35, recogiendo información en Twitter y en los medios económicos Expansión, Invertia, Intereconomía, El Economista y Cinco Días. En este caso sólo son de interés las entidades de tipo COMPANY o COMPANY GROUP, para resolver ambigüedades (p.ej., descartar "Santander" si se refiere a la ciudad).

Inicialmente la interfaz web muestra la información de la reputación de cada una de las empresas durante el día actual. Sobre esta tabla se pueden aplicar filtros: por rango temporal (día, semana o mes), tipo de empresas (tecnológica, financiera) o por el origen de la fuente de datos (Twitter o noticias RSS).

Figura 3: Información general

Por ejemplo, la Figura 3 muestra la información de las entidades financieras: Bankia, BBVA, Santander y Bankinter. La tabla presenta la reputación general y por categorías

y el número de menciones. Los colores indican la polaridad: rojo oscuro (N+), rojo (N), amarillo (NEU), verde (P) y verde oscuro (P+).

A partir de ella, se puede obtener información de detalle. La Figura 4 muestra la información específica de Bankinter: el número de menciones y la reputación asociada por cada categoría detectada. El gráfico circular indica la distribución de los textos por categoría y el gráfico de barras muestra la distribución de la polaridad de los textos en cada categoría. Además se puede obtener información de detalle de cada texto concreto.

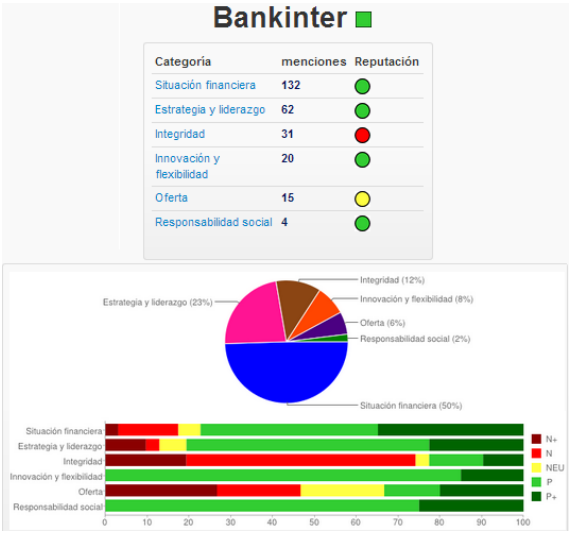

Figura 4: Información detallada

El sistema también ofrece información comparativa de las entidades seleccionadas, general y para cada categoría (Figuras 5 y 6).

Figura 5: Vista comparativa general

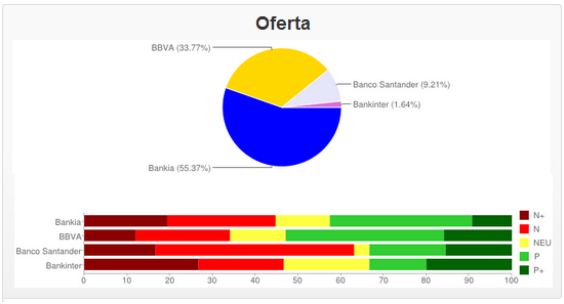

Figura 6: Vista comparativa de una categoría

8 *Conclusiones y Trabajos Futuros*

El sistema automatiza la recogida de información en la red y permite la elaboración de análisis de reputación. Se encuentra en fase beta y en proyecto de implantación en diferentes escenarios. En el piloto del IBEX35 se han recogido más de 10,2 millones de piezas de información desde agosto de 2013, con 500 mil de entidades y 17 millones de etiquetas.

De cara a un futuro sería interesante ampliar su funcionalidad a un mayor número de idiomas, recoger diferentes tipos de fuentes, por ejemplo, blogs o sitios web de opiniones, y calcular la reputación considerando no sólo las menciones de la empresa, sino también sus productos y servicios.

También se podría incluir la relevancia que puede tener un mensaje, estableciendo prioridades. Esto permitiría detectar alertas y dar avisos a la empresa para que puedan gestionar problemas rápidamente.

Bibliografía

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 Paper.*

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