Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 163-166 recibido 15-04-14 revisado 02-06-14 aceptado 10-06-14

TrendMiner: Large-scale Cross-lingual Trend Mining Summarization of Real-time Media Streams[1]

TrendMiner: Large-scale Cross-lingual Trend Mining Summarization of Real-time Media Streams

Paloma Martínez, Isabel Segura
Departamento de Informática
Universidad Carlos III de Madrid
pmf@inf.uc3m.es,
isegura@inf.uc3m.es

Thierry Declerck
Language
Technology Lab
DFKI
Saarbrucken, Germany
thierry.declerck@dfki.de

José L. Martínez
DAEDALUS- DATA,
DECISION and
LANGUAGE S.A.
jmartinez@daedalus.es

Resumen: El reciente crecimiento masivo de medios on-line y el incremento de los contenidos generados por los usuarios (por ejemplo, weblogs, Twitter, Facebook) plantea retos en el acceso e interpretación de datos multilingües de manera eficiente, rápida y asequible. El objetivo del proyecto TredMiner es desarrollar métodos innovadores, portables, de código abierto y que funcionen en tiempo real para generación de resúmenes y minería cross-lingüe de medios sociales a gran escala. Los resultados se están validando en tres casos de uso: soporte a la decisión en el dominio financiero (con analistas, empresarios, reguladores y economistas), monitorización y análisis político (con periodistas, economistas y políticos) y monitorización de medios sociales sobre salud con el fin de detectar información sobre efectos adversos a medicamentos.
Palabras clave: tecnologías del lenguaje en medios sociales, salud y finanzas, generación automática de resúmenes.

Abstract: The recent massive growth in online media and the rise of user-authored content (e.g weblogs, Twitter, Facebook) has led to challenges of how to access and interpret the strongly multilingual data, in a timely, efficient, and affordable manner. The goal of this project is to deliver innovative, portable open-source real-time methods for cross-lingual mining and summarization of large-scale stream media. Results are validated in three high-profile case studies: financial decision support (with analysts, traders, regulators, and economists), political analysis and monitoring (with politicians, economists, and political journalists) and monitoring patient postings in the health domain to detect adverse drug reactions.
Keywords: language technologies in health social media, financial analysis in social media, summarization, social media streams

1 Descripción General

TrendMiner (http://www.trendminer-project.eu/) es un proyecto europeo dedicado al análisis de medios sociales en distintos idiomas que comenzó en el año 2012 con los socios DFKI (Alemania) coordinador del proyecto, Universidad de Sheffield (UK), Ontotext AD, (Bulgaria), Universidad de Southampton (UK), Internet Memory Research (Francia), Eurokleis (Italy), Sora Ogris & Hofinger (Austria). En el 2014 se ha ampliado el consorcio con los siguientes socios: Grupo LaBDA de la Universidad Carlos III de Madrid (España), Nyelvtudomanyi Intezet, Magyar Tudomanyos Akademia (Hungría), Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polonia) y la empresa DAEDALUS-DATA, DECISIONS AND LANGUAGE, S.A. (España).

[1] FP7-ICT287863

2 Objetivos

El proyecto Trendminer se plantea los siguientes retos científicos:

Modelado de conocimiento y extracción de información multilingüe basada en ontologías. Se trata de desarrollar métodos de extracción de información novedosos que sean capaces de analizar streams de medios sociales caracterizados por ser cortos, ruidosos, coloquiales y contextualizados. El objetivo es identificar tendencias y sentimiento en múltiples lenguajes así como de extraer entidades y eventos relevantes y almacenarlos en una base de conocimiento. Los participantes DFKI y Universidad de Sheffield trabajan en esta línea generando ontologías ampliadas con elementos de opinión que proporcionan el conocimiento extralingüístico necesario en los métodos de extracción.

Es precisamente en esta línea donde el trabajo del Grupo LaBDA está desarrollando un recurso ontológico para representar información sobre fármacos así como sus indicaciones y efectos adversos que se incorporará al repositorio de ontologías del proyecto. Esta ontología se probará en tareas de extracción de información en blogs relacionados con salud donde los pacientes informan sobre sus tratamientos y problemas con su medicación (como por ejemplo www.forumclinic.org y www.enfemenimo.com)

Modelos basados en aprendizaje automático para minería de tendencias en medios sociales. Se trabaja en desarrollar enfoques de aprendizaje automático para la identificación de mensajes o posts importantes y para la extracción de fragmentos de texto a partir de grandes volúmenes de texto generados por los usuarios en medios sociales como Twitter. No existen datos de entrenamiento para el aprendizaje supervisado y crearlos consume muchos recursos. Por ello, se investiga en diversas formas para hacer uso de datos en forma de movimiento de precios del mercado y resultados de encuestas para su utilización en los casos de uso a modo de supervisión ligera para inferir la importancia de distintas características de los textos.

En esta línea se han llevado a cabo trabajos para la predicción de voto a partir de Twitter y la predicción de la tasa de desempleo también a partir de Twitter, ver los trabajos (Lampos, Preotiuc-Pietro & Cohn, 2013).

Generación de resúmenes cross-lingüe a partir de medios sociales. El objetivo es definir nuevos enfoques para la generación automática de resúmenes en una línea temporal para observar la evolución de los eventos. En esta línea la Universidad de Sheffield ha desarrollado trabajos como los presentados en (Rout, et al., 2013)

Plataforma para la recolección, análisis y almacenamiento de colecciones de medios sociales en tiempo real. Desarrollada por Ontotext. El objetivo es facilitar la recogida y la minería de conocimiento a partir de medios sociales (tales como Twitter, Facebook, blogs sobre salud y periódicos, etc.). Como principales retos están el procesamiento en tiempo real de grandes volúmenes de posts, la agregación y el almacenamiento así como el procesamiento distribuido y basado en la nube según arquitecturas elásticas para minería de textos. Esto hace posible que las empresas puedan utilizar los datos sin un coste prohibitivo y sin necesidad de invertir en infraestructuras privadas o centros de datos.

Minería de tendencias multilingüe y generación de resúmenes en el caso de uso de ayuda a la toma de decisiones en finanzas. El objetivo es poner en práctica en un desarrollo real a través de la plataforma del socio Ontotext las técnicas desarrolladas en el proyecto en el dominio financiero proporcionando métodos en tiempo real para inversores, analistas, consejeros, agentes de bolsa, en particular para analizar la influencia en los precios en el mercado.

Extensiones de lenguajes y dominios: Con la entrada de nuevos socios en el proyecto se buscaba ampliar los idiomas y los dominios de trabajo. En el caso del español se ha propuesto incrementar la cobertura de las ontologías y las herramientas en el dominio de la salud, en particular, para los fármacos, efectos terapéuticos y reacciones adversas. El objetivo es aplicar las técnicas utilizadas para la extracción de información en textos científicos biomédicos a la extracción en posts y contribuciones en medios sociales tales como *Forumclinic* y *Portales Médicos*. Sin embargo ninguno de estos blogs y foros es tan sofisticado como *Patientslikeme* (http://www.patientslikeme.com/), una plataforma on-line que integra datos de pacientes en inglés.

La figura 1 muestra la arquitectura general del proyecto.

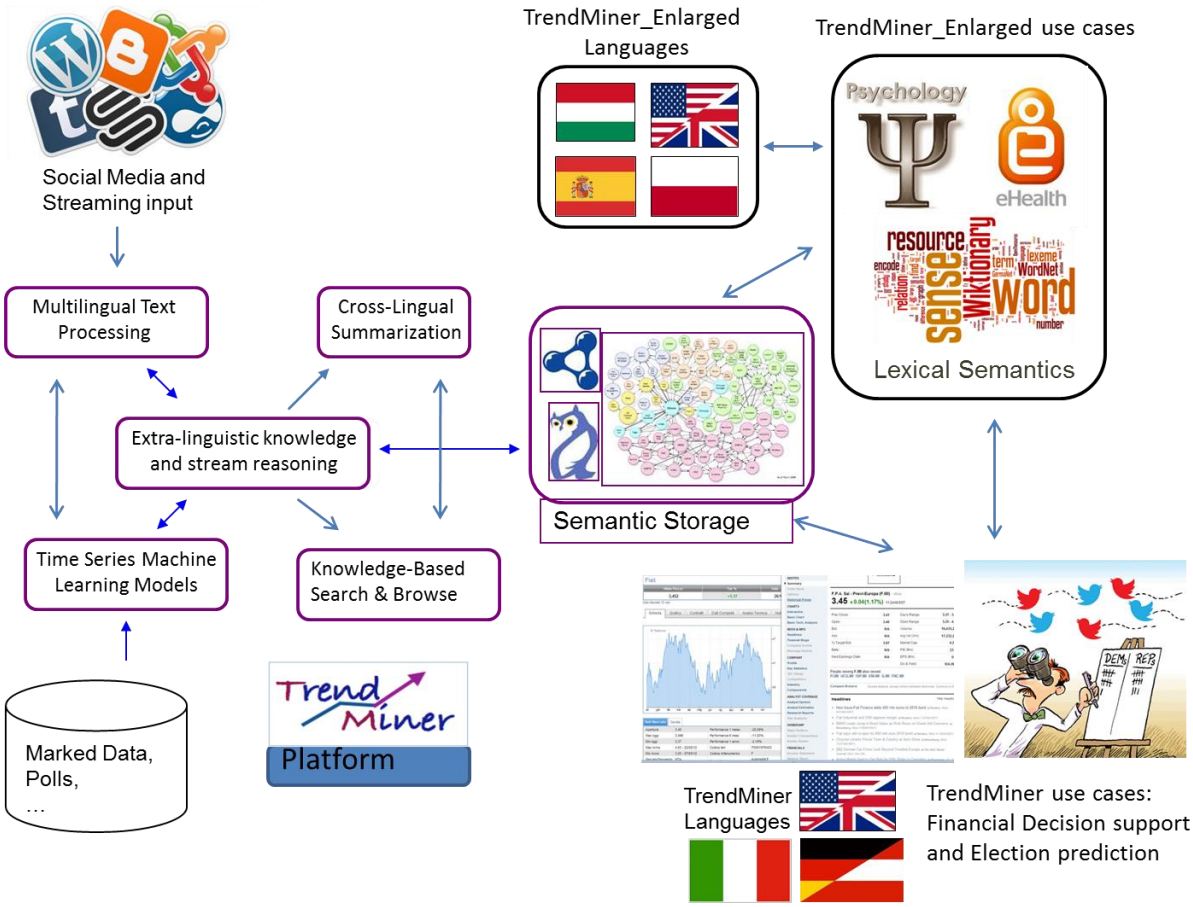

Figura 1: Arquitectura general del proyecto TrendMiner

3 *Situación actual*

En relación con el trabajo de los socios españoles, en la actualidad se trabaja en el dominio de salud en el análisis de consumo de antidepresivos en relación con distintos parámetros como la legislación actual en materia de trabajo, la tasa de desempleo, etc. Se están recogiendo tweets relacionados con diversos eventos usando diversas keywords relacionadas con fármacos antidepresivos, estados emocionales y términos relacionados con la economía con el fin de relacionarlos con el consumo de fármacos

Por otro lado, también se trabaja en la detección de efectos adversos de fármacos en medios sociales usando tecnología del lenguaje para analizar el contenido de los posts (reconocimiento de entidades y relaciones semánticas), (Segura-Bedmar, Revert y Martínez, 2014). Se trata de analizar si los pacientes reportan sobre estos efectos en blogs y foros. El objetivo es estudiar si estos medios pueden ser una fuente de conocimiento adicional a los sistemas de notificación que tienen las agencias europeas de medicamentos y productos sanitarios para que los pacientes y personal sanitario informen sobre sospechas de reacciones adversas que tienen poco uso. En (Segura-Bedmar, Peña-González y Martínez, 2014) se describe un recurso desarrollado para almacenar los fármacos y sus efectos (indicaciones y efectos adversos) relacionados que se integra en el prototipo bajo desarrollo. Hasta la fecha se disponía de información de fármacos y efectos pero de manera aislada.

En esta línea son varias las dificultades para extraer las menciones de fármacos y efectos adversos a partir de los comentarios de los usuarios. Además de abordar los problemas específicos de los medios sociales (como son el uso de abreviaturas, slangs, emoticonos, faltas ortográficas, etc.) son necesarios recursos que no existen en la actualidad. Por ejemplo, no se dispone en español de un diccionario orientado a pacientes como el que existe en inglés,

llamado Consumer Health Vocabulary[2]. Los pacientes no se refieren a sus problemas usando la terminología propia de los profesionales sanitarios

Por último también es interesante analizar la evolución de nuevos fármacos en el mercado analizando cómo evolucionan los comentarios de los pacientes considerando también las opiniones que manifiestan al respecto. Esto podría incluso servir de ayuda a la detección de nuevos efectos que no han sido descubiertos en los ensayos clínicos, una tarea importante en fármaco vigilancia. En el caso español, la Agencia Española de Medicamentos publica anualmente una lista con los fármacos de reciente aprobación a los que hay que hacer un especial seguimiento.

En la actualidad se dispone de un prototipo que analiza comentarios de pacientes en español disponible en http://163.117.129.57:8090/gate/ basado en un pipeline de procesos construido sobre GATE[3] que incorpora una API del procesador lingüístico Textalytics[4], software de DAEDALUS. Para el análisis de opinión se trabajará también con Sentimentalytics[5] del mismo socio.

En relación con el escenario financiero DAEDALUS está trabajando en el caso de uso de Responsabilidad Social Corporativa (RSC), para ayudar a las empresas a obtener una visión clara sobre su reputación on-line entre la gente. Para este propósito se ha trabajado en un prototipo que recoge tweets y sitios de noticias on-line accedidos a través de RSS para empresas del IBEX35. Esta información se analiza de acuerdo a un modelo de reputación predefinido similar a los existentes Merco©[6] y RepTrack©[7] El modelo incluye 7 dimensiones: estrategia y liderazgo, innovación y flexibilidad, integridad oferta responsabilidad social, situación financiera y trabajo. Además, cada una de estas categorías se divide en diferentes características. Dado un texto, un proceso de detección de entidades reconoce si se menciona una de las empresas del IBEX35 y si es así, un proceso de clasificación de textos es capaz de establecer cuáles de las dimensiones mencionadas se cubre. Finalmente, un proceso de análisis de sentimiento puede distinguir entre textos positivos y negativos con el fin de conocer si la gente tiene una idea buena o mala de una empresa en una determinada dimensión.

[2] http://www.consumerhealthvocab.org/
[3] http://gate.ac.uk/
[4] http://textalytics.com
[5] https://sentimentalytics.com/
[6] http://www.merco.info
[7] http://www.reputationinstitute.com

Bibliografía

Moreno, J., Declerck, T., Martínez Fernández, J.L. & Martínez, P. 2013. Prueba de Concepto de Expansión de Consultas basada en Ontologías de Dominio Financiero. Procesamiento del Lenguaje Natural, 51, 109-117.

Lampos, V., Preotiuc-Pietro, D., & Cohn, T. (2013). A user-centric model of voting intention from Social Media. In Proc 51st Annual Meeting of the Association for Computational Linguistics, 993-1003

Rout, D., Preotiuc-Pietro, D., Bontcheva, K., & Cohn, T. Where's @wally? A Classification Approach to Geolocating Users Based on their Social Ties. 24th ACM Conference on Hypertext and Social Media, HT, May 2013, Paris.

Segura-Bedmar, I., Revert, R. & Martínez, P. 2014. Detecting drugs and adverse events from Spanish social media streams. En Proceedings of the 5th International Louhi Workshop on Health Document Text Mining and Information Analysis (Louhi 2014).

Segura-Bedmar, I., Peña-González, S., & Martínez, P. (2014). Extracting drug indications and adverse drug reactions from Spanish health social media, Proceedings of the BioNLP 2014, June, 2014,