Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 127-130 recibido 20-03-2018 revisado 28-04-2018 aceptado 03-05-2018

Plataforma inteligente para la recuperación, análisis y representación de la información generada por usuarios en Internet

Intelligent framework for retrieving, analysing and representing user generated content on Internet

Yoan Gutiérrez, José M. Gómez, Fernando Llopis, Lea Canales, Antonio Guillén
GPLSI - Universidad de Alicante San Vicente del Raspeig, s/n, 03690, Alicante
{ygutierrez, jmgomez, llopis, lcanales, aguillen}@dlsi.ua.es

Resumen: Este proyecto viene motivado por la necesidad de definir una plataforma basada en Tecnologías del Lenguaje Humano que sea capaz de procesar la información de manera inteligente y de forma automática, combinando múltiples técnicas y herramientas. Dicha plataforma flexibilizará el modo de mostrar visualmente los datos resultantes para ser adaptados a las necesidades de los usuarios desde un punto de vista analítico. El avance científico de cada una de las tecnologías involucradas en la creación de la plataforma propuesta, así como su combinación e integración en una única infraestructura, supondrá un paso importante dentro de las tecnologías del lenguaje humano, siendo a su vez, de valiosa utilidad para la sociedad actual y futura.
Palabras clave: Tecnologías del Lenguaje Humano, Plataforma Inteligente, Tecnologías Integradas, Analitícas

Abstract: This project is motivated by the need of defining a platform based on Human Language Technologies capable of intelligently processing textual information, by combining multiple techniques and tools. In addition, the way of displaying the obtained results will be adapted to the users needs from an analytical point of view. The scientific progresses of each technology involved, as well as their combination and integration in a single infrastructure, will contribute to the progress of human language technologies, being in turn of valuable use for the current and future society.
Keywords: Human Language Technologies, Intelligent Platform, Processing Textual Content, Integrating Technologies, Analytics

1 Introducción

Actualmente, Internet cuenta con más de 4,156 millones de usuarios[1], dato que indica que el 54,4 % de la población mundial está conectada a la red de redes y por consiguiente consumiendo y generando información.

La web 2.0 o web social supone uno de los mayores atractivos para los usuarios de internet. Si se analizan los datos de dos de las redes sociales más conocidas hoy en día, Twitter[2] y Facebook[3], encontramos que Twitter cuenta con más de 310 millones de usuarios activos, que generan 500 millones de tweets al día[4], mientras que Facebook cuenta con más de 1.860 millones de usuarios[5] y más de 78 millones de páginas[6]. Si a ello le sumamos el resto de los sitios Web, incluyendo otros tipos de redes sociales, páginas Web, enciclopedias, blogs, foros, contenido multimedia, etc. encontramos más de 4.26 billones de páginas Web indexadas[7].

Sin embargo, el principal inconveniente de toda esta gran cantidad de información disponible es la complejidad para poder analizarla, sobre todo si la persona interesada desea obtener información precisa sobre da-

[1] http://www.internetworldstats.com/stats.htm (Marzo 2017)
[2] https://twitter.com (Marzo 2017)
[3] https://www.facebook.com (Marzo 2017)
[4] http://www.internetlivestats.com/twitter-statistics (Marzo 2017)
[5] https://www.trecebits.com/2017/02/02/facebook-ya-tiene-1-860-millones-de-usuarios/ (Febrero 2017)
[6] http://www.statisticbrain.com/facebook-statistics/ (Marzo 2018)
[7] http://www.worldwidewebsize.com/ (Febrero, 2018)

ISSN 1135-5948. DOI 10.26342/2018-61-14

tos formulados en lenguaje natural que necesiten ser interpretados.

Un modo de reducir el tiempo invertido por los usuarios en analizar grandes cantidades de información es mediante el uso de las Tecnologías del Lenguaje Humano (TLH).

Las herramientas y recursos de TLH desarrollados en los últimos años han permitido mejorar los procesos de búsqueda, recuperación y extracción de información (El-Helw, Farid, y Ilyas, 2012) (Irfan et al., 2015), clasificación de textos (Iglesias, Seara Vieira, y Borrajo, 2013) (Zhang, Zhao, y LeCun, 2015), detección y minería de opiniones (Fernández et al., 2013) (Ravi y Ravi, 2015) o síntesis de información (Cadilhac et al., 2015) (Moen et al., 2016) así como los procesos intermedios involucrados en cada una de estas tareas tales como el análisis semántico (Li y Joshi, 2012) (Gutiérrez, Vázquez, y Montoyo, 2017) que son clave para su interpretación.

Por tanto, se hace necesario aunar esfuerzos en las distintas tareas hacia la creación de una plataforma capaz de identificar el tipo de información que necesita el usuario, recuperarla, procesarla y presentársela de manera adecuada y flexible.

2 Estado del arte

Hoy en día existen algunas herramientas y sistemas informáticos que de una manera u otra son capaces de incorporar tecnologías de TLH para proporcionar infraestructuras analíticas. Por ejemplo Atribus[8] que es capaz de rastrear, buscar, recoger, filtrar y devolver todo lo que se está diciendo de un cliente en la red a partir de las palabras clave para cada uno en tiempo real; Natural Opinions[9] que analiza todo lo que se está diciendo en cada momento en Internet sobre una persona, una marca, una institución o un producto, y detectar automáticamente las entidades, conceptos y opiniones más relevantes; Textalytics[10] el cual se presenta como un motor de análisis de texto que extrae elementos con significado de cualquier contenido y lo estructura para que puedas procesarlo y gestionarlo fácilmente; Sentimentviz[11] propone un medios estimar y visualizar el sentimiento asociado a textos cortos e incompletos, Tweet Reach[12] permite obtener informes estadísticos a partir del análisis de Twitter, SocialBro[13] que propone una solución avanzada para la gestión y el análisis de comunidades de Twitter, permitiendo a los profesionales del Marketing y el Social Media analizar a fondo sus contactos, gestionarlos y definir sus estrategias en función de ello; SumAll[14] entre otras cosas obtiene estadísticas sobre seguidores de varias las redes sociales, como son: números de *me gusta*, cantidad de mensajes, localidades, etc. y los muestra por medios de gráficas de intervalos de tiempo (días, semanas, meses).

Nuestra propuesta estaría básicamente más alineada con las soluciones de Atribus y Natural Opinions. Además seríamos capaces de aportar con la plataforma de TLH valores añadidos como los que siguen a continuación.

3 Características distintivas de la propuesta de proyecto

Las características que se añaden en este proyecto a diferencia de las tecnologías ya existentes son: Nube de conceptos vs nubes de palabras/etiquetas; Dominios relevantes; Mapas de emociones (clasificación de tipos de emociones vs. Simple clasificación de polaridad); Extracción de mensajes de usuarios más relevantes en un intervalo de tiempo (resumen de tweets); Mapas de polaridad donde geográficamente se puedan representar las opiniones expresadas o inferidas de los usuarios de las redes sociales; Detección de conjuntos de términos que caracterizan e indican una localidad (e.g. postiguet, hogueras, arroz, Alicante) vs. Simple geolocalización que proporciona Twitter; Tratamiento multilinge de la información; y Flexibilidad para establecer métricas de análisis de reputación de entidades digitales.

4 Propuesta

Nuestra propuesta de plataforma de TLH se ilustra en la Figura 1. Dicha plataforma permite a los usuarios extraer información que se encuentre dispersa en la Web Social y representarla visualmente desde un punto de vista analítico, tras un intenso procesamiento de datos estructurados y no estructurados.

[8] https://www.atribus.com (Marzo, 2018)
[9] https://www.bitext.com (Marzo, 2018)
[10] https://www.meaningcloud.com/es/ (Marzo, 2018)
[11] https://www.csc2.ncsu.edu/faculty/healey/tweet_viz (Marzo, 2018)
[12] https://tweetreach.com/ (Marzo, 2018)
[13] http://es.socialbro.com (Dec, 2017)
[14] https://sumall.com (Marzo, 2018)

Las herramientas y procesos de TLH son el elemento central de este proyecto, ya que en él se tienen en cuenta tres roles fundamentales.

El primero es relacionado con la **extracción y recuperación** de Contenidos Generados por Usuarios (UGC). La plataforma debe ser capaz de ofrecer mecanismos para que los usuarios puedan definir sus propias búsquedas de información y de este modo el sistema pueda recuperar y extraer la información precisa.

El segundo rol es concerniente a la posibilidad de considerar **diferentes tipos de tecnologías de TLH** con el fin de poder aplicar **Minería de Textos** y obtener diversos rasgos de caracterización.

Y por último y no menos importante, debe ser capaz de ofrecer mecanismos para **mostrar de modo visual y sencillo analíticas** resultantes tras procesar la información obtenida.

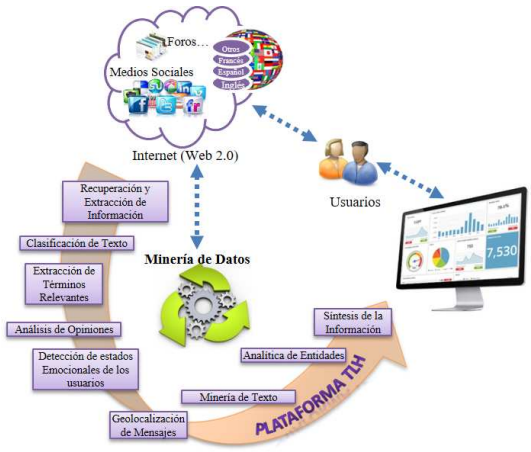

Figura 1: Plataforma de TLH

5 *Objetivo general*

El objetivo general de este proyecto es analizar, proponer y evaluar diferentes enfoques novedosos para el procesamiento de UGC desde un punto de vista analítico, creando una plataforma que combine, integre y visualice la información resultante de distintos procesos de TLH.

La información resultante puede materializarse de distintas formas dependiendo de las necesidades y preferencias del usuario. Por ejemplo, pueden ser sintetizadas en forma de resúmenes, de tweets, valoración de opiniones, términos más relevantes, pasajes, recopilación de fuentes relevantes, geolocalización de los mensajes, autores, etc. Siendo conscientes que dicha información procede de la Web 2.0.

El núcleo del proceso tanto de recuperación y extracción como de procesamiento de la información estaría formado por técnicas y herramientas que conforman las TLH. Para ello, se integrarán tecnologías tales como el análisis semántico, recuperación y extracción de información, minería de opiniones, clasificación de textos, computación afectiva (o análisis de emociones), síntesis de textos y otras que puedan ser de utilidad durante el transcurso del proyecto. Aunque la plataforma no está limitada a la integración de otras tecnologías, sí que es cierto que estas tareas serán las que conformen su núcleo central, y por tanto, serán cruciales para el correcto desarrollo del proyecto.

Por tanto, el proceso de clasificación, análisis y presentación del contenido social implicaría en primer lugar decidir qué información se debe recuperar y seleccionarla. Posteriormente habrá que ser capaces de procesar dicha información. Para ello, será necesari: identificar el tipo de información; clasificarla; detectar lo realmente importante y discriminar aquello que no es relevante; determinar información redundante, complementaria y/o contradictoria; e integrar y combinar todo el conocimiento obtenido. Finalmente, todo el conocimiento obtenido quedará almacenado en un repositorio de minería de textos capaz de indexar toda aquella información que el usuario considere relevante para ser mostrado desde una óptica analítica mediante interfaces visuales.

6 *Oportunidades de explotación*

En la actualidad muchas empresas se preocupan considerablemente por su reputación en la Web 2.0, ya que las redes se han convertido en las vías más populares, rápidas y efectivas de *marketing*. Es por ello, que nuevos perfiles laborales surgen de la mano de las nuevas tecnologías. Por ejemplo, podemos mencionar la figura del analista social[15], que entre otras funciones: evalúa y propone mejoras para la estrategia en los medios sociales y campañas comerciales; monitoriza y recolecta información sobre marca, productos, competencia y sector; clasifica las consultas de los clientes

[15] http://www.concepto05.com/2011/01/que-es-un-social-media-analyst-i-un-nuevo-puesto-de-trabajo (pub Enero, 2011)

para mejorar los sistemas de atención técnica; analiza la reputación online de marcas; realiza análisis sectoriales y comparativas con la competencia; etc.

7 Enlace al proyecto y resultados

En la web del proyecto[16] podéis encontrar la relación de artículos científicos que sustentan las tecnologías desarrolladas, así como registros de software y herramientas de demostración.

Agradecimientos

Este proyecto con referencia GRE16-01: Plataforma inteligente para recuperación, análisis y representación de la información generada por usuarios en Internet, está financiado por las Ayudas a Proyectos Emergentes de la Universidad de Alicante.

Bibliografía

Cadilhac, A., A. Chisholm, B. Hachey, y S. Kharazmi. 2015. Hugo: Entity-based News Search and Summarisation. En *Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval - ESAIR '15*, páginas 51–54.

El-Helw, A., M. H. Farid, y I. F. Ilyas. 2012. Just-in-time information extraction using extraction views. En *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, páginas 613–616, New York, NY, USA. ACM.

Fernández, J., Y. Gutiérrez, J. M. Gómez, P. Martínez-Barco, A. Montoyo, y R. Muñoz. 2013. Sentiment Analysis of Spanish Tweets Using a Ranking Algorithm and Skipgrams. *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013)*, páginas 133–142.

Gutiérrez, Y., S. Vázquez, y A. Montoyo. 2017. Spreading semantic information by Word Sense Disambiguation. *Knowledge-Based Systems*, 132:47–61.

Iglesias, E. L., A. Seara Vieira, y L. Borrajo. 2013. An HMM-based oversampling technique to improve text classification. *Expert Systems with Applications*, 40(18):7184–7192.

Irfan, R., C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Z. Wang, D. Chen, A. Rayes, N. Tziritas, C. Z. Xu, A. Y. Zomaya, A. S. Alzahrani, y H. X. Li. 2015. A survey on text mining in social networks. *Knowledge Engineering Review*, 30(2):157–170.

Li, Y. y K. D. Joshi. 2012. The state of social computing research: A literature review and synthesis using the latent semantic analysis approach. En *18th Americas Conference on Information Systems 2012, AMCIS 2012*, volumen 1, páginas 33–40.

Moen, H., L. M. Peltonen, J. Heimonen, A. Airola, T. Pahikkala, T. Salakoski, y S. Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37.

Ravi, K. y V. Ravi. 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.

Zhang, X., J. Zhao, y Y. LeCun. 2015. Character-level convolutional networks for text classification. En *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, páginas 649–657, Cambridge, MA, USA. MIT Press.

[16] https://gplsi.dlsi.ua.es/gplsi13/es/node/396