Metodología de inteligencia de negocio para análisis social en la infraestructura de datos enlazados SLOD-BI

Indira Lanza Cruz
Doutoranda em Informática na Universitat Jaume I de Castellón, Espanha, Mestre em Sistemas Inteligentes pela mesma instituição, orcid.org/0000-0003-2413-2799, lanza@uji.es.

María José Aramburu Cabo
Docente e pesquisadora do Departamento de Ingeniería y Ciencia de los Computadores na Universitat Jaume I de Castellón, Espanha, Doutorado em Computer Science pela University of Birmingham, Inglaterra, orcid.org/0000-0002-8026-8351, aramburu@uji.es

Rafael Berlanga Llavori
Docente e pesquisador do Departamento de Lenguajes y Sistemas Informáticos da Universitat Jaume I de Castellón, Espanha, Doutor em Physics pela Universidad de Valencia, orcid.org/0000-0002- 9155-269X, berlanga@uji.es

Submetido em: 22/08/2017. Aprovado em: 24/10/2017. Publicado em: 22/02/2018.

RESUMEN

Presenta una nueva metodología basada en infraestructuras de datos abiertos vinculados (Linked Open Data LOD) al realizar tareas de análisis en redes sociales. Esta metodología sigue las típicas fases de un proyecto de inteligencia de negocios (Business Intelligence BI), en el que a partir de un conjunto de fuentes de datos se obtienen métricas e indicadores relevantes para los objetivos estratégicos de la organización (Key Performance Indicators KPI). En el nuevo escenario, las fuentes de datos son redes sociales, y los objetivos estratégicos están relacionados con el desempeño de las organizaciones en esas redes sociales. El artículo demuestra los beneficios de representar toda la información relevante para el análisis bajo una misma infraestructura de datos abiertos vinculados, y como métricas e indicadores pueden ser obtenidos y publicados en la misma infraestructura. Estos beneficios destacan la capacidad de compartir entre los miembros de una comunidad todos los elementos relevantes en un análisis social a partir de los datos de origen para los indicadores sociales.

Palabras clave: Datos abiertos vinculados. Análisis de redes sociales. Indicadores sociales.

Business intelligence methodology applied to social analysis in the SLOD-BI linked data infrastructure

ABSTRACT

It presents a new methodology based on Linked Open Data LODs when performing analysis tasks in social networks. This methodology follows the typical phases of a Business Intelligence BI, in which metrics and indicators that are relevant to the organization's Key Performance Indicators (KPIs) are obtained from a set of data sources. In the new scenario, the data sources are social networks, and the strategic objectives are related to the performance of the organizations in these social networks. The article demonstrates the benefits of representing all information relevant to the analysis under a single linked open data infrastructure, and how metrics and indicators can be obtained and published on the same infrastructure. These benefits highlight the ability to share among members of a community all the relevant elements in a social analysis from source data to social indicators.

***Keywords**: Open linked data. Analysis of social networks. Social indicators.*

Metodologia de inteligência de negócios para análise social da infraestrutura de dados relacionados SLOD-BI

RESUMO

Apresenta uma nova metodologia baseada em infraestruturas de dados abertos vinculados (Linked Open Data LOD) ao executar tarefas de análise em redes sociais. Esta metodologia segue as típicas fases de um projeto de inteligência de negócios (Business Intelligence BI), em que a partir de um conjunto de fontes de dados são obtidos métricas e indicadores relevantes para os objetivos estratégicos da organização (Key Performance Indicators KPI). No novo cenário, as fontes de dados são redes sociais, e os objetivos estratégicos estão relacionados com o desempenho das organizações nessas redes sociais. O artigo demonstra os benefícios de representar toda informação relevante para a análise sob uma mesma infraestrutura de dados abertos vinculados, e como métricas e indicadores podem ser obtidos e publicados na mesma infraestrutura. Esses benefícios destacam a capacidade de compartilhar entre os membros de uma comunidade todos os elementos relevantes em uma análise social a partir dos dados de origem para os indicadores sociais.

***Palavras-chave**: Dados abertos vinculados. Análise de redes sociais. Indicadores sociais.*

INTRODUCCIÓN

La inteligencia de negocio (*Business Intelligence*, BI) tiene como objetivo principal extraer el conocimiento estratégico contenido en la información y los datos de una organización. Este conocimiento se suele representar en forma de indicadores estratégicos calculados a partir de las medidas de interés, que a su vez son tomadas de los datos recolectados desde diferentes fuentes e integrados en un esquema multidimensional. Frecuentemente, las medidas tienen carácter corporativo (ventas, costes, clientes, etc.) y son generadas dentro de la misma empresa. Sin embargo, hoy en día, buena parte de la información estratégica relevante para una organización reside en fuentes externas, incluyendo las redes sociales (Zhou, Lei, Wang, Fan, & Wang, 2015) (Fan & Gordon, 2014). En la actualidad, el análisis social es esencial para cualquier negocio y necesita ser monitorizado como cualquier otro sistema de comercialización.

SLOD-BI es una nueva infraestructura semántica diseñada para capturar y publicar datos de opinión con el fin de facilitar tareas BI y análisis de información social. La infraestructura proporciona la funcionalidad necesaria para realizar un análisis masivo de foros de opinión, incluyendo la extracción automática de datos de sentimientos para después enlazarlos semánticamente. Como resultado, los usuarios pueden incorporar dimensiones de opinión en su análisis, así como analizar los datos corporativos y sociales de manera integrada, lo que está fuera del alcance del BI tradicional.

En este trabajo, consideraremos como información social a toda la información colectiva producida por los clientes y consumidores de un mercado al participar en actividades sociales en línea. También nos referiremos a los datos extraídos de la información social por las herramientas de análisis, como son los datos de sentimiento o hechos de opinión. La cantidad de datos extraídos es masiva por lo que los foros sociales pueden ser considerados como fuentes de Big Data (i.e. fuentes de datos con gran volumen, heterogeneidad y escalabilidad) (Chen, Chiang, & Storey, 2012).

En el presente artículo definimos indicador clave de desempeño o KPI (*Key Performance Indicator*) como un valor medible de forma cualitativa o cuantitativa, generalmente expresado como un porcentaje, que permite evaluar el progreso hacia la consecución de los objetivos planteados en una empresa. La consecución de estos objetivos es revisada periódicamente por los responsables de los departamentos de una empresa. Su reto es encontrar el indicador más idóneo que esté ligado a lo que estén monitorizando (Parmenter, 2015). Los KPI también sirven para saber cómo dinamizar los

canales de la empresa en redes sociales ya que los resultados darán pistas sobre cómo seguir aplicando su estrategia en dichos medios (Guideline Key Performance Indicators, 2015) (Peterson, 2006).

Un indicador social es una medida temporal que permite a una organización medir dinámicamente el impacto de sus actividades en las redes sociales así como determinar el retorno de la inversión (ROI). Por otro lado, las acciones en las redes sociales son definidas en términos de objetivos estratégicos de la organización, que en este caso también tiene un carácter social. Los grandes desafíos de esta línea de investigación residen en la definición dinámica de buenos indicadores sociales, los cuales deben ser capturados y seguidos a partir de la gran cantidad de información que se publica constantemente en las redes sociales. Por tanto, esta línea combina tanto desafíos del BI (definición de objetivos e indicadores estratégicos a partir de almacenes de datos) como los inherentes al manejo de Big Data.

CONTEXTO Y REQUISITOS DEL TRABAJO

A pesar del gran interés comercial que existe en la creación de técnicas analíticas para las redes sociales, existen pocas aproximaciones en la literatura que aborden el tema dentro del área del BI. Algunos trabajos pioneros han sido recientemente revisados en (Berlanga & Nebot, 2015), y básicamente se plantean una correlación entre entidades externas (tales como noticias u opiniones) y entidades internas (los hechos a analizar). Otros trabajos se han centrado en crear modelos multidimensionales para el análisis de opiniones vertidas en las redes sociales acerca de un producto o compañía (Berlanga, y otros, 2015) (García-Moya, 2016). Muchos trabajos del área crean directamente procesos *ad-hoc* que miden algún tipo de indicador sobre un tema determinado en una red social, principalmente de carácter topológico (Wang, Jiao, Abrahams, Fan, & Zhang, 2013), de producto (Yan, Xing, Zhang, & Ma, 2015)(Abrahams, Jiao, Wang, & Fan, 2012) (Chae, 2015), o de sentimiento (polaridad) (Dai, Han, Dai, & Xu, 2015)(He, Wu, Yan, Akula, & Shen, 2015). Actualmente, el análisis de redes sociales está alcanzando un grado de madurez suficiente como para abordarlo desde un punto de vista más metodológico, tal y como se ha abordado el tema del BI tradicional en los almacenes de datos (*Data Warehouse*s, DW) (Diamantini, Potena, & Storti, 2016)(Horkoff, y otros, 2014) (Maté, Trujillo, & Mylopoulos, 2012). Sin embargo, existen peculiaridades en este dominio que no permiten adaptar de forma directa las técnicas BI tradicionales (Berlanga & Nebot, 2015) (Berlanga, y otros, 2015): los datos sociales que deben analizarse se consideran Big Data, y los indicadores sociales son dinámicos, volátiles y menos predecibles en su comportamiento.

Recientemente, se ha producido un elevado interés en la publicación de datos estratégicos como una nube de *datos abiertos y enlazados* (*Linked Open Data*, LOD) (Heath & Bizer, 2011). La iniciativa LOD tiene como principal objetivo crear una infraestructura global de datos semánticos a escala web, la denominada Web 3.0. Basándose en los protocolos web existentes, esta iniciativa propone publicar datos bajo los mismos principios que los documentos web, es decir, deben ser identificados a través de un Identificador Único de Recursos (URI) con el que cualquier usuario o máquina pueda acceder a sus contenidos, y los datos pueden vincularse entre sí a través de sus URI. Para gestionar la red de datos resultante, los datos deben estar provistos de una semántica bien definida que permita a usuarios y máquinas interpretarlos correctamente. Con este propósito, el consorcio W3C ha propuesto varios estándares para publicar y describir semánticamente los datos, principalmente el *Resource Description Framework* (RDF) y el *Ontology Web Language* (OWL). En este trabajo nos referimos como infraestructuras de datos semánticos a las redes de datos resultantes de la publicación y vinculación de datos con los formatos estándar RDF.

A pesar de que proyectos como *Schema.org* están permitiendo la publicación masiva de ofertas de productos como micro-datos, hoy en día aún no existe una infraestructura de datos abierta que permita a los usuarios y aplicaciones realizar directamente tareas de análisis en una gran

cantidad de opiniones publicadas en los medios sociales. La propuesta de SLOD-BI es que dicha infraestructura de datos se ajuste a los principios de la iniciativa LOD (Berlanga, y otros, 2015). Si los datos sociales basados en la Web se migran a la Web 3.0 como datos enlazados con el fin de ser compartidos, validados y finalmente integrados con datos corporativos, entonces se conseguirá habilitar un nuevo escenario de BI global para el análisis social. Además, la mayoría de los datos y vocabularios utilizados por los investigadores y las empresas para realizar el análisis de opiniones podrían ser mejor utilizados si son compartidos, contrastados y validados por la comunidad.

En cuanto a la naturaleza de los datos a publicar en la infraestructura de datos SLOD-BI, hemos identificado un conjunto de requisitos globales que aún no están cubiertos por las propuestas actuales, a saber:

- La infraestructura debe soportar la generación masiva de información social a partir de las contribuciones de los usuarios de los foros de opinión (por ejemplo, revisiones, *tweets*, etc.). Este gran volumen de datos debe ser procesado y los datos de sentimiento deben ser expresados como datos enlazados LOD.
- Los datos de sentimiento publicados en la infraestructura deben ser representados semánticamente bajo vocabularios bien controlados y relaciones taxonómicas útiles. Además, la infraestructura debe garantizar la calidad y homogeneidad de datos, tratando los posibles problemas de multilingüismo e interpretaciones dependientes del contexto. Todos los datos que no sean relevantes deben ser filtrados y descartados.
- El análisis de datos sociales puede implicar la generación masiva de datos de opinión de fuentes sociales (es decir, Big Data). En consecuencia, la infraestructura debe soportar el procesamiento masivo y la distribución de datos, proporcionando particiones óptimas con respecto al uso de datos. Igualmente, la infraestructura debe disponer siempre de los últimos datos, asegurando el rápido procesamiento de la información social más reciente.
- La infraestructura debe soportar operaciones de análisis complejas que integren datos con dos propósitos diferentes. En muchos casos, los datos sociales serán utilizados durante las operaciones de análisis BI para contextualizar los indicadores corporativos con datos del sentimiento externo, requiriéndose para ello enlazar e integrar previamente ambas categorías de datos. En otros casos, las aplicaciones analizarán la información social como tarea de seguimiento de ciertos eventos de la empresa como campañas de marketing o periodos de promoción. Aunque para estas aplicaciones el objeto de análisis serán los datos sociales, también necesitarán integrarse datos relevantes de las bases de datos corporativas.

La metodología de inteligencia de negocio presentada en este artículo se basa en la utilización de la infraestructura SLOD-BI para la definición de KPI adecuados para el seguimiento de objetivos estratégicos de una organización que opera en cierto domino y redes sociales. Desde el punto de vista práctico, esta metodología hará uso de los mecanismos y herramientas provistos por la infraestructura para recolectar, almacenar, monitorear, analizar y resumir indicadores sociales de utilidad, a través de los datos publicados por los usuarios de las redes sociales.

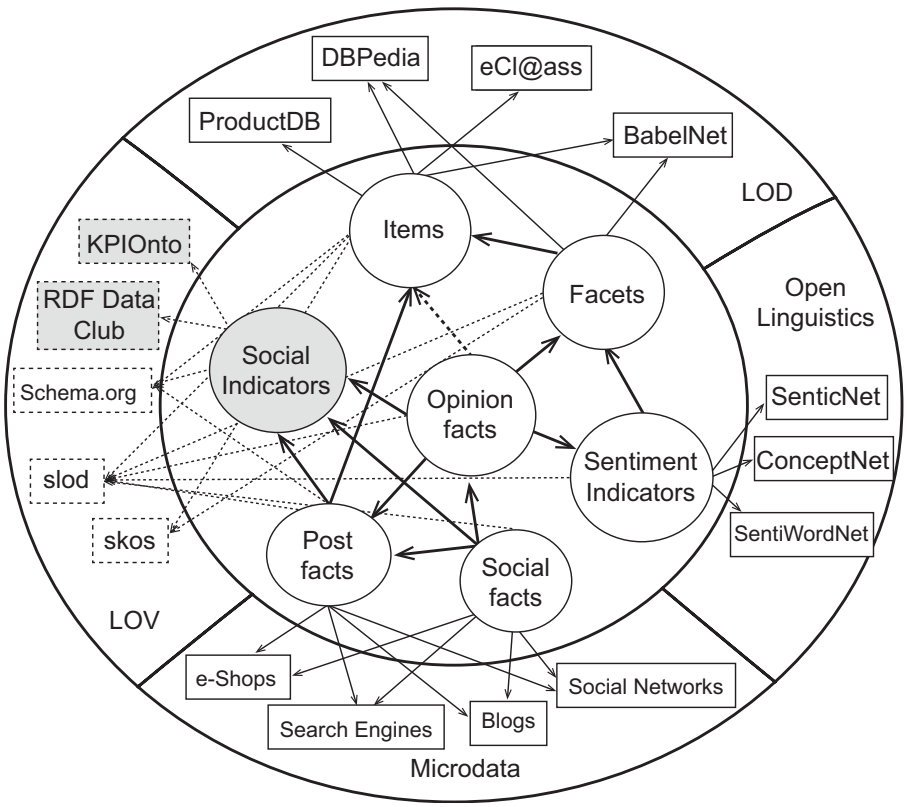
El resto del artículo se organiza de la siguiente manera. Primero se presenta brevemente la infraestructura SLOD-BI para el análisis BI de datos sociales y después se desarrollan las fases de nuestra metodología para el análisis social, brevemente: definición de los objetivos estratégicos, plan de marketing e indicadores clave; patrones de análisis de datos y herramientas LOD disponibles tras la parametrización de dicha infraestructura; y una fase final de validación y reajuste. El artículo termina con las secciones de conclusiones y referencias.

INFRAESTRUCTURA SLOD-BI

La figura 1 presenta los principales componentes utilizados por la infraestructura SLOD-BI para el análisis BI de datos sociales. Como puede verse, el conjunto de datos involucrados se divide en dos capas: en el anillo interno de la figura se agrupan los vocabularios y conjuntos de datos (*datasets*) de la infraestructura, mientras que el anillo externo comprende, por una parte, un grupo de vocabularios externos expresados como datos abiertos y enlazados (*Linked Open Vocabularies*, LOV) y, por otra parte, otro grupo de *datasets* externos directamente relacionados con la infraestructura como, por ejemplo, DBpedia y productDB. Cada componente SLOD-BI consiste en uno o varios *datasets* representando alguna de las perspectivas que se consideran relevantes para el análisis BI de los datos de opinión.

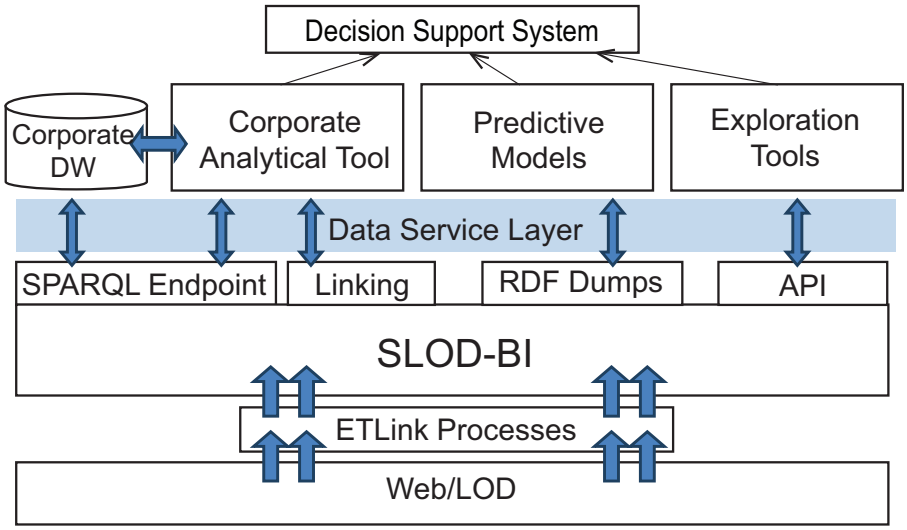
Los enlaces entre las componentes más internas de la infraestructura, representados en la figura 1 con las líneas más gruesas, deben ser semánticamente coherentes, ya que son frecuentemente utilizados para realizar tareas de análisis. Entre otras cosas, esto significa que la infraestructura debe favorecer la composición (*join*) de las tripletas de estos *datasets*. Por otro lado, los enlaces entre las componentes de la infraestructura y los conjuntos de datos externos, representados con las líneas más finas, indican las posibles conexiones entre las entidades de la infraestructura y los *datasets* externos. Estos conjuntos de datos externos son utilizados, principalmente, para realizar operaciones de análisis exploratorio, es decir, para buscar en ellos nuevas dimensiones de análisis (Abelló, y otros, 2014). También desempeñan un papel muy relevante en la adquisición automática de los datos de opinión. Por ejemplo, los datos sociales referentes a algún producto que ya existe en DBpedia podrán ser automáticamente vinculados con el URI de dicho producto en el correspondiente *dataset* de SLOD-BI.

Figura 1 − Infraestructura de datos SLOD-BI



Fuente: Elaborada por los autores, 2017.

Figura 2 – Arquitectura funcional SLOD-BI



Fuente: Berlanga et al.(2015).

En la figura 2 se resumen las unidades funcionales de la infraestructura SLOD-BI necesarias para dar soporte a las anteriores componentes de datos. En la capa inferior, se localizan unos procesos (ETLink Processes) que continuamente seleccionan y supervisan las fuentes de datos sociales para extraer, transformar y vincular su contenido con los *datasets* de la infraestructura. La denominada *Data Services Layer* aloja todos los servicios encargados de producir, a partir de los datos de sentimientos, los datos requeridos por las herramientas analíticas. Estos servicios se implementan sobre una serie de servicios básicos proporcionados por la infraestructura, a saber: (i) un acceso SPARQL para realizar directamente consultas sobre datos de sentimientos, (ii) el servicio *Linking* para enlazar los datos corporativos con los datos de infraestructura (por ejemplo, nombres de productos, ubicaciones, etc.), (iii) un servicio de volcado RDF para los servicios de procesamiento *batch*, (iv) una API para realizar operaciones específicas sobre la infraestructura (por ejemplo, registrar, implementar restricciones de acceso, etc.) y finalmente, (v) herramientas visuales para la exploración de datos.

Obsérvese que en la estructura funcional propuesta, los datos de sentimiento se integran con los datos corporativos en la herramienta *Corporate Analitycal Tool*, haciendo uso para ello de algún servicio intermedio que proporcione los datos adecuados para cada tarea de análisis específica. Adicionalmente, los modelos predictivos y las herramientas de exploración permitirán la ejecución de procesos complejos sobre los datos de sentimientos. En ambos casos, la capa de servicio de datos facilitará la recuperación pertinente de los datos corporativos según sea necesario. Una ventaja importante de usar la capa de servicios de datos para consultar el DW corporativo es que ayuda a mantener el nivel de administración y seguridad de datos necesario para un análisis preciso y confiable (Carey, Onose, & Petropoulos, 2012).

METODOLOGÍA DE ANÁLISIS SOCIAL

A grandes rasgos, la práctica BI abarca dos grandes etapas iterativas: la provisión y la explotación de datos. La etapa de provisión de datos es la que consume más tiempo y la que requiere mayor parte de los recursos financieros y de mano de obra dentro del ciclo de vida de BI (Olszak & Ziemba, 2007). Consiste en el procesamiento de grandes cantidades de información para extraer los datos y métricas con los que calcular los indicadores a analizar. La etapa de explotación de datos está

asociada fundamentalmente a la aplicación del usuario final. Se centra en prácticas relacionadas con el seguimiento y análisis de datos orientado al alcance de los objetivos estratégicos y metas. En el ámbito de la analítica social, es importante añadir las siguientes fases fundamentales: la planificación y programación de las actividades en los medios sociales; el acceso, seguimiento y análisis de hechos e indicadores sociales; la evaluación y desarrollo de decisiones alternativas; y por último, la mejora del desempeño de la organización en los medios sociales.

Tomando como base este modelo y la infraestructura de datos sociales SLOD-BI, proponemos una metodología para el análisis social BI que consta de los siguientes fases: 1) definición de objetivos estratégicos de la organización y su plan de marketing en los medios sociales; 2) definición de indicadores clave KPI para el seguimiento de los objetivos estratégicos en los medios sociales; 3) identificación de patrones de datos para análisis social; 4) selección de herramientas LOD para análisis social en SLOD-BI; 5) parametrización de la infraestructura SLOD-BI; 6) validación y reajuste de la infraestructura. En las siguientes secciones se detalla cada una de las fases de esta metodología.

OBJETIVOS ESTRATÉGICOS Y EL PLAN DE MARKETING EN LOS MEDIOS SOCIALES

Las acciones que una organización ejecuta en las redes sociales tienen que ser diseñadas en términos de sus objetivos estratégicos. Es decir, a partir de los objetivos a los que se quiere llegar (por ejemplo, aumentar el número de ventas de un producto, el número de clientes de un servicio o aumentar la cuota de mercado sobre un público específico), para cada servicio o producto afectado hay que definir un plan de marketing en los medios sociales (González, Menéndez, & C. Seoane, 2013). Entre otras cosas, disponer de un plan de marketing hace posible determinar las métricas e indicadores sociales que se deben monitorizar.

Más detalladamente, se puede decir que una vez se han identificado los servicios, productos y/o perfiles del cliente potencial a los que se quiere llegar, se deben establecer los objetivos estratégicos y el plan de marketing a ejecutar en los medios sociales. Estos objetivos han de ser específicos, alcanzables y medibles (Muñoz Vera & Elósegui, 2011). Para cada objetivo, el plan de marketing debe planificar una estrategia de consecución, así como la manera en que se va a monitorizar sus actuaciones en los medios sociales, o sea, qué indicadores sociales se van a calcular y con qué periodicidad se van a medir.

INDICADORES CLAVE PARA EL SEGUIMIENTO DE LOS OBJETIVOS ESTRATÉGICOS

El cálculo de indicadores sociales requiere procesar Big Data procedente de los medios sociales para obtener unas métricas predefinidas. Es fundamental analizar con detenimiento cada una de las métricas que obtenemos de las redes sociales y evaluar si se correlacionan con el cumplimiento de alguno de nuestros objetivos estratégicos, cómo puede ser "aumentar el número de ventas". El objetivo final es utilizar los datos adquiridos para mejorar los esfuerzos de marketing, sin embargo, el desafío que afrontan los vendedores es que la naturaleza de la conversación social además de tener un volumen enorme se encuentra dispersa, lo que dificulta el agregado y consolidación de la información en una visión con significado para la empresa.

IDENTIFICACIÓN DE PATRONES DE DATOS PARA SOCIAL BI

En esta etapa se deben identificar y especificar las fuentes y patrones de datos que son necesarias para poner en marcha el sistema de BI para análisis social. Estas estructuras permiten la representación y almacenamiento de los datos sociales para su integración con los sistemas internos de una compañía. En la figura 3, se representa la relación entre los principales patrones de BI que intervienen durante el análisis social. Los patrones de análisis en el lado de datos corporativos de la figura

corresponden al modelo multi-dimensional (MD) tradicional de un almacén de datos típico (Codd, Codd, & Salley, 1993) Los patrones en el lado de los datos sociales constituyen la principal contribución de nuestra propuesta metodológica.

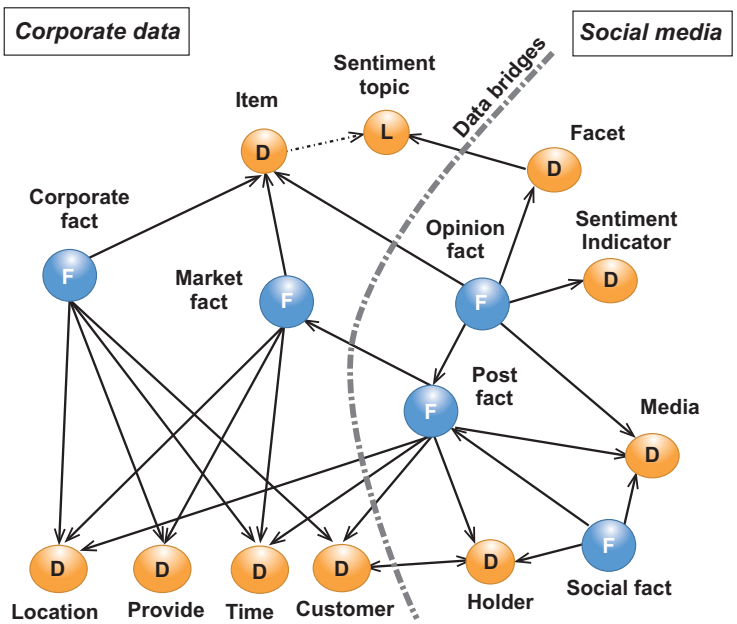
En la figura 3 los hechos están etiquetados con la letra "F", representan observaciones espacio-temporales de alguna medida (por ejemplo unidades vendidas, número de comentarios positivos o índice de impacto de un producto), mientras que las dimensiones representadas por la letra "D", representa el contexto de dichas observaciones.

Los hechos que están relacionados con los datos sociales son hechos de opinión, hechos de mensajes, hechos sociales y hechos de indicadores. Los hechos de opinión (*Opinion Facts*) son observaciones sobre sentimientos expresados por usuarios concernientes a facetas concretas de un *item* (i.e. producto o servicio objeto del análisis), junto con sus indicadores de sentimiento. Los hechos de mensaje (*Post Facts*) son observaciones de la información publicada sobre un *item* objetivo, el cual puede incluir una serie de hechos de opinión. Los hechos sociales (*Social Facts*) proveen información relevante sobre los usuarios y sus opiniones en el contexto de la comunidad a la que pertenecen (Berlanga, y otros, 2015).

En este trabajo se incluyen los hechos de indicador (*Indicator Fact*) que constituye el patrón que modela las observaciones asociadas a los indicadores sociales definidos. Una observación estará conformada por el valor calculado en cierto medio para un indicador determinado sobre las dimensiones producto, área y tiempo. Un ejemplo concreto de un hecho de indicador podría ser el valor calculado para el indicador "índice de impacto" para un producto determinado (marca de automóvil), en área geográfica especificada (España), en la red social Twitter, en un período dado (enero 2017). Para el cálculo de una observación se utilizará la fórmula del indicador asociado, cuyas variables corresponden a las medidas obtenidas en los patrones de los hechos de mensaje, opinión y sociales antes mencionados.

El hecho de indicador es el patrón fundamental durante el análisis social, ya que resume y registra los resultados calculados para cada indicador social definido, facilitando así su seguimiento, análisis y la toma de decisiones. A su vez sirven de entrada para otros procesos relacionados con la toma de decisiones y emisión de alertas, entre otras funciones en la organización.

Figura 3 – Patrones de análisis SLOD-BI



Fuente: Berlanga et al.(2015)

SELECCIÓN DE MÉTODOS LOD PARA EL BI SOCIAL

La plataforma SLOD-BI ofrece la infraestructura necesaria para implementar el análisis social BI y constituye la herramienta fundamental de nuestra propuesta metodológica. Sobre SLOD-BI se propone incorporar nuevas estructuras semánticas para modelar el procesamiento de indicadores sociales: Indicador, Fórmula, Hechos de Indicador (*Indicator Facts*) y Dimensión. A su vez, se define un catálogo básico de indicadores sociales y las fórmulas para su cálculo. Por último, se definen nuevos mecanismos de extracción, transformación y carga para la gestión de las observaciones asociadas a los indicadores definidos.

Para representar los indicadores sociales se propone hacer uso de la ontología KPIOnto del proyecto SemPI (Diamantini, Potena, & Storti, 2016). Su característica distintiva es que permite la representación lógica de las fórmulas asociadas a los KPI, permitiendo hacer explícitas las relaciones algebraicas entre los indicadores. Un indicador puede ser un dato atómico (por ejemplo, una info-métrica en SLOD-BI) o compuesto (la combinación de otros indicadores). Las dependencias de un indicador compuesto se definen por medio de una expresión algebraica $f(I_1,...,I_n)$, siendo a su vez $I_1,..., I_n$ fórmulas de otros indicadores.

Por otra parte, para el registro y seguimiento de las observaciones asociadas a los indicadores sociales (*Indicator Facts*) se hará uso del vocabulario RDF Data Cube propuesto por la W3C (https://www.w3.org/TR/vocab-data-cube/), ya que su modelo es compatible con el modelo de cubo que subyace al SDMX (Statistical Data and Metadata eXchange), norma ISO para intercambiar y compartir datos estadísticos y metadatos entre organizaciones. Este vocabulario permite especificar las unidades de medida, los factores de escala y los metadatos, así como el estado de la observación.

La extensión propuesta para la infraestructura SLOD-BI original se representa en la figura 1 de forma sombreada. Esta extensión incluye un nuevo conjunto de datos (*Social Indicators*) el cual contiene las nuevas estructuras de datos propuestas en este trabajo: indicadores sociales (*Social Indicator*), dimensiones de indicador (*Dimension*) y sus observaciones asociadas (*Indicator Fact*). Se incorporan también nuevos vínculos semánticos con los vocabularios Schema.org, KPIOnto y RDF Data Cube.

PARAMETRIZACIÓN DE LA INFRAESTRUCTURA SLOD-BI PARA EL ANÁLISIS SOCIAL

En esta sección describiremos los aspectos más relevantes de los nuevos componentes para definir indicadores de análisis social, a saber: *Indicator*, *Dimension*, *Formula* e *Indicator Fact*. Los *datasets* se han definido siguiendo los criterios de la W3C. Las características de los *datasets* asociados a los componentes *Item*, *Facets*, *Sentiment*, *Post* and *Opinion Facts* pueden consultarse en el artículo donde se propuso SLOD-BI (Berlanga, y otros, 2015).

CLASE INDICADOR SOCIAL

Este componente modela la estructura de un indicador social y sus instancias describen las métricas que permiten seguir el desempeño en las redes sociales. La clase que lo representa es slod:SocialIndicator, que heredará todas las propiedades de la clase kpi:Indicator (Diamantini, Potena, & Storti, 2016). Las propiedades básicas que utilizaremos son: nombre, identificador, acrónimo, definición, dimensiones compatibles, fórmula, unidad de medida (referida como Measurement Units Ontology3 en idi.fundacionctic.org/muo/muo-vocab.html) y la función de agregado. Por otro lado, la propiedad *maps-to* del vocabulario RTM (http://www.ontopia.net/doc/5.2.1/misc/rdf2tm.html) se aplica para vincular cada indicador atómico con el valor de una medida social capturada en SLOD-BI (por ejemplo, el número de seguidores o el número de veces que se compartió una opinión). La tabla 1 muestra las principales propiedades de la clase slod:IndicadorSocial.

También proponemos un catálogo básico de indicadores sociales para la puesta en marcha del sistema de análisis social, representado en la figura 4. Es importante destacar que debe existir

compatibilidad entre los indicadores participantes en las fórmulas, es decir, deben tener exactamente las mismas dimensiones que pueden especificarse explícitamente o heredar las de sus vecinos en las fórmulas y superclases.

CLASE INDICATORFACT

La clase slod:indicatorFact representa las observaciones asociadas a un indicador determinado para diferentes dimensiones (ej.: producto, espacio, tiempo). Para describir una observación se hace uso de la estructura *Observation* del esquema RDF Data Cube (espacio de nombre qb). Este vocabulario cubre gran parte de los aspectos necesarios para conformar los metadatos del componente *Indicator Facts*, al que añadimos nuevas extensiones que referencian al producto analizado y al indicador calculado. La tabla 2 muestra las propiedades principales asociadas a la clase slod:indicatorFact que hereda de qb:Observation.

La figura 5 muestra, mediante un ejemplo para el cálculo del indicador social de impacto, PKI_Impact, las relaciones semánticas de todos los componentes que intervienen en la infraestructura para el análisis social: indicadores sociales (SocialIndicator), dimensiones (Dimension) y las observaciones o hechos (IndicatorFact). En la figura, las relaciones de tipo *es-un* están representadas con flechas azules, y el resto de relaciones con flechas etiquetadas.

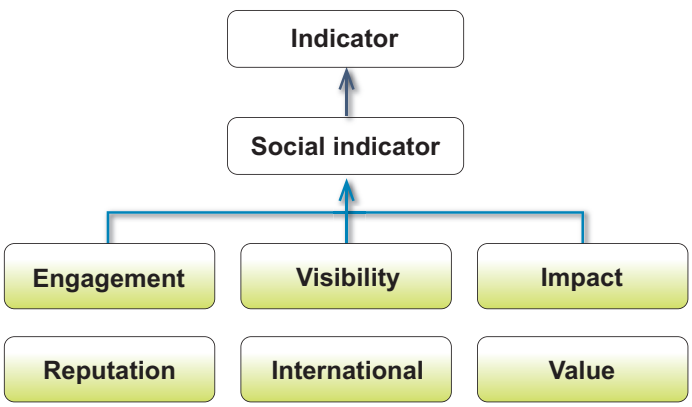
La relación (slod:PKI_indegree, maps-to, slod:indegree) representa la vinculación del valor de un indicador social con la métrica capturada en SLOD-BI.

Tabla 1 − Propiedades para describir un indicador social

Propiedad	Rango	Descripción
kpi:hasUnitOfMeasure	xsd:string	Indica la unidad de medida del indicador.
kpi:hasFormula	rdf:resource	Indica la fórmula para el indicador definida en SLOD-BI.
kpi:hasAggrFunction	rdf:resource	Indica la función de agregado del indicador.
kpi:hasDimension	rdf:resource	Indica una dimensión que es compatible con un indicador.
rtm:maps-to	Data and Object Properties	Hace una equivalencia entre propiedades. Hace un mapeo con una propiedad compatible del componente Social Fact de SLOD-BI.
slod:hasPeriodicity	Time	Periodicidad de evaluación del indicador.

Fuente: Elaborada por los autores, 2017

Figura 4 − Catálogo básico de indicadores sociales para SLOD-BI



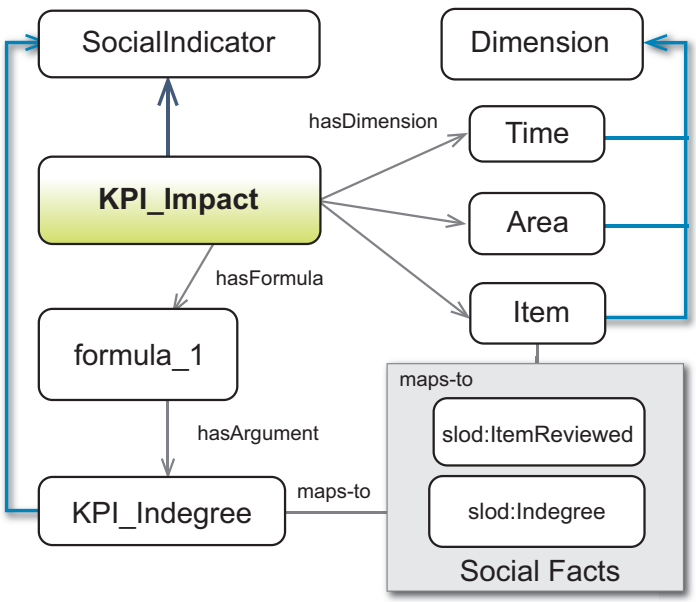
Fuente: Elaborada por los autores, 2017

Tabla 2 – Propiedades para describir un IndicatorFact

Propiedad	Rango	Descripción
sdmx-dimension:timePeriod	xsd:dateTime	Dimensión tiempo utilizada para el indicador asociado
sdmx-dimension:refArea	slod:Community	Dimensión área
s:itemReviewed	slod:Item	Dimensión que especifica el elemento sobre el cual se hace la consulta
slod:hasIndicatorAssociated	slod:SocialIndicator	Indica el indicador social vinculado a esta observación
sdmx-measure:obsValue	xsd:Double	Valor que toma el indicador que se está midiendo en la observación
s:dateCreated	s:Date	Fecha de la observación

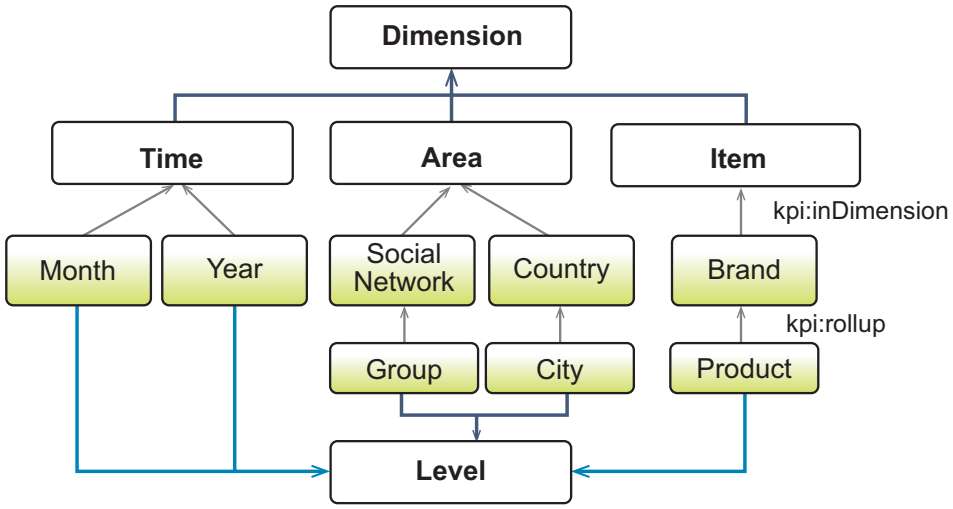
Fuente: Elaborada por los autores, 2017

Figura 5 – Un ejemplo gráfico para la representación de indicadores en SLOD-BI.



Fuente: Elaborada por los autores, 2017

Figura 6 – Fragmento del catálogo de dimensiones propuesto para SLOD-BI



Fuente: Elaborada por los autores, 2017.

CLASE DIMENSION

En el contexto del proyecto, para describir las dimensiones se utiliza un enfoque similar al propuesto en (Diamantini, Potena, & Storti, 2016). Partiendo de la superclase kpi:Dimension, proponemos un catálogo de clases que describen las dimensiones básicas: Time, Area e Item. Los niveles estarán representados por las subclases primitivas disjuntas de la dimensión a la que pertenecen. Por ejemplo, la dimensión

Time puede agrupar niveles asociados a períodos como mes, trimestre o año; la dimensión Area representa el ámbito de los datos sociales recuperados por ejemplo a nivel de país, comunidad o red social; mientras para Item los niveles pueden ser marca, organización o un producto analizado. Algunos ejemplos de instancias de cada nivel podrían ser para marca "Marca X de automóvil", para mes "2017-01" y para red social "Twitter". En la figura 6 se muestra un fragmento del catálogo propuesto.

ETLINK PARA LA POBLACIÓN DE INDICADORES

La infraestructura SLOD-BI implementa una serie de procesos de extracción, transformación y carga de datos (ETL) para poblar los conjuntos de datospropuestos. Los procesos ETL en SLOD-BI difieren notablemente de los implementados en los almacenes de datos tradicionales, motivo por el que se denominan procesos ETLink. En primer lugar, los nodos de extracción de datos deben procesar formatos semi-estructurados (e.g., JSON, XML, etc.) y conectarse a servicios web para extraer los datos sociales relevantes (e.g., API REST de Twitter, API REST de Facebook, etc.). Por otra parte, la transformación de datos depende de operaciones de procesamiento de textos para generar datos de sentimiento, y requieren consultas frecuentes a la infraestructura de datos para vincular los datos producidos. Finalmente, la fase de carga de datos debe ser expresada en formato RDF. El flujo de datos es implementado a través de *scripts* en Python, que hacen uso del paquete de clases *pygrametl* diseñado para la gestión de almacenes de datos (dimensiones y hechos). En el contexto de esta solución, en lugar de usar SQL, se utilizan primitivas de RDF (RDFLib library) para generar datos intermedios, y SPARQL para consultarlos. Los operadores del flujo de trabajo son capaces de consumir y producir datos en forma tabular (CSV) o de tripletas RDF (Berlanga, y otros, 2015).

En el contexto de una solución para el análisis de indicadores sociales, se propone implementar un nuevo proceso ETLink, a modo de servicio web, que permita la consulta y publicación de datos de las nuevas estructuras definidas en las secciones anteriores. Brevemente, se trata de un proceso periódico que consiste en tres fases fundamentales: (i) definir la consulta SPARQL para la captura de datos asociados al indicador definido, (ii) crear nodos para la interpretación y cálculo de las fórmulas definidas en los indicadores, y por último (iii) publicar las observaciones generadas en la infraestructura de datos, de acuerdo a la periodicidad de cada indicador definido.

VALIDACIÓN/EVALUACIÓN A PARTIR DE UN CASO DE USO

Con el propósito de validar la metodología propuesta, se desarrolló un prototipo sobre la infraestructura SLOD-BI para la definición y seguimiento de indicadores sociales en el dominio de alquiler de vehículos. Actualmente la infraestructura está poblada con vocabularios y datos RDF generados a partir de opiniones publicadas en Twitter y en varios medios sociales especializados en el tema (datos disponibles en http://krono.act.uji.es/SLOD-BI/sparql).

A continuación, se detalla cómo se implementa el proceso para la obtención de un indicador de análisis social mediante la resolución de un caso de uso concreto. Una empresa de alquiler de coches se ha planteado como objetivo estratégico aumentar el número de clientes y a su vez repercutir en un incremento del número de alquileres. Para ello la empresa se ha trazado un plan de marketing digital cuyo objetivo es aumentar el impacto de la marca en las redes sociales y los medios digitales. Como objetivos específicos se plantea ampliar el alcance de la marca hacia ciertas comunidades de usuarios (por zona geográfica), y aumentar el número de seguidores e interacciones en las redes sociales. Así, la empresa decide definir un indicador social de Impacto que mida para cada mes la repercusión de sus acciones en estos medios.

La forma de trabajar en la infraestructura propuesta consistiría en tres fases. En la primera fase, la empresa define el indicador social Impacto en función de los patrones de análisis proporcionados por SLOD-BI. Concretamente, debe definir la fórmula del indicador, las dimensiones y la periodicidad con la que se realizarán sus observaciones. Una vez definido y validado el indicador, se ejecutarían de forma automática los procesos ETLink encargados de generar las observaciones según van recogiéndose datos en la infraestructura. Finalmente, mediante consultas SPARQL, la empresa puede visualizar los valores del indicador Impacto en su cuadro de mando. De este modo, la empresa podrá contrastar estos datos con los indicadores internos de la empresa (e.g., ventas), y sacar conclusiones sobre la evolución del objetivo estratégico planteado. A continuación, se desarrollan cada una de las etapas para el caso de uso propuesto.

DEFINICIÓN DE LOS INDICADORES SOCIALES

Los indicadores sociales se definen como clases expresadas en RDF. Como se dijo anteriormente, estas clases utilizan vocabularios existentes como Schema.org (prefijo sc), KPIOnto (prefijo kpi) y RDF Data Cube (prefijo qb) además de los propios de SLOD-BI (slod). Estos últimos han sido extendidos con las clases y propiedades propuestas en este artículo (socialIndicator e indicatorFact).

En la figura 7 se muestra un fragmento del indicador de Impacto propuesto por la empresa, el cual se basa a su vez en otros indicadores atómicos que se enlazan de forma directa a métricas de SLOD-BI (*reposting*).

Figura 7 − Fragmento de la definición del KPI de Impacto en formato Turtle RDF

```
…
slod:KPI_Impact
        a slod:SocialIndicator;
        kpi:acronym “ENG” ;
kpi:hasAggregationFunction slod:Sum ;
        kpi:hasDimension slod:Time, slod:Item, slod:Area
;
 kpi:unitOfMeasure “#”;
 slod:hasPeriodicity slod:Monthly;
        kpi:hasFormula [
                kpi:hasFunction om:divide;
                kpi:has Argument [ kpi:hasFunction
om:plus;
 kpi:hasArgument [ kpi:hasArgumentPosition “1”^^xsd:int ;
                        kpi:hasArgumentName
“addend” ;
                        kpi:hasArgumentValue
slod:KPI_indegree ],
                                [
kpi:hasArgumentPosition “2”^^xsd:int ;

kpi:hasArgumentName “addend” ;

kpi:hasArgumentValue slod:KPI_Reposting ]
                                ],
                                [
kpi:hasArgumentPosition “3”^^xsd:int ;
                kpi:hasArgumentName “addend” ;
                        kpi:hasArgumentValue
slod:KPI_TotalPosts ]
        ].
slod:KPI_Reposting
a slod:SocialIndicator;
        rtm:maps-to slod:reposting.
…
```

Fuente: Elaborada por los autores, 2017

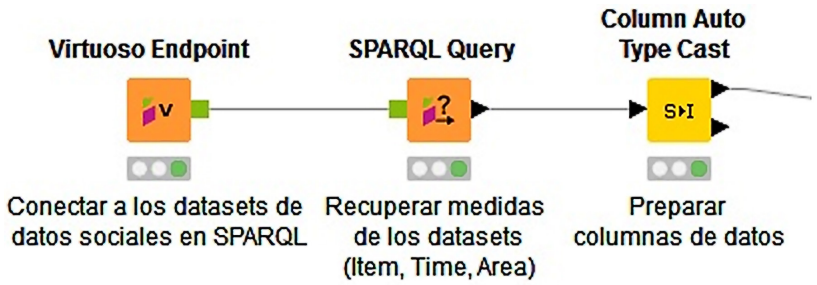
PROCESOS ETLINK PARA OBTENER LAS OBSERVACIONES DE LOS INDICATORFACTS

Una vez definido y publicado el indicador, se generaría la consulta SPARQL al punto de acceso o *endpoint* de SOLD-BI que permite obtener los datos primarios (métricas y agregado de datos). Por ejemplo, para el cálculo de KPI_Impact, se necesitan las métricas de *reposting* (veces que un post es compartido), *indegree* (promedio de seguidores observado en un post), y número total de posts. Estas métricas se contextualizan en las dimensiones de tiempo, ítem y comunidad. La consulta SPARQL correspondiente a este indicador trabajaría con el patrón *Social Fact* (ver figura 3). Cabe mencionar que dicha consulta está parametrizada por la dimensión de tiempo, ya que las observaciones se agrupan según la periodicidad definida en el indicador (en este caso mensual). La construcción de la consulta SPARQL es guiada por los predicados “maps-to” que vinculan de forma directa los elementos del indicador con las métricas concretas de la infraestructura.

A modo ilustrativo mostramos la implementación del proceso ETLink asociado a este indicador con la herramienta KNIME (http://www.knime.org), especialmente adecuada para la definición de procesos de carga (ETLs) y el análisis y visualización de datos.

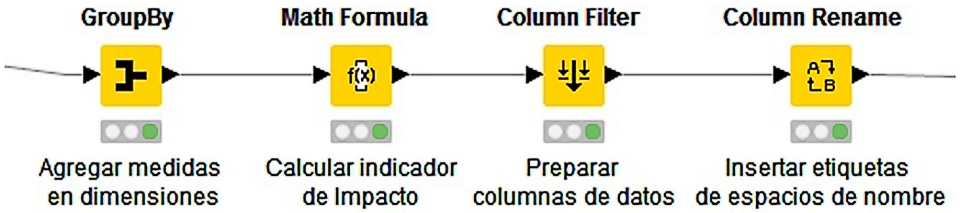
Las figuras 8-10 muestran el flujo de trabajo implementado en KNIME: (i) consulta de datos primarios sobre la infraestructura de SLOD-BI; (ii) preparación de datos mediante un agregado tipo suma, cálculo de la fórmula asociada al indicador KPI_Impact, preparación de datos para generar las tripletas de cada observación; generación de tripletas de cada observación (*IndicatorFact*); (iii) almacenado/publicación de las observaciones. La figura 11 muestra un fragmento de las tripletas asociadas a las observaciones generadas en KNIME.

Figura 8 – Captura de pantalla de la simulación del proceso ETLink en Knime (etapa i)



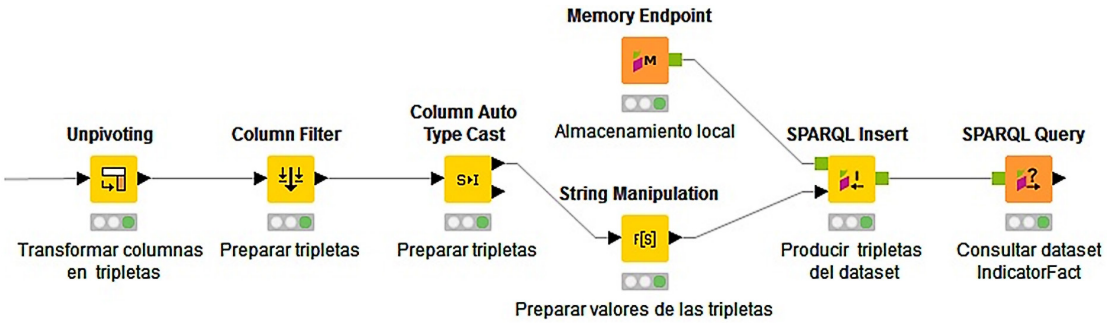
Fuente: Elaborada por los autores, 2017.

Figura 9 – Captura de pantalla de la simulación del proceso ETLink en Knime (etapa ii)



Fuente: Elaborada por los autores, 2017

Figura 10 – Captura de pantalla de la simulación del proceso ETLink sobre Knime (etapa iii)



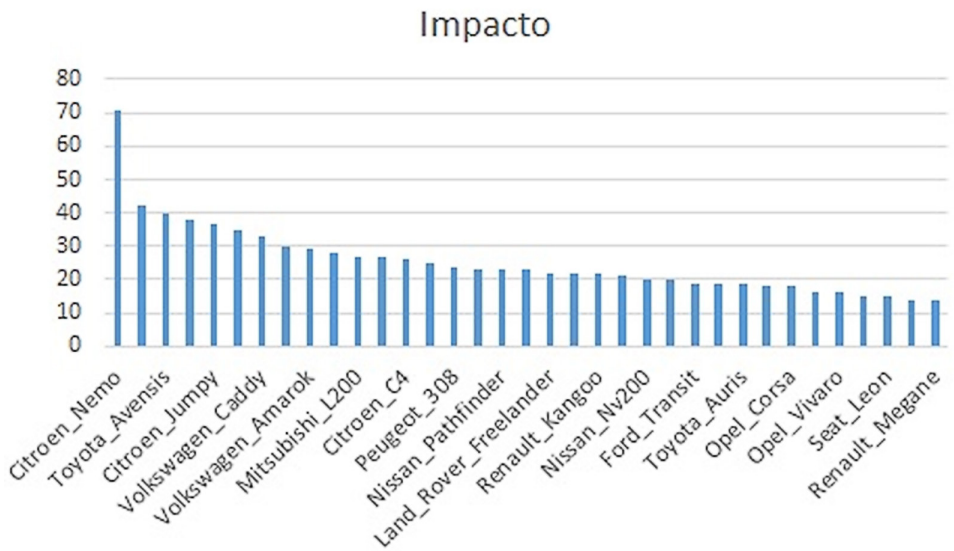
Fuente: Elaborada por los autores, 2017

Figura 11 – Resultados obtenidos en el flujo de ETLink asociado al indicador social Impact

sub	pred	obj
http://krono.act.uji.es/slodbi/IndicatorFact/Row29	slod:hasIndicatorAssociated	<http://krono.act.uji.es/SocialIndicator/KpiImpact>
http://krono.act.uji.es/slodbi/IndicatorFact/Row29	sdmx-measure:obsValue	14
http://krono.act.uji.es/slodbi/IndicatorFact/Row29	sdmx-dimension:timePeriod	<http://reference.data.gov.uk/id/month/2014-10>
http://krono.act.uji.es/slodbi/IndicatorFact/Row29	s:itemReviewed	http://krono.act.uji.es/datasets/cars/Renault_Megane
http://krono.act.uji.es/slodbi/IndicatorFact/Row29	s:dateCreated	2016-11-20T15:45:31.777322+01:00
http://krono.act.uji.es/slodbi/IndicatorFact/Row3	slod:hasIndicatorAssociated	<http://krono.act.uji.es/SocialIndicator/KpiImpact>
http://krono.act.uji.es/slodbi/IndicatorFact/Row3	sdmx-measure:obsValue	71
http://krono.act.uji.es/slodbi/IndicatorFact/Row3	sdmx-dimension:timePeriod	<http://reference.data.gov.uk/id/month/2014-10>
http://krono.act.uji.es/slodbi/IndicatorFact/Row3	s:itemReviewed	http://krono.act.uji.es/datasets/cars/Citroen_Nemo
http://krono.act.uji.es/slodbi/IndicatorFact/Row3	s:dateCreated	2016-11-20T15:45:31.777322+01:00
http://krono.act.uji.es/slodbi/IndicatorFact/Row9	slod:hasIndicatorAssociated	<http://krono.act.uji.es/SocialIndicator/KpiImpact>
http://krono.act.uji.es/slodbi/IndicatorFact/Row9	sdmx-measure:obsValue	22
http://krono.act.uji.es/slodbi/IndicatorFact/Row9	sdmx-dimension:timePeriod	<http://reference.data.gov.uk/id/month/2014-10>
http://krono.act.uji.es/slodbi/IndicatorFact/Row9	s:itemReviewed	http://krono.act.uji.es/datasets/cars/Land_Rover_Freelander
http://krono.act.uji.es/slodbi/IndicatorFact/Row9	s:dateCreated	2016-11-20T15:45:31.777322+01:00

Fuente: Elaborada por los autores, 2017

Gráfica 1 − Gráfica con las observaciones del indicador Impacto para octubre de 2014.



Fuente: Elaborada por los autores, 2017

La gráfica 1 muestra los valores calculados para el indicador social Impacto, sobre distintas marcas de coche, en el mes de octubre del 2014, en el ámbito de la red social Twitter. Los resultados están ordenados decrecientemente para identificar, a simple vista, la marca con más/menos impacto en el período observado. Estos valores pueden ser comparados con los resultados de otros indicadores sociales (por ejemplo, el uso de diversas fórmulas para estimar un mismo indicador social). También pueden ser contrastado con datos corporativos (ej. volumen de alquileres por marca) para una mejor comprensión de los resultados y facilitar la toma de decisiones.

CONCLUSIONES

En este artículo se ha presentado una metodología para la definición y seguimiento de indicadores sociales sobre la infraestructura de datos enlazados y abiertos denominada SLOD-BI. Esta metodología se ha basado también en los principios de los datos LOD, creando y publicando como datos semánticos tanto las definiciones de los indicadores sociales como las observaciones realizadas sobre la propia infraestructura de datos.

Entre los principales beneficios de esta metodología destaca el hecho de que los indicadores estén directamente enlazados con las métricas y los datos que las alimentan, de forma que es posible identificar de forma sencilla el origen de los valores de dichos indicadores. Por otro lado, el hecho que los indicadores sean también datos semánticos permite aplicar técnicas de validación sobre sus fórmulas. Como trabajo futuro se estudiará la creación automática de las descripciones y consultas SPARQL asociadas a las fórmulas KPIs, así como el descubrimiento de métricas adecuadas a determinados objetivos estratégicos.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el Ministerio de Economía y Comercio con el proyecto del Plan Nacional de I+D con número de contrato TIN2014-55335-R.

REFERENCIAS

ABELLÓ, A. et al. Using Semantic Web Technologies for Exploratory OLAP: A Survey. *IEEE Trans. on Knowledge and Data Engineering*, p.571 – 588, 2014.

ABRAHAMS, A. S. Vehicle defect discovery from social media. *Decision Support Systems*, v. 54, n.1, p.87-97, 2012.

BERLANGA, R.; NEBOT, V. Context-Aware Business Intelligence. *Lecture Notes in Business Information Processing,* v. 253, p. 87-110, 2015.

BERLANGA, R. et al. SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence. *International Journal on Data Warehousing and Data Mining*, v.11, n.4, p. 1-28. 2015.

CAREY, M. J.; ONOSE, N.; PETROPOULOS, M. Data Services. Commun. ACM , v.55, n. 6, p. 86-97, 2012.

CHAE, B. K. Insights from hashtag# supplychain and Twitter analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Production Economics*, v. 165, p. 247-259, 2015.

CHEN, H.; CHIANG, R. H.; STOREY, V. C. Business intelligence and analytics: from Big Data to big impact. *MIS Q.*, v. 36, n.4, p.1165-1188, 2012.

CODD, E. F.; CODD, S. B.; SALLEY, C. T. Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. E. F. Codd and Associates. 1993.

DAI, W. et al. Emotion recognition and affective computing on vocal social media. *Information & Management*, v. 52, n.7, p. 777-788, 2015.

DIAMANTINI, C.; POTENA, D.; STORTI, E. SemPI: A semantic framework for the collaborative construction and maintenance of a shared dictionary of performance indicators. *Future Generation Comp. Syst.*, v.54, p. 352-365, 2016.

FAN, W.; GORDON, M. D. The Power of Social Media Analytics. *Communications of the ACM*, v. 57, n.6, p. 74-81, 2014.

GARCÍA-MOYA, L. Modeling and analyzing opinions from customer reviews. 2016. Tesis (Doctoral) - Universidad Jaume I, Castellón de La Plana, 2016.

GONZÁLEZ, N.; MENÉNDEZ, J. L. ; C. SEOANE. Revisión y propuesta de indicadores (KPI) de la Biblioteca en los medios sociales. *Revista Española de Documentación Científica*, v.36, n.1, 2013.

PRICEWATERHOUSECOOPERS . *Guideline Key Performance Indicators*. Melbourne: Public Record Office Victoria. 2015.

HE, W. et al. A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, v.52, n.7, p. 801-812, 2015.

HEATH, T.; BIZER, C. *Linked Data*: Evolving the Web into a Global Data Space. US: Morgan & Claypool, 2011.

HORKOFF, J. et al. Strategic business modeling: representation and reasoning. *Software and System Modeling*, v.13, n.3, p. 1015-1041, 2014.

MATÉ, A.; TRUJILLO, J.; MYLOPOULOS, J. Conceptualizing and Specifying Key Performance Indicators in Business Strategy Models. In: Atzeni, P.; Cheungm, D.; Ram, S. (Ed.). *Lecture Notes in Computer Science.* Berlin: Springer, 2012. p. 282-291.

MUÑOZ VERA, G.; ELÓSEGUI, T. *El arte de medir:manual de analítica web*. Barcelona: Profit, 2011.

OLSZAK, C. M. & ZIEMBA, E. Approach to Building and Implementing Business. *Interdisciplinary Journal of Information, Knowledge, and Management*, 2007.

PARMENTER, D. *Key Performance Indicators*: developing, implementing, and using Winning KPIs. Hoboken: John Wiley & Sons, 2015.

PETERSON, E. T. The Big Book of Key Performance Indicators. Web Analytics Demistyfied. 2006.

WANG, G. A. et al. ExpertRank: a topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems*, v.54, n.3, p. 1442-1451, 2013.

YAN, Z. et al. EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. *Information & Management*, v. 52, n. 7, p. 850-858, 2015.

ZHOU, M. et al. Social Media Adoption and Corporate Disclosure. *Journal of Information Systems*, v. 29, n.2, p. 23-50, 2015.