Revista Española de Documentación Científica
42(2), abril-junio 2019, e234
ISSN-L:0210-0614. https://doi.org/10.3989/redc.2019.2.1599

ESTUDIOS / *RESEARCH STUDIES*

Implementación de un sistema de detección de señales débiles de futuro mediante técnicas de minería de textos

Israel Griol-Barres*, Sergio Milla*, José Millet**

*Vicerrectorado de Empleo y Emprendimiento, Universitat Politècnica de València
Correo-e: igriol@ideas.upv.es | ORCID iD: https://orcid.org/0000-0002-2197-9161
Correo-e: sermilmo@inf.upv.es | ORCID iD: https://orcid.org/0000-0001-9461-0165

** Instituto ITACA, Universitat Politècnica de València
Correo-e: jmillet@eln.upv.es | ORCID iD: https://orcid.org/0000-0002-8879-003X

Recibido: 27-06-2018; 2ª versión: 05-10-2018; Aceptado: 08-10-2018.

Cómo citar este artículo/Citation: Griol-Barres, I.; Milla, S.; Millet, J. (2019). Implementación de un sistema de detección de señales débiles de futuro mediante técnicas de minería de textos. *Revista Española de Documentación Científica*, 42 (2): e234. https://doi.org/10.3989/redc.2019.2.1599

Resumen: Actualmente, una de las mayores amenazas para las empresas es no ser capaces de hacer frente a los cambios constantes que se dan en el mercado, por no predecirlos con la suficiente antelación. Por ello, el desarrollo de nuevos procesos que faciliten la detección de fenómenos y cambios futuros significativos es una componente clave para una correcta toma de decisiones que marque un rumbo correcto para la empresa. Por esta razón, se propone un sistema basado en una arquitectura de inteligencia de negocio que permite detectar cambios discretos o señales débiles (*weak signals*) en el presente, pero que son indicativos de fenómenos más significativos y cambios trascendentales en el futuro. Frente a los trabajos actuales que se centran en fuentes de información estructuradas, o como mucho, con un único tipo de fuente de datos, en este trabajo la detección de estas señales se realiza de forma cuantitativa a partir de documentos heterogéneos y no estructurados de diversa índole (artículos científicos, periodísticos y redes sociales) sobre los que se aplican técnicas de minería de textos. El sistema ha sido testeado para estudiar el futuro del sector de los paneles solares, habiéndose obtenido resultados prometedores para ayudar a expertos en el reconocimiento de nuevos factores de peso en sus mercados y en el desarrollo de nuevas oportunidades.

Palabras clave: señales débiles de futuro; arquitectura de inteligencia de negocio; información no estructurada; minería de textos; toma de decisiones.

System implementation for detection of future weak signals using text mining

Abstract: Nowadays, one of the biggest threats for companies is not being able to cope with the constant changes occurring in the market by not predicting them well in advance. For this reason, the development of new processes that facilitate the detection of future phenomena and significant changes is a key component for correct decision making that can mark a correct course in the company. A business intelligence based architecture system is proposed to allow discrete changes or weak signals detection in the present that are indicative of more significant phenomena and transcendental changes in the future. In contrast with current available works, which are focused on structured information sources or, at most, with only a single type of data source, in this paper the detection of these signals is done quantitatively from various kinds of heterogeneous and unstructured documents (scientific articles, journalistic articles and social networks) on which text mining techniques are applied. The system has been tested in the study of the future of solar panels sector, obtaining promising results that can help business experts in the recognition of new driving factors of their markets and the development of new opportunities.

Keywords: weak signal of the future; business intelligence architecture; unstructured information; text mining; decision-making.

1. INTRODUCCIÓN

Una de las mayores amenazas para las empresas es el constante ritmo de cambios en los mercados. En multitud de casos, las empresas demuestran incapacidad para gestionar y prever estas evoluciones a tiempo (Eisenhardt y Brown, 1999). Los mercados han demostrado ser entornos complejos en los que puede llegar a ser muy complicado tomar la decisión correcta en el momento acertado, pero, sin duda, hacerlo puede marcar el buen devenir de la empresa.

Por ello, cada día cobra mayor importancia el desarrollo de nuevos procesos que faciliten la toma de decisiones en organizaciones, que consideren datos provenientes de diferentes fuentes internas y externas. Como el volumen de datos disponible es cada vez mayor, estos procesos deben involucrar técnicas de captura, transformación, almacenamiento y análisis automático que reduzcan el tiempo y los medios necesarios de análisis, a la vez que proporcionen una gran fiabilidad. Entre ellas, destaca la minería de datos en procesos de inteligencia empresarial o *Business Intelligence* (Dedić y Stanier, 2017).

En la actualidad, las tecnologías de la información asumen un papel importante en los negocios, debido a su papel fundamental en la creación de inteligencia empresarial. Este término incluye un amplio abanico de metodologías para recopilar, procesar y analizar enormes cantidades de datos almacenados en la base de datos de una empresa (Conesa-Caralt y Curto-Diaz, 2010). Además, su objetivo es generar un conocimiento nuevo a partir de información almacenada (Khan, 2012) que permita interpretar, predecir y responder apropiadamente al mundo exterior (Fischler y Firschein, 1987).

Por otro lado, la minería de datos (Witten y Frank, 2005) se basa en extraer conocimiento relevante a partir de documentos y datos de distintas fuentes. En el mundo empresarial, la palabra "futuro" representa la base para la identificación de nuevas oportunidades de negocio potenciales (Yoo y otros, 2009), y multitud de expertos están trabajando en distintos métodos de análisis: evolutivos, detección de patrones, métodos de minería, teorías de innovación disruptiva o de detección de señales de futuro.

Un tipo de señal de futuro es la señal débil o *weak signal* (Ansoff y McDonnell, 1990). Este término puede ser definido como la detección de una evidencia de un cambio emergente dentro de un proceso continuo de exploración en un medio concreto (Ansoff, 1975). Es decir, se trata de eventos externos o internos que están todavía demasiado incompletos como para permitir una estimación precisa de su impacto y/o para desarrollar una respuesta frente a ellos (Cooper y otros, 2011).

Sin embargo, estos cambios enmascaran el potencial para que se desarrollen fenómenos más significativos y cambios trascendentales en el medio, de ahí la importancia de poder identificarlos y monitorizarlos lo más pronto posible. Estos fenómenos, si evolucionan hasta hacerse relevantes (strong signals), tienen el potencial de reforzar un plan de actuación o de obstruirlo. Otra definición del término de weak signal refuerza esta misma idea: "*factores de cambio difícilmente perceptibles en el presente, pero que constituirán una fuerte tendencia en el futuro*" (Godet, 1994).

En conclusión, nos encontramos con tres términos interrelacionados entre sí: la inteligencia empresarial como disciplina, la minería de datos como procedimiento, y la identificación de señales débiles como objetivo.

El resto del artículo se estructura de la siguiente forma. En la Sección 2 se presentan los tipos de trabajos que han servido como antecedente para este estudio y se establecen sus objetivos. En la Sección 3 se explica detalladamente el diseño e implementación del sistema de detección de señales débiles propuesto. En la Sección 4 se interpretan y evalúan los resultados obtenidos aplicando el detector en el sector de los paneles solares. Finalmente, en la Sección 5 se presentan las conclusiones y las líneas de trabajo futuro.

2. ANTECEDENTES Y OBJETIVOS

Aunque las señales débiles o weak signals han ido ganando interés entre los trabajos recientes sobre la toma de decisiones y la predicción de cambios futuros, todavía no hay un uso extendido del término, y hay autores que han utilizado otros sinónimos como, por ejemplo, "semillas de cambio" (Molitor, 2003), "hechos emergentes" (Dator, 2005), "señales de estrategia" (Nikander, 2002) y "señales de aviso precoz" (Mannermaa, 1999).

Todos estos términos similares se engloban dentro del concepto de señal de futuro o *future sign* (Hiltunen, 2008), que, a su vez, surgió a partir del modelo semiótico del signo (Peirce, 1868). Aunque ambos se definen mediante un marco conceptual compuesto de tres dimensiones, el modelo de Hiltunen es justamente específico para señales de futuro.

El modelo en triada de Peirce consiste en un "objeto", es decir, la "porción" de la realidad a la que se accede a través del signo en sí, un "representamen", o la representación simbólica de algo, y de un "interpretante", que se relaciona con la

interpretación que una cultura hace del signo mediante su propio conocimiento.

Por otro lado, la triada del signo de futuro de Hiltunen consta de estas tres dimensiones: "tema" (grado de difusión), "señal" (grado de visibilidad) e "interpretación" (diversidad de fuentes), con el objetivo de profundizar en la diferencia entre señal débil y fuerte (Kuusi y Hiltunen, 2007). En la Figura 1, podemos ver la comparación entre los modelos de triada de Peirce y Hiltunen. Generalmente, una señal débil tendrá un valor absoluto muy bajo en una, dos o en las tres componentes. Por esta razón, pueden pasar desapercibidas.

Un ejemplo sobre el uso de este modelo es una historia publicada en el periódico más importante de Finlandia (Helsingin Sanomat, 2010) en la que era noticia que la cadena de ropa sueca Hennes & Mauritz (H&M) estaba vendiendo ropa vieja, al precio de nueva, bajo la etiqueta de "vintage". En realidad, únicamente el 1% de sus tiendas estaban realizando esta acción. Desde el punto de vista de señal de futuro, el valor de "señal" (es decir, su visibilidad) era enorme. Pero la realidad, el tema o su grado de difusión, es que el 1% de las tiendas es poco representativo. La componente de interpretación es también poco significativa, puesto que únicamente la prensa, y en concreto este diario, había publicado la noticia. Se podría considerar que esto representa una señal débil puesto que únicamente una de las tres componentes era fuerte y, por lo tanto, a priori puede resultar difícil predecir si esta acción de H&M representaría una tendencia fuerte en el futuro o no.

Una organización empresarial que opera en un entorno complejo e impredecible tiene que ser flexible para poder detectar este tipo de información. Aunque las implicaciones de una señal débil son muy difíciles de definir en una primera etapa, toda organización se ve forzada a tomar decisiones con cada vez mayor antelación, y con datos disponibles cada vez más emergentes. Por lo tanto, el tiempo disponible para reaccionar se acorta, a la vez que la organización se vuelve más compleja, siendo la detección de estas señales débiles de cambio una prioridad para poder tomar decisiones acertadas.

Para ello, la organización debe escanear y analizar su entorno de forma frecuente. Los primeros sistemas teóricos de análisis de entornos estaban basados en multitud de filtros (Ansoff y McDonnell, 1990), basados en tres conceptos: (i) Un filtro de vigilancia que incluye metodologías de análisis para la adquisición de información, (ii) un filtro de mentalidad para la percepción de la organización acerca de esa información adquirida, y (iii) un filtro de potencia, para evitar que la señal de futuro pase desapercibida. (Ilmola y Kuusi, 2006). La Figura 2 muestra cómo estos tres filtros interactúan en un sistema de detección de señales débiles.

Estos filtros están conectados con el modelo de Hiltunen. El filtro de vigilancia está identificado con la componente de "tema" que mide el grado de difusión de una señal, es decir, el número de fuentes que se hacen eco de ella. El valor absoluto de la señal a detectar identifica el filtro de potencia con la componente "señal". Este filtro, por lo tanto, vigila la visibilidad de la señal, es decir, la cantidad de veces que aparece esa señal (independientemente del número de fuentes o

Figura 1. (a) Modelo semiótico del signo de Peirce, y (b) Modelo semiótico del signo futuro de Hiltunen

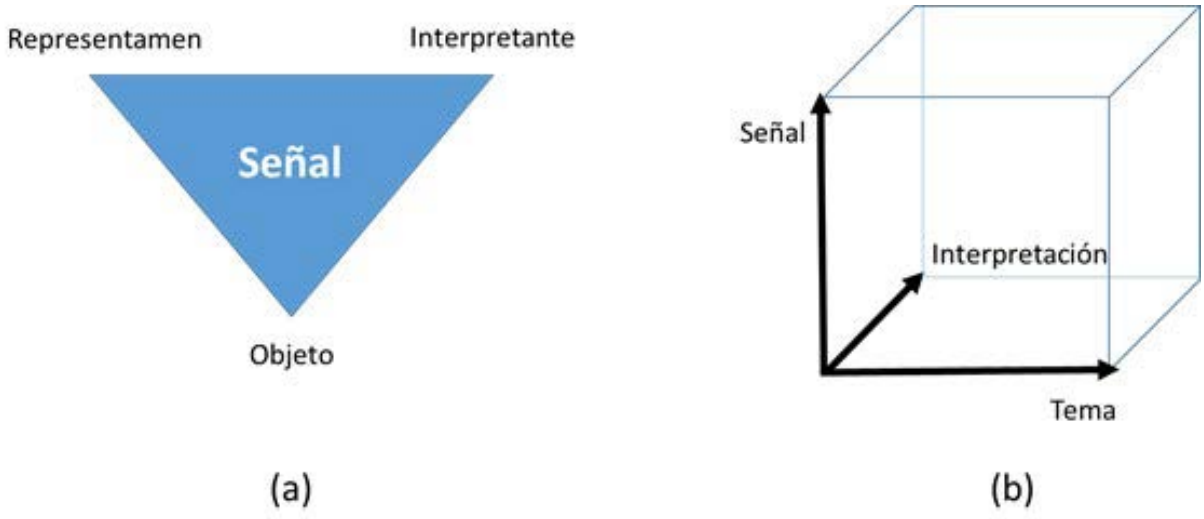
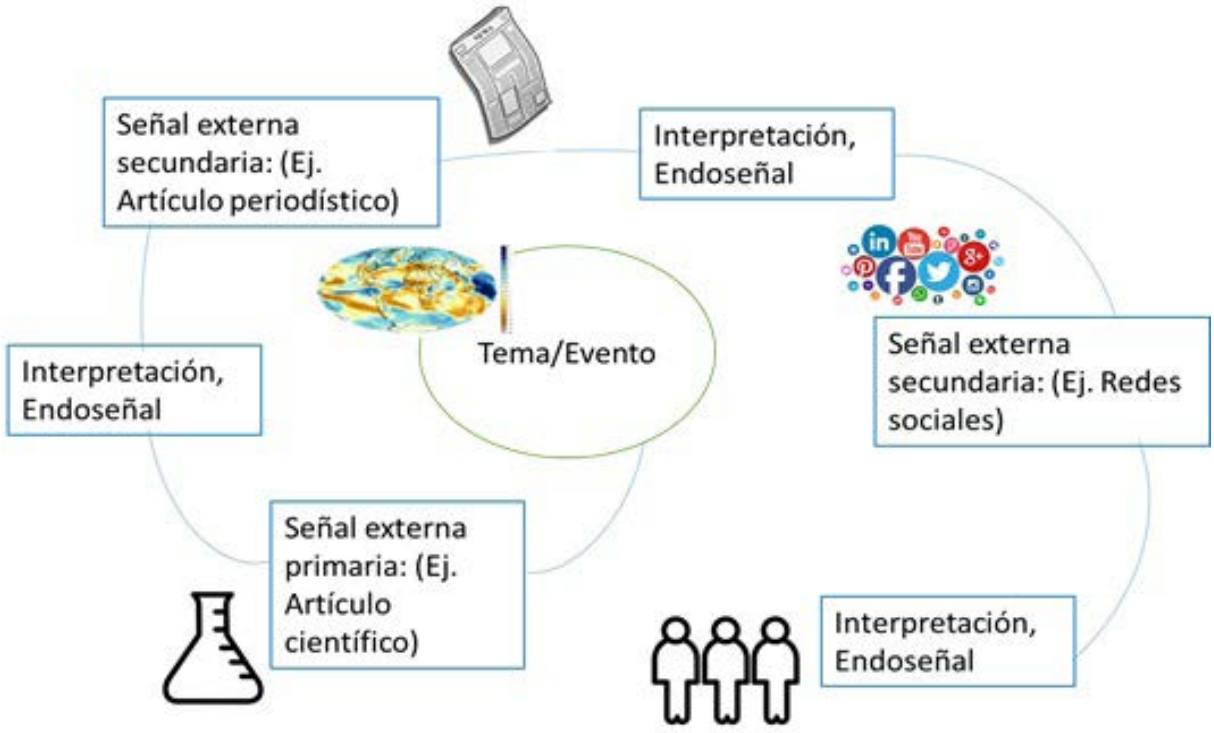


Figura 2. Los filtros de Ansoff para analizar entornos y detectar señales débiles

documentos en las que aparece). Por último, cada tipo de fuente interpretará la señal de una manera distinta, a veces con mayor o menor importancia, estando por lo tanto conectado el filtro de mentalidad con la componente "Interpretación".

Otro concepto importante es el proceso de significación (Kuusi y Hiltunen, 2007). Un evento genera una primera exoseñal (una señal externa a la organización) que podría ser un resultado de investigación publicado en forma de artículo en una revista científica. En el siguiente paso, si supera el filtro de mentalidad, esta exoseñal puede ser interpretada por un periodista como un contenido mediático, transformándose de nuevo en una endoseñal para él, para luego codificarla en forma de una exoseñal secundaria, un artículo periodístico. Del mismo modo, esta señal puede ser codificada por otros actores que la interpretan y comparten en redes sociales, como Twitter. Por lo tanto, interpretar es una actividad en la que hay actores que formulan endoseñales propias basadas en las exoseñales del tema en sí, seguidas de exoseñales secundarias que son a la vez interpretadas y transformadas en nuevas exoseñales. Este proceso está ilustrado en la Figura 3.

Figura 3. El proceso de significación en el que se aprecia la difusión en el tiempo de las señales



Es importante conocer y comprender el proceso de significación, puesto que tiene relevancia en las componentes del signo de futuro. Por un lado, es importante para la componente "tema" puesto que el proceso de significación explica la difusión de una señal en fuentes de distintos tipos. La "interpretación" depende en buena medida de la fuente de la que provienen los datos. Por último, indirectamente también una mayor difusión de la señal se verá reflejada en una mayor componente de "señal".

La mayoría de las publicaciones acerca de la detección de señales débiles están principalmente relacionadas con temas específicos y con análisis cualitativos que tienen en cuenta estas tres componentes y en algunos casos, el proceso de significación. Por ejemplo, hay trabajos que realizan análisis cualitativos acerca de la identificación de señales débiles relacionadas con el terrorismo en ataques de transporte (Koivisto y otros, 2016) o en la influencia en redes sociales (MohamadiBaghmolaei y otros, 2017).

Uno de los escasos estudios que existen sobre la detección de señales débiles que utilizan un análisis cuantitativo (Yoon, 2012) realiza una aproximación al problema desde la perspectiva de una única fuente: noticias web. La cuantificación utilizada en este trabajo consiste en la medida de la presencia de una señal débil a través del grado de visibilidad y del grado de difusión de un término.

Para detectar señales débiles es necesario un proceso sutil de observación y análisis, porque la información sobre estas señales está codificada en multitud de fuentes de datos. Por esta razón, cualquier sistema diseñado para la detección de estas señales requerirá el uso de procesos que consumen en general muchos recursos de memoria y tiempo.

Con estos antecedentes, se plantea el diseño de un sistema de detección de señales débiles con los objetivos siguientes:

Objetivo 1. Confeccionar repositorios eficientes que almacenen una vasta cantidad de datos desestructurados de diversas fuentes sobre una temática en particular, sobre la cuál se está realizando el estudio.

Objetivo 2. Diseñar un sistema de detección de señales débiles que tenga en cuenta el máximo número de componentes del modelo semiótico del signo futuro como sea posible.

Objetivo 3. Diseñar un sistema de detección de señales débiles que se base en un análisis cuantitativo de las fuentes de entrada para obtener resultados más precisos.

Objetivo 4. Demostrar que es posible incluir técnicas de procesado del lenguaje natural, lo que requiere interfaces complejas (Griol y otros, 2016) y comprobar que esto facilita la obtención de resultados que ayuden a expertos y emprendedores en la toma de decisiones.

Objetivo 5. Asegurar que el diseño del sistema sea lo suficientemente eficiente para que su uso sea viable, con tiempos de ejecución razonables, capaz de obtener resultados válidos para expertos y emprendedores, independientemente de la temática bajo estudio.

3. DESARROLLO DE UN SISTEMA PARA LA DETECCIÓN DE SEÑALES DÉBILES

En esta sección se definen las proposiciones y la metodología del sistema de detección de señales débiles implementado. Además, se definen las fases seguidas en la detección y los métodos de minería de textos aplicados.

3.1. Definición del sistema realizado

Empresas de todo tipo han ido paulatinamente adoptando sistemas y herramientas de inteligencia empresarial debido a que les otorga una mejor gestión para aprender del pasado y predecir el futuro (Siegel, 2013). Estos sistemas pueden abastecerse de datos de muchas fuentes diferentes, como sistemas de información, informes, internet, bases de datos corporativas, clientes, proveedores, organismos gubernamentales, o el conocimiento de los empleados.

Por esta razón, se propone un sistema que se nutra de documentos de tres tipologías distintas: artículos científico-técnicos, artículos periodísticos y publicaciones en redes sociales. Dada la gran cantidad de documentos accesibles sobre una determinada temática, se requiere de varios repositorios para almacenarlos y gestionarlos.

La componente temporal es importante para estudiar la evolución de cada señal, por lo que los repositorios internos de información obtenidos de numerosas fuentes online se organizarán por años. Por lo tanto, en primer lugar, se requiere un algoritmo que pueda recopilar y almacenar una gran cantidad de documentos de distintas fuentes, relacionados con una temática, en numerosas bases de datos documentales, una por cada año del periodo que se ha determinado para poder realizar el análisis. Normalmente, una organización ya dispone previamente de bases de datos internas con información de sus actividades, pero en este caso, al tratarse de datos externos, es necesario crear esta recopilación como un paso previo.

El siguiente paso es la implementación de un sistema para gestionar el gran volumen de información del cual extraer el conocimiento necesario para la toma de decisiones. Se debe transformar la información de estos repositorios para almacenar únicamente los datos necesarios, en un formato útil, en un almacén de datos, o *data warehouse* (Giovinazzo, 2000). Para ello, se debe crear un algoritmo que extraiga la información, la transforme y la cargue en dicho almacén de datos (Kimball y Ross, 2002).

En el siguiente paso, una vez que la información está guardada en el almacén, se requiere de otro algoritmo que seleccione y transforme la información para ser tratada por modelos matemáticos, mediante la minería de texto. Luego, se necesita una técnica concreta para la detección de *weak signals* en textos.

En definitiva, se trata de diseñar un sistema que mida de forma cuantitativa señales débiles futuras mediante técnicas de minería de textos, es decir, crear una herramienta que facilite el análisis de un experto sobre un sector, asumiendo las siguientes proposiciones:

Proposición 1. Las palabras clave con muchas ocurrencias en una colección de documentos son importantes (Jung, 2010).

Proposición 2. Las apariciones recientes de palabras clave son más importantes o relevantes que las apariciones pasadas.

Proposición 3. Las dos primeras proposiciones son ciertas para cualquier tipo de fuente de datos externos utilizada.

Proposición 4. Se obtienen resultados más fidedignos cuando se utilizan distintas fuentes de datos, teniendo en cuenta el proceso de significación orientado al futuro y el procesado natural del lenguaje.

3.2. Metodología

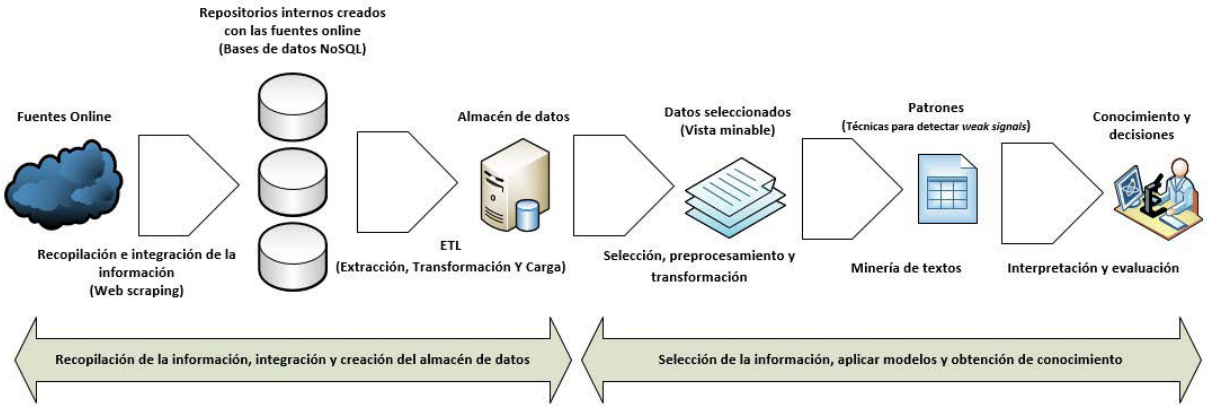
Para la realización del sistema y poder realizar un estudio del comportamiento de las palabras a lo largo de un periodo de tiempo apoyándose en grandes volúmenes de información bases de datos, se ha utilizado la metodología ***Knowledge Discovery in Databases*** (KDD), que se refiere al proceso no trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información (Han et al., 2001). La Figura 4 ilustra las etapas del proceso KDD que se han llevado a cabo en el sistema propuesto en este trabajo.

Como se puede observar en la figura anterior, se ha dividido el proceso en dos partes, la primera consiste en la recopilación de un gran volumen de información que será guardada en un almacén de datos y una segunda parte donde se seleccionará la información y se aplicarán modelos matemáticos para extraer conocimiento de ella.

Como también se puede observar, el proceso consiste en 5 etapas, las cuales son:

1. Recopilación e integración de la información
2. Fase ETL (Extraer, transformar y cargar)
3. Selección, preprocesamiento y transformación
4. Minería de textos enfocada a detectar *weak signals*
5. Interpretación y evaluación

Figura 4. Proceso KDD seguido en la implementación del sistema propuesto



El sistema debe tener capacidad para administrar el conocimiento, almacenarlo en un repositorio de conocimientos y herramientas que puedan aplicar ese conocimiento para una mejor toma de decisiones, y así dotar de una mayor inteligencia empresarial a las organizaciones.

3.3. Fase 1: Recopilación e integración de la información

Durante esta fase, se ha realizado un estudio previo de las fuentes de las que se obtiene la información necesaria. Como se ha explicado anteriormente, los tipos de fuentes son de tipo técnico-científico, periodístico y de redes sociales, y en el caso de este estudio, en inglés.

Una vez analizadas las fuentes online, se ha realizado una selección para determinar de cuales se extraería la información. Para ello, se tuvo en cuenta un factor primordial, los datos gestionados por dichas fuentes tendrían que ser relevantes dentro de su tipo, pero a la vez, tener formatos fáciles de manipular y así poder ser extraídos con mayor facilidad para su almacenamiento. Las fuentes seleccionadas con estos criterios fueron tres: DirectScience (Science Direct, 2018), de tipo técnico-científico, New York Times (New York Times, 2018), de tipo periodístico y por último Twitter (Twitter, 2018a), como fuente de redes sociales (Finger y Dutta, 2014).

Para la recopilación de la información se desarrolló un algoritmo en Python para extraer la información de documentos HTML y tweets y los almacenaría en bases de datos de tipo NoSQL.

En el algoritmo generado, se ha utilizado Beautiful Soup (Beautiful Soup, 2018), para la extracción de documentos técnicos, científicos y periodísticos. Beautiful Soup es una biblioteca de Python diseñada para analizar documentos HTML y que es muy útil para realizar "web scraping", es decir, la extracción de información de sitios web. Sin embargo, para la extracción de tweets se utilizó la propia API de Twitter (Twitter, 2018b).

Para los artículos científicos y técnicos se extrajo la siguiente información: título, autor, abstract o resumen, palabras claves, contenido, conclusiones y año de publicación. Esta información se almacena en una base de datos de tipo NoSQL documental. La base de datos de tipo documental elegida en nuestro caso ha sido MongoDB (MongoDB, 2018). MongoDB es una base de datos orientada a documentos y desarrollada con el concepto de código abierto, uno de los motivos para su elección, además de que guarda estructuras de datos en documentos similares a JSON haciendo que la integración sea más fácil y rápida (Connolly y Begg, 2005).

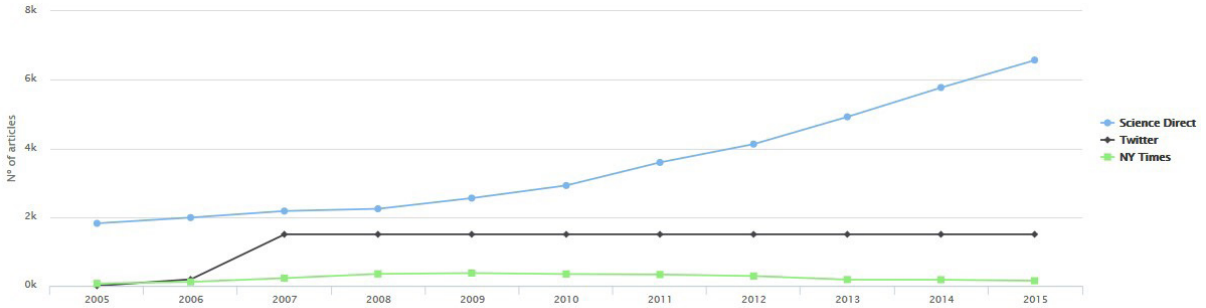
En el presente experimento centrado en la búsqueda de señales débiles se han seleccionado documentos relacionados con la temática de paneles solares. En la Figura 5 se muestra una gráfica con el número de documentos obtenidos por años en los distintos tipos de fuentes: 38.669 documentos científicos extraídos de ScienceDirect, 1.234 artículos periodísticos del New York Times y cerca de 45.191 tweets.

3.4. Fase 2: ETL (Extraer, transformar y cargar)

Una vez creados los repositorios internos, se trató el diseño e implementación del almacén de datos, que soporta el almacenamiento de un gran volumen de información. Se debe tener en cuenta que dicho almacén de datos debe estar orientado a temas, ser variable en el tiempo y no puede perder información almacenada (Inmon, 2005).

En el almacén de datos creado se guardan las palabras extraídas de los documentos almacenados en los repositorios internos del sistema de información, así como las siguientes propiedades relacionadas con las palabras: documento al que pertenece, frecuencia de aparición en el documento, año de aparición y fuente a la que pertenece el documento.

Figura 5. Número de documentos por año y fuente utilizados en el estudio sobre paneles solares



Para guardar esta información, se diseñó un algoritmo que lista los documentos almacenados por años y detecta la frecuencia de apariciones de las palabras en el documento. En esta fase se producen las acciones de extraer, transformar y cargar la información en el almacén. Este algoritmo realiza unas comprobaciones previas, para evitar insertar números, símbolos extraños y *stopwords*, es decir, palabras que no aportan nada ni a la semántica ni al significado del texto, y, por lo tanto, podemos descartar como *weak signals*. Para ello se ha utilizado un algoritmo de eliminación de *stopwords* incluido en la librería de tratamiento de lenguaje natural NLTK (Natural Language Toolkit, 2018). Esta etapa también incluye la lematización, es decir, quedarse con el lema, la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra. (Elmasri y Navathe, 2011).

3.5. Fase 3: Selección, preprocesamiento y transformación

Una vez almacenada la información en el almacén de datos, el siguiente paso es una etapa de "limpieza" en la que deben eliminarse datos inconsistentes, que no presentan información útil, con la finalidad de obtener una estructura de datos adecuada para la fase de transformación. Para este proceso, ha sido necesaria la creación de una base de datos de "palabras clave" (*keywords*), usando como referencia un listado de factores estándares facilitados por la UNESCO (UNESCO World Heritage Centre, 2008). Este listado está formado por una serie de categorías y subcategorías de distintas temáticas, como se puede observar en la Tabla I.

Tras este paso, es posible seleccionar las palabras sobre las cuales se realizarán las operaciones necesarias para obtener la frecuencia absoluta de aparición en documentos y frecuencia absoluta de aparición en el documento en cada año del periodo de tiempo analizado. Estas son variables necesarias para saber si una *keyword* señala una *weak signal* o no. De esta forma, se normalizan las *keywords* por año preparando los datos para la siguiente fase.

3.6. Fase 4: Minería de textos enfocada a detectar weak signals

La minería de textos (*text mining*) es una variación de la minería de datos (*data mining*) que se aplica en el proceso de obtener información de alta calidad de documentos de texto (Hernández y otros, 2004). Este proceso se caracteriza por estructurar los datos de entrada, construir modelos de análisis y analizar los resultados obtenidos. La gran diferencia con respecto a la minería de datos es que los patrones en la minería de textos se obtienen procesando el lenguaje natural en lugar de procesando bases de datos estructuradas. En el caso de este estudio, además, se han generado las bases de datos de documentos desde cero.

Tabla I. Listado de algunas de las cualidades para categorizar y subcategorizar palabras claves

Biological resource use

• Fishing aquatic resources
  • *Trawling*
  • *Netting*
  • *Line fishing*
• Aquaculture
  • *Marine*
  • *Freshwater*
• Land conversion
  • *Agriculture*
  • *Rural*
  • *Forestry*

Buildings and Development
• Housing • Commercial development • Industrial areas
Climate change and severe weather events
Health
Information and communication technologies
Invasive/alien species or hyper-abundant species
Local conditions affecting physical fabric
Management and institutional factors
Other factor(s)
Other human activities
Physical resource extraction
Pollution
Social/cultural uses of heritage
Sudden ecological or geological events
Transportation Infrastructure
Utilities or Service Infrastructure

Para el emparejamiento de *keywords* con las palabras del texto, y para saber si dicha palabra es una *weak signal*, se debe estudiar el incremento o decremento de apariciones en total y el número de documentos en los que aparece.

De esta manera, se tienen en cuenta dos de las tres dimensiones del modelo semiótico de Hiltunen. La dimensión de "señal" de un signo futuro se relaciona con la visibilidad de la señal futura. Para ello, se utiliza la frecuencia de ocurrencia de cada palabra en un conjunto de documentos (de la misma o de distintas fuentes), para definir el grado de visibilidad (DoV) de la *keyword i* en el periodo *j* se emplea la siguiente ecuación:

$$DoV_{ij} = \left(\frac{TF_{ij}}{NN_j}\right) \times \{1 - tw \times (n - j)\}$$

donde *TFij* es el número total de ocurrencias de la palabra i en el periodo *j* (considerando todos los documentos), *NNj* es el número total de documentos en el periodo *j*, *n* es el número de periodos y *tw* es un peso de tiempo, de acuerdo con la proposición de que las nuevas apariciones son más relevantes, que ha sido definido como 0,05 por un grupo de expertos en negocios relacionados con paneles solares (Yoon, 2012).

La dimensión "tema" indica el grado de difusión de los temas relacionados con las *weak signals.* Esta dimensión se relaciona directamente con la frecuencia de ocurrencia de cada palabra en cada documento puesto que esta frecuencia se adopta generalmente para medir cómo de general es un término en una colección de información textual (Salton y Buckley, 1988). Para medir el grado de difusión (DoD) de la palabra *i* en el periodo *j* se aplica la fórmula de cuantificación aplicada en un estudio previo sobre *weaks signals*:

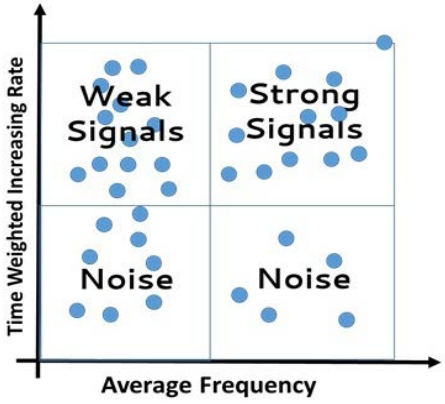
$$DoD_{ij} = \left(\frac{DF_{ij}}{NN_j}\right) \times \{1 - tw \times (n - j)\}$$

donde *DFij* es el número de documentos en los que aparece la palabra i en el periodo j.

Con estas fórmulas, se obtiene la media de ratios de incremento (DoD y DoV) de cada palabra encontrada en la multitud de documentos analizados con respecto a las frecuencias obtenidas por años. Con estos datos podemos generar dos gráficas, una para representar un mapa de apariciones de palabras clave y otra para representar un mapa de difusión de palabras clave.

Cada una de estas gráficas de mapas de *keywords* están compuestas de cuatro cuadrantes, por encima de un umbral de ratio de incremento ponderado con el tiempo, tenemos dos áreas, la de "*Strong Signals*" por encima de un umbral de frecuencia media, y la de "*Weak Signals"* por debajo de ese umbral, tal y como podemos ver en la Figura 6.

Figura 6. Estructura de un mapa de palabras clave



4. INTERPRETACIÓN Y EVALUACIÓN DE LOS RESULTADOS OBTENIDOS

En el presente análisis cuantitativo, podemos concluir que los temas que contienen señales débiles están relacionados con *keywords* de baja frecuencia de ocurrencia en valor absoluto, pero con un alto rango de fluctuación en el incremento de frecuencia de ocurrencia. Por otro lado, serán señales fuertes aquellas palabras clave con alta frecuencia de ocurrencia y con alto grado de incremento. El mapa de apariciones de palabras clave (*keyword emergence map*), representado en la Figura 7, se construye usando la media geométrica de ratio de incremento ponderado con el tiempo de cada palabra clave (DoD) frente a la media absoluta de apariciones de ese término.

Como se ha expresado anteriormente, el clúster de términos situados en la región de una frecuencia de aparición baja en nivel absoluto, pero con una media alta de ratio de incremento, representan *weak signals*. Estudiando la gráfica obtenida, podemos extraer la siguiente Tabla II en la que representamos los términos de *weak signals* más relevantes encontrados.

De forma similar, señales futuras que tienen posibilidad de ser *weak signals* son términos que presentan un patrón anormal de ratio de incremento, pero raramente difundidos, es decir, con una frecuencia baja en valor absoluto de documentos en los que aparece ese término. El mapa de difusión de palabras clave (*keyword issue map*) se calcula usando la media geométrica de la ratio de incremento de frecuencia de documentos (DoV) ponderado con el tiempo, y la media de frecuencia de documentos de cada elemento clave (Figura 8).

Del mismo modo, estudiando esta segunda gráfica obtenida, podemos extraer la siguiente Tabla III en la que representamos los términos de *weak signals* más relevantes encontrados.

Una de las palabras clave descubierta en ambas listas obtenidas es, curiosamente, el nombre del continente África que, por lo tanto, podemos concluir, tiene un comportamiento de una *weak signal*. En la tabla IV podemos ver las cifras obtenidas para este término.

Como consecuencia y para cerciorarnos de la posibilidad de que África es una palabra clave relacionada con una *weak signal* en el ámbito de los paneles solares, estudiamos los documentos que contienen esa palabra clave. De esta manera, pudimos encontrar documentos que mostraban proyectos futuros relacionados con este continente, que a partir de 2016 se están convirtiendo en realidad, como muestran los artículos destacados en la Figura 9 (Cembrero, 2011), (Cooke, 2015) y (Elcacho, 2014).

Figura 7. Mapa de apariciones de palabras clave (Keyword Emergence Map)

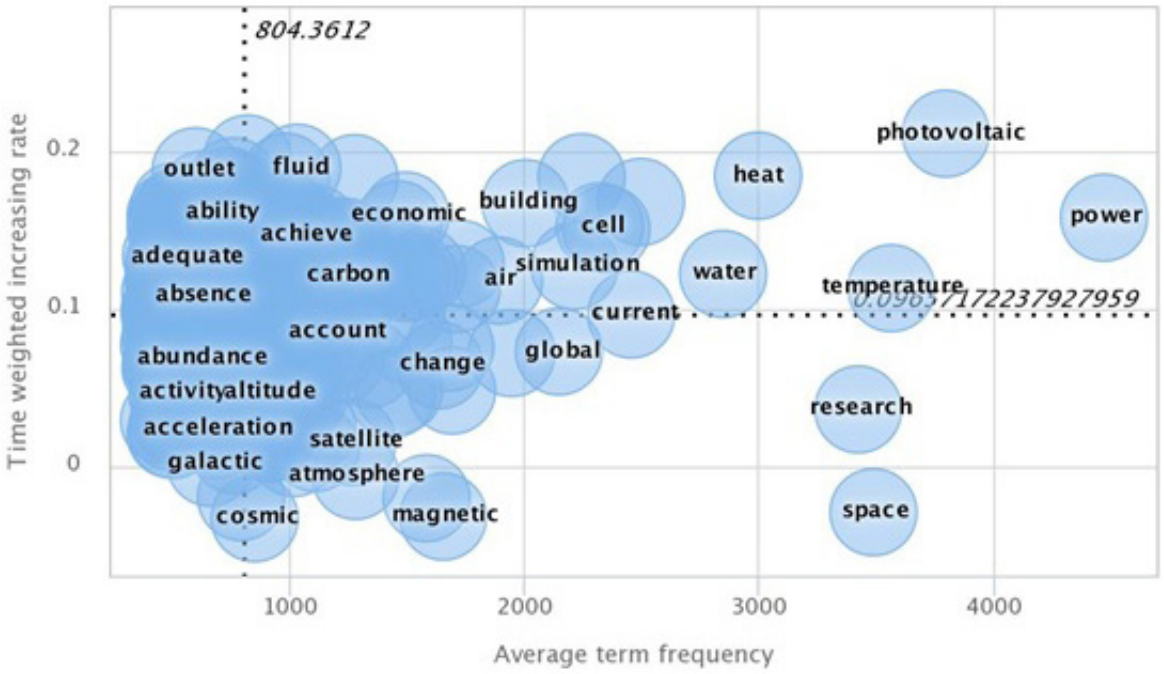


Tabla II. Listado de las *keywords* de *weak signals* más relevantes encontradas con respecto al mapa de apariciones de palabras clave

keyword	Category	Increasing rate	frequency
outlet	Management and institutional factor	0,18768145	593
concentrator	Information and communication technologies	0,18133683	754
outdoor	Buildings and Development	0,17502268	738
optimize	Other factor	0,17474913	721
prices	Management and institutional factor	0,17295696	623
mechanical	Utilities or Service Infrastructure	0,17205154	775
capital	Management and institutional factors	0,16823917	650
airconditioning	Utilities or Service Infrastructure	0,16553072	504
adoption	Other human activities	0,16541124	659
strategies	Management and institutional factor	0,16486441	759
actual	Other factor	0,16233475	758
analytic	Other factor	0,16177472	514
ability	Other factor	0,16093376	701
consumers	Management and institutional factors	0,16023764	597
fabricated	Buildings and Development/ Utilities or Service Infrastructure	0,15983769	774
flatplate	Other factor	0,15966764	506
nano	Information and communication technologies	0,15938169	487
decision	Management and institutional factors	0,15852001	587
influencing	Management and institutional factor /Other human activities	0,15820524	538
periodic	Management and institutional factor / Other factor	0,15560535	489

Figura 8. Mapa de difusión de palabras clave (Keyword Issue Map)

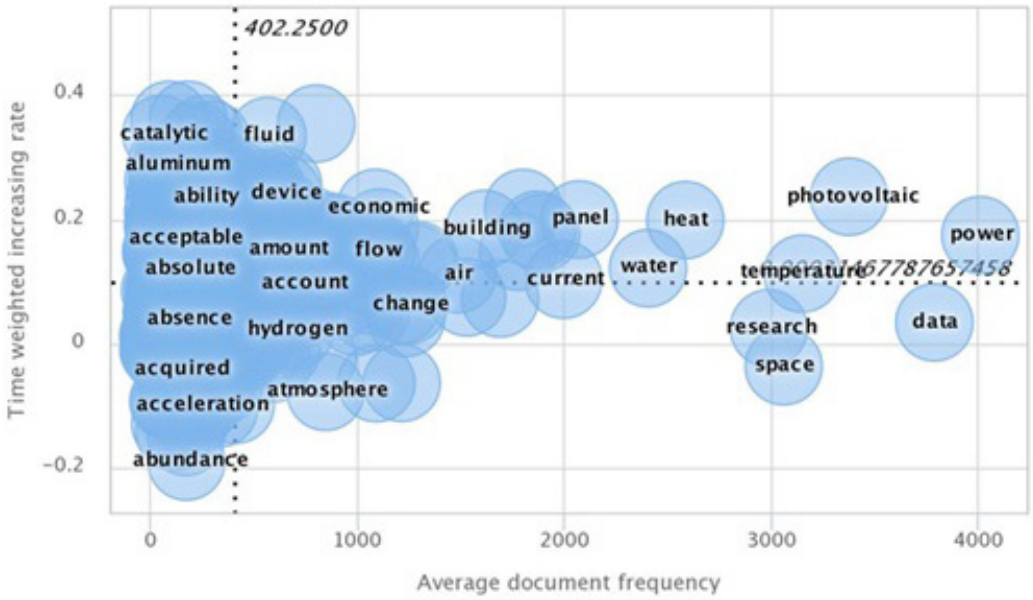


Tabla III. Listado de las *keywords* de *weak signals* más relevantes encontradas con respecto al mapa de difusión de palabras clave

keyword	Category	increasing_rate	frequency
por	Management and institutional factors /Other human activities	0,35980059	179
nano	Information and communication technologies	0,35980059	89
catalytic	Information and communication technologies	0,33663203	50
prices	Buildings and Development/ Management and institutional factors	0,33663203	253
efficiency	Management and institutional factors/Other factor	0,32803338	291
performance	Management and institutional factors / Other factor	0,32625195	220
flatplate	Other factor	0,29872736	109
capital	Management and institutional factors	0,28979791	206
contributed	Management and institutional factor	0,28747598	74
aluminum	Physical resource extraction	0,28747598	117
benefit	Management and institutional factor	0,28236966	210
tank	Utilities or Service Infrastructure	0,28205423	325
root	Biological resource use	0,28027481	170
fabrication	Buildings and Development / Management and institutional factor	0,26649519	269
farms	Biological resource use	0,26189761	60
assembled	Other factor	0,26189761	78
mobility	Transportation Infrastructure	0,26189761	142
manufacturers	Buildings and Development / Management and institutional factor	0,26189761	109
parabolic	Information and communication technologies / Utilities or Service Infrastructure	0,25954105	329
attractive	Other factor	0,25714414	254

Tabla IV. Cifras obtenidas de media geométrica de DoD y DoV de la palabra clave "África"

Keyword	type	increasing_rate	frequency
Africa	dod	0,264674169	196
Africa	dov	0,156120208	581

Figura 9. Algunos artículos periodísticos con la palabra clave "África"

Otras palabras clave encontradas en ambas listas de DoD y DoV son, entre otras: aceleración, ácidos, actuadores, adsorción, aerosol, asequible, agricultura, Ahmed, aire acondicionado, Argelia, alcalino, aleación, alteración, alternativas, aluminio, ángulos, argón, árido, arte, asiático, astronomía, atmósfera, automático, barrera, *benchmark*, cuenca, billón, negro, binario, Boston, caja, cátodo, calcopirita, chimenea, ciudades, circular, nuboso, revestimiento, coloidal, columna, confort. Como vemos, las señales débiles identificadas son de muchos tipos como diferentes materiales, geografías, aspectos relacionados, nombres propios, etcétera.

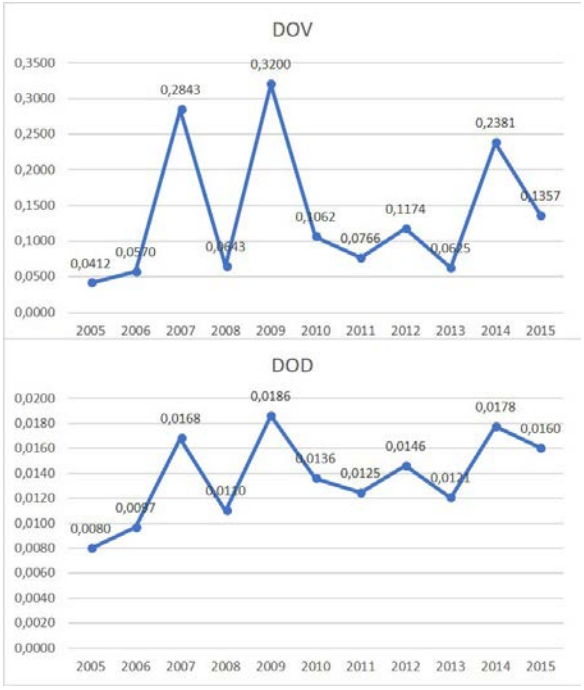
Se ha observado una alta correlación entre los resultados obtenidos mediante nuestra arquitectura en comparación con otros estudios similares (Yoon, 2012). Sin embargo, en este estudio se validan los resultados obtenidos utilizando el mismo set de documentos que ha sido utilizado en los repositorios de entrada, por lo que básicamente están confirmando sus hipótesis con los mismos documentos con los que se ha realizado el estudio, y no con nuevos.

La manera convencional para comprobar si un sistema de clasificación funciona de una manera correcta, es dividir el conjunto de datos disponible en tres más sencillos que realicen el entrenamiento, test y validación. En el caso particular de la detección de señales débiles, hemos podido comprobar que muchas de estas señales pasarían desapercibidas utilizando esta metodología de evaluación de resultados. Las *weak signals* tienen una frecuencia absoluta generalmente muy baja, por lo que una señal débil detectada en un set de test sería raramente vuelta a detectar en otro. Por esta razón, este método de evaluación fue descartado.

El usar fuentes de diferentes tipos provee un mecanismo mejor para comprobar los resultados, por lo que la comprobación de que las señales débiles detectadas no son falsos positivos ha sido a través de comprobar su detección (y difusión) en el estudio independiente de cada uno de los tres tipos de fuentes (ScienceDirect, New York Times y Twitter), y en el resultado global sin discriminar el tipo de fuente.

Otra última manera para evaluar los resultados obtenidos es sabiendo que, desde el punto de vista de un estudio cuantitativo, las señales débiles tienen un patrón anormal en su comportamiento. Aunque son palabras con frecuencia de aparición baja, suelen presentar un alto rango de fluctuación en la ratio de incremento de frecuencia. En la Figura 10 se puede observar el comportamiento fluctuante de los patrones DoV y DoD de la palabra África. Un método de comprobación implementado fue comprobar esta clase de comportamientos en las señales detectadas por el sistema.

Figura 10. Gráficas DoV y DoD de la palabra "África"



5. CONCLUSIONES Y TRABAJO FUTURO

En este artículo, se ha descrito la implementación de un sistema para detectar señales débiles de futuro bajo una aproximación cuantitativa, con un análisis sobre el sector de los paneles solares. En contraste con otros trabajos actuales que se basan en informaciones estructuradas, con análisis cualitativos, o que simplemente se basan en datos de una única fuente, la detección de estas señales se ha basado en documentos heterogéneos de varios tipos.

El método propuesto puede encontrar señales débiles de una manera más eficiente que los humanos expertos en la materia bajo estudio, en el proceso de análisis de una inmensidad de documentos relacionados con esa temática.

Podemos concluir que se han cumplido los objetivos planteados al haber sido demostrado que es posible construir un sistema eficiente de detección de señales débiles, comenzando desde la confección de repositorios de documentos de entrada en lenguaje natural, que siga el modelo semiótico de Hiltunen y que procese los datos de forma cuantitativa.

Por lo tanto, el sistema está plenamente preparado para obtener resultados que pueden ayudar a expertos en negocios en el reconocimiento de nuevos factores clave de sus mercados y en el desarrollo de nuevas oportunidades.

Una posible mejora para evitar la detección de falsos positivos es considerar expresiones de más de una palabra, lo que mejoraría sensiblemente el análisis semántico y daría resultados más interesantes de interpretar. También en el futuro, el sistema deberá testearse con documentos de entrada relacionados con otras temáticas distintas a la de los paneles solares, área empleada en este estudio. A priori, el sistema está preparado para trabajar con cualquier set de documentos almacenados en los repositorios de entrada.

Por otro lado, aunque en cada fase del proyecto se ha optimizado el código y la arquitectura empleados en el sistema, ejecutándose multitud de procesos en paralelo, se seguirá mejorando la eficiencia del sistema mediante mejoras en su arquitectura hardware mediante la ejecución en paralelo que permite las arquitecturas CUDA y la optimización correspondiente en su código. CUDA son las siglas de Compute Unified Device Architecture (Arquitectura Unificada de Dispositivos de Cómputo) y hace referencia a una plataforma de computación en paralelo creada por nVidia para codificar algoritmos en unidades de procesamiento gráfico (GPU). La eficiencia del modelo implementado se puede mejorar aprovechando el gran paralelismo y el alto ancho de banda de la memoria en las GPU.

6. AGRADECIMIENTOS

Este trabajo está parcialmente apoyado por EIT Climate KIC de la Unión Europea (proyecto Accelerator - TC2018B_2.2.5-ACCUPV_P066-1A).

Esta investigación forma parte del programa de Doctorado del Departamento de Ingeniería Electrónica de la Universitat Politècnica de València.

ACKNOWLEDGMENTS

This work is partially supported by EIT Climate KIC of the European Union (project Accelerator - TC2018B_2.2.5-ACCUPV_P066-1A).

This research is also part of the PhD programme of the Departamento de Ingeniería Electrónica of the Universitat Politècnica de València.

7. REFERENCIAS

Ansoff, H.I. (1975). Managing Strategic Surprise by Response to Weak Signals. *California Management Review*, 18 (2), 21-33. https://doi.org/10.2307/41164635

Ansoff, H. I.; McDonnell, E. J. (1990). *Implanting strategic management.* Cambridge: Prentice Hall.

Beautiful Soup (2018). Beautiful Soup Documentation. Disponible en: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ [Fecha de consulta: 1/09/2018].

Cembrero, I. (2011). El gigantesco proyecto solar del Sáhara abastecerá a España en 2015. *El País*. Disponible en: http://elpais.com/diario/2011/11/09/sociedad/1320793203_850215.html [Fecha de consulta: 1/09/2018].

Conesa-Caralt, J; Curto-Diaz, J. (2010). *Introducción al Business Intelligence.* Barcelona: Editorial UOC.

Connolly, T.; Begg, C. (2005). *Database Systems: A Practical Approach to Design, Implementation, and Management* (4th ed.). London: Addison-Wesley.

Cooke, R. (2015). África podría convertirse en la nueva esperanza para la producción de energía solar. Vice News. Disponible en: https://news-old-origin.vice.com/es/article/africa-convertirse-nueva-esperanza-produccion-energia-solar [Fecha de consulta: 1/09/2018].

Cooper, A.; Voigt, C.; Unterfrauner, E.; Kravcik, M.; Pawlowski, J.; Pirkkalainen, H. (2011). *TELMAP. Report on Weak Signals Collection.* Bolton: European Commission Seventh Framework Project (IST-257822).

Dator, J. (2005). Universities without quality and quality without universities. *On the Horizon*, 13 (4), 199-215. https://doi.org/10.1108/10748120510627321

Dedić, N.; Stanier C. (2017). Measuring the Success of Changes to Existing Business Intelligence Solutions to Improve Business Intelligence Reporting. *Journal of Management Analytics*, 4 (2), 130-144. https://doi.org/10.1080/23270012.2017.1299048

Eisenhardt K.M.; Brown S.L. (1999). Patching: restitching business portfolios in dynamic markets. *Harvard Business Review,* 77 (3), 72–82.

Elcacho, J. (2014). Megaproyecto para llevar energía solar desde el Sáhara hasta Europa. *La Vanguardia*. Disponible en: http://www.lavanguardia.com/natural/20141022/54417391167/megaproyecto-tunur-energia-solar-electricidad-sahara-europa.html [Fecha de consulta: 1/09/2018].

Elmasri, R.; Navathe, S.B. (2011). *Fundamentals of Database Systems (6th ed.).* Atlanta: Addison-Wesley.

Finger, L.; Dutta, S. (2014). *Ask, Measure, Learn: Using Social Media Analytics to Understand and Influence Customer Behavior.* Sebastopol: O'Reilly Media.

Fischler, M. A.; Firschein, O. (1987). *Intelligence: The Eye, The Brain and The Computer.* Menlo Park: Addison-Wesley.

Giovinazzo, W. (2000). *Object-Oriented Data Warehouse Design: Building a Star Schema.* Santa Ana: Prentice-Hall.

Godet, M. (1994). *From Anticipation to Action, A Handbook of Strategic Prospective.* Paris: UNESCO Publishing.

Griol, D.; Patricio, M.A.; Molina, J.M. (2016). CALIMACO: desarrollo de un servicio de bibliotecario virtual para la interacción multimodal con dispositivos móviles. *Revista Española de Documentación Científica*, 39 (2), e129. http://dx.doi.org/10.3989/redc.2016.2.1262

Han, J.; Kamber M; Pei, J. (2001). *Data Mining: Concepts and Techniques*. Waltham: Morgan Kaufmann Publishers.

Helsingin Sanomat. (2010). Hennes & Mauritz comenzará a comercializar ropa usada bajo la etiqueta de Vintage. *Helsingin Sanomat.*

Hernández, J.; Ramírez, M.J.; Ferri, C. (2004). *Introducción a la minería de datos.* Valencia: Pearson.

Hiltunen, E. (2008). The future sign and its three dimensions. *Futures*, 40 (3), 247–260. https://doi.org/10.1016/j.futures.2007.08.021

Ilmola, L.; Kuusi, O. (2006). Filters of weak signals hinder foresight: Monitoring weak signals efficiently in corporate decision-making. *Futures*, 38 (8), 908-924. https://doi.org/10.1016/j.futures.2005.12.019

Inmon, W.H. (2005). *Building the Data Warehouse (4th ed.).* Indianapolis: John Wiley.

Jung, K. (2010). *A study of foresight method based on text mining and complexity network analysis.* Seoul: KISTEP.

Khan, R.A. (2012). KDD for Business Intelligence. *Journal of Knowledge Management Practice*, 13 (2).

Kimball, R.; Ross, M. (2002). *The Data Warehouse Toolkit: the complete guide to dimensional modelling.* Indianapolis: John Wiley.

Koivisto, R.; Kulmala, I.; Gotcheva, N. (2016). Weak signals and damage scenarios. Systematics to identify weak signals and their sources related to mass transport attacks. *Finland Technological Forecasting and Social Change* 104, 180-190. https://doi.org/10.1016/j.techfore.2015.12.010

Kuusi, O.; Hiltunen, E. (2007). *The Signification Process of the Future Sign.* Turku: Finland Futures Research Centre ebook 4/2007.

Mannermaa M. (1999). *Tulevaisuuden hallinta skenaariot strategiatyoskentelyssa. (Managing the future, Scenarios in strategy work)*. Provoo: WSOY.

MohamadiBaghmolaei, R.; Mozafari, N.; Hamzeh, A. (2017). Continuous states latency aware influence maximization in social networks. *AI Communications*, 30 (2), 99-116. https://doi.org/10.3233/AIC-170720

Molitor, G.T. (2003). Molitor Forecasting Model: Key Dimensions for Plotting the Patterns of Change. *Journal of Future Studies*, 8 (1), 61-72.

MongoDB (2018). Documentación de MongoDB. Disponible en: https://www.mongodb.com/es [Fecha de consulta: 1/09/2018].

Natural Language Toolkit (2018). NLTK 3.3 Documentation. Disponible en: https://www.nltk.org/ [Fecha de consulta: 1/09/2018].

New York Times (2018). *New York Times*. Disponible en: http://www.nytimes.com [Fecha de consulta: 1/09/2018].

Nikander, I.O. (2002). Early Warnings, *A Phenomenon in Project Management,* Dissertation for the degree of Doctor of Science in Technology. Helsinki: Helsinki University of Technology.

Peirce, C.S. (1868). Some Consequences of Four Incapacities. *Journal of Speculative Philosophy*, 2 (3), 140-157. https://www.jstor.org/stable/i25665647

Salton, G.; Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. http://dx.doi.org/10.1016/0306-4573(88)90021-0

ScienceDirect (2018). Science Direct. Disponible en: http://www.sciencedirect.com/ [Fecha de consulta: 1/09/2018].

Siegel, E. (2013). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die.* New Jersey: John Wiley.

Twitter (2018a). Twitter. Disponible en: http://www.twitter.com [Fecha de consulta: 1/09/2018].

Twitter (2018b). Twitter API Documentation. Disponible en: https://dev.twitter.com/rest/public [Fecha de consulta: 1/09/2018].

UNESCO World Heritage Centre (2008). List of factors affecting the properties. Disponible en: http://whc.unesco.org/en/factors/ [Fecha de consulta: 1/09/2018].

Witten, I.H.; Frank E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco: The Morgan Kaufmann Series in Data Management Systems.

Yoo, S.H.; Park, H.W.; Kim, K.H. (2009). A study on exploring weak signals of technology innovation using informetrics. *Journal of Technology Innovation*, 17(2), 109-130.

Yoon, J. (2012). Detecting weak signals for long-term business opportunities using text mining of Web news. *Expert Systems with Applications*, 39 (16), 12543-12550. http://dx.doi.org/10.1016/j.eswa.2012.04.059