¿De qué se habla cuando se habla de homeopatía en la web?: un análisis de contenido

What do we mean when we speak about homeopathy on the web?: a content analysis

María-José Baños-Moreno (1), Eduardo R. Felipe (2), Zuriñe Piña-Landaburu (3), Mauricio Almeida (4)

(1) Facultad de Comunicación y Documentación, Universidad de Murcia, Campus de Espinardo, s/n 30100 Murcia, mbm4963@um.es (2) Instituto de Engenharia de Sistemas e Tecnologia da Informação, Universidade Federal de Itajubá, Rua Irmã Ivone Drumond Distrito Industrial II, 35903087 - Itabira, MG – Brasil, eduardo.felipe@unifei.edu.br (3) Facultad de Ciencias de la Documentación, Universidad Complutense de Madrid, C. de la Santísima Trinidad, 28010 Madrid, zuri-ne.p.l@gmail.com (4) Escola de Ciência da Informação, Universidade Federal de Minas Gerais (UFMG), Av. Antônio Carlos, 6627, Pampulha, 31270-010 - Belo Horizonte, MG – Brasil, mba@eci.ufmg.br

Resumen

Se realiza un análisis del contenido de las páginas web que tratan sobre la homeopatía con el objetivo de determinar su tendencia, tipología, dominios más destacados y términos más utilizados. Para ello, se han identificado los términos más representativos del ámbito y se han seleccionado un grupo de semillas. Ambos elementos son el punto con que arranca "Crawler by domain", una aplicación desarrollada para la recolección de páginas del dominio. Los resultados muestran que la inmensa mayoría de páginas del sector tienen una visión positiva de la homeopatía. Algo que es lógico teniendo en cuenta que gran parte tiene como fin la venta de productos y/o servicios de homeopatía o se tratan de portales especializados en ésta. Como conclusión general, la tendencia de las fuentes a ofrecer un contenido con sesgo positivo y fácilmente comprensible por el usuario medio, junto con la relativa escasez de páginas con un sentido crítico o incluso sin tendencia puede ser un factor que impulse a los usuarios a inclinarse por la utilización de esta pseudoterapia, ya que es posible que sea interpretada como prueba efectiva de sus beneficios.

Palabras clave: Homeopatía. Análisis de contenido. Web. Crawler by domain. Crawlers.

Abstract

An analysis of the content of the web pages that deal with homeopathy is carried out. The aim is to determine their trends, typology, most prominent domains, and most used terms. To do this, the most representative terms in the field have been identified and a group of seeds has been selected. Both elements are the starting point for "Crawler by domain", an application developed to collect web pages. The results show that many pages in the sector have a positive view of homeopathy. That is logical considering that a large part has as its purpose the sale of homeopathy products and/or services or they are specialized portals in it. As a general conclusion, the tendency of these sources to offer content with a positive bias and easily understood by the average user, together with the relative scarcity of pages with a critical sense or even without bias, may be a factor that encourages users to lean towards for the use of this pseudotherapy, since it is possible that it is interpreted as effective proof of its benefits.

Keywords: Homeopathy. Web. Content analysis. Crawler by domain. Crawlers.

1. Introducción

Desde hace un tiempo, se da la paradoja de que se genera más información que nunca y, al mismo tiempo, crecen las dificultades para acceder a fuentes de información fiables. Prueba de ello es el auge de las redes sociales como puerta de acceso de información o la popularización de los llamados contenidos virales que, a menudo, pueden dar acceso a contenidos poco fiables, manipulados o directamente falsos. En muchas ocasiones, el usuario no atiende a la calidad de sus fuentes. Esto es clave para el éxito de las llamadas noticias falsas o "fake news", que son un elemento fundamental para influir en la opinión pública e incluso explican el triunfo de ciertos movimientos políticos (Alcott y Gentzklow, 2017).

La información poco contrastada no es un fenómeno nuevo. Es parte de la cultura popular, siendo el origen de historias basadas en la literatura oral. Desde el chisme o la historia que antes corría de boca en boca, hasta los actuales contenidos "virales" que llegan a millones de personas en un breve periodo de tiempo, estamos ante algo que, simplemente, ha aprovechado las ventajas de las Tecnologías de la Información para expandirse. Como un organismo vivo, la información poco contrastada ha evolucionado adaptándose al entorno.

Se trata de un fenómeno que, en muchas ocasiones, no tiene mayor relevancia al margen de su interés sociológico o cultural. Sin embargo, esto cambia cuando comienzan a extenderse ideas y/o prácticas que pueden resultar de riesgo tanto para uno mismo como para terceros o, sin llegar a ese punto, cuando el ciudadano tiene acceso a información que, al no tener una base científica real, puede llevarlo a abandonar terapias con evidencia médica en favor de otras basadas en pseudociencias. Uno de los ejemplos más claros es el auge de las medicinas alternativas, como la homeopatía, del griego homoios (igual) y pathos (sufrimiento), en la que se centra este trabajo como objeto de estudio.

La homeopatía surge en 1796, de manos del doctor alemán Samuel Hahnemann, basándose en la idea de que, si una sustancia causa en personas sanas los síntomas de una enfermedad, esa sustancia curaría estos síntomas en personas enfermas (Nayernouri, 2017). Es lo que se conoce como la Ley de los similares. El otro principio en el que se asienta esta pseudociencia es la Ley de los infinitesimales, según la cual, un remedio se vuelve más efectivo al diluirse (Young, 2014). La práctica de la homeopatía se extendió a lo largo del siglo XIX en Europa y cayó en desuso a principios del siglo XX. Hasta que, en los años 70, volvió de forma significativa (Nayernouri, 2017) como terapia alternativa, casi cien años después.

La creciente concienciación del consumidor y la presión ejercida por los sectores críticos con las pseudociencias han hecho que la Real Academia de la Lengua Española haya enmendado la acepción de este concepto, que ha pasado de ser definido como un sistema curativo a una mera práctica que "consiste en administrar a alguien, en dosis mínimas, las mismas sustancias que, en mayores cantidades, producirían supuestamente en la persona sana síntomas iguales o parecidos a los que se trata de combatir" (Redacción Médica, 2019). No es la única institución, el Ministerio de Sanidad continúa revisando la categorización de otras pseudoterapias por carecer de evidencia científica, tales como la magnetoterapia estática o la dieta macrobiótica (Europa Press, 2021).

En esta investigación se combina el uso de técnicas propias de las Ciencias de la Documentación, como la indización y el análisis de contenido, y de la Informática, mediante el diseño e implementación de un rastreador que analiza el contenido de páginas web relacionadas con la homeopatía.

El crawler es un tipo de agente computacional (Chowdhury, 2004) que se comporta como un usuario (buscador) y visita y descarga el contenido de páginas web para su procesamiento posterior (Chakrabarti, 2003). Este tipo de herramienta ya ha sido empleada antes para la recuperación de información filtrada para el análisis o la creación de bases de conocimiento específicas. Además, existen diversas técnicas para la realización de estos programas, como Chakrabarti et. al (1999), que empleó técnicas de aprendizaje automático y taxonomías. Por su parte, Safran et. al (2012) hizo uso de términos próximos a las URLs relevantes, marcadas inicialmente como semillas, para construir un conjunto de entrenamiento para el aprendizaje automático. Hegade et. al (2021) trabaja con un modelo de inferencia contextual, basado en similitudes y reglas de inferencia.

Por otro lado, De Groc (2011) propone un modelo de categorización de páginas web a partir de un léxico basado en filtros temáticos. Aggarwal (2019) hace uso del aprendizaje automático basado en un conjunto de premisas que pueden incluir URLs, páginas padre, textos de anclaje y textos próximos a los elementos relevantes. Dahiwale (2014) adopta un modelo de análisis de contenido basado en determinadas etiquetas HTML, con el fin de computar la relevancia de la página. Bedi et al. (2012) trabaja con un mecanismo multihilo asociado a las ontologías de dominio. Farag et al. (2018) estudia el concepto de eventos para filtrar datos específicos. Du et al. (2015) elabora un camino de similitud a partir da modelo vectorial.

2. Justificación y objetivos

El objetivo general de este trabajo es realizar un análisis de contenido de las páginas web relacionadas con la homeopatía con el propósito de conocer de qué tratan las páginas web relacionadas con la homeopatía y su intencionalidad. De manera más específica, se pretende realizar el seguimiento de páginas y recuperación de información que permita determinar qué términos son empleados para representar los conceptos que giran en torno a esta pseudociencia.

Para satisfacer los aspectos indicados, se ha desarrollado un rastreador, "Crawler by domain", que emplea una estrategia centrada exclusivamente en una lista de términos del área de la homeopatía, configurada por documentalistas. En cada visita del crawler, consulta un vocabulario controlado con el fin de identificar uno o más términos que determinen la posible relevancia del contenido en relación con la temática analizada.

3. Metodología

La investigación arrancó con un análisis bibliográfico y documental con el propósito de obtener una primera lista de términos asociados con la temática. Se recurrió a bibliografía especializada en homeopatía disponible en Web of Science

(WoS) y Google Scholar. También se empleó la taxonomía DeCS, como instrumento terminológico validado. Con ello, se obtuvo una lista de las palabras clave más representativas del resumen de cada publicación, que dieron lugar a un vocabulario controlado. Este producto es empleado en dos ocasiones en este trabajo.

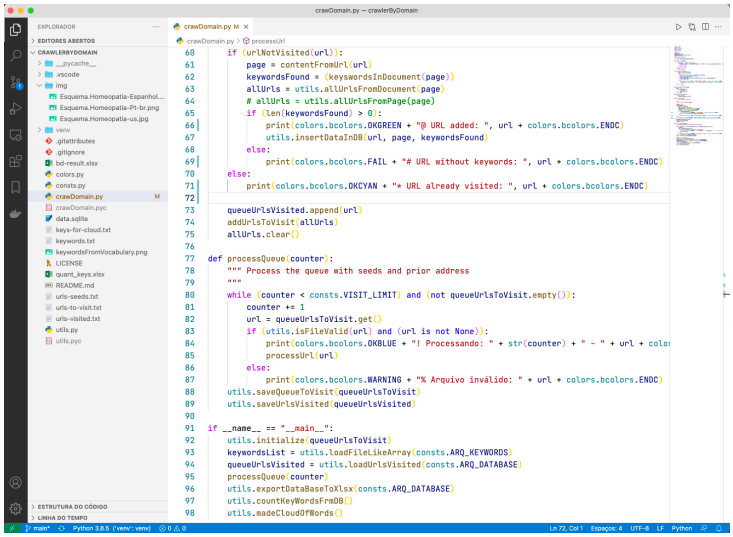

Figura 1. Código fuente en que se muestra uno de los principales procesos del programa

Así, fue empleado en diversas búsquedas en Google, con el propósito de obtener las páginas-semilla que se utilizarán después en el Crawler. Se seleccionaron los 10 primeros resultados de cada búsqueda realizada, descartando aquellas URLs que no estaban relacionadas con el tema de estudio o no resultaban útiles para el propósito del Crawler. Por ejemplo, diccionarios en línea no especializados (The Free Dictionary o WordReference), así como repositorios o bases de datos especializadas en información científica, como Elsevier. Además, varias páginas tenían un origen en común y fueron contabilizadas como un resultado único. Estos términos de búsqueda son empleados después en el crawler desarrollado.

Simultáneamente, se desarrolló "Crawler by domain", con el fin de permitir filtros específicos para la captación de páginas web de la temática escogida. Un esbozo del funcionamiento de este software se muestra en la Figura 1, donde se puede visualizar un proceso importarte del programa para su funcionamiento. Esta herramienta permite utilizar las palabras clave detectadas inicialmente como representativas del tema "homeopatía" y las páginas-semilla obtenidas para recolectar otras páginas web del dominio analizado.

Finalmente, se analizó la información obtenida por el software por documentalistas, excluyendo miles de materias diversas que no estaban relacionadas con la temática propuesta. Así, se revisaron las palabras clave del total de páginas web captadas con el crawler, el tipo de página web y algunas de sus características, clasificándolas por su intencionalidad (divulgación, formativa, venta, etc.).

3.1. Crawler by domain

Los rastreadores son un software especializado que permite la recuperación automática de páginas web empleando un mecanismo de filtro para seleccionar sólo aquellas páginas web del dominio analizado. En otras palabras, a partir de un sitio y sus hipervínculos, se genera una lista de visitas para los demás sitios, hasta que un criterio de parada es satisfecho (Yu, 2011).

La herramienta "Crawler by domain" (https://github.com/erfelipe/crawlerByDomain), es sucesora de MetadadosHTML, un programa para la extracción de metadatos de páginas web (Baños-Moreno et al., 2017). Se trata de una nueva herramienta, escrita en Python (versión 3), orientada a la recolección del código fuente de páginas web especializadas en una temática para, entre otros, extraer datos, como el título o las palabras clave. Entre las diferencias que encontramos con su antecesor, destaca su comportamiento: visitar páginas web a partir de una lista de URL iniciales (semillas), filtrando la recolección de páginas a partir de un vocabulario controlado propio del tema de estudio, eliminado así horas de evaluación humana. Además, el programa es capaz de mostrar al usuario toda la información recolectada y facilita unos resultados más pertinentes. En la Figura 2 se resume su funcionamiento.

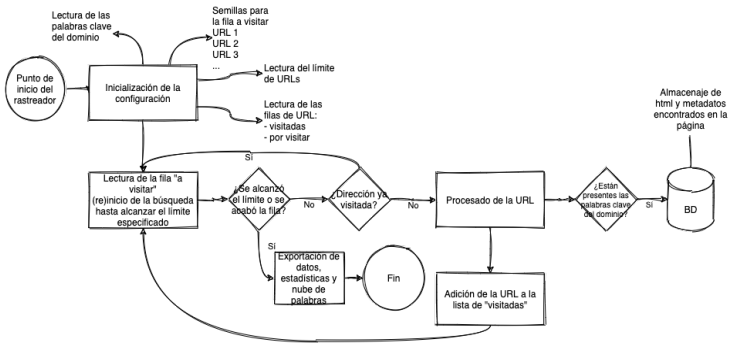

Figura 2. Funcionamiento de la herramienta

Aunque del software se haya incidido particularmente en el modelo de crawler, su parte analítica es similar a un programa orientado al web scraping. Esta técnica permite extraer datos mediante algoritmos específicos con el fin de identificar información relevante para el contexto que se estu-

dia. Este proceso hace posible que, mediante análisis computacional, se pueda analizar grandes volúmenes de datos de forma estandarizada en escenarios donde la evaluación humana se demoraría durante mucho tiempo y, además, podría conducir a errores de interpretación o de análisis.

En el programa, se emplearon diversas bibliotecas de software (documentadas en el github del proyecto), destacando: 1) *requests* para solicitar el contenido de la URL y permitir el análisis de código fuente local; 2) *urllib* para verificar el código fuente de las páginas, con el propósito de alimentar la fila de visitas con las urls de la página visitada; 3) *bs4 (beautifulsoup)* es una biblioteca multifuncional para el parseo y análisis de código html. En todo momento, se procuró escribir el código siguiendo las buenas prácticas del "Clean Code" (Martin, 2009).

En cuanto a sus componentes, destacan los siguientes ficheros:

1. *consts.py:* permite indicar dónde se ubican los archivos de las URLs de las semillas iniciales, la lista de palabras clave y la ruta en la que se deben guardar los ficheros resultantes. Por defecto, todos los archivos estarán en el mismo directorio del programa. También permite definir el límite de visitas para el procesamiento del crawler (en la constante VISIT_LIMIT), que por defecto será de 1.000 páginas recuperadas y el número de grabación parcial (en la constante RECORD_EACH), que permite detener la recolección en un momento dado y retomarla posteriormente.
2. *urls-seeds.txt:* en este fichero se especifican las semillas. El proceso recolecta las direcciones a las que apuntan estas URLs. El proceso se repite con cada nueva colección de páginas, que entran en una cola de espera para su análisis. Así hasta que se detenga de forma activa o se alcance el límite especificado.
3. *keywords.txt:* es la lista de palabras clave que enriquece la recolección. Este fichero, si bien no es obligatorio, es muy recomendable. Su inclusión en la actividad del crawler evita analizar páginas descontextualizadas, que no tienen nada que ver con el tema de estudio. El rastreador busca esos términos en el contenido de cada página web. Si no encuentra ninguno, considera que es ruido y lo rechaza. Para asegurar un buen funcionamiento del proceso, las palabras clave, en el caso de que utilicemos terminología en castellano y/o en portugués, deben introducirse con y sin tildes.
4. *black-list-domain.txt:* es una lista editable de términos que harán que una URL sea desechada por incluir alguno de ellos. Por ejemplo, se puede eliminar todas las referencias a *youtube, facebook, pinterest, twitter, linkedin, github*, entre otras.

A continuación, se describe cómo se configura y funciona la aplicación.

3.1.1. Configuración y puesta en marcha de Crawler by domain

Antes de iniciar su primera iteración, el usuario debe especificar las semillas (en el archivo url-seeds.txt) y el vocabulario especializado en la temática que se estudia (en keywords.txt). También se puede definir el número de visitas que realizar en la constante VISIT_LIMIT del fichero consts.py.

Una vez instaladas las bibliotecas por el administrador PIP y configurados los archivos anteriores, se puede lanzar el Crawler. Para ello, basta con escribir en la línea de comando: *Python crawDomain.py*. Éste puede funcionar durante el tiempo que consideremos, pero, como mínimo, para que recolecte una cantidad relevante de páginas, tendría que alcanzar las 3.000 URLs de contexto. El inicio del procesamiento permite que se generen una base de datos relacional sqlite y ficheros relacionados.

Si es un proceso nuevo, se partirá de las semillas definidas por el usuario. Si se detiene el procesamiento de páginas, después se puede continuar con la recolección respondiendo "N" a la pregunta de construir un nuevo proceso. Las visitas continúan a partir de la última URL tratada, dando continuidad a iteraciones pasadas y permitiendo agregar más páginas al conjunto. Podemos hacerlo en momentos diferentes, con el propósito de incrementar el número de páginas durante un periodo.

Para cada página web visitada se verifica si posee al menos un término del vocabulario controlado entre las palabras clave (etiqueta <meta keywords>), en el cuerpo (<body>) o en el título (<title>) del código fuente. En caso afirmativo, se procesa la web, extrayendo sus metadatos y almacenándolos en el banco de datos junto con su contenido. En caso contrario, la página es descartada, aunque la URL es almacenada de cara a futuros accesos. La iteración continúa hasta que la fila que se procesa quede vacía (aunque es improbable que esto ocurra), o bien se alcance el número de visitas anteriormente definido como límite.

Al acabar, el programa consulta la base de datos generada y exporta su contenido en un documento con formato de hoja de cálculo, que facilita el análisis de los datos obtenidos, que incluye las páginas filtradas y diversos metadatos (se indican en la Tabla I).

Metadato	*Descripción*
domain	Dominio base de la dirección analizada
url	Dirección completa analizada
keys	Palabras clave del vocabulario recuperadas de la página analizada
quantKeys	Cantidad de palabras clave encontradas con respecto al vocabulario controlado
keysFromPage	Palabras clave de la página, declaradas en “meta keywords”
langFromPage	Idioma de la página, declarado en “meta lang” o “language”
descFromPage	Descripción de la página, declarado en “meta description”
authorFromPage	Autor de la página, declarado en tag “meta author”

Tabla I. Metadatos proporcionados por "Crawler by domain"

ID	*Referencia bibliográfica*
1	Bornhöfta, Gudrun; Wolf, Ursula; Ammon, Klaus von; et al. (2006). Effectiveness, Safety and Cost-Effectiveness of Homeopathy in General Practice – Summarized Health Technology Assessment. // Research in Complementary Medicine. 13 (2), 19-29. pp. https://www.ncbi.nlm.nih.gov/pubmed/16883077
2	Cano-Orón, Lorena; Mendoza-Poudereux, Isabel; Moreno-Castro, Carolina (2018). Perfil sociodemográfico del usuario de la homeopatía en España. // Atención Primaria. 1615. https://doi.org/10.1016/j.aprim.2018.07.006
3	Descriptores en Ciencias de la Salud: DeCS. ed. 2018. São Paulo: BIREME / OPS / OMS [actualizado en abril de 2018]. http://decs.bvsalud.org/E/homepagee.htm
4	Ernst, Edzard (2018). Homeopathy: what does the “best” evidence tell us? // Systematic review. 192:8, 458-460. https://www.ncbi.nlm.nih.gov/pubmed/20402610
5	López Espinosa, José Antonio (1999). Notas para la historia de la homeopatía. Revista Cubana de Medicina General Integral, 1999. 15:5, 587-590. http://www.bvs.sld.cu/revistas/mgi/vol15_5_99/mgi17599.htm
6	Nayernouri, Touraj (2017). Homeopathy, Ritual and Magic. Archives of Iranian Medicine. 20:11, 718-722. http://www.aimjournal.ir/PDF/aim-1701
7	Ochoa Ortega, Max Ramiro (2018). Análisis sobre la homeopatía como ciencia o pseudociencia. //Archivo Médico Camagüey. 22:3, 381-392. http://ref.scielo.org/gnx7g9
8	Young, Pablo (2014). La farsa de la homeopatía. // Revista Médica de Chile. 142:2. http://dx.doi.org/10.4067/S0034-98872014000200021

Tabla II. Bibliografía analizada para la selección de términos relacionados con la homeopatía

La herramienta también puede proporcionar el número total de palabras clave y una nube de palabras en formato gráfico.

4. Resultados

Se muestran los resultados obtenidos, organizados por bloques para facilitar su interpretación.

4.1. Obtención de la lista de palabras clave

Como se indicaba en la metodología, primero se realizó una revisión bibliográfica relacionada con la homeopatía. Se recoge la selección de publicaciones obtenida tras la búsqueda bibliográfica en WoS y Google Scholar, así como la taxonomía DeCS en la Tabla II. Esta búsqueda fue realizada en enero y marzo de 2019.En segundo lugar, mediante indización humana se extrajo de la bibliografía anterior una lista de palabras clave relacionadas con la homeopatía. Se indican a continuación, estos términos, organizados por categorías:

- Términos genéricos: Medicina alternativa, Pseudociencia, Medicina natural, Terapias alternativas, Terapias complementarias, Práctica médica alternativa, Sistema terapéutico.
- Términos especializados: Nosode.
- Términos de ámbito organizativo: Industria homeopática, Boiron, COFENAT, Homeopathy Research Institute, Observatorio de Terapias Naturales, Swiss Association of Homeopathic, Physicians, SAHOP/SVHA.
- Personas: Samuel Hahnemann (creador); Homeópata (profesional).
- Principios de la homeopatía: Ley de los similares, Ley de los infinitesimales, Agua polimerizada, Memoria del agua, Dilución homeopática, Número de Avogadro, Constante de Avogadro, Moléculas activas.
- Relativos a tratamientos homeopáticos: Prueba homeopática, Repertorios homeopáticos, Homeopathic Materia Medica, Productos homeopáticos, Tratamientos homeopáticos, Medicina homeopática, Remedios homeopáticos.

Estas relaciones son importantes para conocer con mayor profundidad los elementos relacionados con la homeopatía. Así, por ejemplo, podemos ver en qué principios está basada o qué tratamientos pueden establecer sus especialistas. También permiten determinar cuál es su marco general como pseudociencia o terapia alternativa.

4.2. Obtención de las semillas

Se realizaron las siguientes búsquedas en Google, a partir de los términos recogidos en el apartado anterior: “Homeopatía”; “Nosode”; “Samuel Hahnemann”; “Industria homeopática y Boiron”; “Industria homeopática y COFENAT”; “Industria

homeopática y Homeopathy Research Institute"; "Industria homeopática y Observatorio de Terapias Naturales"; "Industria homeopática y Swiss Association of Homeopathic Physicians"; "Industria homeopática y SAHOP/SVHA"; "Miasmas"; "Ley de los similares"; "Ley de los infinitesimales"; "Agua polimerizada"; "Memoria del agua"; "Dilución homeopática"; "Número de Avogadro"; "Constante de Avogadro"; "Prueba homeopática"; "Repertorios homeopáticos"; "Homeopathic Materia Medica"; "Productos homeopáticos"; "Tratamientos homeopáticos"; "Medicina homeopática"; "Remedios homeopáticos"; "Homeópata".

Estás búsquedas fueron lanzadas entre junio y julio de 2020. Los términos genéricos, como "Medicina alternativa", fueron excluidos de las búsquedas por ser menos específicos que "homeopatía". Tras el análisis de los datos, se consiguieron 12 semillas distintas relacionadas con el tema de estudio (250 en bruto). Como requisito se impuso que el dominio de las URLs debía aparecer como mínimo en tres consultas diferentes (ver la Tabla III).

ID	*Página semilla y búsqueda en que aparece*
001	https://cuidateplus.marca.com (01, 18, 19, 21, 22, 23)
002	http://semh.org (01, 22, 23, 25)
003	https://www.boiron.es (01, 04, 15, 21, 22, 24, 25)
004	http://queeslahomeopatia.com/ (01, 18, 21, 22, 23, 24, 03, 15)
005	https://elpais.com/ (01, 18, 23, 04, 09, 05, 07, 18, 01, 09)
007	http://www.homeopatia.net/ (01, 15, 18, 22, 24, 25)
016	https://es.wikipedia.org (03, 04, 10, 13, 14, 15, 16, 17, 18, 19, 20, 23, 25)
028	https://www.redaccionmedica.com/ (04, 05, 18, 21)
032	https://www.elmundo.es/ (04, 09, 18)
033	https://www.eldiario.es/ (04, 05, 09)
148	https://www.portalfarma.com/ (21, 22, 23, 24)
159	https://www.hablandodehomeopatia.com/ (22, 23, 25)

Tabla III. Semillas y búsquedas en que aparecen

Se desecharon sitios de venta de productos con escasa repercusión, blogs basados en la opinión particular de personas, etc. Un caso excepcional es Boiron (https://www.boiron.es), un laboratorio dedicado exclusivamente a la venta de productos homeopáticos, que apareció en casi un tercio de las búsquedas (7/25) y que indica, sin dudas, un buen posicionamiento en Google, pero también la especificidad de los términos empleados por dicho sitio web.Posteriormente, con el objetivo de incrementar el número de semillas, se utilizó el propio crawler hasta recolectar 238 URLs que, tras su revisión, fueron incorporadas al archivo url-seeds.txt.

4.3. Recolección de páginas sobre homeopatía

Con la lista de palabras clave y semillas obtenidas en los procesos descritos en los apartados 4.1 y 4.2 respectivamente, se configuró la aplicación Crawler by domain para el rastreo de páginas web sobre homeopatía.

El rastreador estuvo funcionando durante varias horas en distintos periodos de tiempo, entre el 19 y el 23 de diciembre de 2021, visitando un total de 141.702 páginas. 5.108 páginas fueron procesadas como válidas para el objeto de estudio, descartando todas aquellas que no tuvieran al menos uno de los términos recogidos en el archivo keywords.txt del crawler en su código fuente (palabras clave, título o cuerpo). El producto resultante fue exportado en formato Excel.

Tras un análisis inicial, se desecharon: páginas sin acceso, de los que no se puede acreditar su contenido; páginas repetidas con distintas casuísticas (URLs recortadas; páginas móviles aceleradas; versiones para impresión, sin certificado de seguridad, etc.); referencias a esquemas, como XML o JSON; agregadores de contenido (*feeds*); información de suscripciones web; avisos de cookies; recuperación de contraseñas; gestión de cuentas de usuario; avisos legales; contacto; políticas de privacidad, carritos de compra o listas de deseos vacías; listas de deseos vacías; búsquedas sin resultados; páginas de bancos de imágenes, etc.

Además, se encontraron dos tipos de falsos positivos: páginas con el término "homeopatía" (o cualquiera de los identificados durante la indización humana) en un enlace, pero no en el propio contenido; páginas con palabras clave polisémicas, como "memoria del agua", que nada tienen que ver con el área en que se centra este trabajo.

Finalmente, la muestra quedó compuesta por 2715 páginas recolectadas unívocas.

4.4. Análisis de los datos obtenidos

4.4.1. Tendencia de la página web

Por tendencia se hace referencia a la consideración que se hace de la homeopatía en la página web. En la Figura 3 se identifican las siguientes tendencias:

1. Positiva: corresponde a páginas en las que se habla bien de la homeopatía, ya sea con intención de vender un producto homeopático, ofrecer un servicio o simplemente convencer

de sus beneficios). Se atribuye esta tendencia a 2238 páginas (82,43%).

2. Negativa: se refiere a aquellas páginas que analizan la homeopatía desde un punto de vista crítico y centrado en la exposición de problemas asociados al lector común. Constituyen el segundo tipo de fuente más abundante (15,69%), aunque muy lejos de la anterior.
3. Sin inclinación: estas páginas se centran prácticamente en un aporte histórico de sus inicios o la descripción de algunos de sus referentes básicos, como qué es la ley de los similares, por ejemplo (1,88%).

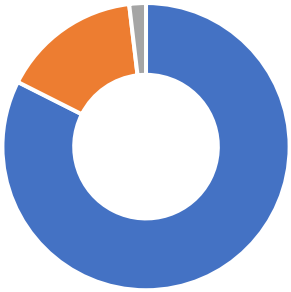

Figura 3. Páginas web según la intencionalidad de sus contenidos

Los datos muestran una clara tendencia de las fuentes a ofrecer una visión de la homeopatía como una terapia con visos de fiabilidad. Teniendo en cuenta que su público objetivo son los usuarios con un conocimiento medio o bajo de medicina, pueden ser un punto de apoyo para el crecimiento que la homeopatía ha vivido en las últimas décadas.

4.4.2. Tipos de página

La tipología de las páginas es diversa, desde aquellas dedicadas a la venta hasta páginas de asociaciones de distintos colectivos.

La Tabla IV agrupa las páginas analizadas por tipos. Los tipos de la tabla anterior fueron identificados *ad hoc* para este estudio.

Destacan muy especialmente los contenidos enfocados a la venta de productos y/o la oferta de servicios propios de la homeopatía. De hecho, constituyen el 50% de las URLs analizadas. En una cantidad bastante inferior (17,5%), también se recolectaron páginas de portales especializados en homeopatía, dedicadas principalmente a informar de sus características, de los beneficios que aportaría, de sus bases, etc.

El tercer tipo de página es el blog (8,18%), generalmente de tipo personal y casi siempre orientado a proporcionar una opinión sobre la homeopatía en un sentido negativo (96%).

También se detectó que un 6,7% del total de páginas se centraban en la formación en homeopatía y pseudoterapias relacionadas, proporcionada por instituciones de todo tipo (universidades, institutos, etc.). Se ha considerado interesante separar este dato de la oferta de servicios, aunque ciertamente, se podría haber considerado un servicio más. Que casi una décima parte de las URLs estudiadas se centren en formar homeópatas o profesionales afines y otro 50% en la venta y servicios da una idea del volumen de negocio de este tipo de producto.

Tipos de páginas	*Frec.*	*%*
Venta de productos y prestación de servicios	1359	50,05
Portal especializado en homeopatía	476	17,53
Blog	222	8,18
Entidad educativa	182	6,70
Prensa generalista, deportes y economía	128	4,71
Asociación de profesionales de homeopatía o terapias alternativas	98	3,61
Publicación o sitio de información científica	54	1,99
Prensa especializada en medicina y salud	52	1,92
Prensa especializada en ciencia, cultura y tecnología	40	1,47
Portal especializado en medicina, farmacia y/o salud	24	0,88
Otro tipo de página	22	0,81
Portal de información pública oficial	16	0,59
Asociación de profesionales de farmacia	15	0,55
Otro tipo de asociación	8	0,29
Asociación de empresas de homeopatía o terapias alternativas	8	0,29
Asociación de profesionales de medicina	6	0,22
Venta de otro tipo de productos	4	0,15
Prensa especializada en terapias alternativas	1	0,04

Tabla IV. Tipología de páginas recolectadas

La siguiente figura muestra los datos de la Tabla IV, agrupando en bloques mayores algunos de los tipos detectados. Así, es interesante observar que algo más de un 8% de las URL recuperadas (221 ítems) corresponden a noticias publicadas en prensa generalista o especializada, formando parte de esa variedad de temas que cubren los medios de comunicación (Baños-Moreno, 2017). Esto indica que el tema objeto de estudio constituye una fuente de interés para la sociedad. Casi

un 79% de las noticias tenían una tendencia negativa, desmitificando los beneficios de la homeopatía. Sería de interés para trabajos futuros profundizar en la información que ofrecen los medios generalistas, pero también los especializados en salud y medicina.

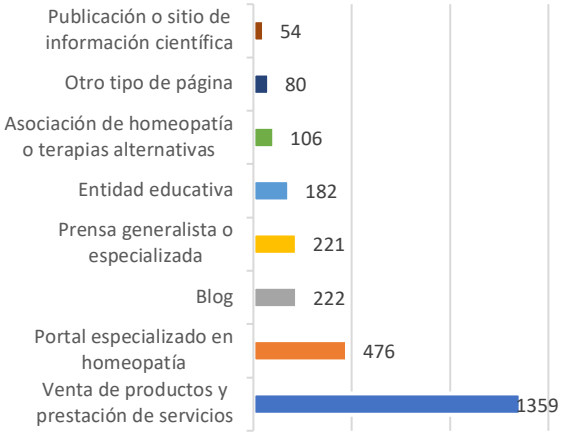

Figura 4. Tipología de páginas recolectadas

En cuanto a publicaciones científicas (casi 2%), se han identificado varios trabajos que analizan los efectos de la homeopatía ante diferentes afecciones. Según algunos estudios, el remedio funcionó positivamente (66,6%), para otros tuvo consecuencias negativas (casi 15%) y para el porcentaje restante no hizo nada (18,51%).

4.4.3. Dominios de las páginas recolectadas

Las 2715 páginas recolectadas corresponden a 196 dominios distintos. Esto podría indicar que las páginas sobre homeopatía tienden a referenciar contenidos de su propio dominio.

En la Tabla V se recogen 16 dominios con mayor frecuencia en la muestra (una ocurrencia mínima de 50), que representan más de un 71% del total de páginas recolectadas. La mayor parte de dominios se centran, como era lógico, a la vista de datos de apartados anteriores, en la venta y/o servicios de productos homeopáticos, o bien a la promoción de la homeopatía.

De los 16 dominios, sólo tres son ajenos al área y, además, tienen una tendencia negativa respecto a esta pseudociencia: "La lista de la vergüenza (blog Naukas)" (3,87%), "El País" (2,25%) y "El blog de búho gris" (1,84%). Resulta llamativo que un medio de comunicación generalista y de gran alcance, como es el periódico español "El País", aparezca más veces que la página web de Boiron, la citada empresa que ofrece productos homeopáticos.

Dominio	*Frec.*	*%*
Hablando de homeopatía	289	10,64
Quanta. Farmacia homeopática y boutique de la vida saludable	239	8,80
B JAIN	211	7,77
Clínica Medicina Integrativa	178	6,56
Consultorio médico-homeopático Doctores Eizayaga	162	5,97
Satisfarma	151	5,56
La lista de la vergüenza (blog Naukas)	105	3,87
Farmàcia Coliseum	97	3,57
Elaesi CDMX	89	3,28
Jacomart Farmacia	79	2,91
Homeopatía Pura	67	2,47
El País	61	2,25
Sociedad Española de Medicina Homeopática	55	2,03
Homeopartía suma. Asamblea Nacional de Homeopatía (ANH)	51	1,88
El blog de búho gris	50	1,84
Instituto Argentino de Terapia Neural (Neuralterapia)	50	1,84

Tabla V. Dominios con mayor frecuencia

4.4.4. Palabras clave de las páginas

Se localizaron 6223 palabras clave, de las cuales, "homeopatia" (2320) es la más frecuente, seguida de "homeopatía" (1523).

Palabra clave	*Frec.*	*%*
homeopatia	2320	37,28
homeopatía	1523	24,47
homeópata	884	14,21
medicina homeopática	461	7,41
productos homeopáticos	366	5,88
nosode	353	5,67
remedios homeopáticos	88	1,41
memoria del agua	83	1,33
tratamientos homeopáticos	62	1,00
miasmas	30	0,48
dilución homeopática	20	0,32
ley de los infinitesimales	16	0,26
ley de los similares	8	0,13
agua polimerizada	4	0,06
homeopathic materia medica	3	0,05
constante de avogadro	1	0,02
samue hahnemann	1	0,02

Tabla VI. Palabras clave identificadas

Estos dos términos constituyen casi el 62% del total. De hecho, juntas ("homeopatia, homeopatía") aparecen en 1227 ocasiones. También se emplean otras palabras clave relevantes para el área objeto de estudio, como "homeópata" (14,21%), "medicina homeopática" (7,41%) o "productos homeopáticos" (5,88%).

Se observa que los términos propios de la jerga homeopática, como "Ley de los infinitesimales" o "Miasmas", son menos representativos en las páginas recolectadas. Se trata de un resultado lógico si se tiene en cuenta que la gran mayoría de éstas se orientan a la venta o divulgación, reafirmando una imagen positiva de la homeopatía. Los conceptos más complejos relacionados con esta pseudoterapia están fuera de su ámbito de interés. Se entiende que el usuario medio desconoce las bases teóricas de la homeopatía, por lo que es comprensible que este tipo de conceptos no aparezcan en las páginas analizadas.

Una excepción es el término "Nosode" (5,67%), utilizado fundamentalmente en las páginas categorizadas como publicaciones o sitios de información científica. Es decir, se inserta en un ámbito científico y muy especializado.

Un posible trabajo futuro podría ser el análisis del vocabulario de las páginas tanto de tendencia positiva como negativa, para conocer su posible influencia en la construcción de la opinión de los usuarios.

5. Conclusiones y trabajos futuros

La homeopatía resulta atractiva para un gran número de pacientes (López Espinosa, 1999) que, si bien se sirven de amigos y conocidos como principal elemento de información, acuden a Internet como segunda fuente (Cano-Orón et al., 2018). Ahora bien, ni la abundancia de páginas especializadas y tampoco un uso extendido aseguran que la información sea de calidad ni fiable.

Este trabajo ofrece información acerca de las posibles vías que los usuarios pueden tener para obtener información sobre este tipo de pseudoterapias. Los resultados muestran una tendencia de las fuentes a ofrecer un contenido con sesgo positivo y fácilmente comprensible por el usuario medio. La relativa escasez de páginas con un sentido crítico o incluso sin tendencia puede ser un factor que impulse a los usuarios a inclinarse por la utilización de esta pseudoterapia, ya que es posible que sea interpretada como prueba efectiva de sus beneficios.

Si bien resulta lógico que las páginas web creadas expresamente para apoyar la homeopatía muestren esa tendencia positiva, la capacidad de los medios para influir en la opinión pública hace que su responsabilidad ética y social sea mayor. De hecho, como se ha observado, sus publicaciones tienen casi siempre una postura crítica con la homeopatía. Resulta evidente que los medios generalistas influyen de forma decisiva en la opinión pública; un usuario que quizás no se haya sentido atraído por la homeopatía o la desconozca, es posible que cambie de punto de vista si una fuente a priori fiable le ofrece una visión positiva o neutra de este tipo de productos.

Tanto los términos más frecuentes (muy poco especializados), como las fuentes recolectadas con mayores ocurrencias revelan la predisposición a ofrecer información con la finalidad de influir con un sesgo positivo que de ofrecer contenidos objetivos y basados en la evidencia científica.

Un análisis pormenorizado de los términos empleados en los contenidos de páginas recolectadas podría mejorar el propio vocabulario controlado, enriqueciendo así los procesos de captación de nuevas páginas visitadas. La extracción de palabras clave del dominio más allá de las páginas visitadas también podría contribuir a mejorar esa lista de términos.

Teniendo en cuenta los riesgos de la popularización de las pseudoterapias, es clara la necesidad de una alfabetización informacional en los usuarios para mejorar su capacidad de detectar fuentes sesgadas. Asimismo, sería interesante un "contraataque" informativo por parte de organismos oficiales y medios de comunicación generalistas que, en un lenguaje fácilmente comprensible y con un vocabulario cercano, explicasen la realidad de la homeopatía como pseudoterapia y los beneficios de la medicina basada en la evidencia.

Por otro lado, el comportamiento del rastreador fue el esperado. Las páginas que no pertenecían a la temática de este estudio, filtradas gracias al empleo de un vocabulario controlado, fueron descartadas. Las páginas captadas presentaban, como mínimo, un término de la lista de palabras relevantes para el dominio analizado y fueron almacenadas en un formato que hace posible su análisis, junto con sus metadatos. Todavía se puede hacer evolucionar el algoritmo del software, programando su mecanismo de captación de páginas hacia una modalidad *multiprocessing*, donde varias páginas se pueden visitar "al mismo tiempo", acelerando o proceso de evaluación y captación (Python, 2022).

"Crawler by domain" ha resultado útil para el propósito de recolección de páginas cuyo contenido trata algún aspecto relacionado con la homeopatía. El proceso de ejecución es sencillo, aunque puede mejorarse y la obtención de un documento en formato de hoja de cálculo facilita el análisis

de los datos. Sin embargo, hay dos mejoras obvias que hay que realizar: a) durante el procesamiento, se deben desechar las páginas que son recolectadas cuando el término del vocabulario controlado se localiza en el cuerpo de la página web pero como parte de un enlace ya que, en realidad, no forman parte de su contenido; b) es necesario que, de alguna forma, se comprueben si los enlaces recortados o que hacen referencias a partes de una página web han sido ya incluidos en la lista de URLs recolectadas, pues estaríamos hablando de una misma repetición.

Finalmente, esta aplicación podría utilizarse para analizar otros temas de interés social, como puede ser el SARS-CoV-2. Esta cuestión está teniendo repercusiones mundiales y, desde hace tiempo se observan movimientos negacionistas en contra de las evidencias científicas. En España, diversos equipos, integrados por documentalistas y periodistas, como VerificaRTVE se encargan de analizar los datos y arrojar luz sobre la veracidad de la información aportada. Un estudio basado en el análisis conceptual de fuentes sobre este tema, en combinación con la "Crawler by domain", aportaría nuevos datos sobre la intencionalidad de las fuentes en línea sobre dicho virus y sus consecuencias para el ser humano.

Referencias

Alcott, H.; Gentzklow, M. (2017). Social Media and Fake News in the 2016 Election. // Journal of economic perspectives. 31:2, 211-36. https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211

Baños-Moreno, M.-J. (2017). Propuesta de modelado de una ontología de dominio para la representación de acciones en política-economía. Universidad de Murcia. http://hdl.handle.net/10201/56661

Baños-Moreno, M.J.; Felipe, E. R.; Pastor-Sánchez, J. A; Lima, G.; Martínez-Béjar, R. (2017). Análisis de metadatos de noticias para la extracción de información del código fuente: El software METADADOSHTML. // Information Research. 22:1. http://InformationR.net/ir/22-1/paper740.html

Bedi, P.; Thukral, A.; Banati, H.; Behl, A.; Mendiratta, V. A. (2012). Multi-Threaded Semantic Focused Crawler. // Journal of Computer Science and Technology. 27:6, 1233–1242. https://doi.org/10.1007/s11390-012-1299-8.

Cano-Orón, L.; Mendoza-Poudereux, I.; Moreno-Castro, C. (2018). Perfil sociodemográfico del usuario de la homeopatía en España. // Atención Primaria. 51:8, 499-505. https://doi.org/10.1016/j.aprim.2018.07.006

Chakrabarti, Soumen (2003). Mining the Web: discovering knowledge from hypertext data. San Francisco, CA: Morgan Kaufmann Publishers.

Chowdhury, G. G. (2004). Introduction to modern information retrieval. 2nd ed. London: Facet.

Dahiwale, P.; Raghuwanshi, M. M.; Malik, L. (2014). PDD Crawler: A Focused Web Crawler Using Link and Content Analysis for Relevence Prediction. // Third International Conference on Advanced Information Technologies & Applications, 7 nov. 2014. Comp. Science & Information Technology (CS & IT). 245–253. https://doi.org/10.5121/csit.2014.41123.

Du, Y.; Liu, W.; LV, X.; Peng, G. (2015). An improved focused crawler based on Semantic Similarity Vector Space Model. // Applied Soft Computing. 36, 392–407. https://doi.org/10.1016/j.asoc.2015.07.026.

Europa Press (2021). Sanidad califica como pseudoterapias a la dieta macrobiótica, al masaje tailandés y a la magnetoterapia estática. https://www.infosalus.com/actualidad/noticia-sanidad-califica-pseudoterapias-dieta-macrobiotica-masaje-tailandes-magnetoterapia-estatica-20210219132911.html (19/02/21).

Farag, M.; Lee, S.; Fox, E.A. (2018). Focused crawler for events. // International Journal on Digital Libraries. 19:1, 3–19. https://doi.org/10.1007/s00799-016-0207-1.

Groc, Clement, Babouk de (2011). Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. // 2011 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), ago. 2011. Lyon, France: IEEE, ago. 2011. 497–498. https://doi.org/10.1109/WI-IAT.2011.253. http://ieeexplore.ieee.org/document/6040719/.

Hegade, P.; Lingadhal, N.; Jain, S.; Khan, U.; Vijeth, K. L. (2021). Crawler by Contextual Inference. // SN Computer Science. 2:3, 216. https://doi.org/10.1007/s42979-021-00574-z.

Lee, J. G.; Bae, D.; Kim, S.; et al. (2020). An effective approach to enhancing a focused crawler using Google. // The Journal of Supercomputing. 76:10, 8175–8192. https://doi.org/10.1007/s11227-019-02787-9.

López Espinosa, J.A. Notas para la historia de la homeopatía. (1999). Revista Cubana de Medicina General Integral. 1999. 587-590. http://www.bvs.sld.cu/revistas/mgi/vol15_5_99/mgi17599.htm

Martin, Robert C. (Org.). (2009). Clean code: a handbook of agile software craftsmanship. Upper Saddle River, NJ: Prentice Hall.

Nayernouri, T. (2017). Homeopathy, Ritual and Magic. // Archives of Iranian Medicine. 20:11, 718-722. http://www.aimjournal.ir/PDF/aim-1701

Python (2022). multiprocessing — Process-based parallelism. // Documentation. https://docs.python.org/3/library/multiprocessing.html

Safran, M. S.; Althagafi, A.; Dunren C. Improving Relevance Prediction for Focused Web Crawlers. In: 2012 IEEE/ACIS 11TH International Conference on Computer and Information Science (ICIS), maio 2012. 2012 IEEE/ACIS 11th International Conference on Computer and Information Science [...]. Shanghai: IEEE, maio 2012. p. 161–166. DOI 10.1109/ICIS.2012.61. http://ieeexplore.ieee.org/document/6211091/.

Souza, R. R.; Tudhope, D.; Almeida, M. B. (2012). Towards a Taxonomy of KOS: Dimensions for Classifying Knowledge Organization Systems. // Knowlegde Organization. 39:3, 180. https://www.nomos-elibrary.de/10.5771/0943-7444-2012-3-179.pdf

Redacción Médica. (2019). La homeopatía no cura ya ni según la definición de la RAE. https://www.redaccionmedica.com/virico/noticias/la-homeopatia-no-cura-ya-ni-segun-la-definicion-de-la-rae-2633 (28-12-2019).

Young, P. (2014).La farsa de la homeopatía. // Revista Médica de Chile. 142:2. http://dx.doi.org/10.4067/S0034-98872014000200021

Yu, Liyang (2011). A developer's guide to the semantic web. Berlin: Springer.

Enviado: 2021-03-27. Segunda versión: 2022-05-01.
Aceptado: 2022-05-13.