Nutrición Hospitalaria

Nutr Hosp 2012;27(Supl. 2):59-66
ISSN (Versión papel): 0212-1611
ISSN (Versión electrónica): 1699-5198
CODEN NUHOEQ
S.V.R. 318

Diseño de un marco semántico para la recuperación contextualizada de documentos científicos en el ámbito sanitario

J. M. Santos Gago, L. M. Álvarez Sabucedo, M. J. Fernández Iglesias, R. Míguez Pérez, V. M. Alonso Roris y F. Mikic Fonte

Departamento de Ingeniería Telemática. Universidad de Vigo. Vigo. España.

Resumen

La atención médica personalizada requiere combinar información pública de diversas fuentes con información disponible sobre un paciente o grupo de pacientes. Un problema bien conocido en el ámbito de la gestión de la información es la enorme cantidad de información disponible. Además, las soluciones actuales no aprovechan las ventajas de las últimas aportaciones en el campo del procesamiento semántico. Este problema es especialmente relevante en el ámbito de la salud, ya que sus procesos clave dependen de manera determinante del acceso a información de alta calidad, completa, actualizada y relevante. Esta propuesta tiene como objetivo proporcionar soluciones novedosas para la gestión y recuperación de información en el ámbito de ciencias de la salud para hacer frente a la situación descrita. Para ello, hemos desarrollado un modelo semántico para representar perfiles de salud y caracterizar fuentes de información relevante y así poder completar un repositorio semántico con referencias de contenido y sus propiedades. Además, proponemos las herramientas necesarias para consultar esta Base de Conocimiento a partir de los perfiles semánticos de los pacientes. La solución propuesta, presentada aquí como una prueba de concepto, pretende contribuir al avance de las tecnologías aplicadas a la salud personal y la medicina basada en la evidencia. Las herramientas desarrolladas también pueden ser utilizadas con el fin de hacer uso del conocimiento existente para dar soporte a la revisión sistemática de informes, estudios y análisis relevantes según las condiciones de salud de los pacientes individuales o perfiles de los pacientes.

Nutr Hosp 2012; 27 (Supl. 2):59-66

DOI:10.3305/nh.2012.27.sup2.6275

Palabras clave: *Informática médica. Bases de conocimiento. Sistemas de gestión de información avanzada. Procesado de datos automático. Semántica.*

Correspondencia: Luis Álvarez Sabucedo.
Departamento de Ingeniería Telemática.
Escuela de Ingeniería de Telecomunicación.
Universidad de Vigo.
36310 Vigo, España.
E-mail: lsabucedo@det.uvigo.es

Recibido: 1-VIII-2012.
Aceptado: 3-IX-2012.

DESIGN OF A SEMANTIC FRAMEWORK FOR CONTEXTUALIZED RETRIEVAL OF SCIENTIFIC DOCUMENTS IN THE HEALTH DOMAIN

Abstract

Personalized healthcare requires recombining heterogeneous publicly available data with a patient's or group of patient's profile. A well-known problem in state-of-the-art information management is the overwhelming amount of information available. Besides, state-of-the-art solutions do not take advantage of modern semantic processing to adequately transform data into knowledge. This issue is especially relevant in the health domain, as key processes depend dramatically on the access to high quality, complete, up-to-date, and relevant content (e.g. diagnostics, risk assessment, public health interventions, etc.). This proposal aims to provide novel information management and retrieval solutions in the domain of health sciences to address the situation discussed above. More specifically, we introduce semantic reasoning to retrieve the most relevant knowledge available according to the health profile of a given person. For this, we developed a semantic model to represent health profiles of people and to characterize existing sources of relevant information in order to crawl them to populate a semantic repository with content references and properties. We outline the tools needed to query the knowledge base using the semantic profiles of individuals to get the most relevant content. The proposed solution, discussed here as a proof-of-concept, aims to contribute to the realm of personal health and evidence-based medicine technologies. The tools developed could also be used to take advantage of existing knowledge to facilitate a systematic review of reports, studies and analysis that may be relevant to the health conditions of single patients or patient profiles.

Nutr Hosp 2012; 27 (Supl. 2):59-66

DOI:10.3305/nh.2012.27.sup2.6275

Key words: *Medical informatics. Knowledge bases. Integrated advanced information. Management systems. Automatic data processing. Semantics.*

Introducción

Los artículos y documentos de investigación y experimentación en el ámbito de la salud son un instrumento de trabajo fundamental en la práctica médica diaria. Sin lugar a dudas, constituyen una fuente de información de incalculable valor tanto para los investigadores como para los profesionales de la salud en el desempeño de muchas de sus actividades. La realización de estudios sistemáticos para casos específicos es una tarea tan importante como tediosa al tener que tratar con una cantidad tan amplia de información y de modo tan habitual. En esta línea, este trabajo presenta los fundamentos de un marco tecnológico que tiene por objetivo sentar las bases para proveer a los investigadores y profesionales de la salud de agentes inteligentes de búsqueda y localización de aquellos documentos científicos relevantes que están más relacionados con el perfil sanitario y la patología concreta de un determinado paciente.

Este marco se plantea como una prueba de concepto base para el desarrollo de aplicaciones de búsqueda de información que puedan utilizar como soporte los profesionales de la salud en sus tareas de atención (diagnóstico y tratamiento) a pacientes. El modelo definido se basa en el empleo de técnicas y herramientas propias del campo de las tecnologías semánticas, las cuales facilitan, por un lado, la integración de información con fuentes de datos externas y, por otro lado, su extensibilidad y adaptabilidad al contexto en el que finalmente se vaya a utilizar.

En los siguientes apartados se discuten los elementos principales que conforman la solución propuesta. En primer lugar, con el fin ofrecer al lector una visión general del sistema propuesto, se presenta la arquitectura de alto nivel subyacente, delineando brevemente los módulos principales de la misma y la función que desempeñan. Algunos de los artefactos y mecanismos fundamentales utilizados en estos módulos se describen con mayor detalle en secciones siguientes. En particular, se describe el modelo semántico utilizado para representar los conceptos y relaciones manejados por el sistema. Se describen, también, las características fundamentales del motor de enriquecimiento, elemento encargado de rastrear la web para complementar la información existente en la Base de Conocimiento, así como los principios en los que se sustenta el indizador de documentos. Finalmente, se presentan las conclusiones extraídas del diseño del marco junto con una discusión de las posibles líneas de trabajo futuras.

Arquitectura del Marco Semántico

El marco definido se basa en el empleo de mecanismos propios del campo de las tecnologías semánticas y, más concretamente, en el empleo de las tecnologías y herramientas para la representación y manipulación del conocimiento que han surgido bajo el paraguas de la Web Semántica. La figura 1 muestra los elementos funcionales de la arquitectura técnica propuesta. Es una arquitectura que conforma un Sistema Basado en Conocimiento[1] escalable y adaptable, basada en modelos arquitectónicos propuestos anteriormente por los autores en el campo de las aplicaciones y la inferencia semántica[2,3]. Hace uso de ontologías específicas de dominio y de reglas de inferencia declaradas por expertos que pueden ser refinadas sin cambios estructurales en la infraestructura. Los módulos más importantes de la arquitectura son los siguientes (fig. 1):

- *Base de Conocimiento.* Es el módulo básico y central del sistema. Aquí se encuentra disponible toda la información definida, recogida e inferida por el sistema. Se trata de un almacén donde se guardan:
  - *Ontologías*: Un conjunto de ontologías representadas en OWL (el lenguaje de definición del ontologías propuesto por el World Wide Web Consortium —W3C—) que incluye una taxonomía de dominio definida en base a vocabularios y tesauros existentes en la actualidad y ontologías con los elementos necesarios para la caracterización de pacientes y estudios sistemáticos.
  - *Reglas de Inferencia*: Son utilizadas por varios de los elementos funcionales del sistema con diferentes propósitos: integración de información, transformación entre esquemas de metadatos heterogéneos, deducción de nuevos conocimientos, detección de inconsistencias, heurísticos para el cálculo de relevancia, etc.
  - *Descripciones de documentos científicos y pacientes*. Declaraciones, expresadas como triplas RDF[4], definidas en base a los términos de las ontologías sobre los documentos científicos indizados y los pacientes.
- *Razonador y Motor de Búsqueda.* El razonador tiene como función inferir nuevos hechos (es decir, nuevas declaraciones RDF) a partir de un documento RDF/OWL base haciendo uso de la información presente en alguna de las ontologías y de un subconjunto de las reglas de inferencia. El motor de búsqueda proporciona los servicios para la realización de consultas semánticas en la Base de Conocimiento.
- *Indizador.* Accede a almacenes de documentos científicos para indizarlos en la Base de Conocimiento. En la mayoría de los casos, los textos científicos se indizan utilizando un conjunto pequeño de términos pertenecientes a tesauros desarrollados específicamente para tal fin (e.g. MeSH). Con el fin de aumentar el número de términos que indizan un determinado texto científico, se aplican algoritmos de extracción de ontologías, obteniendo así un conjunto de términos relevantes que caracterizan con mayor granularidad el texto científico y facilitan su clasificación bajo criterios semánticos.

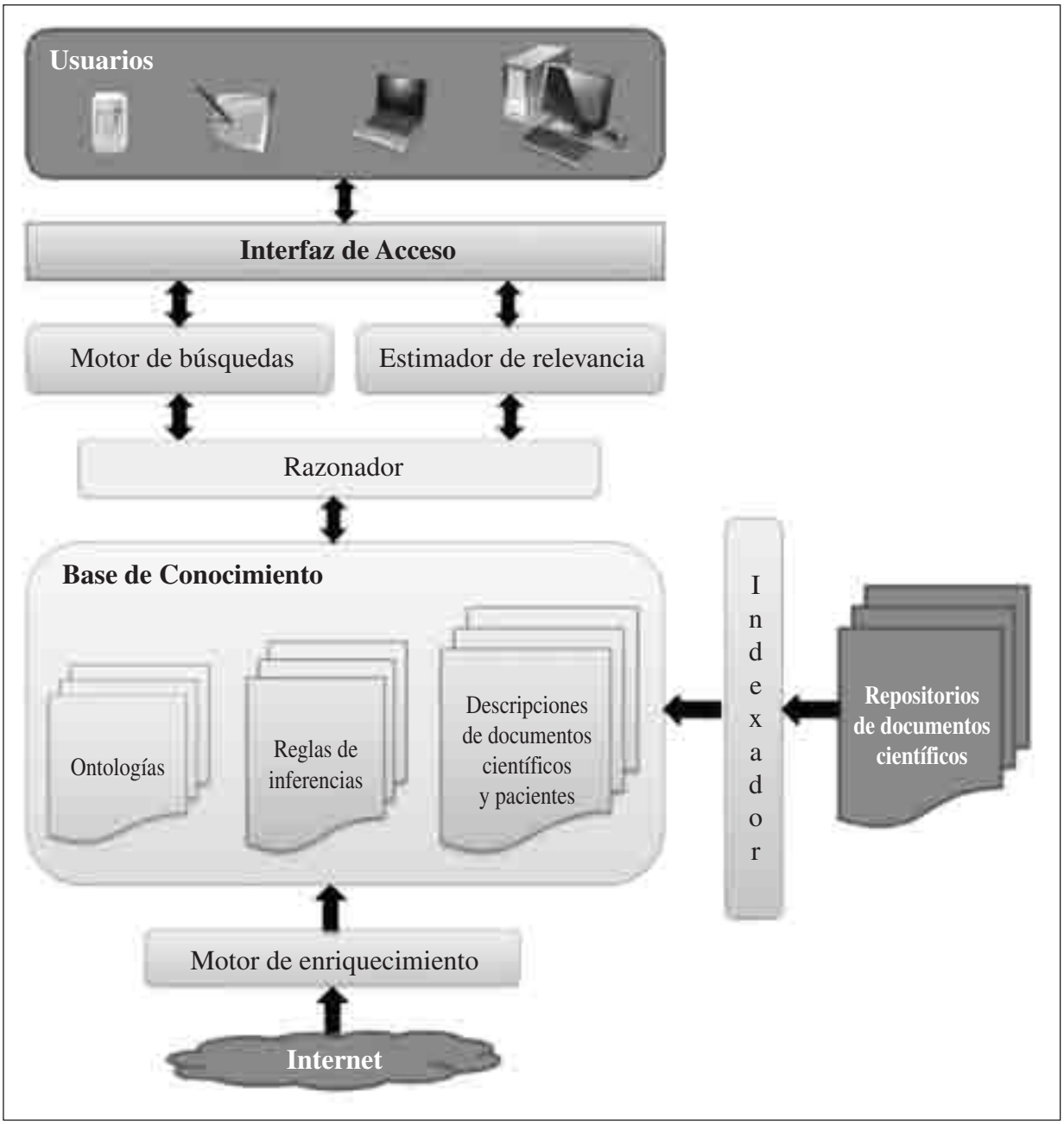

Fig. 1.—Arquitectura del Marco Semántico.

- *Motor de Enriquecimiento.* Recoge de diferentes fuentes disponibles en Internet información relativa a los datos ya existentes en la Base de Conocimiento para enriquecerla. En particular, se encarga de expandir los elementos del tesauro de dominio.
- *Estimador de relevancia.* Se ha definido un algoritmo de cálculo de relevancia que permite ordenar los textos científicos relacionados con un paciente en función de su importancia estimada, permitiendo así asesorar la atención de dicho paciente de un modo semi-automatizado, en función de los últimos documentos disponibles en la literatura.
- *Interfaz de acceso.* Proporciona servicios de acceso para los agentes software que se comunican con los usuarios finales, i.e., los investigadores y los profesionales sanitarios.

Base de Conocimiento

Para realizar el filtrado y recomendación de documentos científicos de un modo riguroso y sistemático es necesario disponer de una descripción detallada y consistente de los datos manejados por el sistema[5,6]. Tal y como se ha mencionado en puntos anteriores, esta

propuesta se basa en la utilización de técnicas semánticas de representación del conocimiento para estructurar la información, haciéndola de esta forma procesable e "interpretable" por sistemas computerizados.

Siguiendo las guías definidas por las principales metodologías del área de la Ingeniería del Conocimiento (e.g. Methontology[7], UPON[8], On-To-Knowledge[9]), el diseño del modelo semántico se basa en dos principios básicos: modularidad y reutilización. Así, en lugar de desarrollarse una conceptualización única y global del sistema, se ha dividido el modelo semántico en tres grandes sub-modelos: categorías taxonómicas de conocimiento, documentos científicos y pacientes; cada uno con un alcance y objetivos claramente delimitados. El proceso de diseño de cada sub-modelo ha tomado como punto de partida estándares y modelos ya vigentes en el dominio que han sido utilizados como bloques constructivos básicos de nuestra conceptualización. A través de diferentes iteraciones se ha ido refinando cada uno de estos modelos, añadiendo nuevos términos y propiedades en aquellos casos en que los vocabularios externos se han mostrado insuficientes para caracterizar de forma completa nuestro dominio de interés. Dado que cada uno de estos modelos se desarrolla de forma independiente, es posible reutilizarlos en distintos contextos, combinarlos y establecer equivalencias *(mappings)* con modelos externos, creándose una arquitectura final interoperable, en línea con los principios rectores de la Web de los Datos[10].

Por otro lado, además de cada conceptualización individual, la Base de Conocimiento incluye un conjunto de reglas heurísticas, específicas de nuestra aplicación, en las que se codifica parte de la lógica subyacente en el dominio. Se han definido reglas para diferentes propósitos (integración de información, detección de inconsistencias, etc.), aunque cabe destacar aquellas que permiten relacionar el perfil de un paciente y su entorno con aquellos textos científicos que tratan en mayor o menor medida las posibles patologías del paciente (ejemplo de regla heurística: "el cólico del lactante es un trastorno de los bebes"). Si bien son reglas simples, la suma de ellas aporta una gran potencia al modelo. En nuestro sistema, se contempla que este tipo de conocimiento heurístico se vaya incorporando paulatinamente por expertos en el dominio.

Taxonomía de Dominio

La construcción de un modelo ontológico que recoja los principales conceptos en el ámbito de la salud no es un problema nuevo, existiendo diferentes iniciativas formales que abordan este proceso de forma rigurosa y consensuada. Como consecuencia de este esfuerzo se han desarrollado modelos como el vocabulario MeSH (Medical Subject Headings)[11], utilizado para indizar los trabajos citados en el portal PubMed.gov, o UMLS (Unified Medical Language System)[12]. Asimismo, existen trabajos que tratan de alcanzar una representación ontológica del dominio más informal, como se propone en el trabajo de Chun et al.[13], donde se pretende recoger e integrar la información generada en diferentes aplicaciones propias de la Web 2.0 (redes sociales y foros de discusión de pacientes, portales web de salud, blogs médicos, etc.). En la revisión realizada por Fernández-Luque et al.[14] se enumeran diferentes técnicas para extraer información de la Web Social para la personalización de servicios de salud. Sin embargo, este tipo de fuentes de datos informales y desestructuradas no han sido consideradas en nuestro caso como fuentes válidas para la extracción de los descriptores debido a la dificultad de obtener de ellas información con un nivel de fiabilidad lo suficientemente alto[14].

En nuestro caso, el modelado de los descriptores biomédicos se realiza siguiendo una aproximación fuertemente basada en el uso de vocabularios, tesauros y taxonomías normalizadas y, por tanto, altamente estructuradas, de amplio uso entre la comunidad científica. En particular, la base de este trabajo se fundamenta en el uso de los descriptores MeSH, así como de términos definidos en glosarios publicados en fuentes reputadas[15]. No obstante, y a diferencia de otras propuestas, se contempla también la expansión semántica del conjunto inicial de términos identificados en estos modelos. Esta expansión puede ser manual, mediante la incorporación de nuevos términos y actualizaciones por parte de los profesionales e investigadores del ámbito de la salud o bien semi-automática, lo que resulta verdaderamente interesante, utilizando para ello técnicas de enriquecimiento semántico.

Los descriptores obtenidos se estructuran como un sistema de clasificación formal del conocimiento, donde las características de cada concepto son expresadas en forma de triplas RDF[4] utilizando propiedades de SKOS[16] (modelo diseñado por el W3C para la representación de tesauros). SKOS ha sido empleado tanto para establecer relaciones estructurales *(skos:broader*, *skos:narrower)* como asociativas *(skos:related)* entre los términos, además de permitir ligarlos con definiciones equivalentes en sistemas externos *(skos:exactMatch*, *skos:closeMatch)* como DBpedia o Wordnet. Esta solución técnica capacita a nuestro sistema para trabajar, cuando sea preciso, bajo más de un sistema de clasificación, haciendo uso de mecanismos de inferencia semántica para hacer búsquedas complejas sobre repositorios que utilicen sistemas de descriptores heterogéneos.

Caracterización de documentos científicos

hDOC, el modelo semántico que hemos diseñado para la descripción de documentos científicos, combina términos y vocabularios propios de los modelos de gestión documental (e.g. Dublin Core y BibTex) con otros más concretos del ámbito de la medicina y la

salud (e.g. PubMed, HL7 CDA[17], LOINC CDO[18]). Este modelo híbrido nos permite aprovechar al máximo la interoperabilidad aportada por modelos de datos genéricos como Dublin Core[19], al tiempo que nos garantiza contar con el potencial expresivo de los estándares médicos. A diferencia de modelos como CDA, el objetivo básico de la ontología documental es caracterizar el contenido de un documento y no la forma en que se estructura.

Las propiedades más importantes definidas en hDOC, desde el punto de vista del marco objeto de nuestra propuesta, son las que permiten indizar un documento en base a los descriptores de la taxonomía antes discutida. En este sentido, hDOC establece un mecanismo de definición de los contenidos en base a descriptores donde, frente a propuestas como PubMed, dotadas de una única propiedad *(meshHeading),* se especializa esta relación en cinco propiedades únicas:

- *primaryTopic*: tema o temas centrales del documento (e.g. apendicitis).
- *mentions*: temas secundarios que son tratados por el documento (e.g. lupus).
- *tag*: temas no existentes en el tesauro de dominio antes mencionado.
- *patientGroup*: tipología de paciente (e.g. mediana edad, hombre, recién nacido, etc.).
- *docType*: tipo de documento (e.g. prueba clínica, artículo científico).

La asignación de valores a estas propiedades para un documento particular se lleva a cabo de forma automática por el indizador (fig. 1).

Perfil de pacientes

En nuestro marco también se contempla, como elemento fundamental, la caracterización semántica de las particularidades de un paciente. En la actualidad existen en el ámbito de las ciencias de la salud diferentes perfiles que permiten describir las características relevantes de un paciente desde una perspectiva médica. En particular, no podemos dejar de mencionar iniciativas como openEHR[20] o los modelos propuestos por HL7[21].

Esta última, HL7, es una propuesta holística que contempla diferentes aspectos del registro electrónico del paciente a lo largo de todo su ciclo de vida como paciente. De hecho, se define en el ámbito de una propuesta global para la gestión de todo el sistema sanitario.

Esta propuesta, si bien es completa y ha sido convenientemente probada, no se adapta a nuestras necesidades. Nuestro sistema persigue confrontar el perfil del paciente con los documentos médico-sanitarios existentes para comprobar la pertinencia de estos últimos a la sintomatología actual del paciente. Es por ello que se optó por una aproximación más sencilla, fundada en estos perfiles, pero orientada a facilitar los objetivos del marco semántico propuesto. Se mantienen puentes que permiten una fácil integración con modelos preexistentes, pero nos centramos en el descubrimiento de evidencias y documentos útiles para la praxis médica, contextualizada al paciente, por medio de motores y técnicas de inferencia propias de las tecnologías semánticas.

Así, el modelo semántico propuesto para la caracterización de pacientes se basa en reutilizar las bases proporcionadas por FOAF[22] y complementarlas con propiedades que incluyan los aspectos principales observados por su médico, y relevantes para su contraste con los estudios médicos disponibles y ya acondicionadas para este fin.

Este modelo se articula alrededor de los siguientes módulos:

- *FOAF*. Se trata del módulo básico del modelo donde se define la información básica del paciente como persona, es decir, incluye información como nombre, dirección de correo, dirección postal, relaciones con otras personas, etc.
- *PersonalCondition*. Incluye información sobre diversos parámetros biomédicos tales como peso, altura, presión arterial, etc. Es importante notar que, a diferencia de otros modelos, en este módulo se almacena la información actual del paciente, sin considerar históricos o control de cambios sobre estos datos.
- *Symptomatological*. Este módulo se encarga de describir las diferentes situaciones de carácter sintomático que caracterizan al paciente y que son o pueden ser de relevancia para la búsqueda de información sobre su persona.

Motor de enriquecimiento

El motor de enriquecimiento es un módulo autónomo del sistema que permite complementar, de forma semi-automática, la Base de Conocimiento con información obtenida de fuentes externas.

Los Sistemas Basados en Conocimiento actuales mantienen y actualizan sus bases de conocimiento mediante las aportaciones de sus usuarios. Esto condiciona la capacidad y potencia del sistema a la implicación y esfuerzo de la comunidad. Sin embargo, en la web actual existe gran cantidad de conocimiento que puede ser libremente reutilizado para enriquecer y actualizar esta información de forma automática y transparente a los usuarios finales. Esta filosofía de enriquecimiento ya ha sido recientemente planteada en trabajos previos[23,24] para otros ámbitos de investigación.

El motor de enriquecimiento de nuestro sistema obtiene información de fuentes externas asociadas a la iniciativa Linked Open Data (LOD). En esta iniciativa se establece una metodología para la publicación de la información especialmente ideada para el procesado por agentes software. Esto permite que el módulo de

enriquecimiento acceda de forma estándar, mediante consultas semánticas, a una enorme cantidad de información interesante expresada de forma homogénea mediante el lenguaje RDF.

El motor de enriquecimiento está compuesto por un agente inteligente que tiene como componente principal la herramienta SILK[25]. Esta herramienta permite buscar en los repositorios externos de LOD los registros que hagan referencia a los mismos conceptos que nosotros tenemos registrados en nuestra Base de Conocimiento. Esto nos permite obtener nueva información de registros específicos del ámbito sanitario ya presentes en nuestra Base de Conocimiento. Por ejemplo, si tenemos registrado localmente el concepto "influenza" es posible recuperar de un repositorio externo información complementaria, como sus sinónimos y su traducción y uso en otros idiomas (grippe, flu, grypa, influensa, gripe, etc.). Este proceso de ligar registros diferentes que hacen referencia al mismo objeto en el mundo real se denomina Record Linkage[26]. Para llevarlo a cabo, la herramienta SILK debe ser configurada por un experto para definir las entidades y propiedades a comparar, los algoritmos de comparación y los umbrales de similitud aceptados.

En nuestro marco las principales fuentes empleadas son:

- DBPedia, Factforge, MediCare, Bio2RDF y Diseasome, entre otros, son repositorios semánticos de LOD con información del ámbito sanitario que nos permiten obtener información general sobre diferentes conceptos (e.g. sus diferentes nombres, terminología científica asociada a cada uno, genes asociados, medicamentos posibles, sintomatología, etc.). Además, estos repositorios organizan cada concepto en una estructura jerárquica que permite encontrar y descubrir nuevos conceptos sanitarios que tengan algún tipo de relación (por ejemplo, en DBPedia el concepto "influenza" es una enfermedad clasificada como "respiratory diseases" al igual que el concepto "bronchitis").
- WordNet, es una enorme base de datos léxica que, además del significado de cada uno de los términos que tiene almacenados, publica los sinónimos, hiperónimos e hipónimos para cada uno de ellos. Esto permite ampliar el repertorio de nombres asociados a cada uno de nuestros descriptores.
- Sindice es un buscador semántico, funcionalmente similar al buscador Google, que indiza información de gran cantidad de repositorios LOD. Mediante este servicio es posible encontrar de forma directa nuevos registros en repositorios que no habían sido contemplados inicialmente. Para llevar a cabo la recuperación de los registros se utiliza la búsqueda basada en los nombres de los descriptores, los originales y los complementados en los procesos de enriquecimiento antes mencionados.

El motor de enriquecimiento se ejecuta periódicamente para así mantener continuamente actualizada la Base de Conocimiento de nuestro sistema de forma totalmente transparente.

Indizador de documentos

Este módulo se encarga de recuperar de repositorios de documentos científicos la información sobre los mismos para indexarlos en el sistema. Además de los datos puramente bibliográficos (título, fecha de publicación, autores, etc.), este módulo obtiene la temática de cada documento. La temática se adquiere principalmente a partir de las palabras clave asignadas por los propios autores, pero también a partir de otros términos relevantes obtenidos del propio contenido del documento.

Para la obtención de términos adicionales se hace uso de diferentes técnicas. En primer lugar, se trata de localizar en el documento a indizar descriptores considerados de especial relevancia en nuestra taxonomía de dominio (principalmente descriptores propios de MeSH). Esta técnica ha sido utilizada con frecuencia en la literatura[27,28]. A continuación se utilizan técnicas para obtener términos, en principio menos relevantes, en base a algoritmos de extracción automática de ontologías, tales como "Semantic Elements Extracting Algorithm"[29] o "Textpresso"[30].

Como resultado de estos procesos obtenemos un conjunto de términos que podemos utilizar para indizar el documento con una gran granularidad. El siguiente paso consiste en asignar estos términos a alguna de las propiedades, antes mencionadas, diseñadas para la caracterización de documentos. En concreto, aquellos términos considerados de mayor relevancia se asignan a la propiedad *primaryTopic*, los considerados de menor importancia, pero existentes en la taxonomía de dominio, se asignan a la propiedad *mentions* y, por último, los restantes se asignan a la propiedad *tag*. Además, estos últimos quedarán registrados para que un experto los valore y, si lo estima oportuno, los incorpore a la taxonomía del dominio y los relacione con descriptores ya existentes en la misma. También se trata de buscar términos relacionados con la tipología de pacientes mencionados (que se asignan a la propiedad *patientGroup)* ayudándose de las reglas de inferencia definidas.

Estimador de relevancia

El estimador de relevancia es el módulo encargado de obtener un valor de utilidad de un documento para un determinado perfil de paciente. En el marco propuesto se contempla la obtención de listados de documentos científicos relevantes para un perfil de paciente en base a la aplicación de una técnica de dos etapas: la etapa de filtrado y la etapa de ordenación de resultados (basada en el cálculo de la relevancia de los mismos).

El filtrado consiste en ejecutar una consulta semántica sobre la Base de Conocimiento del sistema construida en base a la información disponible del paciente y de la información adicional proporcionada por el personal sanitario. Esto permite reducir el conjunto de documentos indizados por el sistema a aquellos que tienen un potencial interés para ese paciente.

En la etapa de ordenación se le asigna a cada documento un valor numérico que representa el grado de adecuación o utilidad estimada del documento para el paciente considerado. En nuestro caso, esta función de utilidad, siguiendo la aproximación propia de los recomendadores multicriterio[31], es una suma ponderada de funciones de utilidad marginal, donde cada una de ellas evalúa un documento atendiendo a un determinado factor. Algunos de los factores considerados son: tópicos primarios (a mayor número de temas relevantes coincidentes con el perfil del paciente mayor relevancia del documento), tópicos secundarios (igual que el anterior, pero aplicable a los temas asignados como *mentions* y *tag),* tipología paciente (si el documento ha sido clasificado dentro del grupo del paciente aumenta su relevancia), fecha de publicación (documentos más recientes tienen mayor relevancia), etc.

Conclusiones y líneas futuras

En los últimos tiempos estamos asistiendo a un proceso de convergencia donde las Ciencias de la Salud, las Tecnologías de la Información, y las Ciencias del Conocimiento se unen para ofrecer una atención médica personalizada, moderna y eficaz. Esta medicina personalizada requiere la combinación de datos heterogéneos disponibles en el ámbito de la investigación médica, la práctica clínica, y datos generales sobre un paciente o grupo de pacientes. Un problema bien conocido es la abrumadora cantidad de información disponible en un contexto donde todavía no se aprovecha al máximo el potencial del procesamiento semántico para transformar adecuadamente toda esa información en conocimiento. Este problema es especialmente relevante en el ámbito de las ciencias de la salud, ya que hay procesos clave en este dominio que dependen de manera dramática de la disponibilidad de conocimiento de alta calidad, completo, actualizado y relevante. Podemos decir que la generalización del concepto de medicina personalizada necesita de nuevas soluciones que permitan generar, de las múltiples fuentes de información disponibles, conocimiento relevante para un paciente específico o un grupo de pacientes.

En este artículo presentamos un marco de referencia original, asociado a una prueba de concepto, basado en la introducción del razonamiento semántico para recuperar el conocimiento disponible más relevante de acuerdo con el perfil de salud de una persona determinada. Este trabajo se encuadra dentro de los trabajos de los autores relacionados con la aplicación de las tecnologías propias de la Web Semántica a escenarios de aplicación concretos. Para ello, se desarrolló un modelo semántico para representar los perfiles de salud de los pacientes, y para identificar y caracterizar semánticamente fuentes existentes de información relevante como bases de datos médicas, colecciones de artículos, fuentes genéricas públicas con contenido de calidad en el ámbito de las ciencias de la salud.

El marco propuesto pretende contribuir a la mejora y evolución de las técnicas aplicadas en el ámbito de la salud personal y de la medicina basada en la evidencia. Las herramientas desarrolladas también podrían ser utilizadas para aprovechar el conocimiento existente con el fin facilitar una revisión sistemática de informes, estudios y análisis que pueden ser relevantes para las condiciones de salud de los usuarios individuales.

Agradecimientos

La publicación del monográfico del que forma parte este artículo está financiada por el Proyecto PCI-AECID (A1/037839/11), dentro de la convocatoria del Programa de Cooperación Interuniversitaria e Investigación Científica (PCI), perteneciente a las ayudas para Acciones Integradas para el Fortalecimiento Científico e Institucional de la Agencia Española de Cooperación Internacional para el Desarrollo (AECID).

Conflicto de intereses

Los autores declaran que no existe ningún compromiso o vínculo con la entidad financiadora que pueda ser entendido como un conflicto de intereses.

Referencias

1. Akerkar R, Sajja P. Knowledge-Based Systems. Ontario, Canada: Jones & Bartlett Learning; 2010.
2. García-Sánchez F, Alvarez-Sabucedo L, Martínez-Béjar R, Anido-Rifón L, Valencia-García R, Gómez J. A knowledge technologies-based multi-agent system for egovernment environments. *Lecture Notes in Computer Science (LNCS)* 2008; 5006: 15-30.
3. Santos JM, Llamas M, Anido L. Applying Semantic Techniques to Integrate Course Catalogues. Computers and Education - E-learning, from Theory to Practice. 2007; (Chapter 7): 77-87.
4. Manola F, Miller E. RDF Primer. W3C Recommendation [sede Web]. W3C; [actualizada 10 feb 2004; citado 25 jul 2012]. Disponible en: http://www.w3.org/TR/rdf-primer/
5. Planas M, Rodríguez T, Lecha M. La importancia de los datos. *Nutr Hosp* 2004; 19 (1): 11-3.
6. Wanden-Berghe C, Sanz-Valero J, Culebras JM; Red de Malnutrición en Iberoamérica Red MeI-CYTED. Información en Nutrición Domiciliaria: la importancia de los registros. *Nutr Hosp* 2008; 23 (3): 220-5.
7. Fernández M, Gómez-Pérez A, Juristo, N. Methontology: From Ontological Art Towards Ontological Engineering. Actas del Symposium on Ontological Engineering of AAAI; Marzo 1997, Stanford, California, Estados Unidos: 33-40.
8. De Nicola A, Missikoff M, Navigli R. A software engineering approach to ontology building. *Information Systems* 2009; 34: 258-75.

9. Sure Y. On-To-Knowledge - Ontology based Knowledge Management Tools and their Application. *Kuenstliche Intelligenz* 2002; 1 (02): 35-7.
10. Bizer C, Heath T, Berners-Lee T. Linked Data - The Story So Far. *Int J Semant Web Inf Syst* 2009; 5 (3): 1-22.
11. Camps D, Recuero Y, Avila RE, Samar ME. Herramientas para la recuperación de la información: Los términos MeSH (Medical Subject Headings). *MedUNAB* 2006; 9 (1): 58-62.
12. Lindberg DA, Humphreys BL, McCray AT. (1993). The Unified Medical Language System. *Methods Inf Med* 1993; 32 (4): 281-91.
13. Chun S, MacKellar B. Social health data integration using semantic Web. Actas del 27th Annual ACM Symposium on Applied Computing; 26-30 Marzo 2012, Riva del Garda (Trento), Italy: 392-397.
14. Fernandez-Luque L, Karlsen R, Bonander J. Review of Extracting Information From the Social Web for Health Personalization. *J Med Internet Res* 2011; 13 (1): e15.
15. Moreno Villares JM, Álvarez Hernández J, Wanden-Berghe Lozano C, Lozano Fuster M. Glosario de términos y expresiones frecuentes de Bioética en la práctica de la Nutrición Clínica. *Nutr Hosp* 2010; 25 (4): 543-8.
16. Miles A, Bechofer S. SKOS Simple Knowledge Organization System - Reference. W3C Recommendation [sede Web]. W3C; [actualizada 18 ago 2009; citado 25 jul 2012]. Disponible en: http://www.w3.org/TR/2009/REC-skos-reference-20090818/
17. Robert H, Dolin RH, Alschuler L, Beebe C, Biron PV, Boyer SL et al. The HL7 Clinical Document Architecture. *J Am Med Inform Assoc* 2001; 8 (6): 552-69.
18. Hyun S, Shapiro JS, Melton G, Schlegel C, Stetson PD, Johnson SB et al. Iterative Evaluation of the Health Level 7—Logical Observation Identifiers Names and Codes Clinical Document Ontology for Representing Clinical Document Names: A Case Report. *J Am Med Inform Assoc* 2009; 16: 395-9. [DOI: 10.1197/jamia.M2821]
19. Dublin Core Metadata Initiative. DCMI Metadata Terms [sede Web]. DCMI; [actualizada 14 jun 2012, citado 25 jul 2012]. Disponible en: http://dublincore.org/documents/dcmi-terms/
20. The openEHR Foundation. openEHR: future proof and flexible EHR specifications [sede Web]. openEHR; [citado 25 jul 2012]. Disponible en: http://www.openehr.org/home.html
21. hl7.org. Health Level Seven International. HL7; [citado 25 jul 2012]. Disponible en: http://www.hl7.org
22. foaf-project.org. The Friend of a Friend Project. FOAF; [citado 25 jul 2012]. Disponible en: http://www.foaf-project.org
23. Alonso-Rorís VM, Míguez-Pérez R, Santos-Gago JM, Álvarez-Sabucedo L. A Semantic Enrichment Experience in the Early Childhood Context. Actas del Frontiers in Education Conference.
24. Ruiz-Calleja A, Vega-Gorgojo G, Asensio-Pérez JI, Bote-Lorenzo ML, Gómez-Sánchez E, Alario-Hoyos C. A Linked Data approach for the discovery of educational ICT tools in the Web of Data. *Computers & Education* 2012; 59 (3): 952-62.
25. Volz J, Bizer C, Gaedke M, Kobilarov G. Silk - A Link Discovery Framework for the Web of Data. Actas del WWW2009 Workshop on Linked Data on the Web; 20 Abril 2009, Madrid, Spain.
26. Winkler WE. Overview of Record Linkage and Current Research Directions, Technical Report. U.S. Bureau of the Census; [actualizada 8 feb 2006, citado 25 jul 2012]. Disponible en: http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf
27. Uramoto N, Matsuzawa H, Nagano T, Murakami A, Takeuchi H, Takeda K. A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal* 2004; 43 (3): 516-33.
28. Rak R, Kurgan L, Reformat M. Multi-label associative classification of medical documents from MEDLINE. Actas del Fourth International Conference on Machine Learning and Applications; 15-17 Diciembre 2005, Los Angeles, California, Estados Unidos. [DOI: 10.1109/ICMLA.2005.47]
29. Dung TQ, Kameyama W. A Proposal of Ontology-based Health Care Information Extraction System: VnHIES. Actas de la IEEE International Conference on Research, Innovation and Vision for the Future; 5-9 Marzo 2007, Hanoi, Vietnam: 1-7.
30. Müller H-M, Kenny EE, Sternberg PW. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLoS Biol. 2004;2(11):e309. [DOI: 10.1371/journal.pbio.0020309]
31. Manouselis N, Costopoulou C. Analysis and Classification of Multi-Criteria Recommender Systems. Multi-channel Adaptive Information Systems on the World Wide Web. 2007; 10 (4): 415-41. [DOI: 10.1007/s11280-007-0019-8].