
	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED	DATE	28/06/2017

Recomendaciones de artículos a través de PUBMED

AUTOR: Nicolás Alejandro Passadore

TUTORES: Carlos Luis Sanchez Bocanegra

Dr. Luis Fernandez Luque


	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED	DATE	28/06/2017

AGRADECIMIENTOS

Es mi deber agradecer a todas aquellas personas que hicieron posible la culminación de este Trabajo Final de Master. En primer lugar a quienes día a día hacían silencio en casa para que pudiera concentrarme, mi familia. En segundo lugar a mis tutores y su paciencia infinita y a mi compañero de ruta Juan Pedro Pereira Carrillo, quién me permitió con el desarrollo de su trabajo de fin de grado, cuyo tutor fue el Dr. Rafael Pastor Vargas, en la Universidad Nacional de Educación a Distancia (UNED), resolver aspectos de mi TFM.


Una mención especial para Dr. Julian Sanchez Viamonte, especialista en infectología, quien casi sin conocerme, accedió a colaborar en mi trabajo y ayudó en el desarrollo del estado del arte de mi trabajo.

Gracias a todos, lo que viene me toca a mi solo, los voy a extrañar...

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED	DATE	28/06/2017

ÍNDIX

1	Resumen Ejecutivo	4
2	Resumen Técnico.....	5
3	Introducción.....	6
3.1	Descripción de los Objetivos.....	7
4	Estado del arte.....	7
4.1	Antecedentes históricos del registro médico y la toma de decisiones.....	7
4.2	Historia del registro médico.....	8
4.3	Medicina basada en evidencia.....	8
4.4	MBE. Definición.....	8
4.5	Búsqueda de información en ciencias de la Salud. Principios.....	9
4.6	Fuentes de información.....	9
4.6.1	PUBMED (Public Medline).....	9
4.6.2	BCP (Biblioteca Cochrane Plus).....	10
4.6.3	UpToDate.....	10
4.6.4	Embase.....	10
4.6.5	Google Scholar (Académico).....	10
4.7	Problemática.....	11
5	Procesamiento de lenguaje natural.....	12
5.1	UIMA (Unstructured Information Management Architecture).....	12
5.2	CTAKES (clinical Text Analysis And Knowledges Extraction System).....	14
5.2.1	Componentes principales.....	15
5.2.1.1	Preprocesador de Documentos (Document Preprocesor.....	15
5.2.1.2	Núcleo (core).....	15
5.2.1.3	Detector de oraciones (Sentence boundary detector).....	15
5.2.1.4	Segmentador (Tokenizer).....	15
5.2.1.5	Detector de aserciones (Assertion).....	15
5.2.1.6	Analizador sintáctico (Chunker).....	15
5.2.1.7	Segmentador dependiente del contexto (Context dependent tokenizer).....	16
5.2.1.8	Analizador de dependencia (Dependency parser).....	16
5.2.1.9	Anotador de búsqueda de diccionario (Dictionary lookup annotator).....	16
5.2.1.10	Generador de variantes léxicas (LVG).....	16
5.2.1.11	Etiquetador gramatical (Part-of-speech tagger).....	16
5.2.2	Sistema para la detección de terminología clínica de textos en español.....	17
6	Metodología.....	17
6.1	Descripción macro del método.....	20
6.2	Método: Esquema de funcionamiento.....	22
6.2.1	En funcionamiento.....	22
7	Resultados.....	24
8	Discusión.....	28
9	Trabajos Futuros.....	29
10	Referencias y Bibliografía.....	30
11	Anexo: Consultas a PUBMED con E-Utilities y SPARQL.....	31

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

1 Resumen Ejecutivo

El uso de Internet como un gran repositorio de información, incluida información en salud, ha generado que pacientes, médicos e investigadores hagan uso de la gran red de redes para realizar diferentes búsquedas de interés. Si bien, el acceso es público y masivo, dar con la información precisa en este gran repositorio, no es tarea fácil. El contenido en Internet es una combinación, de material de gran calidad, con material precario, engañoso y desactualizado en relación a los avances que se van produciendo en distintas áreas del conocimiento. Los distintos criterios y métodos de búsqueda de información en salud en Internet, arrojan resultados dispares. Tanto los tradicionales buscadores como Google o Bing, o bien los buscadores que ofrecen las fuentes de información en salud conocidas, emiten resultados poco claros y generalmente encontrar los que se busca, es como buscar “una aguja en un pajar”.

Sin ir más lejos, solo hace falta establecer como parámetros en cualquier buscador, alguna palabra clave sobre salud, para verificar que la cantidad de resultados que se muestran son un problema para el interesado, el cual se vé obligado a descartar material por estar imposibilitado de revisar todos los resultados obtenidos.

Este Trabajo Final de Master, hace foco en la necesidad de establecer un método de búsqueda en fuentes de información publicadas en Internet, para este trabajo PUBMED, más efectivo y preciso, en relación a los intereses del usuario.


Para este trabajo se han investigado diversas fuentes de información, desde fuentes privadas a fuentes públicas, como también los tradicionales buscadores, Google y Bing. Se han realizado múltiples pruebas sobre los métodos de búsquedas que establecen dichas fuentes, para determinar los porqué de los resultados arrojados y encontrar de esta forma un método adecuado y certero para constituir resultados más efectivos. Para lograr el objetivo se utilizaron distintos métodos de investigación, desde revisión de bibliografía, análisis de datos en la web y entrevistas con profesionales médicos.

El primer paso de este trabajo fue buscar fuentes de información en salud y realizar búsquedas de diferentes temas en cada una de ellas. Para este paso las entrevistas con los profesionales médicos, principales interesados, fueron determinantes. El segundo paso fue estudiar los métodos de búsqueda establecidos por las fuentes de información en salud, además de Google y Bing.

Con todo el material encontrado e investigado, más los conocimientos adquiridos durante todos estos meses, se definió un nuevo método de búsqueda contextualizada, que parte desde los datos vertidos por un profesional médico en la historia clínica electrónica y deriva en una serie de recomendaciones de artículos de PUBMED.

El motivo para desarrollar un método de búsqueda contextualizado, que utiliza la historia clínica electrónica para recuperar palabras claves y datos del contexto, se basa en que dichos datos sirven para parametrizar la consulta final que se hará sobre PUBMED y de esta manera se puede ajustar la consulta y hacer ésta más precisa, personalizando la consulta y mejorando los resultados de la búsqueda.

El método fue probado con diferentes consultas realizadas de manera manual, mostrando una mejora notoria en los resultados obtenidos.

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

2 Resumen Técnico

Actualmente existe una gran cantidad de bibliografía científica médica, distribuida en diferentes sitios de Internet y entidades tanto públicas como privadas. A través de suscripciones en algunos casos o simplemente accediendo al sitio WEB correspondiente, se puede consultar sobre este material [BMJ 2006;333:1283].

Para realizar la búsqueda de determinado contenido, cada sitio y entidad contenedora de material bibliográfico implementan diversas maneras de llegar al artículo de interés. Ejemplos típico son el motor de búsqueda Google¹ y Bing². También existen portales que brindan servicios de búsqueda sobre varias de estas bibliotecas virtuales, intentando hacer una suerte de agrupamiento para dar mayor cobertura de forma integral en cuanto a exploración de material bibliográfico se refiere [DBSI].

La problemática de la búsqueda de artículos científicos relacionados con salud, ha sido sujeto de investigación en los últimos años [MBSSI].


Los problemas mayormente reportados pueden ser clasificados de la siguiente manera:

1. Complejidad o incomodidad que se evidencian en ciertos sitios a la hora de buscar un tema de interés.
2. Gran cantidad de material que se presenta en forma de resumen y que obligan a suscripciones para obtener acceso total a todo un artículo.
3. La efectividad de las búsquedas también es motivo de análisis. En estos casos se necesita de la pericia y experiencia de los usuarios para realizar una búsqueda que arroje resultados certeros, sin tener que seleccionar de millones de artículos posibles.
4. Necesidad de obtener respuestas precisas en cuanto al contenido del material, evitando información antigua que ha sido ya reemplazada con nueva evidencia científica.
5. Cada biblioteca virtual tiene su propio método de búsqueda y conocer cada uno de ellos sugiere tiempo y esfuerzo que el interesado no necesariamente tiene.
6. Termina siendo un requisito aprender conceptos referentes a cada método utilizado, que no siempre están al alcance de cualquier persona, sino que apunta a un público muy específico, dejando de lado a muchos otros interesados.

Aún así, con métodos de búsqueda bien definidos, la precisión de los mismos, en relación a la recuperación de recursos relevantes, sigue siendo motivo de discusión [Lawrence, Giles 1999 Bollacker]. Más cuando quienes los utilizan son profesionales médicos, los cuales se ven obligados a aprender métodos conceptualmente lejos de su actividad.

¹El **buscador de Google** o **buscador web de Google** es un motor de búsqueda en la web. [http://conceptodefinicion.de/google/]

²**Bing** es un buscador web de Microsoft. [https://diccionarioactual.com/bing/]

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED	DATE	30/06/2017

Por lo descrito anteriormente, el médico, además de recibir como respuesta en su búsqueda, material que no siempre es de su interés y la mayoría de las veces, en un volumen imposible de ser revisado, recibe también referencias desactualizadas.

No obstante, la tecnología a avanzado lo suficiente como para poder plantear proyectos que faciliten la búsqueda de material bibliográfico, incluso que integren este tipo de soluciones en ámbitos clínicos o componentes del mismo, como por ejemplo, en el registro médico electrónico.

El objetivo de este TFM es llevar a cabo una investigación que derive en una metodología de recomendaciones de artículos científicos clínicos publicados en PUBMED, en base a diagnósticos, hallazgos clínicos, resultados anormales, signos o síntomas, descritos por el profesional médico en un registro clínico electrónico. Desligando de esta manera al médico de tener que realizar la búsqueda por su cuenta.

3 Introducción

La gran cantidad de material bibliográfico médico existente en Internet, puede ser de mucha utilidad tanto en la práctica asistencial, como en la docencia e investigación [FICSI]. Pero lamentablemente mucho de este material se pierde de ser consultado, dado la ineficacia en la búsqueda realizada por los interesados o por el método de consulta empleado por los diferentes portales o bibliotecas virtuales.

La librería PUBMED³ por ejemplo, contiene al rededor de 27 millones de citas a artículos científicos sobre diferentes temas. Además implementa algunos métodos de búsqueda para facilitar el acceso a dichos artículos. El gran volumen de bibliografía que se encuentra en esta librería y en otras tantas, requiere de un esfuerzo a veces imposible de realizar para llegar a resultados óptimos de consulta sobre temas específicos de salud.

De esta forma, el problema planteado se puede visualizar desde varios aspectos:

1. Por un lado, mucha información disponible pero a su vez inaccesible y mezclada con material antiguo.
2. Formas de búsqueda poco confiables en relación al resultado arrojado.
3. La cantidad de resultados otorgados es un problema, dado que es imposible tener tiempo de revisar cada uno, generando esto, tener que discriminar qué leer y qué no leer. Esto sugiere que no siempre se opta por la mejor opción a la hora de seleccionar con que quedarse.

Lo mencionado anteriormente plantea la necesidad de encontrar mecanismos de búsqueda más eficientes y eficaces de material bibliográfico referente a conceptos de

³“PubMed comprises more than 27 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites”. [<https://www.ncbi.nlm.nih.gov/pubmed/>]

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

salud. Mecanismos que ayuden a los interesados a llegar de manera más directa a la información deseada en el momento preciso, es decir, llegar a los artículos más relevantes según su interés.

Es de destacar que el estudio realizado en este trabajo final de master, tratará sobre recomendaciones de artículos científicos referentes a aspectos de salud, en base a los datos vertidos en un registro clínico electrónico, por un profesional médico durante la atención sanitaria a un paciente.

3.1 Descripción de los Objetivos

Este trabajo busca definir un método de recomendaciones de artículos científicos de salud, capturando los datos vertidos por un profesional médico en un registro clínico electrónico mediante técnicas de procesamiento de lenguaje natural.

El objetivo primario es:

- a) Dar apoyo y soporte al profesional médico.


Los objetivos secundarios son:

- b) Aplicar métodos de búsqueda de artículos científicos médicos sobre la librería PUBMED.
- c) Capturar conceptos clínicos desde el registro clínico electrónico y tomarlos como entrada para generar las recomendaciones pertinentes. Para este objetivo en particular, se recibió la ayuda de Juan Pedro Pereira Carrillo, el cual desarrolló un algoritmo para la detección de terminología clínica de textos en español, el cual he utilizado para mi trabajo.

4 Estado del arte

4.1 Antecedentes históricos del registro médico y la toma de decisiones

En todos los ámbitos de la sociedad, el uso de sistemas informáticos a causado una gran revolución. Particularmente, en el campo de la salud, la tecnología ha incursionado de manera gradual. En los comienzos, solo aquellos procesos que pertenecen a la capa administrativa, eran alcanzados por la informática dentro de un sanitario u hospital, dejando de lado la capa clínica. Luego, con el correr del tiempo, el registro clínico tomó la ventaja y los esfuerzos se centraron en informatizar todos aquellos campos que producen información acerca del paciente. Este proceso de informatización centrado en el paciente y orientado al acto clínico del profesional médico, es el que actualmente mayor cantidad de datos genera.

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

4.2 Historia del registro médico

En diferentes momentos de la historia de la humanidad, se han presentado modelos de registro médico [ERMHI]. En principio un médico asentaba todos los síntomas de un enfermo a lo largo de su enfermedad. Luego, la lógica y conocimiento del médico pasaron a ser los datos registrados sobre la enfermedad de un paciente. Pero con el paso del tiempo y la aparición de algunas técnicas de exploración y el mayor conocimiento, lo que se registraba eran las observaciones que el médico realizaba. En los años 70, el diseño del registro médico cambia y aparece lo que hoy se denomina como, historia clínica orientada a problemas. Básicamente, es estructurar las anotaciones de los médicos bajo una lista de problemas mandatoria, que ordena las anotaciones del médico asociando las observaciones a un problema de la lista. Este nuevo diseño más adelante, sirvió para informatizar el registro clínico.

4.3 Medicina Basada en evidencia

La Medicina moderna occidental, de carácter experimental, funda sus principios en el método científico, la Medicina basada en la Evidencia (MBE).

4.4 MBE. Definición


La MBE, término acuñado por Guyatt, se define como un proceso cuyo objetivo es el de obtener y aplicar la mejor evidencia científica en el ejercicio de la práctica médica cotidiana. Para eso se requiere la utilización concienzuda, juiciosa y explícita de las mejores “evidencias” disponibles en la toma de decisiones sobre el cuidado sanitario de los pacientes [CEBM].

En principio el concepto de MBE se explicaba como la determinación de la mejor evidencia derivada de la investigación científica para la resolución de problemas clínicos. Pero esta definición conceptual, tuvo muchas críticas [BMJ 1996; 312:71-2] y en el año 1996 se decidió replantear el concepto de MBE.

Bajo esta definición “la mejor evidencia científica” es aquella investigación clínicamente relevante, que se realiza sobre las pruebas diagnósticas (incluida la exploración física), el poder de los marcadores pronósticos o sobre la eficacia y seguridad de los regímenes terapéuticos, rehabilitadores y preventivos.

Aún así, la MBE además de comprender la evidencia externa (entiéndase por esto mejor evidencia científica), también comprende la evidencia interna (experiencia práctica individual o personal de cada médico), de esta forma, aún si la evidencia científica no puede ser aplicada a un paciente, siempre existirá la experiencia y conocimiento del médico tratante.

La MBE supone una serie de ventajas, que abarcan al profesional médico, al paciente y al sistema sanitario. Constantemente surgen nuevos tipos de evidencias que, cuando son

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

conocidas y comprendidas, crean cambios importantes y frecuentes en la forma de cuidar a los pacientes. Aún así, debe destacarse, que no todos los profesionales pueden acceder a estas nuevas evidencias y mantenerse actualizados para ejercer su práctica diaria.

4.5 Búsqueda de información en ciencias de la Salud. Principios

La información científica disponible se duplica en menos de cinco años y su producción es de tal magnitud que se estima que en el mundo se publican anualmente alrededor de 20.000 publicaciones periódicas y 17.000 nuevos libros. En este mar de información, que crece de forma exponencial, los profesionales tienen dos desafíos, en primer lugar, encontrar en el menor tiempo posible literatura actualizada sobre un tema de interés y en segundo lugar, poder adoptar frente a la información seleccionada una lectura crítica que permita discriminar entre aquellas publicaciones que resultan válidas para la toma de decisiones y las que no son pertinentes [IEMORDSC].

4.6 Fuentes de Información

Existen diversas fuentes de información biomédicas en Internet. Cada una de estas fuentes puede contener tanto datos estructurados, por ejemplo, datos contenidos en base de datos relacionales, como también datos no estructurados, por ejemplo, publicaciones científicas. Como una característica no menos importante, estos datos pueden estar en diferentes idiomas. Este gran ecosistema de información, es el que permite pensar en la necesidad de diseñar e implementar métodos de búsquedas inteligentes que permitan asociar la información y la necesidad real de un usuario, poniendo a disposición de los interesados la información requerida. Por este motivo, la estrategia de búsqueda no debe ser subestimada. Hay que identificar bien el contexto, los términos que se van a utilizar, la base de datos a consultar y recién tomadas todas esas decisiones, realizar la búsqueda pertinente.

Actualmente existen varias fuentes de información que se describen a continuación. Cabe aclarar que se destacaran las más relevantes para esta investigación.

4.6.1 PUBMED (Public Medline)

Principal base de datos de referencias bibliográficas biomédicas. Cuenta con más de 27 millones de referencias a artículos de salud. La Biblioteca Nacional de los Estados Unidos (NLM), pone a disposición dicha base de consulta. Proyecto desarrollado por el Centro Nacional de Biotecnología de los Estados Unidos (NCBI), PUBMED da acceso un repositorio de múltiples base de datos de bibliografía que son mantenidas por la NLM. Éstas son algunas de las bases de datos que se encuentran en dicho repositorio:

1. MESH.
2. PMC.
3. MEDLINE.

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

4. PUBMED.

PUBMED permite ejecutar búsquedas de dos maneras, la primera es la más trivial, solo ingresando los términos que se desean buscar y luego esperar el resultado. La segunda opción es más compleja y completa, permite utilizar ciertas reglas, como restricciones y combinaciones de términos a través de operadores lógicos como AND y OR.

El enlace a Pubmed es: <https://www.ncbi.nlm.nih.gov/pubmed/>

4.6.2 BCP (Biblioteca Cochrane Plus)

Fuente de información de MBE. Se constituye de varias fuentes de información. El principal objetivo es elaborar revisiones sistemáticas a partir de ensayos clínicos controlados y a evidencia de otras fuentes.

En enlace a BCP es: <http://www.bibliotecacochrane.com/>

4.6.3 UpToDate

Es un base de datos, cuyo contenido es original, dado que incluye material escrito y revisado por un conjunto de médicos. Tiene más de 10.000 temas de interés y un poco más de 20 especialidades cubiertas. Los enlaces que ofrece son a artículos completos y además cuenta con otros servicios como por ejemplo el de interacciones entre drogas y una propia base de datos de medicamentos. Cuenta también con información al paciente y actualizaciones diarias. UpToDate, es un servicio pago, que puede ser adquirido por un profesional de la salud o por una institución en particular. Este servicio ofrece la posibilidad de realizar la búsqueda de información de diferente maneras. Por un lado, introduciendo un término y procediendo a la búsqueda. Desde la propia historia clínica del paciente, incluyendo algunos datos de contexto, como la edad por ejemplo. Siempre el resultado de la búsqueda trae las referencias a los artículos más recientes dentro de la base de datos de UpToDate.

El enlace a UpToDate es: <http://www.uptodate.com/es/home>

4.6.4 Embase

Generada por la editorial Elsevier, esta base de bibliografía de biomedicina, es una versión automatizada del repertorio Excerpta Medica, editada por dicha fundación con origen en Ámsterdam. En su base de datos, se encuentran además de artículos científicos, tesis, libros y demás documentos.

El enlace a Embase es: <https://www.elsevier.com>

4.6.5 Google Scholar (Google Académico)

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

Buscador de Google que se centra en información científico-académica. Además de referenciar libros, artículos, informes etc., también referencia repositorios, base de datos de información académica y artículos científicos. Como dato interesante en la forma de búsqueda, si bien el algoritmo de búsqueda es similar al usado puramente desde Google, agrega un parámetro que define la calidad de los resultados a buscar. Basándose, en la calidad de la revista, libro etc. que se encuentra referenciado.

El enlace a Google Académico es: <https://scholar.google.com/>


4.7 Problemática

Como se mencionó en el apartado anterior, existe una gran cantidad de fuentes de datos que contienen artículos, libros y documentos científicos.

Toda estas fuentes aplican métodos de búsquedas similares, algunas contienen más de una forma de buscar un término o determinada información, haciendo una clara división, entre lo que es un método simple y un método avanzado. Pero todas estas fuentes, sumando los tradicionales buscadores de Internet, como Google y Bing, más allá del método empleado para buscar información, generan un número de resultados de difícil manejo para quien necesita dicha información. Sumado a esto, cuando un profesional médico busca algún término de interés, las referencias resultantes de la búsqueda presentan una serie de aspectos que pueden clasificarse confusos y que poco ayudan a la necesidad real del interesado. Resultados con información fiable y no fiable, para expertos de una materia o para público en general. Ya sea que el usuario tenga o no experiencia en determinada área, las dificultades en las búsquedas, están a la orden del día. La búsqueda de información sobre casos clínicos por parte de un profesional médico, se hace tediosa debido a la gran variedad de herramientas de búsquedas que brindan el acceso a las fuentes de información [ASSEHS], dificultando esto el proceso de comprensión y la recopilación de datos concretos que ayuden a la resolución del caso concretamente.

La práctica clínica diaria exige un conocimiento actualizado de los abordajes diagnósticos y terapéuticos más eficaces y eficientes para cada paciente. Los profesionales médicos se encuentran diariamente con serias dificultades para acceder a información útil.

Se sabe que los médicos con experiencia utilizan alrededor de 2 millones de fragmentos de información para manejar a sus pacientes [BMJ, 313 (1996)], y la mayor parte de esa información es anticuada o errónea. Pero la utilización de las diferentes fuentes de información, incluso las más avanzadas tecnológicamente, fallan con frecuencia y no permiten resolver las múltiples lagunas de conocimiento que aparecen durante la práctica clínica diaria, probablemente porque fueron diseñadas sin la realización previa de un estudio en el que se describiera cuáles son las necesidades de información de los médicos [Aten Primaria 2005;35:419-22].

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

Lo descripto anteriormente a generado el interés de este Trabajo Final de Master, diseñar una metodología que especifique una manera óptima de realizar búsquedas en Internet sobre material científico relacionado con la salud.

5 Procesamiento de Lenguaje Natural

Las herramientas para el procesamiento de la información terminológica en la historia clínica electrónica se basan en lenguajes documentales que permitan clasificar, gestionar y ordenar las enfermedades. Las terminologías clínicas (como UMLS, SNOMED-CT, etc.) permiten especificar conceptos médicos (diagnósticos, procedimientos, etc.) precisos, evitando las ambigüedades y redundancias que se pueden introducir al utilizar el texto libre, mejorando el proceso de registro y reduciendo las posibilidades de error. Es importante que la recogida de la información terminológica en la historia clínica electrónica se defina previamente para poder procesarla y explotarla con posterioridad.

La tendencia para el futuro será la compilación de las diferentes ontologías médicas que permitan al profesional no sólo navegar a través de la historia clínica, sino también **acceder a bases de datos bibliográficos y herramientas de ayuda para la toma de decisiones.**

Sin embargo, una parte importante de la información clínica almacenada actualmente en estos sistemas es información textual no estructurada, texto libre. El texto libre ofrece un gran potencial de recuperación de información, pero no facilita dicha labor de recuperación, dificultando el proceso de explotación de la información recopilada⁴.


El *procesamiento de lenguajes naturales* (PLN) es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. El procesamiento de lenguaje natural tiene múltiples aplicaciones, entre las que figuran los traductores automáticos o el aprendizaje de idiomas online. En la actualidad, las técnicas utilizadas en este ámbito se encuentran ya en un estado muy avanzado para lenguas como el inglés, pero en el caso de otras –como el español–, la tecnología disponible hasta la fecha es mucho más limitada.

5.1 UIMA (Unstructured Information Management Architecture)

Como se ha mencionado con anterioridad, este Trabajo Final de Master, ha hecho uso del trabajo realizado por Juan Pedro Pereira Carrillo, estudiante de la UNED, que ha trabajado sobre la herramienta Ctakes, adaptando la misma para permitir la detección de terminología clínica en español. La mencionada herramienta se describe a continuación. Antes se definirán algunos conceptos necesarios para comprender el funcionamiento de la misma.

4

(Teresa Romá-Ferri & Palomar, 2008).

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

UIMA son sistemas de software que analizan grandes volúmenes de información no estructurada con el fin de descubrir la información relevante para el usuario final. Actualmente, UIMA es un estándar OASIS para análisis de contenido. Proporciona una arquitectura de componentes de software para el desarrollo, descubrimiento, composición y despliegue de análisis multimodales sobre información no estructurada y para su integración con las tecnologías de búsqueda.

UIMA es una arquitectura basada en datos en la que significa que sus componentes se comunican entre sí mediante el intercambio de datos anotados integrados en una estructura.

El siguiente diagrama muestra de forma esquemática una aplicación UIMA y los componentes que la integran.

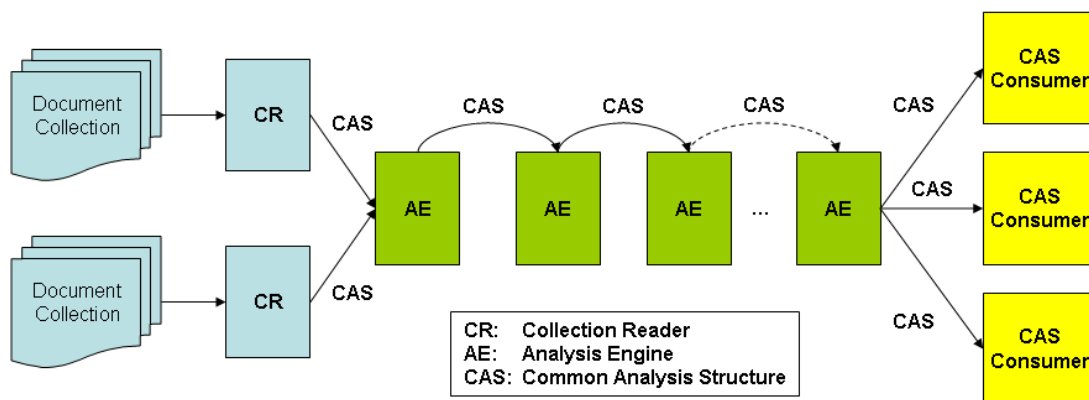



imagen 1 - Representación esquemática de una aplicación UIMA

- **Common Analysis Structure (CAS).** Es la estructura de datos primaria que UIMA utiliza para representar y compartir los resultados del análisis. Contiene:
 - El artefacto. Este es el objeto que se está analizando (documento de texto, audio o video). El CAS proyecta una o más vistas del artefacto. Cada vista se denomina *SofA (Subject of Analysis)*.
 - Una descripción de los tipos de datos del sistema. Nos indica los tipos, subtipos y sus características que pueden tener las anotaciones.
 - Metadatos de análisis. Son anotaciones de separación que describen el artefacto en si o una región del artefacto.
 - Un repositorio de índices para apoyar el acceso e iteración eficiente sobre los resultados del análisis.
- **Collection Reader (CR).** Este componente es el encargado de procesar el documento original e inicializar el CAS que será objeto del análisis. Por ejemplo, un analizador de ficheros de textos planos toma el fichero e inicializa un CAS en el que el artefacto sería el texto contenido del fichero.

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

- *Analysis Engine (AE)*. Son los encargados de analizar los documentos y de realizar las anotaciones. Por ejemplo, un detector de frases (*Sentence Detector*) toma como entrada el CAS con el texto y añade tantas anotaciones como frases contiene el texto, indicando la posición de inicio y de fin de cada frase. La salida de un AE puede ser el mismo CAS de entrada con nuevos anotadores o un nuevo CAS.
- *CAS Consumer (CC)*. Son los procesos encargados de almacenar y/o visualizar las anotaciones obtenidas durante el proceso de análisis. Por ejemplo, un CC puede encargarse de tomar los anotadores y almacenarlos en una base de datos.

Un programa simple de UIMA funciona de la siguiente manera: se le pasa como entrada una fuente de datos de información no estructurada a la cual denominaremos SoFA (*Subject of Analysis*). El motor de análisis lo procesa y, en función de la lógica que se haya establecido en el mismo, da como salida un archivo XMI que contiene enlazadas las anotaciones del texto. La combinación de varios analizadores permite obtener varios tipos de anotaciones sobre un mismo texto en el mismo archivo de salida. Cada tipo de anotación crea su propia lista enlazada de elementos. De esta forma, es fácil extraer características o propiedades del texto, combinarlas o analizarlas.

5.2 CTAKES (clinical Text Analysis And Knowledge Extraction System)

Apache cTakes® es un sistema de procesamiento de lenguaje natural especializado en la extracción de información a partir de textos clínicos desarrollado en Java y de software libre. Permite procesar notas clínicas y reconocer entidades específicas del ámbito clínico como son enfermedades, signos, síntomas, estructuras anatómicas, procedimientos diagnósticos y/o terapéuticos, etc.

Cada entidad incluye atributos que permiten ubicar y delimitar el concepto en el texto analizado. Cabe destacar los siguientes atributos entre otros:

- Código que permiten identificar el concepto en una determinada ontología médica.
- Información contextual que permite determinar si:
 - El concepto se refiere a la historia familiar.
 - El concepto dato es actual o se refiere a otro espacio temporal al pasado.
 - El concepto está relacionado con el paciente o con otro sujeto.
- El dato está negado (es decir, lo que se indica en el texto es la ausencia del concepto).

Apache cTakes® ha sido construido utilizando UIMA⁵. Sus componentes están especialmente entrenados en el dominio clínico, y crean *ricas anotaciones lingüísticas y se-*

5

UIMA (en castellano, Aplicaciones para la administración de información no estructurada) son sistemas de software que analizan grandes volúmenes de información no estructurada con el fin de descubrir la información relevante para el usuario final. Actualmente, UIMA es un estándar OASIS para análisis de contenido

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

mánticas que pueden ser utilizados tanto en sistemas de investigación clínica como en sistemas de ayuda a la decisión.

5.2.1 Componentes principales

5.2.1.1 Preprocesador de Documentos (Document Preprocesor)

Permite transformar un documento CDA (*Clinical Document Architecture*) en un documento de texto plano. El CAS resultante incluye un nuevo Sofa con anotadores que permiten diferenciar las diferentes secciones que componen el documento original.

5.2.1.2 Núcleo (Core)

Incluye los componentes básicos de un procesador de lenguaje natural: Detector de oraciones y Tokenizador. Los componentes han sido implementado a partir de componentes de Apache OpenNLP.

5.2.1.3 Detector de oraciones (Sentence boundary detector)

Ha sido implementado a partir del detector de oraciones de Apache OpenNLP. Las anotaciones se basan en la localización del carácter de fin de línea y ha sido entrenado a partir de datos clínicos anotados. Normalmente, la detección de oraciones se realiza antes de que el texto sea segmentado.

5.2.1.4 Segmentador (Tokenizer)


Este componente se encarga de partir una secuencia de caracteres en elementos denominados *tokens*. Los *tokens* son unidades indivisibles. Normalmente es el primero de los componentes que se utiliza en el procesamiento de un texto. La lista de tokens generados se usa en las siguientes etapas de esta tarea.

5.2.1.5 Detector de aseercciones (Assertion)

El módulo de aseercción considerará si una entidad o evento nombrado es negado, incierto o condicional. Además, el sujeto de la declaración puede ser alguien que no sea el paciente (por ejemplo, el padre del paciente, por "padre tiene diabetes") o el uso de un término podría ser genérico (por ejemplo, "dio un folleto de diabetes"). Cada uno de estos atributos ilustra cómo se puede marcar el valor de "aseercción" de una entidad con nombre.

5.2.1.6 Analizador sintáctico (Chunker)

Permite dividir un texto en partes de palabras sintácticamente correlacionadas, como grupos nominales, grupos verbales, si especificar su estructura interna, ni su rol específico en la oración principal.

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED	DATE	30/06/2017

5.2.1.7 Segmentador dependiente del contexto (Context dependent tokenizer)

Crea anotaciones de uno o más tokens, usando tokens circundantes como pistas. Un ejemplo de una anotación creada a partir de múltiples tokens es un intervalo que incluye dos números y un guion (por ejemplo, 2-3).

5.2.1.8 Analizador de dependencia (Dependency parser)

Los analizadores de dependencia proporcionan información sintáctica sobre oraciones. Encuentran las dependencias entre palabras. Por ejemplo, la estructura de dependencia del texto "terapia de reemplazo hormonal" mostraría que la palabra "hormona" depende de la palabra "reemplazo" y esta, a su vez, depende de "terapia".

5.2.1.9 Anotador de búsqueda de diccionario (Dictionary lookup annotator)

El anotador de búsqueda de diccionario encuentra las entradas de uno o más diccionarios que coinciden con el texto del documento de alguna manera. Puede buscar coincidencias donde las palabras de las entradas del diccionario aparecen en el mismo orden que las palabras del texto del documento, o puede buscar permutaciones de las palabras del diccionario. Además, puede buscar sólo coincidencias exactas de las palabras, o también puede buscar coincidencias con las formas canónicas de las palabras.


cTakes® incluye diccionarios UMLS (SNOMED CT y RxNorm). Para utilizar esos diccionarios, se dispone de un nombre de usuario y una contraseña UMLS y una conexión a Internet (para verificar su nombre de usuario y contraseña UMLS). Como alternativa, es posible crear específicos dependiendo del ámbito en el que se trabaje.

5.2.1.10 Generador de variantes léxicas (LVG)

Este anotador basa su funcionamiento en la utilización de las herramientas léxicas del ESPECIALISTA de la Biblioteca Nacional de Medicina (NLM). Genera una forma canónica para las palabras, agregando variantes que la búsqueda del diccionario utilizará para intentar descubrir términos cuya forma en el texto no esté presente en la base de datos del diccionario. Por ejemplo, las variantes singulares de formas plurales, variantes de mayúsculas, etc.

5.2.1.11 Etiquetador gramatical (Part-of-speech tagger)

Es el componente encargado de asignar a cada palabra del texto su categoría gramatical (nombre, verbo, etc). El etiquetador gramatical de cTakes® está implementado a partir del mismo componente de Apache OpenNLP al cual se le ha proporcionado un envoltorio UIMA que permite facilitar su utilización. Además, se ha adaptado para funcionar con el sistema de tipos de cTakes® agregándoles características y componentes adicionales de soporte.

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	<i>1.0</i>
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

El siguiente diagrama muestra la mayoría de componentes que integran el sistema y la relación de dependencia existentes entre ellos.

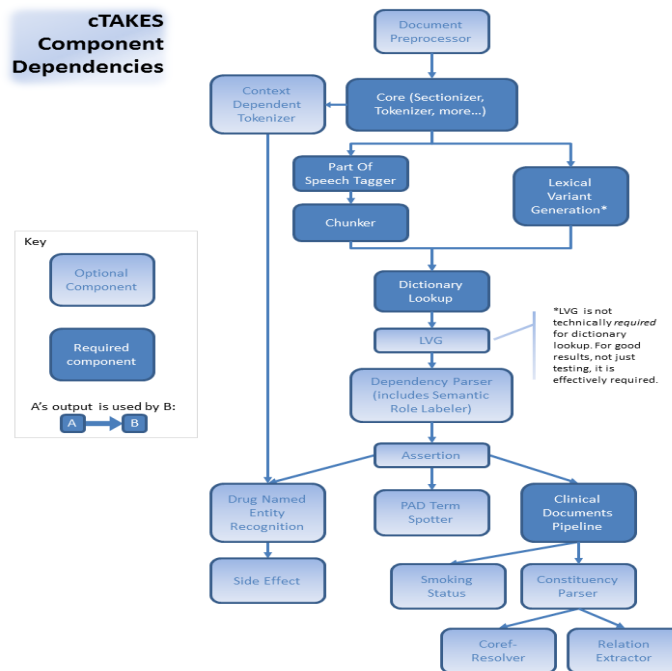


Imagen 2 - Componentes que integran cTakes®

5.2.2 Sistema para la detección de terminología clínica de textos en español

El sistema es una implementación basada en Apache Ctakes©. Tomando como punto de partida algunos de sus componentes, se realiza una adaptación de los mismos para realizar el análisis de textos clínicos en español. El resultado final es un componente agregado que produce una estructura de datos que incluye todos los términos clínicos reconocidos en el texto.

6 Metodología

Antes de detallar el método implementado, se hará un breve resumen del camino recorrido hasta llegar a la mismo.

El proceso de investigación tuvo su comienzo con el estudio de diferentes ontologías⁶, como SNOMED CT⁷, MESH⁸, GALEN⁹, ICD¹⁰ y la relación entre ellas, buscando

⁶(Gruber,1993): “Una ontología es una especificación explícita de una conceptualización”

⁷SNOMED CT: actualmente es la mejor terminología clínica multilingüaje.

[<http://www.snomed.org/snomed-ct/what-is-snomed-ct>]

⁸“MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed.” [<https://www.ncbi.nlm.nih.gov/mesh>]

⁹GALEN: Ontología que contiene conceptos elementales y complejos junto con las relaciones entre ellos. [http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0213-91112008000500006]

¹⁰ICD: Es una clasificación de enfermedades y sirve también para indexar y recuperar información clínica de algunas fuentes bibliográficas.

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

encontrar la mejor manera de representar terminología clínica que luego pueda servir de entrada como parámetros de búsquedas en distintas bibliotecas virtuales de interés.

En principio el estudio derivó en SNOMED CT. El mapeo de conceptos clínicos a términos en SNOMED CT fue factible, pero la segunda instancia, utilizar estos mapeos como parámetros de entrada en las consultas que recuperan información en fuentes bibliográficas, no fue posible. El motivo de esta imposibilidad, es que los métodos de búsqueda de las fuentes investigadas, no aceptan términos de SNOMED CT como entrada para una búsqueda. Existe un intento de Pubmed para realizar este mapeo, pero la solución propuesta, a través de una planilla de cálculo que mapee términos de Pubmed con términos de SNOMED CT, no es de utilidad para este trabajo.

La continuación del estudio presentó a MESH. La elección de MESH tuvo que ver con la elección de PUBMED como fuente de información a ser consultada. Por definición MESH es el tesoro de la Biblioteca Nacional de los Estados Unidos¹¹, un vocabulario controlado que contiene descriptores para indexar artículos en PUBMED.

Luego de haber definido que codificación se utilizaría como parámetros de entrada para las consultas que más adelante arrojarían como resultados una lista de artículos de PUBMED, se investigó, las diferentes maneras de consultar el material de esta fuente de información, intentado encontrar el método más efectivo en cuanto a resultados arrojados.


En este sentido se hizo hincapié en primera instancia en el lenguaje SPARQL¹². Éste es un lenguaje estandarizado que permite consultar grafos *Resource Description Framework*¹³. Las consultas realizadas desde SPARQL sobre fuentes de información, como por ejemplo PUBMED, permiten acceder a la información publicada en dicha librería. El estándar *Resource Description Framework (RDF)* es un lenguaje de descripción que permite capturar la semántica de los datos representados en la WEB. Esta forma de estructurar datos permitió establecer consultas utilizando el lenguaje semántico SPARQL. Con una estructura similar a SQL, SPARQL, contiene herramientas que permiten gestionar datos de múltiples fuentes distribuidas. Además, permite realizar conjunciones y disyunciones de grafos. Otra característica importante es que acepta colocar restricciones sobre valores que pueden figurar como resultados o establecer búsquedas con restricciones a grafos de recursos determinados.

El trabajo con SPARQL se basó en buscar editores de consultas de SPARQL, como por ejemplo VIRTUOSO¹⁴, que permiten ejecutar consultas SQL a PUBMED y estudiar los resultados obtenidos.

¹¹La **Biblioteca Nacional de Medicina de Estados Unidos**, con sede en Rockville Pike, Bethesda, Maryland, Estados Unidos es la biblioteca con el material médico más grande del mundo. [<https://www.nlm.nih.gov/>]

¹²Lenguaje estandarizado que permite consultar grafos que cumplan el estándar *Resource Description Framework*. [<http://sparql.org/>]

¹³RDF es un lenguaje que permite capturar la semántica de los datos representados en la web [<https://www.w3.org/RDF/>]

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

La siguiente es una consulta realizada con SPARQL, utilizando su sintaxis:

```


PREFIX base: <http://RDFEutilsWrapper#>
SELECT ?pubmedUID ?geneUID ?pubmedUID2
WHERE {
  ?pubmed base:pubmed_UID ?pubmedUID.
  ?pubmed base:pubmed_TITL ?title.
  ?pubmed base:pubmed_gene ?gene.
  ?gene base:gene_UID ?geneUID.
  ?gene base:gene_pubmed ?pubmed2.
  ?pubmed2 base:pubmed_UID ?pubmedUID2.
  FILTER (?title = "wilms tumor").
}

```

En este ejemplo, la búsqueda se centra en artículos de Pubmed cuya temática sean los genes.

Durante la investigación realizada para este trabajo, se pusieron a prueba varios EndPoint de SPARQL. Las respuestas obtenidas fueron desalentadoras. En una gran cantidad de casos, la metodología de trabajo con estos EndPoint se basaba en instalar la plataforma de manera local, es decir, en una computadora propia, y luego descargar los fuentes RDF de PUBMED, para así hacer consultas sobre esta fuente. La principal desventaja de esta metodología es que los fuentes RDF se encuentran siempre desactualizados y las consultas no se hacen online. También muchos EndPoint, a pesar de manifestar que soportan PUBMED entre las fuentes a consultar, no emitían resultados alguno ante las consultas realizadas o desplegaban algún tipo de error por no soportar PUBMED, a pesar de mencionar lo contrario. Sumada a esta problemática, muchos EndPoint, contienen escasa información o RDF incompletos que impiden realizar consultas. Si bien la bibliografía muestra varios proyectos en donde SPARQL tiene participación, muchos están discontinuados y otros en versiones beta, dado este panorama se decidió descartar momentáneamente SPARQL para este trabajo final, aunque la línea de investigación continuará con un artículo incluido.

¹⁴“The Virtuoso SPARQL query service originally implemented the SPARQL Protocol for RDF (W3C Recommendation, 15 January 2008), and has been updated to support SPARQL 1.1, providing SPARQL query-processing for RDF data available on the open Internet. The query service implementation extends the standard protocol by providing multiple output-formats alongside the standard XML results serialization. At present this uses additional query parameters, although content-negotiation will be supported soon”.
[\[https://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSSparqlProtocol\]](https://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSSparqlProtocol)

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

Descartado SPARQL, se procedió a buscar alternativas para determinar métodos de búsquedas de evidencia científica en PUBMED. Así se llegó a Entrez Programming Utilities ([E-utilities](#)), un conjunto de programas que permiten consultar información sobre varias fuentes de datos, incluida PUBMED. Lo interesante de este conjunto de programas es que, permite enviar parámetros, como por ejemplo término buscado, rango de edad, fuente de datos, etc., desde la URL, lo cual quiere decir que, la consulta se arma a través de una serie de parámetros preestablecidos que se especifican por método GET en la URL que luego se ejecuta, dando como resultado un archivo XML¹⁵ con una lista de índices a los artículos encontrados en PUBMED. Más adelante se mostrará un ejemplo del funcionamiento.


Definido el camino recorrido de manera muy breve, a continuación se describe la metodología diseñada en este Trabajo Final de Master. Como se mencionó en apartados anteriores, el objetivo de este trabajo final, es mostrar una metodología de recomendaciones de artículos científicos publicados en PUBMED, tomando como parámetros de búsqueda los datos vertidos por los profesionales médicos en el registro clínico electrónico.

6.1 Descripción Macro del Método

- *Registro clínico electrónico.* El profesional médico hace una evaluación del paciente y en lenguaje natural expresa diagnósticos, hallazgos clínicos, resultados anormales, signos o síntomas encontrados durante la evaluación al paciente. Todos estos aspectos se registran en el registro clínico electrónico.
- *Extracción de hallazgos clínicos.* Mediante herramientas de procesamiento de lenguaje natural, se extraen los hallazgos clínicos redactados por el médico en el registro clínico electrónico. Los mismos son devueltos en un archivo XML.
- *Obtención de términos MESH.* Los términos reconocidos luego de aplicar la herramienta de procesamiento de lenguaje natural, son utilizados como parámetros de entrada en bioontology¹⁶. La salida es una lista de términos MESH.
- *Parametrizar la consulta.* Con los hallazgos clínicos determinados, se procede a crear la URL que será ejecutada en pos de buscar los artículos en PUBMED.
- *Ejecución de la consulta.* Utilizando esearch.fcgi, del conjunto de programas Entrez Programming Utilities, se procede a ejecutar la consulta pertinente.
- *Obtención de resultados.* Los resultados de ejecutar la consulta sobre PUBMED son enviados en formato XML (existe la posibilidad de obtenerlo en otros formatos). El archivo contiene una lista de índices de artículos a PUBMED.

¹⁵Meta-lenguaje que permite definir lenguajes de marcado adecuados a usos particulares. Permite definir etiquetas personalizadas para descripción y organización de datos.

¹⁶Es un repositorio de ontologías médicas que permite encontrar las referencias de unas con otras, por ejemplo se podría buscar MELANOMA dentro de SNOMED CT y luego ver con que ontologías se relaciona desde SNOMED CT. En este caso es utilizado para buscar mediante el cui el término MESH correspondiente. [<http://bioportal.bioontology.org/>]

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	<i>1.0</i>
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

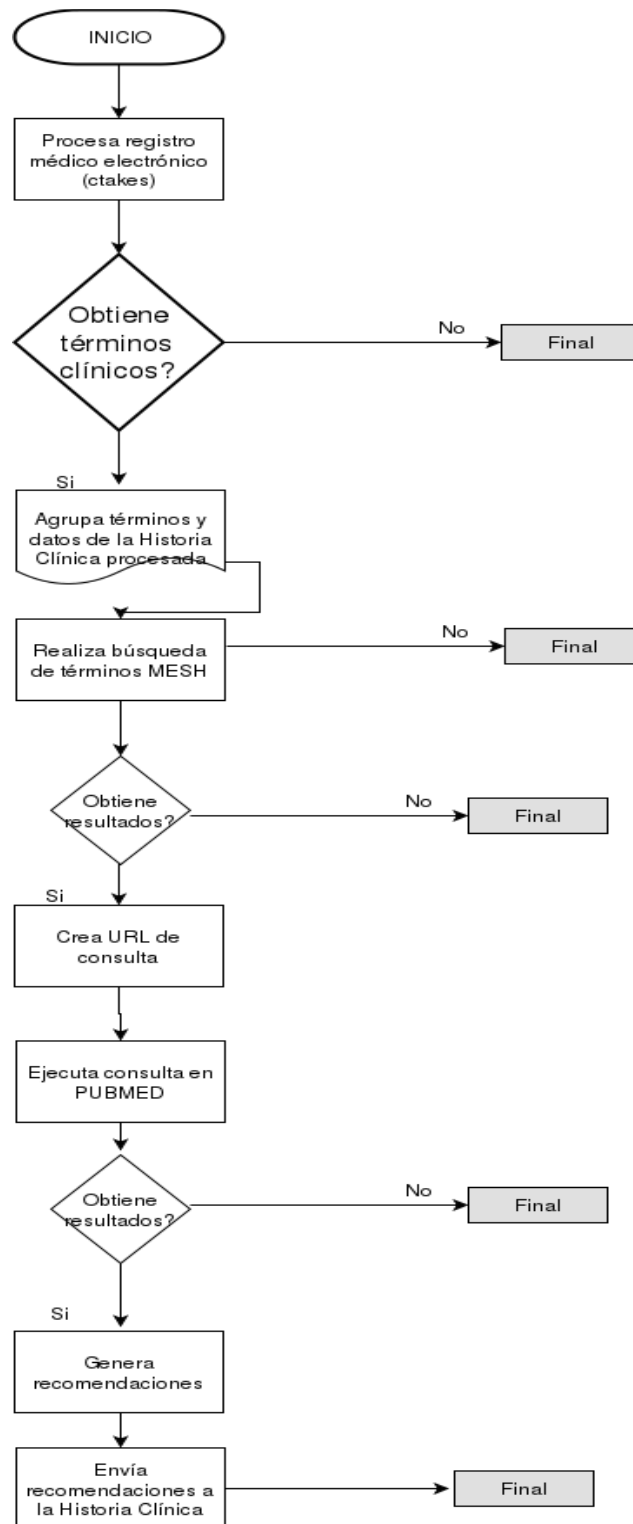

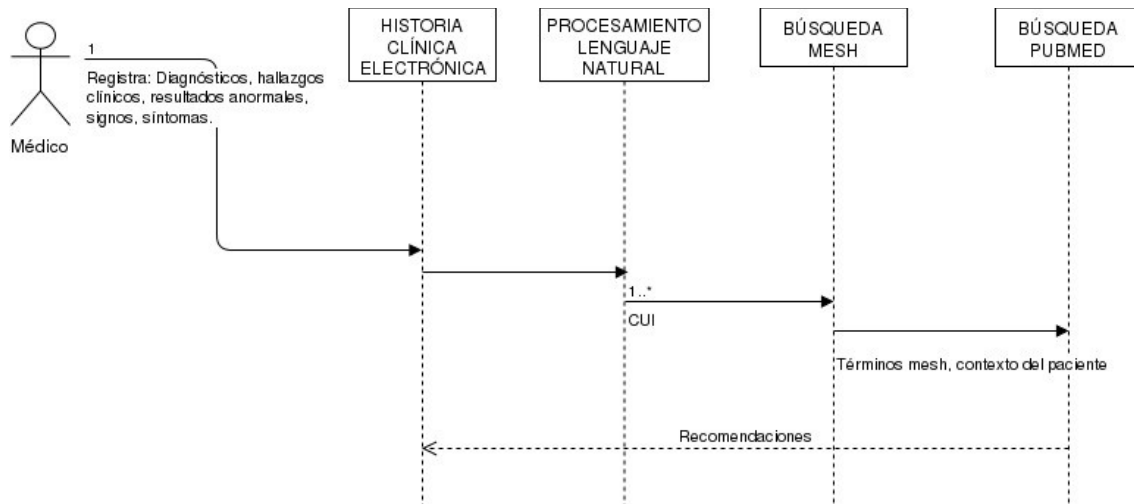


Diagrama de flujo desde la Historia Clínica a las recomendaciones

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

6.2 Método: Esquema de funcionamiento



6.2.1 En funcionamiento

En este apartado se hará una descripción más detallada de la metodología empleada en base a un caso real establecido. Se parte de la base, que los términos vertidos por el profesional médico en el registro clínico electrónico ya han sido capturados y solo falta obtener el término MESH que mejor se adapta al dato introducido por el médico.

Pasos:


1. El profesional médico registra en el registro clínico electrónico la evaluación realizada al paciente. Para el ejemplo entre los hallazgos realizados por el profesional se encuentra el término MELANOMA.
2. Del resultado devuelto por la herramienta de procesamiento de lenguaje natural, se obtiene un código denominado *cui* (Common Unique Identifier)¹⁷. Para el ejemplo el cui es: C0025202.
3. Con el identificador *cui* más la terminología seleccionada MESH, se busca el mejor término MESH asociado a ese identificador *cui*. Para esto se ejecuta lo siguiente:

¹⁷Los cui son los identificadores utilizados en UMLS para representar conceptos. UMLS permite representar términos clínicos sin importar la fuente desde donde provengan, de una única manera, por lo tanto, cada término en UMLS tiene asignado un único cui.

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED	DATE	30/06/2017

1. <http://data.bioontology.org/search?q=C0025202&ontologies=MESH&categories=http://data.bioontology.org/categories/Health&apikey=ae7a2cb3-d230-445b-ba34-d71cb94a5b5c>.
4. El resultado de lo anterior es devuelto en formato JSON [xx <https://es.wikipedia.org/wiki/JSON>] y en el campo ['collection'][0]['prefLabel'] se encuentra el término MESH más apropiado.
5. Luego con el término MESH adecuado se procede a ejecutar la consulta haciendo uso de E-Utilities de la siguiente manera:
 1. [https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&term=". \\$data\['collection'\]\[0\]\['prefLabel'\]](https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&term=)¹⁸
6. Los resultados obtenidos de la consulta anterior son devueltos en formato XML y los mismos representan una lista con índices a artículos de PUBMED.

¹⁸La consulta final puede contener más parámetros, como rango de edad, actualidad de los artículos, autor, títulos, referencias etc.

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED	DATE	30/06/2017

7 Resultados

La prueba de resultados se hizo de manera manual, dejando para más adelante la puesta en marcha de un piloto, cuyos resultados se verán reflejados en un próximo artículo.

Para las pruebas manuales se identificaron una serie de términos MESH los cuales estaban asociados a conceptos clínicos seleccionados de una historia clínica. Luego de armar la consulta pertinente y posteriormente por medio del programa *esearch.fcgi*, realizar la prueba para obtener artículos de pubmed, se concluye que dichas evaluaciones han sido satisfactorias. Se ha comprobado una reducción de artículos reportados, dado que la indexación por medio de MESH tiene un grado de precisión mayor que si se busca por palabras claves que no indexan.

Además, las restricciones introducidas como parámetros, han mostrado no solo menos artículos referenciados, descartados por no ser el tema de interés buscado, sino también mayor cantidad de artículos actualizados, descartando muchos otros artículos que ya han sido reemplazados en cuanto a evidencia científica refiere. Esto se ha logrado colocando las restricciones correspondientes en consulta de búsqueda.

Se detecta un aumento en la cantidad de artículos de relevancia entregados en la consulta. Al afinar la estrategia de búsqueda se logra extraer mayor cantidad de artículos de relevancia. Los datos adquiridos desde la historia clínica, contexto del profesional médico, colaboran con la mejora de la consulta a realizar.

Por último simplemente mencionar que se montará un piloto para realizar más pruebas con esta nueva modalidad que se describirá incluso en un artículo. La idea del piloto será simular un entorno asistencial real y ver que recomendaciones realiza el método descrito.

A continuación de su muestra un ejemplo de dos consultas armadas de manera manual, para mostrar los resultados obtenidos con cada una.


En la primera consulta se puede ver la búsqueda del término melanoma, sin ningún tipo de restricción, solo el término. Los resultados obtenidos superan los 110,000 artículos.

Consulta:

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=melanoma&retmode=json>

Resultado:

```
"header": {
  "type": "esearch",
  "version": "0.3"
},
```


	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

```

"esearchresult": {
  "count": "113653",
  "retmax": "20",
  "retstart": "0",
  "idlist": [
    "26389388",
    "28632887",
    "28632800",
    "28632796",
    "28632148",
    "28631824",
    "28631119",
    "28631017",
    "28630682",
    "28630590",
    "28630054",
    "28630039",
    "28629895",
    "28629701",
    "28629629",
    "28629479",
    "28629213",
    "28626656",
    "28628832",
    "28628690"
  ]
}

```

Esta segunda consulta, incorpora dos aspectos mencionados durante el desarrollo de este trabajo. El primero, indica que el término que se está utilizando es un término MESH. Por otro lado, solicita que los resultados de la consulta, solo involucren artículos de Melanoma en adultos. Los resultados obtenidos son un poco más de 19,000

Consulta:


[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=melanoma\[mesh\]%20AND%20Adult\[sb\]&retmode=json](https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=melanoma[mesh]%20AND%20Adult[sb]&retmode=json)

Resultados:

```

"header": {
  "type": "esearch",
  "version": "0.3"
},
"esearchresult": {
  "count": "19671",
  "retmax": "20",
  "retstart": "0",
  "idlist": [
    "28591523",
    "28466787",
    "28413057",
    "28408015",
    "28404758",
  ]
}

```

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

"28396940",
 "28369155",
 "28360400",
 "28335082",
 "28331853",
 "28314304",
 "28314298",
 "28314272",
 "28289863",
 "28284557",
 "28283736",
 "28274670",
 "28268064",
 "28253243",
 "28237198"

Todavía son muchos resultados, pero también se puede ajustar aún más la estrategia de búsqueda, para ser más específicos en los requerimientos y más exactos en los resultados.

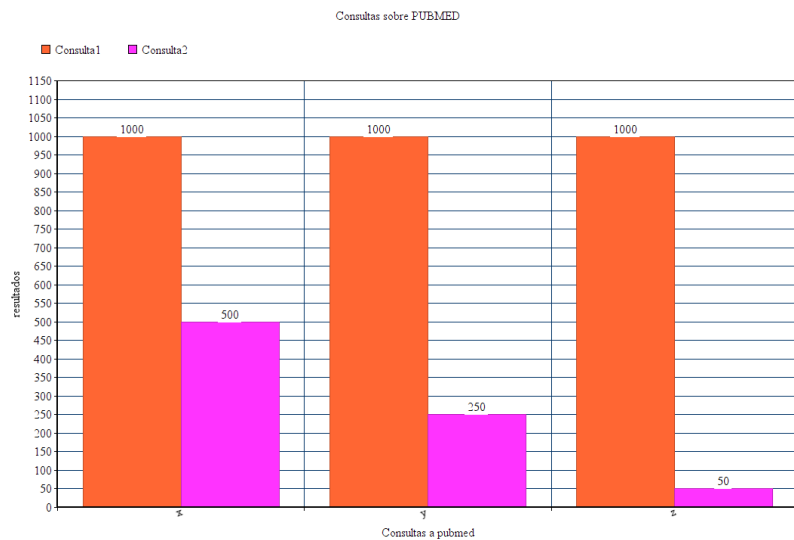




Diagrama generado de manera manual a modo explicativo

El diagrama anterior muestra dos consultas realizadas a PUBMED, la diferencia entre ambas consulta son, en el caso de la consulta1 (color anaranjado), la misma se realiza siempre con el mismo conjunto de palabras claves utilizando el método tradicional que PUBMED ofrece. La segunda consulta, consulta2 (color violeta), la consulta se realiza utilizando el método descripto, añadiendo datos en la historia clínica que luego sirven como parámetro para la consulta. Los resultados obtenidos, muestran una disminución de artículos en cada momento de la consulta a medida que los parámetros para

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	<i>1.0</i>
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED	DATE	<i>30/06/2017</i>

parametrizar la consulta2 aumentan. Lo que representa es un acercamiento a los artículos de interés del profesional médico.

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

8 Discusión


El objetivo de este trabajo ha sido diseñar una metodología de búsqueda de recomendaciones de artículos científicos a través de Pubmed partiendo de términos extraídos de la historia clínica del paciente, en combinación con otros datos de esta misma fuente. El colectivo al que está dirigido son profesionales médicos, dedicados a la docencia, investigación y atención sanitaria.

Desde el punto de vista tecnológico, el estudio se ha basado en el uso de la librería E-utilities, puesta a disposición por Pubmed, como herramienta de búsqueda y el uso de ctakes, como herramienta de procesamiento de lenguaje natural, para extraer términos clínicos vertidos por el médico en las evoluciones que realiza en la historia clínica del paciente. También a nivel tecnológico, se ha hecho uso de servicios como el que proporciona bioontology.org, con el cual se es capaz de obtener los términos MESH que mejor se adapten a los conceptos encontrados en la historia clínica.

Una vez planteados los objetivos y alcances, se ha llevado a cabo una investigación de las diferentes fuentes de información y métodos de búsquedas que ofrecen, como también de las necesidades de los médicos y las problemáticas a la que están sometidos dado el contexto planteado. Se han tenido en cuenta fuentes de información tanto públicas como privadas y se ha hecho uso de los distintos métodos de búsquedas encontrados.

Incluido en este análisis y estudio se encuentra, la investigación sobre SPARQL. Este lenguaje, que ha presentado características interesantes, como la posibilidad de realizar consultas SQL sobre varias fuentes de información, incluso combinando diversas fuentes en una misma consulta. La introducción de restricciones sobre ciertos parámetros para ajustar aún más la consulta y lograr resultados más precisos y acotados según la búsqueda requerida, también son aspectos relevantes de este lenguaje. Además de las características mencionadas sobre SPARQL, los grafos RDF, utilizados por este lenguaje, tienen un rol importante sobre las consultas que se pueden especificar sobre fuentes como Pubmed. Estos grafos, formados por lo que se denomina triples RDF (cada RDF está formado por un sujeto, un predicado y un objeto), se los puede ver como un conjunto de ternas y tiene un enfoque similar de modelado conceptual de un modelo entidad relación. Esta característica, permite realizar consultas sobre fuentes de datos con una estructura similar al lenguaje SQL, con toda la potencia que este aspecto refiere.


A partir de los resultados obtenidos se determina que Pubmed ofrece una gran cantidad y variedad de artículos científicos de mucha utilidad para el profesional médico y que la herramienta de búsqueda E-utilities da la posibilidad de expresar con mayor precisión la consulta sobre determinado tema de interés. Además, permite parametrizar la búsqueda, pudiendo declarar datos importantes que provienen de la historia clínica electrónica, como por ejemplo la edad del paciente y el sexo. Esta solución dispone de todos los componentes necesarios para desarrollar una metodología de búsquedas de

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED	DATE	30/06/2017

recomendaciones para el profesional médico desde la historia clínica electrónica del paciente.


9 Trabajos Futuros

Como líneas futuras para este trabajo, actualmente se está trabajando en una mayor evaluación del método establecido, trabajando con un piloto y se está evaluando la presentación de este trabajo y sus resultados en BMC Medical Informatics (Technical Advance Paper), JAIMA y Methods.

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED	DATE	30/06/2017

10 Referencias y Bibliografía:

- BMJ 2006;333:1283: Dean Giustini, How Web 2.0 is changing medicine, <http://www.bmj.com/content/333/7582/1283>
- DBSI: Yanetsys Sarduy Domínguez, Directorio de buscadores de salud en Internet, http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21252006000100005
- MBSSI: Reinaldo Rodríguez Camiño, Motores de búsqueda sobre salud en Internet, http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352003000500002
- Lawrence, Giles 1999 Bollacker: K. Bollacker, S. Lawrence, and C. L. Giles., Discovering relevant scientific literature on the web, IEEE Intelligent Systems 15(2), pp. 42-47.
- FICSI: Rafael Aleixandre-Benavent, Fuentes de información en ciencias de la salud en Internet, <http://www.medtrad.org/panacea/IndiceGeneral/n33-Ponencias-Aleixandre.pdf>
- ERMHI: Luna D, Otero P, Gómez A, González Bernaldo de Quirós F, El Registro Médico: de Hipócrates a Internet, https://www.hospitalitaliano.org.ar/multimedia/archivos/servicios_attachs/1151.pdf
- CEBM: , CEBM, ,
- BMJ 1996; 312:71-2: Sackett DL, Rosenberg W, Muir JA, Haynes RB, Richardson WS, Evidence based medicine: what it is and what it isn't, IEMORDSC: Arndt KA, Information excess in medicine. Overview, relevance to dermatology, and strategies for coping, Arch Dermatol. 1992 Sep;128(9):1249-56
- ASSEHS: Trivedi, M, A study of search engines for health sciences. Library and Information Science, 1(5), 069-073

	MASTER UNIVERSITARIO DE TELEMEDICINA		VERSIÓN	1.0
	TÍTULO	TRABAJO FIN DE MASTER Recomendaciones de artículos a través de PUBMED		DATE

11 Anexo: Consultas a PUBMED con E-Utilities y SPARQL

11.1 Consultas con SPARQL

Despliega los artículos PUBMED que hacen referencia al descriptor D017966 (Pyramidal Cells)

```
select ?article ?title
where
{
  { ?pubmed sc:has-as-major-mesh mesh:D017966 .
    ?article sc:identified_by_pmid ?pubmed.
    ?article dc:title ?title
  }
}
```

Despliega artículos de PUBMED del autor Arjona

```
PREFIX v: <http://vocabulario:>
SELECT ?article ?article_title
{
  ?article ?p ?author .
  ?article rdfs:label ?article_title .
  ?article a v:PubMedRecord .
  ?author a v:Author .
  ?author v:last_name ?ln .
  ?author v:initials ?in .
  ?ln bif:contains "Arjona" .
}
```

11.2 Consultas con E-Utilities

Despliega las referencias a artículos de PUBMED del término Pyramidal Cells

[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=Pyramidal%20Cells\[mesh\]%20AND%20Adult\[sb\]&retmode=json](https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=Pyramidal%20Cells[mesh]%20AND%20Adult[sb]&retmode=json)

Despliega las referencias a artículos de PUBMED del término Pyramidal Cells publicados entre el año 2016 y Junio del 2017.

[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=Pyramidal%20Cells\[mesh\]%20AND%20Adult\[sb\]&retmode=json&mindate=2016/01/01&maxdate=2017/06/27](https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=Pyramidal%20Cells[mesh]%20AND%20Adult[sb]&retmode=json&mindate=2016/01/01&maxdate=2017/06/27)