ARTÍCULO ORIGINAL

Toma de decisiones inteligente a partir de registros médicos almacenados en CDA-HL7

Intelligent decision-making from medical records stored on the CDA-HL7

Ivett E. Fuentes Herrera,[I] Damny Magdaleno Guevara,[II] María Matilde García Lorenzo[II]

I Centro de Estudios de Informática, Facultad de Matemática-Física-Computación, Universidad Central "Marta Abreu" de Las Villas, Cuba. E-mail: ivett@uclv.cu, Carretera a Camajuaní, Km 5½. Santa Clara, Villa Clara, CP 54830
II Centro de Estudios de Informática, Universidad Central "Marta Abreu" de Las Villas, Cuba.

RESUMEN

Debido al incremento exponencial de la información almacenada en las organizaciones, la Sociedad de la Información está siendo superada por la necesidad de nuevos métodos capaces de procesar la información y asegurar su uso productivo. Esto se hace lógicamente extensible a los centros hospitalarios, a partir del uso extendido de las Historias Clínicas en formato electrónico. Disponer de información sistematizada, gestionarla de forma eficiente y segura es esencial para garantizar mejores prácticas en salud. A esto se le añade la necesidad de soportar estándares que permitan el intercambio entre las instituciones de salud; específicamente HL7 se ha convertido en uno de los más utilizados debido a que proporciona el intercambio a partir del metalenguaje XML. En este trabajo se propone una metodología para el descubrimiento de conocimiento implícito en Historias Clínicas en formato semi-estructurado utilizando el contenido y la estructura de los mismos. Los principales resultados son: (1) La metodología para el agrupamiento de Historias Clínicas; (2) La interpretación de los resultados del agrupamiento para asistir la toma de decisiones diagnósticas; (3) La implementación del estándar HL7, para la manipulación de documentos médicos a partir de CDA.

Palabras Clave: agrupamiento, descubrimiento de Conocimiento, HCE, XML, CDA.

ABSTRACT

Due to the exponential increase of stored information the organizations, the information society is being overtaken by the need for new methods capable of processing information and ensuring its productive use. This is logically extended to the hospitals, from the widespread use of clinical histories in electronic format. To have systematized information, manage it efficiently and securely is essential to ensure better health practices. In addition, there is the need for standards to support the exchange among health institutions; specifically hl7 has become one of the most widely used because it provides the exchange from xml. In this paper is presented a methodology for the discovery of implicit knowledge in medical records with semi-structured format, using their content and structure. The main results are: (1) the methodology for the clustering of medical records; (2) the interpretation of the results of the clustering to assist diagnostic decision-making; (3) the implementation of the hl7 standard for handling medical records from cda.

Key words: clustering, knowledge discovering, EMR, XML, CDA.

INTRODUCCIÓN

En el actual siglo XXI, el mundo desarrollado se ha propuesto lograr la globalización del acceso a los enormes volúmenes de información existentes en medios cada vez más complejos, con capacidades exponencialmente crecientes de almacenamiento y en soportes cada vez más reducidos, por lo que nadie pone en duda el papel que juegan las nuevas Tecnologías de la Información y las Comunicaciones (TIC) en las organizaciones. Esto se hace, lógicamente, extensible a la Gestión Documental (GD), que administra el flujo de documentos e información en las instituciones . Surge así, la necesidad de capturar y conservar también documentos que nacen, viven y mueren en formato electrónico, principalmente en las organizaciones en las que se pueden realizar estudios basados en experiencias anteriores. Particularmente, disponer de herramientas que permitan relacionar documentos y darles una semántica común, constituye una posibilidad de sistematizar la enorme riqueza de datos e información que reside en los sistemas hospitalarios en entornos educativos y de investigación.[1]

A lo anterior se añade que actualmente en nuestro país, los sistemas de registro de Historias Clínicas (HC) en las instituciones de salud son ineficientes, no son automatizados, ocupan mucho espacio físico y no vinculan la información del paciente con los demás hospitales, clínicas u otros organismos de atención de salud pública. Es decir, las HC que se registran no son universales. Por lo tanto, es imperativo que el Sistema de Salud del país haga la transición del registro de HC en formato duro a un registro de Historias Clínicas Electrónicas (HCE), donde cada paciente tenga una HC única y con carácter universal. Además en el contexto mundial, los usos de la HCE cada día impactan de manera creciente y favorable en la investigación clínica, en la investigación farmacéutica (dígase ensayos clínicos, fármaco-epidemiológicos) y en las investigaciones en salud pública (dígase

informes electrónicos de casos, bases de datos poblacionales), entre otros.[2,3] Como consecuencia, la creación de repositorios de HCE y el volumen de información generada desde estos, debe aumentar continua y exponencialmente. La automatización de la información clínica proporciona enormes ventajas tanto para los profesionales de la salud, como para los pacientes, y el estado. Es así que surgen nuevas oportunidades para adoptar herramientas que apoyen la toma de decisiones en la práctica clínica y proporcionar mejores condiciones de trabajo a los médicos, contribuir al ejercicio de una medicina basada en la evidencia y asegurar el uso productivo de la información almacenada.[4,5] En este sentido, aunque se han desarrollado varios sistemas con el propósito de lograr una rápida y eficiente manera de compartir información, la heterogeneidad de ella determina que extraer conocimiento relevante se convierta en un proceso complejo y desafiante.[5]

Es reconocido por varios investigadores[1,4,6,7] la importancia de la estandarización y codificación de datos almacenados en la HCE, así como la necesidad de migrar a una recopilación de la información clínica mediante texto estructurado. Particularmente en el contexto de la HC, la propia distribución de sus elementos hace posible concebirla como un documento XML, debido a la estructura jerárquica y auto-descriptiva implícita en cada uno de los factores que la componen. De hecho, Health Level Seven (HL7), el conjunto de estándares informáticos de salud más desarrollado y de mayor cobertura internacional para dar soporte a la HCE está sustentado en el metalenguaje XML.[4]

Extensible Markup Language (XML) es un metalenguaje desarrollado por el consorcio W3C2 proveniente de Generalized Markup Language (GML) que surgió ante la necesidad de la gran empresa de almacenar grandes cantidades de información. Un documento XML es una estructura jerárquica auto-descriptiva de información, que consiste en un conjunto de átomos, elementos compuestos y atributos.[2] A esto se añade que los documentos XML contienen su información en forma semi-estructurada, al incorporar estructura y datos en una misma entidad. Son extensibles, con estructura de fácil análisis y procesamiento, por lo que XML se ha convertido en el formato de intercambio de datos estándar entre las aplicaciones Web.[2,8,9] Las etiquetas existentes en los documentos XML permiten la descripción semántica del contenido de sus elementos. De este modo, la estructura de los documentos puede ser explotada para realizar recuperación de documentos relevantes.[5]

Por lo anteriormente planteado se hace inevitable crear técnicas para el análisis eficiente de grandes colecciones de este tipo de documentos y extraer conocimiento relevante. Existen tres enfoques para abordarlo: la clasificación, la categorización y el agrupamiento; en este último varios investigadores se han concentrado, debido a que exclusivamente, el agrupamiento de documentos XML permite organizar la información, delimitar la información relevante y descubrir nuevo conocimiento a partir de la información disponible en una colección obtenida como resultado de un proceso de recuperación de información.[3,10-12]

Un algoritmo de agrupamiento intenta encontrar grupos naturales de datos, basándose principalmente en la similitud y las relaciones de los objetos, de forma tal que se obtenga una distribución interna del conjunto de datos en grupos. Cuando el agrupamiento se basa en la similitud de los objetos, se desea que los objetos que pertenecen al mismo grupo sean tan similares como se pueda y los objetos que pertenecen a grupos diferentes sean tan disímiles como sea posible. El análisis de grupos permite descubrir una estructura previamente oculta en los datos, sin embargo, la asignación de los objetos a las clases y la descripción de esas clases son desconocidas.[13] El desarrollo de sistemas que faciliten a los usuarios gestionar grandes colecciones de documentos, mediante la organización y

extracción del conocimiento es una necesidad real. Explotar la estructura específica que tienen las HC puede ofrecer resultados favorables en el agrupamiento de este tipo de documentos, y contribuir de manera significativa a la gestión del conocimiento.

Cuando se trata de documentos XML, los algoritmos de agrupamiento se clasifican principalmente en tres grupos: los que se centran sólo en el contenido de los documentos, sin embargo un buen proceso de agrupamiento no puede descartar el uso de la estructura, por lo que están los algoritmos que utilizan sólo la estructura, considerando que esta juega un papel importante en el agrupamiento para ciertas aplicaciones específicas y los que combinan ambas componentes: estructura y contenido, lo cual constituye un nuevo desafío, ya que la mayoría de los enfoques existentes no utilizan estas dos dimensiones dada su gran complejidad.[14]

Lo antes expuesto ratifica la siguiente problemática:

Aunque como ya se mencionó, existen varias formas de gestionar el conocimiento: la categorización, la clasificación y el agrupamiento, los sistemas de información hospitalarios no implementan mecanismos que garanticen el uso productivo de la información clínica para asistir la toma de decisiones diagnósticas.

En este trabajo se presentan los resultados siguientes: (1) La metodología para el agrupamiento de Historias Clínicas; (2) La interpretación de los resultados del agrupamiento para asistir la toma de decisiones diagnósticas; (3) La implementación del estándar HL7 para la manipulación de documentos médicos a partir de CDA.

El valor práctico del trabajo está enfocado a:

Disponer de un sistema de recuperación de información que soporte la metodología, que permita procesar grandes volúmenes de datos y obtener conocimiento relevante a partir de la información recuperada, con el propósito de asistir a los expertos de salud en el proceso de toma de decisiones diagnósticas, inferir áreas que deben ser exploradas y conducir al desarrollo de investigaciones clínicas, valorar la efectividad terapéutica y la evolución favorable de los pacientes ante un tratamiento.

Información clínica y agrupamiento documental

Cada día más datos electrónicos son presentados debido al crecimiento continuo de información desde múltiples campos y la automatización de gran parte de los procesos de la sociedad. Esto se hace extensible a la gestión de la información clínica, debido al criterio extendido de brindar soporte a la práctica médica facultativa de una medicina basada en la evidencia.[1]

La HC es un documento válido desde el punto de vista clínico y legal a todos los niveles de atención en salud, que recoge información de tipo asistencial, preventivo y social. Es una fuente esencial de datos y constituye el documento principal de un Sistema de Información Hospitalaria (HIS). Es una herramienta básica para las investigaciones biomédicas, la formación de estudiantes y la educación médica postgraduada en la consecución de investigaciones. Constituye, además, el registro completo de la atención prestada al paciente durante su enfermedad, de ahí su trascendencia como documento legal. Es la fuente que, además de recoger todo un informe de salud, comunica el pensamiento médico, registra observaciones, diagnósticos e intervenciones que reflejan uno o varios problemas; sin embargo, su formato tradicional enfrenta diversas dificultades, que se han hecho evidentes

durante la práctica diaria como son: su deterioro o pérdida, debido a que la historia convencional, en su formato de papel, sólo puede existir en un lugar y en un momento y condiciones determinados, así como la presencia de una escritura pobre, ilegible e incompleta que dificulta la interpretación del mensaje que se pretende enviar. Otra de sus limitaciones es que sólo puede contribuir de forma pasiva a la toma de decisiones y esto dificulta el análisis con fines científicos o de planeamiento de estrategias de salud.[6]

La HC incluye la información clínica relacionada con los datos del paciente, antecedentes personales y familiares, hábitos tóxicos y todos los elementos relacionados con su salud biopsicosocial; el proceso evolutivo, el tratamiento y la recuperación.[7] La HC es un documento donde el paciente deja registrado su consentimiento para ser utilizado en la toma de decisiones del profesional de la salud.

Para gestionarlo existen varios modelos según el lugar donde se genera: cronológico, la historia clínica orientada a problemas (HCOP) y la protocolizada. Algunos componentes de los modelos clásicos de la HC, como la orientada hacia el problema, se consideran especialmente adecuados para su uso con fines educativos y científicos.[6] En este trabajo se propone utilizar el modelo cronológico generado en los hospitales.

En este sentido, se evidencia la necesidad de incorporar las TIC en el núcleo de la actividad hospitalaria, lo que implica ofrecer soporte a la HCE. De manera que, cada HC se convierta en un simple registro de la información, que se integra al HIS de la institución. Sin embargo, la conceptualización y aplicación de las TIC en este ámbito no es un proceso acabado, debido a que existen problemas que limitan el uso productivo de la información: la integración efectiva de la HCE y el uso de herramientas de aprendizaje automático de la inteligencia artificial. A esto se añaden los problemas relacionados con la codificación de términos y el uso de estándares.[1,4]

Usos actuales de HCE en la investigación

Los registros informáticos de los servicios de admisión hospitalarios se utilizan en el desarrollo de investigaciones clínicas y epidemiológicas, en ausencia de otras fuentes de datos clínicos bien estructuradas capaces de generar conocimiento [4]. Por lo que debe potenciarse el desarrollo de herramientas que permitan almacenar adecuadamente esta información y hacerla accesible.[4,7]

Esto implicaría inferir áreas que deben ser interpretadas por los expertos en salud a partir de la información disponible y garantizar su uso productivo con el fin de llevar a cabo investigaciones clínicas, lograr nuevas soluciones diagnósticas y terapéuticas, la evaluación del uso de las tecnologías empleadas en casos de difícil control, el estudio de la evolución de los pacientes, la eficacia y la eficiencia de la atención, la identificación de poblaciones de riesgo, y el análisis de la eficiencia del proceso.

Interoperabilidad y el estándar CDA-HL7

Para lograr este intercambio, es necesario que los sistemas de información utilizados por las instituciones de salud, pongan en práctica las normas reconocidas internacionalmente, especialmente las establecidas por el estándar HL7, por lo que la recopilación de información clínica debe ir migrando al uso controlado de texto estructurado. Uno de los principales aportes de este trabajo es concebir la HCE según el estándar Clinical Document Architecture (CDA) de HL7 que especifica la

estructura y semántica de los documentos clínicos con el propósito de facilitar su intercambio en un entorno de interoperabilidad.

CDA-HL7 es un estándar de marcaje, realizado por el comité Structured Documents Technical Committee (SDTC) de HL7, que permite definir la estructura y la semántica de un documento clínico. Es una especificación que facilita el intercambio entre los diferentes sistemas en las organizaciones al utilizar XML.

CDA logra que los documentos sean computacionalmente más legibles. Gracias a la utilización de XML, Reference Information Model (RIM), la metodología de la versión v3 de HL7 y los vocabularios codificados, los documentos clínicos pueden ser interpretados y procesados automáticamente. Las colecciones de documentos CDA pueden ser presentadas directamente a los navegadores Web compatibles con XML. Es posible crearlos y validarlos mediante una plantilla XML o Schema. Y su diseño tiene como principal propósito ofrecer a los pacientes un mejor servicio. Permite una implementación efectiva y eficiente en su costo en un espectro amplio de sistemas heterogéneos, siendo independiente de la plataforma. Soporta el intercambio de documentos legibles entre los usuarios, permitiendo presentar la información de forma adecuada a usuarios con diferentes requisitos y conocimiento. Por su diseño, facilita un amplio rango de procesamiento, al ser fácilmente compatible con muchas aplicaciones de creación y GD.

Un documento CDA contiene una cabecera y un cuerpo. La cabecera sigue una estructura común, que identifica y clasifica el documento, provee información acerca de la autenticación, paciente, autor y actores involucrados. Por lo que al seguir una estructura común, bien definida, la consulta de estos campos de forma automatizada es fácil.

El cuerpo del documento, puede contener tres niveles de implementación: el nivel más bajo, implica una implementación más sencilla pero no utiliza muchas de las ventajas de la arquitectura CDA. El más alto ofrece una verdadera interoperabilidad semántica, pero implica un esfuerzo más amplio en la implementación y requiere una madurez en los sistemas que generan y capturan los datos de los documentos. El nivel 2 sigue una estructura XML bien definida con secciones de información identificadas, cuyo contenido es libre, lo cual facilita al actor del documento realizar una descripción lógica de cada uno de los elementos que refiere. El nivel 3 agrega a cada sección, y a cada dato dentro de estas secciones (diagnósticos, unidades de medición, medicamentos, etc.) Este nivel tiene muchas ventajas, ya que garantiza la verdadera interoperabilidad semántica, permitiendo que los documentos sean procesables, mediante búsquedas y técnicas de Minería de Datos .

Estas ventajas hacen que CDA sea uno de las especificaciones más utilizadas regionalmente en los sistemas de gestión de información clínica. Por ello en este trabajo, proponemos concebir la HCE como un documento XML, que tiene implícitas secciones.[4] En la figura 1 se observa un ejemplo de una HCE correspondiente a un documento XML definido en el estándar CDA.

```
<?xml version="1.0" encoding="UTF-8"?>
<ClinicalDocument xmlns="urn:hl7-org:v3">
  <recordTarget>
    <patientRole>
      <id extension="12345"
      root="2.16.840.1.113883.3.933"/>
      <patientPatient>
        <name>
          <given>Henry</given>
          <family>Levin</family>
          <suffix>the 7th</suffix>
        </name>
        <administrativeGenderCode
          code="M"
          codeSystem="2.16.840.1.113883.5.1"/>
        <birthTime value="19320924"/>
      </patientPatient>
      <providerOrganization>
        <id extension="M345"
        root="2.16.840.1.113883.3.933"/>
      </providerOrganization>
    </patientRole>
  </recordTarget>
...
</ClinicalDocument>
```

Fig. 1. Ejemplo de un documento XML basado en el estándar CDA-HL7

Debido a su estructura jerárquica y auto-descriptiva un documento XML es más natural tratarlo como un conjunto de partes o una serie de secciones (que se puede dividir en varias subsecciones, etc.). Como consecuencia la HCE puede verse como una colección {D1, ..., Dm}, donde cada Di contiene un conjunto de Unidades Estructurales (UE) UE = {UE1, ..., UEn}, con lo cual desaparece el concepto de documento como unidad indivisible.[5] Las diferentes UE identificadas semánticamente en la HCE cronológica, basado en criterios de expertos, se muestra en la figura 2.

Fig. 2. UE identificadas semánticamente en la HCE cronológica, basado en criterios de expertos

Esta propuesta nos permite obtener una representación estandarizada del conocimiento implícito en una colección de HC y proporcionar soporte a la toma de decisiones.

Acerca del agrupamiento de HCE

Un agrupamiento de documentos intenta encontrar una estructura del conjunto de datos en grupos naturales basado principalmente en la similitud y las relaciones de los objetos, para obtener una distribución interna que logre homogeneidad dentro de los grupos y heterogeneidad entre ellos.[13,14] De manera, que los objetos que pertenezcan a un mismo grupo sean tan similares como sea posible y los objetos que pertenezcan a grupos diferentes, sean tan disímiles como se pueda. En este trabajo, se propone realizar el agrupamiento de una colección de HCE basado en el agrupamiento de documentos XML, debido a que, al implementar CDA, una HCE es un documento XML. Así, los resultados obtenidos pueden ser analizados y explicar la lógica de las acciones tomadas por un médico ante situaciones similares.[15-17] Además, múltiples HC agrupadas por signos, síntomas, incidencia y prevalencia, diagnóstico diferencial, antecedentes personales y familiares, hábitos tóxicos, tratamientos y diagnóstico pueden contribuir a mejores prácticas de salud.

Agrupamiento de HCE basado en una metodología para el agrupamiento de documentos XML

Debido a que un documento XML contiene información semi-estructurada, se han propuesto varios trabajos relacionados con el agrupamiento de documentos XML teniendo en cuenta que existen tres variantes:[2,10,18] las que consideran sólo el contenido,[9,13,14] las que utilizan sólo su estructura[8,10,11,18-20] y las que combinan ambas dimensiones .[21-25] No obstante, la mayoría de los enfoques existentes no combinan sus dos dimensiones: estructura y contenido, dada su complejidad; sin embargo, para mejores resultados en el agrupamiento es esencial utilizar ambas.[26] La tabla 1 muestra un resumen de algunos trabajos relacionados con el diseño de algoritmos para el agrupamiento de documentos XML.

En esta sección, se presenta una metodología para el agrupamiento de documentos XML, así como una nueva función de similitud, OverallSimSUX[27] que facilita capturar el grado de similitud entre los documentos.

La relación estructural existente entre los documentos XML puede aportar mejores resultados al agrupamiento, cuando se utiliza el contenido en función de la relación existente entre sus UE. En este trabajo, se propone un conjunto de UE para gestionar la HCE definidas a partir del criterio de expertos, vale recordar que una UE se corresponde con un conjunto de etiquetas con significado semántico.

UE = {Datos personales , antecedentes, síntomas, signos, incidencia, prevalencia, diagnóstico diferencial, complementarios, tratamientos, evolución, diagnóstico}.

Tabla 1. Técnicas para el agrupamiento de documentos XML

Sólo Contenido
Usan alguna variante de VSM: *Kurgan, L. "Semantic mapping of xml tags using inductive machine learning." [14]* *Shen, Y. "Clustering schemaless xml document." [13]*
Sólo Estructura
Representación del árbol XML para calcular alguna variante de la distancia *tree-edit*: *Dalamagas, T. "A Methodology for Clustering XML Documents by Structure." [8]* *Flesca, S. "Fast detection of XML structural similarities." [10]* *Lesniewska, A. "Clustering XML documents by structure." [11]*
Considera la estructura del XML basado en el uso de *Edit Graph*: *Chawathe, S.S. "Comparing Hierarchical Data in External Memory." [18]*
Enfoque jerárquico: *Costa, G. "Hierarchical clustering of XML documents focused on structural components. [19]*
Enfoque del agrupamiento en dos pasos: *Aïtelhadj, A. "Using structural similarity for clustering XML documents." [20]*
Estructura y Contenido
Uso de *Closed Frequent Sub-Trees*: *Kutty, S. "Combining the structure and content of XML documents for clustering using frequent subtrees." [22]*
Análisis comparativo de los documentos XML basado en una variante de VSM: *Yang, W. "A semi-structured document model for text mining." [23]*
Uso de *edit-distance* para comparar los documentos XML semántica y estructuralmente: *Tekli, J.M. "A Novel XML Document Structure Comparison Framework based-on Subtree Commonalities and Label Semantics." [24]*
Uso del algoritmo iterativo *K-Star* en un proceso de agrupamiento recursivo: *Pinto, D. "BUAP: Performance of K-Star at the INEX'09 Clustering Task." [25]*

La construcción de la matriz de similitud basada en el cálculo de la medida de similitud OverallSimSUX facilita capturar el grado de similitud entre los documentos. Esta función analiza la relación existente entre los documentos, cada uno de los cuales se corresponde con una HCE de la colección, tratando simultáneamente los documentos como unidades indivisibles y cada colección de UE como colecciones independientes. Una vista gráfica del modelo del esquema para construir la matriz de similitud OverallSimSUX en este trabajo se muestra en la figura 3.

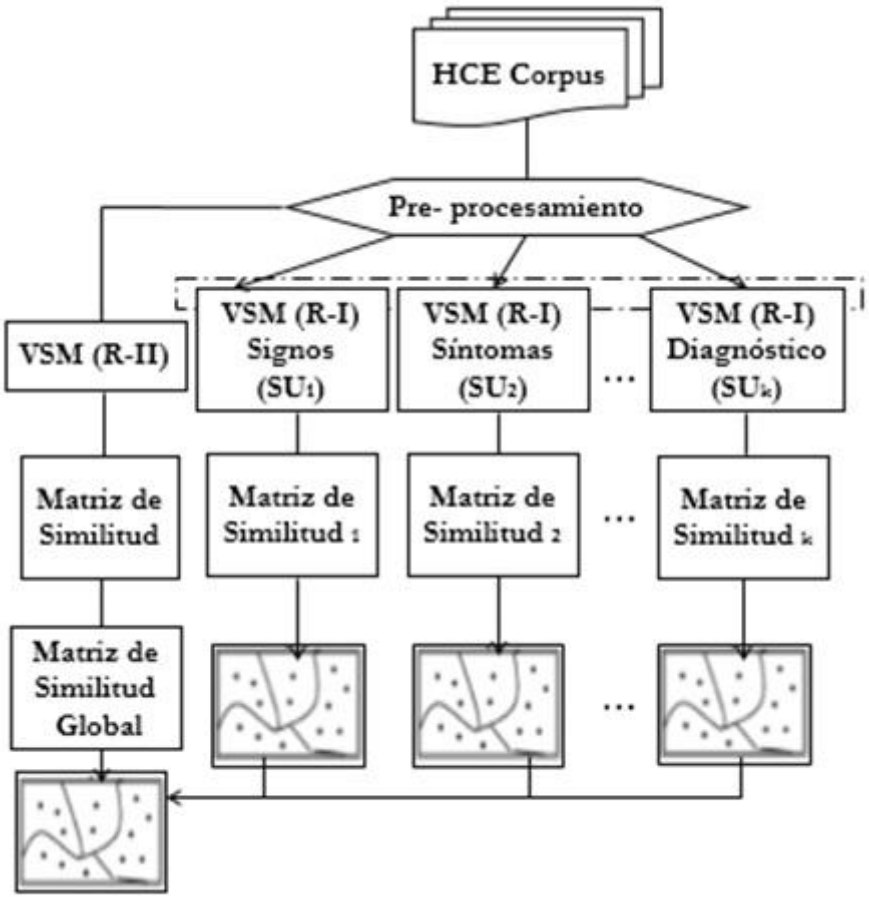

Fig. 3. Esquema de la metodología propuesta

En la figura 4, se detalla el procedimiento general para la construcción de esta matriz, para lo cual son necesarios tres pasos: (1) Pre-procesamiento de toda la colección, identificando cada UE; (2) Representación Textual, a partir de la Representación I (R-I) y Representación II (R-II) y (3) Proceso de Agrupamiento Final.

```
Entrada: Corpus D de documentos de HCE
Salida: Grupos, calidad de los grupos, documento
más representativo por grupo
 Inicio
 1.Pre-procesamientos /*análisis léxico,
 eliminación de palabras de parada,
 segmentación…*/
 2.Construir todas las k-colecciones (corpus D)
 3.For each DSUk
     - Rep-I← Hacer Representación-I (DSUk)
              mediante TF-IDF
     - Matriz_Sim← Calcular la matriz de
                   similitud R-I usando la
                   similitud Coseno
     - Grupos← Aplicar método de agrupamiento
              K-Star a Matriz_Sim
   End For
 4.Rep-II← Hacer R-II corpus D complete usando
   ecuación (1) para calcular la frecuencia
   /*Ver Tabla II */
 5. Matriz_SimII← Calcular la matriz de
                  similitud para R-II usando la
                  similitud Coseno
 6. Matriz_O_Sim ← Calcular matriz de similitud
                   usando la medida
                   OverallSimSUX teniendo en
                   cuenta todos los
                   agrupamientos para cada
                   DSUk y Matriz_SimII
 7.Obtener el agrupamiento Final aplicando el
   método de agrupamiento K-Star[19] a
   Matriz_O_Sim
 Fin
```

Fig. 4. Procedimiento General para el agrupamiento de documentos XML utilizando *OverallSimSUX*

Pre-procesamiento (Paso 1)

La propuesta presentada en este trabajo responde a la necesidad de desarrollar herramientas para la gestión de la información clínica y brindar soporte al descubrimiento de conocimiento. Con el fin de disponer de datos fiables para estandarizar los términos con la misma semántica, en la etapa de pre-procesamiento deben garantizarse el filtrado de la información y la unificación de la terminología utilizada por el personal médico.

Representación Textual (Paso 2)

Obtener las representaciones: Representación I (R-I), a partir de las UE definidas para los documentos HCE, cada colección independiente se corresponde con la colección que contiene las k-ésimas UE de cada uno de los documentos de la colección de HCE original; Representación II (R-II), se realiza teniendo en cuenta la colección completa y los resultados de los agrupamientos realizados utilizando las representaciones R-I.

Representación I

La colección original de documentos se divide en k-colecciones. El concepto de k-colección[27] refleja la correspondencia entre la colección y la UE. Para cada k-colección la Representación I se realiza utilizando la representación clásica VSM. En particular, la construcción de esta matriz se lleva a cabo mediante el cálculo de la frecuencia TF-IDF.[6] TF-IDF es una medida estadística de peso utilizada en el procesamiento del lenguaje natural para determinar la importancia de un término en una colección dada, mediante el uso de la representación vectorial. La importancia de cada término aumenta proporcionalmente al número de veces que aparece este término en el documento (frecuencia), y se compensa con la frecuencia del término en la colección.

Representación II

En este trabajo la estructura de la HCE se adiciona al análisis, por lo tanto, la Representación II es una modificación de la clásica representación VSM y la frecuencia se pondera teniendo en cuenta la UE en la que aparece el término analizado, según se define en la ecuación 1 para un término ti y un documento dj, donde n es la cantidad de UE presentes en dj , frik es la frecuencia de ti en la UEk y wk es el peso de la UEk en la HCE dj. El cálculo del peso de la UEk para cada documento dj se realiza a partir de la ecuación 2; LUE es la longitud de la UEk, LDoc es la longitud del documento dj y la pot es un valor dado. De esta manera la idea queda formalizada correctamente. Sin embargo, si hay elementos que tienen en la HC mayor valor diagnóstico o interés que otros, estos pueden ser tratados como términos difusos, a partir de un conjunto de valores de membresía establecidos por el experto. (tabla 2)

$$tf_{dj}(t_i) = \sum_{k=1}^{n} (w_k * fr_{ik}$$

$$w_{kj} = (e^{(-L_{SU}/L_{Doc})})^{pot}$$

Tabla 2. Representación II. Dónde $tf_{di}(t_i)$ es la frecuencia absoluta de aparición del término t_i en el documento d_j

	$Term_1$	$Term_2$	...	$Term_m$
HCE 1	$tf_{d1}(t_1)$	$tf_{d1}(t_2)$	...	$tf_{d1}(t_m)$
HCE 2	$tf_{d2}(t_1)$	$tf_{d2}(t_2)$	...	$tf_{d2}(t_m)$
...	...	...	...	...
HCE n	$tf_{dn}(t_1)$	$tf_{dn}(t_2)$	...	$tf_{dn}(t_m)$

Agrupamiento de las k-colecciones

A partir de la Representación I se calcula la matriz de similitud, que compara dos documentos utilizando la medida coseno; a partir de la ecuación 3. Se realiza un agrupamiento independiente para cada k-colección. Para llevar a cabo el agrupamiento, se utiliza el clásico algoritmo K-Star.[22]

$$S_{\text{Coseno}}(d_i, d_j) = \frac{\sum_{k=1}^{m} (d_{ik}}{\sqrt{\sum_{k=1}^{m} d_{ik}^2}}$$

Cálculo de la Matriz OverallSimSUX

La medida de similitud OverallSimSUX, es especificada formalmente a través de la ecuación 4. Se inicia con los resultados del agrupamiento realizado para cada k-colección y la matriz de similitud basada en el cálculo de la medida coseno, a partir de la Representación II. OverallSimSUX considera m como la cantidad de UE identificadas en los documentos de HCE. Esta función de similitud se normalizó por la suma de los pesos de las m UE y el valor máximo de similitud global de sg (es decir, 1). Por tanto, su máximo valor (es decir, 1) se alcanza cuando los documentos de HCE i, j pertenecen al mismo grupo en todos los k-agrupamientos y el valor de sg es máximo.

$$f(C, s_g, i, j) = \frac{\sum_{k=1}^{m} (w_k * \lambda_{k(i,j)} + s_{g(i,}}{\sum_{k=1}^{m} w_k + 1}$$

Agrupamiento Final (Paso 3)

Para llevar a cabo el agrupamiento final, se utiliza nuevamente el algoritmo de agrupamiento K -Star.

INTERPRETACIÓN DEL AGRUPAMIENTO DE HCE

El agrupamiento de HCE basado en los síntomas o signos, y no sólo teniendo en cuenta el diagnóstico final, permite al especialista evaluar objetivamente el valor diagnóstico de una prueba en particular, antes de inferir el resultado de las demás pruebas. Lograr una visión coherente de la historia del paciente, lo que realmente se ha hecho y por qué. La relación de pacientes y la correspondencia de ellos atendiendo a sus antecedentes patológicos personales y familiares, la reacción adversa a ciertos medicamentos en casos similares, permitiría explicar la causa de las acciones tomadas y decidir la conducta a seguir en el tratamiento de ciertos pacientes.

Por otra parte, el beneficio de tener pacientes similares, con iguales diagnósticos permite a los estudiantes en menos tiempo completar la HCE de un paciente. Utilizando la metodología propuesta y las UE de la HCE donde se concentran sus dudas, le permitiría obtener grupos similares de casos, y realizar recomendaciones sobre el uso correcto de un tratamiento, un complementario, entre otras opciones.

Los resultados obtenidos por las colecciones de HCE agrupadas deben interpretarse teniendo en cuenta criterios de expertos. El uso de reglas de asociación permite explicar las relaciones entre las HCE de los pacientes, que pertenecen a un mismo grupo. Los centroides o HCE más relevante de cada grupo pueden permitir a los expertos estudiar casos similares a estos basados en criterios aplicados en ocasiones pasadas y emitir diagnósticos finales en un menor tiempo.

Para evaluar los efectos de esta metodología en las colecciones de HCE, se propone el uso de una muestra de HCE de los ingresos hospitalarios del Servicio de Admisión del Hospital "Arnaldo Milián Castro", asociados a 45 enfermedades diferentes.

La metodología ha sido validada a partir de la aplicación de una encuesta a 13 médicos del Hospital "Celestino Hernández Robau". Cada uno de los expertos seleccionados utilizó el sistema con las colecciones de HCE propias de sus ramas de desempeño. Algunos de los temas abordados son: cardiología, hematología y gastroenterología.

Los usuarios siempre utilizaron colecciones conocidas para poder valorar los resultados del sistema. Los usuarios aplicaron el sistema tanto a colecciones heterogéneas como homogéneas. Por otra parte, la interpretación de los resultados obtenidos por la metodología evaluado por expertos, asegura la viabilidad de la metodología propuesta para la gestión de la información clínica y el descubrimiento del conocimiento implícito en ella.

CONCLUSIONES

En este trabajo se evidencia la importancia del agrupamiento documental para la gestión del conocimiento a partir de la información clínica, debido a la necesidad de un conocimiento pertinente a través del uso productivo de la información contenida en HCE y la necesidad del uso de estándares que garanticen la interoperabilidad de los HIS. Fue propuesta una metodología para el agrupamiento de HCE concebidas como documentos XML, brindando soporte a la arquitectura CDA del estándar HL7. Esta metodología combina las dimensiones: estructura y contenido presentes en los documentos clínicos lo que aporta mejores resultados al agrupamiento. La función de similitud OverallSimSUX que incluye la metodología facilita capturar el grado de similitud entre éstos, tomando como génesis la relación entre sus UE. Varias UE fueron propuestas utilizando criterios de expertos, lo cual permite gestionar colecciones de HCE utilizando la metodología.

REFERENCIAS BIBLIOGRÁFICAS

1. Engelbrecht R. K4Health. Knowledge for Health. Integrating EHR and Knowledge for better health care. Status of the EoI and work items. EUROREC. Berlin. [citado 2002 dic 14]. Disponible en: http://www.eurorec.net/EUROREC_2002_presentations.html

2. Brau B. Extensible Markup Language(XML) 1.0., in W3C Recommendation. 1998.

3. C.D M, Raghan P, Schütze H. Introduction to Information Retrieval. 2008 Cambridge University Press.

4. Zwaanswijk M, Verheij F. J, Wiesman R. D. Benefits and problems of electronic information exchange as perceived by health care professionals: an interview study. BMC Health Services Research, 11:256. Doi:10.1186/1472-6963-11-256. (2011)

5. Dalamagas T, Cheng T, Winkel K-J, Sellis T. A. Methodology for Clustering XML Documents by Structure. Information Systems (2006).

6. Dick R. S, Oteen E. B, Detmer D. E. (eds). The computer-based patient record: An essential technology for health care. Revised Edition Washington, D.C: The Nacional Academy Press. 1997. Capítulo 2. p. 74-99. [citado 2002 dic 14]. Disponible en: http://books.nap.edu/books/0309055326/html/R1.html

7. Gérvas J. La historia clínica electrónica: muchas promesas y pocos hechos. Aten Primaria. 2008;40(Supl 1):13

8. Guerrini G, Mesiti M, Sanz I. An Overview of Similarity Measures for Clustering XML Documents. 2006.

9. Wilde E, Glushko R.J. XML fever. Comm. ACM, 2008. 51(7): p. 40-46. doi: 10.1145/1364782.1364795

10. Wang G. RPE query processing and optimization techniques for XML databases. J. Comput. Sci. Technol, 2004. 19(2): p. 224-237.

11. Bertino E, Ferrari E. XML and data integration. IEEE Internet Comput, 2001. 5(6): p. 75-76. doi: 10.1109/4236.968835

12. Algergawy A. XML Data Clustering: An Overview, in ACM Computing Surveys. 2011. doi: 10.1145/1978802.1978804

13. Kruse R, Döring C, Lesor M.-J. Fundamentals of Fuzzy Clustering, in Advances in Fuzzy Clustering and its Applications, J.V.d. Oliveira and W. Pedrycz, Editors. 2007, John Wiley and Sons: Est Sussex, England. p. 3-27.

14. Ji T, X Bao, D. Yang. FXProj - A Fuzzy XML Documents Projected Clustering Based on Structure and Content. LNAI 7120, 2011: p. 406-419.

15. Yousuke W, Hidetaka K , Haruo Y. Similarity search for office XML documents based on style and structure data. International Journal of Web Information Systems, 2013. 9(2): p. 100-117. doi: 10.1108/IJWIS-03-20 13-0005

16. Kaufman L, Rousseeuw P.J. Finding groups in data: an introduction to cluster analysis. Wiley Series in probability and mathematical statistics. 1990: John Wiley and Sons.

17. Martín C. Aprendizaje Automático Y Minería De Datos Con Modelos Gráficos Probababilísticos. DEA, Universidad De Granada. 2007

18. Tekli J.M, Chbeir R. A Novel XML Document Structure Comparison Framework based-on Subtree Commonalities and Label Semantics. Elsevier, 2011. doi: 10.1016/j.websem.2011.10.002

19. Pinto D, Tovar M, Vilariño D . BUAP: Performance of K-Star at the INEX'09 Clustering Task. in INEX 2009 Workshop Pre-proceedings. 2009. Woodlands of Marburg, Ipswich, Queensland,Australia. doi: 10.1007/978-3-642-14556-8_43

20. Vries C. Overview of the INEX 2010 XML mining track: clustering and classification of XML documents, in In Lecture Notes in Computer Science, Springer: Amsterdam. 2011

21. Kurgan L, Swiercz W, Cios K.J. Semantic mapping of xml tags using inductive machine learning. in 11th International Conference on Information and Knowledge Management. 2002. Virginia, USA.

22. Shin K, Han S.Y. Fast clustering algorithm for information organization., in In:Proc. of the CICLing Conference. 2003, Lecture Notes in Computer Science.Springer-Verlag (2003). p. 619-622. doi: 10.1007/3-540-36456-0_69

23. MacQueen J.B. Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967: Berkeley, University of California. p. 281-297.

24. Sim I, Gorman P, Greenes R. A, Haynes R. B, Kaplan B, Lehmann H, Tang P. C. Clinical decision support systems for the practice of evidence-based medicine. J. Am Med Inform Assoc 2001; 8: 527-34.

25. Xiong H, J, Chen K. K-means clustering versus validation measures: a data distribution perspective. in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2006). 2006. Philadelphia, PA, USA: ACM Press. doi: 10.1109/TSMCB.2008.2004559

26. Costa G. Hierarchical clustering of XML documents focused on structural components. Data & Knowledge Engineering, 2013. 84: p. 26-46. doi: 10.1016/j.datak.2012.12.002

27. Magdaleno D, Fuente sI.E, García M.M. Clustering XML Documents using Structure and Content Based in a Proposal Similarity Function (OverallSimSUX). Computación y Sistemas, 2015.

Recibido: 22 de marzo de 2016.
Aprobado: 12 de mayo de 2016.