Ciencias Matemáticas, Vol. 29, No. 2, Pag. 99-105, 2015
Recibido 06-2015

Metodología para asistir la toma de decisiones diagnóstica a partir del descubrimiento del conocimiento implícito en Historias Clínicas
Methodology for diagnostic decision making starting from knowledge discovery in clinical records

Ivett E. Fuentes[1]*, Damny Magdaleno[1], María M. García[1]

Resumen La proliferación de información disponible en los centros hospitalarios a partir del uso extendido de las historias clínicas en formato electrónico (HCE) es arrolladora. Disponer de información sistematizada, gestionarla de forma efectiva y segura es esencial para garantizar mejores prácticas de salud. La gestión de información clínica se vuelve cada vez más compleja y desafiante; debido a que los repositorios de HCE son heterogéneos, grandes, diversos y dinámicos; lo que dificulta compartir la información y reutilizarla. Si bien los medios tecnológicos actuales y las necesidades impuestas por modelos emergentes de gestión clínica favorecen el uso extendido de HCE; la llamada Sociedad de la Información está siendo superada por la necesidad de nuevos métodos capaces de procesar esta información de forma eficiente y eficaz. En este trabajo se analiza la importancia que tiene el agrupamiento documental en la gestión del conocimiento desde la información clínica. Se presenta una metodología para el agrupamiento de HCE teniendo en cuenta los diferentes factores de la HC y los datos recogidos en cada uno de ellos a partir de la anamnesis o interrogatorio y el examen físico; con el propósito de identificar automáticamente la relación de los pacientes a través de sus síntomas o signos. Se propone una variante para asistir la toma de decisiones diagnóstica de un nuevo paciente, mediante una clasificación supervisada que utiliza la información relevante proporcionada por la metodología presentada. Finalmente la interpretación de los resultados muestra la factibilidad de la propuesta.

Abstract The proliferation of available information in hospitals that result from the widespread use of electronic medical records (EMR) is overwhelming. Having systematized information, manage effectively and safely is essential to ensure better health practices. In this paper, it's analyzed the importance of document clustering in Document Management in order to discover hidden knowledge in the clinical information. It is proposed a methodology for automatic clustering of EMR taking into account different factors and data collection in physical examination. It is presented a variant for to assistant the diagnosis decision making of a new patient, by means of a supervised classification that it uses the relevant information provided by the presented methodology. Interpretation of the results of EMR clustering showed the feasibility of is proposed.

Palabras Clave
Agrupamiento, descubrimiento de conocimiento, HCE, XML, clasificación

[1] *Universidad Central "Marta Abreu" de las Villas, Cuba, ivett@uclv.cu, dmg@uclv.edu.cu, mmgarcia@uclv.edu.cu*
***Autor para Correspondencia**

1. Introducción

Nadie pone en duda el papel que juegan las nuevas Tecnologías de la Información y las Comunicaciones (TIC) en las organizaciones y esto se hace, lógicamente, extensible a la Gestión Documental (GD), que administra el flujo de documentos e información en las instituciones de salud impuesto por el uso creciente de las historias clínicas en formato electrónico. Con ello, surgen nuevas oportunidades para la utilización de la enorme riqueza de datos e información que reside en los sistemas hospitalarios en entornos educativos y de investigación. Por otro lado, 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 de salud pública (dígase: informes electrónicos de casos, bases de datos poblacionales), entre otros [1]. Como consecuencia, la

creación de repositorios de HCE y el volumen de información generada desde estos, aumenta continua y exponencialmente. Adoptar herramientas de soporte a la toma de decisiones en la práctica clínica es necesario para brindar a los médicos mejores condiciones de trabajo, contribuir al ejercicio de una medicina basada en pruebas y asegurar el uso productivo de la información almacena [2,3]. En este sentido, aunque se han desarrollado 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 [4, 5].

La importancia de la estandarización y codificación de datos almacenados en la HCE es reconocido por varios investigadores [1, 4, 6, 7]. Como consecuencia, la recopilación de la información clínica debe ir migrando hacia el uso controlado de textos estructurados. En efecto, la propia distribución de los elementos de la HC, hacen 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) es el conjunto de estándares informáticos de salud más desarrollado y de mayor cobertura internacional para dar soporte a la HCE [4]. HL7 facilita el intercambio electrónico de información clínica, mediante la notación formal de modelado UML y el metalenguaje XML [2].

XML se ha convertido en el formato de intercambio de datos estándar entre las aplicaciones Web; debido a su extensibilidad y estructura de fácil análisis y procesamiento [2, 8, 9]. Un documento XML es una estructura jerárquica de información que incorpora estructura y datos en una misma entidad. De este modo, la estructura de estos documentos puede ser explotada para realizar recuperación de documentos relevantes [5]. Aunque existen varias formas de gestionar el conocimiento: la categorización, la clasificación y el agrupamiento; exclusivamente, el agrupamiento de documentos XML nos 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, 11, 12]. Por tal motivo, en el presente trabajo se estudia la importancia del agrupamiento documental y su interpretación en la gestión del conocimiento implícito en HCE. La organización del artículo es la siguiente: en la sección 2 se analizan las técnicas para abordar el agrupamiento de la información clínica. En la sección 3 se propone una metodología para el agrupamiento de HCE en formato XML combinando su contenido y estructura a partir de la definición del concepto de Unidad Estructural (SU[1]), lo que posibilita al médico una mejor práctica de salud. En la sección 4 se plantea como asistir la toma de decisiones diagnóstica a partir de una clasificación supervisada que utiliza el conocimiento descubierto. En la sección 5 se presentan las conclusiones.

[1] Siglas en inglés: *Structural Unit* (SU)

2. Información clinica y agrupamiento documental

Cada día más datos electrónicos son presentados debido a: el continuo crecimiento de información desde múltiples esferas y la automatización de gran parte de los procesos de la sociedad. Esto se hace extensible a la gestión de la información clínica, debido a que la medicina incorpora, cada vez en mayor medida, el soporte de la evidencia clínica en las decisiones de la práctica facultativa habitual [1].

La Historia Clínica (HC) es una fuente de datos fundamental y construye el documento principal en un sistema de información hospitalario (HIS). Es una herramienta básica para la investigación biomédica, la formación de estudiantes y la educación médica continuada. Así como, un documento legal que surge del contacto entre el médico y el paciente, válido desde el punto de vista clínico y legal debido a que engloba información de tipo asistencial, preventivo y social. Esta información incluye datos clínicos relacionados con la situación del paciente, datos de sus antecedentes personales y familiares, hábitos tóxicos y todo aquello vinculado con su salud biopsicosocial; el proceso evolutivo, tratamiento y recuperación [7]. Es un documento donde el paciente deja registrado y firmado su reconocimiento para utilizarlo en la toma de decisiones. Existen varios modelos atendiendo al lugar donde se genera: cronológicos, orientada a problemas de salud (POMR) y protocolizada. Algunos componentes de modelos clásicos de HC, como la orientada a problemas, han sido considerados especialmente adecuados para los usos docentes y científicos de la HCE [6]. En este trabajo se propone utilizar el modelo cronológico que se genera en los centros hospitalarios.

Incorporar las TIC en el núcleo de la actividad sanitaria, supone brindar soporte a la HCE. Así, la HC deja de ser un simple registro de la información generada, para formar parte de un HIS integrado. No obstante, en el proceso de conceptualización y de implementación de las TIC existen problemas que limitan el uso productivo de la información e impiden lograr su impacto positivo en la calidad de la atención clínica, en la morbilidad y en la mortalidad, en la integración efectiva de la HCE y en el uso de las herramientas de aprendizaje automático de inteligencia artificial. A esto se le añade, los problemas respecto a la codificación, las normas y los estándares [1, 4].

2.1 Usos actuales de la HCE en Las investigaciones

Los registros informatizados del servicio de admisión de los hospitales son utilizados para realizar investigaciones clínicas y epidemiológicas, al no disponer de otras fuentes de datos bien estructurados en los servicios clínicos capaz de obtener conocimiento [4]. Por lo que almacenar adecuadamente esta información, hacerla accesible y reutilizarla en la forma más conveniente es un proceso todavía en potencia [4, 7].

En este sentido, inferir áreas que deben ser interpretadas por los expertos de la medicina a partir de la información

disponible, garantizaría el uso productivo de información contenida en la HCE; con el propósito de realizar investigaciones clínicas que permitan: nuevas soluciones diagnósticas y terapéuticas, valoración del uso de tecnología de punta, estudio de los resultados en pacientes, efectividad y eficacia de la atención médica, identificación de poblaciones de riesgo, desarrollo de registros y análisis de la eficacia de procesos. Por lo que, se hace necesario que los sistemas de información que utilizan las instituciones de prestación de servicios de salud, implementen estándares informáticos internacionalmente reconocidos, como HL7. De ahí, la necesidad de migrar de manera controlada hacia información clínica almacenada de forma estructurada. En este trabajo se propone concebir la HCE como un documento XML, en el que existen implícitamente secciones que resulta más natural tratarla como un conjunto de partes o una serie de secciones (que pueden dividirse en varias subsecciones y así sucesivamente). Consecuentemente un conjunto dado de HCE se corresponde con la colección $D = \{D_1, \ldots, D_m\}$, donde cada D_i contiene a su vez un conjunto de unidades estructurales $SU = \{SU_1, \ldots, SU_n\}$. Así, desaparece el concepto de documento como unidad indivisible [5]. Las *SU* identificadas semánticamente en la HCE cronológica basado en el criterio de expertos se muestran en la figura 1. Esta concepción garantiza lograr una representación del conocimiento estandarizada y brindar soporte a la toma de decisiones.

2.2 Enfoque para agrupar HCE

El agrupamiento documental concibe encontrar una estructura de grupos que se ajuste al conjunto de datos, logrando homogeneidad dentro de los grupos y heterogeneidad entre ellos, siendo una alternativa para describir automáticamente el significado científico y clínico de la información biomédica desde grandes volúmenes de información. Específicamente, un algoritmo de agrupamiento intenta encontrar grupos naturales de datos, basándose principalmente en la similitud y las relaciones de los objetos, para obtener una distribución interna del conjunto de datos mediante su particionamiento en grupos [13, 14]. 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 [15, 16, 17].

En este trabajo se propone un enfoque para agrupar HCE al concebirla como un documento XML. El aprendizaje obtenido puede ser interpretado y brindar soporte a las acciones realizadas por los expertos, rara vez explicadas. Además, al disponer de múltiples HC agrupadas por signos, patologías o síntomas, incidencia y prevalencia, diagnóstico diferencial, pruebas y tratamientos efectuados; el profesional de la salud podría particularizar experiencias implícitas en las HCE procesadas.

3. Agrupamiento de HCE basado en una metodologia para el agrupamiento de documentos XML

Debido a que un documento XML contienen su información en forma semiestructurada, varios trabajos han sido propuesto teniendo en cuenta las tres variantes de abordar el agrupamiento de XML [2, 10, 11]: los que consideran solo el contenido [9, 13, 14], los que utilizan solo su estructura [8, 10, 11, 18, 19, 20] y los que combinan ambas dimensiones [22, 23, 24, 25]. La mayoría de los enfoques existentes no combinan sus dos dimensiones: estructura y contenido, dado su gran complejidad; sin embargo, para obtener mejores resultados en el agrupamiento es esencial utilizar ambas [26]. La tabla 1 muestra un resumen de algunos de los algoritmos para el agrupamiento de documentos XML.

En esta sección se propone un método para el agrupamiento de documentos XML, a partir de la nueva función de similitud, *OverallSimSUX* [27], que facilita capturar el grado de similitud entre los documentos. La metodología general para el agrupamiento de la información contenida en HCE, combinando las dimensiones: estructura y contenido; con el propósito de contribuir al descubrimiento de conocimiento clínico relevante es presentada. Una visión gráfica del esquema del modelo general presentado en este trabajo se muestra en la figura 2.

La relación estructural existente entre los documentos XML aporta mejores resultados al agrupamiento cuando el contenido es utilizado en función de la relación existente entre las SU. En este trabajo, se presenta un conjunto de SU identificadas en la HCE utilizando el criterio de expertos, *SU*= {*Datos Personales*[2]*, Antecedentes, Síntomas, Signos, Incidencia, Prevalencia, Diagnóstico Diferencial, Pruebas, Tratamiento, Recuperación, Diagnóstico Final*}.

La construcción de la matriz de similitud basada en el cálculo de la medida de similitud propuesta facilita capturar el grado de similitud entre los documentos. Esta función analiza la relación existente entre los documentos de HCE, tratando simultáneamente los documentos como unidades indivisibles y cada colección de SU como colecciones independientes. La figura 2 muestra, como obtener la matriz de similitud *OverallSimSUX* a partir de una colección de documentos de HCE.

En el algoritmo 1 se detalla el procedimiento general para la construcción de esta matriz, a partir de tres pasos fundamentales:

1. Pre-procesamiento de toda la colección, identificando cada unidad estructural.
2. Representación Textual utilizando la Representación I y la Representación II.
3. Proceso de Agrupamiento Final.

[2]Basado en {*Sexo, Edad, Lugar de Nacimiento, Lugar de Residencia, Grupo Sanguíneo*}

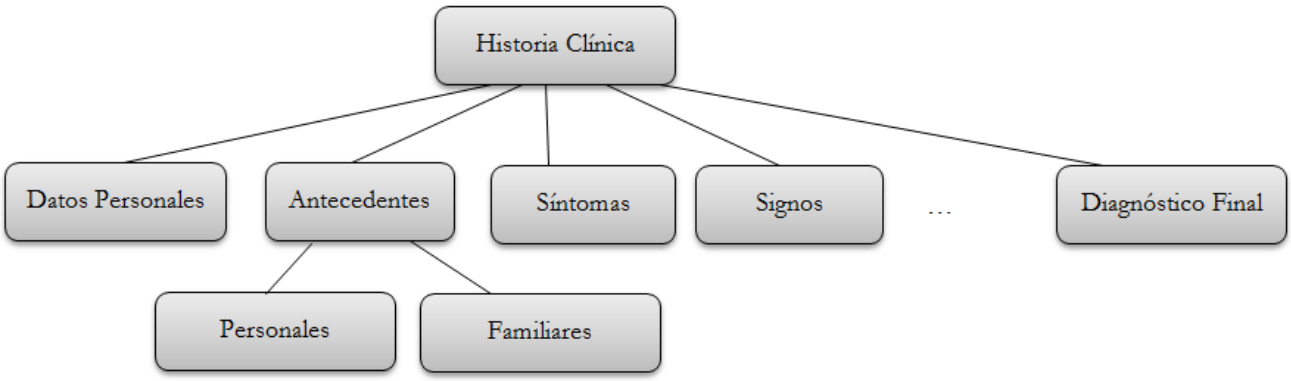

Figura 1. Unidades Estructurales semánticamente identificadas en la HCE cronológica utilizando criterio de expertos.

Tabla 1. Resumen de algoritmos para el agrupamiento de documentos XML

Agrupamiento por	Autor	Método
Solo Contenido	Kurgan et al [21] Shen [13]	Una variante de VSM
Solo Estructura	Dalamagas et al [5] Flesca [10] Lesniewska [11]	XML como árbol para calcular distancia tree-edit
Solo Estructura	Chawathe [18] Costa [26] Aïtelhadj [20]	Uso de Edit Graph Enfoque jerárquico Enfoque two-step
Estructura y Contenido	Kutty [22] Yang [23] Tekli y Chbeir [18] Pinto et al [19]	Usan Closed_Frequent_Sub-Trees Variante de VSM Uso de la similitud semántica y distancia tree-edit Uso del algoritmo K-Star en proceso recursivo

Pre-procesamiento La propuesta realizada en este trabajo responde a la necesidad de desarrollar herramientas para gestionar la información clínica y brindar soporte al descubrimiento de conocimiento. Con el propósito de estandarizar los términos con igual significado semántico, el preprocesamiento de los datos incluye la unificación de la terminología usada por el personal médico.

Representación Textual Obtener las representaciones, Representación I, usando las colecciones de HCE de las SU tratadas como colecciones independientes; Representación II, considerando la colección completa; realizar los agrupamientos de las colecciones asociadas a cada US utilizando la Representación I.

Representación I La colección original de documentos es dividida en k-colecciones. El concepto de k-colección [27] refleja de la correspondencia entre la colección y la SU. Para cada k-colección la Representación I se construye utilizando la VSM clásica. En particular, la construcción de esta matriz se realiza utilizando la medida Frecuencia del Término y Frecuencia Inversa del Documento (TF-IDF) [6]. TF-IDF es una medida estadística que determina cuán importante es un término, usando el vector de representación. La importancia de cada término aumenta proporcionalmente al número de veces que este término aparece en el documento (la frecuencia), pero se compensa con la frecuencia del término en la colección.

Representación II En este trabajo la estructura de la HCE es adicionada al análisis, por consiguiente la Representación-II es una modificación de la VSM clásica, donde la frecuencia es pesada teniendo en cuenta la importancia de la SU a la que pertenece el término analizado, definida en la ecuación (1) para un término t_i y un documento d_j. Donde n es la cantidad de SU de d_j, fr_{ik} es la frecuencia de t_i en la SU_k y w_k es el peso de la importancia de la SU_k en el documento d_j. El cálculo del peso de SU_k para cada documento d_j se realiza según la ecuación (2); donde L_{SU} es la longitud de SU_k, L_{Doc} es la longitud del documento d_j y pot es un valor dado. De esta manera la idea queda formalizada. Aunque, si existen elementos de la HCE con mayor interés diagnóstico que otros, pueden tratarse como términos borrosos atendiendo a un conjunto de grados de pertenencia fijados por el experto.

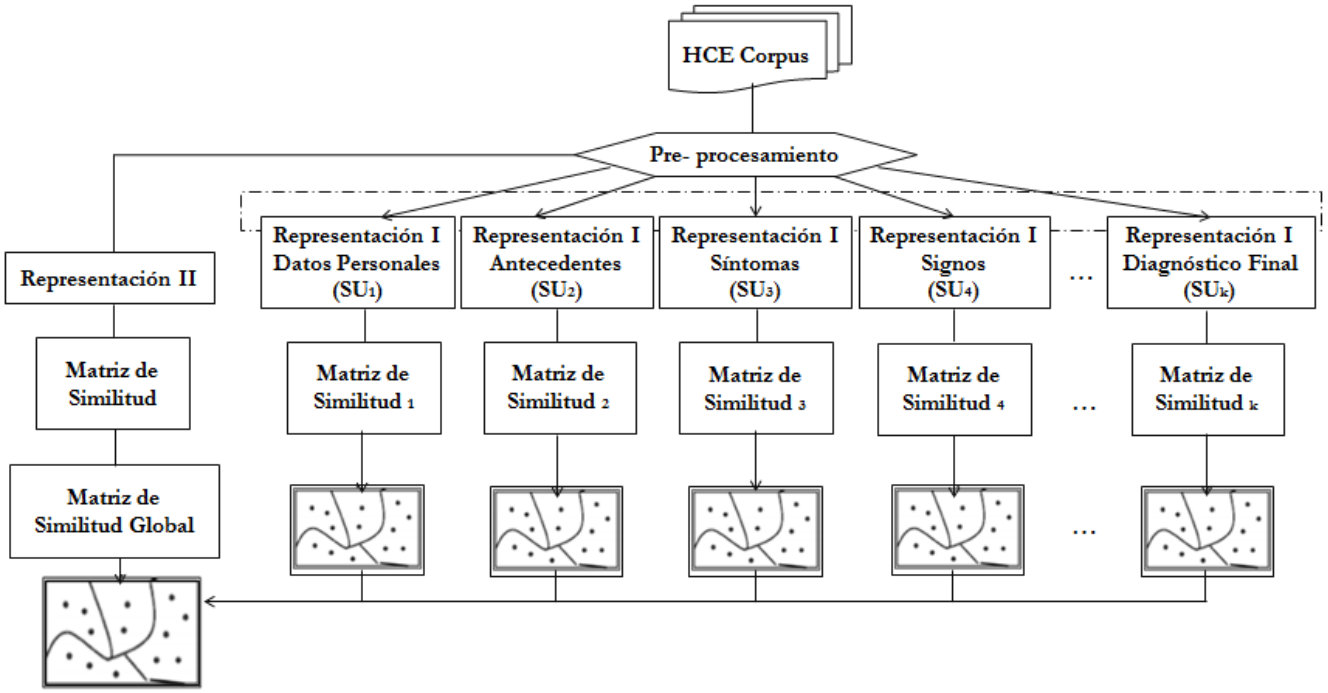

Figura 2. Esquema que muestra el modelo general propuesto.

Algorithm 1 Construcción de la matriz de similitud *Overall-SimSUX*

Input: Corpus D de documentos de HCE
Output: Grupos, calidad de los grupos, documento más representativo por grupo

1: Pre-procesamientos; /* análisis léxico, eliminación de palabras de parada, segmentación ...*/
2: Construir todas las k-colecciones (corpus D);
3: **for all** D_k **do**
4: Rep_I ← Hacer Representación I (DSU_k) mediante TF-IDF;
5: Matriz_Sim ← Calcular la matriz de similitud Rep_I usando la similitud Coseno;
6: Grupos ← Aplicar método de agrupamiento *K-Star* a Matriz_Sim;
7: **end for**
8: Rep_II ← Hacer Representación II corpus D completo usando ecuación (1) para calcular la frecuencia; /* Ver Tabla 2 */
9: Matriz_SimII ← Calcular la matriz de similitud para Rep_II usando la similitud Coseno;
10: Matriz_O_Sim ← Calcular matriz de similitud usando la medida *OverallSimSUX* teniendo en cuenta todos los agrupamientos para cada DSU_k y Matriz_SimII;
11: Obtener el agrupamiento final aplicando el método de agrupamiento *K-Star* a Matriz_O_Sim;

Tabla 2. Representación I donde $tf_{dj}(t_i)$ es la frecuencia de aparición absoluta del término t_i en el documento d_j.

	Término$_1$	Término$_2$	...	Término$_m$
HCE$_1$	$tf_{d1}(t_1)$	$tf_{d1}(t_2)$	...	$tf_{d1}(t_m)$
HCE$_2$	$tf_{d2}(t_1)$	$tf_{d1}(t_2)$	...	$tf_{d2}(t_m)$
...	...	...	...	...
HCE$_n$	$tf_{dn}(t_1)$	$tf_{dn}(t_2)$	...	$tf_{dn}(t_m)$

Se emplearán las siguientes ecuaciones

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

$$w_{kj} = \left(e^{-\frac{L_{SU}}{L_{Doc}}}\right)^{pot}, \tag{2}$$

$$S_{\text{coseno}}(d_i, d_j) = \frac{\sum_{k=1}^{m} d_{ik} d_{jk}}{\sqrt{\sum_{k=1}^{m} d_{ik}^2 \sum_{k=1}^{m} d_{jk}^2}}, \tag{3}$$

$$f(C, s_g, i, j) = \frac{\sum_{k=1}^{m} (w_k \lambda_{k(i,j)} + s_{g(i,j)})}{\sum_{k=1}^{m} w_k + 1}. \tag{4}$$

Agrupamiento de una k-colección A partir de la Representación I se obtiene una matriz de similitud que compara dos documento utilizando la medida coseno; calculada según la ecuación (3). Como resultado se obtiene para cada k-colección un agrupamiento independiente aplicando el algoritmo *K-Star*

clásico [22].

Cálculo de la matriz *OverallSimSUX* La similitud *OverallSimSUX*, se especifica formalmente en la ecuación (4). Esta se calcula teniendo en cuenta los resultados de los agrupamientos de cada *k*-colección y la matriz de similitud basada en el cálculo de la medida coseno para la Representación-II. *OverallSimSUX* considera *m* como la cantidad de SU identificadas en los documentos de HCE. Esta función de similitud es normalizada por la suma de los pesos de las *m* SU y el máximo valor de similitud global (e.g. 1). Por consiguiente, su máximo (e.g. 1) se alcanza cuando el documento de HCE *i* y *j* pertenecen al mismo grupo en todos los *k*-agrupamientos (e.g. k = 1) y el valor del s_g es máximo.

Agrupamiento Final Para el agrupamiento final se aplica el algoritmo *K-Star* a la matriz de similitud *OverallSimSUX*.

3.1 Clasificación Supervisada de un nueva HCE

A continuación se expone una variante para asistir la toma de decisiones diagnóstica ante la llegada un nuevo paciente. La idea no es la de dar un diagnóstico completo y definitivo, sino ayudar a los expertos a realizar el análisis diferencial de posibles enfermedades a partir del conocimiento descubierto por la metodología en colecciones de HCE representativas, con el propósito de conseguir un diagnóstico óptimo. La clasificación de la nueva HCE se realiza a partir del cálculo de la similitud del nuevo documento con cada uno de los $d_j \in D$, (D colección de HCE agrupada); se aplica el algoritmo de clasificación supervisada KNN para $k = 7$, con lo cual se tendrá los *k* ejemplos más cercanos. Estos casos más similares al caso analizado permitirán enfocarse en la US diagnóstico y realizar un análisis completo del paciente y su relación con los pacientes más similares a este y no descartar información menor que pudieran escapar del análisis humano, resaltando la información relevante que ayudará al especialista a no pasar por alto ningún detalle y llegar a un diagnóstico óptimo.

3.2 Interpretación del agrupamiento de HCE

Agrupar las HCE atendiendo a sus síntomas o signos y no únicamente teniendo en cuenta los diagnósticos finales, permiten al especialista estimar objetivamente el valor diagnóstico de una prueba determinada sin interferir resultados de otras pruebas. Es decir, inferir una vista coherente de la historia del paciente, de lo que realmente se ha hecho, por qué y qué ha sucedido. Los antecedentes patológicos familiares y personales y la reacción adversa ante determinados medicamentos en casos similares, permitiría explicar el porqué de las acciones realizadas por el médico para tratar a ciertos pacientes. A su vez ante la presencia de un nuevo caso sin un diagnóstico definitivo, encontrar pacientes similares a él, permitiría al médico valorar si los estudios de revisión basados en pruebas aplicados a estos pacientes similares, serían factibles aplicarlos a su actual paciente [4].

Por otra parte, el beneficio de disponer de pacientes similares con iguales diagnósticos finales, permitiría a los estudiantes en un menor tiempo completar la HC de un paciente. Utilizando la metodología propuesta y las SU de las HCE donde se concentran sus dudas, obtendría grupos de casos similares que constituyen recomendaciones respecto al uso correcto de un complementario, un plan de tratamiento, entre otras opciones.

Los resultados obtenidos mediante el agrupamiento de documentos de HCE pueden ser interpretados teniendo en cuenta el criterio de expertos. El uso de reglas de asociación permitirá explicar las relaciones entre HCE de pacientes que pertenecen a un mismo grupo. Los centroides o HCE más relevantes de cada grupo permitirá a los expertos estudiar casos similares con evoluciones favorables. Para verificar los efectos de la metodología en colecciones de HCE, se propone utilizar una muestra de 1.5 millones de HCE del archivo del servicio de admisión del Hospital .Arnaldo Milián Castro". La interpretación de los resultados obtenidos por la metodología debe evaluarse por expertos lo que evidencia la viabilidad de la metodología propuesta para la gestión de la información clínica y el descubrimiento de conocimiento implícito en ellas.

4. Conclusiones

En este trabajo se analizó la importancia del agrupamiento documental para el descubrimiento de conocimiento desde la información clínica. Debido a la necesidad de obtener conocimiento relevante que garantice el uso productivo de la información contenida en HCE, se propone una metodología para el agrupamiento de HCE concebidas como documentos XML, que combina sus dos dimensiones: estructura y contenido. Se muestra la función de similitud *OverallSimSUX* entre las HCE tomando como génesis la relación entre sus SU. Varias unidades estructurales se proponen para asegurar manejar una colección de HCE usando la metodología.

Referencias

[1] Engelbrecht R. K4Health. Knowledge for Health. Integrating EHR and Knowledge for better health care. Status of the EoI and work items. EUROREC. Berlin. 13-14 December.

[2] Brau, B., et al., 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., R. A. Verheij, F. J. Wiesman y R. D. Friele. 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. 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., M. Mesiti, and I. Sanz, An Overview of Similarity Measures for Clustering XML Documents. 2006.

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

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

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

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

[13] Kruse, R., C. Döring, and M.-J. Lesor, 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, and D. Yang, FXProj – A Fuzzy XML Documents Projected Clustering Based on Structure and Content. LNAI 7120, 2011: p. 406-419.

[15] Yousuke, W., K. Hidetaka, and Y. Haruo, 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. and P.J. Rousseeuw, Finding groups in data: an introduction to cluster analysis. Wiley Series in probability and mathematical statistics. 1990: John Wiley and Sons.

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

[18] Tekli, J.M. and R. Chbeir, A Novel XML Document Structure Comparison Framework based-on Subtree Commonalities and Label Semantics. Elsevier, 2011. 11(2011). doi: 10.1016/j.websem.2011.10.002

[19] Pinto, D., M. Tovar, and D. Vilariño. 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. et al. (2011). Overview of the INEX 2010 XML mining track: clustering and classification of XML documents, in In Lecture Notes in Computer Science, Springer: Amsterdam.

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

[22] Shin, K. and S.Y. Han, 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. Wu, and J. Chen. 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., et al., 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., I.E. Fuentes, and M.M. García, Clustering XML Documents using Structure and Content Based in a Proposal Similarity Function (OverallSimSUX). Computación y Sistemas, 2015.