Universidad Central “Marta Abreu” de Las Villas
Facultad de Matemática, Física y Computación

Trabajo para optar por el Título Académico Licenciado en Ciencia de la Computación

Razonador Basado en Casos para la inferencia de hipótesis diagnósticas, basado en la teoría de los conjuntos aproximados

Autor
Katterine Nodarse Morales

Tutores
MSc. Ivett Elena Fuentes Herrera
Dra. María Matilde García Lorenzo

Consultante
Dra. Leticia Arco García

Junio, 2017

Hago constar que el presente trabajo fue realizado en la Universidad Central “Marta Abreu” de Las Villas como parte de la culminación de los estudios de la Licenciatura en Ciencias de la Computación, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma total como parcial y que además no podrá ser presentado en eventos ni publicado sin previa autorización de la Universidad.

_____

Firma del autor

Los abajo firmantes, certificamos que el presente trabajo ha sido realizado según acuerdos de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

_____	_____
Firma del tutor	Firma del jefe del Seminario de Inteligencia Artificial

.

A mis padres.

A mi novio.

A mi tío y a mi tía.

A mis abuelos.

AGRADECIMIENTOS

A Dios.

A mis padres y novio por apoyarme en todo y confiar siempre en mí.

A mis tutoras por su entrega y apoyo incondicional todo el tiempo. A Ivett por enseñarme tanto, por su dedicación y guía en cada etapa de este trabajo. A la profesora Marilyn por su orientación, su visión y sabios consejos.

A mis abuelos por todo el apoyo y amor que me han dado desde que nací.

A mi tío y a mi tía por estar siempre pendientes de mí a la par de mi mamá y mi papá.

A mis dos grandes amigas Beatriz y Anllelys por siempre estar ahí cuando me ha hecho falta.

A mis amigos con los cuales he pasado estos 5 años en especial a Francisco por sacarme de tantos apuros.

A todos mis profesores, en especial al grupo de Inteligencia Artificial de quienes he aprendido mucho.

A todos, gracias.

RESUMEN

La información generada en los centros hospitalarios, a partir del uso extendido de la Historia Clínica electrónica es arrolladora. A esto se le añade, el crecimiento continuo del número de estudiantes vinculados a la asistencia médica, como consecuencia del envejecimiento de la población y el incremento de las enfermedades. Surge así, la necesidad de disponer de herramientas que permitan asegurar el uso productivo de la información almacenada y contribuyan a la formación de profesionales con un mayor nivel de competencia en la práctica del método clínico.

En este trabajo se implementó el sistema *CDARS* que soporta un modelo para el descubrimiento de conocimiento implícito en Historias Clínicas Electrónicas en formato semi-estructurado, basado en la hibridación de técnicas de agrupamiento y clasificación. El modelo concebido para la deducción de posibles hipótesis diagnósticas, en esta investigación ha sido implementado, considerando la propuesta *Rough Text*[1] basada en la teoría de conjuntos aproximados para enfrentar diversas tareas de la minería de textos. El sistema propuesto que implementa el modelo permite la manipulación de Historias Clínicas a partir del estándar CDA, integrando agrupamiento y clasificación automática y facilita el entrenamiento asistencial de estudiantes en los centros hospitalarios. Los resultados obtenidos luego de aplicar los experimentos, muestran como RST contribuye satisfactoriamente al comportamiento del modelo en la inferencia realizada a partir del método de solución de problemas RBC.

[1] Propuesta en la investigación doctoral de la Dra. Leticia Arco García

ABSTRACT

The information obtained from hospital centers by the extended use of the electronic medical histories is overwhelming. Besides, the constant growth of the number of students involved in the medical assistance due to the aging of the population, and the increase of diseases should be mentioned. Therefore, it is necessary to have at one's disposal tools that allow to guarantee the efficient use of the stored information, and that contribute to the training of professionals with a higher competence level in the practice of the clinical method.

In this project, the system CDARS was implemented. This system adopts a model for discovering implicit knowledge in electronic medical histories in a semi-structured format, based on the hybridization of clustering and classification techniques. The conceived model for deducting the possible diagnosis hypothesis has been implemented in this research taking into account the proposal Rough Text based on the Rough Set Theory, to face multiple tasks while text mining. The proposed system, which implements the model, allows the manipulation of medical histories from the CDA standard, integrating clustering and automatic classification, and facilitates the assistance training of students in hospital centers. The results obtained after applying the experiments, show that RST satisfactorily contributes to the model performance in the inference carried out from the problem solving method RBC.

Tabla de Contenidos

INTRODUCCIÓN **1**

1. RESEÑA DE LOS MÉTODOS DE DESCUBRIMIENTO DE CONOCIMIENTO DESDE INFORMACIÓN CLÍNICA SEMI-ESTRUCTURADA Y LA TEORÍA *ROUGH TEXT* **7**

1.1 Gestión de la información inteligente *7*

1.2 Documentos clínicos semi-estructurados *8*

1.3 Técnicas de minería de textos para documentos *10*

1.3.1 Agrupamiento de documentos **11**

1.3.2 S-Space **13**

1.4 Modelo híbrido para la deducción de hipótesis diagnósticas *14*

1.4.1 Organización de la Base de Casos **15**

1.4.2 Motor de Inferencia **15**

1.5 Rough Text ..... ***¡Error! Marcador no definido.***

1.6 Consideraciones finales del capítulo *18*

2. EXTENSIÓN DE UN MODELO DE DESCUBRIMIENTO DE CONOCIMIENTO A PARTIR DE REGISTROS CLÍNICOS, BASADO EN LA TEORÍA DE LOS CONJUNTOS APROXIMADOS **21**

2.1 Construcción de la Base de Casos *21*

2.1.1 Conocimiento para la construcción de la Base de Casos ..... ¡Error! Marcador no definido.

2.1.2 Base de Casos **22**

2.1.3 Organización de la Base de Casos **23**

2.1.3.1 Sistema de decisión necesario para aplicar RST *24*

2.1.3.2 Refinamiento del agrupamiento *26*

2.1.3.3 Determinación de prototipos *28*

2.2 Motor de Inferencia *29*

2.2.1 Módulo recuperador **29**

2.2.2 Módulo adaptador **30**

2.2.3 Módulo Revisor **31**

2.3 Procedimiento general del RBC que soporta CDARS *32*

2.3.1 Módulo 1: Recuperación y creación de índices a partir del corpus de HCE **33**

2.3.2 Módulo 2: Representación de las UE **33**

2.3.3 Módulo 3: Similitud y Agrupamiento para cada UE **34**

2.3.4 Módulo 5: Razonamiento Basado en Casos **35**

2.3.4.1 Recuperación *35*

2.3.4.2 Adaptación *35*

2.3.4.3 Revisión *36*

2.4 Consideraciones finales del capítulo *37*

3. EVALUACIÓN DE LA EXTENSIÓN PROPUESTA Y DESCRIPCIÓN DEL SISTEMA CDARS **39**

3.1 Evaluación del RBC *39*

3.1.1 Experimento 1 ..... ¡Error! Marcador no definido.

3.1.2 Experimento 2 ..... **39**

3.1.3 Experimento 3 ..... ¡Error! Marcador no definido.

3.1.4 Experimento 4 ..... **43**

3.1.5 Experimento 5 ..... **44**

3.2 Implementación del modelo *45*

3.2.1 *Recuperación de la información* **45**

3.2.1.1 Extracción de la información *45*
3.2.1.2 Indexación y búsqueda *45*
3.2.2 *Pre-procesamiento* **46**
3.2.3 *Representación* **46**
3.2.4 *Agrupamiento* **46**
3.2.5 *Recuperación* **47**
3.2.6 *Adaptación* **47**
3.2.7 *Interfaz de usuario* **47**
3.3 Diseño del Sistema CDARS *47*
3.4 Interfaz de usuarios del sistema CDARS para la recomendación de hipótesis diagnósticas *50*
3.5 Conclusiones finales del capítulo *55*

CONCLUSIONES **56**

RECOMENDACIONES **57**

REFERENCIAS BIBLIOGRÁFICAS **58**
Anexo 1. Algunas medidas de calidad de términos ..... ***¡Error! Marcador no definido.***
Anexo 2. Similitudes, distancias más usadas para comparar objetos y medidas de calidad *62*
Anexo 3. HCE basada en el estándar CDA-HL7 ..... *64*
Anexo 4. Encuesta realiza a los expertos *66*
Anexo 5. Resultados de la encuesta realiza a los expertos *68*
Anexo 6. Caracterización de los expertos *66*
Anexo 7. Variantes para el cálculo del umbral de similitud entre objetos *65*
Anexo 8. Cálculo de la Pertenencia Aproximadd *64*

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 Tecnologías de la Información y las Comunicaciones (TIC) en las organizaciones. 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 (Sebastián et al., 2006).

En el contexto de la actividad hospitalaria, la Historia Clínica (HC) constituye la fuente esencial de datos y el documento principal del Sistema de Información Hospitalario (*Hospital Information System*, HIS). La propia distribución de los elementos en el documento clínico permite concebir la HC como un documento XML, debido a la estructura jerárquica y auto-descriptiva implícita en cada uno de los factores que la componen. Los HIS apuestan por la implementación de estándares de codificación semi-estructurados debido, a la ventaja que brinda etiquetar los documentos para el acceso a partes específicas de estos. Específicamente, *Health Level Seven* (HL7), el conjunto de estándares informáticos de salud más desarrollado y de mayor cobertura internacional para dar soporte a la HCE, se sustenta en el metalenguaje XML (Zwaanswijk et al., 2011).

El estándar *Clinical Document Architecture* (CDA) de HL7 especifica la estructura y semántica de los documentos clínicos con el propósito de facilitar su intercambio en un entorno de interoperabilidad (Aruquipa, 2014). CDA-HL7, estándar de marcaje, realizado por el comité *Structured Documents Technical Committee* (SDTC) de HL7, 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, logrando 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 presentarse directamente a los navegadores Web compatibles con XML. Es posible crearlos y validarlos mediante una plantilla XML o *Schema* (Alonso, 2014). 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 (Dolin et al., 2001). Por lo que al seguir una estructura común, bien definida, la consulta de estos campos de forma automatizada es fácil (Denny et al., 2009).

Debido a que gestionar el conocimiento a partir de la información almacenada es fundamental en la práctica clínica (Zwaanswijk et al., 2011); se hace inevitable crear técnicas para el análisis eficiente de grandes colecciones de este tipo de documentos y extraer conocimiento relevante. Existen varias funciones para descubrir conocimiento entre las que se encuentran 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 (C.D. et al., 2008).

Por otra parte, cuando se desea considerar la clasificación basado en la inferencia de posibles alternativas diagnósticas a considerar a la llegada de un nuevo caso, constituye el Razonamiento Basado en Casos (RBC) uno método adecuado, al ser de los que más se asemeja al procedimiento que emplea el médico mediante el método clínico tradicional para emitir un diagnóstico, ya que basa su funcionamiento en experiencias anteriores, mediante una asociación de los casos similares en la construcción de la nueva solución.

A esto se añade que, la valoración de los resultados de los agrupamientos, sobre todo en dominios textuales, está influenciada por la subjetividad. Por otro lado, la Teoría de los Conjuntos Aproximados (*Rough Set Theory*, RST) al no requerir de información adicional o preliminar sobre el conjunto de datos, ni suposición sobre éstos, incluso en circunstancias caracterizadas por vaguedad e incertidumbre, constituye una técnica que puede contribuir positivamente al proceso de validación, etiquetamiento y fusión de los grupos resultantes del

agrupamiento de documentos; así como para obtener los subconjuntos de elementos más representativos de los grupos de documentos identificados.

Ante la necesidad de incorporar estos métodos de aprendizaje automatizados en el núcleo de la actividad hospitalaria, ofreciendo soporte a la HCE; en el Centro de Estudios de Informática (CEI) de la Universidad Central "Marta Abreu" de las Villas (UCLV) se propuso un modelo para el descubrimiento de conocimiento implícito en Historias Clínicas Electrónicas en formato semi-estructurado, que define una variante para asistir la toma de decisiones diagnóstica de un nuevo paciente, mediante una clasificación supervisada que utiliza RBC para la clasificación de documentos. El modelo integra el agrupamiento y la clasificación automática en dependencia de los criterios del decisor y con ello, asiste al personal médico en la toma de decisiones y propicia el desarrollo de herramientas para el entrenamiento asistencial de estudiantes en los centros hospitalarios.

Por otra parte se dispone de una propuesta basada en la RST para enfrentar diversas tareas de la minería de textos *Rough Text* que a su vez puede contribuir al comportamiento del modelo en la inferencia realizada a partir del método de solución de problemas RBC y brindar amplias ventajas que permitan mejorar la calidad de las soluciones a partir de los elementos antes mencionados.

Lo antes expuesto ratifica la problemática que justifica el siguiente **planteamiento de investigación**:

La necesidad de disponer de una partición del conjunto de objetos en grupos que sea adecuada y contribuya a realizar inferencias a la llegada del paciente. Superar estos desafíos es esencial para dar a los especialistas y estudiantes mejores condiciones para el ejercicio de una medicina basada en la práctica del método clínico tradicional.

El **objetivo general** de esta investigación consiste en:

Desarrollar una variante de RBC para contribuir a la toma de decisiones y el entrenamiento asistencial a partir de registros clínicos basado en la teoría de conjuntos aproximados.

Este objetivo se desglosa en los siguientes **objetivos específicos**:

1. Refinar grupos de documentos a través del uso de RST para la obtención de un agrupamiento de mayor calidad.

2. Definir una organización de la base de casos.
3. Definir un enfoque de revisión que permita manejar casos atípicos.
4. Evaluar el sistema basado en casos a partir de un caso de estudio de colecciones de HCE.

Las **preguntas de investigación** planteadas son:

1. ¿Cómo determinar cuáles son los documentos más representativos de un grupo que utiliza en modelo en la fase de recuperación?
2. ¿Cómo refinar la organización de la BC que favorezca retroalimentar una inferencia realiza ante casos de difícil diagnóstico?
3. ¿Cómo combinar los resultados obtenidos de una inferencia realizada a la estructura definida para la BC?
4. ¿Cómo validar el efecto del modelo integrado en las soluciones obtenidas?

Como respuestas a las preguntas de investigación y después de haber realizado el marco teórico se formuló la siguiente **hipótesis de investigación**:

H1: La combinación del modelo para el descubrimiento de conocimiento en HCE en formato semi-estructurado, combinado la relación estructura-contenido y la propuesta basada en la teoría de conjuntos aproximados para enfrentar diversas tareas de la minería de textos, facilita eficazmente la toma de decisiones para el personal médico y el entrenamiento asistencial de estudiantes en los centros hospitalarios.

La **novedad científica** principal que aporta esta investigación, radica en la combinación de los métodos de aprendizaje automático para la manipulación de HCE en formato semi-estructurado que conforman el modelo y la propuesta basada en la teoría de conjuntos aproximados *Rough Text* para refinar la inferencia.

La implementación del sistema *CDARS* tiene un **valor práctico**, pues permite asistir la toma de decisiones diagnóstica y el entrenamiento asistencial de los estudiantes en los centros de asistencia a partir de las inferencias realizadas por el modelo propuesto, contribuyendo de esta forma a la formación de profesionales con un mayor nivel de competencia en la práctica del método clínico.

El **valor teórico** de la investigación está directamente vinculado con su novedad científica.

La **tesis** está **estructurada** en tres capítulos. En el Capítulo 1 se realiza una reseña de los principales elementos que conforman el modelo para el descubrimiento propuesto que se implementa en esta tesis y la propuesta basada en la teoría de los conjuntos aproximados *Rough Text*. Se hace énfasis en las bondades de las aproximaciones inferiores y superiores, para validar y refinar el agrupamiento y la búsqueda de prototipos. En el Capítulo 2 se detalla sobre la extensión realizada del modelo combinando los elementos de la teoría Rough Text. En el Capítulo 3 se presentan los resultados experimentales y una descripción del sistema a nivel de usuario con el propósito de explicar cómo utilizar el entrenador CDARS. Este documento culmina con las conclusiones, las recomendaciones, referencias bibliográficas y los anexos.

1

RESEÑA DE LOS MÉTODOS DE DESCUBRIMIENTO DE CONOCIMIENTO DESDE INFORMACIÓN CLÍNICA SEMI-ESTRUCTURADA Y LA TEORÍA *ROUGH TEXT*

1. RESEÑA DE LOS MÉTODOS DE DESCUBRIMIENTO DE CONOCIMIENTO DESDE INFORMACIÓN CLÍNICA SEMI-ESTRUCTURADA Y LA TEORÍA *ROUGH TEXT*

La gran cantidad de información clínica generada en los centros hospitalarios y la riqueza de información que estas poseen, hace que sea difícil para los motores de búsqueda identificar la información relevante (Aruquipa, 2014) y asegurar su uso productivo (Denny et al., 2009). Crear un conjunto de normas técnicas y prácticas para administrar el flujo de documentos y establecer una organización entre ellos, constituye una de las tareas fundamentales en el área de la gestión del conocimiento y que ofrece ventajas en la formación del profesional de la salud desde la propedéutica clínica y la medicina interna (Cruz Hernández et al., 2012).

En este capítulo, se realiza una reseña de las principales técnicas de aprendizaje automatizado que facilitan el descubrimiento de conocimiento, enfatizando en un modelo para el descubrimiento de conocimiento propuesto en el CEI. Se hace énfasis en las bondades de *Rough Text* para enfrentar diversas tareas de la minería de textos.

1.1 Gestión de la información inteligente

Las enormes cantidades de datos generados cada día en las instituciones, imponen un gran reto, lograr una gestión de los datos que los transforme en información inteligente y viabilice el proceso de toma de decisiones. Por lo que, sin importar cuánto se han desarrollado los sistemas automatizados para el manejo de información, se reconoce que la complejidad mayor reside en tomar las mejores decisiones y extraer conocimiento relevante desde la información disponible; al ser la información el medio y el entorno que permite el conocimiento (Visval and Sara, 2009).

La Gestión Documental (GD) define que elementos considerar en el proceso de administrar el flujo de documentos de las organizaciones(Grossman and Frieder, 2012), de este modo los Sistemas de GD (SGD), tienen el propósito de implementar técnicas que permitan relacionar los documentos y darles una semántica común; posibilitando la búsqueda de información dentro de las colección y ofreciendo documentos relacionados; que conformen una base operativa de colaboración. Esto significa que una aplicación de GD está orientada a un contexto operacional que tenga relevancia para cualquier organización. Los SGD por lo general, se refieren a las

siguientes áreas: almacenamiento, recuperación, agrupamiento y clasificación, seguridad, custodia, distribución, creación y autenticación de los documentos. En general, un sistema de GD pretende: facilitar el trabajo con los documentos, facilitar que la documentación y la información que contienen se comparta y se aproveche como un recurso colectivo y conservar la memoria de la organización (Grossman and Frieder, 2012).

Gestionar el conocimiento requiere de la integración de varias áreas del saber: el descubrimiento de conocimiento, la minería de datos y de textos, las que pretenden lograr una visión selectiva y perfeccionada de la información contenida en documentos escritos. Específicamente, el descubrimiento de conocimiento integra la recuperación y extracción de información, el análisis de textos, el resumen, la categorización, la clasificación, el agrupamiento, la visualización, entre otras (Dixon, 1997, Tan, 1999). De todos ellos, es el agrupamiento una técnica esencial para organizar la información, determinar información relevante y crear nuevo conocimiento a partir de la información existente en las colecciones de documentos (C.D. et al., 2008). Estos últimos, se clasifican de acuerdo a su estructura como: estructurados, no estructurados y semi-estructurados (Magdaleno et al., 2015, Magdaleno, 2013). Particularmente los documentos con formato semi-estructurado han tenido un crecimiento exponencial en las instituciones ya que permiten una mejor manipulación aumentando de esta manera la eficacia de los sistemas de GI (Lemon et al., 2005).

1.2 Documentos clínicos semi-estructurados

La HC constituye el documento principal de un SIH, 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 (Barreto 2015). 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.

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. (Pathak et al., 2013, Zwaanswijk et al., 2011, Denny et al., 2009).

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, está sustentado en el metalenguaje XML (Zwaanswijk et al., 2011). Específicamente, *Clinical Document Architecture* (CDA) de HL7 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.

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 (Dolin et al., 2006). 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.) Estos dos últimos garantizan interoperabilidad, permitiendo que los documentos sean procesables, mediante búsquedas y técnicas de Minería de Datos[2].

[2] *ISO/HL7 27932:2009 - Data Exchange Standards -- HL7 Clinical Document Architecture, Release 2*

Estas ventajas hacen que CDA sea uno de las especificaciones más utilizadas regionalmente en los sistemas de gestión de información clínica. En la Figura 1.1 se observa un ejemplo de una HCE correspondiente a un documento XML definido en el estándar CDA.

En todo el mundo se encuentran en desarrollo varias iniciativas sobre estándares de información para el área salud. Por su parte HL7, al ser un estándar no propietario y reconocido por el Comité Técnico de Información para Salud de la Organización Internacional para la Estandarización (ISO) es muy utilizado (Ripoll Garrido, 2011).

```
<?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>
```

Figura. 1.1: Ejemplo de un documento XML basado en el estándar CDA-HL7

1.3 Técnicas de minería de textos para documentos

Con el desarrollo de la Inteligencia Artificial y específicamente la Minería de Textos en la actualidad pueden ser procesados grandes volúmenes de información con el objetivo de extraer patrones que residan en los datos almacenados (El-Sayed, 2013). Las técnicas más utilizadas se dividen en tres grandes grupos: la clasificación, la categorización y el agrupamiento.

La *clasificación* comprende la distribución de los objetos de cualquier género de clases, sobre la base de rasgos diferenciales correspondientes. Al clasificar documentos se realiza un análisis de su contenido y forma, y se sitúan en grupos mediante un sistema de clasificación desarrollado

con estos fines (Patidar et al., 2012). Una amplia gama de clasificadores están disponibles, cada uno con sus fortalezas y debilidades y su rendimiento depende en gran medida de las características de los datos que deben clasificarse. Entre estos podemos distinguir, el algoritmo de los *k*-vecinos más cercanos y el Razonamiento Basado en Casos.

La *categorización* de documentos es un tipo de problema perteneciente a la familia de problemas asociados a encontrar agrupamientos entre objetos de cualquier tipo. Si bien la categorización de documentos tiene características particulares que surgen de las propiedades de los documentos como objetos a agrupar, los principios generales coinciden con los que se aplican para categorizar cualquier otro tipo de elementos (Taghva and Veni, 2010).

El *agrupamiento* es descrito como una herramienta para el descubrimiento porque tiene la potencialidad de revelar relaciones basadas en datos complejos no detectadas previamente (Kruse et al., 2007, Anderberg, 1973, Shankar, 2012). El agrupamiento es fundamental para una eficiente organización y recuperación de los documentos relevantes.

A continuación, se realiza una panorámica sobre el agrupamiento de documentos, enfatizando las técnicas especializadas en los XML.

1.3.1 Agrupamiento de documentos

Un algoritmo de agrupamiento, intenta encontrar grupos naturales basándose principalmente en la similitud y relaciones de los objetos, de forma tal que se obtenga una distribución interna del conjunto de datos mediante su particionamiento en grupos, es decir, sigue el principio de maximizar la similitud dentro del grupo y minimizar la similitud entre ellos (Anderberg, 1973, Kruse et al., 2007).

Una medida de similitud o función de similitud es una función real que cuantifica la similitud entre dos objetos. Aunque no existe una definición única, por lo general las medidas de similitud son en algún sentido la inversa de la disimilitud o distancia métrica (Aggarwal and Reddy, 2013).

Algunas de las funciones más conocidas para el trabajo con documentos son: *Coseno* (Huang, 2008), *Jaccard* (Taghva and Veni, 2010) y *Dice* (Strehl et al., 2000), entre otras (Chahal and

Singh, 2013). En el Anexo 1 se muestra una selección de similitudes o distancias más usadas para comparar objetos.

Varios enfoques de agrupamiento continúan emergiendo con la finalidad de obtener agrupamientos con mayor precisión, por lo que esta área es reconocida por varios investigadores como un campo abierto (Vries et al., 2011). Entre ellos los algoritmos de agrupamiento duros y borrosos (Höppner, Klawonn et al. 1999). Las técnicas duras pueden ser deterministas o con solapamiento. Las deterministas crean una partición, donde los grupos son mutuamente excluyentes y exhaustivos del universo de objetos. Dentro de esta clasificación pueden incluirse los de tipos jerárquicos que crean un dendograma de clasificación que permite seleccionar una partición atendiendo a la calidad basado en la composición de los grupos en un nivel de la jerarquía. Estos a su vez, pueden clasificarse en algoritmos de agrupamiento jerárquicos aglomerativos (Hierarchical Agglomerative Clustering; HAC) o divisivos (Manning et al. 2008).

De manera que, dado un conjunto de *N* objetos para agrupar y una matriz de distancia o similitud el algoritmo básico de agrupamiento jerárquico consiste en:

1. Asignar a cada objeto su propio grupo (conformar *N* grupos, cada uno con un solo objeto).
2. Encontrar el par de grupos (más similar) y mezclarlos en un único grupo.
3. Repetir el paso 2 hasta que todos los objetos son agrupados en un único grupo.

El agrupamiento de enlace promedio calcula la similitud entre dos grupos como el promedio de la similitud entre los pares de objetos de un grupo y otro. Este proceso de enlace promedio es más lento que el agrupamiento de enlace único porque se necesita determinar la similitud promedio entre una gran cantidad de pares de objetos para determinar la similitud del grupo. Por otra parte, es más robusto que enlace único en cuanto a la calidad del agrupamiento (Aggarwal & Zhai 2012). Se ha recomendado como mejor algoritmo para el agrupamiento de documentos en representaciones vectoriales (Manning et al. 2008)

Así, esta técnica puede contribuir a una eficiente organización y recuperación de los documentos relevantes en la gestión hospitalaria.

No obstante, esta gestión requiere de otras funciones para emitir recomendaciones que contribuyan a asistir la toma de decisiones diagnóstica ante la admisión de un nuevo paciente.

En este sentido, algunos sistemas han sido propuestos utilizando un enfoque de RBC, tales como: CASEY (Bonillo, 2003); PROTOS (Bareiss Jr, 1988) y PARIS (Bateman, 1990). En (Lorenzo and Bello, 1996) se avalan las potencialidades del RBC para sistemas de diagnóstico médico, debido a que el RBC es un método de aprendizaje que responde a la forma de razonar de muchos especialistas entre los que se encuentran los médicos. En el próximo epígrafe se muestra una propuesta que híbrida técnicas de clasificación y agrupamiento para inferir hipótesis diagnósticas a la llegada del paciente.

1.3.2 **S-Space**

La herramienta *S-Space Package* es una biblioteca de código abierto y libre para desarrollar y evaluar algoritmos de espacio palabra. Los algoritmos son divididos en cuatro categorías basándose en su similitud estructural:

- Modelos basados en documentos: dividen el corpus en documentos discretos y construyen un VSM a partir de las frecuencias de las palabras en los documentos. Por ejemplo: VSM, LSA.
- Modelos basados en coocurrencia: construyen el espacio vectorial usando la distribución de palabras coocurrentes en un contexto, el cual puede ser definido como una región alrededor de una palabra o caminos en un árbol gramatical. Por ejemplo: HAL, COALS (Rhode, 2009).
- Modelos basados en aproximación: aproximan datos de coocurrencia para lograr mejor escalabilidad de grandes conjuntos de datos. Por ejemplo: Random Indexing y RRI (Cohen, 2010).
- Modelos basados en inducción del sentido de las palabras: intentan descubrir sentidos diferentes de las palabras mientras construyen un espacio vectorial. Por ejemplo: Purandare and Pedersen (Purandare and Pedersen, 2004).

En esencia, la idea de estos modelos es que los rasgos de palabras se extraen de un corpus y la distribución de estos rasgos es usada como base para la similitud semántica. Para las matrices se utilizan esquemas de peso como TF-IDF y PMI. Poseen, además, algoritmos de agrupamiento de tipo aglomerativo jerárquico, agrupamiento espectral y es posible la integración con la

biblioteca CLUTO[3]. Algunas de las medidas de similitud que posee son la medida coseno, Euclidiana, Jaccard y KL divergence (Jurgens and Stevens, 2010).

1.4 Modelo híbrido para la deducción de hipótesis diagnósticas

El modelo propuesto se basa en la hibridación de técnicas de clasificación y agrupamiento. El mismo parte de una colección de HCE en formato semi-estructurado, la cuales deben ser pre-procesadas, etapa en la cual se identifican la Unidades Estructurales (UE), secciones de la HCE con significado semántico. Consecuentemente los documentos pertenecientes a una colección de HCE de entrada $D = \{D_1,\ldots, D_m\}$, constan con un conjunto de Unidades Estructurales (UE) $UE = \{U_1,\ldots, U_n\}$.

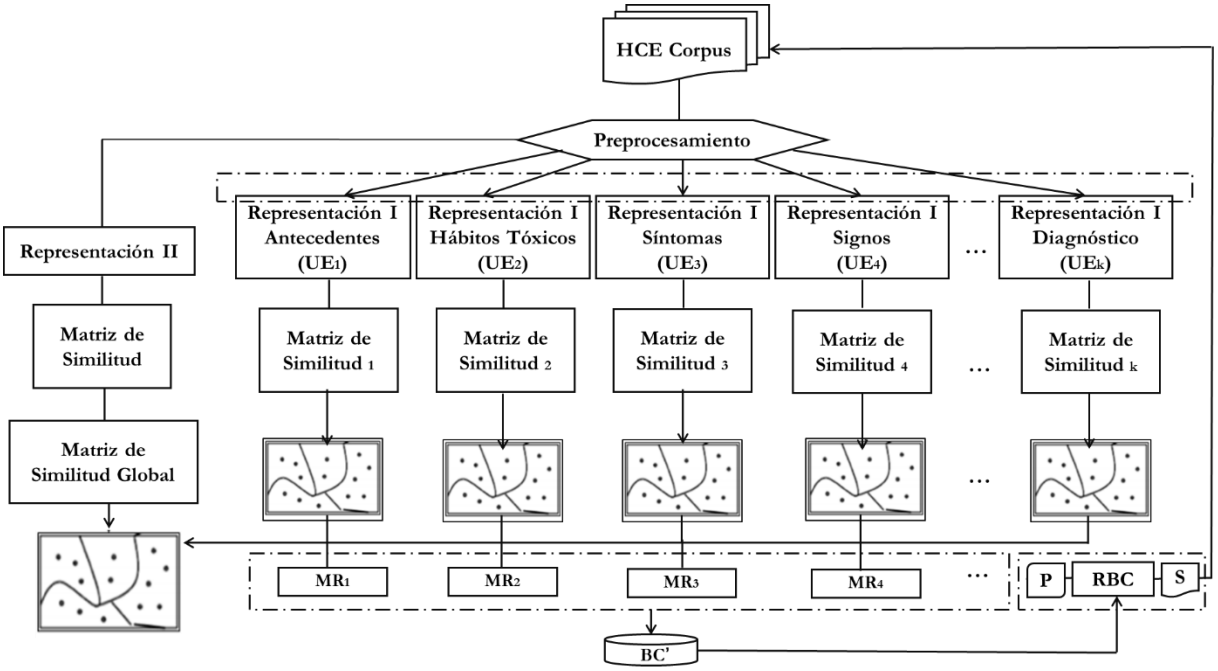

Figura 1.2. Modelo para el descubrimiento de conocimiento a partir HCE semi-estructuradas para la toma de decisiones

Para la aplicación del modelo se transforma el corpus de HCE convirtiendo cada uno los ficheros de entrada CDA en una secuencia de *tokens*[4] de palabras. En el paso subsecuente a la extracción

[3] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

[4] En este trabajo se tratarán indistintamente los vocablos términos, token o palabras.

de términos, estos *tokens* se usan para generar rasgos significativos (índices de términos). En esta etapa de trasformación del corpus se determinan las UE a las que pertenece cada término que aparece en el documento, además se identifican *k* colecciones independientes, dividiendo la colección original en *n* colecciones, donde *n* es el número de UE en un documento. El concepto de *k-colección* (Magdaleno et al., 2015) refleja la correspondencia entre la colección y la UE.

En la etapa de representación cada documento clínico se identifica como un vector de rasgos en un espacio en el cual cada dimensión corresponde a términos indexados distintos. Un vector documento, en cada componente tiene un valor numérico para indicar su importancia. En el modelo la *Representación I* se realiza para cada para cada *k-colección*.

Luego se obtiene la matriz de similitud aplicando una medida de similitud Coseno(Fuentes et al., 2016) a partir de la *Representación I* para cada UE, y para realizar los agrupamientos a utilizar en la organización de la BC sobre la cual realiza la inferencia el RBC.

1.4.1 Organización de la Base de Casos

Debido a que, el número de casos de la BC puede incrementarse considerablemente en el tiempo y el sistema de RBC debe ser capaz de obtener respuestas en tiempo razonable, como consecuencia la forma de organización y recuperación de la BC determinará el tiempo requerido para obtener los casos más similares al nuevo problema(Fuentes et al., 2016).

El modelo propone una variante que independientemente del hardware utilizado permite obtener respuestas en un tiempo razonable, al reducir el espacio de búsqueda, con lo cual no se limita su aplicación independientemente de las características de infraestructura informática en las unidades del Sistema de Salud Cubano y posibilitar su empleo en ambientes educativos.

Este modelo propone el uso de una estructura a partir de conformación de grupos de documentos semi-estructurados de HCE, obtenidos a partir del agrupamiento de las UE correspondientes a las colecciones independientes que constituyen rasgos predictores.

1.4.2 Motor de Inferencia

En la etapa de recuperación, se seleccionan el o los casos más similares al nuevo caso, considerando como más similares aquellos casos cuya similitud supera el umbral de similaridad. Este proceso se realiza teniendo en cuenta los prototipos de los grupos de documentos obtenidos

por el agrupamiento de las colecciones de HCEs. La recuperación se basa en una similitud local que determina la analogía entre valores de un mismo rasgo y una similitud global, que combina los resultados de las similitudes locales a todos los rasgos de los casos a comparar. La similitud global es el resultado de la suma ponderada de las similitudes entre las UE correspondientes de cada caso y el caso problema, según se observa en la ecuación 1.1.

$$\mathrm{SimGlobal}(\mathrm{X},\mathrm{Y}) = \sum_{\mathrm{i}=0}^{\mathrm{m}} W_i * \mathrm{SimCoseno}_{\mathrm{i}}\left(\mathrm{X}_{\mathrm{UE}_i}, Y_{\mathrm{UE}_i}\right)/\mathrm{n} \qquad \textbf{1.1}$$

$$\text{donde: } \sum W_i = 1$$

Luego se emiten consideraciones a partir de los criterios aportados a los casos recuperados como similares para elaborar una solución, que implica emitir posibles hipótesis diagnósticas del que se analiza.

El enfoque basado en casos del modelo propuesto, para un nuevo problema infiere hipótesis diagnósticas que deben ser revisadas por el experto luego de las pruebas complementarias, que permiten confirmar la veracidad de un diagnóstico. Por lo que, la etapa de revisión puede ser empleada cuando se presenta un nuevo problema al razonador, en los casos en que el experto exija una crítica de la sugerencia de diagnósticos posibles o cuando se infiere que no se tiene suficiente información a partir de los prototipos identificados y es necesaria una retroalimentación.

1.5 Conjuntos Aproximados

RST fue introducida por Z. Pawlak en 1982 (Pawlak, 1982a). Se basa en aproximar cualquier concepto a un subconjunto duro del dominio (por ejemplo, una clase de un problema de clasificación o un grupo resultante de un proceso de agrupamiento), por un par de conjuntos exactos, llamados aproximación inferior y superior del concepto aproximado.

Los conjuntos aproximados consideran que, todo objeto *x* de un universo *U* tiene asociada una cierta cantidad de información, expresada por atributos que describen el objeto(Bazan et al., 2004). La estructura de información básica de esta teoría es el sistema de información; par (*U*,

A) donde *A* es el conjunto de atributos o rasgos y *U* es un conjunto no vacío llamado universo de objetos descritos usando los atributos.

Particularmente, *Rough Text* constituye una propuesta basada en esta teoría para enfrentar diversas tareas de la minería de textos(Arco et al., 2006a).

La representación del corpus y el resultado del agrupamiento como atributo de decisión, conforman el sistema de decisión necesario para aplicar RST, según la propuesta. Se utiliza una relación de similitud *R*' debido a que dos documentos del universo (*U*) pueden ser similares pero improbablemente iguales. Así para determinar cuán similares son los documentos se utiliza una de las medidas de similitud propuestas en la literatura para comparar documentos (Anexo 1).

Luego, si *D* el conjunto de documentos del corpus textual y *s*: $D \times D \rightarrow \mathbf{R}$ una función que mide la similitud entre los objetos de *D*, se considera la relación *R*' (clase de similitud del objeto *x*) como sigue: $R'(x)=\{y \in U : yR'x$, *y* está relacionado con *x* sí y solo si $s(x,y)>\xi\}$, ξ es el umbral de similitud. Es necesario calcular *R*'(*x*) para cada documento en *U*. Así, es posible calcular las aproximaciones inferior y superior de un grupo de documentos según las expresiones 1.2 y 1.3, respectivamente.

$$R'_*(C_j) = \{D_i \in C_j : R'(D_i) \subseteq C_j\} \qquad \textbf{(1.2)}$$

$$R'^*(C_j) = \bigcup_{D_i \in C_j} R'(D_i) \qquad \textbf{(1.3)}$$

Los documentos en $R'_*(C_j)$ pueden ser clasificados con certeza como miembros del grupo C_j y son los más representativos de ese grupo, mientras que los documentos en $R'^*(C_j)$ pueden ser clasificados como posibles miembros del grupo y son aquellos que posiblemente pertenezcan a él.

La teoría de los conjuntos aproximados también tiene gran utilidad para evaluar la calidad del agrupamiento, en este sentido varias medidas internas han sido propuestas(Arco et al., 2006b).

Una forma de medir la pertenencia de un documento a un grupo, es la función de pertenencia aproximada (Pawlak and Skowron 1994). En la expresión (1.4) se muestra que ésta cuantifica el grado de solapamiento relativo entre $R'(x)$ y el grupo al cual el documento x pertenece. Esta función puede interpretarse como una estimación basada en frecuencias de la probabilidad

condicional de que el documento *y* pertenezca al conjunto X, dados los valores del documento x con respecto al conjunto de atributos. El valor $\mu_x(x)$ mide el grado de inclusión del documento x en el grupo X (Skowron 2000; Grabowski 2004).

$$\mu_x(x) = \frac{|X \cap R'(x)|}{|R'(x)|} \qquad \textbf{1.4}$$

$$w_i(x) = \frac{\sum_{x \in X_i} \mu_{X_i}(x)}{|X_i|} \qquad \textbf{1.5}$$

Otra expresión para medir la pertenencia aproximada de un documento *x* a un grupo *X*, cuantificando en qué grado la clase de similitud de *x (R'(x))* es capaz de cubrir el grupo *X* se utiliza en (Arco et al., 2006b, Caballero et al., 2007). Las expresiones (1.6) y (1.7) muestran su forma de cálculo y la variante de ponderación que se deriva de ella, respectivamente.

$$V_X(x) = \frac{|X \cap R'(x)|}{|X|} \qquad \textbf{1.6}$$

$$w_i = \frac{\sum_{x \in X_i} V_{X_i}(x)}{|X_i|} \qquad \textbf{1.7}$$

A partir del análisis de las dos variantes comentadas para el cálculo de la pertenencia de los documentos, la pertenencia de los documentos a los grupos se puede realizar combinando las dos expresiones anteriores. La idea es considerar de forma integrada el grupo *X* y el conjunto de documentos relacionados con *x*. Así, en el denominador del cociente se calcula la cardinalidad de la unión de estos conjuntos. La expresión (1.8) muestra otra forma de cálculo de la pertenencia aproximada y la expresión (1.9) su uso para ponderar los grupos.

$$\varpi_X(x) = \frac{|X \cap R'(x)|}{|X \cup R'(x)|} \qquad \textbf{1.8}$$

$$w_i = \frac{\sum_{x \in X_i} \varpi_{X_i}(x)}{|X_i|} \qquad \textbf{1.9}$$

1.6 Consideraciones finales del capítulo

La proliferación de información disponible en los centros hospitalarios, evidencia la necesidad de desarrollar herramientas que permitan convertir esa información en conocimiento e inferir

áreas que deben ser interpretados por los expertos, a partir de la información disponible garantizando su uso productivo y contribuyendo al entrenamiento asistencial del personal en formación.

El modelo para el agrupamiento propuesto que incorpora la hibridación de técnicas de clasificación y agrupamiento permite lograr inferir hipótesis diagnósticas en etapas tempranas.

Rough Text constituye una propuesta que puede contribuir a lograr un mejor comportamiento del modelo en la inferencia realizada a partir la combinación de los elementos de las aproximaciones inferior y superior cuando se necesita una retroalimentación de la inferencia realizada.

2

RBC PARA LA INFERENCIA DE HIPÓTESIS DIAGNÓSTICAS BASADO EN LA TEORÍA DE LOS CONJUNTOS APROXIMADOS

2. RBC PARA LA INFERENCIA DE HIPÓTESIS DIAGNÓSTICAS BASADO EN LA TEORÍA DE LOS CONJUNTOS APROXIMADOS

En este capítulo, se presenta un Razonador Basado en Casos a partir de registros clínicos para la toma de decisiones; el cual integra los elementos de la teoría de los conjuntos aproximados para validar y refinar los agrupamientos que propone el modelo en la organización de la base de casos, así como la inferencia a partir de la búsqueda de prototipos que faciliten la recuperación, y la retroalimentación en la deducción de posibles hipótesis diagnósticas ante la presencia de un nuevo caso.

2.1 Construcción de la Base de Casos

Tomando como principio que, problemas nuevos se resuelven a partir de problemas similares a estos, se implementa el razonamiento basado en casos como método de solución de problemas. Para lo cual es necesario construir la base de casos y su organización, ya que dependiendo de esta última se facilita el acceso y la recuperación de los casos semejantes {Herrera, 2016 #182} Una visión gráfica del esquema se muestra en la Figura 2.1.

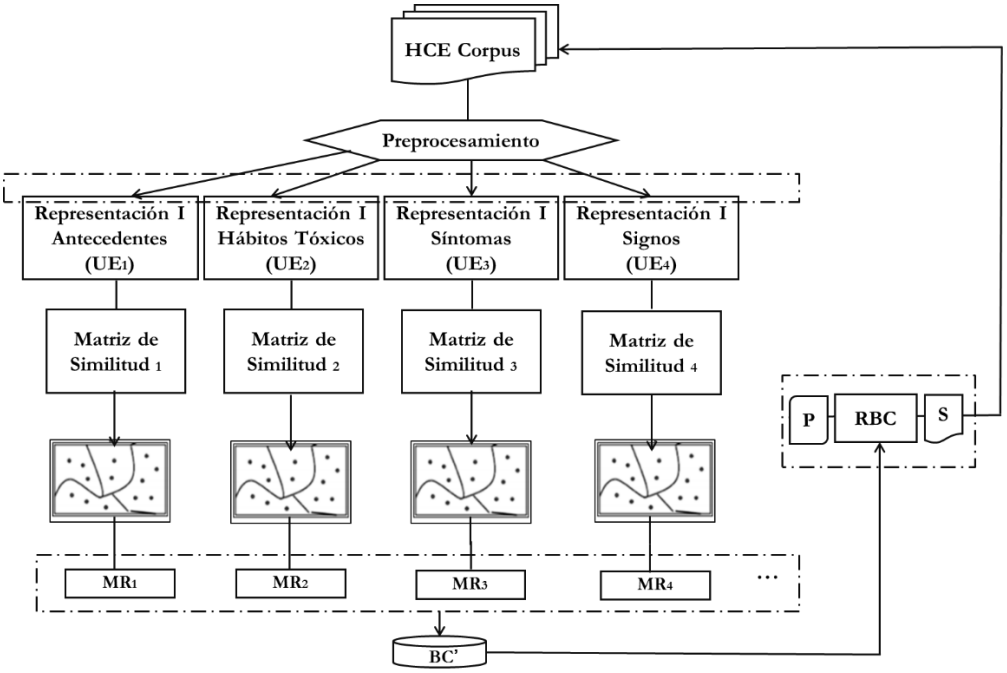

Figura 2.1 Esquema que muestra el modelo propuesto para el agrupamiento de HCE

2.1.1 Base de Casos

Del total de rasgos que componen la HCE ante la admisión de un nuevo paciente, se consideran sólo los cinco primeros como rasgos predictores, debido a que son los completados por el médico o el estudiante durante el interrogatorio, siguiendo el esquema del método clínico detallado al inicio. El modelo de razonador propuesto en (Fuentes et al., 2016) utiliza estos cinco rasgos para la predicción.

La Figura 2.1 muestra la estructura de un caso teniendo en cuenta los rasgos predictivos y objetivos que se consideran en la conformación de los casos en la BC.

HCE						
AP	HT	ST	SG	HEA	DF	DN
Rasgos Predictores					Rasgos Objetivos	

Figura 2.1: Estructura de un caso.

Un caso almacena los rasgos predictores siguientes:

- **Antecedentes (AP)**: Antecedentes patológicos personales y familiares, se refieren a antecedentes de diferentes patologías, modos de vida y características del paciente, dígase antecedentes de nacimiento, infancia y pubertad, quirúrgicos.
- **Hábitos Tóxicos (HT)**: Hábitos Tóxicos, detallando si es consumidor de alcohol, café, etc. la frecuencia y la cantidad. Si es fumador desde cuándo y la frecuencia con que lo realiza diariamente.
- **Síntomas (ST)**: Síntomas que refiere el paciente durante el interrogatorio que permite al experto realizar el resumen sindrómico.
- **Signos (SG)**: Signos clínicos a cualquier manifestación, detectados por el médico durante el examen físico por aparatos realizado al paciente.
- **Historia de la Enfermedad Actual (HEA):** Historia de la Enfermedad Actual, constituye la parte fundamental de la HC. Es en esta sección, en la cual se precisa la enfermedad que está cursando el paciente al momento de consultar. Recoge los síntomas y manifestaciones de enfermedad que él o la paciente ha presentado, cómo han evolucionado en el tiempo, y en la práctica, qué ha ocurrido.

Los rasgos objetivos:

- **Diagnóstico Diferencial (DF):** Diagnóstico Diferencial, conjunto de enfermedades que pueden ser las causantes de los síntomas y signos que sufre el paciente, una vez que se ha realizado la anamnesis y la exploración física, y antes o después de obtener pruebas diagnósticas complementarias. También se refiere a la argumentación del médico sobre la mayor o menor probabilidad de sufrir unas u otras enfermedades ante el cuadro clínico del paciente; con el fin de orientar las pruebas complementarias que deben realizarse hasta el diagnóstico de certeza. Sería la presunta hipótesis que debe ser posteriormente discriminada.
- **Diagnóstico (DN):** Diagnóstico positivo.

Del contenido que almacenan las HCE basadas en el estándar CDA-HL7 representadas en forma de documentos semi-estructurados, el enfoque basado en casos propuesto en(Fuentes et al., 2016) utiliza la información textual asociada a cada UE definida como rasgo predictor. Es importante señalar que, aunque la jerarquía de especificación del metalenguaje XML que implementa el estándar CDA-HL7 reconoce otras etiquetas de codificación de términos; en este trabajo se realiza el análisis sobre la información textual, ya que en la ingeniería del conocimiento realizada los expertos refieren que la descripción detallada de cada uno de los elementos que son completados a la llegada del paciente en la HCE, constituye un elemento esencial en la conducta que sigue el médico para emitir diagnósticos sospechosos.

Por lo que, utilizar sólo términos codificados podría limitar el razonamiento, debido a que situaciones en las que se tienen pacientes con iguales síntomas pero con características de origen y forma diferentes, la probabilidad de que sean recuperados por el sistema como casos similares sería mayor. En el caso específico de la UE Diagnóstico Diferencial en la jerarquía de especificación XML para enumerar los diagnósticos posibles, se utiliza la etiqueta *<item>*.

Una vez descritos los elementos que permitieron decidir qué información contienen los casos y como representarla, en el sub-epígrafe siguiente se describen los elementos a considerar para organizar los casos y facilitar el acceso.

2.1.2 Organización de la Base de Casos

Este modelo propone el uso de una estructura que se concibe a partir de conformación de grupos de documentos de HCE. Estos se obtienen a partir del agrupamiento de las UE correspondientes a las colecciones independientes que constituyen rasgos predictores, según se detalló

anteriormente, los cuales son validados y refinados utilizando los elementos de *RST*, como resultado se obtiene una partición refinada del conjunto de documentos. Posteriormente, se determina el elemento típico del grupo o prototipos quedando en un nivel superior. La Figura 2.2 muestra el esquema de los pasos que se siguen para la organización de los casos de la BC.

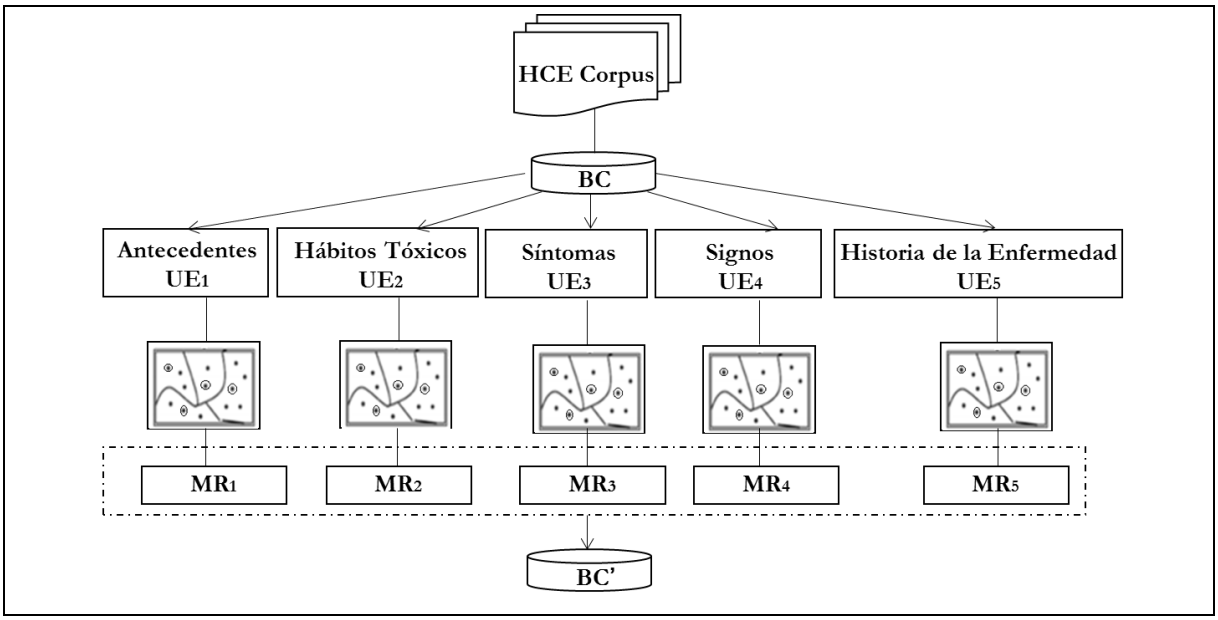

Figura 2.2: Esquema general para la construcción del modelo de la BC.

El modelo de la BC propuesto se resume en el procedimiento de la Figura 2.3.

```
Entrada: BC de HCE, UE que constituyen rasgos predictores
Salida: BC' (Particiones: Grupos homogéneos de HCE afines, HCE prototipos
cada grupo y calidad del agrupamiento.)
Inicio
      Para cada DUEk∈ Rasgos_Predictores
         1. Construir matriz de semejanza utilizando la Similitud_Coseno.
         2. Estimación del umbral de similitud β0.
         3. Agrupar y Refinar (Ver epígrafe 2.3.1.1 y 2.).
         4. Determinar Prototipo(s) de cada grupo (Ver epígrafe 2.3.1.3).
      Fin_Para
Fin
```

Figura 2.3: Procedimiento general para el esquema de construcción del modelo de la BC.

2.1.2.1 *Sistema de decisión necesario para aplicar RST*

Debido a que el modelo para el descubrimiento de conocimiento utilizado, parte de los documentos más representativos de los grupos, obtenidos como resultado de los agrupamientos asociados a cada unidad estructural, en esta propuesta se utiliza la teoría de los conjuntos

aproximados para refinar los resultados obtenidos por una partición inicial del conjuntos de documentos, ya la teoría de los conjuntos aproximados tiene gran utilidad para evaluar la calidad del agrupamiento a partir del cálculo de las aproximaciones inferiores, las cuales a su vez facilitan el proceso de selección de los prototipos o documentos más representativos de cada grupo, de lo cual en buena medida depende de la calidad del agrupamiento.

A continuación se describen cada uno de los elementos necesarios para aplicar RST en esta etapa.

La representación VSM del corpus y el resultado del agrupamiento como atributo de decisión, conforman el sistema de decisión necesario para aplicar RST.

Sea D el conjunto de documentos del corpus de HCE y sea $s{:}D\times D\rightarrow\mathbf{R}$ la función que mide la similitud entre los documentos de D, se considera la propuesta(Pawlak, 1982b) que define la relación R' según la expresión (2.1), donde $s(x, y)$ retorna el valor de similitud entre los documentos x e y, considerando el umbral de similitud ξ. Las variantes para su cálculo se describen más adelante.

$R'(x)=\{y\in U : yR'x$, es decir y está relacionado con x sí y solo si $s(x,y)>\xi\}$ **2.1**

Para determinar cuan similares son dos documentos en R' se utiliza el coeficiente Coseno(Cruañes Vilas et al., 2012).

En las expresiones (2.2) y (2.3) se muestran las formas de cálculo a partir de la relación R' para cada documento en U de las aproximaciones R'-inferior $R'_*(C_j)$ y R'-superior $R'^*(C_j)$, respectivamente.

$$R'_*(C_j)=\{D_i\in C_j : R'(D_i)\subseteq C_j\} \qquad \textbf{2. 2}$$

$$R'^*(C_j)=\bigcup_{D_i\in C_j} R'(D_i) \qquad \textbf{2.3}$$

Así los documentos de la aproximación inferior $R'_*(C_j)$ del grupo C_j, son aquellos que pueden ser clasificados con certeza como miembros del grupo C_j, ya que solamente se relacionan con documentos del mismo grupo y son los más representativos de ese grupo, mientras que los documentos en $R'^*(C_j)$ son la unión de todos los documentos que pertenecen al grupo C_j con los

documentos que ellos se relacionan según la relación *R*', de manera que los documentos que pertenecen a $R'^{*}(C_j)$ pueden ser clasificados como posibles miembros del grupo y son aquellos que posiblemente pertenezcan a él.

A continuación, se describe la medida de pertenencia aproximada (Pawlak and Skowron 1994) utilizada para el refinamiento del agrupamiento.

Pertenencia Aproximada:

En este trabajo se propone calcular la pertenencia de los documentos a los grupos combinando las expresiones A2.1 y A2.3 del Anexo 2 que permiten el cálculo del grado de inclusión de un documento a un grupo. La idea es considerar de forma integrada el grupo *X* y el conjunto de documentos relacionados con *x*. Así, en el denominador del cociente se calcula la cardinalidad de la unión de estos conjuntos. La expresión (2.4) muestra la forma de cálculo de la pertenencia aproximada y la expresión (2.5) su uso para ponderar los grupos (Magdaleno, 2008).

$$\varpi_X(x) = \frac{|X \cap R'(x)|}{|X \cup R'(x)|} \qquad \textbf{2.4}$$

$$w_i = \frac{\sum_{x \in X_i} \varpi_{X_i}(x)}{|X_i|} \qquad \textbf{2.5}$$

2.1.2.2 *Refinamiento del agrupamiento*

Varios algoritmos de agrupamiento tienden a crear muchos grupos unitarios o mal formados, sobre todo cuando el umbral de similitud entre los documentos no es adecuadamente seleccionado. En tales casos, es posible obtener aproximaciones inferiores vacías para esos grupos, lo cual significa que todos los documentos del grupo se relacionan con documentos de otros, indicando que el grupo no es consistente, por lo que no está adecuadamente conformado. Este elemento puede utilizarse como otro indicador de la calidad del grupo y puede ser utilizado para refinar los resultados del agrupamiento (Magdaleno, 2008).

En este trabajo se realiza un refinamiento del agrupamiento inicial obtenido utilizando RST a través de los resultados de la aproximación inferior. Para esto se analiza la colección de documentos *D* del grupo que se quiere eliminar y de este modo el documento d_i es asignado al grupo d_i que tenga mayor relación utilizando para esto la expresión (2.4).

A continuación, se muestra el procedimiento para realizar el refinamiento de los grupos a partir de los resultados obtenidos al aplicar RST a la colección de grupos. Las entradas a este algoritmo son: el resultado del agrupamiento inicial, así como los resultados de calcular las aproximaciones inferiores y superiores a cada grupo. La salida del algoritmo es el agrupamiento refinado.

```
Entrada: Aproximaciones inferiores (2.2) y superiores (2.3),
         Calidad del agrupamiento
Salida: Agrupamiento refinado
Mientras existan grupos con aproximación inferior vacía(A_* = Ø)
hacer
      Seleccionar el primer grupo i tal que A_*(Grupo_i) = Ø
      Para j=1 hasta Cantidad de Documentos del grupo_i
            Mientras k<=Cantidad total de grupos hacer
                    Calcular la pertenencia aproximada del documento j
                    (Doc_j) al grupo k (ϖ_grupok(Doc_j)), a través de la
                     expresión (2.10)
           Asignar Doc_j al grupo hacia al cual tuvo mayor pertenencia.
     Eliminar el grupo_i
```

Figura 2.5 Aplicación de RST en el refinamiento del agrupamiento.

Este algoritmo crea nuevos grupos, logrando refinar el agrupamiento obtenido en aras de una mayor calidad. Así, se muestra una potencialidad adicional de los Conjuntos Aproximados en otra etapa post-agrupamiento. Al finalizar este algoritmo se puede volver a aplicar el Algoritmo descrito en la Figura 2.4 para obtener las aproximaciones y resultados de validación que se ajusten al agrupamiento refinado deseado.

Breves consideraciones sobre el umbral de similitud

Estimar adecuadamente el umbral de similitud a utilizar es fundamental, así como la selección de la similitud a emplear. En esta tesis se utiliza la similitud Coseno porque ha sido una de las más efectivas para comparar vectores documentos. En (García 1999) se han propuesto expresiones que permiten estimar umbrales a partir de una matriz de similitudes o distancias y que no requieren información adicional del conjunto de datos que se procesen. Una de estas variantes es la media de las similitudes entre todos los pares de documentos posibles; según la expresión (2.6). Aunque es necesario tener en cuenta que la media es sensible a observaciones

extremas, y en su defecto, se pudiera utilizar la media recortada para eliminar los valores fuera de rango.

$$\bar{X} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s(D_i, D_j) \tag{2.6}$$

También se puede utilizar la media de los valores máximos de las similitudes entre cualquier par de objetos; expresión (2.7). Esta forma de cálculo puede provocar la obtención de un umbral muy alto, conduciendo a que exista coincidencia del grupo y sus aproximaciones. Esta situación puede arrojar valores de precisión y calidad cercanos a 1, cuando en realidad el resultado del agrupamiento no sea tan bueno (A*(Grupo) = Grupo = A*(Grupo) => Precisión = calidad = 1).

$$\overline{X}_{max} = \frac{1}{n} \sum_{i=1}^{n} max\{s(D_i, D_j)\} \tag{2.7}$$

Por el contrario, la media de los valores mínimos de las similitudes entre cualquier par de objetos permite obtener umbrales de similitud muy bajos; expresión (2.8). De esta forma, la aproximación inferior de un grupo será mucho menor que el grupo, y éste a su vez, mucho menor que su aproximación superior. Esto provoca que se obtengan valores muy bajos de precisión y calidad cuando en realidad el resultado del agrupamiento no sea tan malo (A*(Grupo) << Grupo << A*(Grupo) => Precisión = calidad = 0).

$$\overline{X}_{min} = \frac{1}{n} \sum_{i=1}^{n} \min_{\substack{j=1..n \\ i \neq j}} \{s(D_i, D_j)\} \tag{2.8}$$

La última variante de cálculo a utilizar es la media ponderada de la media de las similitudes y la media de los máximos; expresión (2.9).

$$\overline{X}_w = \alpha \overline{X} + (1 - \alpha) \overline{X}_{max} \tag{2.9}$$

La descripción de la notación utilizada es la siguiente: *n* es la cantidad de objetos de la colección, $s(D_i, D_j)$, es el valor de la similitud entre los vectores documento *Di* y *Dj*, y α es un valor entre 0 y 1 que permite ponderar la media y la media de los máximos en la expresión (2.9). En esta tesis el valor de α que se utilizó fue 0.5, según (Magdaleno, 2008) arroja los mejores resultados.

2.1.2.3 *Determinación de prototipos*

Después de haber realizado el agrupamiento y refinarlo, para cada grupo de casos obtenido, se determina el caso o los casos "representantes del grupo" o "prototipos", que se corresponden con los casos que pueden ser clasificados con certeza como miembros del grupo. La determinación de prototipos se basa en el cálculo de la aproximación inferior luego de obtener los grupos refinados. Para el cálculo de la aproximación inferior se utiliza la expresión (2.2).

$$R'_*(C_j) = \{D_i \in C_j : R'(D_i) \subseteq C_j\} \qquad \textbf{(2.10)}$$

Si existe más de un elemento en la aproximación inferior, los cuales serán prototipos del grupo, estos son ordenados decrecientemente atendiendo al valor de la pertenencia promedio de este con los elementos del grupo al cual pertenece. La siguiente expresión muestra cómo obtener la pertenencia promedio de un documento o_k en el grupo G, siendo m la cantidad de documentos del grupo.

$$S(o_k, G) = \frac{\sum_{\substack{i=1 \\ i \neq k}}^{m} SimCoseno(o_k, o_i)}{m-1} \qquad \textbf{(2.11)}$$

2.2 Motor de Inferencia

El **motor de inferencia** es la máquina de razonamiento del sistema, el cual compara el problema insertado con los que están almacenados en la BC, como resultado infiere una respuesta con el mayor grado de semejanza a la que se busca, adaptando los casos más similares recuperados.

2.2.1 **Módulo recuperador**

En la etapa de recuperación, se seleccionan el o los casos más similares al nuevo problema. Este proceso se realiza teniendo en cuenta los prototipos de los grupos de documentos obtenidos por el agrupamiento refinado de las colecciones de HCE. La recuperación se basa en una similitud local que determina la analogía entre valores de un mismo rasgo y una similitud global, que combina los resultados de las similitudes locales a todos los rasgos de los casos a comparar. La similitud global es el resultado de la suma ponderada de las similitudes entre las UE correspondientes de cada caso y el caso problema, según se observa en la ecuación 2.12.

$$\mathrm{SimGlobal}(\mathrm{X}, \mathrm{Y}) = \sum_{\mathrm{i}=0}^{\mathrm{m}} W_i * \mathrm{SimCoseno}_{\mathrm{i}}\left(\mathrm{X}_{\mathrm{UE}_i}, Y_{\mathrm{UE}_i}\right)/\mathrm{n} \qquad \textbf{(2.12)}$$

$$\text{donde: } \sum W_i = 1$$

Para el cálculo de los pesos de los rasgos predictores se consultó a los expertos, los cuales concluyeron que los rasgos Historia de la Enfermedad Actual, Signos y Síntomas son los más importantes, la suma del peso de ellos debe ser 0.75, y la suma total de todos los pesos tiene que ser 1. Al aplicar los experimentos se obtuvieron los mejores resultados para los siguientes pesos: HEA=0.25 , SG=0.25, ST=0.25, HT=0.13, AP=0.12 (Valdés, 2017)].

La Figura 2.6, muestra el procedimiento empleado en el proceso de recuperación de los *k* casos más similares al nuevo paciente, basado en el cálculo de la similitud *SimGlobal* (Ver Ecuación 2.12).

Entrada: P; problema a resolver

$\mathrm{BC}' = \{\mathrm{bc}'_1, \mathrm{bc}'_2, \ldots, \mathrm{bc}'_n\}$; sub-conjunto de casos obtenidos como prototipos.

Salida: MS'

Inicio

$MS' = \emptyset$

$MS = \{p_i | i = \overline{1,k}\}$, k casos más similares, $MS \subset \mathrm{BC}'$.

$\forall bc_i \in MS$

Si $SimGlobal(bc_i, P) > \gamma$

$MS' = MS' \cup bc_i$

Si no

$g = maxSimG\{g_{ik} | \forall C_{UEk}, bc'_i \in g_{ik}\}$

$LA = Upper_Aproximation(g)$

$LA' = \{d_i \in LA | max\{SimGlobal(d_i, P)\}\}$

$MS' = MS' \cup LA'$

Fin

Figura 2.6: Algoritmo para la recuperación de los *k* casos más similares.

2.2.2 Módulo adaptador

En este paso, se consideran los casos recuperados en la etapa anterior para elaborar una solución, que constituye un conjunto de diagnósticos posibles a considerar ante la admisión de un nuevo paciente. En adición a ello, de los diagnósticos sugeridos, el sistema propone un primer diagnóstico (*Hipótesis*) como el diagnóstico posible tomando en consideración los diagnósticos presentados para los casos más similares al nuevo problema *P*, y emite además un conjunto de diagnósticos diferenciales que deben descartarse. Para presentar el conjunto de diagnósticos

diferenciales posibles se discurren los diagnósticos de los casos más similares y sus diagnósticos diferenciales, realizando una adaptación que elimina los diagnósticos o diagnósticos diferenciales ya considerados los cuales se ordenan atendiendo al nivel de ocurrencia.

En la Figura 2.7 se describe como se realiza la adaptación al nuevo problema.

Entrada: $MS' = \{p_1, p_2, \dots, p_m\}$ subconjunto de casos obtenidos del módulo recuperador.

Salida: DF, DN; Rasgos objetivos.

Inicio

Para cada $p_{\mathrm{i}} \in MS'$

$\boldsymbol{ListDNUNF} = \{\mathrm{ld}_k = \{\mathrm{DN}_i \text{ U } \mathrm{DF}_i\}\}.\quad k=\overline{1,m}$

$\mathbf{R(i,j)} = \begin{cases} 1 & ,\mathrm{ld}_i \cap \mathrm{ld}_j \neq \emptyset \\ 0 & ,e.o.c. \end{cases}$

Fin Para

$D = MAJORITY\ (p_{i\ DNUDF}, \mathrm{R})$

Se obtiene MS'' donde $MS'' \subseteq MS'$, casos que más se relacionan en R

$\forall\, p_{\mathrm{i}} \in MS''$, construir $D_1 = \{d_i | d_i \in DN(pi)\}$.

$DN = \max\{OCCUR(d_i), \mathrm{d_i} \in D_1\}$

construir $D_2 = \{d_j | d_j \in DN(pi) \cup List(DF(pi)\,) \wedge d_j \notin D_1\}$.

$DF = List(\max_k\{OCCUR(d_i), \mathrm{d_i} \in D_2\})$

Fin

Figura 2.7: Algoritmo para conformar la solución inicial.

2.2.3 **Módulo Revisor**

La revisión se realiza cuando los casos recuperados fueron los *k* más similares iniciales, por tanto, no se explora la aproximación superior de los grupos de ninguno de los casos obtenidos como los más similares, es decir, la cantidad de casos son insuficientes para el completamiento del problema y no se pudieron discriminar atendiendo a la similitud con que estos fueron recuperados. Se realiza un proceso de selección basado en la comparación jerárquica por rasgos predictores y la importancia de estos en la recuperación con el propósito de obtener el caso más similar al nuevo problema

En la Figura 2.8, se describe como se obtiene el DN a partir de la revisión de la adaptación.

```
Entrada: MS' = {p1, p2, ..., pk} k casos iniciales.
Salida: conjunto diagnósticos
Inicio
    Si los casos recuperados fueron los k iniciales
       1. Ordenar los Rasgos_Predictores de mayor a menor
        atendiendo al peso
       2.  Para k=0 hasta Cantidad de UE ,UEk∈ Rasgos_Predictores
           Para j=0 hasta Cantidad de Documentos de MS'
                  Calcular s_UEk(pj, P)
           Fin Para
          Ordenar similitudes de los pj de mayor a menor
          Si la cantidad de pj con similitud máxima es 1
                     Retornar el DN del pj que su similitud es
                máxima
              Sino
                Si k==Cantidad de UE
                 Para j=0 hasta Cantidad de Documentos de MS'
                        Adicionar a listDN el DN de pj
                 Fin Para
                Sino
                 Para i=0 hasta cantidad de pi con similitud máxima
                        Adicionar pi a MS''
                        MS' = MS''
                 Fin Para
           Fin Para
```

Figura 2.8: Algoritmo para conformar el DN a partir de la revisión de la adaptación.

2.3 Procedimiento general del RBC que soporta CDARS

En la Figura 2.7 se muestra los módulos principales que conforman el procedimiento general del RBC implementado.

Módulo 1(Pre-procesamiento). Recuperación de la información identificando en cada documento recuperado las Unidades Estructurales (UE) que constituyen los rasgos predictores para el RBC.

Módulo 2(Representación). Representación del corpus textual obtenido para cada UE, obtener una representación VSM clásica.

Módulo 3(Similitud y Agrupamiento). Cálculo de la matriz de similitud utilizando la medida de similitud *Coseno* y un agrupamiento para cada UE

- **3.1** Valoración (validación de los grupos obtenidos, utilizando el cálculo de la pertenencia aproximada.
- **3.2** Refinamiento del agrupamiento a partir de la calidad de los grupos.
- **3.3** Determinación de prototipos a partir del cálculo de la aproximación inferior.

Módulo 4(Razonamiento). Realizar la inferencia de posibles hipótesis diagnósticas para un nuevo problema *P*.

- **4.1** Recuperación (determinar los prototipos más similares al nuevo problema)
- **4.2** Adaptación (a partir del cálculo de la aproximación superior se revisa y se construye una solución final)

Figura 2.9 Módulos principales del RBC propuesto.

2.3.1 Módulo 1: Recuperación y creación de índices a partir del corpus de HCE

La entrada al modelo lo constituye la colección de HCE que no son más que documentos XML basados en el estándar CDA-HL7. A partir de esta especificación se comienza el proceso de recuperación utilizando primeramente el API *Jdom* de Java destinada al trabajo con documentos XML, que permite identificar las UE que se incorporan al índice creado introduciéndose las facilidades de *Lucene* (Hatcher et al., 2009).

Se reutilizan las facilidades de *Lucene* para la representación del corpus: análisis léxico, eliminación de palabras vacías, segmentación, contenido en esta poderosa biblioteca de recuperación de información.

2.3.2 Módulo 2: Representación de las UE

En esta etapa se obtiene la representación para cada UE. Debido a la necesidad que el modelo realice un análisis léxico del texto y/o permita representar la información, se ha considerado la forma de representación VSM. Se seleccionó el VSM porque es un modelo que no es costoso computacionalmente como ocurre con los modelos probabilísticos, y se puede adaptar al contexto de los criterios que emiten los especialistas en la HCE de la relatoría del paciente en su mayoría textos cortos. Se obtiene la representación VSM para cada UE; a partir de la matriz

VSM clásica (Salton et al., 1975), que contiene en sus filas el índice de términos construido utilizando *Lucene* y los documentos de la colección en sus columnas, las celdas representan la frecuencia de aparición de cada término en la UE del documento que se procesa. A partir de cada la cual el sistema realiza el proceso de normalización, esta razón se calcula, utilizando la frecuencia absoluta de aparición del término y la longitud del documento (se asume como longitud del documento la longitud de la UE, ya que estas son tratadas como colecciones independientes).

2.3.3 Módulo 3: Similitud y Agrupamiento para cada UE

Para cada representación resultante se calcula una matriz de similitud utilizando como medida la similitud coseno, esta se muestra en la expresión 2. 3. Se genera un agrupamiento para cada UE a partir de la similitud asociada a la *Representación*.

$$S_{coseno(o_i,o_j)} = \frac{\sum_{k=1}^{m}(o_{ik} * o_{jk})}{\sqrt{\sum_{k=1}^{m} o_{ik}^2 * \sum_{k=1}^{m} o_{jk}^2}} \qquad 2.13$$

Como resultado en este módulo se realiza un agrupamiento inicial que permite obtener la partición p_0 que luego será refinada en la siguiente fase. El agrupamiento tiene por objetivo crear grupos coherentes internamente, es decir, que los documentos de un mismo grupo sean lo más similares posibles y a la vez que sean disimilares a los documentos de otros grupos (Manning et al. 2008). Como se abordó en el Capítulo 1, los algoritmos de agrupamiento jerárquicos construyen una partición del conjunto de objetos, específicamente HAC que es el algoritmo seleccionado para construir la partición inicial, ya que no requiere conocimiento previo del dominio como la cantidad inicial de grupos a formar siendo capaz de determinarla, por tanto, constituye un buen punto de partida como método interior a aplicar para obtener la conformación de grupos final.

Los algoritmos HAC construyen una partición en cada nivel de la jerarquía hasta obtener un solo grupo donde se incluyen todos los objetos; sin embargo, en la presente investigación se necesita obtener una partición inicial que luego será refinada con el propósito de obtener grupos de alta calidad. Luego al aplicar un algoritmo HAC es necesario cortar la jerarquía en un nivel para obtener dicha partición inicial.

El enfoque que se empleó en esta tesis fue el de realizar el corte según un nivel pre-especificado de similitud entre objetos en un mismo grupo. Este enfoque conlleva a aplicar un umbral que permita agrupar comparando las medidas de similitud de los grupos con dicho umbral; es decir, los grupos serán agrupados hasta que la mayor similitud de un grupo sea menor que el umbral especificado, teniendo en cuenta que la similitud de grupos se obtiene a partir del cálculo de la similitud entre sus elementos más disímiles, este umbral en nuestro caso debe ser bajo para evitar pasos a un nivel mayor en la jerarquía, elemento que se realiza al aplicar el refinamiento con RST, que tiene el propósito de obtener los grupos más compactos para la organización de la Base de Casos. Se seleccionó la medida coseno para hallar la matriz de similitud en el algoritmo de agrupamiento (Ver expresión 2.13).

Para realizar el agrupamiento se empleó la herramienta *S-Space Package* que es una biblioteca de código abierto y libre para desarrollar estos algoritmos de espacio palabra.

2.3.4 **Módulo 5: Razonamiento Basado en Casos**

2.3.4.1 *Recuperación*

Después de haber obtenido los prototipos que conforman la BC se determinan los casos más similares al nuevo problema *P*. Para determinar los casos más similares se parte de los prototipos almacenados en la BC y se computa la similitud de estos con el nuevo problema, seleccionando los *k* más similares al nuevo problema. Luego para cada caso más similar recuperado se comprueba si su similitud global con el nuevo caso es mayor que un umbral de similitud basado en la expresión (2.24).Si la similitud global supera el umbral se recupera el caso. De lo contrario se retroalimenta la búsqueda con el grupo más similar al que pertenece el caso. A partir de la aproximación superior del grupo más similar (Ver expresión (2.3)), se explora dicha aproximación superior y se recupera el documento o los documentos de la aproximación superior tal que la $SimGlobal$ es máxima, es decir todos aquellos que tengan el máximo valor de similitud global con el nuevo paciente.

2.3.4.2 *Adaptación*

En la elaboración de la solución inicial se consideran los casos más similares recuperados, tomando como punto de partida una combinación de las soluciones recuperadas según el

procedimiento de la Figura 2.6, a partir de este subconjunto de casos se obtiene una solución inicial en función de los rasgos objetivos definidos.

Este procedimiento comienza con la conformando la lista ***ListDNUDF*** donde se guarda a su vez una lista para cada caso recuperado los diagnósticos y diagnósticos diferenciales, después se procede a la búsqueda del valor del rasgo objetivo DN y DF, utilizando el operador MAJORITY (Baldizzoni, 2013), la ocurrencia tiene en cuenta la relación semántica R que determinan si dos casos tienen al menos un diagnóstico en común. Luego se seleccionan los casos que más se relacionan en R para transformar el conjunto de k casos en un subconjunto MS''; donde cada $pi \in MS'$, satisface que su rasgo objetivo DN y/o DF contiene diagnósticos relacionados según R. El rasgo objetivo DF es una lista de diagnósticos diferenciales ordenados por los diagnósticos de mayor ocurrencia entre los DN y DF de los casos en MS''.

2.3.4.3 *Revisión*

A partir de los resultados obtenidos en la fase de recuperación si los casos recuperados fueron los k más similares iniciales se ordenan los rasgos predictores de mayor a menor atendiendo al peso asignado a cada rasgo y se procede a calcular la similitud de cada caso p_i con el nuevo caso P para cada rasgo, construyendo un ordenamiento decreciente que permita discriminar los casos obtenidos, se termina cuando en alguna UE predictora exista un solo p_i cuya similitud sea máxima. Este proceso permite realizar una comparación jerárquica por rasgo predictor hasta encontrar un p_i con mayor similaridad al menos una UE. La cantidad de casos que pertenecen a MS' se reduce ya que en MS' solo permanecen los casos que hayan tenido similitud máxima, es decir todos aquellos que tengan el máximo valor de similitud en la UE tratada en cada iteración. Si al analizar el último rasgo predictor no se ha encontrado un único caso, se devuelve una lista con todos los diagnósticos de los casos de MS'con los cuales tiene máxima similitud. Por tanto, en la etapa de Revisión se selecciona el DN o los DN de los casos más similares al nuevo problema basado en la comparación jerárquica de las UE predictoras.

2.4 Consideraciones finales del capítulo

El modelo propuesto para la toma de decisiones diagnóstica en la admisión de un nuevo paciente integra el agrupamiento de HCE combinando estructura-contenido, el razonamiento basado en casos y la teoría de los conjuntos aproximados.

En la implementación del modelo se concibe una organización de la base de casos a partir de los agrupamientos de UE predictoras de las HCE; facilita la recuperación y el acceso a los casos más semejantes a un problema nuevo.

El conocimiento del dominio de aplicación utilizado permite inferir como solución diagnósticos posibles ordenados, como una guía de posibilidades que deben descartarse en ese orden.

El enfoque presentado obtiene hipótesis diagnósticas en tiempo real, ya que sólo utiliza la información obtenida por el experto a la llegada del paciente.

El modelo puede aplicarse a múltiples contextos, brinda la facilidad de inferir áreas que pueden ser fuente de investigaciones médicas. Puede emplearse para asistir el proceso de entrenamiento asistencial de estudiantes en los centros hospitalarios; así como para gestionar la información en repositorios de estudios de casos, discusiones diagnósticas, etc.

3

EVALUACIÓN DEL RBC Y DESCRIPCIÓN DEL SISTEMA CDARS

3. EVALUACIÓN DEL RBC Y DESCRIPCIÓN DEL SISTEMA CDARS

En el presente capítulo se realiza una evaluación de la efectividad del razonador basado en casos propuesto, para la determinación de posibles diagnósticos en la admisión de nuevos pacientes. Además se brinda una descripción del sistema desarrollado a nivel de usuario con el propósito de explicar cómo utilizar *CDARS* para el entrenamiento asistencial, y se describe cómo se implementan las diferentes etapas del modelo y las herramientas que se emplean en cada fase.

3.1 Evaluación del RBC

La evaluación del sistema se realiza mediante la verificación y validación. Se verifica que el sistema fue implementado y diseñado correctamente y que efectivamente es el producto que satisface los requerimientos de los usuarios. El proceso de verificación se realizó utilizando criterio de expertos. Para ello se dispuso de los casos de estudio suministrados y revisados por los expertos (Anexo 3)

Como caso de estudio se utilizan colecciones de HCE del Servicio de Admisión del Hospital Provincial "Celestino Hernández Robau" asociadas a diferentes enfermedades, en el área de la medicina oncológica, cardiovascular, enfermedades infecciosas, respiratorias, etc. Los datos utilizados fueron datos reales de pacientes ingresados en la institución, por lo cual el manejo de la información se realizó bajo la supervisión del personal, con el propósito de velar por la confidencialidad y la privacidad legal en el marco de la ética profesional, que establece el derecho del paciente a la reserva de toda la información relacionada con su proceso y con su estancia en la institución.

3.1.1 **Experimento 1**

El primer experimento se realiza para verificar las bondades de la RST para refinar el agrupamiento y obtener grupos más compactos. Para ello, se muestra que existen grupos resultantes del proceso de agrupamiento inicial con aproximación inferior vacía y que después de refinar el agrupamiento esto no sucede, lo cual favorece el resultado final. Este experimento se realiza sobre una colección pequeña de 20 HCE para ilustrar mejor los resultados.

Las aproximaciones inferiores y superiores de cada grupo resultante de un proceso de agrupamiento pueden sugerir como fusionar grupos, de forma tal que se refinen resultados de agrupamientos, donde, por ejemplo, exista la presencia de muchos grupos unitarios o con muy pocos documentos. En el experimento 2 se muestra una situación real donde se requiere la fusión e ilustra como las aproximaciones inferiores y superiores pueden contribuir al refinamiento de los resultados.

Dada una colección de 20. HCE a la UE se le aplicó el algoritmo de agrupamiento HAC y posteriormente en la etapa de post-agrupamiento se calcularon las aproximaciones inferiores y superiores al resultado del agrupamiento. En las tablas 3.1, 3.3, 3.5, 3.7, 3.9 se pueden observar los grupos resultantes del proceso de agrupamiento con sus aproximaciones inferiores y superiores y en las tablas 3.2, 3.4, 3.6, 3.8, 3.10 se observan los grupos resultantes del refinamiento del agrupamiento con sus aproximaciones inferiores y superiores de las unidades estructurales AP, HT.ST.SG y HEA respectivamente.

Grupo	**Documentos**	**Aproximación Inferior**	**Aproximación Superior**
1	[0]	[0]	[0]
2	[4]	[]	[4, 6]
3	[15]	[15]	[15]
4	[16]	[16]	[16]
5	[2, 5, 11, 13]	[2, 5, 11, 13]	[2, 5, 11, 13]
6	[6, 14],	[]	[3, 4, 6, 8, 9, 14, 17, 18, 19]
7	[9, 12]	[12]	[9, 12, 14]
8	[1, 17, 18, 3, 19, 7, 8, 10]	[1, 7, 10]	[1, 3, 6, 7, 8, 10, 14, 17, 18, 19]

Tabla 3.1. Grupos de documentos y sus aproximaciones inferiores y superiores de la UE Antecedentes Patológicos.

Grupo	**Documentos**	**Aproximación Inferior**	**Aproximación Superior**
1	[0],	[0],	[[0],
2	[4, 15],	[15],	[4, 6, 15],
3	[16],	[16],	[16],
4	[2, 5, 11, 13],	[2, 5, 11, 13],	[2, 5, 11, 13]
5	[6, 9, 12, 14],	[9, 12],	[3, 4, 6, 8, 9, 12, 14, 17,18,19]
6	[1, 17, 18, 3, 19, 7, 8, 10]	[1, 7, 10]	[1, 3, 6, 7, 8, 10, 14, 17,18,19]

Tabla 3.2. Grupos refinados y sus aproximaciones inferiores y superiores de la UE Antecedentes Patológicos.

Grupo	Documentos	Aproximación Inferior	Aproximación Superior
1	[17, 4, 5, 6, 8-11, 13, 15]	[4, 5, 6, 8-13, 15, 17]	[4, 5, 6, 8-11, 13, 15, 17]
2	[0, 18, 19]	[0, 18, 19]	[0, 18, 19]
3	[16, 1, 2, 3, 12, 14]	[1, 2, 3, 12, 14, 16]	[1, 2, 3, 12, 14, 16]
4	[7]	[7]	[7]

Tabla 3.3. Grupos de documentos y sus aproximaciones inferiores y superiores de la UE Hábitos Tóxicos

Grupo	Documentos	Aproximación Inferior	Aproximación Superior
1	[4]	[4]	[4]
2	[7]	[7]	[7]
3	[10]	[10]	[10]
4	[13]	[13]	[13]
5	[15]	[15]	[15]
6	[16]	[]	[16, 17]
7	[1, 2, 14]	[2]	[0, 1, 2, 3, 11, 14, 18, 19]
8	[0, 18, 3, 19, 5, 9, 11]	[5, 9]	[0, 1, 3, 5, 9, 11, 14, 18, 19]
9	[17, 6, 8, 12]	[6, 8, 12]	[6, 8, 12, 16, 17]

Tabla 3.4. Grupos de documentos y sus aproximaciones inferiores y superiores de la UE Síntomas.

Grupo	Documentos	Aproximación Inferior	Aproximación Superior
1	[4]	[4]	[4]
2	[7]	[7]	[7]
3	[10]	[10]	[10]
4	[13]	[13]	[13]
5	[15]	[15]	[15]
6	[1, 2, 14]	[2]	[0, 1, 2, 3, 11, 14, 18, 19]
7	[0, 18, 3, 19, 5, 9, 11]	[5, 9]	[0, 1, 3, 5, 9, 11, 14, 18, 19]
8	[16, 17, 6, 8, 12]	[6, 8, 12, 16, 17]	[6, 8, 12, 16, 17]

Tabla 3.5. Grupos refinados y sus aproximaciones inferiores y superiores de la UE Síntomas.

Grupo	Documentos	Aproximación Inferior	Aproximación Superior
1	[1]	[1]	[1]
2	[2]	[2]	[2]
3	[4]	[4]	[4]
4	[5]	[5]	[5]
5	[9]	[9]	[9]
6	[10]	[10]	[10]
7	[13]	[]	[0, 3, 7, 13, 18, 19]
8	[14]	[14]	[14]
9	[15]	[15]	[15]
10	[16, 17, 12]	[]	[6, 12, 16, 17]
11	[6, 8]	[8]	[6, 8, 12, 16, 17]
12	[0, 18, 3, 19, 7, 11]	[11]	[0, 3, 7, 11, 13, 18, 19]

Tabla 3.6. Grupos de documentos y sus aproximaciones inferiores y superiores de la UE Signos.

Grupo	Documentos	Aproximación Inferior	Aproximación Superior
1	[1]	[1]	[1]
2	[2]	[2]	[2]
3	[4]	[4]	[4]
4	[5]	[5]	[5]
5	[9]	[9]	[9]
6	[10]	[10]	[10]
7	[14]	[14]	[14]
8	[15]	[15]	[15]
9	[16, 17, 6, 8, 12]	[6, 8, 12, 16, 17]	[6, 8, 12, 16, 17]
10	[0, 18, 3, 19, 7, 11, 13]	[0, 3, 7, 11, 13, 18, 19]	[0, 3, 7, 11, 13, 18, 19]

Tabla 3.7. Grupos refinados y sus aproximaciones inferiores y superiores de la UE Signos.

Grupo	Documentos	Aproximación Inferior	Aproximación Superior
1	[7]	[7]	[7]
2	[11]	[11]	[11]
3	[16]	[16]	[16]
4	[9, 13]	[]	[2, 5, 9, 13, 15, 19]
5	[5, 15]	[]	[0, 5, 9, 15, 19]
6	[1, 2]	[]	[1, 2, 3, 10, 12, 13, 17, 18]
7	[18, 3, 6]	[6]	[0, 1, 2, 3, 6, 12, 18, 19]
8	[4, 8]	[]	[4, 8, 12, 17]
9	[0, 19, 14]	[14]	[0, 3, 5, 13, 14, 18, 19]
10	[17, 10, 12]	[]	[1, 2, 3, 4, 8, 10, 12, 17, 18]

Tabla 3.8. Grupos de documentos y sus aproximaciones inferiores y superiores de la UE Historia de la Enfermedad Actual.

Grupo	Documentos	Aproximación Inferior	Aproximación Superior
1	[17, 4, 7, 8, 10, 12],	[4, 7, 8]	[1, 2, 3, 4, 7, 8, 10, 12, 17, 18]
2	[11]	[11]	[11]
3	[16, 5, 9, 13, 15]	[9, 15, 16]	[0, 2, 5, 9, 13, 15, 16, 19]
4	[18, 3, 6]	[6]	[0, 1, 2, 3, 6, 12, 18, 19]
5	[0, 1, 2, 19, 14]	[14]	[0-3, 5, 10, 12, 13, 14, 17-19]

Tabla 3.9. Grupos refinados y sus aproximaciones inferiores y superiores de la UE Historia de la Enfermedad Actual.

En las Tablas 3.1, 3.4, 3.6, 3.8 se observan los grupos resultantes del agrupamiento de las UE AP, ST, SG y HEA respectivamente, se puede observar que existen grupos con aproximación inferior vacía y al analizar la aproximación superior de estos, todos los documentos de los mismos están relacionados con los documentos de otros grupos. En las Tablas 3.2, 3.5, 3.7, 3.9 se observan los grupos refinados para cada UE respectivamente y se comprueba que después de refinarlos no existen grupos con aproximación inferior vacía. En el caso de la UE Hábitos Tóxicos se puede observar en la Tabla 3.3 que al calcular la aproximación inferior para cada grupo siempre el conjunto es no vacío y por tanto no fue necesario refinar el agrupamiento

3.1.2 **Experimento 2**

Debido a que una de las etapas del modelo para recuperar los casos más similares al nuevo problema presentado necesita de la determinación de un umbral que permita estimar cuan similares son los objetos al nuevo problema, en este trabajo se propone utilizar las variantes. Variante 1 (Ecuación 2.8), Variante 2 (2.9) y Variante 3 (2.10) descritas en el Capítulo 2. Con el propósito de verificar el comportamiento del modelo de RBC propuesto a partir de la validez de las respuestas obtenidas, dependiendo de la selección de una u otra variante, se realizó el cuarto experimento que consistió en emplear el conjunto de casos correspondientes a HCE del caso de estudio y realizar una validación cruzada (LOOCE)(Cawley, 2006), para cada iteración se considera un caso de prueba y el resto de entrenamiento. Los resultados se observan en la tabla

Como se observa en la Tabla 3.11 los porcientos de validación evidencian el buen comportamiento de todas las variantes para el cálculo del umbral, no obstante las variantes 1 y 3 obtiene resultados superiores con respecto a la variante 2. En este trabajo se propone utilizar la variante 3 ya que en (Magdaleno, 2008) se aborda que generalmente esta obtiene los mejores resultados.

	Variante 1	Variante 2	Variante 3
% RBC	100.0	98.5663	100.0

Tabla 3.11. Porcientos de validación para cada variante de cálculo de umbral.

3.1.3 **Experimento 3**

Finalmente para el proceso de validación del RBC se diseñó el quinto experimento. Este consistió en emplear el conjunto de casos correspondientes a HCE del caso de estudio y realizar una validación cruzada como se describió anteriormente. Los resultados se observan en la Tabla 3.12.

En la literatura se sugiere que la recuperación de los *k* casos más similares, se debe realizar sobre la base de considerar valores de *k=1*, *k=3*, *k=5* ó *k=7*. En este trabajo se propone utilizar en la recuperación el valor *k=3*, debido a que además de reportar los mejores resultados, los expertos refieren que al completar la HCE del paciente y realizar una interpretación adecuada de sus manifestaciones debe sugerirse un diagnóstico como hipótesis sin obviar otros pocos que pudieran descartarse, para no limitar la verificación complementaria. Al validar los resultados del RBC el valor *k=3* para la recuperación es el que ha reportado los mejores resultados según se observa en la Tabla 3.10.

	k=1	k=3	k=5	k=7
% RBC	100.0	100.0	91.8919	90.0901

Tabla 3.12. Validación del RBC

3.2 Implementación del modelo

3.2.1 *Recuperación de la información*

Para la implementación del modelo extendido se utilizaron varias herramientas que facilitan el trabajo con la colección de entrada.

3.2.1.1 *Extracción de la información*

Para leer el contenido de las HCE almacenadas en formato XML se utilizó la API JDOM que proporciona una variedad de funciones de utilidad para examinar el contenido y la estructura de un documento. Las clases que se utilizaron de JDOM son las siguientes: *Document* que representa todo el documento XML, *Element* que representa un elemento XML, un objeto *Element* tiene métodos para manipular sus elementos secundarios, sus atributos de texto, y espacios de nombres, *Attribute* representa un atributo de un elemento y tiene métodos para obtener y establecer el valor del atributo, *Text* que representa el texto de la etiqueta XML y para parsear el XML a partir de un fichero se usa el *SAXBuilder*.

3.2.1.2 *Indexación y búsqueda*

Para la indexación y búsqueda de los documentos del corpus se utilizó la herramienta *Lucene*. Las principales clases utilizadas en el proceso de indexación son : *IndexWriter* es el componente principal del proceso de indexación, esta clase crea y añade documentos a un índice, es decir, da permisos de escritura sobre un índice pero no permite realizar búsquedas ni leer el índice, la estructura de datos que conforma el índice es también llamada índice inverso, este tipo de estructura permite hacer eficiente el uso de espacio en disco y permite realizar búsquedas rápidas por palabras clave, ya que no trata a los documentos como el elemento central de la búsqueda, sino que busca por palabras claves, *Directory* la cual indica el directorio de la ubicación de un índice en *Lucene*, los datos antes de ser indexados deben pasar a través de un objeto de tipo *Analyzer*. El analizador, el cual es especificado en el constructor de *IndexWriter*, es el encargado de extraer los *tokens* del texto que debe ser indexado y de eliminar el resto transformando la entrada en objetos de tipo *Term*. El proceso de análisis se realiza cuando añadimos un documento al índice con la función *addDocument (Document)* de *IndexWriter*, de manera que se analizan los datos para hacerlos más manejables para indexar, *Document* es una clase que

almacena campos (*Field*) con la meta información que queremos indexar y/o guardar en el índice (emails, rutas, autor, fecha, etc.) y *Field* que son los fragmentos de información de cada documento que deseamos almacenar en el *Document*.

La principal ventaja de *Lucene* es que ofrece a los usuarios capacidad para búsquedas totalmente textuales, las principales clases que se usaron para la búsqueda del núcleo de *Lucene* son: *IndexSearcher* que es la clase que accede al índice y tiene permiso para leerlo, pero no puede modificarlo ,esta clase tiene un método *search* (*Query*) que es el encargado de realizar la búsqueda en el índice, *Query* es una clase abstracta que permite definir la consulta sobre la que queremos realizar la búsqueda y *QueryParser* es la clase que procesa una expresión dada por un usuario y la convierte en un objeto de la clase *Query* , indica por cual campo realizaremos la búsqueda.

3.2.2 *Pre-procesamiento*

En la etapa de Pre-procesamiento se reutilizaron las facilidades de *Lucene* para pre-procesar cada UE, debido a que los datos antes de ser representados deben pasar a través de un objeto de tipo *Analyzer* para que se les haga un análisis léxico, se eliminen las palabras vacías y se extraigan las raíces o lemas.

3.2.3 *Representación*

Para representar cada UE se implementó la clase VSM, la cual empaqueta los siguientes métodos: *getTerms()* que crea una lista con todos los términos de la colección sin incluir repetidos , *createVSM()* que es donde se llena la matriz VSM con los pesos de los términos en cada documento, en este caso se utilizó la normalización de la frecuencia**,** *obtenerMatrixVSM()* que es el encargado de crear la matriz VSM que contiene en sus filas el índice de cada documento y en sus columnas cada término de la colección.

3.2.4 *Agrupamiento*

Para obtener un agrupamiento por cada UE se utilizó la *HierarchicalAgglomerativeClustering* de la biblioteca *S-Space*, la cual implementa la interfaz *Clustering* que ofrece varia propiedades para controlar el agrupamiento.

3.2.5 *Recuperación*

Para recuperar los documentos más similares al nuevo problema se implementó la clase *Retrive*, la cual contiene los siguientes métodos: *createCase()*, encargado de instanciar los casos que van a conformar la BC a partir de los prototipos de cada grupo, *SimGlobal()*calcula la similitud entre un caso *k* y el nuevo problema , *retrive()* recupera los casos más similares al nuevo problema.

3.2.6 *Adaptación*

Para realizar la adaptación de la respuesta se implementó la clase *Adapter* en la cual se implementan los métodos *buildDN()* y *buildDF()*, los cuales devuelven los DN y los DF para el nuevo problema.

3.2.7 *Interfaz de usuario*

Para el diseño de la interfaz de usuario del sistema se usó *Yii*, framework *PHP* basado en programación orientada a objetos que implementa el patrón de diseño MVC (Modelo-Vista-Controlador) el cuál es ideal para la programación Web. Para la implementación de la aplicación web se empleó además el framework *Bootstrap*, diseñado para facilitar la creación de sitios web que permite crear de forma sencilla webs de diseño adaptable, que sean agradables y atractivas al usuario.

3.3 ***Diseño del Sistema CDARS***

Para el diseño del sistema CDARS fue necesario la creación de un modelo que integra cuatro módulos fundamentales: Analizador, Representación, Agrupamiento y RBC , una Interfaz de Usuario y un Controlador que comunica la Interfaz de Usuario con el modelo .En la Figura 3.1 se muestra el diseño general del sistema CDARS.

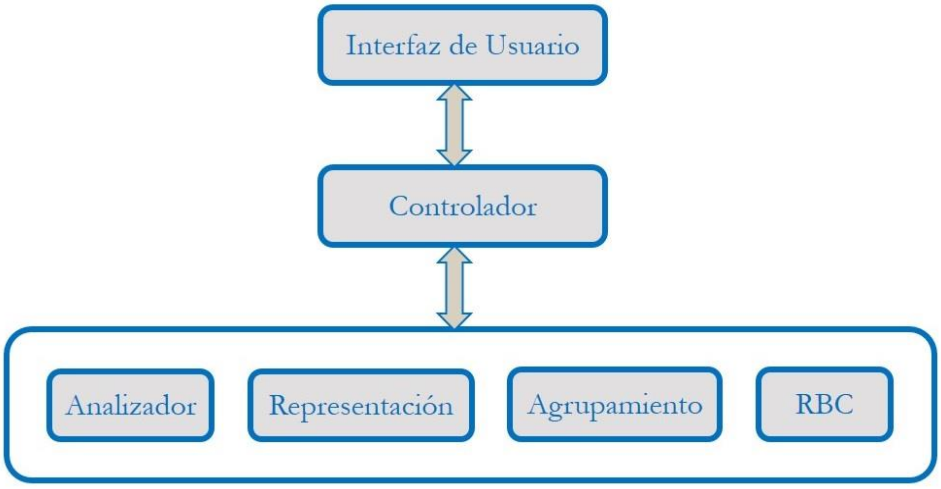

Figura 3.1 Diseño general del sistema CDARS

En el módulo Analizador se encuentran las clases que permiten la recuperación de la información identificando en cada documento recuperado las UE y el pre-procesamiento de la misma. En el módulo de Representación se encuentran las clases que permiten la representación del corpus textual obtenido para cada UE. El módulo de Agrupamiento está formado por las clases que permiten realizar el agrupamiento jerárquico y refinar dicho agrupamiento utilizando RST, el cálculo de los umbrales de similitud y de los prototipos de grupos. El módulo de RBC lo forma la clase que permite la recuperación de los casos más similares a un nuevo problema y la clase que permite realizar la adaptación de la respuesta, así como la revisión de la misma en caso que sea necesario.

La interfaz de usuario es una aplicación web desarrollada en PHP. Se empleó *Yii*, framework genérico de programación Web que puede ser utilizado para todo tipo de aplicaciones Web. Este framework implementa el diseño de patrón MVC el cual tiene por objetivo separar la lógica del negocio de las consideraciones de la interfaz de usuario para que los desarrolladores puedan modificar cada parte más fácilmente sin afectar a la otra. En la figura 3.2 se muestra la estructura estática de una aplicación Yii.

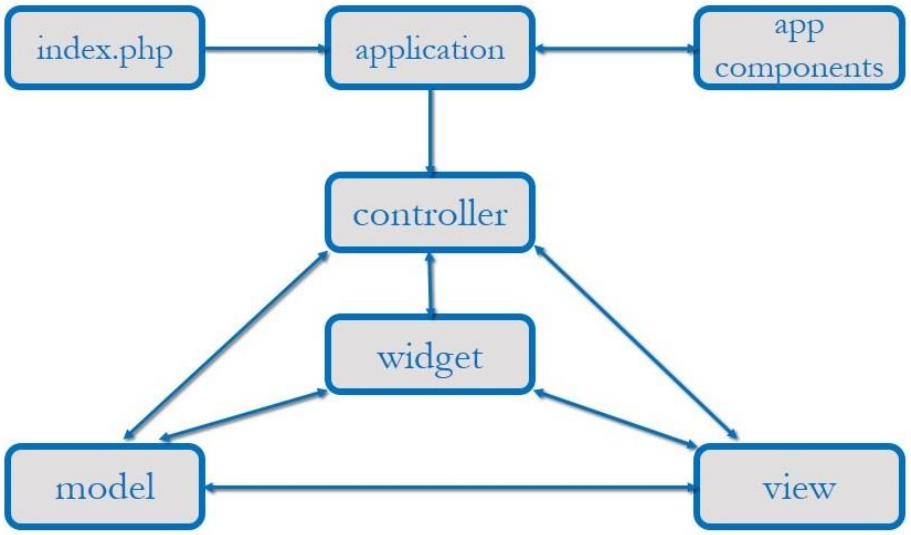

Figura 3.2 Estructura estática de una aplicación Yii.

En el desarrollo de la Interfaz de Usuario se emplearon recursos de la programación web como el trabajo con imágenes, enlaces, formularios, menús, etc., los cuales facilitan el funcionamiento del sitio web y la interacción con los usuarios.

El Controlador es el que empaqueta todas las clases controladoras y es el encargado de establecer la comunicación entre el modelo y la Interfaz de Usuario. *Yii* permite ejecutar aplicaciones implementadas en java, por lo que se usó ésta facilidad del *framework* para incorporar el modelo propuesto en este trabajo.

El diagrama de la Figura 3.3 muestra un flujo de tareas de una aplicación *Yii* cuando resuelve un pedido de usuario:

1. Un usuario realiza un pedido y el servidor Web se encarga de la solicitud mediante la ejecución del script de arranque en *index.php*.
2. El script de entrada crea una instancia de aplicación y la ejecuta.
3. La aplicación obtiene la información detallada del pedido del usuario del componente de aplicación *request*.
4. El controlador determina la acción (pedido) con ayuda del componente de aplicación llamado *urlManager*. Para este ejemplo el controlador es *post* que refiere a la clase *PostController* y la acción es *show* que su significado es determinado por el controlador.

5. La aplicación crea una instancia del controlador pedido para resolver el pedido del usuario. El controlador determina que la acción *show* refiere al nombre de método *actionShow* en la clase controlador. Entonces crea y ejecuta los filtros asociados con esta acción (ejemplo: control de acceso). La acción es ejecutada si los filtros lo permiten.
6. La acción lee el modelo *Post* cuyo ID es 1 de la base de datos o repositorio.
7. La acción realiza la vista llamada *show* con el modelo *Post*
8. La vista lee y muestra los atributos del modelo *Post.*
9. La vista ejecuta algunos *widgets*.
10. El resultado realizado es embebido en un esquema (*layout*).
11. La acción completa la vista realizada y se la muestra al usuario.

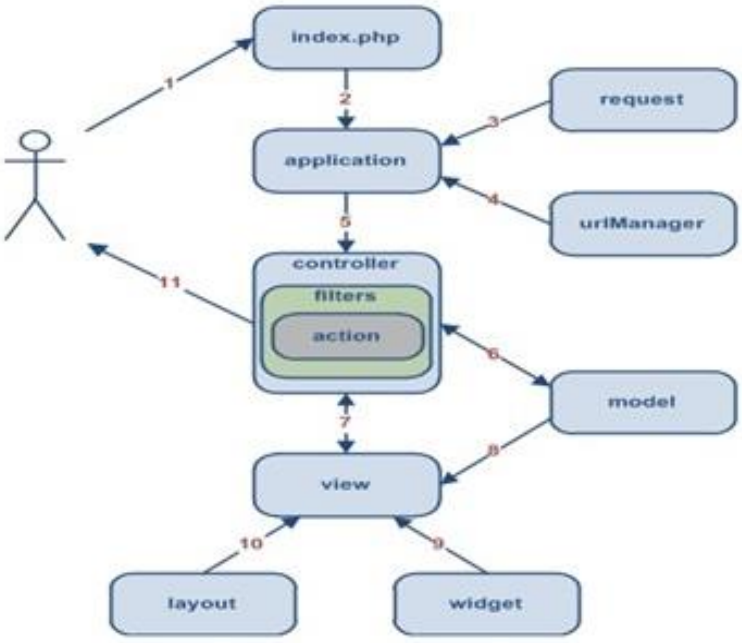

Figura 3.3. Típico flujo de tareas de una aplicación *Yii*.

3.4 Interfaz de usuarios del sistema CDARS para la recomendación de hipótesis diagnósticas

La interfaz de usuario cuanta con una página de inicio, la cual se muestra en la Figura 3.4, que permite a los usuarios acceder a sitios de interés sobre la medicina cubana e internacional.

Figura 3.4. Página de inicio de la interfaz web.

La aplicación permite acceder a la información referente a las HCE que se encuentran ubicadas en un repositorio. La Figura 3.5 muestra la página que permite acceder a las HCE.

	Paciente	Médico	Fecha	Hora	Diagnóstico	Evolución
1	R PL	D U	2016/03/12	10:12:51	Adenocarcinoma de colón	Satisfactoria a cirugía
2	N N	D U	2016/03/12	10:12:51	carcinoma escamoso bien diferenciado	Reservado
3	A MT	D U	2016/03/12	10:12:51	angina inestable	satisfactorio
4	N N	R LM	2016/03/12	10:12:51	bronconeumonía bacteriana intrahospitalaria	satisfactoria
5	JD RN	R LM	2016/03/12	10:12:51	EPOC más sepsis respiratoria	reservado
6	R BA	R LM	2016/03/12	10:12:51	linfoma no Hodking	favorable
7	N N	R C	2016/03/12	10:12:51	dengue sin signos de alarma	satisfactorio
8	N N	D U	2016/03/12	10:12:51	infarto agudo del miocardio	reservado
9	N N	R LM	2016/03/12	10:12:51	neumopatía inflamatoria por neumococo	Satisfactoria
10	N N	R C	2016/03/12	10:12:51	dengue con signos de alarma	satisfactorio

Figura 3.5. Página para acceder a de la información de las HCE.

La aplicación cuenta además con un menú Recomendaciones que le da la posibilidad al estudiante de Entrenar para adquirir conocimientos y mejorar su rendimiento intelectual y Consultar la información almacenada. La Figura 3.6 muestra la página del menú Recomendaciones que le permite al estudiante Entrenar y Consultar.

Figura 3.6.Página del menú Recomendaciones.

¿Cómo puede entrenar un estudiante de medicina utilizando CDARS?

CDARS facilita el proceso de entrenamiento que permite obtener hipótesis diagnósticas a partir de los rasgos predictores. En este contexto se establece una asociación entre campo y unidad estructural de un documento XML. La Figura 3.7 muestra la página donde se puede realizar dicho entrenamiento.

Figura 3.7.Página para realizar el entrenamiento.

¿Cómo realizar búsquedas a partir una colección de documentos?

CDARS facilita el proceso de búsqueda y consulta de la colección, permitiendo organizar los resultados por la relevancia y el criterio del agrupamiento. El sistema admite realizar búsquedas por múltiples campos, al estilo *Lucene*. En este contexto se establece una asociación entre campo y unidad estructural de un documento XML. La Figura 3.8 muestra la página que permite editar las UE que se desean incluir en el procesamiento de una consulta.

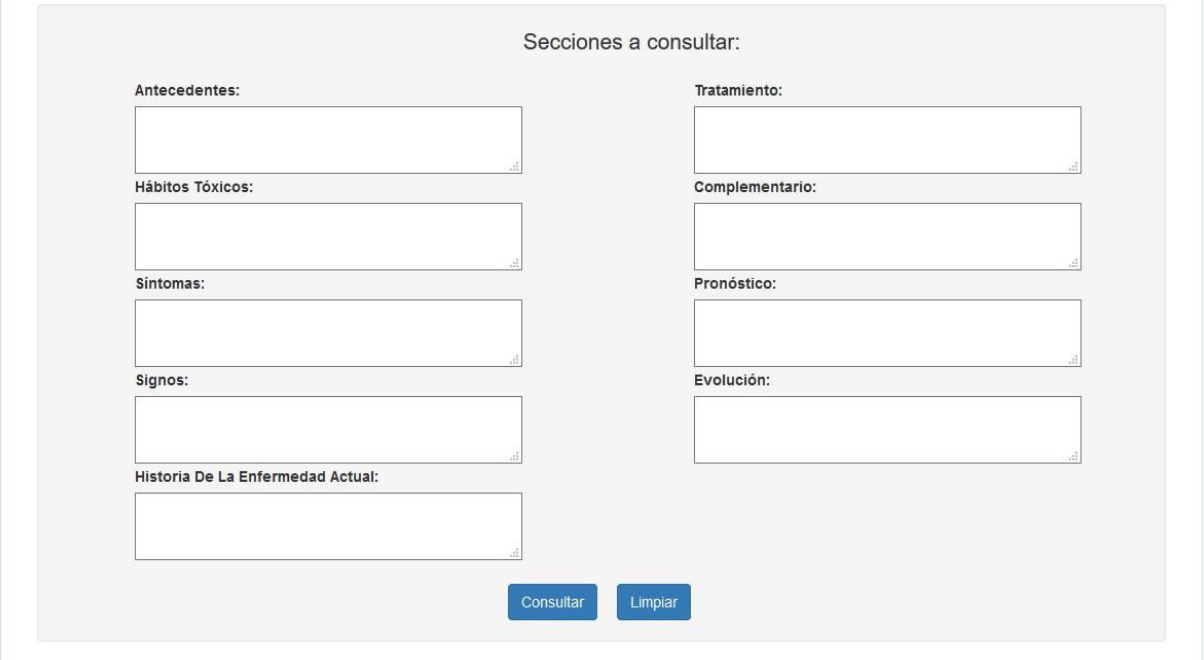

Figura 3.8.Página que permite consultar la información.

Para iniciar por primera vez el sistema, el usuario administrador debe acceder a la opción Actualizar Base de Casos lo cual le permite conformar la BC o actualizarla en una etapa determinada. Esta configuración se mantendrá mientras no se indique nuevamente la ejecución de la acción. La Figura 3.9 muestra la página que permite actualizar la BC.

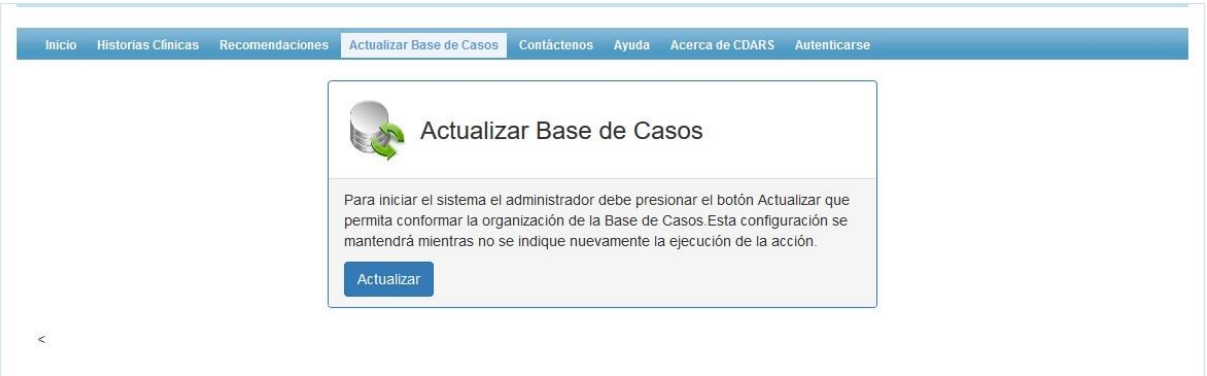

Figura 3.9.Página que permite actualizar la BC.

Además, la aplicación cuenta con una vista donde se puede obtener información acerca del sistema, una vista que permite contactar con el sitio, una vista para autenticarse como uno de los usuarios de prueba creados y una ayuda.

3.5 Conclusiones finales del capítulo

Los casos de estudio definidos permitieron demostrar la factibilidad del modelo propuesto basado en RST para el agrupamiento de documentos XML.

Los resultados de los experimentos realizados muestran la precisión del RBC propuesto y su valor en la inferencia de posibles hipótesis diagnósticas.

La interfaz de usuario del sistema CDARS que incluye la implementación del procedimiento general propuesto es amigable y facilita el acceso a la información referente a las HCE, el entrenamiento asistencial, el proceso de búsqueda y consulta de la colección permitiendo organizar los resultados por la relevancia, así como la conformación y actualización de la BC.

CONCLUSIONES

Como resultado del presente trabajo se desarrolló un sistema basado en casos para la inferencia de hipótesis diagnósticas, basado en la teoría de los conjuntos aproximados utilizando el estándar CDA-HL7.

- EL refinamiento del agrupamiento basado en el cálculo de la aproximación inferior permite obtener particiones más compactas a partir del agrupamiento utilizando HAC.
- La BC se organiza jerárquicamente partiendo de la selección de los prototipos en un nivel superior como resultado de los grupos refinados reduciendo el espacio de búsqueda explorado para obtener los casos más semejantes a un problema nuevo.
- El módulo de Revisión del RBC garantiza el completamiento del nuevo problema cuando los casos recuperados son insuficientes garantizando el entrenamiento asistencial a los estudiantes.
- El sistema CDARS fue validado usando criterios de expertos a partir de casos de estudios propuestos por especialistas del Hospital Provincial “Celestino Hernández Robau”; el cual contribuye a la toma de decisiones que realiza el especialista, constituye un entrenador para estudiantes de Medicina en la Propedéutica Clínica y cuenta con una interfaz amigable. A partir de la validación se obtienen los mejores resultados para k=3 y considerando como variante de cálculo de umbral la media de los máximos y los mínimos.

RECOMENDACIONES

- Analizar otras formas de cálculo de umbral para construir las aproximaciones inferiores y superiores.
- Determinar otras variantes de obtención de prototipos.
- Estudiar otras formas de refinar agrupamiento para determinar su efecto a la hora de conformar las particiones más adecuadas.

REFERENCIAS BIBLIOGRÁFICAS

AGGARWAL, C. C. & REDDY, C. K. 2013. *Data clustering: algorithms and applications*, CRC Press.

ALONSO, O. E. 2014. Implementación del modelo RIM de HL7 v3 en orientación a objetos y su uso en procesos de interoperabilidad semántica.

ANDERBERG, M. R. 1973. *Clustering Analysis for Applications*, New York: Academic.

ARCO, L., BELLO, R. & ARTILES, M. Un nuevo enfoque del uso de los conjuntos aproximados en la solución de problemas de la minería de textos. VII Conferencia Científica Internacional de la Universidad de Ciego de Ávila (UNICA2006), 2006a.

ARCO, L., BELLO, R. & GARCIA, M. M. On clustering validity measures and the Rough Set Theory. Artificial Intelligence, 2006. MICAI'06. Fifth Mexican International Conference on, 2006b. IEEE, 168-177.

ARUQUIPA, C. M. G. 2014. Modelo para la Recuperación de Datos de Expedientes Clínicos mediante HL7. *Revista del Postgrado en Informática***,** 148.

BAREISS JR, E. R. 1988. Protos: a unified approach to concept representation, classification, and learning.

BARRETO , P. J. 2015. La historia clínica: documento científico del médico.

BATCHELOR, B. 1978. *Pattern Recognition: Ideas in Practice,* New York, Plenum Press.

BATEMAN, J. A. 1990. Upper modeling: Organizing knowledge for natural language processing. DTIC Document.

BAZAN, J., NGUYEN, H. S. & SZCZUKA, M. 2004. A view on rough set concept approximations. *Fundamenta Informaticae,* 59**,** 107-118.

BONILLO, M. L. 2003. *Razonamiento Basado en Casos aplicado a Problemas de Clasificación.* Ph. D. thesis, Universidad de Granada.

C.D., M., RAGHAN, P. & SCHÜTZE, H. Introduction to Information Retrieval. 2008 Cambridge University Press.

CABALLERO, Y., ARCO, L., BELLO, R., SALGADO, Y., MÁRQUEZ, Y. & LEÓN, P. Nuevas medidas de la Teoría de los Conjuntos Aproximados para la evaluación de sistemas de información en Bioinformática. presentado en II Congreso Internacional de Bioinformática y Neuroinformática Informática, 2007.

CAWLEY, G. C. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. Neural Networks, 2006. IJCNN'06. International Joint Conference on, 2006. IEEE, 1661-1668.

CHAHAL, M. & SINGH, J. 2013. Importance of Similarity Measures in Query Optimization: A Review. *International Journal of Science, Engineering and Computer Technology,* 3**,** 103.

COHEN 2010.

CRUAÑES VILAS, J., ROMÁ FERRI, M. T. & LLORET, E. 2012. Análisis del uso de métodos de similitud léxica con conocimiento semántico superficial para mapear la información de enfermería en español.

DENNY, J. C., SPICKARD, A., JOHNSON, K. B., PETERSON, N. B., PETERSON, J. F. & MILLER, R. A. 2009. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association,* 16**,** 806-815.

DIXON, M. 1997. An overview of document mining technology. http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/dixm97_dm.ps.

DOLIN, R. H., ALSCHULER, L., BEEBE, C., BIRON, P. V., BOYER, S. L., ESSIN, D., KIMBER, E., LINCOLN, T. & MATTISON, J. E. 2001. The HL7 clinical document architecture. *Journal of the American Medical Informatics Association,* 8**,** 552-569.

DOLIN, R. H., ALSCHULER, L., BOYER, S., BEEBE, C., BEHLEN, F. M., BIRON, P. V. & SHABO, A. 2006. HL7 clinical document architecture, release 2. *Journal of the American Medical Informatics Association,* 13**,** 30-39.

DUCH, W. 2002. Similarity-based methods: a general framework for classification. *Control and Cybernetics,* 29**,** 937-968.

EL-SAYED, M. A. 2013. NEW SIMILARITY MEASURE BASED ON MINKOWSKI DISTANCE AND ITS APPLICATIONS IN FACE RECOGNITION. *Advances in Computer Science and Engineering,* 11**,** 1.

FRAKES, W. B. & BAEZA-YATES, R. 1992. *Information Retrieval. Data Structure & Algorithms,* New York, Prentice Hall.

FUENTES, I. E., GUEVARA, D. M. & GARCÍA, M. M. 2016. Toma de decisiones inteligente a partir de registros médicos almacenados en CDA-HL7. *Revista Cubana de Informática Médica,* 16.

FUENTES, I. E., MAGDALENO, D. & GARCÍA, M. M. 2015. Metodología para asistir la toma de decisiones diagnóstica a partir del descubrimiento del conocimiento implícito en Historias Clínicas. 29.

GROSSMAN, D. A. & FRIEDER, O. 2012. *Information retrieval: Algorithms and heuristics*, Springer Science & Business Media.

HAND, D. J. 1981. *Discrimination and classification*, John Wiley and Sons.

HATCHER, E., GOSPODNETIC, O. & MCCANDLESS, M. 2009. *Lucene in Action*.

HUANG, A. Similarity measures for text document clustering. Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, 2008. 49-56.

JURGENS, D. & STEVENS, K. The S-Space package: an open source package for word space models. Proceedings of the ACL 2010 System Demonstrations, 2010. Association for Computational Linguistics, 30-35.

KRUSE, R., DÖRING, C. & LESOR, M.-J. 2007. Fundamentals of Fuzzy Clustering. *In:* OLIVEIRA, J. V. D. & PEDRYCZ, W. (eds.) *Advances in Fuzzy Clustering and its Applications.* Est Sussex, England: John Wiley and Sons.

LEMON, M. J., CASTELLANOS, M. & STINGER, J. R. 2005. Method for content mining of semi-structured documents. Google Patents.

LORENZO, M. M. G. & BELLO, R. E. P. 1996. A model and its different applications to case-based reasoning. *Knowledge-based systems,* 9**,** 465-473.

MAGDALENO, D. 2008. *Refinamiento, evaluación y etiquetamiento de grupos textuales basados en la teoría de los conjuntos aproximados.* Máster en Ciencia de la Computación, Universidad Central "Marta Abreu" de Las Villas.

MAGDALENO, D., FUENTES, I. E. & GARCÍA, M. M. 2015. Clustering XML Documents Using Structure and Content based on a New Similarity Function OverallSimSUX. *Computación y Sistemas,* 19**,** 151-161.

MAGDALENO, D., I.E. FUENTES, AND M.M. GARCÍA 2013. Sistema para el Agrupamiento de Artículos Científicos en formato XML usando Lucene (LucXML). N.R. 3267-2013, Editor.

MICHALSKI, R. S., STEPP, R. E. & DIDAY, E. 1981. A recent advance in data analysis: clustering objects into classes characterized by conjuntive concepts. *Progress in Pattern Recognition,* 1**,** 33-56.

PATHAK, J., BAILEY, K. R., BEEBE, C. E., BETHARD, S., CARRELL, D. S., CHEN, P. J., DLIGACH, D., ENDLE, C. M., HART, L. A. & HAUG, P. J. 2013. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *Journal of the American Medical Informatics Association,* 20**,** e341-e348.

PATIDAR, A. K., AGRAWAL, J. & MISHRA, N. 2012. Analysis of different similarity measure functions and their impacts on shared nearest neighbor clustering approach. *International Journal of Computer Applications (0975–8887) Vol,* 40.

PAWLAK, Z. 1982a. Rough sets. *International Journal of Computer & Information Sciences,* 11**,** 341-356.

PAWLAK, Z. 1982b. Rough sets. *International Journal of Parallel Programming,* 11**,** 341-356.

PURANDARE, A. & PEDERSEN, T. Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces. CoNLL, 2004. 41-48.

RHODE 2009.

RIPOLL GARRIDO, F. 2011. Tecnologia de la informacion al servicio de una gestion y salud de calidad. *Informatica Medica.*

SALTON, G., WONG, A. & YANG, C.-S. 1975. A vector space model for automatic indexing. *Communications of the ACM,* 18**,** 613-620.

SEBASTIÁN, M.-Á. A., SANTASUSAGNA, L. B., MARÍN, A. P. & GARCÍA, A. T. 2006. *Gestión diaria del hospital,* Masson.

SHANKAR, R. 2012. *Evolutionary Document Clustering and Summarization of Scientific Articles using Frequent Itemsets.* International Institute of Information Technology Hyderabad.

STREHL, A., GHOSH, J. & MOONEY, R. Impact of similarity measures on Web-page clustering. Proceedings of the 17th National Conference on Artificial Intelligence

(AAAI 2000): Workshop of Artificial Intelligence for Web Search, 2000 Austin, Texas. 58-64.

TAGHVA, K. & VENI, R. Effects of similarity metrics on document clustering. Information Technology: New Generations (ITNG), 2010 Seventh International Conference on, 2010. IEEE, 222-226.

TAN, A. Text Mining: The state of the art and the challenges. Proceedings of the Conference Knowledge Discovery and Data Mining (PAKDD'99): Workshop Knowledge Discovery from Advanced Databases, 1999 Pacific Asia. 65-70.

VALDÉS, B. 2017. *Razonador Basado en Casos para la deducción de hipótesis diagnósticas ,basado en la integración de recursos lingüísticos.*

VISVAL, A. & SARA, M. 2009. La gestión documental, de información y el conocimiento en la empresa. El caso de Cuba. *Acimed,* 19.

VRIES, C. M. D., NAYAK, R., KUTTY, S., GEVA, S. & TAGARELLI, A. 2011. Overview of the INEX 2010 XML mining track : clustering and classification of XML documents. *In Lecture Notes in Computer Science, Springer.* Amsterdam.

WILSON, D. R. & MARTÍNEZ, T. R. 1997. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research,* 6**,** 1-34.

ZWAANSWIJK, M., VERHEIJ, R. A., WIESMAN, F. J. & FRIELE, R. D. 2011. Benefits and problems of electronic information exchange as perceived by health care professionals: an interview study. *BMC health services research,* 11**,** 1.

Anexo 1. Similitudes, distancias más usadas para comparar objetos y medidas de calidad

Sean los objetos O_i y O_j descritos por m rasgos, donde O_i= (o_{i1}, …, o_{im}) y O_j=(o_{j1}, …, o_{jm})

Distancia Euclidiana

$$D_{Euclideana}\left(O_i, O_j\right) = \sqrt{\sum_{k=1}^{m}\left(o_{ik} - o_{jk}\right)^2} \qquad \text{(A1.1)}$$

Distancia <u>Minkowski</u> (Batchelor, 1978)

$$D_{Minkowski}\left(O_i, O_j\right) = \left(\sum_{k=1}^{m}\left|o_{ik} - o_{jk}\right|^{\gamma}\right)^{\frac{1}{\gamma}} \text{ donde } \gamma \geq 1 \qquad \text{(A1.2)}$$

La distancia <u>Minkowsky</u> es equivalente a la distancia <u>Manhattan</u> o <u>city-block</u>,y a la distancia Euclidiana cuando γ es 1 y 2, respectivamente(Batchelor, 1978). Para los valores de $\gamma \geq 2$, la distancia <u>Minkowsky</u> equivale a <u>Supermum</u> (Hand, 1981).

Distancia Euclidiana heterogénea (<u>HeterogenousEuclidean – OverlapMetric</u>; HEOM)

$$D_{HEOM}\left(O_i, O_j\right) = \sqrt{\sum_{k=1}^{m} d_{local}\left(o_{ik}, o_{jk}\right)^2} \text{ , donde}$$

$$d_{local}\left(o_{ik}, o_{jk}\right) = \begin{cases} d_{Overlap}\left(o_{ik}, o_{jk}\right) & si\ k\ simbólico \\ d_{NormEuclidean}\left(o_{ik}, o_{jk}\right) & si\ k\ numérico \end{cases} \qquad \text{(A1.3)}$$

$$d_{Overlap}\left(o_{ik}, o_{jk}\right) = \begin{cases} 0, & si\ o_{ik} = o_{jk} \\ 1, & en\ otro\ caso \end{cases} \text{ y } d_{NormEuclidean}\left(o_{ik}, o_{jk}\right) = \frac{\left|o_{ik} - o_{jk}\right|}{\max_k - \min_k}$$

Distancia <u>Camberra</u> (Michalski et al., 1981)

$$D_{Camberra}\left(O_i, O_j\right) = \sum_{k=1}^{m}\frac{\left|o_{ik} - o_{jk}\right|}{\left|o_{ik} + o_{jk}\right|} \qquad \text{(A1.4)}$$

Correlación de <u>Pearson</u>(Wilson and Martínez, 1997)

$$D_{Pearson}\left(O_i, O_j\right) = \frac{\sum_{k=1}^{m}\left(o_{ik} - \overline{atributo_k}\right)\left(o_{jk} - \overline{atributo_k}\right)}{\sqrt{\sum_{k=1}^{m}\left(o_{ik} - \overline{atributo_k}\right)^2 \sum_{k=1}^{m}\left(o_{jk} - \overline{atributo_k}\right)^2}} \qquad \text{(A1.5)}$$

donde $\overline{atributo_k}$ es el valor promedio que toma el $atributo_k$ en el conjunto de datos.

Las expresiones de Chebychev, Mahalanobis, distancia de Hamming y la máxima distancia son otras variantes de cálculo de distancias entre objetos (Wilson and Martínez, 1997). En(Duch, 2002) se presentan formas de medir la similitud considerando una relación asimétrica entre los objetos y teniendo en cuenta la probabilidad condicional de un objeto respecto al otro.

A partir de estudios realizados, se ha demostrado que al agrupar documentos, los coeficientes Dice, Jaccard y Coseno, han reportado los mejores resultados(Frakes and Baeza-Yates, 1992). Una valoración del impacto de la distancia Euclidiana y los coeficientes Dice, Jaccard y Coseno en dominios textuales se presenta en(Strehl et al., 2000).

Coeficiente Dice

$$S_{Dice}\left(O_i, O_j\right) = \frac{2\sum_{k=1}^{m}\left(o_{ik} \cdot o_{jk}\right)}{\sum_{k=1}^{m} o_{ik}^2 + \sum_{k=1}^{m} o_{jk}^2} \qquad \text{(A1.6)}$$

Coeficiente de Jaccard

$$S_{Jaccad}\left(O_i, O_j\right) = \frac{\sum_{k=1}^{m}\left(o_{ik} \cdot o_{jk}\right)}{\sum_{k=1}^{m} o_{ik}^2 + \sum_{k=1}^{m} o_{jk}^2 - \sum_{k=1}^{m}\left(o_{ik} \cdot o_{jk}\right)} \qquad \text{(A1.7)}$$

Coeficiente Coseno

$$S_{Coseno}\left(O_i, O_j\right) = \frac{\sum_{k=1}^{m}\left(o_{ik} \cdot o_{jk}\right)}{\sqrt{\sum_{k=1}^{m} o_{ik}^2 \cdot \sum_{k=1}^{m} o_{jk}^2}} \qquad \text{(A1.8)}$$

Anexo 2. Cálculo de la Pertenencia Aproximadd

La pertenencia de un documento a un grupo cuantifica el grado de solapamiento relativo entre $R'(x)$ y el grupo al cual el documento x pertenece,. Esta función puede interpretarse como una estimación basada en frecuencias de la probabilidad condicional de que el documento y pertenezca al conjunto X, dados los valores del documento x con respecto al conjunto de atributos. El valor $\mu_x(x)$ mide el grado de inclusión del documento x en el grupo X (Skowron 2000; Grabowski 2004).

$$\mu_x(x) = \frac{|X \cap R'(x)|}{|R'(x)|} \qquad \text{A2. 1}$$

$$w_i(x) = \frac{\sum_{x \in X_i} \mu_{X_i}(x)}{|X_i|} \qquad \text{A2.2}$$

-Otra expresión para medir la pertenencia aproximada de un documento x a un grupo X, cuantificando en qué grado la clase de similitud de x (R'(x)) es capaz de cubrir el grupo X se utiliza en (Arco, Bello et al. 2006; Caballero 2007; Caballero, Arco et al. 2007; Caballero, Arco et al. 2007).Las expresiones (2.6) y (2.7) muestran su forma de cálculo y la variante de ponderación que se deriva de ella, respectivamente

$$V_X(x) = \frac{|X \cap R'(x)|}{|X|} \qquad \text{A2. 3}$$

$$w_i = \frac{\sum_{x \epsilon X_i} V_{X_i}(x)}{|X_i|} \qquad \text{A2.4}$$

En este trabajo se propone calcular la pertenencia de los documentos a los grupos combinando las A2.1 y A2.3 .La idea es considerar de forma integrada el grupo X y el conjunto de documentos relacionados con x. Así, en el denominador del cociente se calcula la cardinalidad de la unión de estos conjuntos. La expresión (A2.5) muestra la nueva forma de cálculo de la pertenencia aproximada y la expresión (A2.6) su uso para ponderar los grupos.

$$\varpi_X(x) = \frac{|X \cap R'(x)|}{|X \cup R'(x)|} \qquad \text{A2.5}$$

$$w_i = \frac{\sum_{x \in X_i} \varpi_{X_i}(x)}{|X_i|} \qquad \text{A2.6}$$

Anexo 3. Caracterización de los expertos

Con la finalidad de realizar una validación del modelo para el descubrimiento de conocimiento propuesto, en este trabajo se utilizaron colecciones de HCE del Servicio de Admisión del Hospital Provincial "Celestino Hernández Robau" asociadas a diferentes enfermedades, en el área de la medicina oncológica, cardivascular, enfermedades infecciosas, respiratorias, etc.; que constituyen las causas más frecuentes de ingreso en la institución. Los datos utilizados fueron datos reales de pacientes ingresados. El manejo de la información se realizó bajo la supervisión de expertos, con el propósito de velar por la confidencialidad y la privacidad legal en el marco de la ética profesional, que establece el derecho del paciente a la reserva de toda la información relacionada con su proceso y con su estancia en la institución. En el proceso de conformación de los casos intervinieron varios especialistas en Medicina Interna, Cirugía, Oncología y Cardiología, cuyos roles estuvieron enmarcados indistintamente en la revisión y evaluación con la finalidad de estimar la calidad de la información a utilizar en el proceso de validación. A su vez, se contó con un conjunto de expertos Especialistas de Segundo Grado en estas áreas, profesores principales de Propedéutica Clínica y Medicina Interna de la Cátedra de Medicina y Atención Primaria de Salud. Los cuales emitieron sus consideraciones sobre las soluciones ofrecidas por el sistema y la reparación de ella en la fase de revisión:

- Dr. Rafael López Machado.
- Dr. Carlos Vázquez Subit.
- Dr. Fernando Aparicio Martínez.
- Dr. Javier Vázquez Roque.
- Dr. Miguel Ángel Rodríguez Gómez.
- Dr. Rolando de la C. Fuentes Morales.
- Dr. Mabel M. Herrera González.

Como resultado del proceso de estimar la calidad de la información de las HC, se detectaron pacientes cuya HC no incluían una diferenciación entre signos y síntomas atendiendo a su forma de presentación, no obstante estos casos fueron utilizados en el estudio con la finalidad de comprobar que el modelo, al igual que el experto para emitir recomendaciones precisas sobre

posibles diagnósticos, debe contar con información detallada en cada una de la etapas del interrogatorio médico-paciente durante la admisión.

Encuesta realiza a los expertos

Esta encuesta se realiza con el propósito de conocer las opiniones que se tienen acerca de la satisfacción que tiene con los resultados obtenidos por el agrupamiento. Los objetivos de la encuesta son investigar y explorar acerca de la eficacia, la eficiencia y la interpretación de los resultados en función de su criterio. Por ello le pediría que contestara las preguntas que contribuyen a medir las variables y objetivos trazados. No le tomará más de 15 minutos. Le pedimos que conteste este cuestionario con la mayor sinceridad posible. Lea las preguntas cuidadosamente.

1. ¿Se encuentra satisfecho con la división de los documentos en grupos al emplear los valores automáticos para el cálculo de la importancia de cada sección en el agrupamiento?

O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no

2. ¿Qué interpretación le confiere a los grupos obtenidos para pacientes que presentan iguales patologías pero diferentes causas?

_____

_____

_____

3. ¿Cuáles son a su juicio las secciones más importantes de la HC en un diagnóstico por comparación?

_____

_____

_____

4. ¿Descubre conocimiento con los resultados del agrupamiento?

O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no

5. ¿Los resultados del agrupamiento contribuyen a la toma de decisiones?

O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no

6. ¿Los documentos que arroja el sistema son a su juicio los más importantes o más representativos del grupo?

O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no

Justifique:_____

_____

_____

Resultados de la encuesta realiza a los expertos

Es importante en esta evaluación conocer el grado de conformidad de los expertos con los resultados obtenidos. Los objetivos de la evaluación en esta etapa son investigar y explorar acerca de la calidad de los resultados alcanzados con el agrupamiento realizado por la metodología, en función de su criterio. Las preguntas incluidas en la encuesta tributan a los objetivos propuestos.

Un elemento importante en esta evaluación fue la satisfacción de los expertos con los resultados obtenidos, determinada por la eficacia, la eficiencia y la interpretación. En este anexo se presenta la encuesta elaborada, siguiendo los lineamientos trazados en (Hernández- Sampieri 2007). La encuesta se aplicó a diferentes especialistas del Hospital Provincial "Celestino Hernández Robau" que emitieron una valoración de los resultados obtenidos por la metodología en colecciones de pacientes con patologías pertenecientes a sus áreas de especialización.

Tabla 7.1 Satisfacción de los usuarios respecto a los resultados del agrupamiento, el descubrimiento de conocimiento, la toma de decisiones y los documentos más representativos de cada grupo. (1) algunas veces, (2) casi siempre y (3) definitivamente sí

	Pregunta 1	**Pregunta 4**	**Pregunta 5**	**Pregunta 6**
(1)	0	0	0	0
(2)	0	20	10	10
(3)	100	80	90	90

Anexo 4. HCE basada en el estándar CDA-HL7

La implementación del estándar, estructura la información de la HCE del paciente en las *UE={Antecedentes, Hábitos Tóxicos, Síntomas, Signos, Historia de la Enfermedad Actual, Diagnóstico Diferencial, Complementarios, Tratamiento, Diagnóstico, Evolución, Pronóstico}* (Fuentes et al., 2015), refleja una correspondencia entre UE y secciones con etiqueta *component.*

```
<?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>
```

```
<component>
  <section>
   <code code="10164-2"
         codeSystem="2.16.840.1.113883.6.1"
         codeSystemName="LOINC"/>
   <title>Síntomas</title>
   <text>Disnea y tos nocturna húmeda …</text>
…
  </section>
</component>
```

```
<component>
  <section>
   <code code="10164-2"
         codeSystem="2.16.840.1.113883.6.1"
         codeSystemName="LOINC"/>
   <title>Signos</title>
   <text>Fiebre de 38 grados, murmullo vesicular
globalmente disminuido estertores crepitantes en
ambos pulmones, expansibilidad torácica
disminuida…</text>
…
  </section>
</component>
```

```
<component>
  <section>
   <code code="10164-2"
         codeSystem="2.16.840.1.113883.6.1"
         codeSystemName="LOINC"/>
   <title>Historia de la Enfermedad Actual</title>
   <text>Alrededor de una semana comienza con tos
seca, falta de aire y fiebre …</text>
  </section>
</component>
```

....

Figura A4.1 Identificar colecciones independientes, utilizando un ejemplo hipotético de HCE.