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 deducción de Hipótesis Diagnósticas basado en la integración de recursos lingüísticos.

Autor
Beatriz Valdés Pérez

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 mi familia

AGRADECIMIENTOS

A mis padres por apoyarme y guiarme siempre.

A mis tutoras Ivett y Marilyn por su gran ayuda y entrega en este proyecto.

A mis amigos por estar siempre presentes.

A mis profesores por sus enseñanzas.

A toda mi familia.

A todos, muchas gracias.

RESUMEN

Debido al incremento exponencial de la información almacenada en las organizaciones, la Sociedad de la Información está siendo superada por la necesidad de nuevos métodos capaces de procesar la información y asegurar su uso productivo. Esto se hace lógicamente extensible a los centros hospitalarios, a partir del uso extendido de las Historias Clínicas en formato electrónico. Surge así, la necesidad de disponer de herramientas que permitan asegurar el uso productivo de la información almacenada y contribuyan al ejercicio de una medicina basada en la evidencia y en la práctica del método clínico tradicional.

En este trabajo se implementó el sistema *CDAIS* 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 considera nuevos recursos lingüísticos para capturar la intencionalidad de los términos utilizados por los especialistas en la relatoría de las manifestaciones patológicas que refiere y presenta el paciente a su llegada, aportando a la fase de representación. El sistema propuesto que implementa el modelo permite la manipulación de Historias Clínicas a partir del estándar CDA, integrando agrupamiento y clasificación automática. Los resultados obtenidos luego de aplicar los experimentos, muestran como estos recursos lingüísticos contribuyen satisfactoriamente al comportamiento del modelo en la inferencia realizada a partir del método de solución de problemas RBC.

ABSTRACT

Due to the exponential increase of the stored information in the organizations, the society of the information is been surpassed by the necessity of new methods capable to process information and assure its productive use. This can be logically extendible to hospital centers, starting from the extensive use of clinical histories in electronic format. It arises this way, the necessity of preparing tools that allow to assure the productive use of the stored information and contribute to exercise a medicine based on the evidence and the practice of the clinic-traditional method.

On this project it was implemented the CDAIS system, that supports a model to the knowledge discovery implicated in electronic clinical histories on a semi-structured format, based on the hybridization of clustering and classification technics the created model for the deduction of possible diagnostic hypothesis, in the investigation considers some new linguistic resources in order to catch the intention of the used terms by the specialists in the telling of the pathological manifestations, that sais and presents the pacient at his arrival, lending to the representation phase. The proposed system that the model implements allows the manipulation of clinical histories, starting from the CDA standard, integrating clustering and automatic classification. The results obtained after applicating the experiments, shows how these linguistic resources, contribute satisfactorily to the model behavior in the deduction already done, starting from the problems solution method RBC.

Tabla de Contenidos

INTRODUCCIÓN

Las instituciones continuamente han demandado un uso de la información íntegra, que permita no sólo caracterizar su significado en su medio, sino además descubrir nuevo conocimiento, compartirlo y tomar decisiones(Lesniewska, 2009); principalmente en las organizaciones en las que se pueden realizar estudios basados en experiencias anteriores.

En el caso de gestión del conocimiento (GD) (Tekli et al., 2011) esto se traduce en la búsqueda, la explotación y evaluación del conocimiento, mediante un proceso de transformación; que comprende el desarrollo de sistemas que faciliten a los usuarios gestionar sus grandes colecciones textuales, mediante la organización y extracción del conocimiento oculto en la información. Estos sistemas a partir de una colección personal presentada como entrada, deben proponer como salida, grupos homogéneos de documentos afines, los términos relevantes y los documentos más representativos de cada grupo, así como las relaciones entre ellos y la calidad con que fueron obtenidos los grupos, proporcionando el control para la evaluación de los resultados del agrupamiento obtenido (Arco, 2009).

Disponer de herramientas que satisfagan este propósito, constituye una posibilidad para sistematizar la enorme riqueza de información que reside en los sistemas hospitalarios en entornos educativos y de investigación (Sebastián et al., 2006); ofreciendo ventajas para los profesionales de la salud, los pacientes, y el estado; al propiciar mejores condiciones de trabajo a los médicos, contribuyendo al ejercicio de una medicina basada en la evidencia y asegurando el uso productivo de la información almacenada (Zwaanswijk et al., 2011, Dalamagas et al., 2006).

A esto se le añade, que actualmente en nuestro país, los sistemas de registro de Historias Clínicas (HC) en las instituciones de salud son ineficientes, y no vinculan la información del paciente con los demás hospitales, clínicas u otros organismos de atención de salud pública. Es decir, las HC que se registran no son universales (Rodríguez, 2006). Por lo tanto, es imperativo que el Sistema de Salud del país haga la transición del registro de HC a un registro de Historias Clínicas Electrónicas (HCE) estandarizado, donde cada paciente tenga una HC única y con carácter universal.

Por otra parte, en el contexto mundial, los usos de la HCE impactan cada día de manera creciente y favorable en la investigación clínica, en la investigación farmacéutica (dígase ensayos clínicos, fármaco-epidemiológicos) y en las investigaciones en salud pública (dígase informes electrónicos de casos, bases de datos poblacionales), entre otros (Eysenbach, 2008, del Río Crespo et al., 2015). Como consecuencia, la creación de repositorios de HCE y el volumen de información generada desde estos, aumenta continua y exponencialmente.

Aunque se han desarrollado varios sistemas con el propósito de lograr una rápida y eficiente manera de compartir información, la heterogeneidad de ella determina que extraer conocimiento relevante sea un proceso complejo y desafiante (Dalamagas et al., 2006).

Varios investigadores (Pathak et al., 2013, Zwaanswijk et al., 2011, Denny et al., 2009) reconocen la importancia de la estandarización y codificación de los datos almacenados en la HCE, así como la necesidad de migrar a una recopilación de la información clínica mediante texto estructurado. Particularmente en el contexto de la HC, la propia distribución de sus elementos hace posible concebirla como un documento XML, debido a la estructura jerárquica y auto-descriptiva implícita en cada uno de los factores que la componen. De ahí que se han desarrollado varios intentos por lograr que los documentos de los Sistema de Información Hospitalaria (HIS) migren hacia documentos con formato semi-estructurados, debido a la ventaja que brinda etiquetar los documentos para el acceso a partes específicas de estos. De hecho, *Health Level Seven* (HL7), es el conjunto de estándares informáticos de salud más desarrollado y de mayor cobertura internacional para dar soporte a la HCE y 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 es un estándar de marcaje, realizado por el comité *Structured Documents Technical Committee* (SDTC) de HL7, que permite definir la estructura y la semántica de un documento clínico. Es una especificación que facilita el intercambio entre los diferentes sistemas en las organizaciones al utilizar XML. CDA logra que los documentos sean computacionalmente más legibles.

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) y garantiza la verdadera interoperabilidad semántica, permitiendo que los documentos sean procesables, mediante búsquedas y técnicas de Minería de Datos[1]. Estas ventajas y especialmente su codificación en el meta-lenguaje XML, hacen que CDA sea uno de las especificaciones más utilizadas regionalmente en los sistemas de gestión de información clínica (Slavov et al., 2013).

Por otra parte, *Extensible Markup Language* (XML) es un meta-lenguaje desarrollado por el consorcio W3C2 proveniente de *Generalized Markup Language* (GML) que surgió ante la necesidad de la gran empresa de almacenar grandes cantidades de información. Un documento XML es una estructura jerárquica auto-descriptiva de información, que consiste en un conjunto de átomos, elementos compuestos y atributos (Brau, 1998). Los documentos XML contienen su información en forma semi-estructurada, al incorporar estructura y datos en una misma entidad. Son extensibles, con estructura de fácil análisis y procesamiento, por lo que XML se ha convertido en el formato de intercambio de datos estándar entre las aplicaciones Web (Guerrini et al., 2006). Las etiquetas existentes en los documentos XML permiten la descripción semántica del contenido de sus elementos. De este modo, la estructura de los documentos puede ser explotada para realizar recuperación de documentos relevantes (Dalamagas et al., 2006).

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

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

partir de la información disponible en una colección obtenida como resultado de un proceso de recuperación de información (C.D. et al., 2008).

Un algoritmo de agrupamiento encuentra grupos naturales de datos, basándose principalmente en la similitud y las relaciones de los objetos, de forma tal que se obtenga una distribución interna del conjunto de datos en grupos. Cuando el agrupamiento se basa en la similitud de los objetos, se desea que los objetos que pertenecen al mismo grupo sean similares y los objetos que pertenecen a grupos diferentes sean disímiles (Pascual et al., 2007).

Este proceso del cálculo de la similitud a partir de una colección de documentos tiene como tarea fundamental realizar el pre-procesamiento y la representación textual de la colección, que en buena parte determina el comportamiento y la calidad del proceso (Hofmann, 2000).

Para el desarrollo de estas aplicaciones resulta importante seleccionar un modelo de representación textual que tenga en cuenta tanto la información sintáctica como semántica (Torres López and Arco García, 2016). Lo antes expuesto se debe a que el lenguaje natural, entendido como la herramienta para expresarse, posee propiedades que merman la efectividad de los sistemas de recuperación de información textual. Estas propiedades son la variación y la ambigüedad lingüística. La variación lingüística se entiende como la posibilidad de utilizar diferentes palabras o expresiones para comunicar una misma idea. En cambio, la ambigüedad lingüística se produce cuando una palabra o frase permite más de una interpretación (Vallez and Pedraza, 2007).

Ambos fenómenos inciden en el proceso de recuperación de información aunque de forma distinta. La variación lingüística provoca el silencio documental, es decir la omisión de documentos relevantes para cubrir la necesidad de información, ya que no se han utilizado los mismos términos que aparecen en el documento. En cambio, la ambigüedad implica el ruido documental, es decir la inclusión de documentos que no son significativos, ya que se recuperan también documentos que utilizan el término pero con significado diferente al requerido. Estas dos características dificultan considerablemente el tratamiento automatizado del lenguaje.

A nivel morfológico una misma palabra puede adoptar diferentes roles morfo-sintácticos en función del contexto en el que aparece, ocasionando problemas de ambigüedad, por lo cual utilizar recursos que permitan la desambiguación de las palabras permite capturar con mayor efectividad el conocimiento oculto en la información.

Como consecuencia, contemplar estos elementos en el desarrollo de herramientas que permitan gestionar la información de la HC, fuente esencial de un HIS a partir de su estructura específica, puede aportar resultados favorables al agrupamiento de este tipo de documentos, y contribuir de manera significativa a la gestión del conocimiento y la toma de decisiones.

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

La necesidad de abordar **otros recursos lingüísticos** que se incluyen en la relatoría que realiza el especialista de las manifestaciones patológicas que refiere y presenta el paciente a su llegada y que influyen en la recuperación de hipótesis tempranas.

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

Desarrollar un Razonador Basado en Casos (RBC) para contribuir a la toma de decisiones y el entrenamiento asistencial a partir de registros clínicos basado en la integración de recursos lingüísticos.

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

1. Analizar el efecto de incorporar el análisis semántico a partir del modelo de representación LSA al RBC.
2. Tratar los problemas presentes en la escritura de los textos clínicos a partir del manejo de la negación la sinonimia y los modificadores.
3. Definir la organización de la base de casos a partir del agrupamiento jerárquico.
4. Implementar el RBC atendiendo a la incorporación del tratamiento de los recursos lingüísticos.
5. Evaluar el sistema basado en casos a partir de un caso de estudio de colecciones de HCE.

Las **preguntas de investigación** planteadas son:

- ¿Cuáles métodos de representación textual son aplicables de manera efectiva en colecciones de documentos clínicos?
- ¿Qué elementos del procesamiento textual considerar y cómo integrarlos para obtener un esquema que permita descubrir conocimiento desde la HCE?

- ¿Qué herramientas de código abierto facilitan la implementación del modelo de descubrimiento de conocimiento en sus diferentes etapas?
- ¿Qué aspectos se deben considerar en las fases de adaptación y revisión para lograr que el razonador basado en casos facilite una clasificación efectiva?
- ¿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: El modelo para el descubrimiento de conocimiento en HCE en formato semi-estructurado, que combina la relación estructura-contenido y nuevos recursos lingüísticos, 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 nuevos recursos lingüísticos para refinar el cálculo de la similitud existente entre los pacientes como parte de la recuperación de los casos más semejantes ante la presencia de un nuevo problema.

La implementación del sistema *CDAIS* tiene un **valor práctico**, pues permite estimar la conducta a seguir en el tratamiento del paciente y las pruebas diagnósticas adecuadas para comprobar y emitir un diagnóstico final en la asistencia, a partir de la deducción de posibles hipótesis diagnósticas emitidas a la llegada del paciente y luego del interrogatorio realizado por el especialista.

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 los recursos lingüísticos que permiten refinar la valoración de los términos en el procesamiento textual. Se hace énfasis en las formas de representación textual sintáctica y semántica. En el Capítulo 2 se detalla sobre el modelo propuesto combinando nuevos recursos lingüísticos. 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 sistema

CDAIS. Este documento culmina con las conclusiones, las recomendaciones, referencias bibliográficas y los anexos.

1

ACERCA DEL DESCUBRIMIENTO DE CONOCIMIENTO DE INFORMACIÓN CLÍNICA SEMI-ESTRUCTURADA

1. ACERCA DEL DESCUBRIMIENTO DE CONOCIMIENTO DE INFORMACIÓN CLÍNICA SEMI-ESTRUCTURADA

La gran cantidad de información clínica generada en los centros hospitalarios y la riqueza de información que estas poseen, evidencia la necesidad de desarrollar nuevas técnicas; que permitan el análisis exploratorio de esta y que capturen eficientemente las relaciones internas en la estructura jerárquica y auto-descriptiva de las Historias Clínicas, en beneficio una asistencia médica efectiva (Aruquipa, 2014). Dos tareas importantes en el desarrollo de aplicaciones de gestión del conocimiento son el pre-procesamiento y la etapa de representación (Abelleira and Cardoso, 2010). 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. Además, se describen diferentes formas de realizar el pre-procesamiento, así como modelos de representación textual que consideran la información sintáctica y semántica.

1.1 Gestión de la información inteligente

Sin lugar a dudas las tecnologías de la información y la comunicación (TIC) influyen decisivamente en el nivel de desarrollo y el bienestar del ser humano. Por lo que, la urgencia en su implementación, asimilación y uso marcan puntos decisivos de cambios sociales, económicos y culturales necesarios a nivel global (Hernandez, 2009)

El gran reto por tanto de las organizaciones, es 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 la 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) se plantea un conjunto de objetivos encaminados a lograr desarrollar productos y servicios que permitan incorporarle un valor añadido a la información (Grossman and Frieder, 2012), hasta convertirla en conocimiento, estos son (Mesa, 2006):

Realizar un diseño normalizado de documentos, que permita controlar su uso.

Evitar la creación de documentos innecesarios, la duplicidad y la presencia de versiones caducadas. Seleccionando aquellos documentos que tengan valor para la gestión y para el futuro.

Organizar (clasificar, ordenar y describir) los documentos para su adecuada explotación al servicio de la gestión y la toma de decisiones; de manera que estos procedimientos se simplifiquen.

Permitir la recuperación de información de una forma mucho más rápida, efectiva y exacta.

Lograr que los archivos sean vistos dentro y fuera de la organización como verdaderas unidades de información útiles no solo para la administración sino también para la cultura.

Específicamente, los Sistemas de GD (SGD), permiten relacionar los documentos entre sí y darles una semántica común. De manera que faciliten la búsqueda de información en toda la colección y provean 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).

La GD no es una actividad aislada dentro de las organizaciones y por lo tanto debe integrarse con la Gestión de la Información (GI) y la Gestión del Conocimiento (GC) teniendo estas como punto común en el proceso de convertir el conocimiento tácito en explícito.

Por otra parte, el conocimiento se puede gestionar de diversas formas y hacerlo 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).

1.2 Historias Clínica Electrónica

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

La HC incluye la información clínica relacionada con los datos del paciente, antecedentes personales y familiares, hábitos tóxicos y todos los elementos relacionados con su salud biopsicosocial; el proceso evolutivo, el tratamiento y la recuperación (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.

Por tanto, incorporar las TIC en el núcleo de la actividad hospitalaria, implica ofrecer soporte a las HCE que integran el HIS de la institución, lo cual se traduce en 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).

Aunque, en todo el mundo se encuentran en desarrollo varias iniciativas sobre estándares de información para el área salud.

Health Level Seven (HL7), se ha convertido en el conjunto de estándares informáticos de salud más desarrollado y de mayor cobertura internacional (Zwaanswijk et al., 2011), al ser un estándar no propietario, reconocido por el Comité Técnico de Información para Salud de la Organización Internacional para la Estandarización (ISO) (Ripoll Garrido, 2011).Para dar soporte a la HCE HL7, ofrece *Clinical Document Architecture* (CDA) de HL7 que especifica la estructura y semántica de los documentos clínicos mediante el metalenguaje XML, facilitando su intercambio en un entorno de interoperabilidad.

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, provee la información de los elementos que ha presentado y referido el paciente durante su enfermedad y la conducta y tratamiento de los especialistas durante su atención (Treins et al., 2006).

En la Figura 1.1 se observa un ejemplo de una HCE correspondiente a un documento XML definido en el estándar CDA.

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

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

1.3 XML

XML es un metalenguaje que permite jerarquizar, estructurar la información y describir los contenidos dentro del propio documento, así como la reutilización de partes del mismo. (Menárguez, 2013).

Por medio del XML es posible definir los documentos con el grado de exhaustividad requerido. Una de sus características principales es organizar jerárquicamente todas las unidades

informativas de un documento mediante estructuras lógicas. En la terminología de XML, estas unidades se denominan entidades (*entities*). XML posee mecanismos que permiten revisar la estructura lógica de los documentos con el propósito de que las computadoras que se interconecten entre sí para operar con estos datos lo hagan de forma fluida. El acceso a los documentos XML se realiza mediante un procesador que revisa la estructura de los documentos e interpreta los contenidos de acuerdo a una gramática.

La gramática de los lenguajes XML, es decir, la estructura y los elementos permitidos en los documentos XML, se define mediante:

- DTD (Document Type Definition): Documento ASII plano que especifica tanto los elementos que forman un tipo de documento dado, como las relaciones que se dan entre ellos.
- XSD (XML Schema Definition): Mejoran los DTD's porque están escritos en XML y permiten nuevas característica.

Entre las ventajas de adoptar XML podemos destacar las siguientes: los autores o proveedores pueden diseñar sus propios documentos a medida, diseñando sus propias etiquetas en dependencia de las funciones que quieran dar a los datos. La información contenida en un documento XML puede ser más rica y fácil de usar, ofrece más facilidades para la representación en los navegadores ya que elimina muchas de las complejidades de SGML en aras de una mayor flexibilidad del modelo, con lo que la escritura de programas para manejar XML es mucho más sencilla. La información será más accesible y reutilizable porque la flexibilidad de las etiquetas de XML puede utilizarse sin tener que amoldarse a las reglas específicas de un fabricante (Cabarcas et al., 2015).

1.4 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.

Específicamente, el *agrupamiento* es descrito como una herramienta para el descubrimiento de conocimiento como se mencionó al inicio del capítulo, porque tiene la potencialidad de revelar

relaciones basadas en datos complejos no detectadas previamente (Kruse et al., 2007, Anderberg, 1973, Shankar, 2012), por lo cual es fundamental para una eficiente organización y recuperación de los documentos relevantes.

1.4.1 **Pre-procesamiento**

Cuando de colecciones textuales se trata para descubrir conocimiento, resulta necesario realizar una fase previa de *pre-procesamiento*, es decir, una serie de transformaciones que producen una Forma Intermedia (FI) del texto original que permiten tratarlo computacionalmente. Entre ellas, se destacan la confección del texto plano, la eliminación de las palabras vacías, la extracción de raíces o lemas, el manejo de la negación y las palabras modificadoras de términos. Esta etapa genera un conjunto de palabras o términos más pequeño y de mayor calidad que el original. El pre-procesamiento es prácticamente inviolable pues influye en gran medida en la calidad del proceso de agrupamiento.

1.4.1.1 Conversión de un documento a texto plano

Desde el punto de vista computacional es necesario hacer más ligeros los documentos; o sea, minimizar el espacio que ocupan en memoria.

Por las razones anteriores se hace necesario realizar la conversión de los documentos a archivos de texto plano, o sea que contengan el contenido que interesa en una cierta etapa del procesamiento. El texto plano texto simple, como también se le conoce, son solo caracteres, texto sin formatear; es decir, sin códigos de tipos de letras, negritas, cursivas, formatos de párrafos, etc. En esta etapa además se sustituyen las mayúsculas por minúsculas y se eliminan los signos de puntuación y los acentos.

1.4.1.2 Reducción de palabras vacías

En el contenido de los textos se pueden encontrar frecuentemente palabras que se consideran carentes de utilidad. Entre ellas están palabras como los artículos que no es conveniente tener en cuenta a la hora de determinar la similitud entre unidades textuales por repetición de términos. A este tipo de palabras se les denomina palabras vacías o stop words.

De este modo resulta indispensable eliminar las palabras vacías del texto antes del agrupamiento, creando una lista de términos vacíos verificando la presencia de cada palabra en la misma. Esta lista está formada por las preposiciones, conjunciones, artículos, pronombres, así

como todas aquellas palabras que suelen ser poco discriminantes por su elevada frecuencia de aparición en el texto.

1.4.1.3 Extracción de raíces o lemas

Como parte del procesamiento de texto y específicamente del procesamiento de lenguaje natural se encuentran la extracción de raíces y la extracción de lemas. El objetivo principal de dichas tareas es obtener, en el mínimo número de caracteres posibles, el máximo de información del término.

Raíz léxica o lexema: es la unidad léxica primaria de una palabra, que lleva los aspectos más significativos del contenido semántico y que no se puede reducir en componentes más pequeños. Por ejemplo, los términos *hablan* y *hablando* se reducirían a la raíz *habl*.

Lema: es cada una de las entradas de un diccionario o enciclopedia. El lema define un conjunto de palabras con la misma raíz léxica, y que pertenece a la misma categoría gramatical (verbo, adjetivo, etc.).

La lematización pretende normalizar los términos pertenecientes a una misma familia y por tanto próximos en significado, reduciéndolos a una forma común o lema, que no coincide necesariamente con la raíz. Por ejemplo, los términos *hablan* y *hablando* se reducirían al lema *hablar*.

1.4.1.4 Manejo de la negación

El tratamiento de la negación es un problema abierto dentro del procesamiento del lenguaje natural, y la minería de opinión en particular. Son muy abundantes las opiniones negativas expresadas con términos positivos negados y viceversa. Por ejemplo, la oración "*No me gusta la carcasa del teléfono"* se puede apreciar una opinión negativa con un término positivo (*gusta*) negado (Zafra et al., 2015).

En lo que respecta al tratamiento de la negación, (Taboada et al., 2011) utiliza información morfológica para identificar el alcance de la negación, mientras que (Yang et al., 2008) considera dicho alcance como los términos situados a la derecha de la negación y en (Fernández Anta et al., 2013) se emplea una heurística que asume que los tres elementos a continuación de una negación son los que deben cambiar su polaridad. Otros como (Vilares et al., 2013) siguen

una estrategia distinta, que se basa en obtener la estructura sintáctica del texto para tratar las construcciones lingüísticas e identificar los elementos de la frase.

Por otra parte, trabajos como (Vilares et al., 2013) limitan el tratamiento de la negación a los términos *no*, *nunca* y *sin*. Otros como (Amores et al.) crean listas de palabras negadoras a partir de diversas listas ya publicadas y en (Zafra et al., 2015) parten de partículas de corte negativo como: "*no", "tampoco", "nadie", "jamás", "ni", "sin", "nada", "nunca" y "ninguno"*.

Estos términos negadores se consideran palabras vacías y al eliminarlos se está descartando la presencia de la negación y asumiendo que la opinión es positiva cuando en realidad no lo es, lo que trae inconsistencias en la etapa de representación y agrupamiento.

1.4.1.5 Manejo de palabras modificadoras

Las palabras modificadoras son aquellas que pueden incrementar, reducir o cambiar la polaridad de otro término dentro de la oración. Especial atención tienen las palabras intensificadoras las cuales aumentan la fuerza de otros términos y las atenuadoras que reducen la fuerza de otros términos. En gramática, un intensificador es una palabra que hace hincapié en otra palabra o frase. También conocido como un refuerzo o un amplificador. Por ejemplo, algo no sólo es bueno, sino muy bueno, o incluso, terriblemente bueno. Los adjetivos intensificadores modifican a los sustantivos; los adverbios de intensificación modifican comúnmente verbos, adjetivos graduables, y otros adverbios (Amores et al.).

La mayoría de estos términos se consideran palabras vacías por lo que se eliminan en esta etapa de pre-procesamiento y no se tienen en cuenta para la etapa de representación. Por ejemplo, no es lo mismo un paciente que tuvo "*poca fiebre"* a uno que tuvo *"mucha fiebre",* o uno que tuvo una "*buena evolución al tratamiento"* a uno que tuvo "*pésima evolución al tratamiento"*. Considerar la presencia de estos términos sería de gran importancia para mejorar la calidad de la representación y del agrupamiento.

1.4.1.6 Manejo de la sinonimia

La preocupación por la sinonimia, como relación de semejanza significativa, es antigua (García-Hernández, 1997).

Griegos y romanos apuntaron a dos características fundamentales de la sinonimia: 1) es una característica del significado de las palabras y atiende a la pluralidad de significantes de un mismo referente. 2) la sinonimia es una relación de semejanza de significados. Si bien los intentos de definir o caracterizar a la sinonimia son antiguos, su estudio sistemático como relación léxica es más reciente. Los diccionarios de sinónimos muestran a la sinonimia como una característica léxica y aproximada. Siguiendo, a veces reglas léxicas, como acontece cuando se prefija una palabra de diferentes maneras. Por ejemplo, se obtienen sinónimos en algún grado con la composición por concatenación de alguno de los siguientes prefijos *{pre, post, ultra, súper, sub}* con la palabra *código (Fernández Lanza and Sobrino Cerdeiriña, 2000).*

La sinonimia no puede definirse de manera clara y concisa dada la diversidad de definiciones y perspectivas en lingüística (Zapico and Vivas, 2014) señala que dentro de la poca atención que ha recibido el tema, cada una de las posturas que se han ido presentado, si bien resultan impecables en sí mismas, no llegan nunca a resultar viables si se busca su aplicación al análisis de situaciones comunicativas reales.

En (Martínez, 1988) se explica que la divergencia en las posturas está dada por puntos de partida diferentes (no se considera la misma definición de sinonimia, algunos analizan la lengua, otros analizan el habla, etc.) y en su propio análisis toma en cuenta que al actualizarse la lengua en el habla, se pierde cualquier neutralidad pretendida en los diccionarios. No hay que dejar nunca de lado que el diccionario en sí es una forma de definir fuera del uso el significado, que si bien está elaborada a partir de estudios sobre el uso, nunca podrá ser totalmente fiel al mismo, puesto que en cada acto de habla, la lengua cambia lentamente.

Así, el manejo de la sinonimia permite reducir la cantidad de términos a representar y a su vez mejorar la calidad de la representación de los mismos.

1.4.2 **Formas de representación textual**

Para el desarrollo de aplicaciones de agrupamiento de documentos resulta importante seleccionar un modelo de representación textual que tenga en cuenta tanto la información sintáctica como semántica. Los modelos de representación se dividen mayormente en tres grupos, los basados en el modelo espacio vectorial, los basados en grafos y los probabilísticos (Torres López and Arco García, 2016).

1.4.2.1 Modelo espacio vectorial

El modelo espacio vectorial (Vector Space Model; VSM) es una forma de representación textual que ha sido utilizada en campos de recuperación de información y el procesamiento de textos de forma general para representar documentos textuales a través de vectores de términos (Salton et al., 1975). Una interpretación de este modelo es: "*En VSM cada documento se identifica como un vector de rasgos en un espacio en el cual cada dimensión corresponde a términos indexados distintos (palabras). Un vector documento dado, en cada componente tiene un valor numérico para indicar su importancia*" (Ochoa and Arco, 2008). De forma general este modelo se basa en dos conceptos fundamentales: el esquema de pesos y la similitud entre dos vectores de términos (Seijo et al., 2011). Este esquema de pesos determina la manera en la que se asignan los pesos a los términos en los documentos, dependiendo de su importancia en el contenido de los mismos. El esquema de pesos más usado es la frecuencia de aparición de los términos en los documentos (Term Frequency / Inverse Document Frequency; TF-IDF) para expresar el peso relativo del rasgo *w* en el vector asociado a un documento dado y se calcula según la expresión (1.1), donde *idf* (*w*) se calcula según la siguiente expresión:

$$tfidf(w,d) = tf(w,d) * idf(w) \quad \textbf{(1.1)}$$

$$idf(w) = log\frac{N}{df(w)} \quad \textbf{(1.2)}$$

Así, *tf* (*w*,*d*) es la frecuencia del término (cantidad de ocurrencias de la palabra *w* en un documento *d*), *idf* (*w*) es la frecuencia inversa de documentos (cantidad de documentos donde aparece la palabra *w* pero de forma inversa, debido a que se le otorga mayor peso a las palabras que ocurren en una menor cantidad de documentos), *df* (*w*) es la frecuencia de documento (cantidad de documentos que contienen la palabra *w*) y *N* representa la cantidad total de documentos en el corpus (Aggarwal and Zhai, 2012, Manning et al., 2008).

La representación VSM se utiliza con frecuencia en el agrupamiento para representar los documentos con el objetivo de determinar la similitud entre ellos (Torres López and Arco García, 2016). En sentido general, la principal dificultad de este modelo es que se basa en una comparación estricta de los términos, por lo que la eficacia se ve afectada por palabras distintas que describen el mismo concepto (sinonimia). Esta forma de representación tiene la desventaja que opera en el plano estadístico de los documentos, es decir, considera los documentos como

bolsas de palabras (Bag of Words; BOW), y sufre de varias limitaciones para capturar estructuras estadísticas de los documentos (Berry and Kogan, 2010).

1.4.2.2 Análisis semántico latente

El análisis semántico latente (Latent Semantic Analysis; LSA) es un modelo que representa conceptos semánticos presentes en los documentos (Deerwester et al., 1990). Para obtener esta representación, primeramente se realiza la representación VSM del corpus textual. Por tanto, se representa el texto como una matriz término-documento, en el que cada fila representa una palabra y cada columna una oración, párrafo o documento. El valor en la intersección de fila y columna es la frecuencia con la cual la palabra aparece en el documento. Este valor se calcula a través de una función de peso que exprese la importancia del término en el documento. Para una colección real, esta matriz puede llegar a tener más de 100000 filas y columnas, por ello se aplica la técnica de factorización de matrices conocida por descomposición de valores singulares (Singular Value Descomposition; SVD), donde la matriz representada anteriormente puede ser descompuesta en el producto de tres matrices como se expresa en la ecuación (1.3):

$$X = T_0 S_0 D_0 \qquad \textbf{(1.3)}$$

La clave de LSA es usar SVD para construir una matriz (concepto x documento) donde el número de conceptos sea varios órdenes de magnitud menor al número de términos en la colección (típicamente para el número de conceptos se escoge un valor bastante inferior a 1000). En la ecuación (1.3) X representa una matriz $M \times N$ (M es el tamaño del vocabulario y N es el tamaño de la colección), T_0 es una matriz $M \times K$ (K es la cantidad de conceptos), S_0 es una matriz $K \times K$ (matriz diagonal de valores singulares), D_0 es una matriz $K \times N$ (la matriz que va a asociar cada documento con sus conceptos). T_0 y D_0 son matrices cuadradas con columnas ortonormales[2] y se denominan las matrices izquierda y derecha de vectores singulares; respectivamente. Todos los elementos de S_0 son positivos. Los valores más altos de S_0 se corresponden a las correlaciones dominantes en la colección. Si los valores singulares de S_0 están ordenados, se pueden tomar los k valores más altos (significa quedarse con los k conceptos más prominentes en la colección), fijar el resto a 0 y obtener una aproximación a X, que es la

[2] Es ortonormal si es a la vez un conjunto ortogonal y la norma de cada uno de sus vectores es igual a 1.

matriz de rango *k* más cercana a *X* según mínimos cuadrados (Seijo et al., 2011). Al descartar las dimensiones (filas y columnas) con valor 0 se obtiene la matriz aproximada $\hat{X}$, donde *S* es *k x k* y *T* (*M x k*) y *D'* (*k x N*) surgen a partir de eliminar en T_0 y D_0 las columnas asociadas a las dimensiones descartadas en S_0. Así se obtiene el modelo reducido expresado en (1.4):

$$X \sim \hat{X} = TSD' \quad \textbf{(1.4)}$$

La elección de *k* es importante pues determina la cantidad de conceptos resultantes en el espacio semántico. Este valor debe ser suficientemente alto para ajustarse apropiadamente a la estructura conceptual que se tiene en la colección pero sin llegar a ser tan alto como para sobreajustar la colección con detalles insignificantes (Seijo et al., 2011). Este modelo representa los vectores de documentos en un espacio dimensional asociado a los conceptos presentes en la colección, y por tanto, se considera una forma de representación textual, no obstante, otros autores clasifican este modelo como un método de reducción de dimensionalidad.

LSA se sobrepone a los asuntos de sinonimia para el agrupamiento (Aggarwal and Zhai, 2012) (Torres López and Arco García, 2016). También presenta como ventaja que reduce la dimensionalidad de documentos al proyectar los vectores BOW en un espacio semántico construido a partir de la matriz término-documento de SVD. LSA tiene como limitación que dada la naturaleza de alta dimensionalidad de los datos textuales, el cálculo de SVD puede ser costoso y las dimensiones resultantes pueden ser difíciles de interpretar debido a que cada dimensión es una combinación lineal de un conjunto de palabras a partir del espacio original (Berry and Kogan, 2010).

1.5 Algoritmos de agrupamiento jerárquicos

Los algoritmos de agrupamiento jerárquicos pueden ser aglomerativos (Hierarchical Agglomerative Clustering; HAC) o divisivos (Manning et al., 2008).

Dado un conjunto de *N* objetos para agrupar y una matriz de distancia o similitud el algoritmo básico de agrupamiento jerárquico consiste en:

Asignar a cada objeto su propio grupo (conformar *N* grupos, cada uno con un solo objeto).

Encontrar el par de grupos (más similar) y mezclarlos en un único grupo.

Repetir el paso 2 hasta que todos los objetos son agrupados en un único grupo.

Típicamente visualizados en un dendrograma, entre ellos se destaca el agrupamiento de enlace único (Single-linkage), agrupamiento de enlace promedio del grupo (Average-linkage) y el agrupamiento de enlace completo (Complete-linkage). El agrupamiento de enlace único calcula la similitud entre dos grupos considerando la similitud entre sus dos objetos más similares. Considera un criterio local, es decir, solo tiene en cuenta las áreas donde dos grupos están más cerca uno del otro; las partes más distantes del grupo no se tienen en cuenta (Manning et al., 2008). 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 and Zhai, 2012). Se ha recomendado como mejor algoritmo para el agrupamiento de documentos en representaciones vectoriales (Manning et al., 2008). El algoritmo de agrupamiento de enlace completo calcula la similitud de dos grupos como la similitud de sus miembros más disimilares. No es un criterio local, porque toda la estructura del agrupamiento puede influenciar decisiones de combinación de grupos. Este criterio favorece la obtención de grupos compactos con diámetros pequeños, pero es sensible a objetos que están lejanos. Un solo objeto lejos del centro puede incrementar el diámetro del grupo y cambiar el agrupamiento final. Tiene como desventaja que es sensible a los puntos que no se ajustan en la estructura global del grupo (Manning et al., 2008).

Los algoritmos HAC construyen la jerarquía hasta obtener un solo grupo donde se incluyen todos los objetos. Si se desea obtener grupos con determinada calidad, es posible utilizando este algoritmo de agrupamiento, cortar la jerarquía en un nivel para obtener la partición. Algunas variantes para obtener esta partición a partir del dendrograma son (Manning et al., 2008):

Cortar según un nivel pre-especificado de similitud entre objetos en un mismo grupo.

Cortar el dendrograma donde el espacio entre dos combinaciones de similitudes sucesivas entre objetos en un mismo grupo es mayor.

Estimar la suma de cuadrados como una función para una cantidad de grupos *K*.

Pre-especificar la cantidad de grupos *K* y seleccionar el punto de corte que produce *K* grupos.

Obtener todas las posibles particiones y seleccionar aquella que ofrezca la mejor calidad del agrupamiento.

Por otra parte, los algoritmos HAC suelen ser útiles cuando no se cuenta con conocimiento a priori que permita especificar los parámetros que requieren la mayoría de los algoritmos planos partitivos en su inicialización.

1.6 Modelo híbrido para la deducción de hipótesis diagnósticas

Debido a la importancia del desarrollo de técnicas efectivas para el descubrimiento de conocimiento en el contexto de la actividad hospitalaria, nuevos enfoques continúan emergiendo con la finalidad de obtener sistemas de mayor precisión (Vries et al., 2011). A ello, se adiciona que en gestión hospitalaria se requiere de emitir recomendaciones que contribuyan a asistir a los especialistas en la toma de decisiones diagnóstica ante la admisión de un nuevo paciente, desde las etapas tempranas del proceso de atención. Para contribuir a ello, en el CEI se ha propuesto un modelo que permite la deducción de posibles hipótesis diagnósticas (Fuentes Herrera, 2016).

Generalidades del modelo

El modelo integra el agrupamiento y la clasificación basada en el método de solución de problemas RBC. Manipula HCE semi-estructuradas en el estándar de codificación CDA, el cual concibe los registros clínicos como documentos XML. Este se inicia a partir de la colección de documentos de HCE de entrada $D=\{D_1,\ldots, D_m\}$, donde cada HCE constan con un conjunto de Unidades Estructurales (UE) $UE=\{U_1,\ldots, U_n\}$ y como resultado se obtienen grupos homogéneos de documentos afines, los documentos más representativos de cada grupo y la calidad del agrupamiento; garantizando el control para la evaluación de los resultados.

Una visión gráfica del esquema del modelo general para el agrupamiento presentado en este trabajo se muestra en la Figura 1.2.

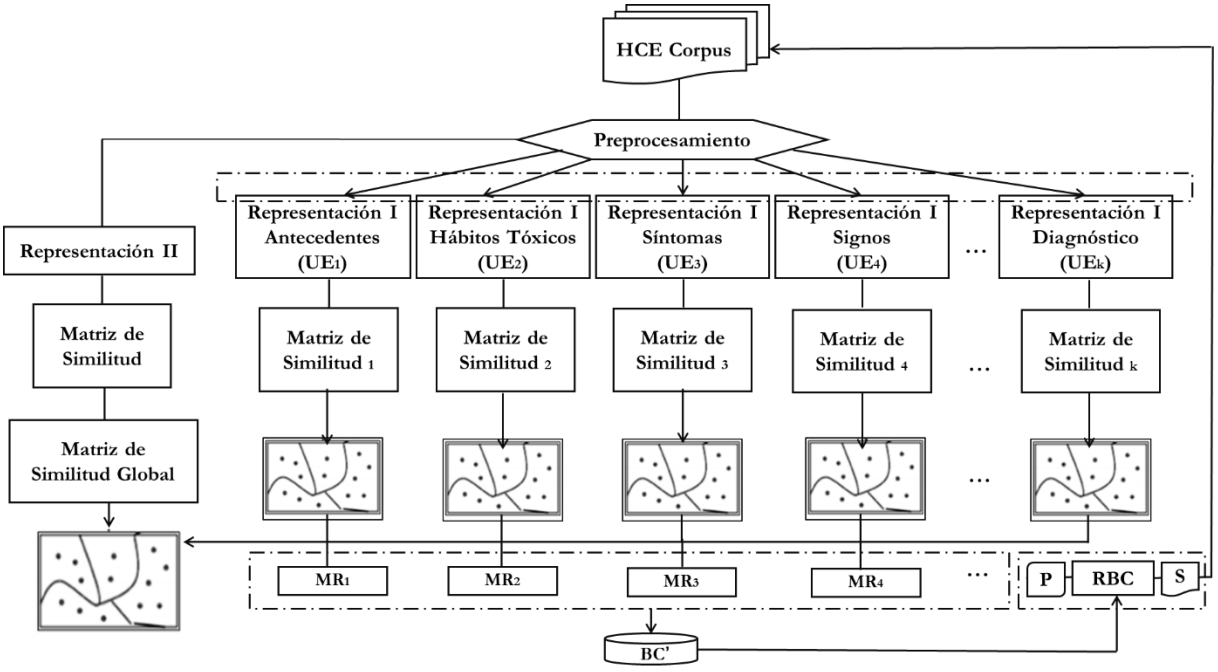

Figura 1.2: Modelo para el descubrimiento de conocimiento a partir HCE semi-estructuradas para la toma de decisiones

Etapa de Pre-Procesamiento

Para la aplicación del modelo se transforma el corpus de HCE convirtiendo cada uno los ficheros de entrada CDA en una secuencia de *tokens*[3] de palabras. En el paso subsecuente a la extracción de términos, estos *tokens* se usan para generar rasgos significativos (índices de términos). En esta etapa de trasformación del corpus se determina 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.

Etapa de Representación

En la fase 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. El modelo cuenta con dos tipos de representaciones (*Representación I*, para cada UE y *Representación II*, tomando en consideración toda la colección).

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

En la *Representación I* los valores numéricos iniciales coinciden con la frecuencia de aparición absoluta de cada término independiente en cada sección de la HCE. Esto significa que, si se tiene el término *"fiebre"*, la frecuencia de aparición para la *Representación I*, contabilizará el número de ocurrencias en la UE *Síntomas* y *Signos* por separado, considerando en cada caso únicamente la cantidad de veces que aparece en la colección tratada.

Esta etapa se resume en el esquema de la Figura 1.3

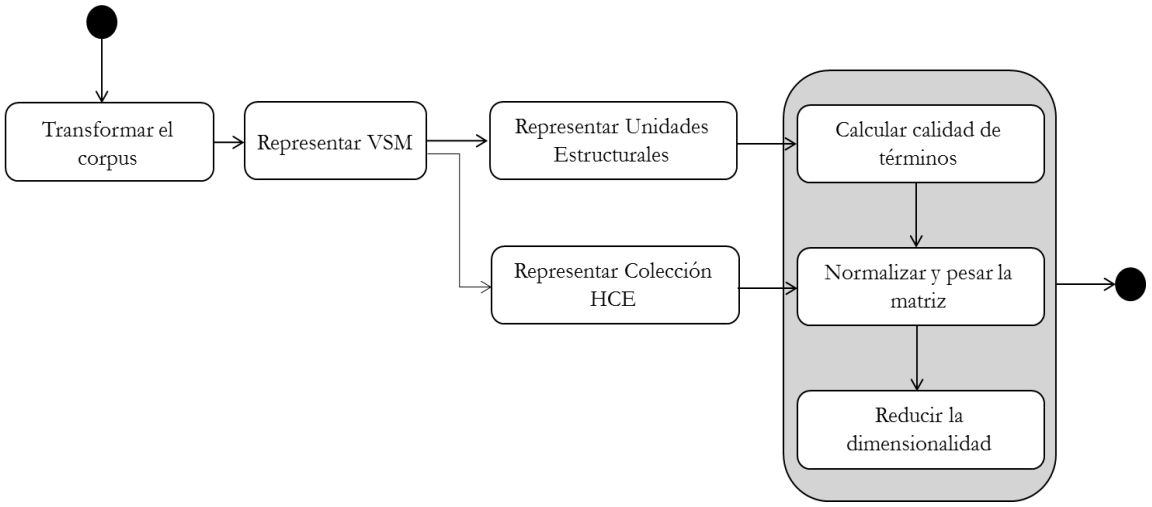

Figura 1.3: Esquema de representación del corpus textual y las *k*-colecciones.

Etapa del cálculo de la similitud y agrupamiento

Se obtiene la matriz de similitud obtenida aplicando la medida de similitud Coseno (Taghva and Veni, 2010) para cada *k*-colección a partir de la *Representación I*, resultados que se utilizarán para realizar los agrupamiento de cada sub-colección a utilizar en la organización de la BC sobre la cual realiza la inferencia el RBC.

Etapa de inferencia

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.

Del contenido que almacenan las HCE basadas en el estándar CDA-HL7 representadas en forma de documentos semi-estructurados, el enfoque basado en casos 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; el modelo 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.

La organización de la BC en el modelo se concibe a partir de conformación de grupos de documentos semi-estructurados de HCE. Estos se obtienen a partir del agrupamiento de las UE correspondientes a las colecciones independientes que constituyen rasgos predictores, antes comentados.

En la etapa de recuperación, se seleccionan el o los casos más similares al nuevo caso. Este proceso se realiza teniendo en cuenta los prototipos de los grupos de documentos obtenidos por el agrupamiento 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 (1.5).

$$\mathrm{SimGlobal}(\mathrm{X},\mathrm{Y}) = \sum_{\mathrm{i}=0}^{\mathrm{m}} W_i * \mathrm{SimCoseno_i}\left(\mathrm{X}_{\mathrm{UE}_i}, Y_{\mathrm{UE}_i}\right)/\mathrm{n} \qquad \textbf{(1.5)}$$

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

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

1.7 Consideraciones finales del capítulo

La proliferación de información disponible en los centros hospitalarios, evidencian la necesidad de desarrollar herramientas que permitan garantizar su uso productivo.

Para descubrir conocimiento desde información textual es necesario emplear modelos de representación textual que capturen tanto la información semántica como sintáctica presente en la secciones de los documentos clínicos que constituyen una relatoría. Aunque, VSM es el

modelo que sentó las bases para la representación de textos y ha sido ampliamente utilizado en algoritmos de Minería de Texto, tiene la desventaja que pierde el orden de los términos en la unidad estructural que se analiza, por lo que se han realizado propuestas donde se incluyen elementos semánticos que permiten obtener mayores relaciones entre las unidades.

Por otra parte, en la etapa de pre-procesamiento los términos negadores, intensificadores y atenuadores pueden ser considerados palabras vacías y como consecuencia, eliminarlos descarta la presencia de estos términos, lo cual puede provocar inconsistencias en la etapa de representación y agrupamiento. El manejo de la sinonimia permite reducir la cantidad de términos a representar y a su vez mejorar la calidad de la representación de los mismos.

2

RAZONADOR BASADO EN CASOS PARA EL DESCUBRIMIENTO DE CONOCIMIENTO A PARTIR NUEVOS RECURSOS LINGÜÍSTICOS

2. RAZONADOR BASADO EN CASOS PARA EL DESCUBRIMIENTO DE CONOCIMIENTO A PARTIR NUEVOS RECURSOS LINGÜÍSTICOS

En este capítulo, se presenta un modelo híbrido de descubrimiento de conocimiento para la toma de decisiones diagnósticas a partir de documentos clínicos, combinando técnicas de agrupamiento y clasificación. Debido a que para la construcción de la base de casos en el razonamiento basado en casos resultan necesarias para la transformación de la información textual de la colección de HCE, las tareas de pre-procesamiento y representación esta propuesta integra nuevos recursos lingüísticos que permitan capturar la mayor cantidad de información posible, así como superar las dificultades semánticas, generadas por la sinonimia (Torres López and Arco García, 2016), y como consecuencia mayor precisión en las inferencias realizadas.

2.1 Construcción de la Base de Casos

Tomando como principio que problemas nuevos se resuelven a partir de problemas similares a estos, el razonamiento basado en casos como método de solución de problemas que propone el modelo parte de una colección de HCE, y considera solamente para la construcción de la base de casos los elementos obtenidos por el especialista a la llegada del paciente para emitir un criterio diagnóstico.

Una visión gráfica del esquema del modelo que se implementa en este trabajo se muestra en la Figura 2.1.

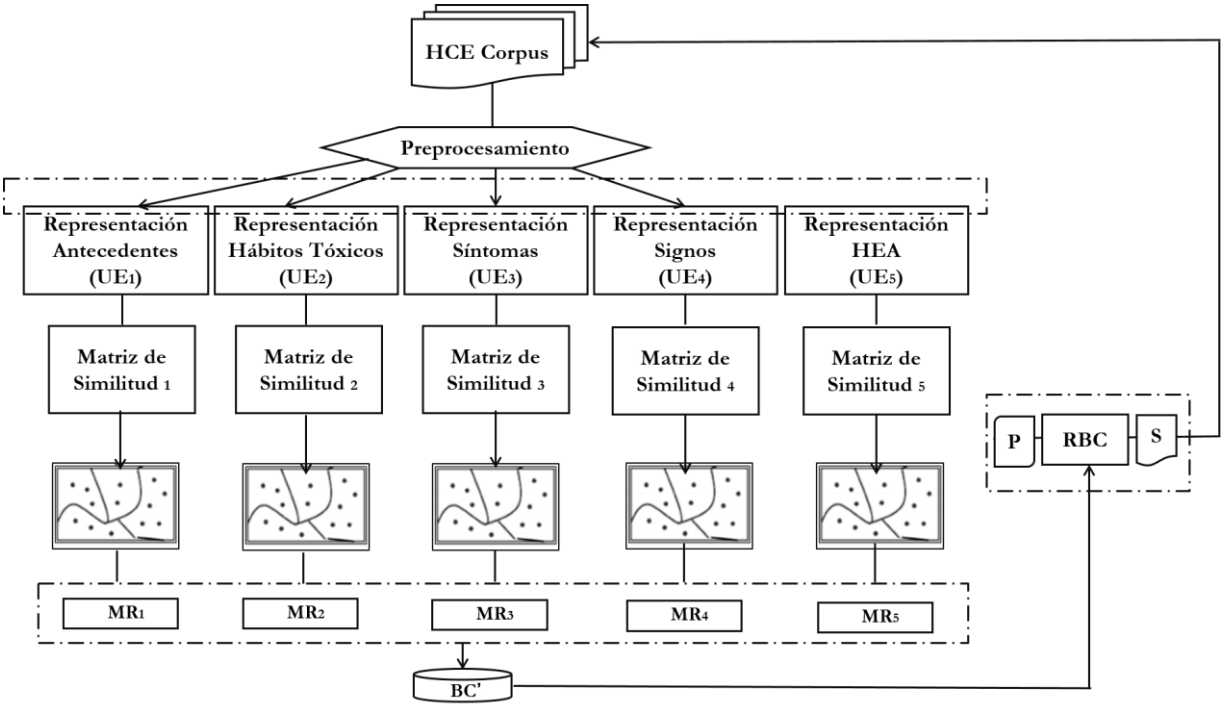

Figura 2.1: Esquema que muestra el modelo propuesto para el agrupamiento de HCE

2.1.1 **Adquisición del conocimiento para la construcción de la Base de casos**

Es conocido por muchos que el diagnóstico es una de las tareas fundamentales de los médicos y la base para una terapéutica eficaz (Ilizástigui 2000). En sí mismo no es un fin sino un medio, e indispensable para establecer el tratamiento adecuado (Díaz Novás, 2008). Algunos lo señalan como la parte más importante del trabajo médico, pero a pesar de eso conlleva muchas dificultades cuando se explica cómo realizarlo (Moreno, 2001, Sáez, 2013).

Varios expertos refieren que en el proceso de búsqueda del diagnóstico los especialistas se auxilian de distintos procedimientos. Pero de ellos, es el método clínico tradicional el más efectivo (Cruz Hernández et al., 2012). Este incluye una serie de pasos ordenados y sucesivos que empiezan con la formulación del problema, (1) la búsqueda de información mediante la anamnesis y el examen físico, (2) la exposición de la hipótesis diagnóstica explicativa basada en la información obtenida del paciente, (3) la cual es contrastada después por la realización de exámenes complementarios o por la evolución del caso. Por último, la comprobación (Moreno, 2001, Sáez, 2013).

El resultado del interrogatorio médico-paciente queda plasmado en la HC, a la llegada de este.

Esta información es almacenada, de forma que cada HCE completada se compone de 11 rasgos o unidades estructurales, en este sentido existe una analogía entre ambos términos debido a que un rasgo es una sección o parte de la HCE que incluye la relatoría realizada por el médico durante las diferentes etapas, dígase: interrogatorio y examen físico a su llegada; tratamiento y evolución durante su ingreso. Puesto que, las HCE a tratar implementan la especificación del estándar CDA-HL7, existe una correspondencia entre cada UE identificada semánticamente y cada sección etiquetada dentro de la estructura jerárquica del XML asociado, dígase *UE={Antecedentes, Hábitos Tóxicos, Síntomas, Signos, Historia de la Enfermedad Actual, Diagnóstico Diferencial, Diagnóstico, Tratamiento, Complementarios, Evolución, Pronóstico}*.

2.1.2 **Base de Casos**

Los niveles de estructuración jerárquica asociados a cada UE identificados en la ingeniería del conocimiento realizada, del total de rasgos que componen la HCE ante la admisión de un nuevo paciente, considera sólo los cinco primeros debido a que son 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 Herrera, 2016) utiliza estos cinco rasgos para la predicción.

La Figura 2.2 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.

<table>
<tr><td colspan="7">HCE</td></tr>
<tr><td>AP</td><td>HT</td><td>ST</td><td>SG</td><td>HEA</td><td>DF</td><td>DN</td></tr>
<tr><td colspan="5">Rasgos Predictores</td><td colspan="2">Rasgos Objetivos</td></tr>
</table>

Figura 2.2: 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 objetivable, 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, dónde 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 Herrera, 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 contienen los casos y la información a utilizar como rasgos predictores en la inferencia, en el sub-epígrafe siguiente se detallan los elementos que se consideran para organizar los casos y facilitar el acceso, ya que el modelo propuesto en (Fuentes Herrera, 2016), parte del hecho que dependiendo de la organización se facilita el acceso y de ahí la recuperación de los casos semejantes.

2.2 Organización de la base de casos

El modelo para la organización de la base de casos propone el uso de una estructura que se concibe a partir de conformación de grupos de documentos semi-estructurados de HCE. Estos se obtienen a partir del agrupamiento de las UE correspondientes a las colecciones independientes que constituyen rasgos predictores, detallado antes. Posteriormente se determina el elemento típico del grupo que queda en un nivel superior. La Figura 2.3 muestra un esquema general de los pasos que se siguen para organizar los casos de la BC.

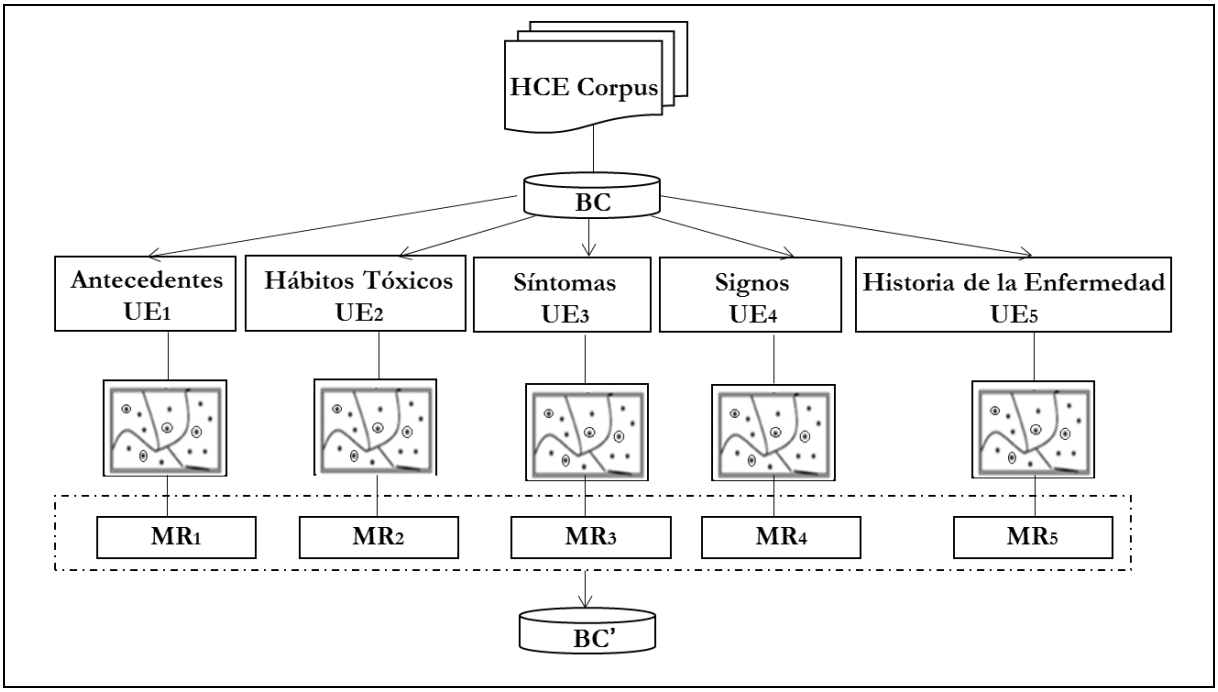

Figura 2.3: Esquema general para la construcción del modelo de la BC.

El modelo de la BC propuesto se resume en el procedimiento de la Figura 2.4.

Para cada grupo de casos se determina el caso "representante(s) del grupo", que se corresponde con el caso que más se parece a los restantes casos del grupo.

Entrada: BC de HCE, UE que constituyen rasgos predictores
Salida: BC' (Particiones: Grupos homogéneos de HCE afines, HCE prototipos de cada grupo y calidad del agrupamiento.)
Inicio
Para cada $\mathbf{DUE_k} \in$ **Rasgos_Predictores**
1. Construir matriz de semejanza utilizando la Similitud_Coseno.
2. Agrupar(β_0).
3. Estimación de la calidad del agrupamiento utilizando mediadas internas
4. Determinar umbral de similitud grupal γ_i
5. Determinar *Prototipo* de cada grupo.

Fin_Para
Fin

Figura 2.4: Procedimiento general para el esquema de construcción del modelo de la BC.

Unido a esta concepción se utiliza una estructura jerárquica, para discriminar los casos de la BC atendiendo al rasgo diagnóstico. Este proceso favorece el acceso y recuperación de los casos más similares al nuevo paciente en la fase adaptación y justificación y crítica que más adelante se detallan. La estrategia presentada permite acceder de forma rápida a las hojas donde se localizan pacientes asociados a una patología determinada

La determinación del umbral de similitud γ_i, se logra siguiendo el criterio γ-semejantes que se describe en (Ruiz-Shulcloper, 1995), basado en el cálculo de las medias de las similitudes, que se aborda en el Anexo 12.

2.3 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.

2.3.1 **Módulo recuperador**

En la etapa de recuperación, se seleccionan el o los casos más similares al nuevo caso. Este proceso se realiza teniendo en cuenta los prototipos de los grupos de documentos obtenidos por el agrupamiento de las colecciones de HCE. La propuesta facilita reducir el espacio de búsqueda

y mejorar el tiempo de respuesta. 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 2.1.

$$\mathrm{SimGlobal}(\mathrm{X},\mathrm{Y}) = \sum_{\mathrm{i}=0}^{\mathrm{m}} W_i * \mathrm{SimCoseno_i}\left(\mathrm{X}_{\mathrm{UE}_i}, Y_{\mathrm{UE}_i}\right)/\mathrm{m} \qquad \textbf{2.1}$$

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

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, aportando en la mayoría de los casos aproximadamente un 75% de la información que se requiere para el diagnóstico, en este sentido la suma del peso de estos rasgos debe ser 0.75, y la suma total de todos 1. Utilizándose las siguientes configuraciones para el cálculo de la importancia de los mismos.

HEA	***SG***	***ST***	***AP***	***HT***
0.25	0.25	0.25	0.12	0.13
0.30	0.25	0.20	0.13	0.12
0.35	0.20	0.20	0.15	0.10

Figura 2.5: Configuraciones posibles para los valores de pesos asociados a los rasgos predictores.

La Figura 2.6, muestra el procedimiento empleado en el proceso de recuperación de los casos más similares al nuevo paciente, basado en el cálculo de la similitud *SimGlobal* (Ver Ecuación 2.1).

Entrada: P; problema a resolver

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

$\forall g_i \in C_{UEk}, i = \overline{1, m_k}$. Sea m_k la cantidad de grupos en el agrupamiento C_{UEk}

$bc_i \in g_i$. Sea bc_i el elemento prototipo en g_i

$\gamma_i \in bc_i$. Sea γ_i el umbral de similitud (calidad grupal) del grupo g_{i_max}, tal que $g_{i\_max} = \max\{SG(bc_i, g_i) \mid \forall g_i, bc_i \in g_i\}$. (Epígrafe 2.2.3.2)

Salida: MS=$\{p_i \mid i = \overline{1, m}\}$, *m* casos recuperados a partir de los más similares.

Inicio

$G = \{bc'_i \mid \text{SimGlobal}(bc'_i, P)\}$. Donde:

$$\text{SimGlobal}(bc'_i, P) = \sum_{j=0}^{n} w_i * \text{SimCoseno}_i\left(bc'_{iUE_j}, P_{UE_j}\right)/n$$

$\forall bc_i \in G$, KS=$\text{máx}_k\{bc'_t \mid \text{SimGlobal}(bc'_t, P)\}$

$$\forall bc_i \in KS,\ Casos = \begin{cases} bc_i & ; \gamma_i < \text{SimGlobal}(bc_i, P) \\ g_{i\_max} & ; e.o.c. \end{cases}$$

$$\forall c_j \in Casos\ \ MS = c_i \quad ; \text{SimGlobal}(c_i, P) \geq \text{SimGlobal}(bc_i, P)$$

Fin

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

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* o *k=7*. El umbral de similitud en el que se basa este algoritmo para la recuperación de los grupos de los *k* casos más similares se obtiene a través del cálculo de la media de la similitud de cada documento en el grupo, como se muestra en la ecuación 2.2.

$$\gamma_i = \frac{\sum_{j=1}^{n} SimCoseno(g_j, G_i)}{n} \qquad \textbf{2.2}$$

2.3.2 **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 por transformación que elimine los diagnósticos o diagnósticos diferenciales ya considerados los cuales se ordenan atendiendo a la ocurrencia de estos en los casos recuperados como más similares.

En la Figura 2.7, se describe como se realiza la adaptación al nuevo problema, se propone una solución inicial a partir de los casos recuperados, que es evaluada y adaptada para que sea lo más factible posible a las condiciones del nuevo paciente. Esto es necesario, debido a que *(…) cada paciente resulta ser un problema nuevo con características y manifestaciones no necesariamente idénticas a otros pacientes tratados (…)*.

Entrada: $MS = \{c_1, c_2, \dots, c_m\}$, m casos obtenidos del módulo recuperador(m>>K).

Salida: DF, DN; Rasgos objetivos.

Inicio

$\forall\, c_i \in MS \quad , LD = \{ld_z = \{DN_i \cup DF_i\}\}.\ \ z=\overline{1,m}$

$$R(i,j) = \begin{cases} 1 & , ld_i \cap ld_j \neq \emptyset \\ 0 & , e.o.c. \end{cases}$$

$D = MAJORITY\ (c_{i\ DN \cup DF}, R)$

Se obtiene MS' donde $MS' \subseteq MS$, casos obtenidos a partir de D

$\forall\, c_i \in MS'$, construir $D_1 = \{d_i | d_i \in DN(ci)\}$.

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

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

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

Fin

Figura 2.7: Algoritmo para conformar la solución inicial.

2.3.3 **Módulo revisor y autoaprendizaje**

La revisión se realiza cuando no se recuperó ningún caso del grupo de ninguno de los casos obtenidos como los más similares, y no se puede discriminar atendiendo a la similitud con la que fueron recuperados estos casos. Se realiza un proceso de selección basado en los 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 cómo se obtiene el DN en el proceso de revisión de la adaptación a partir de los pesos de los rasgos.

Entrada: $MS = \{c_1, c_2, \ldots, c_k\}$, k casos más similares al nuevo problema obtenidos del módulo recuperador.
$W = \{w_i | i = \overline{1,5}\}$, pesos de los rasgos predictores.
Salida: DN; Rasgo objetivo.
Inicio
$Ordenar\_descendente(W)$.
$\forall\ c_i \in MS,\quad \forall\ w_j \in W,\quad DN = DN_{max_{i=1,k}} \left\{ SimCoseno_i \left(c_{i\_UE_j}, P_{UE_j} \right) \right\}$
Fin

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

Para garantizar el autoaprendizaje del sistema, cada caso creado se almacena en un fichero XML-CDA en el repositorio de HCE. El experto revisa los casos creados antes de proceder a la actualización de la BC. Cada cierto período de tiempo se actualiza la BC debido a la complejidad que este proceso conlleva siendo esto una funcionalidad inherente a los expertos.

2.4 Procedimiento general del modelo que soporta CDAIS

En la Figura 2.9 se muestra los módulos principales que conforman el procedimiento general del modelo extendido.

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 y una LSA.

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

3.1 Validación del agrupamiento para cada UE.

3.2 Cálculo de prototipos.

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 de los casos más similares recuperados se revisa y se construye una solución final)

Figura 2.9: Módulos principales del modelo propuesto.

2.4.1 **Módulo 1: Creación de índices a partir del corpus de HCE y pre-procesamiento**

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 el pre-procesamiento del corpus: análisis léxico y eliminación de palabras vacías.

En este trabajo en la etapa de pre-procesamiento se adiciona al análisis el manejo de la negación, intensificación y atenuación, ya que el no considerarlos puede provocar inconsistencias en la etapa de representación y agrupamiento, ya los términos que definen estos recursos al ser palabras vacías son eliminados descartando su presencia. Además se describe el manejo de la sinonimia utilizado con el propósito de permitir reducir la cantidad de términos a representar y a su vez mejorar la calidad de la representación.

2.4.1.1 Tratamiento de la negación

Como se mencionó en el Capítulo 1, propuestas como la de (Vilares et al., 2013) para el manejo de la negación se limita a los términos *no*, *nunca* y *sin*. Este trabajo al igual la propuesta de (Zafra et al., 2015) incluye otras partículas de corte negativo como son: *tampoco*, *nadie*, *jamás*, *ninguno*, *ni* y *nada.*

Este tratamiento se realiza a partir de un proceso que incluye, identificar cada palabra negadora y los términos a los que modifica haciendo uso de las facilidades de *TreeTagger*, herramienta útil para etiquetar textos con información gramatical y léxica. En (Fernández Anta et al., 2013) se emplea una heurística que asume que la negación tiene como alcance los tres elementos a continuación de la palabra negadora. Este trabajo se basa en esta heurística para identificar el sustantivo o el verbo al que modifica la palabra negadora y así poder manejarlo como un término diferente.

Por ejemplo, en las oraciones *"El paciente no tuvo fiebre"* y *"El paciente presentó fiebre alta"* el término *fiebre* a pesar de ser el mismo, está en dos contextos diferentes, en la primera oración aparece negado y en la segunda no. En el caso de la primera oración, al aplicar el procedimiento general para el manejo de la negación se obtendría el término *no_fiebre*, que es diferente al término *fiebre* obtenido en la segunda oración. En la Figura 2.10, se describe el procedimiento general empleado para el manejo de la negación.

Entrada: T; lista de términos que conforman el texto a analizar.
N=$\{n_i | i = \overline{1,k}\}$, k palabras negadoras.
Salida: T´, conjunto de términos que conforman el texto con los términos negados en caso de aparecer palabras negadoras.
Inicio

$$\forall t_i \in T \ , \ T' = \begin{cases} t_i' \ , & si \ t_i \ es \ un \ término \ negano \ por \ n_j \in N \\ t_i \ , & e.o.c. \end{cases}$$

Fin

Figura 2.10: Algoritmo para el manejo de la negación.

2.4.1.2 Tratamiento de palabras modificadoras

Los intensificadores pueden ser adjetivos o adverbios, los adjetivos modifican a los sustantivos y los adverbios de intensificación modifican comúnmente a verbos, adjetivos graduables, y otros adverbios. En este trabajo se manejan solamente intensificadores adjetivales ya que las historias clínicas son documentos con oraciones cortas y poco complejas.

Para ello, se construyen dos listas de términos, una para los intensificadores y otra para los atenuadores, a partir de las consultas manuales a diccionarios y análisis de listas de palabras modificadoras ya existentes.

Para el tratamiento de los modificadores, al igual que para la negación, en (Fernández Anta et al., 2013) se considera que los tres términos a la derecha son los que deben variar su polaridad. Basado en esta idea, para tratar estas palabras modificadoras, se siguió el procedimiento general empleado para el manejo de la negación. Al tratar los términos intensificados como términos diferentes, por ejemplo, en la oración *"el paciente tuvo mucha fiebre",* al aplicarle el procedimiento general se obtendría el término *mucha_fiebre*, así en la etapa de representación

se pueden identificar estos términos y aumentar (en caso de un intensificador) o disminuir (en caso de un atenuador) el valor de la calidad del término en el documento.

La literatura (Vilares et al., 2013) plantea que este aumento y disminución debe hacerse entre un 25% y un 30%. En este trabajo se propone cuando hay presencia de palabras intensificadoras aumentar la calidad del término en un 25% y a su vez restar este mismo porciento de forma equitativa al resto de los términos, pues la suma total de la calidad de los términos del documento es 1. En el caso de la presencia de palabras atenuadoras se hace exactamente lo contrario.

2.4.1.3 Tratamiento de la sinonimia

Para el manejo de la sinonimia se construyó un Diccionario de Sinónimos de Términos Médicos (DSTM) a partir de un manual de la terminología médica (Saldaña, 2012). Este diccionario de sinónimos está compuesto por grupos de sinónimos, cada grupo está formado por un término representante de grupo y varios miembros. Este diccionario tiene la estructura de un documento XML, lo que facilita el indexado y las búsquedas sobre él. En la siguiente Figura, se muestra un ejemplo de cómo estaría formado un grupo.

```
<synonymgroup>
            <prototype>padecer</prototype>
            <synonyms>sufrir</synonyms>
            <synonyms>pasar</synonyms>
            <synonyms>soportar</synonyms>
            <synonyms>sentir</synonyms>
            <synonyms>aguantar</synonyms>
            <synonyms>adolecer</synonyms>
      </synonymgroup>
```

Figura2.11: Sinonimia de términos en el diccionario DSTM.

El procedimiento general empleado para el manejo de la sinonimia a través del uso de este diccionario se describe en la Figura 2.12.

Entrada: T; lista de términos que conforman el texto a analizar.
DSTM; diccionario de sinónimos de términos médicos.
Salida: T´, conjunto de términos que conforman el texto modificando los sinónimos.
Inicio
$\forall t_i \in T$, $r = \text{buscar\_representante}(t_i, \text{DSTM})$.

$$T' = \begin{cases} r \ , & si \ \exists r \\ t_i \ , & e.o.c. \end{cases}$$

Fin

Figura 2.12: Procedimiento general para el manejo de la sinonimia.

2.4.2 **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 información semántica, se han considerado las formas de representación VSM clásica y LSA. De las formas de representación textual existentes 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 las opiniones que emiten los especialistas en la HCE de la relatoría del paciente en su mayoría textos cortos. Debido a que VSM no permite representar relaciones semánticas entre las palabras, se propone experimentar también con LSA que se encuentra disponible en la herramienta *S-Space Package*[4], modelo que también aplica técnicas de reducción de la dimensionalidad y utiliza matrices término-documento para la búsqueda de similitudes semánticas entre los vectores. Los modelos de semántica distribucional[5] fundamentalmente utilizan representaciones de matrices término-término y matrices término-documento (Turney 2010) para representar relaciones entre palabras como por ejemplo la coocurrencia de palabras (Sahlgren 2006). VSM y LSA no representan la coocurrencia de palabras sino la frecuencia de palabras individuales en los documentos, de ahí que pudiera ser útil el uso de otros modelos semánticos que utilicen matrices término-término.

Actualmente existen herramientas que incluyen implementaciones de representaciones espacio vectorial que incorporan elementos semánticos siguiendo los enfoques distribucional y

[4] https://github.com/fozziethebeat/S-Space/wiki/GettingStarted

[5] Se conocen también por modelos espacio-palabra (Word Space Model).

composicional. Para el enfoque distribucional existen dos tipos de herramientas, aquellas orientadas al conteo de vectores y las que se basan en la predicción de contextos. Las primeras se apoyan en cuatro etapas: (1) realizar una representación del texto para extraer la cantidad de coocurrencias; utilizar un esquema de pesos para estimar dichas cantidades;(2) reducir la dimensionalidad de las representaciones y comparar las unidades textuales a través de medidas de similitud (Fuentes Herrera, 2016). Las segundas representan vectores de aprendizaje de palabras usando redes neuronales; por lo que desafortunadamente requieren de un corpus de entrenamiento, de ahí que no se proponen para esta etapa. Por tanto, se propone aplicar los modelos distribucionales semánticos basados en vectores de conteo.

2.4.3 **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 calculada.

$$S_{coseno(o_i,o_j)} = \frac{\sum_{k=1}^{m}(o_{ik} * o_{jk})}{\sqrt{\sum_{k=1}^{m} o_{ik}^2 * \sum_{k=1}^{m} o_{jk}^2}} \qquad \textbf{2. 3}$$

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). De manera que en la organización de la base de casos cada grupo serán pacientes cuyos rasgos coinciden con un cierto nivel de precisión. En este contexto, se pueden emplear fundamentalmente tres tipos de algoritmos de agrupamiento, los algoritmos jerárquicos aglomerativos (Huang et al., 2013) (Cselle et al., 2007), los partitivos (Seo and Sycara, 2004) y los probabilísticos como Expectation-Maximization (Lu et al., 2013).

En este caso se emplearon los algoritmos jerárquicos aglomerativos ya que según la literatura ofrecen los mejores resultados (Day and Edelsbrunner, 1984), particularmente se utilizó un algoritmo de agrupamiento de enlace completo que, aunque es sensible a objetos que están lejanos, este criterio favorece la obtención de grupos compactos con diámetros pequeños, que es lo ideal en este tipo de problemas.

Luego al aplicar un algoritmo HAC es necesario cortar la jerarquía en nivel para obtener una partición. El enfoque que se empleó 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, si es igual o mayor se detiene el agrupamiento. Se seleccionó la medida coseno para hallar la matriz de similitud en el algoritmo de agrupamiento (Ver Anexo 3).

Para realizar el agrupamiento se empleó la herramienta *S-Space Package* que 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 (Rohde et al., 2006).
- Modelos basados en aproximación: aproximan datos de coocurrencia para lograr mejor escalabilidad de grandes conjuntos de datos. Por ejemplo: Random Indexing y RRI (Cohen et al., 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[6]. Algunas de las medidas de similitud que posee son la medida coseno, Euclidiana, Jaccard y KL divergence (Jurgens and Stevens, 2010).

2.4.3.1 Validación de grupos

Para la validación del agrupamiento se emplearon medidas de validación internas y externas. Se utilizó la medida externa *Overall F-Measure* (Ver Anexo 4) para medir la precisión con la que se obtuvieron los grupos, basado en las configuraciones utilizadas para la selección del umbral de corte del agrupamiento HAC a partir de las medidas de calidad interna del agrupamiento. Las medidas de validación interna empleadas son *Overall Similarity*, *Índice Dunn* y *Average Similarity* que se abordan en el Anexo 5.

2.4.3.2 Cálculo de prototipos

Después de haber realizado el agrupamiento, para cada grupo de casos obtenido, se determina el caso "representante del grupo" o "prototipo", que se corresponde con el caso que más se parece a los restantes documentos del grupo. Para ello se calcula la similitud de cada documento en el grupo y se toma como prototipo el que ofrezca una mayor similitud. Para hallar la similitud entre los documentos se emplea la similitud coseno. La ecuación 2.4 muestra cómo obtener la similitud de un documento o_k en el grupo G, donde *m* representa la cantidad de documentos del grupo.

$$SG(o_k, G) = \frac{\sum_{\substack{i=1 \\ i \neq k}}^{m} SimCoseno\,(o_k, o_i)}{m-1} \qquad \textbf{2. 4}$$

2.4.4 **Módulo 4: Razonamiento Basado en Casos**

En esta etapa se construye el RBC, que está formado por un recuperador, un adaptador y un revisor (Epígrafe 2.3). El recuperador compara el problema insertado con los que están almacenados en la BC y recupera los casos más similares al nuevo problema, así como otros casos del grupo donde cada uno de ellos fue más similar y dichos casos tuvieron una similitud igual o mayor que los casos más similares recuperados de la BC. El adaptador parte de los casos

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

recuperados y como resultado infiere una respuesta con el mayor grado de semejanza a la que se busca. Por otro lado, el revisor se encarga de inferir una respuesta cuando los casos recuperados para realizar la adaptación no son suficientes y no se puede discriminar atendiendo a la similitud con la que fueron recuperados estos casos, para ello se realiza un proceso de selección basado en los rasgos predictores y la importancia (pesos) de estos.

2.5 Consideraciones finales del capítulo

El modelo propuesto para el agrupamiento de HCE basadas en el estándar CDA, facilita eficazmente el agrupamiento, lo que contribuye a la gestión de información y conocimiento en esta área.

El manejo de los recursos lingüísticos contribuye al aumento del grado de similitud existente entre las HCE, utilizando las unidades estructurales que caracterizan el proceso lógico evolutivo en el diagnóstico y tratamiento del paciente durante su ingreso.

Las representaciones VSM y LSA permiten la recuperación de información y el procesamiento de las HCE a través de la construcción de vectores de términos, que a su vez son empleados por el HAC para realizar el agrupamiento.

El agrupamiento facilita la recuperación y el acceso a los casos, al reducir el espacio de búsqueda explorado para obtener los casos más semejantes a un problema nuevo con una complejidad $O(kn)$, en la práctica $k \ll n$.

El RBC permite, dado un nuevo problema, recuperar los casos más similares a él e inferir una respuesta con el mayor grado de semejanza posible.

3

EVALUACIÓN DEL MODELO PROPUESTO Y DESCRIPCIÓN DEL SISTEMA CDAIS

3. EVALUACIÓN DEL MODELO PROPUESTO Y DESCRIPCIÓN DEL SISTEMA CDAIS

En el presente capítulo se realiza una evaluación de la efectividad del razonamiento basado en casos realizado, 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 *CDAIS* 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 modelo que soporta CDAIS

La evaluación del modelo tiene como propósito comprobar la validez de los resultados que obtiene el modelo de RBC propuesto en este trabajo a partir de la comprobación de las respuestas obtenidas por éste ante la presencia de un nuevo problema.

Para evaluar el modelo propuesto, se utilizó como caso de estudio 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.

A partir de estas colecciones de HCE se conformaron 15 corpus de prueba, donde en cada corpus se encuentran entre 180 y 200 historias clínicas de al menos dos clases diferentes, dichas clases fueron definidas por los expertos.

3.1.1 **Conformación de grupos con el manejo de los recursos lingüísticos.**

A continuación se ilustra un ejemplo de cómo varía la conformación de los grupos en un corpus de 20 HCE teniendo en cuenta el manejo o no de los recursos lingüísticos, lo cual influye en la conformación de la BC y con ello mejora la calidad de la recuperación de los casos más similares.

Sin manejo de recursos lingüísticos	Con manejo de recursos lingüísticos
[1, 2, 18, 11, 13, 14, 15]	[0, 4, 6, 9, 12]
[0, 4, 6, 9, 12]	**[3]**
[3, 5, 7]	[16, 8, 10]
[16, 8, 10]	[2, 18]
[17, 19]	[17, 19]
	[5, 7]
	[1, 11, 13, 14, 15]

Figura 3.1: Conformación de grupos teniendo en cuenta el manejo o no de los recursos lingüísticos en la UE predictora Hábitos Tóxicos

Como se observa en la tabla anterior, la conformación de grupos varió, por ejemplo, en el caso del grupo formado por los documentos [3, 5, 7], al tener en cuenta los recursos lingüísticos se dividió en dos grupos: [3] y [5, 7], lo cual fue correcto, ya que como se puede ver en la Figura 3.2 el contenido que incluyen estas HCE en la UE Hábitos Tóxicos reflejan información diferente.

Documento 3	**Documento 5**	**Documento 7**
no fumador	fumador	fumador
no alcohol	consumidor de café	consumidor de alcohol diario
si cafeína		

Figura 3.2: Contenido que incluido en la UE Hábitos Tóxicos.

3.1.2 **Selección del modelo de representación**

Como se detalló en el Capítulo 2, una de las etapas por las cuales transita el modelo para la organización de la base de casos, requiere de la representación textual para el cálculo de la similaridad entre las *k*-colecciones de HCE para la obtención del agrupamiento asociado a cada UE predictora. A pesar de los buenos resultados en las etapas pre-agrupamiento obtenidos por el modelo VSM en este trabajo se incluye el modelo de análisis semántico latente LSA, debido a que se sobrepone a los asuntos de sinonimia para el agrupamiento (Aggarwal and Zhai, 2012) (Torres López and Arco García, 2016) y reduce la dimensionalidad de documentos al proyectar

los vectores BOW en un espacio semántico construido a partir de la matriz término-documento de SVD.

Por lo cual, con el propósito de valorar las bondades de uno u otro modelo y su influencia en la obtención de grupos, se diseña un experimento para estimar la calidad del agrupamiento obtenido a partir de una u otra forma de representación.

Este experimento consiste en calcular la precisión del agrupamiento realizado por cada UE predictora, utilizando uno u otro modelo de representación para ver si existen diferencias significativas entre ellos.

Atendiendo a la clasificación de las medidas para la evaluación del agrupamiento de (Höppner et al., 1999, Silberschatz and Tuzhilin, 1996, Kaufman and Rousseeuw, 1990), en este trabajo se seleccionó la medida externa: *Overall F-measure* (OFM) para medir la precisión del agrupamiento. Esta medida se basa en un parámetro $\propto$ que mide la precisión y el cubrimiento de los grupos obtenidos, en este caso se utiliza $\propto = 1$ ya que en este tipo de problemas solo interesa medir la precisión. Estos resultados se pueden ver en el Anexo 6

Como se observa en la tabla A6.1, los valores de precisión para los agrupamientos obtenidos por ambos modelos coinciden en la mayoría de las UE, solo difieren para las UE Hábitos Tóxicos y Antecedentes Patológicos, como se puede ver en las Figuras 3.3 y 3.4 respectivamente. Vale señalar que, aunque los valores de esta tabla para estas UE conducen a pensar que no existen diferencias significativas entre las precisiones obtenidas para ambas UE, para corroborarlo se aplicó el test de Wilcoxon. Los resultados de este test se muestran en la Figura 3.5.

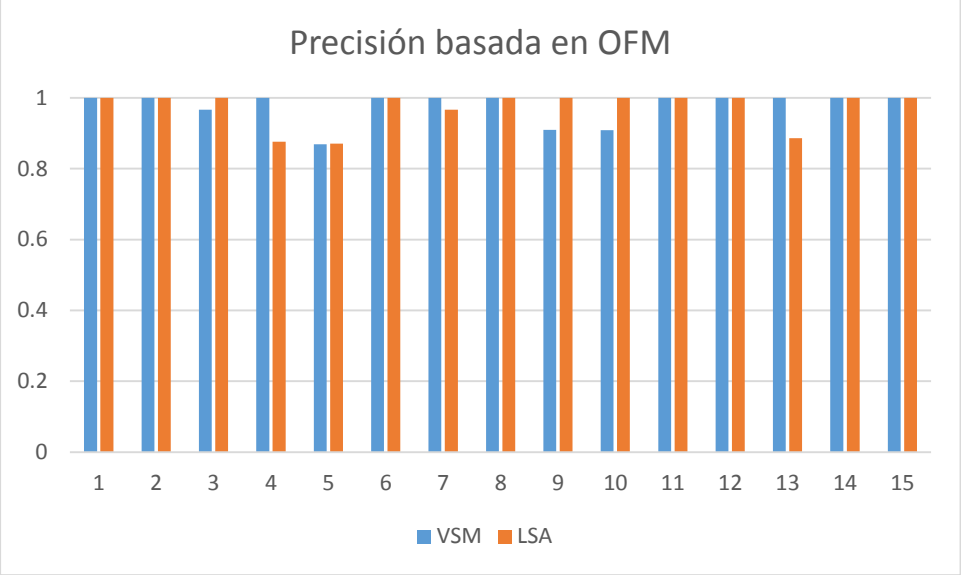

Figura 3.3: Precisión basada en OFM usando los modelos de representación VSM y LSA para los agrupamientos correspondiente a la UE Hábitos Tóxicos.

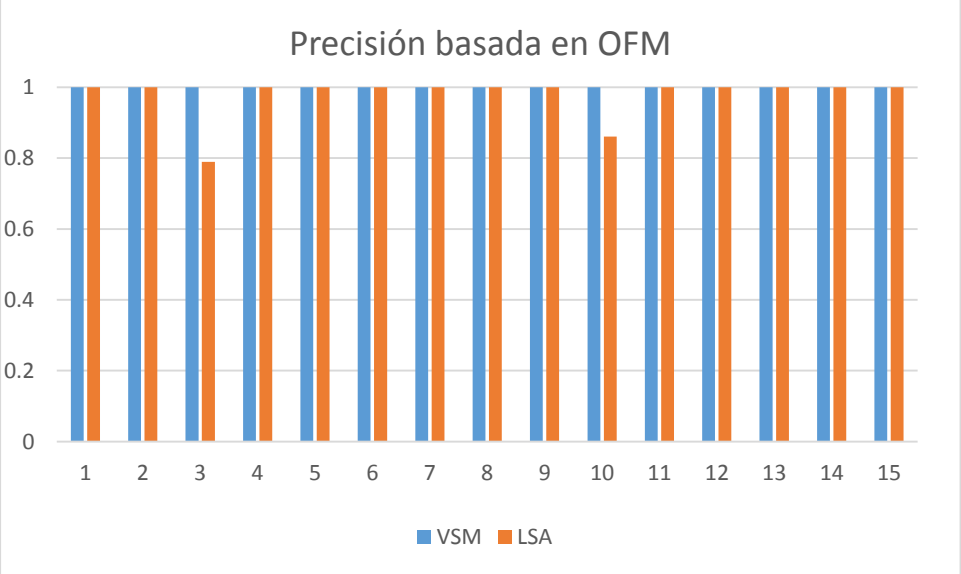

Figura 3.4: Precisión basada en OFM usando los modelos de representación VSM y LSA para los agrupamientos correspondiente a la UE Antecedentes Patológicos.

Rangos					**Estadísticos de contraste[a]**	
		N	Rango promedio	Suma de rangos	Z	Sig. asintót. (bilateral)
HT_LSA - HT_VSM	Rangos negativos	3[a]	5.33	16.00	-.338[b]	.735
	Rangos positivos	4[b]	3.00	12.00		

	Empates	8[c]				
	Total	15				
AP_LSA - AP_VSM	Rangos negativos	2[d]	1.50	3.00	-1.342[b]	.180
	Rangos positivos	0[e]	.00	.00		
	Empates	13[f]				
	Total	15				

a. HT_LSA < HT_VSM
b. HT_LSA > HT_VSM
c. HT_LSA = HT_VSM
d. AP_LSA < AP_VSM
e. AP_LSA > AP_VSM
f. AP_LSA = AP_VSM

a. Prueba de los rangos con signo de Wilcoxon
b. Basado en los rangos positivos.

Figura 3.5: Resultados del test de Wilcoxon para verificar si existen diferencias significativas en la precisión basada en OFM usando los modelos de representación VSM y LSA para los agrupamientos correspondiente a las UE Hábitos Tóxicos y Antecedentes Patológicos.

Los resultados de la tabla anterior reflejan valores de significación superiores a 0.05. Esto indica que no existen diferencias significativas en cuanto a la calidad del agrupamiento utilizando una u otra forma de representación ya que, por lo cual se decide emplear VSM debido a que LSA tiene la limitación que dada la naturaleza de alta dimensionalidad de los datos textuales, el cálculo de SVD puede ser costoso y las dimensiones resultantes pueden ser difíciles de interpretar debido a que cada dimensión es una combinación lineal de un conjunto de palabras a partir del espacio original (Berry and Kogan, 2010). Además en las tablas anteriores se observa que para la UE Antecedentes Patológicos se obtiene para todos los casos el máximo valor de precisión. Por otra parte, para la UE Hábitos Tóxicos se obtienen empates en la mayoría de los casos.

3.1.3 **Evaluación del algoritmo de agrupamiento**

Como se detalló en el Capítulo 2, una de las etapas por las cuales transita el modelo para la organización de la base de casos, requiere de la obtención del agrupamiento para cada *k*-colección asociada a cada UE predictora, tarea que en este trabajo se realiza utilizando un algoritmo jerárquico aglomerativo de enlace completo que, favorece la obtención de grupos compactos con diámetros pequeños, lo cual resulta ideal en el problema que nos ocupa, además de no requerir a priori la especificación de la cantidad de grupos a obtener.

Por otra parte, debido a que se desea obtener una partición del conjunto de HCE en grupos con alta calidad. Luego al aplicar un algoritmo HAC es necesario cortar la jerarquía en nivel para obtener una partición, atendiendo a un nivel pre-especificado de similitud entre objetos en un mismo grupo, lo cual conlleva a aplicar un umbral que permita agrupar comparando las medidas de similitud de los grupos con dicho umbral β_0; es decir, los grupos serán agrupados hasta que la mayor similitud de un grupo sea menor que el umbral especificado.

Se diseñó un experimento para medir la precisión del RBC para un valor de *k=3* a partir de una validación cruzada (LOOCE) (Cawley, 2006), tal que en cada iteración se considera un caso de prueba y el resto de entrenamiento. El caso de prueba se toma como un nuevo problema a resolver y los de entrenamiento conforman la base de casos que emplea el razonador para recuperar los casos más similares e inferir una respuesta con el mayor grado de semejanza a la que se busca.

Para ello, se utilizaron los umbrales de similitud $\beta_0 = 0.3$, $\beta_0 = 0.4$ y $\beta_0 = 0.5$ debido a que este agrupamiento calcula la similitud de dos grupos como la similitud de sus miembros más disimilares y como consecuencia es deseable que la partición obtenga grupos compactos. La Figura 3.6 muestra la precisión que alcanza el RBC para valores fijos del umbral.

Umbral	**Porciento de clasificaciones correctas**
0.3	98.75
0.4	99.07407
0.5	99.76959

Figura 3.6 Precisión que alcanza el RBC para valores fijos del umbral.

A partir de los resultados de la tabla anterior se selecciona para este problema el valor del umbral 0.5 debido a que este reporta los mejores resultados.

3.1.4 **Evaluación del RBC**

En el proceso de validación del sistema propuesto se diseñó un experimento que consistió en emplear la Base de Casos obtenida luego de realizar el pre-procesamiento, representación, agrupamiento y cálculo de prototipos, al caso de estudio descrito en el epígrafe 3.1, y realizar una validación cruzada. El procedimiento empleado se describe en la Figura 3.7.

Figura 3.7: Procedimiento empleado para realizar la validación cruzada del RBC.

Esta validación mostró una precisión del 99.76959% para un valor de *k=3* en la recuperación.

Entrada: BC; base de casos
Salida: precisión del RBC expresada en porciento
Inicio

$\forall \mathrm{bc_i} \in \mathrm{BC}$, $\mathrm{caso_{prueba}} = \mathrm{bc_i}$

$\mathrm{casos_{entrenamiento}} = \{BC \backslash \mathrm{bc_i}\}$

$mas\_similares = Recuperar\_ksimilares;$

$\mathrm{dn_i} = Adaptar(mas_{similares})$

$\mathrm{Certeza_i} = \begin{cases} 1 & ; \mathrm{dn_i} \in DN_{cp} \\ 0 & ; e.o.c. \end{cases}$;

$\mathrm{Presición} = \mathrm{calcular\_porciento(Certeza)}$;

Fin

3.1.5 **Selección del valor de k para la recuperación**

Como se describe en el Epígrafe 2.3.1, la recuperación depende de un parámetro *k*, que expresa la cantidad de casos a recuperar de la Base de Casos, basado en el cálculo de la similitud global que se muestra en la ecuación 2.2.

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* o *k=7*, por lo que se experimentó con estos valores. Para ello se calculó la precisión del RBC usando cada uno de estos valores de *k*. Los resultados se muestran en la Figura 3.8.

Valor de k	Porciento de clasificaciones correctas
k=1	100
k=3	99.76959
k=5	99.53917
k=7	98.61751

Figura 3.8: Precisión del RBC para los distintos valores de *k*.

A pesar de que se obtiene una mejor precisión para *k=1*, en este trabajo se propone utilizar en la recuperación el valor *k=3*, debido a que también reporta buenos resultados y 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, para *k=1* el conjunto de casos recuperados para la adaptación es muy reducido y pudieran obviarse estos otros casos.

3.1.6 **Cálculo de los pesos de los rasgos predictores**

Como se detalló en el Capítulo 2, 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, como se observa en la ecuación 2.2.

El cálculo de esta similitud global depende de los pesos de los rasgos predictores. Para el cálculo de los pesos de estos rasgos se consultó a los expertos, los cuales concluyeron que los rasgos Historia de la Enfermedad Actual, Signos y Síntomas son los más importantes, aportando en la mayoría de los casos aproximadamente un 75% de la información que se requiere para el diagnóstico, en este sentido la suma del peso de estos rasgos debe ser 0.75, y la suma total de todos 1, por tanto se experimentó con las siguientes configuraciones para el cálculo de la importancia de los mismos.

HEA	**SG**	**ST**	**AP**	**HT**
0.25	0.25	0.25	0.12	0.13
0.30	0.25	0.20	0.13	0.12
0.35	0.20	0.20	0.15	0.10

Figura 3.9: Configuraciones posibles para los valores de pesos asociados a los rasgos predictores.

Para realizar este experimento se midió la precisión del RBC en cada una de las configuraciones anteriores, los resultados se muestran en la Figura 3.10.

	Pesos					
Experimentos	**HEA**	**SG**	**ST**	**AP**	**HT**	**Porciento de clasificaciones correctas**
1	0.25	0.25	0.25	0.12	0.13	99.76958525
2	0.3	0.25	0.2	0.13	0.12	99.76958525
3	0.35	0.2	0.2	0.15	0.1	99.76958525

Figura 3.10: Precisión del RBC para las distintas configuraciones de los pesos de los rasgos predictores.

Como se puede ver en la tabla anterior, para las distintas configuraciones de los pesos de los rasgos predictores el porciento de validación del RBC es el mismo por tanto son válidas cualquiera de las configuraciones anteriores para el pesado de los rasgos, en este caso se utilizó la configuración correspondiente al experimento 1: HEA=0.25, SG=0.25, ST=0.25, AP=0.12 y HT=0.13.

3.2 Diseño e implementación del Sistema CDAIS

El diseño del sistema CDAIS se dividió en tres capas fundamentales como se muestra en la Figura 3.11. La primera capa o inferior es la capa del dominio, la segunda o intermedia es la capa controladora y la tercera o superior es la capa de interfaz de usuario.

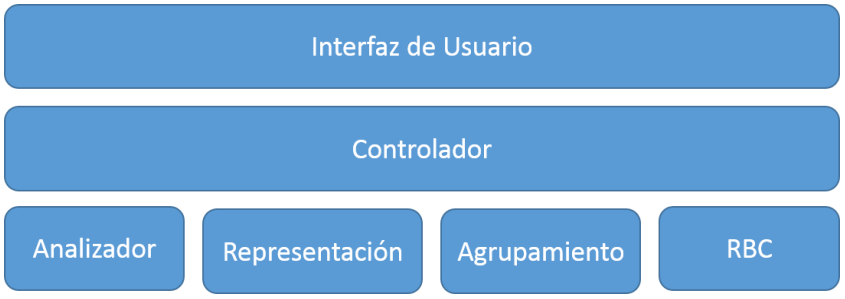

Figura 3.11: Diseño general del sistema CDAIS

En la capa inferior están las clases del dominio las cuales conforman los distintos módulos del modelo. En el módulo Analizador se encuentran las clases que permiten la recuperación de la información y el pre-procesamiento de la misma, así como el manejo de los recursos lingüísticos que se abordan es este trabajo. En el módulo de Representación se encuentran las clases que permiten la representación del corpus textual. El módulo de Agrupamiento está formado por las clases que permiten realizar el agrupamiento jerárquico, el cálculo de los umbrales de similitud y de los prototipos de grupos. El módulo de RBC lo forman 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.

Por otra parte, El sistema CDAIS cuenta con una interfaz de usuario que soporta el modelo implementado en este trabajo, con el objetivo de evidenciar cómo se pueden desarrollar otras herramientas que incorporen el modelo. Ésta no es más que la tercera capa, la cual contiene todas las clases relacionadas con las formas visuales y la interacción con el usuario.

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 modelo-vista controlador (model-view-controller MVC) el cual tiene por objeto 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 3.12 se muestra la estructura estática de una aplicación Yii.

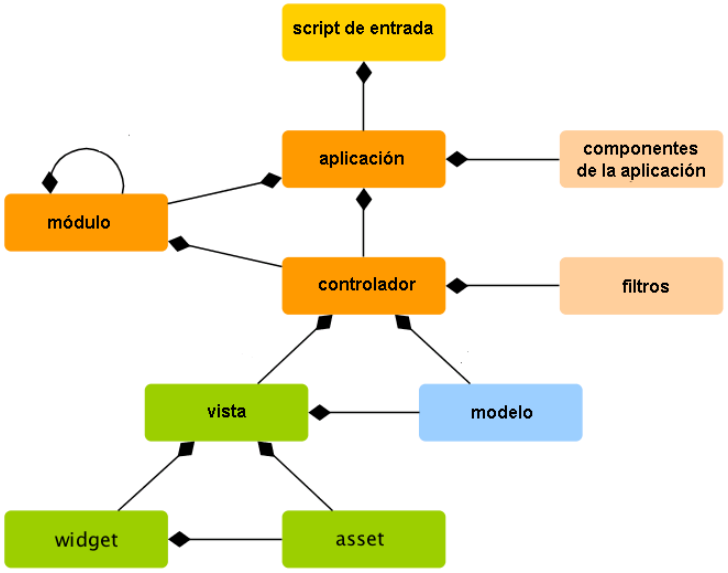

Figura 3.12: Estructura estática de una aplicación Yii.

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.

En el desarrollo de la interfaz de usuario se emplearon recursos de la programación web como el trabajo con formularios, imágenes, enlaces, menús, etc., los cuales facilitan el funcionamiento de sitios web y la interacción con los usuarios.

La capa intermedia es la que empaqueta todas las clases controladoras y es la encargada de establecer la comunicación entre las clases de las dos capas mencionadas. Yii permite ejecutar aplicaciones implementadas en java, por lo que se usó ésta facilidad del framework para incorporar el modelo propuesto en este trabajo.

3.3 Interfaz de usuarios del sistema CDAIS para asistir la toma de decisiones diagnóstica

La interfaz de usuario cuenta con un menú donde se pueden acceder a las funcionalidades que éste permite. La página de inicio de este sitio (ver Figura 3.13) permite el acceso a varios sitos de interés de la medicina cubana y latinoamericana entre los que se encuentran Infomed, la Comercializadora de Servicios Médicos Cubanos S.A, el Formulario Nacional de Medicamentos, la Revista Cubana de Medicina, entre otros, con el fin de potenciar el conocimiento y el desarrollo en este sector.

Figura 3.13: Fragmento de la página de inicio de la interfaz web.

La aplicación permite gestionar la información referente a las historias clínicas que se encuentran ubicadas en un repositorio: ver un resumen, editar y eliminar HCE. La Figura 3.14 muestra la página que permite acceder a estas operaciones.

Figura 3.14: Fragmento de la página para la gestión de la información de las HCE.

Otra de las funcionalidades de la aplicación es la del ingreso de nuevos pacientes, para ello crea una nueva HCE y la incorpora al repositorio. La Figura 3.15 muestra un fragmento de la página que permite hacer dicho ingreso.

Cuando se está haciendo el ingreso del nuevo paciente se pueden obtener las hipótesis diagnósticas para asistir al usuario en el completamiento de los campos Diagnóstico y Diagnóstico Diferencial. La Figura 3.16 muestra un fragmento de la página de ingreso de un nuevo paciente donde se pueden completar la información de las UE predictoras y obtener las hipótesis diagnósticas.

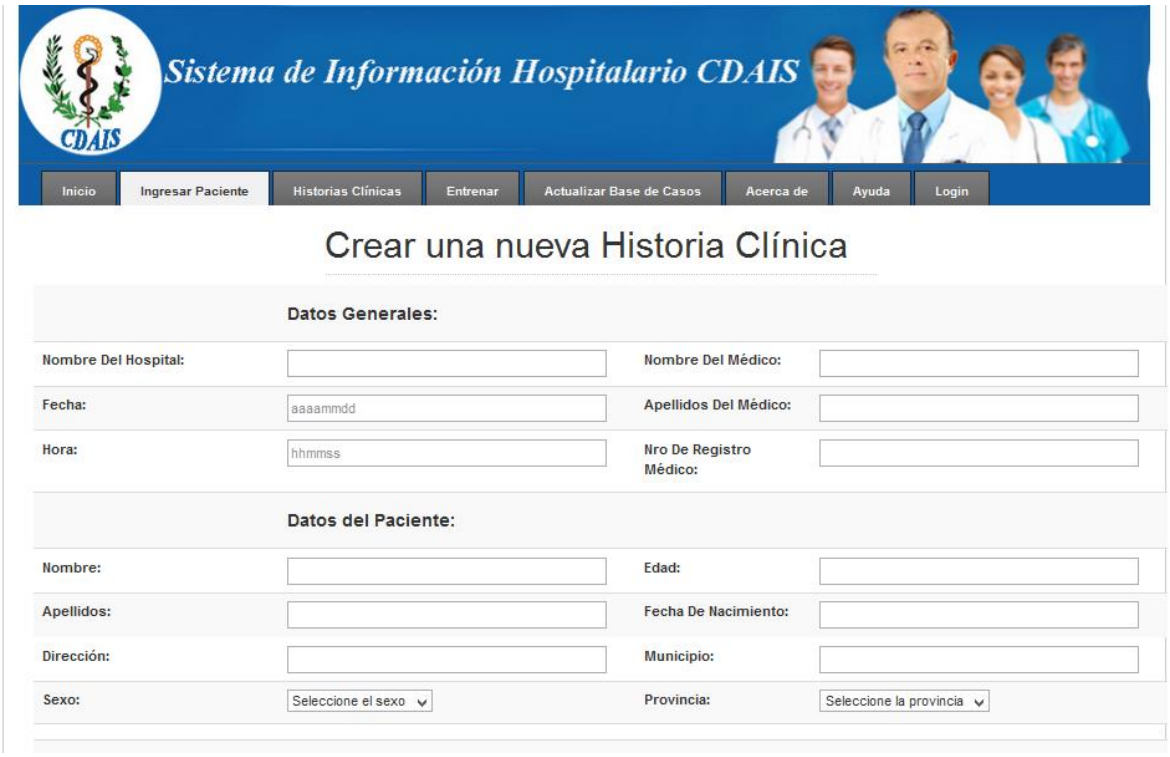

Figura 3.15: Fragmento de la página para el ingreso de un nuevo paciente.

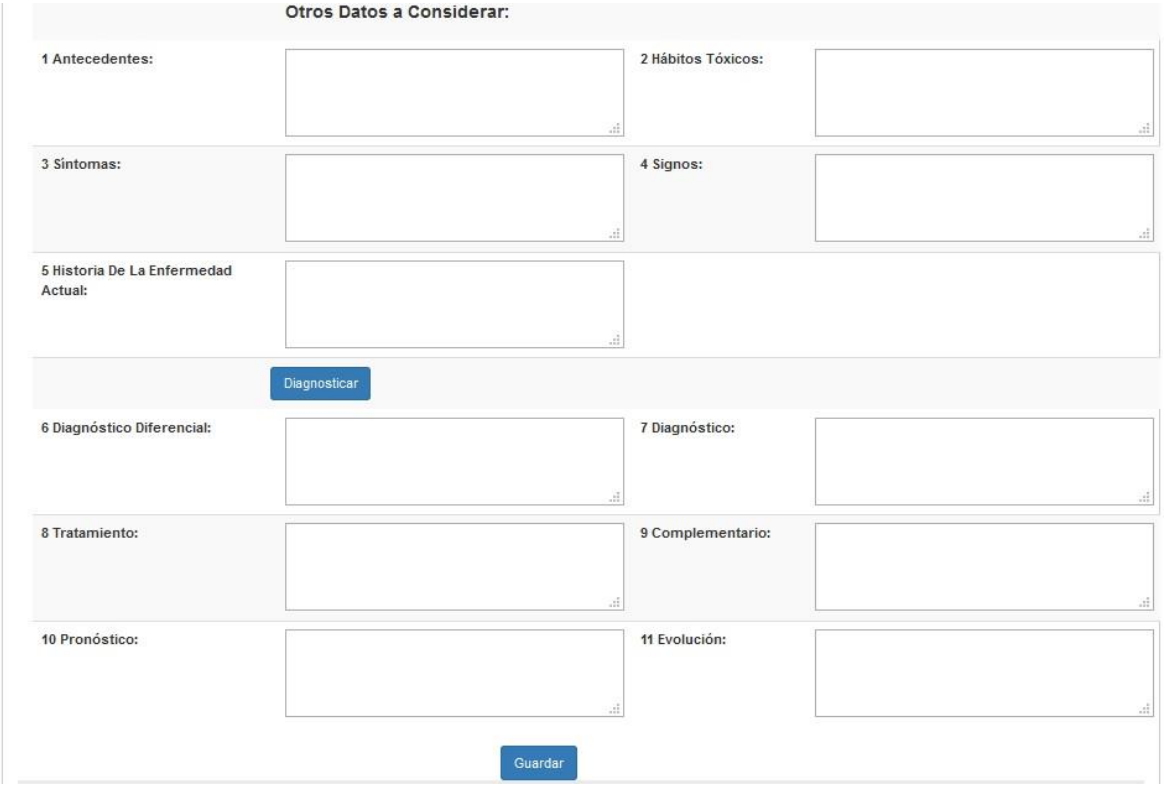

Figura 3.16: Fragmento de la página para el ingreso de un nuevo paciente donde se pueden obtener las hipótesis diagnósticas.

La aplicación cuanta además con otra página que permite obtener hipótesis diagnósticas de nuevos casos a partir de los principales rasgos predictores a modo de entrenamiento sin necesidad de ingresar estos casos en el sistema. La Figura 3.17 muestra la página donde se puede realizar dicho entrenamiento.

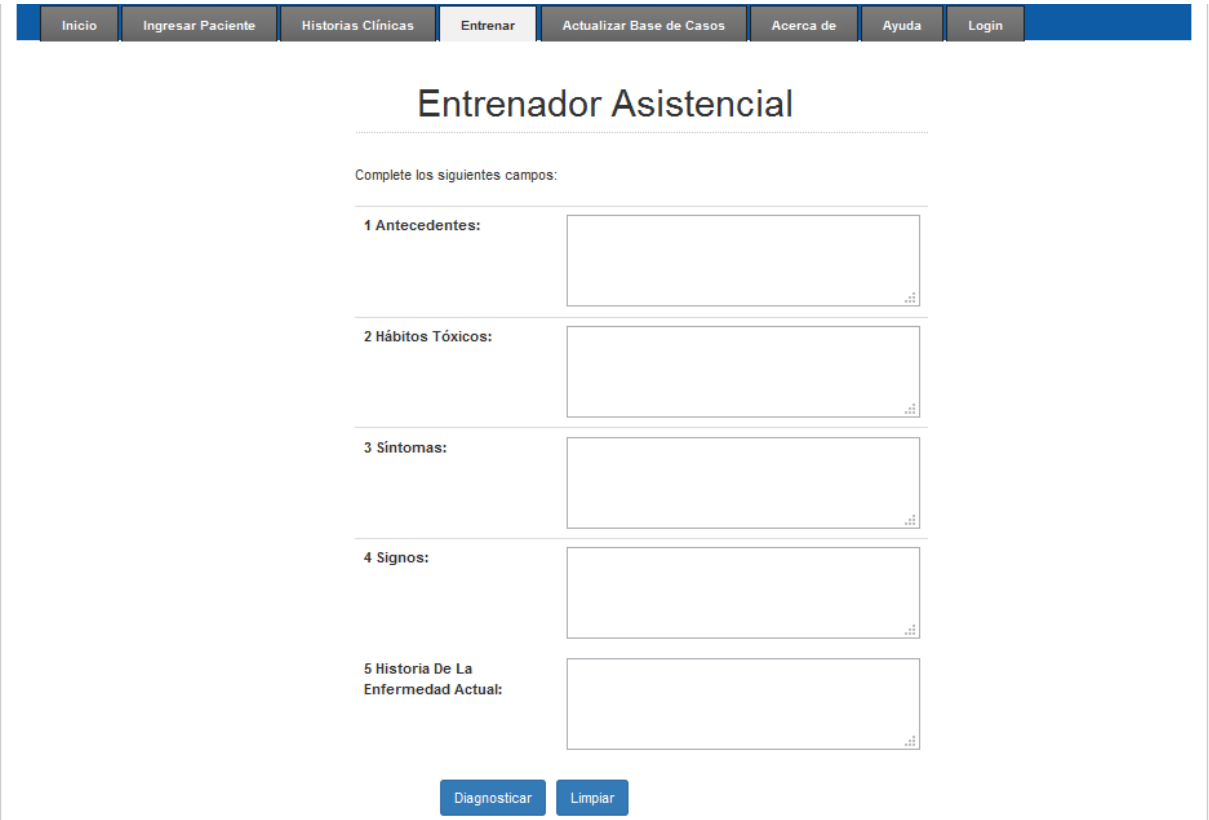

Figura 3.17: Página para realizar un entrenamiento asistencial.

Otra de las funciones que brinda el sitio web es la de actualizar la BC para no perder información relevante de las nuevas HCE creadas en la inferencia. La Figura 3.18 muestra la página donde se puede realizar dicha actualización.

Figura 3.18: Página para actualizar la BC.

Además, la aplicación cuenta con una ayuda para explicar a los usuarios como interactuar con el sitio web, una vista donde se puede obtener información acerca del sistema como se muestra en la Figura 3.19, y una vista para autenticarse como uno de los usuarios de prueba creados como se muestra en la Figura 3.20.

Figura 3.19: Página con información acerca del sistema.

Figura 3.20: Página para la autenticación de usuarios.

3.4 Consideraciones finales del capítulo

Los resultados experimentales permiten concluir que la implementación del enfoque basado en casos propuesto concibe la organización de la base de casos a partir de agrupamientos que utiliza como modelo de representación VSM ya que obtiene resultados similares al modelo de espacio semántico LSA con menor costo computacional.

Los experimentos realizados demuestran que el modelo propuesto logra resultados de precisión del RBC elevados y el nivel de granularidad deseado para un valor de $k=3$.

El RBC alcanza un comportamiento similar para las distintas configuraciones de los pesos sugeridos por los expertos.

El modelo puede aplicarse a múltiples contextos, brinda la facilidad de inferir áreas que pueden ser fuente de investigaciones médicas.

CONCLUSIONES

Como resultado del presente trabajo se desarrolló un sistema basado en casos para la deducción de hipótesis diagnósticas a la llegada de un nuevo paciente incorporando el tratamiento de **recursos lingüísticos** que se incluyen en la relatoría que realiza el especialista de las manifestaciones patológicas que refiere y presenta el paciente.

- Para la representación se implementan en el RBC los modelos VSM y LSA. Los experimentos desarrollados a partir de los casos de estudio evidencian que no existen diferencias significativas en cuanto a la calidad del agrupamiento utilizando una u otra forma de representación, manteniéndose la representación VSM debido a que LSA tiene la limitación de alta dimensionalidad, además de ser costosa computacionalmente.
- El RBC implementado incluye el tratamiento de los recursos lingüísticos negación, modificadores y sinonimia, implementando las herramientas que el contexto demanda, sin embargo, fue necesario incluir un diccionario de sinónimos en español teniendo en cuenta las deficiencias de la terminología médica de los existentes.
- La organización de la base de casos a partir del agrupamiento jerárquico favorece el acceso y la recuperación de los casos más similares a un nuevo problema a partir del análisis de la similitud de los prototipos, lo cual reduce sustancialmente el espacio de búsqueda.
- Ante la llegada de un nuevo paciente el RBC propone el completamiento de los rasgos que involucran el diagnóstico de hipótesis tempranas a partir de la recuperación eficaz de los casos recuperados que incluyen el manejo de los recursos lingüísticos, con una adecuada revisión de los casos completados cuando estos no son suficientes y no se puede discriminar atendiendo a la similitud con la que fueron recuperados.
- El sistema CDAIS es una aplicación web que permite al usuario la obtención de hipótesis diagnósticas tempranas a través de una interfaz amigable. La evaluación del sistema a partir de colecciones de HCE del Servicio de Admisión del Hospital Provincial “Celestino Hernández Robau” permite ajustar los diferentes parámetros del sistema. Evidenciando que no existen diferencias significativas entre los pesos dados al sistema

a partir de los criterios dados por los expertos. El mejor valor de k se alcanza para $k=3$, el valor del parámetro umbral para el agrupamiento HAC resulta 0.5.

RECOMENDACIONES

- Analizar otras variantes para construir el diccionario de sinónimos de términos médicos.
- Estudiar otras formas de agrupamiento para determinar su efecto en la conformación de las particiones deseadas.
- Determinar otras formas para obtener los prototipos empleados en la organización de la base de casos.

REFERENCIAS BIBLIOGRÁFICAS

ABELLEIRA, M. A. P. & CARDOSO, C. A. 2010. Minería de texto para la categorización automática de documentos. *PhD in Computer Science por Carnegie Mellon University, Madrid, España*.

AGGARWAL, C. C. & ZHAI, C. 2012. *Mining text data*, Springer Science & Business Media.

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

AMORES, M., ARCO, L. & BARRERA, A. Efectos de la Negación, Modificadores, Jergas, Abreviaturas y Emoticonos en el Análisis de Sentimiento.

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

ARCO, L. 2009. *Agrupamiento basado en la intermediación diferencial y su valoración utilizando la teoría de los conjuntos aproximados.* Doctorado en Ciencias Técnicas, Universidad Central "Marta Abreu" de Las Villas.

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

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

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

BERRY, M. W. 2004. *Survey of Text mining: Clustering, Classification, and Retrieval,* New York, USA, Springer Verlag.

BERRY, M. W. & KOGAN, J. 2010. *Text mining: applications and theory*, John Wiley & Sons.

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

CABARCAS, A., PUELLO, P. & MARTELO, R. J. 2015. Sistema de Información Soportado en Recuperación XML para Pequeñas y Medianas Empresas (PYME) de Cartagena de Indias, Colombia. *Información tecnológica,* 26**,** 135-144.

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.

COHEN, T., SCHVANEVELDT, R. & WIDDOWS, D. 2010. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of biomedical informatics,* 43**,** 240-256.

CRUZ HERNÁNDEZ, J., HERNÁNDEZ GARCÍA, P., ABRAHAM MARCEL, E., DUEÑAS GOBEL, N. & SALVATO DUEÑAS, A. 2012. Importancia del método clínico. *Revista Cubana de Salud Pública,* 38**,** 422-437.

CSELLE, G., ALBRECHT, K. & WATTENHOFER, R. BuzzTrack: topic detection and tracking in email. Proceedings of the 12th international conference on Intelligent user interfaces, 2007. ACM, 190-197.

DALAMAGAS, T., CHENG, T., WINKEL, K.-J. & SELLIS, T. 2006. A Methodology for Clustering XML Documents by Structure. *Information Systems*.

DAY, W. H. & EDELSBRUNNER, H. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification,* 1**,** 7-24.

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. & HARSHMAN, R. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science,* 41**,** 391.

DEL RÍO CRESPO, É. A., CARVAJAL, P. A. G. & GARCÉS, S. A. C. 2015. con la implementación del uso de soluciones estandarizadas.

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.

DÍAZ NOVÁS, J. 2008. El tratamiento médico: experiencia, base teórica y método. *Revista Habanera de Ciencias Médicas,* 7**,** 0-0.

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.

EYSENBACH, G. 2008. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *Journal of medical Internet research,* 10.

FERNÁNDEZ ANTA, A., NÚÑEZ CHIROQUE, L., MORERE, P. & SANTOS MÉNDEZ, A. 2013. Sentiment analysis and topic detection of Spanish tweets: A comparative study of NLP techniques.

FERNÁNDEZ LANZA, S. & SOBRINO CERDEIRIÑA, A. 2000. Hacia un tratamiento computacional de la sinonimia. *Procesamiento del lenguaje natural, nº 26 (septiembre 2000); pp. 89-96.*

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

FUENTES HERRERA, I. E. 2016. *Descubrimiento de conocimiento en entornos hospitalarios a partir de registros médicos para la toma de decisiones.*

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.

FUKUHARA, T., TAKEDA, H. & NISHIDA, T. Multiple-text summarization for collective knowledge formation. Proceedings of the IEEE International Conference om Systems, Man and Cybernetics, 1999 Tokyo. IEEE Press.

GARCÍA-HERNÁNDEZ, B. 1997. Sinonimia y diferencia de significado. *Revista española de lingüística,* 27**,** 1-31.

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

GUERRINI, G., MESITI, M. & SANZ, I. 2006. An Overview of Similarity Measures for Clustering XML Documents.

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

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

HERNANDEZ, V. S. 2009. *Tecnologias de la informacion y el conocimiento* [Online]. GestioPolis.

HOFMANN, T. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. Advances in neural information processing systems, 2000. 914-920.

HUANG, X., ZHANG, X., YE, Y., DENG, S. & LI, X. 2013. A topic detection approach through hierarchical clustering on concept graph. *Applied Mathematics & Information Sciences,* 7**,** 2285.

ILIZÁSTIGUI , D. F. 2000. El método clínico: muerte y resurrección. *Educación Médica Superior,* 14**,** 109-127.

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.

LESNIEWSKA, A. Clustering XML documents by structure. ADBIS'09 Proceedings of the 13th East European conference on Advances in Databases and Information Systems, 2009. 238-246

LU, Y., ZHANG, P., LIU, J., LI, J. & DENG, S. 2013. Health-related hot topic detection in online communities using text clustering. *Plos one,* 8**,** e56221.

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.

MANNING, C. D., RAGHAVAN, P. & SCHÜTZE, H. 2008. *Introduction to information retrieval*, Cambridge university press Cambridge.

MARTÍNEZ, J. M. G. La sinonimia. Problema metalingüístico. Anales de filología hispánica, 1988.

MENÁRGUEZ, T. M. 2013. Modelos de representación de arquetipos en sistemas de información sanitarios. *Proyecto de investigación:*.

MESA, R. Y. 2006. De la gestión de información a la gestión del conocimiento. *Acimed,* 14**,** 0-0.

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.

MORENO, R. M. 2001. *El arte y la ciencia del diagnóstico médico,* La Habana.

NIU, Z.-Y., JI, D.-H. & TAN, C.-L. Document clustering based on cluster validation. *In:* EVANS, D. A., ed. Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM 2004), 2004 Washington, D.C., USA. ACM Press, 501-506.

NÜRNBERGER, A., KLOSE, A. & KRUSE, R. Clustering of document collection to support interactive text exploration. Proceedings of the 25th Annuals Conference of the Gesellschaft für Klassification. Studies in Classification, Data Analysis and Knowledge Organization. Exploratory Data Analysis in Empirical Research, 2001 Germany. 291-299.

OCHOA, A. & ARCO, L. 2008. Differential betweenness in complex networks clustering. *Progress in Pattern Recognition, Image Analysis and Applications***,** 227-234.

PASCUAL, D., PLA, F. & SÁNCHEZ, S. 2007. Algoritmos de agrupamiento. *Método Informáticos Avanzados*.

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.

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

RIJSBERGUEN, C. J. 1979. *Information Retrieval,* London: Butterworths.

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

RODRÍGUEZ, S. H. 2006. La relación médico-paciente. *Revista Cubana de Salud Pública,* 32**,** 0-0.

ROHDE, D. L., GONNERMAN, L. M. & PLAUT, D. C. 2006. An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM,* 8**,** 627-633.

ROSELL, M., KANN, V. & LITTON, J. E. Comparing comparisons: document clustering evaluation using two manual classifications. Proceedings of International Conference on Natural Language processings ICON, 2004 Hyderabad, India.

RUIZ-SHULCLOPER, J. 1995. *Introducción al reconocimiento de patrones. Enfoque lógico combinatorio,* México, CINVESTAV IPN.

SÁEZ, L. E. M. 2013. Propuesta metodológica para la enseñanza del método clínico. *Edumecentro,* 2**,** 32-38.

SAHAMI, M. 1998. *Using machine learning to improve informatio access.* PhD. Thesis, Stanford University.

SALDAÑA, E. 2012. *Manual de Terminología Médica*.

SALTON, G. & BUCKLEY, C. 1988. Term weighting approaches in automatic text retrieval. *Information Processing and Management,* 24**,** 513-523.

SALTON, G. & MCGILL, M. 1983. *Introduction to Modern Information Retrieval,* New York, USA, McGraw-Hill.

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.

SEIJO, C., LUNA, F., GUADIX, J. M. H., SEIJO, J. F. C., LUNA, J. M. F. & GUADIX, J. F. H. 2011. *Recuperación de información. Un enfoque práctico y multidisciplinar,* Ra-Ma.

SEO, Y.-W. & SYCARA, K. 2004. Text clustering for topic detection.

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

SLAVOV, V., RAO, P., PATURI, S., SWAMI, T. K., BARNES, M., RAO, D. & PALVAI, R. 2013. A new tool for sharing and querying of clinical documents modeled using HL7 Version 3 standard. *Computer methods and programs in biomedicine,* 112**,** 529-552.

STEINBACH, M., KARYPIS, G. & KUMAR, V. A comparison of document clustering techniques. Proceedings of 6th ACM SIGKDD World Text Mining Conference, 2000a Boston. ACM Press, 1-20.

STEINBACH, M., KARYPIS, G. & KUMAR, V. A comparison of document clustering techniques. Proceedings of 6th ACM SIGKDD World Text Mining Conference, 2000b Boston. ACM Press.

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.

TABOADA, M., BROOKE, J., TOFILOSKI, M., VOLL, K. & STEDE, M. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics,* 37**,** 267-307.

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.

TEKLI, J., CHBEIR, R., TRAINA, A. & TRAINA, C. 2011. XML document-grammar comparison: related problems and applications. *Open Computer Science,* 1**,** 117-136.

TORRES LÓPEZ, C. & ARCO GARCÍA, L. 2016. Representación textual en espacios vectoriales semánticos. *Revista Cubana de Ciencias Informáticas,* 10**,** 148-180.

TREINS, M., CURE, O. & SALZANO, G. On the interest of using HL7 CDA release 2 for the exchange of annotated medical documents. Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on, 2006. IEEE, 524-532.

VALLEZ, M. & PEDRAZA, R. 2007. El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines. *Hipertext. net*.

VILARES, D., ALONSO, M. A. & GÓMEZ-RODRÍGUEZ, C. 2013. Clasificación de polaridad en textos con opiniones en espanol mediante análisis sintáctico de dependencias. *Procesamiento del lenguaje natural,* 50**,** 13-20.

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.

YANG, H.-X., WANG, W.-X. & WANG, B.-H. 2008. Asymmetric negotiation in structured language games. *Physical Review E,* 77**,** 027103.

YANG, Y. & PEDERSEN, J. O. A comparative study on feature selection in text categorization. *In:* FISHER, D. H., ed. Proceedings of the Fourteenth International Conference on Machine Learning, 1997 San Francisco, US. Morgan Kaufmann Publishers, 412-420.

ZAFRA, J., SALUD, M., MARTÍNEZ CÁMARA, E., VALDIVIA, M., TERESA, M. & MOLINA GONZÁLEZ, M. D. 2015. Tratamiento de la Negación en el Análisis de Opiniones en Espanol.

ZAPICO, M. & VIVAS, J. 2014. La sinonimia como caso particular de distancia semántica. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação,* 19**,** 253-266.

ZIPF, G. K. 1949. *Human Behaviour and the Principle of Least Effort,* Addison-Wesley.

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. Algunas medidas de calidad de términos

Umbral de frecuencia de términos y Ley de Zpif. Eliminar los términos que tienen o muy alta o muy baja frecuencia de aparición (Sahami, 1998), a partir de un cálculo adecuado del umbral y del estudio de la Ley de Zipf (Zipf, 1949). Términos que raramente aparecen en una colección de documentos tendrán poco poder discriminante y pueden ser eliminados (Rijsberguen, 1979). En contraste, términos con frecuencia de aparición alta se asumen que son comunes y que tampoco tienen poder discriminante[7].

Umbral de frecuencia de documentos. Teniendo en cuenta que $n(t)$ es el número de documentos en los cuales el término t aparece al menos una vez, una heurística simple de selección es excluir todos los términos desde el vocabulario cuya frecuencia de documentos es menor que algún umbral, ya que términos que ocurren en sólo muy pocos documentos improbablemente llevan información que permita distinguir los grupos textuales y tienden a ser ruidosos (Yang and Pedersen, 1997). Además, usar la ocurrencia de términos infrecuentes no es confiable estadísticamente. Al eliminar estos términos se mantiene el poder discriminante y se mejora la efectividad del agrupamiento y clasificación textual.

Frecuencia inversa de documento y TFIDF. La importancia de los términos se asume inversamente proporcional al número de documentos en los cuales el término particular aparece. Después de eliminar las palabras de parada, la importancia de un término se incrementa con su frecuencia de uso. Combinando estas ideas se formuló la medida frecuencia del término / frecuencia inversa de documentos (*tfidf*).

$$tfidf(t) = tf(t) \cdot idf(t), \text{ donde } idf(t) = \log \frac{n}{n(t)} \qquad \text{(A1.1)}$$

Una combinación similar de frecuencia de términos y frecuencia inversa de documentos es se utiliza usualmente para asignar pesos a los términos (Salton and Buckley, 1988).

[7] Términos con alta frecuencia de aparición pueden formar parte de la lista de palabras de parada automáticamente construida desde la colección de documentos. En esta tesis se considera que la lista es suministrada.

Razón de señal a ruido. Medir el poder discriminante que transmite cada término, basado en el ruido R(*t*), como la entropía de la distribución de la probabilidad del término *t* entre los documentos (Salton and McGill, 1983):

$$\text{SNR}(t)=\log tf(t) - \text{R}(t),\ \text{R}(t) = -\sum_{j=1}^{n} \text{P}(d_j,t)\log \text{P}(d_j,t) \text{ y } \text{P}(d_j,t) = \frac{tf_{d_j}(t)}{tf(t)} \qquad \text{(A1.2)}$$

Entropía. Calcular la entropía como una medida de importancia, según Lochbaum y Streeter en 1989 (Nürnberger et al., 2001):

$$\text{Entropía}(t) = 1 + \frac{1}{\ln(n)} \sum_{i=1}^{n} p_i(t) \cdot \ln(p_i(t)) \quad \text{donde } p_i(t) = \frac{tf_{d_i}(t)}{\sum_{j=1}^{n} tf_{d_j}(t)} \qquad \text{(A1.3)}$$

Calidad de términos. Medir la calidad de los términos según las expresiones q_0 y q_1, la segunda constituye una variante de la primera donde n_1 es el número de documentos en los cuales *t* ocurre al menos una vez (Berry, 2004).

$$q_0(t) = \sum_{j=1}^{n} \left(tf_{d\,j}(t)\right)^2 - \frac{1}{n}\left[\sum_{j=1}^{n} tf_{d\,j}(t)\right]^2 \text{ y } q_1(t) = \sum_{j=1}^{n_1} \left(tf_{d\,j}(t)\right)^2 - \frac{1}{n_1}\left[\sum_{j=1}^{n_1} tf_{d\,j}(t)\right]^2 \qquad \text{(A1.4)}$$

Skewness y Kurtosis. Calcular la parcialidad de los términos mediante la combinación de Skewness y Kurtosis según $\text{P}(t)=w_1 \cdot \text{Skewness}(t)+w_2 \cdot \text{Kurtosis}(t)$, donde w_1 y w_2 son pesos positivos y *s* es la desviación estándar de la ocurrencia del término *t* en la colección de documentos (Fukuhara et al., 1999).

$$\text{Skewness}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(tf_{d_i}(t) - \frac{tf(t)}{n}\right)^3}{s^3} \text{ y } \text{Kurtosis}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(tf_{d_i}(t) - \frac{tf(t)}{n}\right)^4}{s^4} - 3 \qquad \text{(A1.5)}$$

Anexo 2. 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.*

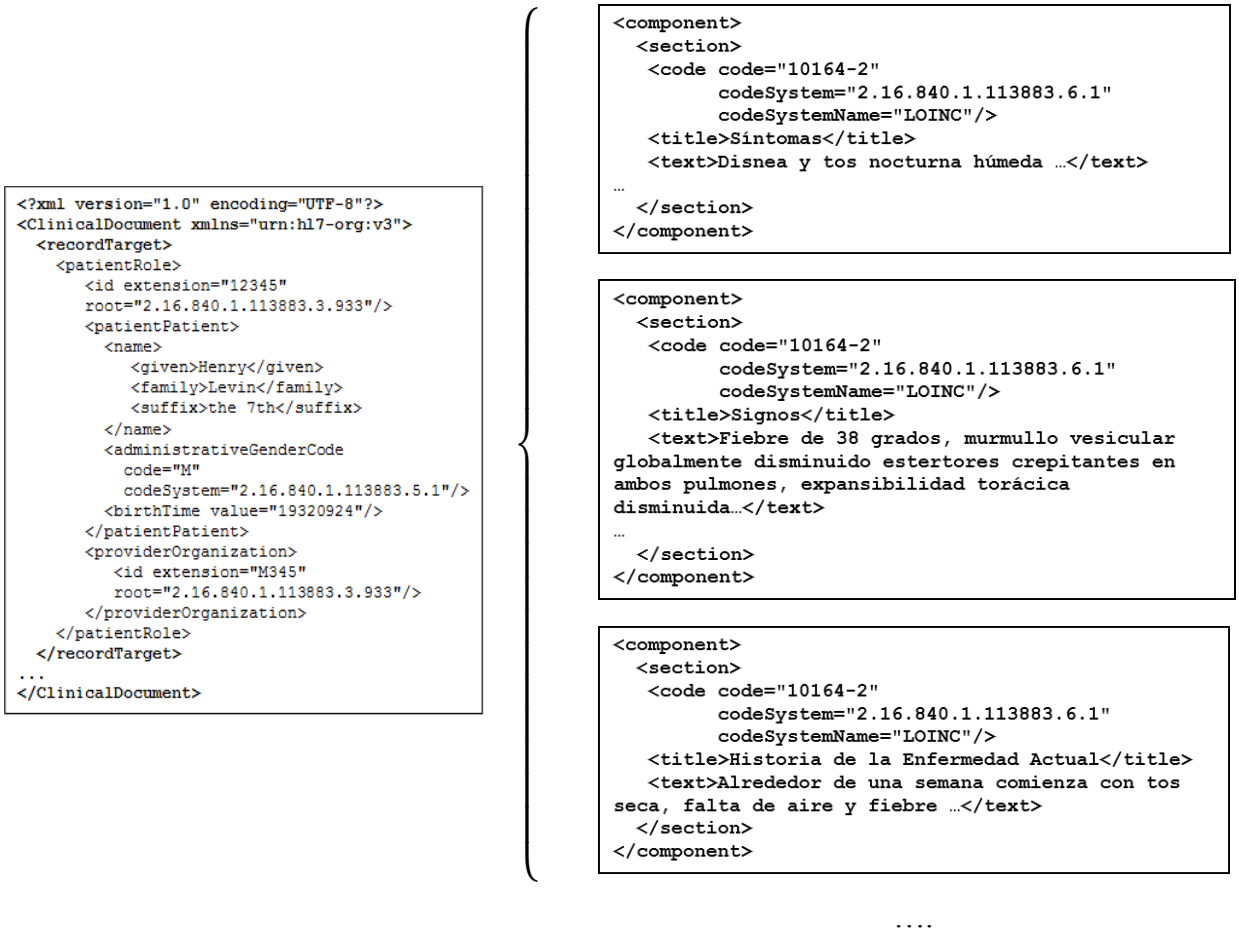

Figura A2.1 Identificar colecciones independientes, utilizando un ejemplo hipotético de HCE.

Anexo 3. 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

$$D_{Euclideana}(O_i, O_j) = \sqrt{\sum_{k=1}^{m}(o_{ik} - o_{jk})^2} \qquad \text{(A3.1)}$$

Distancia Minkowski (Batchelor, 1978)

$$D_{Minkowski}(O_i, O_j) = \left(\sum_{k=1}^{m}|o_{ik} - o_{jk}|^{\gamma}\right)^{\frac{1}{\gamma}} \quad \text{donde } \gamma \geq 1 \qquad \text{(A3.2)}$$

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

Distancia Euclidiana heterogénea (HeterogenousEuclidean – OverlapMetric; HEOM)

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

$$d_{local}(o_{ik}, o_{jk}) = \begin{cases} d_{Overlap}(o_{ik}, o_{jk}) & si\ k\ simbólico \\ d_{NormEuclidean}(o_{ik}, o_{jk}) & si\ k\ numérico \end{cases} \qquad \text{(A3.3)}$$

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

Distancia Camberra (Michalski et al., 1981)

$$D_{Camberra}(O_i, O_j) = \sum_{k=1}^{m} \frac{|o_{ik} - o_{jk}|}{|o_{ik} + o_{jk}|} \qquad \text{(A3.4)}$$

Correlación de Pearson(Wilson and Martínez, 1997)

$$D_{Pearson}(O_i, O_j) = \frac{\sum_{k=1}^{m} \left(o_{ik} - \overline{atributo_k}\right)\left(o_{jk} - \overline{atributo_k}\right)}{\sqrt{\sum_{k=1}^{m} \left(o_{ik} - \overline{atributo_k}\right)^2 \sum_{k=1}^{m} \left(o_{jk} - \overline{atributo_k}\right)^2}} \quad \text{(A3.5)}$$

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

$$S_{Dice}(O_i, O_j) = \frac{2\sum_{k=1}^{m} \left(o_{ik} \cdot o_{jk}\right)}{\sum_{k=1}^{m} o_{ik}^2 + \sum_{k=1}^{m} o_{jk}^2} \quad \text{(A3.6)}$$

Coeficiente de Jaccard

$$S_{Jaccad}(O_i, O_j) = \frac{\sum_{k=1}^{m} \left(o_{ik} \cdot o_{jk}\right)}{\sum_{k=1}^{m} o_{ik}^2 + \sum_{k=1}^{m} o_{jk}^2 - \sum_{k=1}^{m} \left(o_{ik} \cdot o_{jk}\right)} \quad \text{(A3.7)}$$

Coeficiente Coseno

$$S_{Coseno}(O_i, O_j) = \frac{\sum_{k=1}^{m} \left(o_{ik} \cdot o_{jk}\right)}{\sqrt{\sum_{k=1}^{m} o_{ik}^2 \cdot \sum_{k=1}^{m} o_{jk}^2}} \quad \text{(A3.8)}$$

Anexo 4. Algunas medidas externas para la validación del agrupamiento

Medida-*F* Global (Overall *F*-Measure; OFM)(Steinbach et al., 2000b)

$$Overall\,F-Measure=\sum_{i=1}^{k}\frac{n_i}{n}\max\{F-Measure(i,j)\} \qquad \text{(A4.1)}$$

donde k es el número de clases, n_i es el tamaño de la clase i, n es el número total de objetos agrupados y $F\text{-}Measure(i, j)$ se calcula según la expresión siguiente. Si $\alpha = 1$, entonces OFM se nombra Purity(Rosell et al., 2004).

Medida-*F*(*F*-Measure) de la clase *i* respecto al grupo *j*

$$F-Measure(i,j)=\frac{1}{\alpha(1/\Pr(i,j))+(1-\alpha)(1/\mathrm{Re}(i,j))} \qquad \text{(A4.2)}$$

Si $\alpha = 1$ entonces $F\text{-}Measure(i,j)$coincide con precision, si $\alpha = 0$ entonces $F\text{-}Measure(i,j)$coincide con cubrimiento. $\alpha = 0.5$ significa igual peso para precisión y cubrimiento.

Micro-averaged precision y micro-averaged recall (Niu et al., 2004)

$$\text{MA - Pr}=\frac{\sum_{i=1}^{k}\alpha_i}{\sum_{i=1}^{k}(\alpha_i+\beta_i)} \quad \text{y} \quad \text{MA - Re}=\frac{\sum_{i=1}^{k}\alpha_i}{\sum_{i=1}^{k}(\alpha_i+\gamma_i)} \qquad \text{(A4.3)}$$

donde α_i es el número de objetos correctamente asignados a la clase i, β_i es el número de objetos incorrectamente asignados a la clase i y γ_i es el número de objetos incorrectamente no asignados a la clase i. MA-Pr = MA-Re si cada objeto pertenece a sólo un grupo y la clasificación de referencia también tiene una clasificación única para cada objeto.

Medidas propuestas por INEX

$$Purity(k)=\frac{NDMLC_k}{NDC_k} \qquad \text{(A4.4)}$$

$$Micro - Purity(k) = \frac{\sum_{k=0}^{n} Purity(k) * TotalFoundByClass(k)}{\sum_{k=0}^{n} TotalFoundByClass(k)} \quad \text{(A4.5)}$$

$$Macro - Purity(k) = \frac{\sum_{k=0}^{n} Purity(k)}{TotalofCategories} \quad \text{(A4.6)}$$

Donde se asume el total de categorías como la cantidad de grupos encontrados.

Anexo 5. Algunas medidas internas para la validación del agrupamiento

Similitud global (Overall Similarity)(Steinbach et al., 2000a)

$$OverallSimilarity(Grupo) = \frac{1}{|Grupo|^2} \sum_{O_i, O_j \in Grupo} \text{distancia}(O_i, O_j) \tag{A5.1}$$

Índices Dunn

$$I(C) = \frac{\min_{i \neq j}\{\delta(C_i, C_j)\}}{\max_{1 \leq l \leq k}\{\Delta(C_l)\}} \tag{A5.2}$$

donde $C=\{C_1, \ldots, C_k\}$ es el agrupamiento de un conjunto de objetos O, $\delta: C \times C \rightarrow R$ es una medida de distancia de grupo a grupo y $\triangle: C \rightarrow R$ es una medida de diámetro del grupo.

$$\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \text{ y } \Delta(C_i) = \max_{x, y \in C_i} d(x, y) \tag{A5.3}$$

donde $d: C \times C \rightarrow R$ es una función que mide la distancia entre los objetos de O.

Una de las propuestas de Bezdek para el cálculo de $\delta(C_i, C_j)$ **y** $\Delta(C_i)$

$$\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} d(x, y) \text{ y } \Delta(C_i) = 2\left(\frac{\sum_{x \in C_i} d(x, c_i)}{|C_i|}\right) \tag{A5.4}$$

donde c_i es el centro del grupo C_i.

Medidas propuestas por INEX

$$Purity(k) = \frac{NDMLC_k}{NDC_k} \tag{A5.5}$$

$$Micro - Purity(k) = \frac{\sum_{k=0}^{n} Purity(k) * TotalFoundByClass(k)}{\sum_{k=0}^{n} TotalFoundByClass(k)} \tag{A5.6}$$

$$Macro - Purity(k) = \frac{\sum_{k=0}^{n} Purity(k)}{TotalofCategories} \tag{A5.7}$$

Donde se asume el total de categorías como la cantidad de grupos encontrados.

Anexo 6. Calidad del agrupamiento basado en precisión a partir de los modelos VSM y LSA

	VSM-OFM($\propto = 1$)					LSA-OFM($\propto = 1$)				
CORPUS	**ST**	**SG**	**HEA**	**HT**	**AP**	**ST**	**SG**	**HEA**	**HT**	**AP**
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	0.966834	1	1	1	1	1	0.789385
4	1	1	1	1	1	1	1	1	0.87619	1
5	1	1	1	0.868421	1	1	1	1	0.87069	1
6	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	0.966667	1
8	1	1	1	1	1	1	1	1	1	1
9	1	1	1	0.909836	1	1	1	1	1	1
10	1	1	1	0.908197	1	1	1	1	1	0.86089
11	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	0.88587	1
14	1	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1	1	1

Tabla A6.1: Valores de precisión obtenidos con OFM usando los modelos de representación VSM y LSA para cada agrupamiento realizado por UE predictora.

Anexo 7. Calidad del agrupamiento de la colección asociada a la UE Síntomas para diferentes configuraciones del umbral de similitud.

Corpus	*Overall Similarity*	*Índice Dunn*	*Average Similarity*	*Overall F-Measure*
1	0.05539369	0.13240067	0.77988771	1
2	0.03889402	0.14777587	0.77744139	1
3	0.04747028	0.12079861	0.76316177	1
4	0.05265369	0.14955488	0.74855764	1
5	0.03214535	0.13241285	0.796875	1
6	0.03517571	0.15297269	0.78872661	1
7	0.04391309	0.14896455	0.80224981	1
8	0.04229368	0.17241782	0.79289454	1
9	0.0351938	0.12358239	0.77388931	1
10	0.03525149	0.13957889	0.7902901	1
11	0.0282557	0.1963175	0.79953817	1
12	0.04756827	0.15935126	0.7504933	1
13	0.04332113	0.12316848	0.75127153	1
14	0.04150322	0.16552151	0.80436277	1
15	0.05046666	0.13433251	0.7711838	1
Promedio	0.04196665	0.14661003	0.77938823	1

Tabla A7.1: Medidas de calidad en el agrupamiento de la UE Síntomas para $\beta_0 = 0.3$

Corpus	*Overall Similarity*	*Índice Dunn*	*Average Similarity*	*Overall F-Measure*
1	0.05688218	0.11862658	0.8063262	1
2	0.03796452	0.13189914	0.82431204	1
3	0.04858636	0.1109713	0.80637903	1
4	0.05621761	0.13657103	0.77380427	1
5	0.03390678	0.12622818	0.83536803	1
6	0.0381737	0.14293703	0.82041995	1
7	0.04370032	0.14466705	0.81280054	1
8	0.04744637	0.15608004	0.79777097	1
9	0.04274592	0.11300542	0.79417221	1
10	0.0476331	0.1248166	0.81942249	1
11	0.0307788	0.16751506	0.8540092	1
12	0.05414008	0.14999515	0.79097242	1
13	0.05128356	0.09952327	0.79906301	1
14	0.05118847	0.18604626	0.84213665	1
15	0.05006936	0.16038618	0.80614549	1
Promedio	0.04604781	0.13795122	0.81220683	1

Tabla A7.2: Medidas de calidad en el agrupamiento de la UE Síntomas para $\beta_0 = 0.4$

Corpus	*Overall Similarity*	*Índice Dunn*	*Average Similarity*	*Overall F-Measure*
1	0.06612526	0.11417951	0.83660941	1
2	0.04783431	0.13866199	0.82974601	1
3	0.05771396	0.12197161	0.83257923	1
4	0.07817615	0.12053199	0.81476707	1
5	0.03145033	0.15510356	0.85769022	1
6	0.04789092	0.15414789	0.85888148	1
7	0.04588119	0.12951714	0.84513714	1
8	0.05839777	0.16068396	0.81979892	1
9	0.04721473	0.1403325	0.83339104	1
10	0.04489154	0.11198667	0.84964082	1
11	0.03660528	0.15574711	0.88094144	1
12	0.05194461	0.15064712	0.81277596	1
13	0.06245261	0.11564835	0.81412694	1
14	0.05015811	0.18422228	0.84350866	1
15	0.04691343	0.17063819	0.82435453	1
Promedio	0.05157668	0.14160133	0.83692992	1

Tabla A7.3: Medidas de calidad en el agrupamiento de la UE Síntomas para $\beta_0 = 0.5$

Anexo 8. Calidad del agrupamiento de la colección asociada a la UE Signos para diferentes configuraciones del umbral de similitud.

Corpus	*Overall Similarity*	*Índice Dunn*	*Average Similarity*	*Overall F-Measure*
1	0.052794522	0.184573419	0.81248013	1
2	0.023089723	0.222626102	0.872335589	1
3	0.037576555	0.185726993	0.85514826	1
4	0.052572451	0.142321173	0.696962863	1
5	0.035780396	0.194002631	0.768311188	1
6	0.0586506	0.167732237	0.800964723	1
7	0.048478899	0.16145943	0.788427327	1
8	0.049235059	0.167715797	0.752401675	1
9	0.027584678	0.219393278	0.744034082	1
10	0.033381798	0.182125177	0.815717225	1
11	0.030337503	0.233736545	0.867878765	1
12	0.040130346	0.169371796	0.913477721	1
13	0.05820162	0.172224462	0.799945852	1
14	0.041951244	0.168546466	0.79056453	1
15	0.050370603	0.149712427	0.702500746	1
Promedio	0.042675733	0.181417862	0.798743378	1

Tabla A8.1: Medidas de calidad en el agrupamiento de la UE Signos para $\beta_0 = 0.3$

Corpus	*Overall Similarity*	*Índice Dunn*	*Average Similarity*	*Overall F-Measure*
1	0.056882176	0.118626579	0.806326201	1
2	0.037964522	0.131899144	0.824312043	1
3	0.048586364	0.110971296	0.806379034	1
4	0.056217611	0.136571033	0.773804266	1
5	0.033906783	0.126228184	0.835368025	1
6	0.038173704	0.142937032	0.820419953	1
7	0.043700324	0.144667045	0.812800539	1
8	0.047446366	0.156080043	0.797770967	1
9	0.042745923	0.113005423	0.794172212	1
10	0.0476331	0.124816604	0.819422491	1
11	0.030778805	0.167515058	0.854009196	1
12	0.054140076	0.14999515	0.790972422	1
13	0.051283565	0.099523267	0.799063014	1
14	0.051188475	0.186046256	0.842136651	1
15	0.050069364	0.160386181	0.80614549	1
Promedio	0.04604781	0.13795122	0.812206834	1

Tabla A8.2: Medidas de calidad en el agrupamiento de la UE Signos para $\beta_0 = 0.4$

Corpus	*Overall Similarity*	*Índice Dunn*	*Average Similarity*	*Overall F-Measure*
1	0.063675179	0.183001693	0.881821144	1
2	0.023364274	0.179246918	0.923021482	1
3	0.039036853	0.159014592	0.908373117	1
4	0.066092102	0.122338616	0.819847439	1
5	0.041044652	0.151631422	0.858976234	1
6	0.067586166	0.156332211	0.84053865	1
7	0.057602706	0.128321189	0.820676549	1
8	0.06470266	0.14680048	0.81517412	1
9	0.034340319	0.156377303	0.855885485	1
10	0.031331249	0.159625289	0.887823101	1
11	0.030262911	0.227892645	0.949083581	1
12	0.040993678	0.161379774	0.937485156	1
13	0.056284602	0.19360727	0.840206945	1
14	0.036800382	0.152568988	0.876314425	1
15	0.048321589	0.128358559	0.835854757	1
Promedio	0.046762622	0.16043313	0.870072146	1

Tabla A8.3: Medidas de calidad en el agrupamiento de la UE Signos para $\beta_0 = 0.5$

Anexo 9. Calidad del agrupamiento de la colección asociada a la UE Historia de la Enfermedad Actual para diferentes configuraciones del umbral de similitud.

Corpus	*Overall Similarity*	*Índice Dunn*	*Average Similarity*	*Overall F-Measure*
1	0.080041532	0.095638297	0.622040517	1
2	0.03962601	0.132631944	0.697571037	1
3	0.06860413	0.097223948	0.602365228	1
4	0.08237018	0.090975607	0.616325684	1
5	0.035371619	0.177797236	0.668224801	1
6	0.069581063	0.113864007	0.61812319	1
7	0.065416095	0.104513092	0.635040087	1
8	0.080144413	0.082285035	0.592918129	1
9	0.043786126	0.135761822	0.651097165	1
10	0.043436232	0.148853332	0.667182716	1
11	0.034822099	0.175945155	0.72386895	1
12	0.063207618	0.144267179	0.629064671	1
13	0.066880748	0.107688183	0.633969913	1
14	0.057864181	0.131104544	0.688985386	1
15	0.055952234	0.135271747	0.651458606	1
Promedio	0.059140285	0.124921409	0.646549072	1

Tabla A9.1: Medidas de calidad en el agrupamiento de la UE Historia de la Enfermedad Actual para $\beta_0 = 0.3$

Corpus	*Overall Similarity*	*Índice Dunn*	*Average Similarity*	*Overall F-Measure*
1	0.085482051	0.101954784	0.679738153	1
2	0.046974549	0.114638657	0.716071461	1
3	0.078634011	0.099342033	0.653892793	1
4	0.083119906	0.101881316	0.676070566	1
5	0.057036502	0.127365777	0.730417467	1
6	0.094620343	0.128420882	0.657914283	1
7	0.070398612	0.094025915	0.700571619	1
8	0.07710979	0.119806216	0.698303374	1
9	0.057869605	0.109831138	0.695509272	1
10	0.057369413	0.134995479	0.705655446	1
11	0.047443403	0.136864223	0.739063616	1
12	0.068853751	0.124388473	0.69585044	1
13	0.080487769	0.124899843	0.692545409	1
14	0.062026626	0.109387734	0.72439217	1
15	0.067656923	0.119287275	0.699732936	1
Promedio	0.06900555	0.11647265	0.697715267	1

Tabla A9.2: Medidas de calidad en el agrupamiento de la UE Historia de la Enfermedad Actual para $\beta_0 = 0.4$

Corpus	***Overall Similarity***	***Índice Dunn***	***Average Similarity***	***Overall F-Measure***
1	0.063094994	0.098141902	0.793128728	1
2	0.054348354	0.134248275	0.756935133	1
3	0.076103698	0.079114008	0.72940652	1
4	0.070793638	0.085762404	0.765355037	1
5	0.055763725	0.117513324	0.766880978	1
6	0.074305165	0.117292171	0.780647663	1
7	0.075820945	0.092495439	0.75529237	1
8	0.068049354	0.134057172	0.779782769	1
9	0.060046978	0.096211666	0.748267744	1
10	0.057579294	0.118181615	0.742179111	1
11	0.050680358	0.136755758	0.795145853	1
12	0.072350146	0.127109542	0.725675324	1
13	0.070463458	0.112037981	0.774461032	1
14	0.053517198	0.093574064	0.811938846	1
15	0.060631369	0.118673082	0.778002957	1
Promedio	0.064236578	0.11074456	0.766873338	1

Tabla A9.3: Medidas de calidad en el agrupamiento de la UE Historia de la Enfermedad Actual para $\beta_0 = 0.5$

Anexo 10. Calidad del agrupamiento de la colección asociada a la UE Hábitos Tóxicos para diferentes configuraciones del umbral de similitud.

Corpus	*Overall Similarity*	*Índice Dunn*	*Average Similarity*	*Overall F-Measure*
1	0.055482444	0.202872914	0.858306322	1
2	0.047146081	0.234845804	0.938454227	1
3	0.023721004	0.206542809	0.869765912	0.928930366
4	0.056394435	0.24047558	0.912146734	0.876190476
5	0.051194744	0.26205544	0.877958152	0.85625
6	0.057123039	0.261222498	0.895367392	0.838383838
7	0.038845806	0.16446267	0.867057593	1
8	0.064963679	0.14546456	0.863434828	1
9	0.046779488	0.227610628	0.907120065	0.813114754
10	0.023287255	0.272783675	0.870161536	0.885245902
11	0.012976855	0.35391344	0.913541045	0.767803598
12	0.039719689	0.239039821	0.892107559	0.839179842
13	0.076172815	0.210323122	0.863099634	0.829431438
14	0.026871903	0.224788731	0.926884251	1
15	0.04394816	0.193320349	0.883979828	1
Promedio	0.044308493	0.229314803	0.889292338	0.908968681

Tabla A10.1: Medidas de calidad en el agrupamiento de la UE Hábitos Tóxicos para $\beta_0 = 0.3$

Corpus	*Overall Similarity*	*Índice Dunn*	*Average Similarity*	*Overall F-Measure*
1	0.056955251	0.181135175	0.879773637	1
2	0.047146081	0.234845804	0.938454227	1
3	0.022074685	0.184935458	0.893167029	0.928930366
4	0.056394435	0.24047558	0.912146734	0.876190476
5	0.051934248	0.211834149	0.884808072	0.869863014
6	0.062209941	0.219161917	0.896598139	1
7	0.038845806	0.16446267	0.867057593	1
8	0.064963679	0.14546456	0.863434828	1
9	0.05120985	0.202826966	0.90278139	0.837704918
10	0.023287255	0.272783675	0.870161536	0.885245902
11	0.015541463	0.267903476	0.950376974	0.792490119
12	0.038441613	0.207898189	0.903482817	0.885869565
13	0.083029027	0.191713801	0.863654702	1
14	0.026871903	0.224788731	0.926884251	1
15	0.04394816	0.193320349	0.883979828	1
Promedio	0.04552356	0.209570033	0.895784117	0.938419624

Tabla A10.2: Medidas de calidad en el agrupamiento de la UE Hábitos Tóxicos para $\beta_0 = 0.4$

Corpus	***Overall Similarity***	***Índice Dunn***	***Average Similarity***	***Overall F-Measure***
1	0.056381361	0.186587343	0.885690101	1
2	0.047146081	0.234845804	0.938454227	1
3	0.040448185	0.143980133	0.904210569	0.966834171
4	0.04590947	0.210863621	0.941887063	1
5	0.056706721	0.191712908	0.904879413	0.875
6	0.059977631	0.222820588	0.942661248	1
7	0.035455366	0.170700816	0.901726339	1
8	0.074165388	0.171743784	0.868850073	1
9	0.04630953	0.193220584	0.921018972	0.837704918
10	0.028038301	0.237510452	0.906909089	0.885245902
11	0.023941035	0.218894864	0.954864367	1
12	0.05568433	0.182209278	0.918018441	1
13	0.077658889	0.195445913	0.88124201	1
14	0.041490985	0.17425374	0.963454543	1
15	0.068648048	0.183546736	0.925467192	1
Promedio	0.050530755	0.194555771	0.91728891	0.970985666

Tabla A10.3: Medidas de calidad en el agrupamiento de la UE Hábitos Tóxicos para $\beta_0 = 0.5$

Anexo 11. Calidad del agrupamiento de la colección asociada a la UE Antecedentes para diferentes configuraciones del umbral de similitud.

Corpus	*Overall Similarity*	*Índice Dunn*	*Average Similarity*	*Overall F-Measure*
1	0.059832369	0.145299767	0.844307173	1
2	0.028044803	0.135167798	0.83706869	1
3	0.037374806	0.116344646	0.882700705	1
4	0.062557397	0.128634752	0.848031649	1
5	0.030362609	0.142813425	0.83449121	1
6	0.05641264	0.144449865	0.854339519	1
7	0.047660596	0.13519525	0.853556635	1
8	0.050465489	0.119025839	0.774307282	1
9	0.027898699	0.117425459	0.828845488	1
10	0.019569298	0.253204951	0.886181849	1
11	0.032830258	0.179573243	0.834084597	1
12	0.039506389	0.254894967	0.902620706	1
13	0.036443221	0.177999562	0.873459277	1
14	0.039442486	0.149682653	0.859068303	1
15	0.035484109	0.134347241	0.86278433	1
Promedio	0.040259011	0.155603961	0.851723161	1

Tabla A11.1: Medidas de calidad en el agrupamiento de la UE Antecedentes para $\beta_0 = 0.3$

Corpus	*Overall Similarity*	*Índice Dunn*	*Average Similarity*	*Overall F-Measure*
1	0.058105199	0.132337215	0.85903679	1
2	0.026157302	0.115579495	0.849230039	1
3	0.037374806	0.116344646	0.882700705	1
4	0.05664689	0.119663991	0.869029835	1
5	0.038766415	0.131728145	0.84528813	1
6	0.063168785	0.141899742	0.859873153	1
7	0.05978766	0.125837484	0.874750077	1
8	0.05650039	0.17238595	0.811162523	1
9	0.032593263	0.100033908	0.845285001	1
10	0.026111332	0.226345916	0.89465206	1
11	0.0331472	0.175890874	0.844830752	1
12	0.039506389	0.254894967	0.902620706	1
13	0.042657597	0.177820026	0.875433925	1
14	0.038308947	0.174396579	0.879304868	1
15	0.041749043	0.11913252	0.861913103	1
Promedio	0.043372081	0.152286097	0.863674111	1

Tabla A11.2: Medidas de calidad en el agrupamiento de la UE Antecedentes para $\beta_0 = 0.4$

Corpus	***Overall Similarity***	***Índice Dunn***	***Average Similarity***	***Overall F-Measure***
1	0.067303129	0.145571269	0.872061049	1
2	0.030093888	0.106392641	0.887625738	1
3	0.045145574	0.102638356	0.901899704	1
4	0.071141079	0.116925082	0.885985941	1
5	0.037425745	0.14413862	0.874005314	1
6	0.05899861	0.1661092	0.878454842	1
7	0.070335382	0.12192495	0.891055038	1
8	0.053681624	0.140247973	0.861187239	1
9	0.034137236	0.182348881	0.886054078	1
10	0.027724275	0.201288708	0.910179297	1
11	0.03475556	0.144837466	0.874653607	1
12	0.046357766	0.218450989	0.91365739	1
13	0.043977683	0.162166643	0.883466074	1
14	0.035343023	0.154828925	0.893796122	1
15	0.047347027	0.14180575	0.894466113	1
Promedio	0.04691784	0.149978364	0.887236503	1

Tabla A11.3: Medidas de calidad en el agrupamiento de la UE Antecedentes para $\beta_0 = 0.5$

Anexo 12. Variantes para el cálculo del umbral de similitud entre objetos

La forma de medir la similitud y qué umbral utilizar para formar conjuntos de relaciones, es una tarea difícil que depende del dominio donde fue aplicado, cómo fueron descritos los objetos y qué nivel de granularidad se desea evaluar en los resultados. Otros elementos que influyen en la estimación del umbral son la variabilidad en la densidad de los grupos y la varianza y desviación estándar de las similitudes. Por otro lado, el umbral, en algunos casos, constituye una herramienta que tiene el usuario para hacer que el método se ajuste a sus requerimientos y características del problema (Arco, 2009).

Cálculo del umbral de similitud global

A continuación se exponen algunas variantes para el cálculo del umbral de similitud inicial, que requiere el algoritmo de agrupamiento propuesto en la sección anterior. El cálculo en cada uno de los criterios se realiza a partir de la matriz de similitud y no se requiere información adicional del conjunto de datos que se procesa.

Se considera *m* como la cantidad de objetos de la colección y $s(O_i, O_j)$ el valor de similitud entre los objetos O_i y O_j (Ruiz-Shulcloper, 1995).

Definición 2.5 (Umbral de Semejanza). La magnitud γ se denominará umbral de semejanza y puede ser calculada de la siguiente manera:

1. La media de las similitudes entre todos los pares de objetos posibles; expresión A12.1:

$$\bar{X} = \frac{1}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{n} s(o_i, o_j) \qquad \text{(A12.1)}$$

2. La media de los valores máximos de las similitudes entre cualquier par de objetos; expresión A12.2:

$$\bar{X}_{max} = \frac{1}{m} \sum_{i=1}^{m-1} {}^{j=1..m}_{i\neq j}\max[s(o_i, o_j)] \qquad \text{(A12.2)}$$

3. La media de los valores mínimos de las similitudes entre cualquier par de objetos; expresión A12.3:

$$\bar{X}_{min} = \frac{1}{m}\sum_{i=1}^{m-1} {}_{\substack{j=1..m \\ i\neq j}}\min[s(o_i, o_j)] \quad \text{(A12.3)}$$

Cálculo del umbral de similitud grupal

A continuación se exponen algunos criterios para el cálculo del umbral de similitud grupal.

En este contexto se considera *m* como la cantidad de objetos del grupo$_i$.

Definición 2.6. La magnitud $\varphi(d_k, c_i)$ se denomina umbral de similitud grupal y puede calcularse como se muestra a continuación:

1. La media de las similitudes entre todos los pares de objetos posibles que pertenecen al grupo; expresión A12.4:

$$\overline{X(o_k, c_\iota)} = \frac{1}{m}\sum_{i=1}^{m} s(o_i, o_k) \quad \text{(A12.4)}$$

2. El valor máximo de similitud que alcanza con uno de los elementos del grupo; expresión A12.5:

$$\bar{X}_{max(o_k, c_i)} = \max[s(o_i, o_{jk})] \quad \text{(A12.5)}$$

3. El valor mínimo de similitud que alcanza con uno de los elementos del grupo; expresión A12.6:

$$\bar{X}_{min(o_k, c_i)} = min[s(o_i, o_k)] \quad \text{(A12.6)}$$

Anexo 13. 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:_____

_____

_____

Anexo 14. 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.

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

Tabla A14.1: 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

Anexo 15. 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, cardiovascular, 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.