Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Aplicación de técnicas de graph mining para buscar patrones en la lista de problemas de la historia clínica del Hospital Italiano de Buenos Aires.

Tesis presentada para optar al título de
Magister en explotación de datos y descubrimiento del conocimiento

Ing. María del Pilar Ávila Williams

Director: Dr. Marcelo Soria

Buenos Aires, 2018

APLICACIÓN DE TÉCNICAS DE GRAPH MINING PARA BUSCAR PATRONES EN LA LISTA DE PROBLEMAS DE LA HISTORIA CLÍNICA DEL HOSPITAL ITALIANO DE BUENOS AIRES

La lista de problemas es el componente estructural de la historia clínica electrónica del Hospital Italiano de Buenos Aires, en ella se detallan los hallazgos y observaciones de los pacientes. Se presenta un análisis basado en teoría de grafos con el objetivo final de encontrar agrupaciones significativas entre problemas antes del 2016. En este modelo los problemas son los nodos y los enlaces son los vínculos con SNOMED CT y la co-ocurrencia en los pacientes. Este análisis comprende la construcción de subconjuntos de los contextos: servicios de atención de salud, nivel asistencial o ámbito y grupo etario. Para evaluar la capacidad predictiva de las agrupaciones se utiliza las métricas de precisión y exactitud en una lista de 10 predicciones seleccionadas en la lista de problemas de paciente en el año 2017. Los resultados mostraron que realizar la lista de predicciones usando sólo los problemas de los contextos, mejora significativamente la capacidad predictiva de las agrupaciones, especialmente en el contexto de servicio de atención de salud y grupo etario.

Palabras claves: Graph mining, lista de problemas, redes complejas, detección de comunidades, SNOMED CT.

APPLICATION OF GRAPH MINING TECHNIQUES TO LOOK FOR PATTERNS IN THE LIST OF PROBLEMS IN THE MEDICAL HISTORY OF THE HOSPITAL ITALIANO DE BUENOS AIRES

The problem list is the structural component of the electronic medical record of the Hospital Italiano de Buenos Aires, in which the clinical findings and observations of the patients are detailed. An analysis based on graph theory is presented, with the final purpose of finding significant clusters of problems before 2016. In this model, the problems are the nodes, and the links are the connections with SNOMED CT and the co-occurrence in the patients. This analysis includes the construction of subsets with contexts: health care services, level of care or scope and age group. To evaluate the predictive capacity of the clusters, precision and accuracy metrics are used in a list of 10 predictions selected in the problem list of patients for the year 2017. The results showed that making the list of predictions using only the problems of the contexts, significantly improves the predictive capacity of the clusters, especially in the context of health care service and age group.

Keywords: Graph mining, Problem list, Complex networks, Community detection, SNOMED CT.

AGRADECIMIENTOS

Este trabajo fue realizado con el apoyo del Departamento de Informática en Salud del Hospital Italiano de Buenos Aires, agradezco principalmente al Dr. Daniel Luna por crear el espacio necesario para concebir el proyecto y llevarlo a cabo, a la Dra. Sonia Benitez y mis compañeros Hernán Berinsky y Hee Park que fueron fundamentales para enfocar epistemológicamente esta investigación. Al grupo de terminología, por su apoyo constante y por su compañerismo.

Agradezco a muchas otras personas que me ayudaron desde lo técnico, y desde lo emocional. Esas personas hacen posible que uno de mis sueños se haga realidad.

A mi director Marcelo Soria, que reafirma mi teoría: lo más importante al realizar una tesis es tener un excelente director.

A mi familia, que son el motor de mi vida y la fuente de la eterna felicidad. Mi abuela, que me enseñó que estudiar era lo más importante de mi vida.

A mis amigos, los que se fueron, los que dejé en mi país, los que gané en este país. Gracias por hacerme sentir su compañía aún en la distancia, gracias por enseñarme cuáles son las cosas importantes en la vida.

A mi esposa, compañera de todas mis batallas y cómplice de tantos sueños.

Los llevaré en el corazón toda la vida, gracias.

A las mujeres de mi vida: mi abuela, mi madre y mi esposa

Índice general

LISTADO DE FIGURAS

LISTADO DE TABLAS

ABREVIATURAS

cdf distribución acumulativa. 11, 24, 46

CMT Convergent Medical Terminology. 7, 23, 38, 41, 42

CRISP-DM CRoss Industry Standard Process for Data Mining. 19, 22

HCE historia clínica electrónica. 1–3, 7, 23, 29, 30, 33, 40

HIBA Hospital Italiano de Buenos Aires. 1–4, 7, 22, 25, 29, 38–42, 45

IHTSDO International Health Terminology Standards Development Organisation. 3, 5

refset reference set. 6, 7, 16, 23, 29, 36, 38, 39, 41, 42, 45, 56, 64–66

RP Red de problemas. 45–48, 51, 52, 61, 62, 64, 65

RP-SCT Red semántica de problemas. 45–48, 51–54, 61, 62, 64, 65

UMLS Unified Medical Language System. 7

1. CAPÍTULO: INTRODUCCIÓN

1.1. Motivación

Las historias clínicas orientadas a problemas monitorizan y detallan una lista de problemas médicos para cada paciente, que incluyen tanto el diagnóstico final como todos aquellos hallazgos que aún no han sido específicamente diagnosticados (Weed, 1968). La lista está separada en problemas activos y pasivos. Los problemas activos deben representar la situación actual del paciente, dándole al médico una herramienta para la toma de decisiones sobre el tratamiento a seguir. Una vez que se ha establecido la lista actualizada de problemas, todas las subsecuentes órdenes, planes, notas de progreso y datos numéricos deben ser asociados a cada problema. Inherente a este enfoque está la necesidad de completitud y actualización en la formulación de la lista de problemas. Esto requiere que los datos sean apropiadamente recolectados y actualizados de activos a pasivos, de tal manera que las conclusiones que emergen a partir de los datos sean lógicas y relevantes. Otra necesidad es la exactitud e integridad con la que los problemas son definidos inicialmente y esto tiene relación directa con la granularidad con la que se seleccionan los problemas (Luna y cols., 2013).

El *Institute of Medicine (IOM)*[1] y la *Joint Commission*[2] recomiendan que la lista de problemas sea exacta, actualizada y completa. Éstas se han convertido en la característica más importante para medir su calidad, ya que una lista de problemas actualizada y con un nivel adecuado de detalles mejora la comunicación entre profesionales de la salud, y se espera que también mejore la calidad en la atención a los pacientes. La exactitud en la lista de problemas impacta directamente en los sistemas de toma de decisiones del hospital y en la implementación de sus programas de salud. Sin embargo, a pesar de los numerosos beneficios, las listas de problemas son a menudo imprecisas, incompletas y desactualizadas ya que no representan la situación actual del paciente.

La historia clínica electrónica (HCE) del Hospital Italiano de Buenos Aires (HIBA) utiliza listas de problemas como componente estructural, y se han podido identificar deficiencias en la completitud, actualización y exactitud (Otero, 2014). A pesar de que implementa Snomed CT[3] como terminología clínica de referencia para identificar los problemas, su gran tamaño es también un obstáculo para su uso y mantenimiento. Existe evidencia que indica que se está utilizando sólo una pequeña fracción de su contenido (Lezcano y Sicilia, 2011). Además, un estudio concluyó que en las áreas de hospitalización, emergencia y ambulatorios, los problemas registrados por los profesionales en la atención primaria de salud tienen mejor calidad que los registrados por los especialistas, y sugiere que trabajar con interfaces orientadas al contexto mejora la exactitud (Luna y cols., 2013).

[1] El *Institute of Medicine (IOM)* es una organización independiente y sin ánimo de lucro fundada en 1970 como el brazo de la salud de la Academia Nacional de las Ciencias de Estados Unidos. Su objetivo es ayudar a sectores de gobierno y privado a tomar decisiones de salud informadas al proporcionar evidencia confiable

[2] La *Joint Commission* es una organización independiente y sin ánimo de lucro que acredita y certifica más de 20.500 organizaciones y programas de salud en Estados Unidos. La acreditación y certificación *Joint Commission* es reconocida a nivel mundial como un símbolo de calidad que refleja el compromiso de la organización para cumplir con ciertos estándares de desempeño.

[3] http://www.ihtsdo.org/

El desarrollo de esta tesis propone un enfoque basado en la teoría de grafos para agrupar los conceptos de Snomed CT que se utilizan en la composición de las listas de problemas del HIBA. Se analiza y evalúa la calidad de los grupos generados automáticamente en relación a diferente información contextual como el ámbito (internación, ambulatorios, emergencia y atención domiciliaria), el servicio de atención de la salud y el grupo etario del paciente. A partir de las relaciones de estos grupos se construye una taxonomía complementaria para clasificar los problemas de una manera que sea consistente con su uso dentro de la HCE.

La taxonomía de los problemas tiene un impacto directo en la organización de los datos y permite la creación de servicios de búsqueda para recuperar problemas basados en el contexto y en la co-ocurrencia de problemas. Se espera que estos servicios puedan ayudar a mejorar la calidad de la lista de problemas dentro de la HCE del HIBA.

1.2. Marco teórico

En este sección se presenta un marco general enfocado en las definiciones del dominio del problema de investigación. Las principales definiciones son sobre los datos y el análisis de grafos.

Los datos son la lista de problemas del HIBA y Snomed CT, la terminología con la que se codifican los diagnósticos y hallazgos. Snomed CT es una red, que matemáticamente se modela como un grafo, y se usa para representar conocimiento médico. Snomed CT se usa en el registro primario para representar el conocimiento de los problemas de los pacientes, lo que permite pasar de problemas más genéricos a más específicos en una organización jerárquica. En el uso secundario y de tomas de decisiones, Snomed CT tiene múltiples casos de uso con fines epidemiológicos, estadísticos, económicos, soporte para la investigación, etc (Lee, de Keizer, Lau, y Cornet, 2014).

Después de definir los datos, en las siguientes secciones presento los diferentes patrones que se evalúan en el análisis de grafos y de las redes que representan los problemas. El objetivo de estas secciones es hallar comunidades significativas que permitan agrupar los problemas de la red según su información contextual y también detectar valores atípicos.

El análisis de redes y descubrimiento de patrones se enmarcan en las definiciones de *graphmining*, ya que la estructura básica de todos los datos es un grafo. El descubrimiento implica tareas de (a) aprendizaje supervisado encontrando patrones que distinguen un subgrafo de otros, y para lo cual se necesita un conjunto de ejemplos positivos y otro conjunto de ejemplos negativos; (b) aprendizaje no supervisado por medio del agrupamiento o *clustering*; y (c) visualización de grafos para representar el conocimiento. En esta tesis sólo se investiga la aplicación de aprendizaje no supervisado.

1.2.1. Lista de problemas

A fines de la década del sesenta, Weed (Weed, 1968) publicó sus ideas respecto a los **Registros Médicos Orientados a Problemas** que permiten identificar y monitorear cada problema médico. Esta lista de problemas debería ser una tabla dinámica de contenidos que pueden ser actualizados en cualquier momento. El HIBA implementó su HCE a partir de 1998, utilizando la lista de problemas como componente estructural. Este proyecto desarrollado *in-house* permite el registro de toda la información relacionada con la salud de los pacientes para su posterior análisis (Luna y cols., 2013, 2003). La HCE es una aplicación web, orientada a problemas y centrada en los pacientes. Las funcionalidades

de la HCE dependen del ámbito o nivel de asistencia (ambulatorio, internación general, guardia, triage, internación geriátrica, internación domiciliaria, seguimiento domiciliario, episodio externo y episodio ambulatorio).

Durante la implementación de la HCE y la capacitación de los médicos se definió que un **problema** es el motivo de consulta o diagnóstico que genera una acción por parte del sistema de salud (López Osornio y cols., 2004). Cuando un paciente llega a atenderse, el flujo de trabajo de la atención médica requiere que los profesionales ingresen los problemas mediante texto libre, en lugar de navegar por las jerarquías de la terminología buscando los mejores términos que los describan.

El texto narrativo no estructurado es la forma de documentación más frecuentemente usada en medicina. Por medio de la codificación se intenta disminuir la ambigüedad propia del texto libre. Los motivos por los cuales se codifica son múltiples, tales como el económico (para facturar un acto médico), epidemiológico o estadístico (para tener datos sobre incidencia y prevalencia de patologías en una población dada), soporte para la investigación (permite la recuperación de información para estudios científicos), asistencial (permite reclutar candidatos para programas de gestión de enfermedades). En el contexto de una HCE la codificación de texto libre es también útil para el funcionamiento de sistemas de soporte clínico en la toma de decisiones (López Osornio y cols., 2004; López Osornio, Luna, y de Quiros, 2002).

El proceso de la codificación se realiza de manera secundaria y centralizada. Un número reducido de profesionales de la medicina, que concentran el conocimiento de la clasificación a utilizar, son los responsables de asignar secundariamente los códigos correspondientes al texto libre que el personal asistencial registró durante la atención. Esta modalidad asegura una mejor consistencia en la codificación (López Osornio y cols., 2004, 2005). En el HIBA se codificaron más de 1 700 000 textos libres cargados entre 1998 y 2017. La codificación se realizó usando la terminología Snomed CT.

1.2.2. Snomed CT

Snomed es una nomenclatura desarrollada por el Colegio Americano de Patólogos con descripciones de morfología y anatomía. Su primera versión fue lanzada en el año 1965.Desde entonces tuvo diferentes versiones, una de estas versiones fue Snomed RT del año 1997 cuyo desarrollo se basó en **lógica descriptiva**, la cual especificaba su semántica. Esto significa que el conocimiento está representado de una forma estructurada y formalmente bien comprendida. Esta última versión se unió en el año 2002 a *Clinical Terms Version 3* desarrollada por el sistema de salud británico, dando origen a Snomed CT por *Clinical Terms*. En 2007, la *International Health Terminology Standards Development Organisation (IHTSDO)* adquirió los derechos de propiedad intelectual sobre todas las versiones de Snomed. Aunque en un principio se tratara de una nomenclatura sistematizada de medicina, en la actualidad es la terminología más completa y precisa del mundo. (Bhattacharyya, 2016)

Snomed CT permite el almacenamiento y la recuperación de información clínica basada en su significado, es decir que la definición de la información se construye a partir de relaciones semánticas (Rector, Brandt, y Schneider, 2011; Bhattacharyya, 2016; IHTSDO, 2016b). Cuando un solo concepto no es suficiente para definir la información, se puede crear uno nuevo usando especificaciones previamente establecidas según la representación composicional. Este proceso se llama post-coordinación. De lo contrario, cuando un solo concepto es suficiente para definir la información y mapea de manera exacta con un con-

cepto de Snomed CT, se llama pre-coordinación. El HIBA extendió Snomed CT a partir del año 2002. Al año 2017, 1 700 000 descripciones de texto libre estaban agrupadas en 520 000 conceptos post-coordinados y 60 000 pre-coordinados.

1.2.2.1. Componentes de Snomed CT

El modelo lógico de Snomed CT define tipos de componentes y la manera en la cual cada componente y sus derivados están relacionados. Los tipos de componentes son conceptos, descripciones y relaciones:(IHTSDO, 2016b)

Conceptos: Cada concepto representa una única entidad clínica, el cual tiene un identificador único, numérico y fácilmente procesable por un computador. El identificador provee una referencia única y sin ambigüedades a cada concepto y no tiene contenido semántico. El identificador es abreviado con SCTID.(IHTSDO, 2016b)

Descripciones: Un conjunto de descripciones textuales se asignan a cada concepto. De esta manera el usuario humano obtiene una representación del concepto en lenguaje natural.

Relaciones: Una relación representa una asociación entre dos conceptos. Las relaciones se usan para definir lógicamente el significado de un concepto de tal manera que pueda ser procesada por un computador. Hay dos tipos de relaciones:

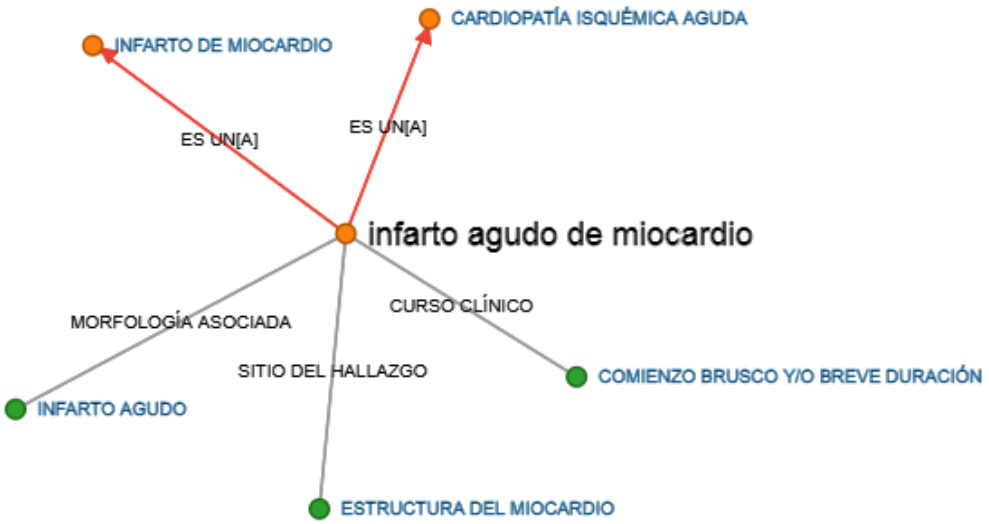
1. Relaciones subtipo: Usa la relación |ES UN|, de tal manera que el concepto P |ES UN| Q, define que el concepto P es un subtipo de concepto Q, o de manera inversa que Q es un supertipo de P

2. Relaciones atributos: Una relación atributo contribuye a la definición del concepto mediante la asociación con otros valores que permiten la caracterización del concepto. Estos valores pueden ser estructuras corporales, sustancias, objetos físicos, etc.

Por consiguiente, el modelo lógico especifica una representación estructurada de los conceptos para definir las entidades clínicas, las descripciones que son usadas para representar las diferentes variaciones léxicas del concepto, y las relaciones entre los conceptos.(IHTSDO, 2016b)

En la figura 1.1 se puede observar un ejemplo de las relaciones entre conceptos. El concepto **Infarto agudo de miocardio** tiene dos relaciones supertipo: **Infarto de miocardio** y **Cardiopatía isquémica aguda**. Contribuyendo a la definición de **Infarto agudo de miocardio**, hay tres relaciones atributos: (1) la morfología asociada, que especifica los cambios morfológicos observados a nivel tisular o celular que son característicos de una enfermedad, en este caso la morfología asociada es **infarto agudo**; (2) el sitio de hallazgo, que especifica el lugar del cuerpo en el que se localiza el hallazgo, para este caso el sitio del hallazgo es la **estructura del miocardio**; y (3) el curso clínico, que representa tanto el comienzo como el curso de una enfermedad, para este caso el curso clínico es **comienzo brusco y/o breve duración**.

Fig. 1.1: Ejemplo de concepto y relaciones de Snomed CT: Infarto agudo de miocardio



1.2.2.2. Modelo de Conceptos de Snomed CT

La organización jerárquica de Snomed CT permite pasar de conceptos más genéricos a más específicos (Bhattacharyya, 2016). El modelo define la forma en que los conceptos están dispuestos dentro de los subtipos de jerarquía y los tipos de relaciones de atributos que se permiten entre conceptos (Bhattacharyya, 2016; IHTSDO, 2016b). Como consecuencia, cada concepto debe pertenecer a una sola jerarquía.

Hay 19 jerarquías, no todas son usadas para representar conceptos clínicos. Las siguientes 8 jerarquías tienen reglas para definir atributos y crear conceptos clínicos según IHTSDO (IHTSDO, 2016a).

- **Hallazgo clínico**, representa los resultados de una observación clínica, evaluación o juicio e incluye estados clínicos normales y anormales. La jerarquía de hallazgo clínico incluye conceptos para representar diagnósticos.

- Los **procedimientos** representan actividades realizadas en la prestación de servicios de salud. Incluyen no sólo los procedimientos invasivos sino también la administración de medicamentos, toma de imágenes, educación, terapias y procedimientos administrativos.

- Los **especímenes** representan entidades que se obtienen, usualmente de pacientes, para el análisis.

- **Estructuras corporales**, representan estructuras anatómicas normales y anormales.

- **Productos biológicos/farmacéuticos**, representan productos farmacéuticos (no dispositivos)

- Las **situaciones de contexto explícito** representan conceptos en los que el contexto clínico se especifica como parte de una definición del mismo concepto. Esto incluye la presencia o ausencia de una condición, ya sea que el hallazgo clínico es actual, hace parte del pasado o está relacionado a alguien diferente al paciente. Ejemplos de estas situaciones de contexto explicito son: sospecha de cáncer de mama (SCTID: 134405005), antecedente de neoplasia maligna de mama (SCTID: 415076002) y antecedente familiar de neoplasia maligna de mama (SCTID: 429740004), respectivamente.
- Los **eventos** representan las ocurrencias que excluyen a los procedimientos y las intervenciones.
- Por último, los **objetos físicos** representan a los objetos naturales o hechos por el hombre.

La tabla 1.1 contiene las jerarquías usadas en la lista de problemas y la frecuencia de conceptos de Snomed CT. Se puede observar que las jerarquías más usadas son hallazgo clínico, situación con contexto explícito y procedimientos.

Tab. 1.1: Jerarquías de la lista de problemas y frecuencias de uso

Jerarquía	Conceptos diferentes	Cantidad de problemas
Hallazgo clínico	82 732	13 594 650
Situaciones de contexto explícito	5086	1 222 531
Procedimiento	14 991	1 138 167
Otras Jerarquías	1122	144 553
Evento	298	40 084
Estructura Corporal	186	11 274
Producto biológico/farmacéutico	35	2040
Objeto Físico	75	1285

1.2.2.3. *Reference set*

Para facilitar el uso de Snomed CT se construyen *reference set (refset)*, que son conjuntos de conceptos en determinados dominios como especialidades, cohortes, tipos de enfermedades, etc. Estos dominios se llaman subconjuntos simples.

Uno de los aspectos claves en el uso significativo de la terminología es la creación de *refset* que puedan ser utilizados como vocabularios controlados para contextos específicos. Estos *refset* facilitan el uso de Snomed CT como terminología de codificación primaria para la lista de problemas u otros niveles de documentación clínica, y maximizan potencialmente la interoperabilidad de datos a través de las instituciones(Dolin y cols., 2004).Snomed CT no define metodologías para construir *refset* con conceptos que se agrupen por un contexto. La metodología propuesta en esta tesis agrupa los conceptos usados en el registro histórico de la lista de problemas según el contexto dado por el área jerárquica, el grupo etario y el nivel de asistencia con el que fue registrado.

Existen *refset* públicos con características similares a los que se analizan en esta tesis. Estos *refset* fueron compilados por la organización Kaiser Permanente con el consen-

so de todos sus centros médicos[4]. Estos *refset* son útiles para hacer comparaciones con los resultantes en los experimentos de este trabajo: *Cardiology; Common Lab Procedures; Emergency Department; Endocrine, Nephrology, and Urology; ENT, Gastroenterology, and Infectious Diseases; Hematology and Oncology; History and Family History; Injury; Mental Health; Miscellaneous Problem List ; Musculoskeletal; Neurology; Obstetrics and Gynecology; Ophthalmology; Orthopedics; Pediatrics; Primary Care; Radiology; Skin/Dermatology and Respiratory; Specimen Source and Specimen Type; Urology & Nephrology; Vaccinations; Vascular Procedures List.*

1.2.3. Lista de problemas en la HCE del HIBA

Cualquier usuario con acceso a la HCE puede registrar un problema, este usuario puede pertenece a una área administrativa o asistencial del HIBA. Por consiguiente, las áreas jerárquicas que tiene el HIBA responden a necesidades operativas de la HCE, y tienen diferentes niveles de agregación, por ejemplo: Sección de neurocirugía vascular hace parte de una área más general llamada Servicio de neurocirugía. También existen áreas jerárquicas con muy pocos registros o que son *ad hoc* por ejemplo: Instituto universitario H.I., Programa de prevención del cáncer de colon hereditario, Laser en rinosinusología.

Además cada registro de la lista de problemas tiene asociado el nivel de asistencia que indica el ámbito en el que fue cargado el problema. Los niveles de asistencia son las diferentes modalidades de contacto que tiene el paciente con el hospital, asegurando una óptima atención en cada situación específica. En la HCE el nivel de asistencia puede ser cualquiera de los siguientes valores:

- Ambulatorio: Se tiene definido cuándo empieza y cuándo termina esa atención. El paciente solicita previamente un turno ambulatorio.
- Episodio ambulatorio: Se sabe cuándo empieza pero no cuándo termina; entonces la modalidad de registro cambia. En el ambulatorio generalmente hay alguien que longitudinalmente ve al paciente en distintos momentos.
- Triage: Se genera cuando un paciente tiene un contacto con el hospital sin un turno previo, consiste en una revisión médica rápida que permite definir la prioridad de atención.
- Guardia: Los pacientes críticos con inminencia de muerte o que ingresan con una patología aguda, de gravedad moderada o severa, pero sin muerte inminente por la misma. En este nivel de asistencia se espera que el alta del nivel se dé entre las 24 y 36 horas de su ingreso, pudiendo trasladarse a otro nivel asistencial del hospital, o a otro hospital, o más rara vez a su domicilio.
- Internación: Tiene un periodo limitado del cuidado, el paciente es tratado por un episodio grave de una enfermedad, cuyas condiciones puedan resultar en un trauma o su muerte. Los tipos de internación son general, domiciliaria y geriátrica.

[4] Kaiser Permanente es la organización de mantenimiento de salud, sin fines de lucro, más grande de los Estados Unidos (integra 28 centros médicos y tiene presencia en 8 estados). Es un sistema integrado de prestación de servicios de salud, que organiza y proporciona o coordina la atención de los miembros de la organización. Kaiser Permanente construyó una solución de terminología médica para toda la organización llamada *Convergent Medical Terminology (CMT)*, que son subconjuntos de terminología adaptada a pacientes y médicos, vinculada a los estándares de interoperabilidad de Estados Unidos e internacionales. Desde el año 2010 donan estos subconjuntos a la *Unified Medical Language System (UMLS)*

1.2.4. Redes

1.2.4.1. Definición

Las redes están presentes en casi cada aspecto de nuestra vida. La tecnología nos ubica en un mundo lleno de redes, las relaciones físicas o lógicas que establecemos con nuestro entorno constituyen una red en sí misma. Al final de la década del noventa, el avance y popularidad de los computadores cambió la manera como se entendían las redes. La capacidad de los computadores hizo posible acumular grandes bases de datos con estructuras de redes y analizarlas rápida y eficientemente. Esto permitió por primera vez, la comparación de datos de redes reales con los modelos existentes, en particular el modelo ER. Otra influencia importante de la revolución de la computación fue la creación y rápido desarrollo de dos redes enormes: la internet y la World Wide Web (WWW).

En consecuencia, el concepto abstracto de una red cubre una amplia variedad de estructuras en las cuales las entidades de los sistemas complejos son representados por vértices o nodos y las relaciones o interacciones entre esas entidades son representadas con aristas o enlaces de la red (Estrada y Knight, 2015).A continuación presento algunos ejemplos de redes usadas en las áreas de biología y medicina, aunque en la literatura siguen surgiendo nuevos.

1.2.4.2. Tipos de redes

Redes sociales: Contempla las redes con interacciones entre individuos. Estos pueden consistir en redes de amigos o conocidos, relaciones de trabajo o sexuales. Además de su importancia en los estudios sociales, entender la estructura de estas redes es también importante para los epidemiólogos, ya que es por medio de estas redes que las epidemias se propagan. (Cohen y Havlin, 2010)

Redes biológicas: Este tipo abarca diferentes tipos de redes, las redes biológicas pueden ser lógicas, representando interacciones entre proteínas, entre genes, o entre proteínas y genes. Interacciones entre moléculas en las vías metabólicas de la célula pueden ser vistas como una red. Aunque las interacciones son físicas, los enlaces no son entidades físicas sino la posibilidad de una interacción entre dos moléculas. Otras redes lógicas son las ecológicas predador-presa (donde los vértices son las especies, y los enlaces dirigidos representan la depredación de una especie por la otra). Otro tipo de redes biológicas son las redes biológicas físicas, como el sistema de nervios, las neuronas en el cerebro y la red de venas en un organismo. (Cohen y Havlin, 2010)

Redes de salud: A continuación enumero algunos trabajos con redes en el área de la salud.

- La red de enfermedades humanas (Goh y cols., 2007) es una red de trastornos y enfermedades genéticas vinculadas por genes en común. Esta red permite explorar el fenotipo y las enfermedades genéticas asociadas, indicando así el origen genético que tienen en común muchas enfermedades. El proyecto nació en el 2007, y para el 2014 se agregó la red de síntomas a partir de metadata de los encabezados de temas

médicos (MeSH[5]) de PubMED. A partir de estas redes se han hecho múltiples trabajos que explican la interacción entre enfermedades, interacciones entre proteínas, interacciones de enfermedades con drogas y comorbolidades.

- La red de medicina (Van Haaren, Davis, Lappenschaar, y Hommersom, 2013; Barabási, Gulbahce, y Loscalzo, 2011), fue creada bajo la premisa de que una enfermedad es rara vez consecuencia de una anormalidad en una sola parte del sistema corporal del paciente. Esto implica que las redes deben ilustrar el impacto y la complejidad de la asociación de las enfermedades, además que estas enfermedades pueden ser agrupadas y segmentadas.

1.2.5. Grafos

Los grafos se usan para describir de forma matemática relaciones entre redes. Los grafos representan las propiedades topológicas esenciales de una red mediante el tratamiento de ésta como una colección de nodos y enlaces. Este enfoque permite usar herramientas y métodos matemáticos para realizar cálculos en redes complejas.

En su definición es un conjunto de pares ordenados $G = (V, E)$ tal que $E \subseteq [V]^2$, así cada elemento de E está asociado a un par de elementos (el mismo o distinto) de V. Los elementos de V se llaman vértices (o nodos) de G, y los elementos de E se llaman aristas de G.(Diestel, 2005; Balakrishnan y Ranganathan, 2000)

1.2.5.1. Conceptos relacionados a los grafos

Se comentan y describen algunos términos acerca de grafos, siguiendo las definiciones de Diestel (Diestel, 2005):

Orden El número de vértices de un grafo, o su cardinalidad, es su **orden**, y se escribe como $|G|$; el número de aristas se denota con $||G||$. Los grafos son finitos o infinitos de acuerdo a su orden. Un **grafo vacio** de orden 0 o 1 se llama **trivial**. Los grafos que se estudian en esta tesis son todos finitos y no triviales.

Adyacencia Dos vértices u y v de G son **adyacentes** o **vecinos** si u y v son los vértices finales de una arista de G. Las dos aristas $e \neq f$ son **adyacentes** si tienen un vértice en común.

Grafo completo y triángulo Si todos los vértices de G son adyacentes entre si, entonces se dice que G es **completo**. Un **grafo completo** de n vértices se denota como K^n. K^3 se llama **triángulo**.

Vértices y aristas independientes Los vértices o aristas que no son adyacentes a ningún otro se llaman **independientes**

Grado El **grado** de un vértice v $|E(v)|$ es el número de aristas de v, es decir el número de vértices adyacentes a v. Un vértice de grado 0 es **independiente**.

[5] MeSH por Medical Subject Headings, es el vocabulario controlado Biblioteca Nacional de Medicina. Este vocabulario se usa para indexar artículos para la base de datos MEDLINE y PubMED. Cada cita de los artículos se asocia con un conjunto de palabras claves MeSH que describen el contenido de la cita.

Subgrafo Sea $G \cup G' := (V \cup V', E \cup E')$ y $G \cap G' := (V \cap V', E \cap E')$. Si $G \cap G' = 0$ entonces G y G' son **disjuntos**. Si $V' \subseteq V$ y $E' \subseteq E$, entonces G' es un **subgrafo** de G (y G es un **supergrafo** de G'), la notación es $G' \subseteq G$. Formalmente se dice que G contiene a G'.

Caminos y ciclos Un **camino** es un grafo no trivial $P = (V, E)$ de la forma

$$V = x_0, x_1, ..., x_k, E = x_0x_1, x_1x_2, ..., x_{k-1}x_k \tag{1.1}$$

donde las x_i son todas distintas. Los vértices x_0 y x_k están enlazados por P y se llaman vértices **terminales**. Los vértices $x_1, ..., x_{k-1}$ son los vértices **internos** de P. El número de aristas de un camino es su **longitud**.

Una notación usual para representar un camino es la secuencia de sus vértices, de tal forma que $P = x_0x_1...x_k$ se denota como camino P de x_0 a x_k (también entre x_0 y x_k).

Si $P = x_0...x_{k-1}$ es un camino y $k \geq 3$, entonces el grafo $C := P + x_{k-1}x_0$ se llama un **ciclo**. Se denomina ciclo por su (cíclica) secuencia de vértices, el ciclo anterior se escribe como $x_0...x_{k-1}x_0$.

Grafos dirigidos Un **grafo dirigido** es un par (V, E) de conjuntos disjuntos de vértices y aristas que juntos forman dos mapas (1) inicial: $E \to V$ y (2) terminal: $E \to V$, asignándole a cada arista e un **vértice inicial** $inicial(e)$ y un **vértice termina**l $terminal(e)$. Se dice que la arista e se dirige desde $inicial(e)$ a $terminal(e)$.

Grafo de Snomed CT Al diseccionar la estructura de Snomed CT, cada concepto se representa con un vértice y cada relación entre los conceptos se representa por una arista. No hay relaciones circulares y todas son unidireccionales sin excepciones, pero un vértice $inicial(e)$ puede tener más de una relación de salida o vértices $terminal(e)$, de esta manera se construye un grafo acíclico dirigido. (Bhattacharyya, 2016)

1.2.5.2. Patrones de grafos en redes

La mayoría de las redes de gran escala comparten patrones que no se notan en redes pequeñas. Entre todos los patrones, la mayoría de las características más conocidas son: **distribución libre de escala, efecto de mundo pequeño y fuertes estructuras de comunidad**.(Tang y Liu, 2010)

Distribución libre de escala (Tang y Liu, 2010). Los grados de los vértices en las redes de gran escala a menudo siguen una distribución de ley de potencia, también conocidas como distribución Zipfian o distribución Pareto.

En su definición, una variable aleatoria X sigue una distribución de ley de potencia si

$$p(x) = Cx^{-\alpha}, x \geq x_{min}; \alpha \geq 1 \tag{1.2}$$

De tal manera que $\alpha \geq 1$ asegura que exista una constante de normalización C. Una distribución que sigue la ley de potencia se llama también distribución sin escalas, ya que la forma de la distribución permanece sin cambios, excepto para una constante multiplicativa general, cuando la escala de unidades se incrementa por un factor. Esto es

$$p(ax) = bp(x) \tag{1.3}$$

Donde a y b son constantes. En otras palabras, no existe una escala característica con la variable aleatoria. La forma funcional es la misma para todas las escalas. La red con una distribución libre de escalas para los grados nodales también se denomina **red libre de escala**.

Mientras que para las distribuciones normales, es extremadamente raro que un evento ocurra con una desviación muy lejana de la media, en las distribuciones de ley de potencias la cola es mucho más larga. De tal manera, es común que algunos vértices de la red tengan alto grado mientras que la mayoría tengan pocas conexiones. La razón es que la cola de una distribución de ley de potencias decae polinomialmente. Esto es asintotáticamente más lento que en la distribución normal que lo hace exponencialmente, resultando en un fenómeno de cola pesada. La curva de la distribución de ley de potencias se convierte en una recta si graficamos la distribución en una escala log-log, ya que

$$\log p(x) = -\alpha \log x + \log C \tag{1.4}$$

Esta asociación se puede usar para verificar graficamente si la distribución sigue la ley de potencias. Una verificación más robusta es aproximar la función de distribución acumulativa (cdf) descripta con la siguiente ecuación:

$$F(X \geq x) \propto x^{-\alpha+1} \tag{1.5}$$

Métricas de comunidades Informalmente, una comunidad es un conjunto de vértices donde cada vértice es más cercano a otros vértices en su comunidad que a los vértices que están fuera de ella. Esta característica ha sido encontrado especialmente en las redes sociales (Tang y Liu, 2010). La figura 1.2 ilustra las comunidades que se forman a partir de las relaciones de los vértices, cada color representa una comunidad.

La fuerza de las comunidades se miden con diferentes métricas, en esta tesis se analiza el coeficiente de agrupamiento global o **transitividad**, el coeficiente de agrupamiento promedio y la longitud media del camino mínimo entre vértices.

Coeficiente de agrupamiento - Transitividad La transitividad (Scott, Wasserman, Faust, y Galaskiewicz, 1996) C^{Δ}, se calcula como una medida de la proporción de triángulos cerrados en un grafo:

$$C^{\Delta} = 3 \cdot \frac{\text{número de triángulos}}{\text{número de tripletas conectadas}} \tag{1.6}$$

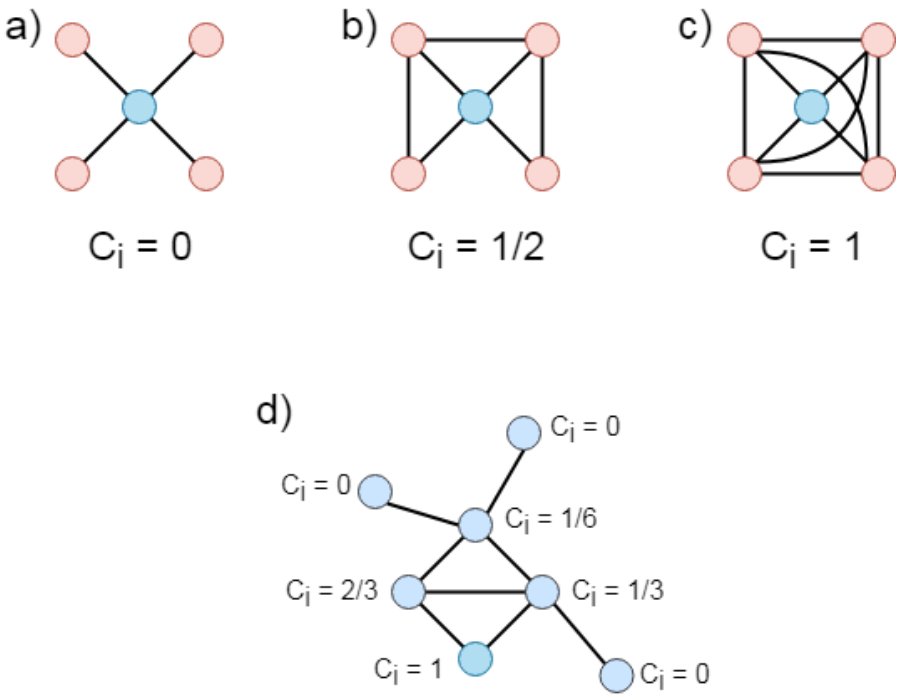
Esta métrica muestra de manera global cuán agrupado es el grafo, ya que representa la probabilidad de que dos vértices estén conectados si comparten un vecino en común. En el contexto de la lista de problemas, si los problemas $P1$, $P2$ co-ocurren en pacientes que también tengan $P3$, esta métrica representa la probabilidad de que $P1$ y $P2$ ocurran juntas.

Coeficiente de agrupamiento - Promedio (Saramäki, Kivelä, Onnela, Kaski, y Kertész, 2007; Kaiser, 2008) El promedio del coeficiente de agrupamiento del grafo G, necesita del cálculo local de los coeficientes de agrupamiento de cada vértice i. En el caso de los grafos dirigidos,

$$c_i = \frac{\Gamma_{v_i}}{|E(v_i)|(|E(v_i)| - 1)}, \tag{1.7}$$

Fig. 1.2: Ejemplo de comunidades

Fig. 1.3: Ejemplos de coeficientes de agrupamiento



donde $|E(v_i)|$ es el grado del vértice v_i y Γ_{v_i} es el número de aristas entre los vecinos de v_i.

Finalmente, el coeficiente de agrupamiento se calcula con la siguiente ecuación.

$$C = \frac{1}{n} \sum_{v \in G} c_v, \tag{1.8}$$

Donde n es el número de vértices evaluados.

El coeficiente de agrupamiento promedio produce una alta varianza para los vértices con menos grados. Por ejemplo, para vértices con grado 2, C_i es 0 o 1.(Tang y Liu, 2010)

La figura 1.3 muestra en los ejemplos a), b) y c) el cálculo del coeficiente de agrupamiento del vértice i en azul. En el ejemplo d) está el cálculo de cada nodo. Siguiendo las ecuaciones 1.8 y 1.6, el coeficiente de agrupamiento promedio es de 0.310 y la transitividad es de 0.375, respectivamente.

Longitud media del camino mínimo entre vértices Cálculo de la media de la distancia entre los vértices en un grafo. Comunidades con longitudes medias más pequeñas indican estructuras de comunidades más fuertes.

Distancia semántica La distancia semántica se usa para calcular similaridad entre conceptos. En diferentes trabajos en el dominio médico (Wang, Day, Visweswaran, y Hogan, 2010; Gan, Dou, y Jiang, 2013; Pedersen, Pakhomov, Patwardhan, y Chute, 2007; Zare y cols., 2015) se presentan diferentes maneras más sofisticadas de calcular la similaridad que la longitud media del camino mínimo. Sin embargo, estos enfoques son también alguna variante de la longitud media del camino mínimo. El modelo de datos define la

arista "|ES UN|"para establecer relaciones entre conceptos ancestros y descendientes, de tal manera que todos los conceptos son un subtipo de otro más general, y un concepto puede tener uno o más descendientes con varios niveles de especificación.

1.2.6. *Graphmining*

La minería de datos se ha convertido en sinónimo de encontrar patrones frecuentes en datos transaccionales, el término más general descubrimiento de conocimiento abarca esta y otras tareas. Descubrimiento o aprendizaje no supervisado en *graphmining* implica no sólo la tarea de encontrar patrones en un conjunto de transacciones, sino también encontrar la posible superposición de patrones en un gran grafo. Descubrimiento también implica la tarea de ***clustering***, la cual intenta describir todos los datos por medio de la identificación de clases de grafos que comparten patrones comunes de atributos y relaciones entre sí. *Clustering* también puede extraer relaciones entre *clusters*, resultando en una organización jerárquica o taxonómica.(Cook y Holder, 2006)

En contraste, el aprendizaje supervisado es la tarea de extraer patrones que distinguen un conjunto de otros. Estos conjuntos se llaman los ejemplos positivos y los negativos. Este conjunto de ejemplos pueden contener muchas transacciones de grafos o un sólo gran grafo. El objetivo es encontrar un patrón de relación entre nodos que aparezca a menudo en los ejemplos positivos pero no en los ejemplos negativos. Tal patrón se usa para predecir la clase (positiva o negativa) de nuevos ejemplos. (Cook y Holder, 2006)

La última tarea en graph mining es la visualización del conocimiento descubierto. La visualización de grafos es la representación de los vértices, aristas y etiquetas de un grafo de tal manera que los humanos entiendan los conceptos representados por el grafo.(Cook y Holder, 2006)

1.2.6.1. Análisis de redes

El análisis de redes involucra una variedad de tareas (Tang y Liu, 2010), a continuación listo las que aplicaré en este trabajo:

1. Análisis de centralidad, ayuda a identificar los vértices (o problemas en el contexto de esta tesis) "más importantes" en la red. Esta importancia puede definirse de distinta manera y cada una ayuda a entender diferentes aspecto de la influencia y el poder del vértice en la red:

    - **Centralidad de grado.** Cuenta el número de conexiones que tiene un vértice. Los problemas con los grados más altos serán los que más ocurren en los pacientes.
    - **Intermediación (*betweenness*).** Mide el número de veces que un vértice en particular es miembro de los caminos más cortos entre otros dos vértices. Usualmente, la medida de intermediación está normalizada en el rango de [0, 1], dónde 0 es un vértice desconectado y 1 un vértice por donde pasan todas las conexiones.(Cook y Holder, 2006; Brath y Jonker, 2015)
    - **Cercanía (*Closeness*).** La cercanía de un vértice mide la distancia promedio a todos los otros vértices. Si pocos vértices son alcanzables o si la distancia entre los vértices aumenta, entonces la medida de cercanía es pequeña. (Cook y Holder, 2006; Brath y Jonker, 2015).

- **Centralidad de autovector (*Eigenvector*).** Dado un vértice, se suma recursivamente todas las distancias a todos los otros vértices. Entre más centrales sean los otros vértices, el vértice que se está midiendo tendrá mayor centralidad(Brath y Jonker, 2015).

2. Clasificación de redes y detección de *outliers*. Algunos problemas están etiquetados con información contextual: servicio de salud, ámbito y grupo etario. Por ejemplo, en una red con algunos problemas pocos comunes identificados como del Servicio Cardiológico, ¿es posible inferir otros problemas poco comunes asociados por su co-ocurrencia en los pacientes?

3. *Clustering*. Según Blonde et al.(Blondel, Guillaume, Lambiotte, y Lefebvre, 2008), el problema de *clustering* requiere la partición de una red en grupos de vértices densamente conectados, donde los vértices que pertenecen a diferentes grupos se conectan escasamente. Las formulaciones exactas de este problema de optimización son computacionalmente intratables. En los últimos años se han propuesto muchos algoritmos para encontrar particiones razonablemente buenas de una manera razonablemente rápida, debido al incremento de la disponibilidad de conjuntos de datos con grandes redes y el impacto de las redes en la vida. Se pueden distinguir los siguientes tipos de algoritmos :

   - **Divisivos:** Se comienza con todo el grafo y se van eliminando iterativamente las aristas, dividiendo así progresivamente el grafo en subgrafos desconectados más y más pequeños. Estos subgrafos se identifican como comunidades. El punto crucial en un algoritmo divisivo es la selección de las aristas que se van a eliminar, las cuales deben ser las que conectan comunidades y no las aristas que están dentro de las comunidades. (Radicchi, Castellano, Cecconi, Loreto, y Parisi, 2004; Girvan, Newman, Cecconi, Loreto, y Parisi, 2002; Newman, 2004)
   - **Aglomerativos:** Para cada uno de los vértices del grafo se calcula un peso, el cual mide qué tan conectados están los vértices. A partir del conjunto de todos los vértices sin aristas, se ordenan decrecientemente los vértices según su pesos, y se agregan iterativamente las aristas entre pares similares de vértices. De esta manera, los vértices se agrupan en comunidades cada vez más grandes, y el árbol se construye hasta la raíz. Este árbol representa a todo el grafo. (Radicchi y cols., 2004; Pons y Latapy, 2005)
   - **Optimización:** Estos algoritmos buscan encontrar resultados usando heurísticas o maximizando una función objetivo con un enfoque voraz (*greedy*) para mejorar la complejidad computacional de los algoritmos clásicos. Estos métodos consisten en unir recursivamente las comunidades buscando optimizar la función objetivo.(Clauset, Newman, y Moore, 2004; Blondel y cols., 2008)

1.2.6.2. Evaluación de estructuras

La detección de comunidades es el tópico que se desarrolla en mayor profundidad en esta tesis. Se usan los tres enfoques principales y se comparan los resultados. En parte, la razón por la cual hay tantas definiciones y métodos, es que no está clara cómo debe ser la estructura de una comunidad en una red real. De tal manera que diferentes métodos se desarrollaron según las necesidades de sus autores (Tang y Liu, 2010).

Para poder comparar los diferentes resultados obtenidos con los algoritmos, utilizaré las siguientes estrategias:

- Cálculo de modularidad
- Comparación de las comunidades con un *gold-standard*

Cálculo de modularidad. Según Newman (Newman, 2006), una buena división de una red en comunidades no es solamente aquella en la que el número de aristas entre los grupos es pequeño, sino en la que el número de aristas entre los grupos es menor al esperado. Sólo si el número de aristas entre los grupos es significativamente más bajo al esperado se puede decir con justificación que se ha encontrado una estructura de comunidad significativa. Equivalentemente, se puede examinar el número de aristas dentro de las comunidades y buscar divisiones de las redes en las cuales este número es más alto del esperado, los dos enfoques son equivalentes. Estas consideraciones constituyen una medida basada en un punto de corte para modificar una función de beneficio Q definida como

$$Q = (\text{Número de aristas en la comunidad}) - (\text{Número de aristas esperadas}) \tag{1.9}$$

Esta función de beneficio se llama **modularidad**. La modularidad de una partición es un valor escalar entre -1 y 1. La modularidad mide la densidad de los enlaces en las comunidades comparándola con los enlaces entre las comunidades(Blondel y cols., 2008). Valores cercanos a 1 indican estructuras de comunidades más fuertes, y valores cercanos a -1 indican que los nodos están agrupados en comunidades a las que no pertenecen(Tang y Liu, 2010).

Comparación de las comunidades con un gold-standard. El término *gold-standard* (Versi, 1992) se utiliza en los trabajos científicos para referirse a valores de referencia que están disponible en condiciones razonables. No es la prueba perfecta, sino la mejor disponible. El *gold-standard* es importante ante la imposibilidad de realizar mediciones directas para determinar si los conceptos que están dentro de una comunidad realmente pertenecen o no a ella. Los *gold-standard* con el que se realizan las mediciones son los públicos y disponibles por la organización Kaiser Permanente (Ver sección 1.2.2.3).

La medición que se realizó se llama cubrimiento, la cual se define como la proporción de los conceptos que se comparten entre los dos conjuntos de datos, al tamaño del *refset* (Ver sección 1.2.2.3) de referencia T, como se define en la ecuación 1.10

$$\text{cubrimiento} = \frac{\text{conceptos mapeados en T}}{\text{Tamaño de T}} \tag{1.10}$$

- *Conceptos mapeados en T*, es el número total de conceptos de Snomed CT que comparten el *refset* que se compara y el *refset T*, y
- el *tamaño de T* es el número total de conceptos que tiene el *refset T*.

1.3. Objetivos

Con el presente trabajo se contribuye a la organización de los datos y el uso significativo de Snomed CT para la recuperación de información en la lista de problemas del Hospital

Italiano de Buenos Aires. Esta contribución es la construcción de una taxonomía de problemas, cuyas relaciones son ponderadas por su co-ocurrencia en las listas de problemas de los pacientes y sus relaciones jerárquicas con Snomed CT.

El producto final de este trabajo es el modelo que representa grupos o clasificaciones y las relaciones significativas de los problemas que co-ocurren en los pacientes. Los experimentos para llegar a este modelo tendrán en cuenta la lista de problemas y se añadirá información de contexto: ámbito o nivel de asistencia, área jerárquica y grupo etario. La comparación entre los modelos y la evaluación de su capacidad predictiva se hará con las medidas de precisión y exactitud.

1.4. Organización del trabajo

En el capítulo 2 se explica la metodología, se describen los algoritmos y los criterios de evaluación.

En el capítulo 3 hago una descripción de los datos y las decisiones en la limpieza de datos. Luego, se aplican las técnicas de *graphmining*.

Por último en el capítulo 4, hablo sobre las conclusiones y futuros pasos.

2. CAPÍTULO: METODOLOGÍA

2.1. Introducción

Tomando como base la metodología *CRoss Industry Standard Process for Data Mining (CRISP-DM)* divido el trabajo en fases, describo los procesos que se realizan en cada una de ellas y los entregables que servirán de insumo para la siguiente fase.

En las primeras fases estudio el problema y los datos. La siguiente fase, que es la preparación de los datos genero el conjunto de entrenamiento con los registros previos al año 2017, y el conjunto de validación con los registros del año 2017. En las fases de diseño y evaluación, ejecuto los algoritmos de aprendizaje no supervisado con el fin de encontrar grupos significativos con alto valor de precisión y exactitud. En la fase final se organizan los datos y redacto el documento final.

2.2. Algoritmos de aprendizaje no supervisado

El paquete igraph(Csardi y Nepusz, 2006) de Python ofrece una variedad de algoritmos para realizar aprendizaje no supervisado. Sin embargo, por el tamaño del grafo algunos algoritmos son imposibles de usar en la práctica. Por ejemplo, el método *leading edge betweenness* para detectar comunidades tiene una complejidad $O(|E||V|^2)$ en el peor caso (donde $|V|$ número de vértices y $|E|$ número de aristas). Para procesar el grafo de la red semántica de problemas se necesitaría en el peor de los casos más de 181.93 años. Los detalles de otros algoritmos para la generación de comunidades del paquete igraph se encuentran en la tabla 2.1.

Seleccioné los siguientes algoritmos para realizar todos los experimentos de aprendizaje no supervisados, teniendo en consideración la complejidad computacional y escogiendo un representante de cada uno de los tipo de algoritmos mencionados en el capítulo anterior (divisivos, aglomerativos y optimización) (Ver sección 3).

2.2.1. Algoritmo: *leading eigenvector*

(Newman, 2006)

- Complejidad: $c|V|^2 + |E|$
- Tipo de algoritmo: Optimización

El algoritmo *Leading Eigenvector* implementado por igraph usa un algoritmo recursivo para detectar estructura de comunidades. Este algoritmo divide la red maximizando la modularidad respecto a la red original.

Los métodos basados en este enfoque han tenido excelentes resultados en test estandarizados. Desafortunadamente, la optimización exhaustiva de la **modularidad** requiere grandes esfuerzos computacionales, incluyendo algoritmos voraces (*greedy*), enfriamiento simulado (*simulated annealing*) y optimización extrema (*EO*). El algoritmo desarrollado por Newman tiene un enfoque diferente. Se reescribe la función de **modularidad** en términos de una matriz, lo cual permite expresar la tarea de optimización como un problema

Tab. 2.1: Algoritmos de aprendizaje no supervisado del paquete igraph

Algoritmo	Observaciones	Complejidad[1]	Cálculo de tiempo de ejecución[2]
spinglass	Comunidades basadas en la física estadística	No especificado	
leading eigenvector	Comunidades basadas en matrices de autovectores	$O(\|E\|+\|V\|^2 * steps)$, donde "steps" es atributo del algoritmo	6.86 segundos
walktrap	Comunidades basados en caminos aleatorios	$O(\|E\|\|V\|^2)$ en el peor caso, $O(\|V\|^2 log\|V\|)$ en promedio,	181.93 años en el peor caso, 38 segundos en el caso promedio
edge betweenness	Comunidades basados en la métrica de centralidad intermediación	$O(\|E\|\|V\|^2)$ en el peor caso	181.93 años en el peor caso
fast greedy	Comunidades basadas en optimización de la modularidad	$O(\|E\|\|V\|log\|V\|)$ en el peor caso, $O(\|E\|+\|V\|log^2\|V\|)$ en promedio	22.84 minutos en el peor caso, 0.0011 segundos en promedio
multilevel	Comunidades mediante la optimización de la modularidad multinivel	En promedio lineal en grafos dispersos	0.00064 segundos en promedio
label propagation	Comunidades basadas en la propagación de etiquetas	$O(\|E\|+\|V\|)$	Depende del tamaño de la matriz resultante.
infomap	Estructuras de comunidades que minimizan el valor esperado	No especificado	

[1] donde $|V|$ número de vértices, $|E|$ número de aristas.
[2] Cálculos basados en números de instrucciones por segundos en un procesador AMD Athlon FX-60 (Dual Core) Reloj 2.6 GHZ = 22150 MIPS.

espectral en el álgebra lineal. Este enfoque lidera una familia de algoritmos rápidos para detectar comunidades que producen resultados que compiten con los mejores métodos previos a él.

2.2.2. Algoritmo: *multilevel*

(Blondel y cols., 2008)

- Complejidad: "lineal" cuando $|V|$ es aproximadamente igual a $|E|$, el algoritmo sugiere que podría ser cuadrática con los grafos completamente conectados.
- Tipo de algoritmo: Aglomerativo

El tamaño típico de las grandes redes se cuenta en millones cuando no miles de millones de vértices. Ejemplos de estas redes son las sociales, las telefónicas móviles o la web. En esta escala se demanda nuevos métodos para recuperar información a partir de su estructura. Un enfoque prometedor consiste en la descomposición de redes en subunidades o comunidades. La identificación de estas comunidades es de crucial importancia ya que pueden ayudar a descubrir módulos funcionales desconocidos a priori, tales como tópicos en redes de información o ciber comunidades en redes sociales. Además, las meta redes resultantes, cuyos vértices son comunidades, pueden ser usadas para visualizar la estructura de la red original.

El algoritmo **multilevel** encuentra particiones con alta **modularidad** en grandes redes en poco tiempo y despliega una completa estructura de comunidad jerárquica para una red, dando así acceso a diferentes resoluciones de una detección de comunidades. El algoritmo se divide en dos fases que se repiten iterativamente. Asume que se empieza con una red de N vértices, y cada uno de los vértices tiene un peso. Primero, se asigna una comunidad diferente a cada vértice de la red. De esta manera, en esta partición inicial hay tantas comunidades como vértices. Entonces, por cada vértice i se evalúa si mejora la **modularidad** al cambiar la comunidad asociada al vértice i por las comunidades en las que están cada uno de vértices adyacentes a él. El vértice i es entonces puesto en la comunidad en la cual la modularidad es máxima y positiva. Si no se obtiene una ganancia positiva, entonces el vértice i se mantiene en su comunidad original. Este proceso se aplica repetidamente y de manera secuencial para todos los vértices hasta que no se puede lograr ninguna mejora.

2.2.3. Algoritmo: *label propagation*

(Raghavan, Albert, y Kumara, 2007)

- Complejidad: $|V| + |E|$
- Tipo de algoritmo: Divisivo

Cada vértice tiene una etiqueta inicial. En cada iteración del algoritmo se usa una función uniforme para determinar las aristas que son eliminadas, en el siguiente paso cada vértice adopta la etiqueta que es mayoritaria en los vértices adyacentes. A medida que las etiquetas se propagan a través de la red, grupos de nodos densamente conectados constituyen un consenso en sus etiquetas.

Hay dos condición posibles de parada. En la primera puede ser que se llegue a la convergencia, es decir que la mayoría de los vecinos de cada vértice tengan la misma etiqueta que dicho vértice. La segunda es que se fija la cantidad máxima de iteraciones.

Al final del algoritmo, los nodos con la misma etiqueta son conectados como comunidad.

Las ventajas de este algoritmo sobre otros métodos es su simplicidad y su efectividad en el tiempo. El algoritmo usa la estructura de la red para guiar su progreso y no optimiza alguna métrica específica. Además, el número de comunidades y sus tamaños no se conocen a priori y se determinan cuando finaliza el algoritmo.

Aunque las redes con una sola comunidad satisfacen el criterio de parada, romper los enlaces aleatoriamente permite que se generen subgrafos disjuntos, evitando así que la misma etiqueta se propague por todo el grafo. En el caso de redes homogéneas que no tienen estructura de comunidades, el algoritmo identifica la red como una sola comunidad.

2.3. Metodología CRISP-DM

En el desarrollo de proyectos de minería de datos y descubrimiento de conocimiento, la metodología que más se usa es *CRISP-DM*. CRISP-DM es una metodología independiente de dominio, por lo que puede utilizarse con cualquier herramienta de minería de datos y se puede aplicar para resolver cualquier problema de minería de datos.(Marbn, Mariscal, y Segovi, 2009)

Este trabajo se divide en seis fases (ver Figura 2.1). Todas las fases usan transversalmente el mismo software: Oracle para el modelo de datos ER, Neo4J[1] para el modelo de datos de grafos, Python y Java para la limpieza, selección de datos y ejecución de modelos, y las librerías de javascript D3.js y linkurious.js para la visualización de los datos. A continuación describo lo que contempla cada fases de esta tesis.

2.3.1. Fase 1: Entendimiento del Dominio.

Esta fase fue previa al trabajo de tesis en sí, y el entregable fue la propuesta de tesis. Se compone de la determinación de los objetivos, de las metas de minería de datos y la evaluación de los recursos.

Como está consignado en el primer capítulo introductorio, el objetivo es la construcción de una taxonomía de problemas, cuyas relaciones están ponderadas por su co-ocurrencia en las listas de problemas de los pacientes y sus relaciones jerárquicas con Snomed CT. Para alcanzar dicho objetivo, establecí que la meta de minería de datos es representar el conocimiento en un enfoque de red, y a partir del análisis de la red realizar la construcción jerárquica de grupos o clasificaciones de los problemas.

Para realizar esta tesis, el HIBA proporcionó el acceso a los datos de la lista de problemas y la terminología Snomed CT. En términos de hardware, el HIBA proporcionó el acceso a un servidor con las siguientes características:

- Sistema Operativo: Ubuntu 16.04.5 LTS (GNU/Linux 4.4.0-138-generic x86_64)
- Memoria RAM: 64 GB

[1] Neo4j (`https://neo4j.com/`) es un proyecto de código abierto que permite implementar el modelo de bases de datos de grafos. Es la solución empresarial que más se usa(Solid IT, 2019), combina la fortaleza del almacenamiento nativo de grafos y una arquitectura escalable y optimizada para asegurar un buen rendimiento en las consultas basadas en relaciones

- Espacio de almacenamiento: 760 GB
- Número de núcleos: 8 núcleos

2.3.2. Fase 2: Entendimiento de los datos.

Esta fase inicia con la recolección de los datos y procede con las actividades que permitan familiarizarse con ellos. El entregable se encuentra en la primera sección del capítulo 3 y contempla un análisis descriptivo para determinar la calidad de los datos de la lista de problemas, la existencia de valores atípicos (*outliers*) que deben ser eliminados de los datos de entrenamiento y validación, y la construcción de contextos interesantes que permitan la formulación de hipótesis.

La construcción de *refset* para agrupar problemas por contextos tiene dos metodologías: (1) determinar por extensión cuáles son los conceptos que los componen. Ya que no es fácil hacerlo por comprensión usando las relaciones subtipo de Snomed CT. (Højen, Sundvall, y Gøeg, 2014; Lee, Cornet, Lau, y de Keizer, 2013); o (2) El uso de *refset* públicos para comparar el cubrimiento con los *refset* generados aquí (Lee y cols., 2013). Esta tesis usa la segunda metodología. Se usa aprendizaje no supervisado para generar grupos con los registros en la lista de problemas y los contextos en los cuales se realizaron los registros. Si hay solapamiento de conceptos entre los contextos, se fusionan los contextos. Finalmente si hay *refset* similares en Kaiser Permanente CMT entonces se evalúa el cubrimiento con ellos (Ver sección 1.2.6.2). Posteriormente las decisiones finales sobre fusión de contextos son evaluados manualmente por la residencia de informática médica.

Los contextos nivel asistencial y grupo etário tienen una cardinalidad pequeña, 8 y 9 respectivamente. Por el contrario, las áreas jerárquicas tienen una cardinalidad grande, inicialmente son 601 áreas jerárquicas. Las áreas jerárquicas tienen un proceso extra para evaluar sus solapamientos y así agruparlas dentro de las más significativas. Este último proceso se realiza con el algoritmo de aprendizaje supervisado StanfordNLP ColumnClassifier (Manning y cols., 2014). La variable objetivo del algoritmo de clasificación es el servicio al que pertenecen los conceptos, un algoritmo de clasificación tendría mayor dificultad para predecir el servicio si hay muchos conceptos similares entre ellos, por lo tanto serían candidatos a fusión. La dificultad se mide con valores de *F1 score*[2] bajos. La residencia de informática médica determina manualmente si concuerda con la fusión de los contextos.

2.3.3. Fase 3: Preparación de los datos.

Esta fase se compone de la extracción de los datos desde el modelo ER, limpieza y transformación de los datos con los que se realizan los modelos. El entregable es el conjunto de datos de entrenamiento y validación. El conjunto de entrenamiento contiene los registros de problemas previos al 2017. El conjunto de validación contiene los registros que ingresaron a la HCE en el 2017 y cuyos pacientes tengan una lista de problemas no vacía.

2.3.4. Fase 4 y 5: Diseño y evaluación.

Esta fase se compone de las siguientes tareas:

[2] *F1-Score* es un valor único que pondera la precisión y la exhaustividad. *F1 score* hace referencia a la efectividad de un modelo y es conocida en la estadística como una proporción de acuerdo específico ya que se aplica a una clase específica, la clase positiva.(Powers, 2011)

- Construcción del grafo: Determinación del modelo de datos y carga de la base de datos de grafos Neo4j. Usando los datos de entrenamiento se obtiene un grafo cuyos vértices son los problemas y las aristas son las co-ocurrencias en los pacientes (Red de Problemas) y las relaciones dadas por la terminología de referencia Snomed CT (Red semántica de problemas). Los vértices son etiquetados con información contextual: ámbito o nivel de asistencia, grupo etario y área jerárquica.

- Caracterización de las redes y búsqueda de patrones en las redes:

  - Determinar si la distribución de la red es libre de escala: La hipótesis nula es que los grados de los vértices se distribuyen según la ley de potencias. Se estima la función de distribución acumulativa cdf, en el caso de que el p-valor de la función no permita rechazar la hipótesis nula entonces se compara el estadístico *log-likelihood ratio* con otras distribuciones.
  - Análisis de estructuras de comunidad: Métricas para comparar cuantitativamente los grafos de la red de problemas y la red semántica de problemas. Se evalúa si los grafos presenta evidencias de tener comunidades: coeficientes de agrupamiento, longitud media del camino mínimo entre nodos y promedio de grados. El cálculo de estas métricas están disponibles en los paquetes igraph(Csardi y Nepusz, 2006) y networkx(Hagberg, Schult, y Swart, 2008) de python.

- Aprendizaje no supervisado: Aplicación de algoritmos de aprendizaje no supervisado a la red de problemas y la red semántica de problemas. Por las limitaciones en hardware no se puede procesar el grafo completo de la red semántica de problemas, se realizan las siguientes tareas con sub-grafos de problemas que co-ocurran en 10 000, 1000, 100, 10 y 5 pacientes:

  - Identificación de grupos o *clusters* significativos que comparten patrones comunes de atributos y relaciones.
  - Cada uno de los vértices que contiene el *cluster* tiene calculada las medidas de centralidad para entender su influencia dentro de la red.
  - A los grupos encontrados se les realiza un test para evaluar la significancia de los *clusters*, los siguientes pasos son realizados 1000 veces:
    - se generan aleatoriamente aristas en el grafo
    - se aplican los algoritmos de aprendizaje no supervisado
    - se evalúa si la modularidad es superior a la modularidad del grafo original.

- Validación de los resultados: Teniendo en cuenta las directrices de recuperación de información (Manning, Raghavan, y Schuütze, 2008; Hersh, Buckley, Leone, y Hickam, 1994), el conjunto de validación tiene tres componentes:

  - **El conjunto de pruebas**: se conforma de los problemas registrados previamente y los valores del contexto (nivel asistencia, grupo etario y área jerárquica), que representan la consulta, y el problema a predecir que representa la respuesta correcta

- **Los conceptos de Snomed CT a ser recuperados**: la lista previa de problemas permite identificar los subgrafos, y los valores de contexto permiten filtrar los vértices. De esta manera, creo una **lista de sugerencias** por cada uno de los registros del conjunto de pruebas.
- una medida de relevancia por cada par de consulta-concepto recuperado, esta medida es la **transitividad** o la **distancia semántica**, y permite un ordenamiento de los resultados. Estas medidas están definidas en el capítulo introductorio (Ver sección 1.2.5.2).

Lo que se evalúa es si los vértices que pertenecen al mismo *cluster* permiten completar los problemas faltantes en la lista de los problemas de los pacientes del conjunto de datos de validación. Esta evaluación se realiza con la lista de problemas previas al 2017 y los contextos para seleccionar y filtrar la **lista de sugerencias**, y predecir los problemas registrados en el 2017.

Se evalúan por separado la capacidad predictiva (precisión y exactitud) de los grupos encontrados a partir del grafo de la red de problemas y del grafo de la red semántica de problemas. También son evaluadas las capacidades predictivas filtrando el contexto: ámbito o nivel de asistencia, grupo etario y área jerárquica, sólo en red de problemas. La precisión y la exactitud (*accuracy*) se definen como:

$$Exactitud = \frac{\text{Verdaderos positivos} + \text{Verdaderos negativos}}{\text{Total de los datos}} \tag{2.1}$$

$$Precision = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos positivos}} \tag{2.2}$$

La exactitud suma los verdaderos positivos, que son los casos en los que la lista de sugerencia contiene la respuesta correcta, y los verdaderos negativos, que son los casos en los que la lista de sugerencias esta vacía y el problema a predecir no está dentro de los conceptos de Snomed CT a ser recuperados, y los divide en la totalidad de los datos de prueba. Se evalúa también una versión más flexible de verdadero positivo, donde se define una distancia tolerable de longitud media del camino mínimo entre la respuesta correcta y alguno de los conceptos de la lista de sugerencias. Esta distancia se estableció de tamaño 3, e indica similitudes entre un concepto y los descendientes que están hasta dos vértices de distancia, o conceptos descendientes del mismo padre, como se ilustra en la figura 2.2.

La precisión evalúa cuántas veces la lista de sugerencias no vacía tiene la respuesta correcta, dividido las veces que se obtuvo una respuesta positiva. Teniendo en cuenta que las listas de sugerencias son de un tamaño n ordenados según la relevancia, las medidas son llamadas precisión al n o $P@n$ y exactitud al n o $Acc@n$.

2.3.5. Fase 6: Implementación.

En la última fase organizo los datos y redacto el documento final de la tesis. El plan de implementación en la organización de los datos y los servicios de terminología del HIBA no hacen parte de los entregables de esta tesis.

Fig. 2.1: Metodología del trabajo de tesis

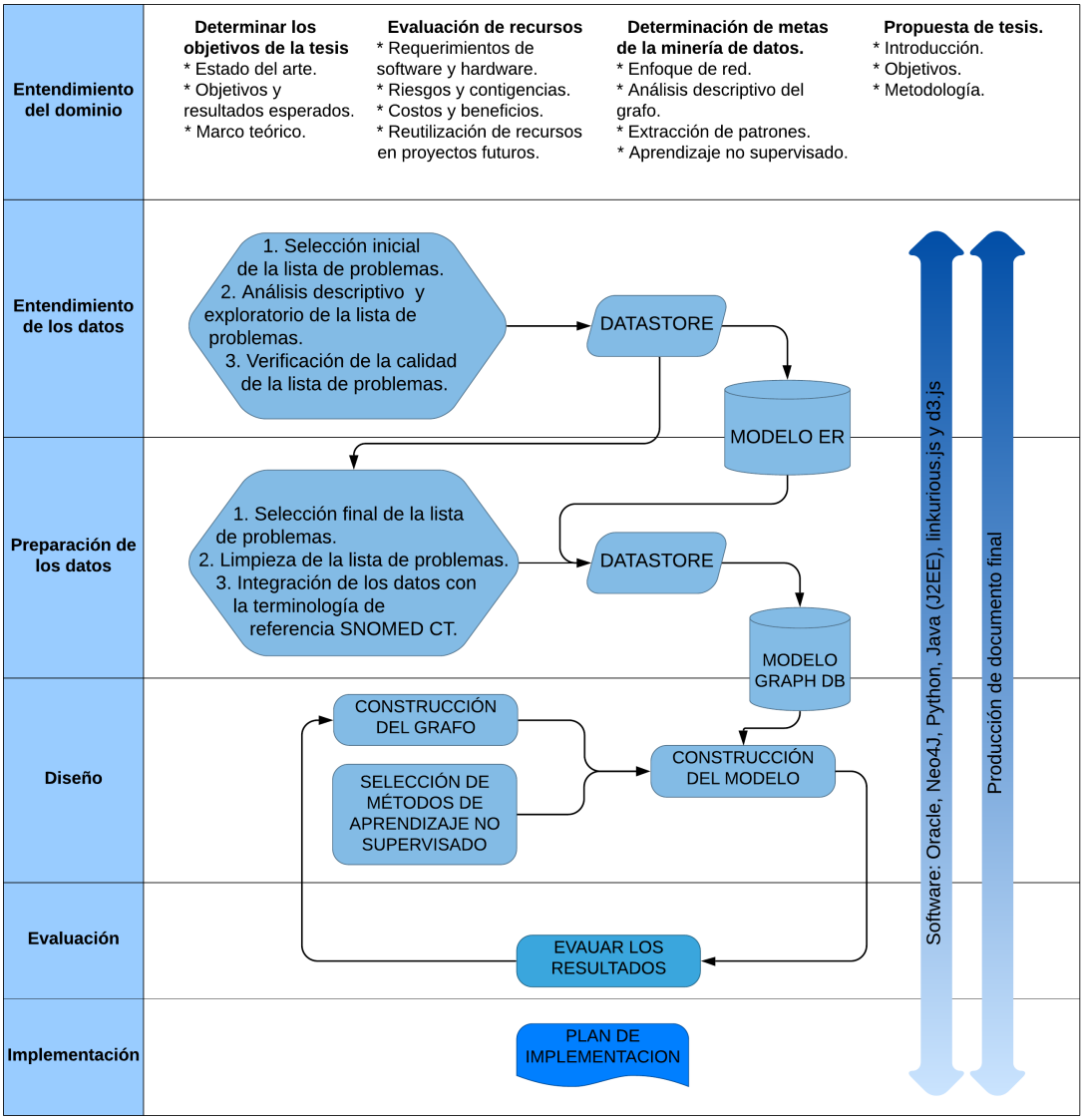
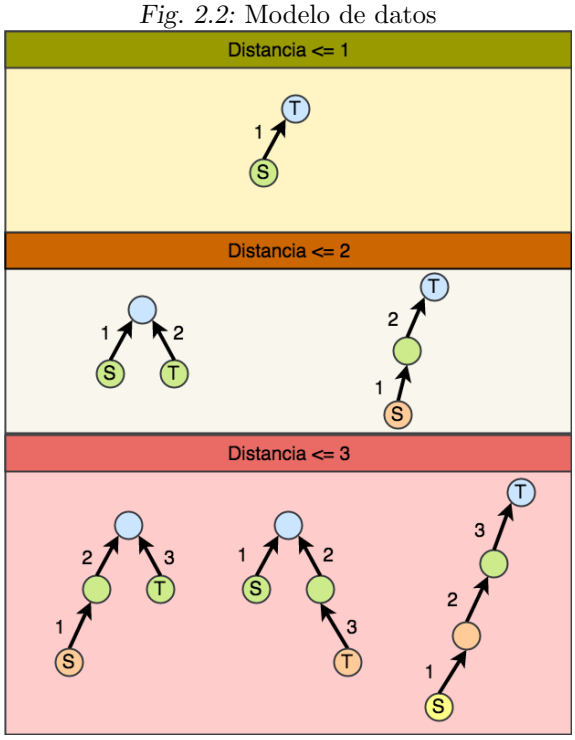


Fig. 2.2: Modelo de datos



3. CAPÍTULO: RESULTADOS DE LA DESCRIPCIÓN Y LIMPIEZA DE DATOS

3.1. Introducción

Dentro de las fases para el desarrollo de un proyecto de minería de datos, después del entendimiento del dominio, se encuentra el entendimiento de los datos. Este capítulo se enmarca en el entendimiento de los datos y describe los resultados producto de las dos tareas siguientes: (1) descripción de la lista de problemas desde cada una de las dimensiones de estudio (paciente, fecha de registro, problema cargado, área jerárquica del especialista, grupo etario del paciente y nivel de asistencia) y (2) construcción de grupos de problemas según el contexto.

La primera tarea permite determinar la calidad de los datos y definir los criterios para la construcción del conjunto de entrenamiento y validación. En la segunda tarea se crean los *refset* que pueden ser utilizados como vocabularios controlados dependientes de contextos. Estas tareas tienen el objetivo de reducir la cardinalidad de algunas de las dimensiones mencionadas y permitir la interoperabilidad de los datos y de esta manera facilitar su posterior comparación con los realizados en otras instituciones.

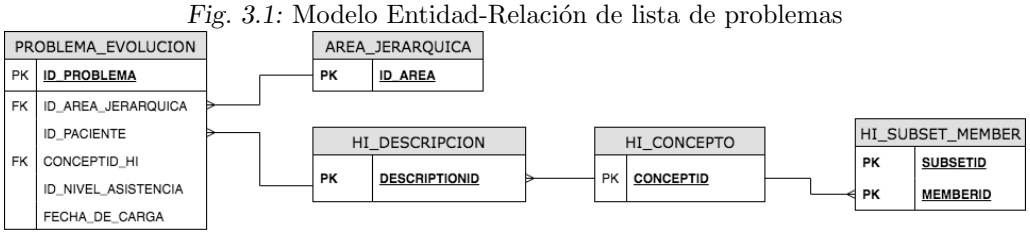
3.2. Comprensión de datos

La lista de problemas registra las observaciones y hallazgos realizados a los pacientes y otra información del contexto. Los problemas de los pacientes son descripciones que se agrupan en conceptos, y los conceptos a su vez pertenecen a uno o varios *refset*.

Cualquier usuario con acceso a la HCE puede registrar un problema, este usuario puede pertenecer a una área administrativa o asistencial del HIBA. Algunas de las áreas tienen sub-áreas, presentando así una estructura jerárquica. Cuando se registra un problema, se asocia el área jerárquica del que registra.

Además cada registro tiene el atributo nivel de asistencia que indica el ámbito en el que fue cargado el problema (1 = Ambulatorio , 2 = Internacion General, 3 = Guardia, 4 = Triage, 5 = Internacion Geriatrica, 6 = Internacion Domiciliaria, 7 = Seguimiento Domiciliario, 8 = Episodio Externo, 9 = Episodio Ambulatorio). La figura 3.1 contiene el modelo entidad relación, donde se muestra cómo se relacionan estos datos.

Fig. 3.1: Modelo Entidad-Relación de lista de problemas



A continuación describo la distribución de los problemas para cada una de las dimensiones de la entidad PROBLEMA_EVOLUCION.

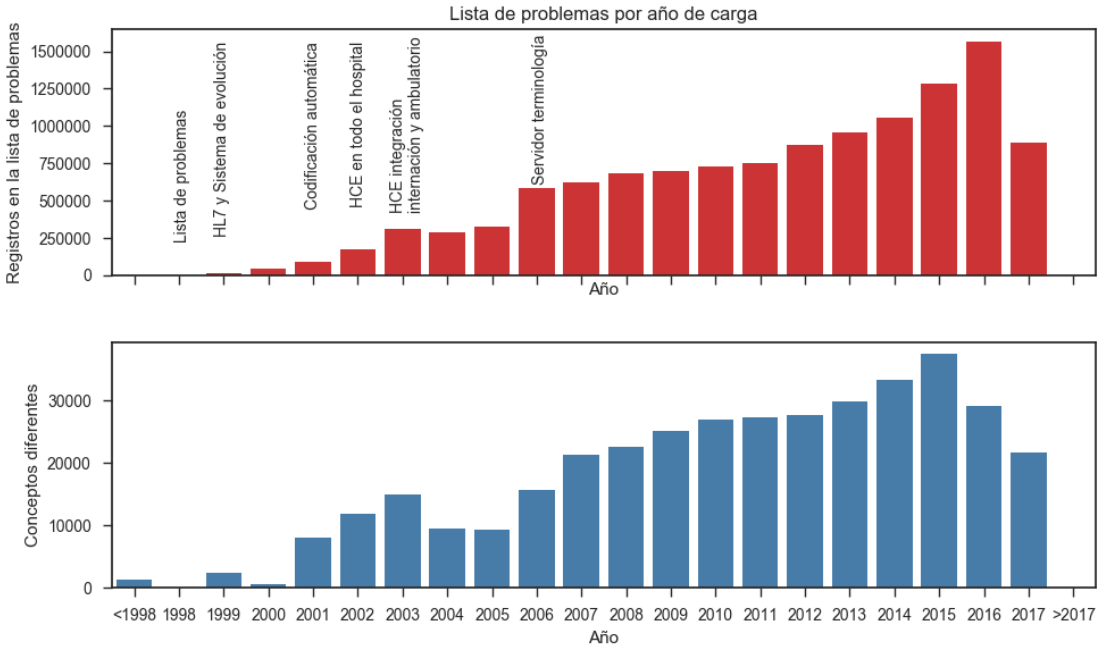
3.2.1. Distribución en el tiempo

El atributo FECHA_DE_CARGA indica la fecha en la que se cargó el problema por primera vez. Para las anotaciones posteriores se vuelve a usar el problema registrado y no se crea uno nuevo. En la figura 3.2, se puede observar que aunque la implementación de la HCE empezó en el año 1998, hay registros anteriores a esa fecha, los cuales se interpretan como errores en el momento de la carga, al igual que los registros con las fechas superiores al 2017. En total son 9149 casos que corresponden al 0.004 % de los datos.

Como se muestra en la figura 3.2 la distribución no es uniforme, su crecimiento se explica por los hitos de implementación dentro de la HC, en el año 2002 se generalizó el uso de la HCE en todo el hospital, y en el año 2006 se implementó el servidor de terminología. Los años 2015 y 2016 muestran un incremento en el registro de la lista de problemas.

En cuanto a los conceptos diferentes registrados por año, se observa una distribución más uniforme, especialmente después del año 2007. Los datos del año 2017 son sólo los del primer semestre.

Fig. 3.2: Distribución de problemas por año de carga



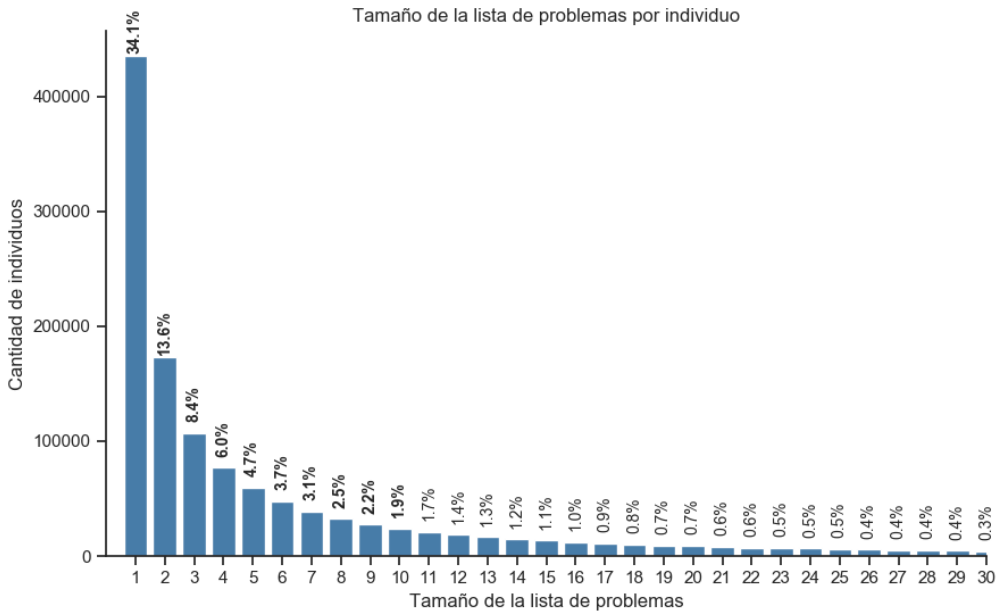
3.2.2. Distribución por paciente (individuo)

Según la figura 3.3 la mayoría de los individuos tiene sólo un problema registrado en su lista de problemas, estos corresponde al 34.1 % de los datos. Los individuos que tienen registrados hasta 10 problemas en la lista son el 80 % de los datos.

3.2.3. Distribución por conceptos

El conjunto de datos contiene 88 869 conceptos distintos usados para identificar problemas de los pacientes. La figura 3.4 representa el uso de estos conceptos en todos los

Fig. 3.3: Distribución del tamaño de la lista de problemas por individuos



registros de la lista de problemas, en ella se observa que un conjunto pequeño de problemas tiene una frecuencia muy alta y el resto que queda en la cola de la distribución fueron usados muy pocas veces. Los conceptos con más frecuencia de uso son Control de salud, Fiebre, Dolor abdominal, Hipertensión arterial, Catarro de las vías áreas superiores, Malestar general, Tos, Embarazo, Lumbalgia, Broncoespasmo, Evaluación inicial del paciente e Infección del tracto urinario. Juntos estos trastornos representan el 20 % del total de registros.

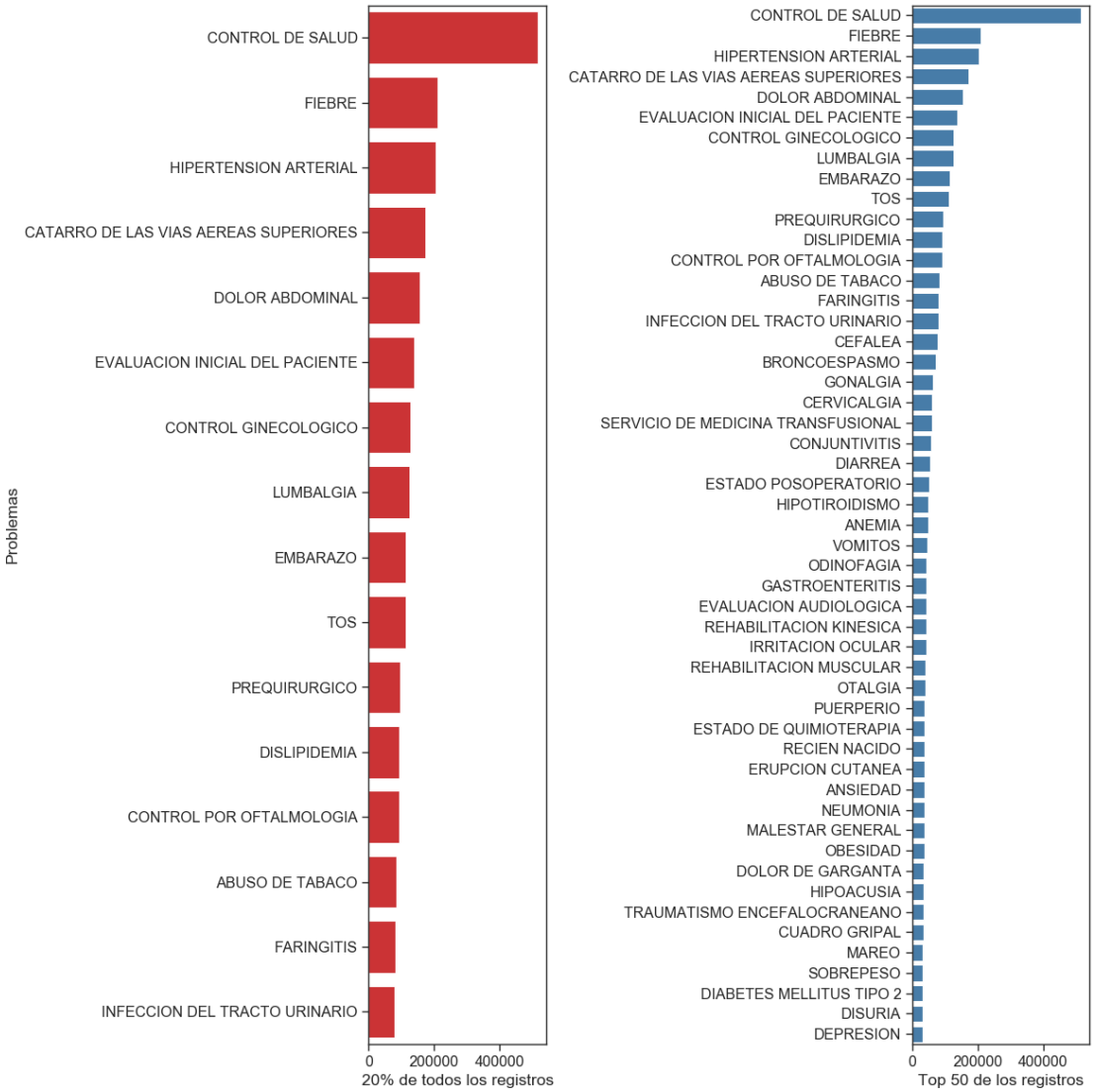
3.2.4. Distribución por contextos

Los contextos que tuve en cuenta en esta tesis fueron área jerárquica, nivel de asistencia y grupo etario, a continuación describo cómo se distribuyen los conceptos y registros de la lista de problemas en los contextos.

3.2.4.1. Contexto: Nivel de asistencia

Según la tabla 3.1 se puede observar que en cuatro niveles de asistencia (Ambulatorio, Guardia, Triage e Internación general (Ver sección 1.2.3)) están el 97 % del total de los registros de la lista de problemas. De estos cuatro niveles de asistencia, en el caso de Ambulatoria por cada 81 registros hay un concepto de Snomed CT diferente asociado a la lista de problemas, en la guardia esta relación es de 105:1. Triage es la relación más baja ya que es 188:1 entre los registros y los conceptos de Snomed CT, e Internación General es la más alta (42:1).

Fig. 3.4: Distribución de todos los problema por su aparición en la lista



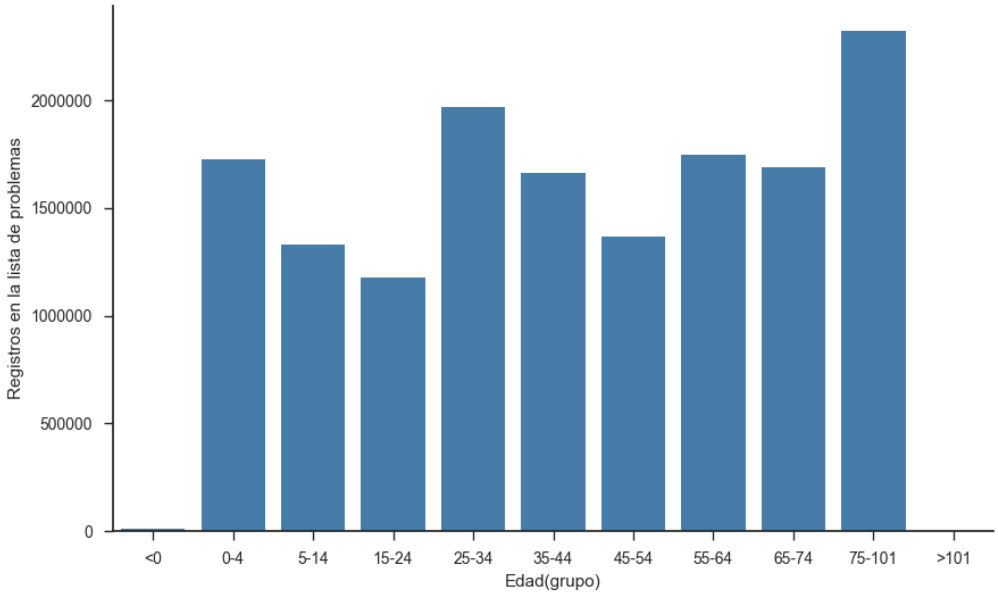
Tab. 3.1: Distribución de registros y conceptos por nivel de asistencia o ámbito

Nivel de Asistencia o ámbito	Registros	Diferentes Conceptos
Ambulatorio	771 096	9543
Guardia	603 412	5759
Triage	556 166	2963
Internación general	267 303	6322
Internación domiciliaria	22 297	1975
Episodio ambulatorio	17 947	1247
Seguimiento domiciliario	14 537	1302
Internación geriátrica	204	115

3.2.4.2. Contexto: Grupo etario

La distribución de los problemas registrados por edad en la figura 3.5 evidencia que los grupos etarios en los que más se consulta son 0-4 años, 25-34 años y mayores de 75 años.

Fig. 3.5: Distribución de todos los problema por edad de los pacientes

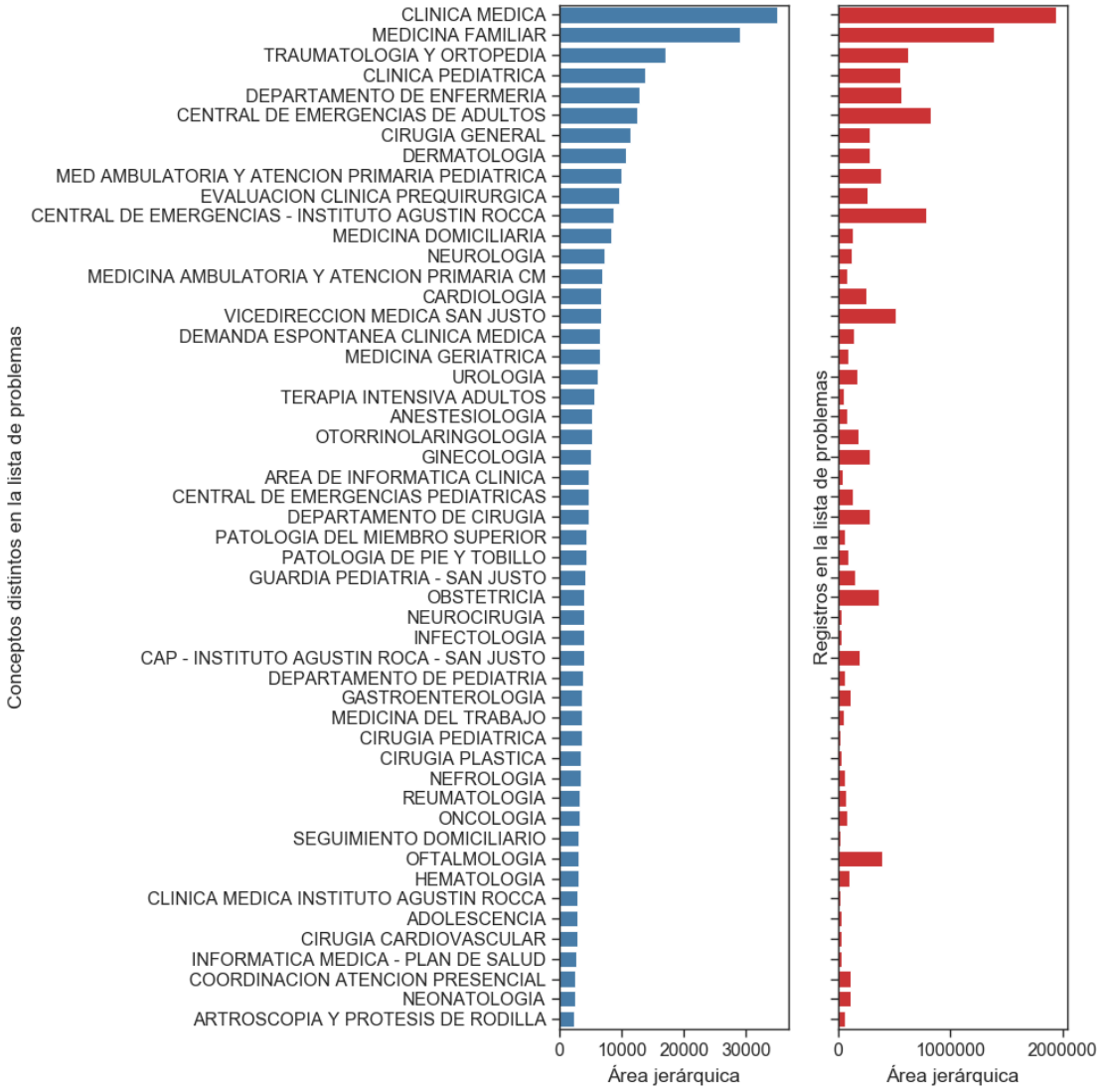


3.2.4.3. Contexto: Área Jerárquica

La HCE tiene 918 áreas jerárquicas, de las cuales 601 tienen usuarios que han registraron problemas dentro de la lista de algún paciente. Hay ocho jerarquías que usan más de 10 000 conceptos (Clínica médica y Medicina Familiar, Traumatología y ortopedia, Clínica pediátrica, Departamento de enfermería, Central de emergencia de adultos, Cirugía general y Dermatología), el resto de jerarquías usan en promedio menos de 5000 conceptos. (Ver figura 3.6). En la siguiente sección, que corresponde a la creación de los contextos,

se evalúa si los datos de la lista de problemas necesitan también de ese nivel de detalle, o si las áreas pueden ser agrupadas.

Fig. 3.6: Registros y conceptos únicos en las top 50 áreas jerárquicas.

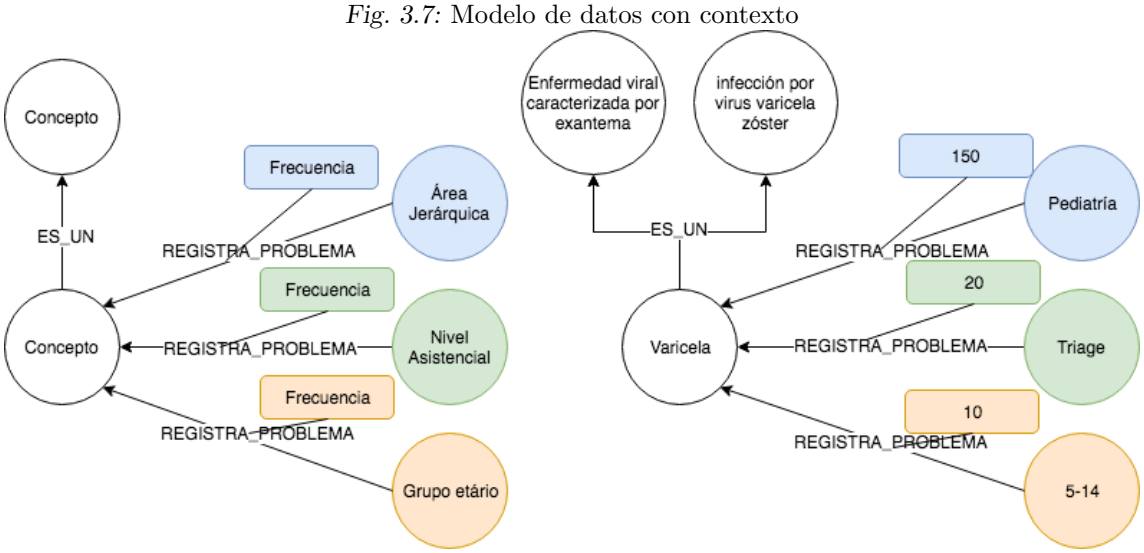


3.3. *Refsets* según el contexto

En la primera fase de preparación de los datos, creé la base de datos de grafos. En el modelo de datos que se observa a la izquierda de la figura 3.7 los vértices son los **conceptos** que está relacionados entre si por la arista **(ES_UN)**, y los **contextos** están relacionados con los **conceptos** con la arista **(REGISTRA_PROBLEMA)**, esta relación tiene el atributo **Frecuencia**. Este modelo representa los diferentes niveles de generalización de los conceptos con la relación (ES_UN). También se representa que un contexto (área jerárquica, nivel asistencial o grupo etario) tiene varios conceptos y un concepto puede estar en varios contextos con la relación (REGISTRA_PROBLEMA). A la derecha de la

figura 3.7 hay una instancia de este modelo. Se representa que contextualmente el concepto varicela está registrado 150 veces en el área jerárquica Pediatría, 20 veces en el nivel asistencial Triage y 10 en el grupo etario de 5 a 14 años. Semánticamente, la varicela es una enfermedad viral caracterizada por exantema y es una infección por virus varicela zoster.

Fig. 3.7: Modelo de datos con contexto



El siguiente paso fue la detección del solapamiento entre los grupos y la posterior generación de los grupos contextuales finales, para realizar este paso se aplicaron algoritmos de aprendizaje no supervisado. A continuación describo los resultados obtenidos por cada uno de los contextos.

3.3.1. Contexto: Nivel de asistencia

Con el fin de encontrar conceptos que estén asociados a los niveles de asistencia apliqué los algoritmos de agrupamiento *leading vector*, *label propagation* y *multilevel* (Ver sección 2.2)). Los valores de modularidad y número de grupos son similares aplicando todos los algoritmos:

- *Leading vector*: modularidad = 0.346, número de grupos = 4, test: P=0.000
- *Label propagation*: modularidad = 0.393 , número de grupos = 5, test: P=0.000
- *Multilevel*: modularidad = 0.397 , número de grupos = 5, test: P=0.000

[1]

La tabla 3.2 contiene los conceptos que se comparten en todos los grupos después de aplicar los algoritmos de agrupamiento. Los niveles de asistencia Internación general e Internación domiciliaria contienen los mismos conceptos. Después de fusionar estos dos niveles, son 3002 conceptos que representan su contexto.

[1] La fase de diseño y evaluación (Sección 2.3.4) contiene los pasos que se siguieron para realizar los test de signficancia de los *clusters*

Tab. 3.2: Contexto de nivel de asistencia y sus conceptos

Nivel de Asistencia o ámbito	Conceptos finales del grupo
Ambulatorio	6304
Guardia	2269
Triage	779
Internación general	3002
Internación domiciliaria	3002
Episodio ambulatorio	370
Seguimiento domiciliario	681
Internación geriátrica	101

3.3.2. Contexto: Grupo etario

Al aplicar los algoritmos de agrupamiento, obtengo resultados similares entre *Leading vector* y *multilevel*. *Leading Vector* dividió el grafo en cinco grupos y la modularidad fue de 0.171. *Multilevel* dividió el grafo en 4 grupos con una modularidad de 0.138. En el caso del algoritmo *label propagation* el valor de la modularidad es de 0, por lo tanto sus resultados son descartados.

Después de combinar los resultados de los algoritmos *Leading vector* y *multilevel*, para identificar los grupos que invariantemente se repitan en ambos resultados, se descarta el grupo etario de 5 a 14 años y fusionan los grupos etarios de 0 a 4, 15 a 24, 25 a 34 y 35 a 44 años. La tabla 3.3 contiene los tamaños finales de los grupos.

Tab. 3.3: Contexto del grupo etario y sus conceptos

Grupo Etario	Conceptos del grupo
0-4	12 532
5-14	0
15-24	12 532
25-34	12 532
35-44	12 532
45-54	6731
55-64	5073
65-74	5165
75-101	4292

3.3.3. Contexto: áreas jerárquicas

Como se definió en la fase del entendimiento de los datos de la metodología (sección 2.3.2), la creación de los *refset* de áreas jerárquicas consiste en la ejecución de las 3 tareas siguientes: (1) en la definición de los contextos significativos y (2) identificación de solapamientos entre *refset* y (3) evaluación final.

Se define como área significativa las que tienen un mapeo a Snomed CT. Su identificación se realiza con *string matching* de las áreas jerárquicas a alguno de los descendientes de los **servicios de la atención de la salud (SCTID: 224891009)** de SNOMED CT.

Si no hay un mapeo, se selecciona el servicio de las áreas que tengan sobreposición en sus conceptos con los *clusters* previamente identificados con el *string matching*. Se excluyen las áreas que tienen un sólo concepto, y las áreas que entran dentro del *cluster* identificado como administrativo.

Diseño del modelo : Ejecuté 3 algoritmos de agrupamiento. como se observa en la tabla 3.4, el algoritmo *leading vector* dividió el conjunto de datos en 2 y su modularidad es la más baja por lo tanto sus resultados se descartaron. Los resultados de los otros dos algoritmos se usaron como criterio para seleccionar la jerarquía final.

Tab. 3.4: Agrupamiento de áreas jerárquicas y conceptos

Algoritmo	Modularidad	Grupos	Test
Label Propagation	0.165	18	0.000
Leading Vector	0.144	2	0.000
MultiLevel	0.447	14	0.000

Resultados de contexto significativos Ejemplos de los resultados de la aplicación de los criterios de exclusión y de selección del primer experimento se encuentran en la tabla 3.5. Los resultados de la aceptación de estos criterios se encuentran en la tabla 3.6, el 63 % de los casos fueron excluidos correctamente. Con respecto a los criterios de selección, el 96 % de los casos fueron seleccionados correctamente según la evaluación realizada por la residencia médica.

Tab. 3.5: Ejemplos de criterios de exclusión y selección

Área Jerárquica	Excluído?	Servicio Seleccionado	Observaciones
Dirección administrativa	Si		Hace Parte Del Cluster De Administrativos
Coordinación de turnos - departamento de enfermería	Si		Tiene un sólo concepto seleccionado.
Servicio de neurocirugía	No	Servicio de neurocirugía	*String Matching* Con un servicio de Snomed CT
Sección diálisis peritoneal continua ambulatoria- Servicio de nefrología	No	Servicio de nefrología adultos	*String Matching* del área más general con un servicio de Snomed CT
Rinología plástica	No	servicio de otorrinolaringología	Comparte cluster con áreas con los mismos conceptos.

En este primer proceso reemplacé las 601 áreas jerárquicas con 44 servicios de la atención de la salud de Snomed CT, y eliminé las que entraron en el criterio de exclusión y fueron aceptadas por el residente.

Tab. 3.6: Resultados de la aplicación de los criterios de exclusión y selección

Criterio	Reportados	Aceptados
Criterio de exclusión	223	142
Selección por *string matching*	291	291
Selección por agrupamiento	116	111

3.3.3.1. Solapamientos entre *refset*

En este proceso se detectan solapamientos entre los 44 servicios mapeados a Snomed CT. El objetivo es evitar la redundancia dentro del contexto y definir servicios cuyos conceptos hace que se diferencie de los otros servicios. Teniendo como variable dependiente el servicio y variable independiente los conceptos, al aplicar un algoritmo de clasificación el resultado de la métrica $F1score$ permite detectar si la variable independiente predice la variable dependiente. Si se presenta un alto solapamiento de conceptos con otros servicios se obtienen valores $F1score$ bajos. Los servicios con $F1score$ bajos son candidatos a fusionarse. El alto solapamiento fue fijado con un valores $F1score < 0{,}65$.

Los datos de entrada son los conceptos y 44 servicios de la atención de la salud. Se aplica el algoritmo de clasificación StanfordNLP ColumnClassifier, con una partición de 50 % de los datos para entrenamiento y 50 % para test.

La matriz de confusión generada por el modelo aporta información sobre qué servicios comparten los mismos conceptos, y se infiere así una fusión entre ellos. La tabla 3.7 contiene los 15 servicios que fueron fusionados según el criterio de selección, y los servicios a los que fueron fusionados. El resultado final es 29 servicios de la atención de la salud, con conceptos que representan su contexto.

Al aplicar el algoritmo de clasificación a los 44 servicios, el modelo creado predice a qué servicio pertenecen los conceptos con el *F1 score* global = 0.500. Al crear un nuevo modelo con los 29 servicios, se obtiene el *F1 score* global = 0.664.

3.3.3.2. Evaluación final

La evaluación final de los contextos de los servicios de atención de la salud se realiza a partir del cálculo del cubrimiento de estos en comparación con los disponibles por *Kaiser Permanente* (Ver sección 1.2.2.3). La tabla 3.8 contiene los servicios que se compararon con CMT. En algunos casos necesité unir varios servicios para comparar con uno sólo de CMT, por ejemplo los *refset* de los servicios de nefrología de adultos, endocrinología, endocrinología pediátrica y urología se unieron para comparar con el *refset Endocrine, Nephrology, and Urology* de CMT.

La tabla 3.9 muestra el cálculo de la cobertura de los *refset* de CMT comparándolos con los servicios de atención de la salud del HIBA, se observa que las mayores coberturas se encuentran en los conjuntos de datos afines relacionados en la tabla 3.8.

Los niveles de granularidad entre los dos conjuntos de datos hace compleja la tarea de comparación. Por ejemplo, en el caso del *refset* del servicio de cardiología del HIBA tenemos el hallazgo Ulcera Arterial, el cual no aparece en el *refset* de cardiología de *Kaiser Permanente*, pero Ulcera Arterial es hijo de Trastorno arterial, el cual es una generalización que si aparece en *Kaiser Permanente*. Por lo tanto, este término debió ser contado como parte de la cobertura. Para contrastar los resultados, calculé el cubrimiento con conceptos

Tab. 3.7: Solapamiento en servicios de atención médica del HIBA

Servicio inicial	F1 Score	TP	Servicios final	F1 Score	Solapados
servicio de anestesia (SCTID: 310001007)	0,048	155	servicio de medicina general (SCTID: 700232004)	0,747	1595
servicio audiológico (SCTID: 310004004)	0,000	0	servicio de otorrinolaringología (SCTID: 310149003)	0,675	10
servicio de cirugía vascular (SCTID: 310168000)	0,178	406	servicio de cardiología adultos (SCTID: 3811000179104)	0,672	909
servicio de colposcopía (SCTID: 310024000)	0,000	0	servicio de ginecoobstetricia (SCTID: 310060005)	0,877	293
servicio de dermatología pediátrica (SCTID: 3821000179109)	0,405	2141	servicio de dermatología (SCTID: 700241009)	0,800	3240
servicio de farmacia (SCTID: 310080006)	0,392	112	servicio de medicina general (SCTID: 700232004)	0,747	186
servicio de fonoaudiología (SCTID: 310101009)	0,612	7402	servicio de otorrinolaringología (SCTID: 310149003)	0,675	6136
servicio de gastroenterología (SCTID: 700433006)	0,233	2716	servicio de medicina general (SCTID: 700232004)	0,747	13836
servicio de gastroenterología pediátrica (SCTID: 3771000175106)	0,000	0	servicio de otorrinolaringología (SCTID: 310149003)	0,675	1
servicio de internación domiciliaria (SCTID: 4291000179105)	0,131	15	servicio de medicina general (SCTID: 700232004)	0,747	86
servicio de patología (SCTID: 310074003)	0,000	0	servicio de imagenología (SCTID: 3851000179100)	0,546	32
servicio de psicología (SCTID: 310123008)	0,000	0	servicio de psiquiatría (SCTID: 310116007)	0,827	32
servicio de reumatología (SCTID: 3621000175101)	0,336	96	servicio de medicina general (SCTID: 700232004)	0,747	226
servicio de terapia intensiva (SCTID: 310032008)	0,275	968	servicio de medicina general (SCTID: 700232004)	0,747	1725
servicio de terapia intensiva pediátrica (SCTID: 310034009)	0,179	156	servicio de medicina general (SCTID: 700232004)	0,747	308

cuya longitud media del camino mínimo es <=1, <=2 y <=3.

Como se muestra en la tabla 3.10, hay una significativa mejora en los valores de cubrimiento, incluso con distancia <=2 todos alcanzan más de un 90 % de cubrimiento de conceptos del *Kaiser Permanente* por el HIBA , con excepción de Oncología. Esta mejora se explica porque el crecimiento del HIBA agrega hasta un nivel más de precisión, muchos conceptos son descendientes del mismo ancestro y al no ser tenidos en cuenta porque no hay un *match* exacto se descartan también todas esas ramas que crecen a lo ancho.

En el caso del cubrimiento de los conceptos de HIBA por *Kaiser Permanente*, el cubrimiento también mejora pero sigue siendo muy pequeño, esto se debe a que el HIBA tiene no sólo más conceptos en los *refset* sino que se evidencia una mayor diversidad en ellos.

3.4. Discusión del capítulo

El objetivo de este capítulo era consolidar el conjunto de datos que se usará para entrenamiento y validación de los modelos. Para poder lograr este objetivo primero realicé un análisis descriptivo de cada una de las dimensiones y luego creé *refset* que pueden ser utilizados como vocabularios controlados dependientes de contextos.

En la primera sección abordé la comprensión de los datos de la lista de problemas. Este análisis me permitió identificar valores atípicos y datos de error al momento de la carga,

revisando la distribución desde diferentes dimensiones: tiempo, individuos, conceptos y contextos.

En el caso del tiempo sólo son válidos los registros con fecha de carga desde el año 1998 hasta el tiempo presente, dado que este ha sido el lapso en el que ha estado activa la HCE.

Desde el punto de vista de los individuos el 34.1 % tiene sólo un problema, estos registros también se descartan porque carecen de co-ocurrencia con otros problemas.

En la dimensión de los conceptos, el 20 % de los registros pertenecen a 12 conceptos y sus variantes lexicográficas. Estos problemas tendrán poca capacidad predictiva dado que se relacionarán con un número grande de conceptos. El uso de los conceptos presenta una distribución de cola larga, esto es consistente por lo reportado por Fung et. al.(Fung y Xu, 2015) donde haciendo una comparación del uso de la lista de problemas de ocho instituciones, encontraron que el 95 % de los registros corresponde a sólo el 22.8 % de términos únicos.

Los contextos analizados fueron el nivel de asistencia y grupo etario y área jerárquica. Aunque los grupos formados tienen modularidades bajas, inferiores al 0.5 en todos los casos, se realizó la combinación de los resultados de los diferentes algoritmos para crear el contexto. Cada uno de estos contextos tiene sus particularidades, descriptas a continuación.

En el nivel de asistencia hay desbalanceo entre los grupos. Los grupos más grandes son el nivel de asistencia ambulatorio (6304 conceptos), las internaciones que se fusionaron en una sola (3002 conceptos), y la guardia (2269 conceptos). Aunque el nivel de asistencia de Triage sea el tercero en cantidad de registros, el número de conceptos en su grupo es muy pequeño, lo que indica la baja diversidad de conceptos seleccionados en este nivel de asistencia.

Los grupos etarios que más registran problemas son entre los 0-4 años, 25 y 34 años y mayores de 75 años, los registros con edades menores a 0 o mayores a 101 fueron descartadas por interpretarse como errores del sistema o casos de prueba. Utilizando los algoritmos de aprendizaje no supervisado, se fusionaron los grupos etarios entre los 15 y 44 años.Los problemas que existen en estos grupos no tienen diferencias significativas. Al mismo tiempo, descarté el grupo etario entre 5 y 14 años, ya que los problemas no se agruparon en un grupo consistente.

El HIBA tiene 601 áreas jerárquicas con diferentes niveles de agregación, algunas de esas áreas corresponden a funciones administrativas. El proceso de definición de los contextos significativos permitió, mediante algoritmos no supervisados, excluir las áreas administrativas y mapear las restantes a 44 servicios de atención de la salud de Snomed CT. En el futuro, otras instituciones podrán mapear sus propias áreas a estos servicios de salud y usar los contextos, facilitando así la interoperabilidad con otros centros médicos.

Con el siguiente proceso se detectaron solapamiento entre aquellos servicios con $F1score$ bajos. Entre los servicios con $F1score$ más bajos están: servicio de anestesia (0,048), servicio audiológico (0,000), servicio de colposcopía (0,000), servicio de internación domiciliaria (0,131), servicio de patología (0,000) y servicio de psicología (0,000). Estos servicios fueron fusionados con los que más presentaban solapamiento en la matriz de confusión.

Los servicios con mejor $F1score$ son: servicio de cardiología pediátrica (0,871), servicio de fisioterapia (0,993), servicio de ginecoobstetricia (0,877), servicio de oftalmología adultos (0,938), servicio de psiquiatría (0,827), servicio de psiquiatría pediátrica(0,945) y servicio de traumatología (0,821). Estos valores muestran que los conceptos diferencian bien estos servicios, y que no deberían ser fusionados aunque uno sea padre del otro como

el caso de servicio de psiquiatría y servicio de psiquiatría pediátrica.

Como resultado de estos procesos obtuve 29 servicios finales. Estos servicios tienen mejor $F1score$ global que los 44 servicios. Para evaluar el cubrimiento de los refset, use como referencia los donados por *Kaiser Permanente* a Snomed CT, en ellos encontré 10 refset de CMT que eran afines a 16 de los 29 servicios finales. Los resultados obtenidos muestran que la mayoría de los servicios tienen una cobertura entre 0,30 y 0,45. Los que se ubican con mejor cobertura son pediatría (0,56), cardiología(0,46) y dermatología (0,45). El de peor cobertura es oncología (0,07), este es el único caso en el que el refset creado por el HIBA es de menor tamaño al de referencia, en todos los otros casos el tamaño de los refset del HIBA es superior en tamaño a los de referencia.

Los niveles de precisión entre los dos conjuntos de datos hace compleja la tarea de comparación. Por lo tanto, contrasté estos resultados con la cobertura calculada con diferentes distancias semánticas, encontrando que hay una significativa mejora en los valores de cubrimiento. Incluso con distancia <= 2, la mayoría de los servicios alcanzan más de un 90 % de cubrimiento de conceptos de CMT .

Una limitación de este trabajo es que no tiene en cuenta el nivel de granularidad de los conceptos que se están comparando, por lo tanto un par de conceptos que estén muy arriba en la jerarquía y sean muy generales tienen igual distancia que dos conceptos que estén muy abajo y sean muy específicos. Sin embargo, en la práctica entre más precisos sean los conceptos que tienen similitud, se puede confiar más en que realmente tienen un significado similar.

Un trabajo similar reportado por Nova Scotia(Kuropatwa y Giannangelo, 2016), en el que desarrollan sus propios refset de especialidades, obtiene coberturas inferiores a las reportadas en esta tesis, incluso sin el cálculo de distancias semánticas (ver tabla 3.11).

3.5. Conclusión del capítulo

En este capítulo he presentado las decisiones en la extracción, transformación y limpieza de la lista de problemas, para definir: (1) el conjunto de datos de entrenamiento y validación , y (2) los contextos que puedan ser usados como vocabularios controlados y limitar el espectro de búsqueda en el siguiente capítulo.

La transformación más compleja fue la de la definición del contexto de los 29 servicios de atención a partir de las 901 áreas jerárquicas. Si bien agrupar los conceptos por contexto es necesario para facilitar el uso significativo de la Snomed CT, son escasos los ejemplos sobre metodologías o refset públicos que puedan ser usados como referencia.

En este capítulo usé modelos de aprendizaje no supervisado y supervisado para definir qué servicios, niveles asistenciales y grupos etarios pueden ser diferenciables de los demás a partir de sus conceptos. En el caso de los servicios, una vez que fueron definidos evalué su cubrimiento con respecto a los refset de referencia de *Kaiser Permanente*, aunque en la mayoría de los casos da un cubrimiento exacto superior al 40 %, si se utilizan las distancias semánticas estos cubrimientos se incrementan significativamente, sobrepasando en la mayoría de los casos el 90 %.

Tab. 3.8: refset afines entre de HIBA y CMT

Servicio HIBA	Subset Kaiser Permanente	Abreviatura
servicio de cardiología adultos (SCTID:3811000179104) + servicio de cardiología pediátrica (SCTID:4381000179100)	*Cardiology Problem List.* Versión 2016	Cardio
servicio de oftalmología adultos (SCTID:4441000179102)	*Ophthalmology Problem List.* Versión 2016	Oftalmo
servicio de psiquiatría (SCTID:310116007)	*Mental Health Subset.* Versión 2016	Psiqui
servicio de oncología clínica (SCTID:310022001)	Hematology and Oncology. Version 2015	Onco
servicio de nefrología adultos (SCTID:3931000179103) + servicio de endocrinologia (SCTID:700434000) + servicio de endocrinología pediátrica (SCTID:3761000175103) + servicio de urología (SCTID:310167005)	*Endocrine, Nephrology, and Urology.* Versión 2015	ENU
servicio de ginecoobstetricia (SCTID:310060005)	*Obstetrics and Gynecology.* Versión 2015	Gineco
servicio de neurocirugía (SCTID:310159002) + servicio de neuropediatría (SCTID:394538003) + servicio de neurología adultos (SCTID:4011000179108)	*Neurology.* Versión 2015	Neuro
servicio de dermatologia	*Skin/Dermatology and Respiratory.* Versión 2015	Derma
servicio de pediatría (SCTID:310066004)	*Pediatrics.* Versión 2014	Pedia
servicio de traumatología (SCTID:4101000179107)	*Orthopedics.* Versión 2014	Orto

Tab. 3.9: Cubrimiento de Kaiser Permanente en Servicios de HIBA

Hiba (No. de conceptos)	**Kaiser permanente (No. de conceptos)**									
	Cardio (880)	**Derma** (2757)	**Gineco** (1307)	**ENU** (1639)	**Neuro** (1792)	**Oftalmo** (3285)	**Onco** (4086)	**Pedia** (3793)	**Psiqui** (1163)	**Orto** (5009)
Cardio(9182)	0,46	0,33	0,11	0,20	0,19	0,09	0,14	0,34	0,13	0,15
Derma(11622)	0,14	0,45	0,08	0,11	0,10	0,06	0,20	0,27	0,08	0,13
Gineco(8319)	0,15	0,18	0,33	0,16	0,09	0,06	0,11	0,27	0,10	0,11
ENU(18075)	0,31	0,42	0,21	0,40	0,42	0,17	0,22	0,46	0,28	0,24
Neuro(10856)	0,26	0,31	0,13	0,19	0,40	0,16	0,17	0,39	0,26	0,21
Oftalmo (4188)	0,06	0,06	0,02	0,04	0,10	0,41	0,03	0,16	0,04	0,09
Onco (721)	0,07	0,03	0,02	0,03	0,03	0,00	0,07	0,05	0,02	0,01
Pedia (17325)	0,26	0,31	0,18	0,27	0,25	0,18	0,18	0,56	0,22	0,28
Psiqui(2373)	0,07	0,13	0,07	0,06	0,08	0,03	0,03	0,16	0,30	0,07
Orto(22555)	0,21	0,37	0,09	0,13	0,24	0,05	0,16	0,34	0,10	0,42

Tab. 3.10: Cubrimientos de servicios con distancias semánticas entre 1 y 3

Servicio	Cubrimientos con Distancia Semántica <=3		Cubrimientos con Distancia Semántica <=2		Cubrimientos con Distancia Semántica <=1		Cubrimientos exactos	
	Kaiser/Hiba	Hiba/Kaiser	Kaiser/Hiba	Hiba/Kaiser	Kaiser/Hiba	Hiba/Kaiser	Kaiser/Hiba	Hiba/Kaiser
Cardio	0,475	0,947	0,320	0,890	0,211	0,782	0,104	0,457
Oftalmo	0,779	0,969	0,647	0,949	0,505	0,787	0,257	0,407
Derma	0,831	0,974	0,683	0,942	0,452	0,824	0,242	0,449
Pedia	0,85	0,973	0,769	0,940	0,594	0,857	0,305	0,560
ENU	0,492	0,979	0,319	0,944	0,174	0,811	0,075	0,403
Onco	0,582	0,879	0,433	0,623	0,307	0,252	0,073	0,173
Psiqui	0,572	0,904	0,476	0,813	0,384	0,627	0,227	0,303
Orto	0,726	0,986	0,586	0,963	0,402	0,860	0,179	0,420
Neuro	0,563	0,975	0,385	0,963	0,227	0,825	0,112	0,404
Gineco	0,597	0,936	0,443	0,904	0,284	0,743	0,119	0,326

Tab. 3.11: Comparación de la cobertura de las especialidades del HIBA vs Nova Scotia

	Nova Scotia			HIBA		
Servicio	# Refset	# Kaiser	Cobertura	# Refset	# Kaiser	Cobertura
Cardiología	886	653	0,18	9182	880	0,46
Dermatoloía	638	691	0,16	11622	2757	0,45
Hematología*	714	330	0,13	721	4086	0,17
Enfermedades infecciosas	2,202	1101	0,08	-	-	
Oftamología	1,327	413	0,34	4188	3285	0.41
Cirugía ortopédica	1,306	167	0,10	22555	1089	0,25
Otoloaringología	1,344	641	0,29	-	-	
Pediatría	3,699	2181	0,34	17325	3793	0,56
Medicina respiratoria	114	511	0,08	-	-	

4. CAPÍTULO: RESULTADOS DEL ANÁLISIS DE REDES

4.1. Introducción

Este capítulo contiene la construcción de dos grafos. El primero con las relaciones a Snomed CT, y el segundo sin las relaciones. Ya que es de interés analizar si las conexiones de Snomed CT mejoran la capacidad de formar comunidades y su capacidad predictiva.

El análisis se compone de la aplicación de las definiciones de los grafos descriptas en los capítulos anteriores, para evaluar si los grafos formados siguen patrones de redes de gran escala. Además del análisis sobre los agrupamientos encontrados luego de aplicar los algoritmos *label propagation*, *leading vector* y *multilevel*.

Finalmente, se evaluará la capacidad predictiva de los conceptos que están en el mismo grupo dada una lista de problema de un paciente. Esta evaluación se realizará usando los refset de los contextos como vocabularios controlados.

4.2. Definición del grafo

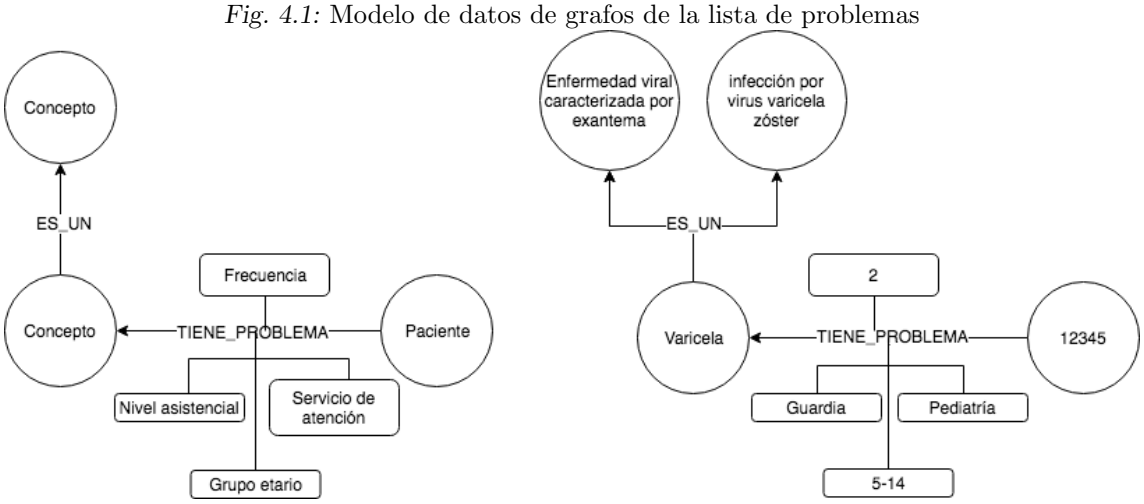
Sea el grafo de la figura 4.1 $G = (V, E)$, donde V denota a los vértices, equivalente a nodos: conceptos de snomed CT y pacientes, y E denota las aristas, equivalente a las relaciones $|TIENE\_PROBLEMA|$, o las relaciones jerárquicas $|ES\_UN|$ entre los conceptos de snomed CT (Hallazgo Clínico, Procedimientos, Eventos, Situación con contexto explícito). Los valores del nivel asistencial, servicio de atención y grupo etario son propiedades de la arista $|TIENE\_PROBLEMA|$, ya que un mismo paciente puede acudir al hospital por el mismo problema en diferentes contextos. Si un mismo paciente acude varias veces con el mismo problema al mismo contexto, esa información se registra en la propiedad frecuencia. A la derecha de la figura4.1 se observa una instancia del modelo, donde el paciente identificado con el id **12345** tiene el problema **Varicela**. El contexto con el que el paciente acudió al hospital es el nivel asistencia de la **Guardia**, el servicio de atención de salud es **Pediatría**, y el grupo etario es entre **5 y 14 años**. El paciente acudió **dos** veces al hospital con el mismo contexto y el mismo problema. Según snomed CT, la **varicela** es una enfermedad viral caracterizada por exantema y es una **infección por virus varicela zoster**.

Los 88 869 conceptos distintos de la lista de problemas del HIBA están relacionados jerárquicamente con 11 946 conceptos de snomed CT. El grafo extendido con Snomed CT tiene en cantidad de vértices y aristas $|V|$ =19 354 y $|E|$ =1 173 234 respectivamente.

Se realiza el mismo análisis sobre el subgrafo generado sólo con las relaciones $|TIENE\_PROBLEMA|$ y los nodos que comparten esta relación. El número de vértices y aristas de este subgrafo son $|V|$ =11 946 y $|E|$ =937 107 respectivamente.

Para diferenciar las dos redes, la primera la nombro como **Red semántica de problemas (RP-SCT)** y la segunda la nombro como **Red de problemas (RP)**

Fig. 4.1: Modelo de datos de grafos de la lista de problemas



4.3. Patrones de la red de la lista de problemas

4.3.1. Red libre de escala

En esta sección evalúo si la redes comparten patrones con redes de gran escala. Realizo dos pruebas, en la primera hago un ajuste a una distribución de ley de potencias usando los grados de cada nodo, y en la segunda hago una comparación con las distribuciones exponencial y logarítmica (Ver sección 1.2.5.2).

En el caso de la **RP-SCT** la función cdf es:

$$F(X \geq 51) \propto x^{-1,82+1} \tag{4.1}$$

El $p - valor(0{,}3188 > 0{,}05)$ no me permite rechazar la hipótesis nula.

En la tabla 4.1 están los resultados de las comparaciones realizadas entre ley de potencias con otras distribuciones. La distribución exponencial fue usada para confirmar que los datos tienen una cola pesada, y la distribución lognormal tiene un mejor ajuste que la ley de potencias, como se muestra también en la figura 4.2.

Tab. 4.1: Comparación de la ley de potencias con otras distribuciones de la **RP-SCT**

Distribución	Resultado
Exponencial	10.64*
Lognormal	-1.5

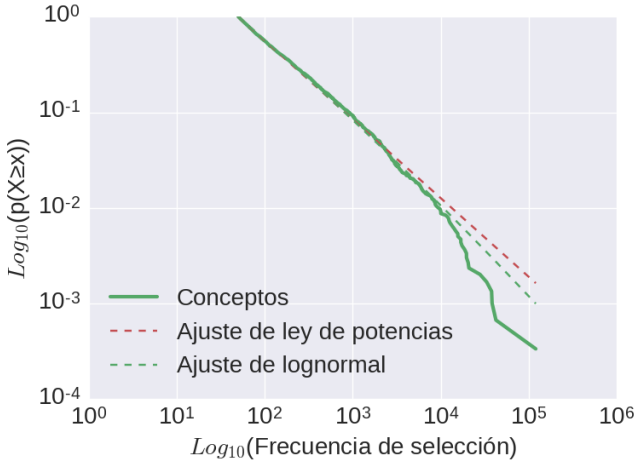
* $p - valor$ <0.05.

En el caso de la **RP** la función cdf es:

$$F(X \geq 279) \propto x^{-1,86+1} \tag{4.2}$$

Con el $p - valor(0{,}03 < 0{,}05)$ se acepta la hipótesis nula según la cual los datos se distribuyen según la ley de potencias. La figura 4.3 y la tabla 4.2 muestra también que esta distribución tiene mejor ajuste.

Fig. 4.2: Gráfico log-log, comparación de ajuste de distribuciones de la red semántica de problemas



Tab. 4.2: Comparación de la ley de potencias con otras distribuciones de la **RP**

Distribución	Resultado
Exponencial	4.64*
Lognormal	-2.41

* p-value<0.05.

4.3.2. Estructuras de comunidad

Existen diferentes métricas para evaluar cuantitativamente los efectos de comunidad en un grafo, las usadas a continuación siguen el trabajo de Boccaletti *et. al*(Boccaletti, Latora, Moreno, Chavez, y Hwang, 2006) y los detalles se encuentran en el marco teórico. (Ver sección 1.2.5.2)

Tab. 4.3: Métricas de efectos de comunidad en redes

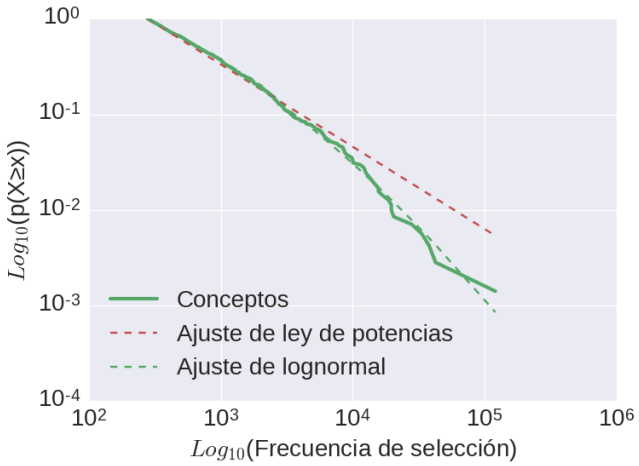
Métrica	Valor en la RP-SCT	Valor en la RP
Coeficiente de agrupamiento - transitividad	0.19	0.21
Coeficiente de agrupamiento - transitividad local promedio	0.41	0.80
Coeficiente de agrupamiento promedio	0.34	0.78
Longitud media del camino mínimo entre nodos	2.80	2.17
Promedio de grados	60.62	78.4

Los valores de coeficiente de agrupamiento cercanos a 1 indican alto grado de agrupamiento, como se puede ver en la tabla 4.3 la **RP** tiene mayores coeficientes que la **RP-SCT**.

4.4. Agrupamiento

Se aplican los algoritmos de agrupamiento *label propagation*, *leading vector* y *multilevel* al grafo de la **RP** y a sub-grafos de la **RP-SCT**, conformados por relaciones que tengan

Fig. 4.3: Gráfico log-log, comparación de ajuste de distribuciones de la **RP**



un mínimo de frecuencia en pacientes (5, 10, 100, 1000 y 10 000). Esto con el propósito de evaluar la estabilidad de los agrupamientos aunque el grafo pierda nodos.

La tabla 4.4 contiene los resultados de la modularidad y la cantidad de grupos encontrados al aplicar los algoritmos de agrupamiento. Se puede observar que en el caso del grafo de la **RP** se generan similares número de grupos aunque la modularidad varía en los algoritmos. La modularidad más alta es el del algoritmo *multilevel* y el más bajo es el de *label propagation*.

En el caso de los sub-grafos **RP-SCT**, el número de grados varía entre algoritmos. *Multilevel* es el que mejores modularidades obtiene, *leading vector* va desmejorando en su modularidad a medida que crece el grafo, *Label propagation* mejora a medida que crece el grafo.

Tab. 4.4: Resultados de agrupamiento de grafos de problemas

Grafo	Algoritmos de agrupamiento					
	Leading Vector		Multilevel		Label Propagation	
	Modularidad	Grupos	Modularidad	Grupos	Modularidad	Grupos
Red de Problemas(RP)	0.36	22	0.43	23	0.08	22
Red Semántica de Problemas (Relaciones en al menos 10000 individuos) (RP-SCT-10.000)	0.50	14	0.59	27	0.27	87
Red Semántica de Problemas (Relaciones en al menos 1000 individuos) (RP-SCT-1.000)	0.50	14	0.59	27	0.27	88
Red Semántica de Problemas (Relaciones en al menos 100 individuos) (RP-SCT-100)	0.50	14	0.59	24	0.29	92
Red Semántica de Problemas (Relaciones en al menos 10 individuos) (RP-SCT-10)	0.29	11	0.61	24	0.37	91
Red Semántica de Problemas (Relaciones en al menos 5 individuos) (RP-SCT-5)	0.32	12	0.62	25	0.46	88

Estos seis grafos con sus 3 algoritmos representan grupos fuertes y débiles, a continua-

ción presento los grupos de problemas que son consistentes en todos los agrupamientos. Además calculo la longitud media del camino mínimo entre los nodos del agrupamiento, esta información permite detectar grupos con conceptos con diferentes significados semánticos y valores atípicos.

4.4.1. Agrupamientos de Red de Problemas

Combinando los resultados de los agrupamientos hay 11 132 posibles combinaciones, encontré 53 grupos de problemas que comparten las mismas agrupaciones. Según la tabla, 4.5 la mayoría (19) de los grupos encontrados tienen dos nodos, los grupos que tiene más nodos tienen en promedio una mediana de grados por nodo más grande, es decir que son nodos altamente conectados.

En la figura 4.4 representa las distancias semánticas entre los conceptos que comparten el mismo grupo. La mayoría de las medianas de las distancias se ubican cercanas a 10. Los grupos con las mayores dispersiones son el cluster_1, cluster_6, cluster_10, cluster_16, cluster_22, cluster_27, cluster_30, cluster_34 y cluster_36.

En la tabla A.1 se encuentra los conceptos con más altas medidas de centralidad (grado, cercanía e intermediación). En esta tabla se encuentran sólo los casos que contienen las mayores dispersiones según la distancia semántica. Se puede observar que coinciden los conceptos con los más altas métricas según el grado y la cercanía, y que difiere en el caso de la intermediación. En la métrica de intermediación se encuentran conceptos que sin tener muchas conexiones son mucho más importantes para la conexión de otros dos vértices del grafo.

Por ejemplo, según lo anterior en el caso del cluster_6 (que se puede interpretar como de enfermedades generales), los problemas con mayor grado (los más populares) son: FIEBRE (SCTID: 386661006) y TOS (SCTID: 49727002), pero el problema con mayor intermediación es RESFRÍO COMÚN (SCTID: 82272006). Es decir, que este último conecta más otros pares de problemas que la Fiebre y la Tos. Otros ejemplos se pueden encontrar en el cluster_10, donde el problema con mayor grado y cercanía es MALESTAR GENERAL (SCTID: 367391008), y el de mayor intermediación es INMUNODEFICIENCIA COMBINADA SEVERA (SCTID: 31323000); y en el grupo de embarazo (cluster_16) donde los problemas con mayor grado son PACIENTE ACTUALMENTE EMBARAZADA (SCTID: 77386006) , ANSIEDAD (SCTID: 48694002) y MAREO (SCTID: 404640003), pero los de mayor intermediación son AMENAZA DE TRABAJO DE PARTO PREMATURO (SCTID: 6383007), INTOXICACIÓN POR FÁRMACO Y/O SUSTANCIA MEDICINAL (SCTID: 7895008) y SANGRADO VAGINAL (SCTID: 268471004).

4.4.1.1. Valores atípicos

Según el boxplot de la figura 4.4 generado con las distancias semánticas, se presentan valores atípicos en los siguientes grupos:

- cluster_6: Este es un agrupamiento de enfermedades generales donde los problemas que tiene mayor distancia semántica en promedio con el resto del mismo grupo son CIRUGÍA DE CATARATAS (PROCEDIMIENTO) (SCTID: 110473004), ATAQUE DE PÁNICO (HALLAZGO) (SCTID: 225624000) y AMIGDALECTOMÍA (PROCEDIMIENTO) (SCTID: 173422009)

- cluster_10: Este es un agrupamiento de administración de medicamentos y procedimientos invasivos, los problemas con mayores distancias son ADMINISTRACIÓN DE ANTICOAGULANTE (PROCEDIMIENTO) (SCTID: 71788004), TRATAMIENTO CON ANTIBIÓTICOS INTRAVENOSOS (PROCEDIMIENTO) (SCTID: 281790008), INYECCIÓN DE GAMMAGLOBULINA (PROCEDIMIENTO) (SCTID: 180191005) y CAMBIO DEL TUBO DE TRAQUEOSTOMÍA (PROCEDIMIENTO) (SCTID: 2267008). Estas distancias se explican porque la mayoría de los conceptos de este grupo están en la jerarquía de Hallazgos.

- cluster_27: En este agrupamiento de enfermedades relacionadas con la edad avanzada, los problemas que tienen mayor distancia son CONFUSIÓN AGUDA(HALLAZGO) (SCTID: 130987000) y PACIENTE EN CAMA (HALLAZGO) (SCTID: 160685001).

- cluster_36: En este agrupamiento de enfermedades cardiopulmonares, las mayores distancias se encuentran en los conceptos BIOPSIA ENDOMIOCÁRDICA (PROCEDIMIENTO) (SCTID: 387829002), ALOTRASPLANTE ORTOTÓPICO DE CORAZÓN (PROCEDIMIENTO) (SCTID: 232974001) y TUMOR DE KLATSKIN (TRASTORNO) (SCTID: 253017000), esta última considerada como una enfermedad huérfana[1].

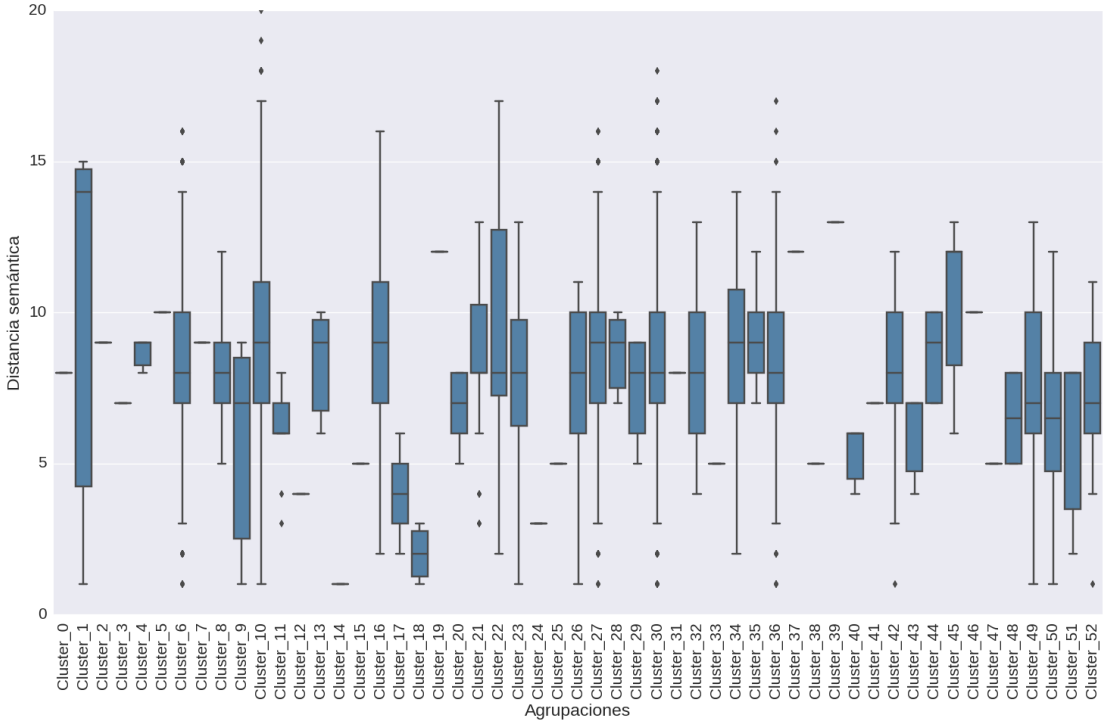
Los valores atípicos en las cotas inferiores, donde las distancias semánticas son cercanas a uno, se debe a relaciones entre conceptos ancestros y sus descendientes. Por ejemplo: en el cluster_42: EPISTAXIS ANTERIOR (SCTID: 232354002) es descendiente de EPISTAXIS (SCTID: 249366005), y ambos conceptos están en el mismo grupo.

Tab. 4.5: Grupos que comparten las mismas agrupaciones en la red de problemas

Número de Nodos	Grupos Encontrados	Mediana de grados por nodo
2	19	1949
3	9	1832
4	5	1833
5	1	1818
6	5	3287
7	1	1590
8	2	5192
10	2	2222
13	1	802
16	1	990
22	1	2198
25	1	3008
34	1	4210
45	1	864
59	1	3700
77	1	1228
84	1	3234

[1] www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=ES&Expert=99978

Fig. 4.4: Distancias semánticas entre conceptos de los agrupamientos de la RP



4.4.2. Agrupamientos de Red Semántica de Problemas

Combinando los resultados de los agrupamientos utilizando cualquier de los algoritmos, hay 2.39×10^{26} posibles combinaciones. Encontré 68 grupos de problemas que comparten las mismas agrupaciones. Según la tabla 4.6 La mayoría (44) de los grupos encontrados tienen 2 nodos.

La figura 4.5 representa las distancias semánticas entre los conceptos que comparten el mismo grupo, se puede observar que las medianas de las distancias tienen un mínimo de 1 y máximo de 10. A diferencia de la de RP, en la RP-SCT las medianas de la distancia semántica no son constantes.

En la tabla A.2 se encuentra los conceptos con las más altas medidas de centralidad (grado, cercanía e intermediación). Estos grupos son más pequeños que en la RP, pero sobresalen los grupos con distancias semánticas grandes cuyos conceptos pareciera que no están muy relacionados, por ejemplo, el cluster_0: INCISIÓN DE LA TRÁQUEA (SCTID: 48387007), AMIGDALECTOMÍA(SCTID: 173422009) y CIRUGÍA DE CATARATAS (SCTID: 110473004), el cluster_2: FÍSTULA TRAQUEOESOFÁGICA (SCTID: 95435007) y ÚTERO UNICORNE (SCTID: 1372004), el cluster_6: HIPERCORTISOLISMO (SCTID: 47270006) y TUMOR DE KLATSKIN (SCTID: 253017000). Sin embargo, al hacer una búsqueda rápida se pueden encontrar evidencias de las relaciones de estas enfermedades en artículos de divulgación científica, como se muestra en la tabla A.3.

Observando las diferencias de los resultados entre las dos redes, entre los grupos que estaban en la RP y desaparecen en la RP-SCT, se encontró evidencia de casos que son distantes semánticamente y que fueron agrupados juntos cuando se usó la red sin las aristas semánticas RP , Ejemplos:

- Distancia 10, Defecto ventilatorio (SCTID: 11483009) y Problema nasal (SCTID: 301199001).
- Distancia 13, Alergia (SCTID: 106190000) y Dermatitis atópica (SCTID: 24079001).
- Distancia 4, Náuseas y vómitos (SCTID:16932000) y Hallazgo relacionado con el vómito (SCTID:300359004)

Por otra parte, hay grupos que se generaron en RP-SCT y no aparecen en RP. Los conceptos tienen distancias semánticas pequeñas, ejemplos:

- Distancia 2, Hipertiroidismo (SCTID: 34486009) y trastorno del cuello (SCTID: 118939000). Hipertiroidismo es un hijo de trastorno del cuello
- Distancia 2, Mucositis ulcerosa de cuello uterino (SCTID: 428193004) y Rinitis (SCTID: 70076002). Ambos son hijos de enfermedad inflamatoria de las membranas mucosas.
- Distancia 2, Enfermedad inflamatoria pélvica aguda (SCTID:237037006) y Absceso agudo de la mama (SCTID:16698000). Ambos son hijos de trastorno inflamatorio agudo.

4.4.2.1. Valores atípicos

Según el boxplot generado con las distancias semánticas en la figura 4.5, se presentan valores atípicos sólo en el grupo cluster_56 formado por: PROCTORRAGIA (SCTID: 12063002), INDIGESTIÓN (SCTID: 162031009), DISFAGIA (SCTID: 40739000), VEJIGA: INCONTINENTE (SCTID: 165232002), CÁLCULO RENAL (SCTID: 95570007), HEMORRAGIA DIGESTIVA BAJA (SCTID: 87763006), SÍNDROME DE ICTERICIA COLESTÍSICA (SCTID: 44018007) y PIELONEFRITIS (SCTID: 45816000). Los problemas con mayores distancias son VEJIGA: INCONTINENTE (SCTID: 165232002) y SÍNDROME DE ICTERICIA COLESTÍSICA (SCTID: 44018007).

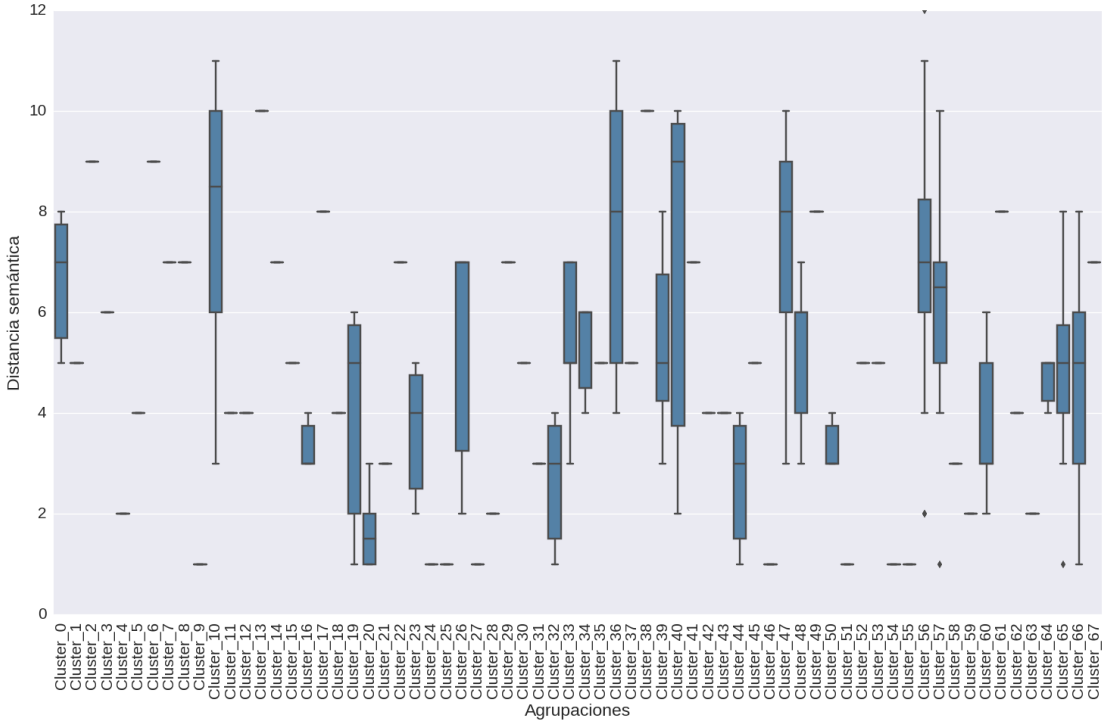
Tab. 4.6: Grupos que comparten las mismas agrupaciones en la red semántica de problemas

Número de Nodos	Grupos Encontrados	Mediana de grados por nodo
2	44	2041
3	12	2144
4	3	1843
5	4	1826
6	2	1226
8	2	3501
9	1	1172

4.4.2.2. Agrupamientos *Label Propagation*

Combinando los resultados de aplicar sólo el algoritmo *label propagation* en la RP-SCT con diferentes cantidades mínimas de individuos que tengan el problema (RP-SCT-10.000,RP-SCT-1.000, RP-SCT-100,RP-SCT-10 y RP-SCT-5) , hay 5.64×10^9 posibles

Fig. 4.5: Distancias semánticas entre conceptos de los agrupamientos de la RP-SCT



combinaciones (Ver tabla 4.4) . De estas combinaciones hay 232 grupos comunes a todos los resultados de agrupamiento.

En los 232 grupos se presenta que más del 50 % de los grupos tienen menos de 50 nodos, estos grupos son muy cohesivos, ya que tienen distancias en promedio de dos saltos entre vértices y con una baja cantidad total de distancias atípicas a los otros vértices del grupo (10 en total en 144 grupos). Los grupos más grandes tienen también la mayor cantidad de vértices con valores de distancias atípicas (Ver tabla 4.7).

Tab. 4.7: Agrupamientos recurrentes en la red RP-SCTcon el algoritmo *Label propagation*

Número de nodos	Estadísticos de distancia entre nodos					
	Número de agrupamientos	Promedio	Desviación estándar	Máximo	Mínimo	Distancias con valores atípicos
Menores de 50 nodos	144	2	1,7	10	1	10
Entre 50 y 100 nodos	25	2	0,8	5	2	34
Entre 100 y 200 nodos	13	3	1,4	8	2	0
Entre 200 y 500 nodos	20	3	1,1	7	2	56
Entre 500 y 1000 nodos	10	4	1,4	7	3	106
Mayor de 1000 nodos	20	6	1,4	9	4	6600

4.4.2.3. Agrupamientos *Leading vector*

Combinando los resultados de aplicar el algoritmo *leading vector* en la RP-SCT con diferentes cantidades mínimas de individuos que tengan el problema (RP-SCT-10.000,RP-SCT-1.000, RP-SCT-100,RP-SCT-10 y RP-SCT-5) , hay 3.62×10^5 posibles combinacio-

nes (Ver tabla 4.4) . De estas combinaciones hay 38 grupos comunes a todos los resultados de agrupamiento.

Según la tabla 4.8 el 50 % de los agrupamientos tienen más de 1000 nodos. Los grupos hasta 1000 nodos son bastante cohesivos, con distancias promedios cercanas a 3 y sin distancias atípicas a los otros vértices del grupo.

Tab. 4.8: Agrupamientos recurrentes en la red RP-SCT con el algoritmo *Leading vector*

Número de nodos	Estadísticos de distancia entre nodos					
	Número de agrupamientos	Promedio	Desviación estándar	Máximo	Mínimo	Distancias con valores atípicos
Menores de 50 nodos	15	3,4	2,8	9	1	0
Entre 50 y 100 nodos	3	3	1,7	5	2	0
Entre 100 y 200 nodos	1	3	-	3	3	0
Entre 500 y 1000 nodos	2	5,5	0,7	6	5	0
Mayor de 1000 nodos	19	7	1,2	9	5	50712

4.4.2.4. Agrupamientos *Multilevel*

Combinando los resultados de aplicar el algoritmo *multilevel* en la RP-SCT con diferentes cantidades mínimas de individuos que tengan el problema (RP-SCT-10.000,RP-SCT-1.000, RP-SCT-100,RP-SCT-10 y RP-SCT-5) , hay 1.05×10^7 posibles combinaciones (Ver tabla 4.4) . De estas combinaciones hay 199 grupos comunes a todos los resultados de agrupamiento.

Según la tabla 4.9 el 50 % de los agrupamientos tienen menos de 50 nodos. La distribución es similar a los del algoritmo *label propagation*, donde los grupos menores a 1000 nodos son cohesivos con distancias de 3, 4 y 6 en promedio y pocos nodos con distancias atípicas a los otros nodos del grupo.

Tab. 4.9: Agrupamientos recurrentes en la red RP-SCT con el algoritmo *Multilevel*

Número de nodos	Estadísticos de distancia entre nodos					
	Número de agrupamientos	Promedio	Desviación estándar	Máximo	Mínimo	Distancias con valores atípicos
Menores de 50 nodos	109	3	2,2	9	1	4
Entre 50 y 100 nodos	21	4	1,5	6	2	2
Entre 100 y 200 nodos	13	4	1,8	8	2	50
Entre 200 y 500 nodos	15	4	1,1	6	3	26
Entre 500 y 1000 nodos	11	6	1,6	9	4	66
Mayor de 1000 nodos	31	6	1,0	8	4	46345

4.4.3. Capacidad predictiva de los agrupamientos

En la sección anterior encontré 590 grupos de problemas que se constituían independientemente del algoritmo de agrupamiento o de la red de entrenamiento. Cada uno de los grupos representa opciones de sugerencias para los usuarios. Es decir, si una lista de problemas tiene alguno de estos conceptos, los otros conceptos pueden ser sugeridos al profesional de la salud para que complete la lista del paciente.

Realicé la evaluación con grupos de menos de 1000 conceptos. Teniendo en cuenta las directrices de recuperación de información:

- Un conjunto de pruebas: La *query* formada por la lista de problemas antes del año 2017 y los problemas del 2017 representan la respuesta correcta,
- Los conceptos de snomed CT a ser recuperados: subgrafos que contengan alguno de los conceptos de la lista de problemas del paciente sin filtros y con filtro por contexto.
- Medida de relevancia por cada par de *query*-concepto recuperado: centralidad y distancia semántica.

4.4.3.1. Directriz: Centralidad es la medida de relevancia y sin filtros de contexto

La tabla 4.10 contiene las métricas de precisión y exactitud de los 10 resultados más relevantes ordenados por centralidad. En la tabla 4.11 no se tienen en cuenta las repeticiones. En ambas tablas se establece la comparación de asumir como verdaderos positivos los casos de predicciones exactas, o los casos cuando la distancia entre las predicciones y el verdadero resultado es menor o igual a dos.

Los nodos excluidos sirven para filtrar los que por su alta medida de centralidad generan muchos falsos positivos, por ejemplo según el grado el concepto más popular es FIEBRE (SCTID: 386661006), pero al aparecer en una lista de problemas como *query* hay 51 606 posibles conceptos diferentes con los que está relacionado. Las medidas de centralidad con los que filtré y ordené los resultados son grado, intermediación, cercanía y autovector.

Se puede observar en ambas tablas, que en el caso de las predicciones exactas, los resultados con precisiones más altas se logran con 20 nodos excluidos en grados , cercanía y autovector. En el caso de intermediación aumenta en la medida que más se filtren datos, con un máximo local en 400 nodos. Los resultados de exactitud más altos se encuentran localmente en el máximo de 400 nodos excluidos.

Cuando las predicciones no son exactas, sino que se toma como verdadero positivo si al menos hay una distancia de hasta 2 entre la respuesta y las predicciones, los valores de precisión y exactitud aumentan, presentándose los máximos en los casos en los que no se excluyen nodos. Estos valores son en ordenamiento por grado y cercanía P@10(0.948) y Acc@10(0.949) y ordenando por autovector P@10(0.947) y Acc@10(0.948). En el caso de intermediación, como en las predicciones exactas, el máximo local está cuando se excluyen 400 nodos, donde se obtienen P@10(0.972) y Acc@10(0.973).

La precisión y exactitud son medidas que dependen de los falsos positivos, verdaderos positivos y verdaderos negativos. En las tablas 4.14 y 4.15 se encuentran las frecuencias de estos valores para los mejores casos con y sin repeticiones. En el caso en el que las predicciones no son exactas hay un incremento significativo de los verdaderos positivos.

4.4.3.2. Directriz: Distancia semántica es la medida de relevancia y sin filtros de contexto.

En las tablas 4.12 y 4.13, que corresponden a ordenamiento por distancia semántica con y sin repeticiones, obtengo los mejores valores de precisión y exactitud en los casos de las predicciones no exactas. Los valores son P@10(0.995) y Acc@10(0.995) con repeticiones y P@10(0.997) y Acc@10(0.723) sin repeticiones en distancias ordenadas de menor a mayor. La eliminación de los top 100 nodos con mayores medidas de centralidad afectan significativamente la capacidad predictiva de los agrupamientos.

Tab. 4.10: Capacidad predictiva de los agrupamientos ordenando por centralidad

Nodos excluidos	Grado				Intermediación				Cercanía				Autovector			
	Predicción Exacta		Distancia a la respuesta <= 2		Predicción Exacta		Distancia a la respuesta <= 2		Predicción Exacta		Distancia a la respuesta <= 2		Predicción Exacta		Distancia a la respuesta <= 2	
	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10
0	0,075	0,106	0,948	0,949	0,029	0,061	0,934	0,936	0,075	0,106	0,948	0,949	0,075	0,106	0,947	0,948
20	0,085	0,295	0,656	0,735	0,028	0,060	0,933	0,936	0,075	0,297	0,637	0,724	0,086	0,296	0,660	0,738
40	0,072	0,306	0,608	0,707	0,028	0,060	0,933	0,935	0,072	0,306	0,618	0,714	0,073	0,304	0,605	0,704
60	0,071	0,317	0,598	0,705	0,028	0,061	0,933	0,935	0,059	0,314	0,582	0,695	0,069	0,316	0,581	0,692
80	0,043	0,309	0,566	0,687	0,028	0,061	0,933	0,935	0,040	0,306	0,569	0,688	0,061	0,315	0,575	0,690
100	0,043	0,312	0,568	0,690	0,028	0,061	0,933	0,935	0,037	0,310	0,564	0,687	0,041	0,310	0,557	0,681
120	0,033	0,309	0,540	0,672	0,019	0,053	0,933	0,935	0,041	0,314	0,568	0,691	0,040	0,316	0,549	0,678
140	0,034	0,317	0,517	0,658	0,019	0,053	0,933	0,935	0,032	0,318	0,518	0,660	0,032	0,316	0,523	0,663
160	0,032	0,320	0,507	0,654	0,019	0,053	0,933	0,935	0,032	0,320	0,515	0,660	0,031	0,318	0,518	0,661
180	0,029	0,320	0,514	0,660	0,018	0,053	0,933	0,935	0,032	0,325	0,509	0,658	0,030	0,323	0,511	0,658
200	0,026	0,324	0,514	0,663	0,019	0,053	0,933	0,935	0,032	0,329	0,503	0,656	0,028	0,324	0,508	0,658
220	0,027	0,334	0,498	0,656	0,018	0,053	0,933	0,935	0,030	0,335	0,504	0,660	0,027	0,329	0,514	0,665
240	0,035	0,342	0,526	0,677	0,037	0,071	0,939	0,941	0,038	0,350	0,517	0,674	0,035	0,341	0,540	0,685
260	0,037	0,352	0,772	0,846	0,049	0,082	0,962	0,963	0,039	0,353	0,776	0,849	0,037	0,350	0,775	0,848
280	0,061	0,371	0,784	0,855	0,076	0,109	0,963	0,964	0,062	0,373	0,791	0,860	0,061	0,369	0,791	0,860
300	0,063	0,378	0,799	0,867	0,082	0,115	0,964	0,966	0,063	0,378	0,796	0,864	0,062	0,373	0,798	0,865
320	0,064	0,382	0,804	0,870	0,091	0,124	0,970	0,971	0,064	0,386	0,794	0,865	0,063	0,376	0,804	0,869
340	0,065	0,386	0,804	0,872	0,097	0,129	0,971	0,972	0,065	0,391	0,793	0,865	0,054	0,373	0,795	0,864
360	0,067	0,393	0,809	0,876	0,100	0,132	0,972	0,973	0,067	0,397	0,808	0,876	0,054	0,380	0,816	0,880
380	0,056	0,393	0,816	0,882	0,102	0,134	0,973	0,974	0,068	0,400	0,812	0,879	0,055	0,383	0,821	0,883
400	0,057	0,397	0,809	0,878	0,102	0,135	0,972	0,973	0,068	0,403	0,810	0,879	0,055	0,389	0,815	0,880

4.4.3.3. Directriz: Filtros de contexto y medida de relevancia según el grado.

En las siguientes secciones, la lista de conceptos predichos están filtrados utilizando tres criterios: el servicio de atención de salud, el ámbito o nivel asistencia y el grupo etario. Del tal forma, que si los conceptos predichos no hacen parte del refset de ese contexto, entonces se elimina de la lista de predicciones.

Agrupamientos con servicio de atención de salud Los servicios de atención que se evaluan a continuación, fueron validados en el capítulo anterior con los refset de *Kaiser Permanente* (Ver sección 3.3.3). Sin utilizar elementos de contexto, el mejor caso es donde se ordenan los conceptos según su distancia semántica, admitiendo repeticiones el valor de P@10 era de 0.174 y Acc@10 era de 0.202 y sin repeticiones el valor de P@10 era de 0.230 y Acc@10 era de 0.244. En contraste, según lo registrado en la tabla 4.16, usando los refset de servicios de atención de salud todos los valores de predicción y exactitud son cercanos a 0.800 con y sin repeticiones.

Los servicios de atención con mejor capacidad predictiva observando las predicciones exactas son:

- Pediatría (P@10=0.893 y Acc@10 = 0.897) con repeticiones y (P@10=0.895 y Acc@10 = 0.896) sin repeticiones,
- Neurología Adultos (P@10=0.880 y Acc@10 = 0.884) con repeticiones y (P@10=0.886 y Acc@10 = 0.888) sin repeticiones y
- Cardiología (P@10=0.872 y Acc@10 = 0.876) con repeticiones y (P@10=0.859 y Acc@10 = 0.860) sin repeticiones

Si la predicción no es exacta pero se admite como verdadero positivo una distancia de dos nodos entre la respuesta y alguna de las predicciones, entonces los valores P@10 y Acc@10 son superiores al 0.950. Estos valores no varían entre los ordenamientos por centralidad o distancias semánticas.

Tab. 4.11: Capacidad predictiva de los agrupamientos ordenando por centralidad sin repeticiones

Nodos excluidos	Grado				Intermediación				Cercanía				Autovector			
	Predicción Exacta		Distancia a la respuesta <= 2		Predicción Exacta		Distancia a la respuesta <= 2		Predicción Exacta		Distancia a la respuesta <= 2		Predicción Exacta		Distancia a la respuesta <= 2	
	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10
0	0,101	0,116	0,963	0,699	0,039	0,056	0,953	0,691	0,101	0,117	0,963	0,699	0,100	0,116	0,963	0,698
20	0,097	0,166	0,709	0,530	0,038	0,054	0,952	0,691	0,085	0,161	0,690	0,519	0,099	0,167	0,714	0,533
40	0,082	0,168	0,661	0,502	0,037	0,054	0,951	0,690	0,081	0,168	0,670	0,508	0,082	0,166	0,657	0,499
60	0,080	0,177	0,648	0,497	0,038	0,054	0,952	0,690	0,067	0,170	0,631	0,487	0,077	0,175	0,630	0,485
80	0,048	0,161	0,614	0,478	0,038	0,055	0,952	0,690	0,045	0,157	0,615	0,479	0,069	0,172	0,624	0,482
100	0,048	0,164	0,615	0,480	0,038	0,055	0,951	0,690	0,042	0,159	0,611	0,477	0,046	0,161	0,603	0,472
120	0,037	0,158	0,588	0,464	0,026	0,043	0,951	0,690	0,045	0,164	0,614	0,480	0,045	0,166	0,594	0,468
140	0,038	0,165	0,564	0,451	0,025	0,043	0,951	0,690	0,036	0,166	0,565	0,452	0,035	0,164	0,570	0,455
160	0,035	0,168	0,553	0,445	0,025	0,043	0,951	0,690	0,036	0,168	0,562	0,451	0,035	0,166	0,565	0,452
180	0,032	0,168	0,560	0,451	0,025	0,043	0,951	0,690	0,035	0,174	0,555	0,449	0,034	0,171	0,556	0,448
200	0,030	0,171	0,560	0,452	0,025	0,043	0,951	0,690	0,036	0,178	0,548	0,446	0,031	0,172	0,553	0,448
220	0,030	0,183	0,542	0,445	0,025	0,043	0,951	0,690	0,033	0,184	0,548	0,448	0,030	0,177	0,558	0,453
240	0,039	0,193	0,571	0,464	0,050	0,067	0,957	0,694	0,042	0,201	0,562	0,460	0,039	0,191	0,585	0,471
260	0,042	0,203	0,781	0,593	0,066	0,083	0,971	0,704	0,043	0,205	0,785	0,595	0,042	0,201	0,783	0,594
280	0,067	0,228	0,794	0,601	0,102	0,119	0,972	0,705	0,068	0,231	0,800	0,605	0,067	0,226	0,801	0,605
300	0,069	0,235	0,808	0,611	0,109	0,126	0,973	0,706	0,069	0,236	0,805	0,608	0,068	0,230	0,807	0,609
320	0,070	0,240	0,812	0,613	0,122	0,139	0,977	0,708	0,070	0,244	0,802	0,608	0,069	0,234	0,813	0,613
340	0,071	0,244	0,812	0,614	0,129	0,146	0,978	0,709	0,071	0,250	0,801	0,609	0,059	0,229	0,804	0,608
360	0,073	0,252	0,817	0,618	0,134	0,150	0,978	0,709	0,074	0,258	0,816	0,618	0,059	0,236	0,824	0,621
380	0,061	0,250	0,823	0,622	0,136	0,152	0,979	0,710	0,074	0,261	0,820	0,621	0,060	0,240	0,828	0,624
400	0,062	0,255	0,815	0,619	0,136	0,153	0,978	0,709	0,074	0,264	0,818	0,620	0,060	0,247	0,822	0,621

Tab. 4.12: Capacidad predictiva de agrupamientos ordenando con distancias semánticas

Agrupamiento	Distancia de menor a mayor				Distancia de mayor a menor			
	Predicción exacta		Distancia a la respuesta <= 2		Predicción exacta		Distancia a la respuesta <= 2	
	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10
Sin eliminar ningún nodo	0,174	0,202	0,995	0,995	0,068	0,099	0,906	0,909
-100 top grado	0,068	0,099	0,985	0,989	0,045	0,314	0,198	0,423
-100 top intermediación	0,066	0,102	0,995	0,995	0,066	0,098	0,907	0,910
-100 top cercanía	0,158	0,396	0,984	0,989	0,048	0,317	0,198	0,425
-100 top autovector	0,158	0,394	0,985	0,989	0,044	0,312	0,196	0,422

Agrupamientos con nivel de asistencia En la tabla 4.17 se encuentran los resultados de la capacidad predictiva del modelo cuando se filtra usando el contexto nivel de asistencia. Al compararlos con los resultados del mejor caso sin información contextual, se desmejora el valor de P@10 en todos los niveles de asistencia.

Los resultados obtenidos son inferiores a los obtenidos con el contexto dado por el área jerárquica.

4.4.3.4. Agrupamientos con grupo etario

En la tabla 4.18 se encuentran los resultados de la capacidad predictiva del modelo, usando el contexto de grupo etario al que pertenece el paciente para filtrar las sugerencias. Los modelos obtenidos tienen mejor capacidad predictiva que los obtenidos en los otros contextos, según los valores de predicción y exactitud en el top 10 (P@10 y Acc@10 respectivamente).

El mejor modelo es el del grupo etario de los pacientes entre 75 y 101 años. Al mismo tiempo, este grupo tiene más registros en la lista de problemas que los otros grupos (ver 3.5).

En el caso de los grupos etarios de 0 a 4 años, 15 a 24 años, 25 a 34 años y 35 a 44 años,

Tab. 4.13: Capacidad predictiva de agrupamientos ordenando con distancias semánticas sin repeticiones

Agrupamiento	Distancia de menor a mayor				Distancia de mayor a menor			
	Predicción exacta		Distancia a la respuesta <= 2		Predicción exacta		Distancia a la respuesta <= 2	
	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10
Sin eliminar ningún nodo	0,230	0,244	0,997	0,723	0,091	0,107	0,924	0,670
-100 top grado	0,177	0,277	0,987	0,716	0,051	0,166	0,220	0,228
-100 top intermediación	0,230	0,243	0,997	0,722	0,088	0,105	0,924	0,670
-100 top cercanía	0,176	0,277	0,986	0,716	0,054	0,170	0,221	0,229
-100 top autovector	0,049	0,164	0,987	0,716	0,049	0,164	0,218	0,227

Tab. 4.14: Mejores resultados de precisión y exactitud del modelo de agrupamiento

Método de ordenamiento	Filtro de nodos	FP	TP	TN
Centralidad	-20 top grado	110159	10295	35774
	-400 top grado	94274	5648	56306
	-20 top intermediación	146780	4201	5247
	-400 top intermediación	135136	15395	5697
	-20 top cercanía	109808	8877	37543
	-400 top cercanía	93310	6793	56125
	-20 top autovector	110041	10412	35775
	-400 top autovector	95451	5547	55230
Distancias Semánticas (menor a mayor)	Sin eliminar ningún nodo	124738	26249	5241
	-100 top grado	140723	10264	5241
	-100 top intermediación	140335	9979	5914
	-100 top cercanía	94375	17679	44174
	-100 top evector	94646	17764	43818
Distancia a la respuesta <= 2	Ordenamiento por grado	7909	143078	5241
	Ordenamiento por intermediación	9954	141033	5241
	Ordenamiento por cercanía	7896	143091	5241
	Ordenamiento por evector	8064	142923	5241
	Ordenamiento Distancia semántica	810	150177	5241

empleé el mismo filtro de problemas, ya que como se definió en el capítulo anterior los algoritmos de agrupamiento no encontraron diferencias significativas entre los problemas de estos grupos etarios (Ver sección 3.3.2). La capacidad predictiva en estos grupos es similar, comparando su precisión y exactitud. El peor rendimiento tiene un mínimo en P@10 0.831 y Acc@10 0.838 en todas las ocurrencias, y en P@10 0.850 y Acc@10 0.853 sin repeticiones, estos valores corresponden al grupo etario de 0 a 4 años.

4.4.4. Visualización de agrupamientos por contextos

En esta sección se describe por medio de visualizaciones cómo los contextos y las grupos están representados. De manera general, los grupos encontrados en la lista de problemas se categorizan como se muestra en la figura 4.6. Un mismo concepto puede pertenecer a varios grupos.

Tab. 4.15: Mejores resultados de precisión y exactitud del modelo de agrupamiento sin repeticiones

Método de ordenamiento	Filtro de nodos	FP	TP	TN
Centralidad	-20 top grado	94393	10193	8646
	-400 top grado	84398	5549	23285
	-20 top intermediación	107068	4180	1984
	-400 top intermediación	95893	15153	2186
	-20 top cercanía	95005	8803	9424
	-400 top cercanía	83360	6694	23178
	-20 top autovector	94268	10306	8658
	-400 top autovector	85302	5449	22481
Distancias Semánticas (menor a mayor)	Sin eliminar ningún nodo	85627	25625	1980
	-100 top grado	81817	17642	13773
	-100 top intermediación	85664	25535	2033
	-100 top cercanía	81841	17497	13894
	-100 top evector	94652	4908	13672
Distancia a la respuesta <= 2	Ordenamiento por grado	4071	107181	5241
	Ordenamiento por intermediación	5249	106003	5241
	Ordenamiento por cercanía	4064	107188	5241
	Ordenamiento por evector	4169	107083	5241
	Ordenamiento Distancia semántica	355	110897	5241

En las visualizaciones de los grafos, los nodos con igual color representan igual grupo, el tamaño de los nodos hace referencia a su grado, y los enlaces son la co-ocurrencia de los nodos en la lista de problemas.[2]

La figura B.1 representa los grupos de la lista de problemas. Mayoritariamente se pueden observar tres grupos, uno azul, otro verde y otro naranja. También mayoritariamente hay una nube densa de nodos fuertemente conectados en el centro, y pequeños grupos se diferencian de ese centro, sin embargo sólo uno de esos pequeños grupos tienen nodos de colores diferente al azul, verde y naranja.

Haciendo el análisis por contexto, el número de nodos disminuye y su comportamiento varía de contexto a contexto. Por el análisis de la sección anterior, se sabe que los servicios de atención a la salud tienen mejor capacidad predictiva que la lista de problemas sin contexto. Haciendo la comparación visual de los grafos que representan las agrupaciones en los servicios de atención de salud, se pueden realizar las siguientes observaciones:

1. Los grafos con mejor capacidad predictiva son los representados con las figuras B.6 (Pediatría), B.5 (Neurología) y B.2 (Cardiología), a su vez son los grafos más grandes y con mayor variedad de agrupaciones (colores) diferentes en los nodos. Cuando el algoritmo tiene en cuenta el contexto, la lista de predicciones se reduce sólo a los conceptos más cercanos en el mismo grupo.

2. El grafo representado por la figura B.4 que corresponde a los servicios de endocrinología, nefrología y urología, tiene mejor capacidad predictiva en los servicios de endocrinología y nefrología, que en urología. Este es un grafo donde hay muchas conexiones entre los nodos, de ahí su densidad en el centro, y los grupos de nodos

[2] Puede encontrar las visualizaciones de todos los grafos en la url `https://piwica.github.io/resultadosTesisDMHIBA/`

Tab. 4.16: Capacidad predictiva de agrupamiento con contexto: Servicio de atención de salud

Servicio de atención de salud	Todas las ocurrencias				Sin repeticiones			
	Predicción exacta		Distancia a la respuesta <= 2		Predicción exacta		Distancia a la respuesta <= 2	
	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10
Cardiología adultos	0,872	0,876	0,953	0,954	0,859	0,860	0,953	0,953
Cardiología pediátrica	0,844	0,854	0,931	0,936	0,851	0,855	0,938	0,940
Dermatología	0,814	0,825	0,921	0,926	0,827	0,830	0,931	0,933
Endocrinología	0,831	0,841	0,933	0,937	0,836	0,839	0,940	0,941
Endocrinología pediátrica	0,829	0,840	0,937	0,941	0,826	0,836	0,928	0,932
Ginecoobstetricia	0,796	0,811	0,910	0,916	0,810	0,814	0,921	0,923
Nefrología adultos	0,894	0,897	0,963	0,964	0,842	0,844	0,941	0,941
Neurología adultos	0,880	0,884	0,954	0,955	0,886	0,888	0,960	0,961
Oftamología adultos	0,868	0,873	0,948	0,950	0,874	0,875	0,954	0,955
Otorrinolaringología	0,853	0,859	0,939	0,942	0,845	0,848	0,939	0,940
Pediatría	0,893	0,897	0,962	0,964	0,895	0,896	0,964	0,965
Psiquiatría	0,864	0,869	0,950	0,952	0,864	0,866	0,953	0,954
Traumatología	0,803	0,814	0,915	0,920	0,786	0,792	0,913	0,916
Urología	0,794	0,806	0,911	0,916	0,742	0,751	0,895	0,899

Tab. 4.17: Capacidad predictiva de agrupamiento con contexto: Nivel asistencial

Nivel de asistencia	Todas las ocurrencias				Sin repeticiones			
	Predicción exacta		Distancia a la respuesta <= 2		Predicción exacta		Distancia a la respuesta <= 2	
	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10
Ambulatorio	0,137	0,593	0,407	0,721	0,137	0,564	0,408	0,701
Episodio Ambulatorio	0,099	0,543	0,347	0,669	0,125	0,411	0,486	0,654
Guardia	0,149	0,546	0,432	0,697	0,165	0,554	0,439	0,700
Internación Domiciliaria	0,181	0,497	0,467	0,673	0,311	0,516	0,568	0,697
Internación General	0,143	0,560	0,408	0,696	0,269	0,467	0,611	0,716
Internación geriátrica	0,189	0,489	0,410	0,628	0,317	0,480	0,513	0,629
Triage	0,146	0,546	0,428	0,696	0,354	0,448	0,732	0,771

que se forman en la periferia tienen relación con los servicios de endocrinología y nefrología y no con urología. Por ejemplo, los problemas asociados a diabetes mellitus, neoplasia de riñón y hematuria.

3. Enfermedad sospechada, dolor y antecedente familiar de trastorno son problemas con un valor de grado muy alto en casi todos los contextos.

4. En el servicio de dermatología, todos los problemas pertenecen al mismo grupo, ver figura B.3. Así mismo es uno de los grafos más pequeños.

Los grafos con el contexto del nivel asistencial o ámbito, permiten hacer las siguientes observaciones:

1. El grafo más simple es el de la figura B.8 (Episodio Ambulatorio), y también el que tiene menor capacidad predictiva.

Tab. 4.18: Capacidad predictiva de agrupamiento con contexto: Grupo etario

Grupo Etario	Todas las ocurrencias				Sin repeticiones			
	Predicción exacta		Distancia a la respuesta <= 2		Predicción exacta		Distancia a la respuesta <= 2	
	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10	P@10	Acc@10
0-4	0,831	0,838	0,960	0,962	0,850	0,853	0,966	0,967
15-24	0,900	0,903	0,978	0,979	0,910	0,911	0,983	0,983
25-34	0,876	0,880	0,974	0,975	0,889	0,891	0,979	0,979
35-44	0,887	0,890	0,975	0,976	0,897	0,898	0,980	0,980
45-54	0,892	0,895	0,977	0,978	0,901	0,902	0,982	0,982
55-64	0,899	0,902	0,980	0,981	0,909	0,911	0,984	0,984
65-74	0,919	0,921	0,983	0,984	0,927	0,928	0,987	0,987
75-101	0,942	0,943	0,989	0,989	0,948	0,948	0,992	0,992

2. El grafo más complejo es el de la figura B.7 (Ambulatorio), pero su capacidad predictiva no es muy diferente a los demás niveles asistenciales.

3. Los grafos de los niveles de asistencia de internación domiciliaria e internación general, son los mismos. Ver figuras B.9 y B.10.

En todos los grafo con el contexto del grupo etario se pueden observar una gran variedad de agrupaciones más pequeñas, que las separan de un centro más complejo y denso de problemas. También se observa una gran variedad de colores, los cuales significan que los métodos de agrupación han encontrado asociaciones significativas entre problemas.

El grafo de la figura B.11 que representa los problemas con pacientes entre los 0 y 4 años es el más grande y complejo. El grafo de la figura B.12 que representa los problemas con pacientes entre los 75 y 101 años aunque es el más simple tiene la mejor capacidad predictiva.

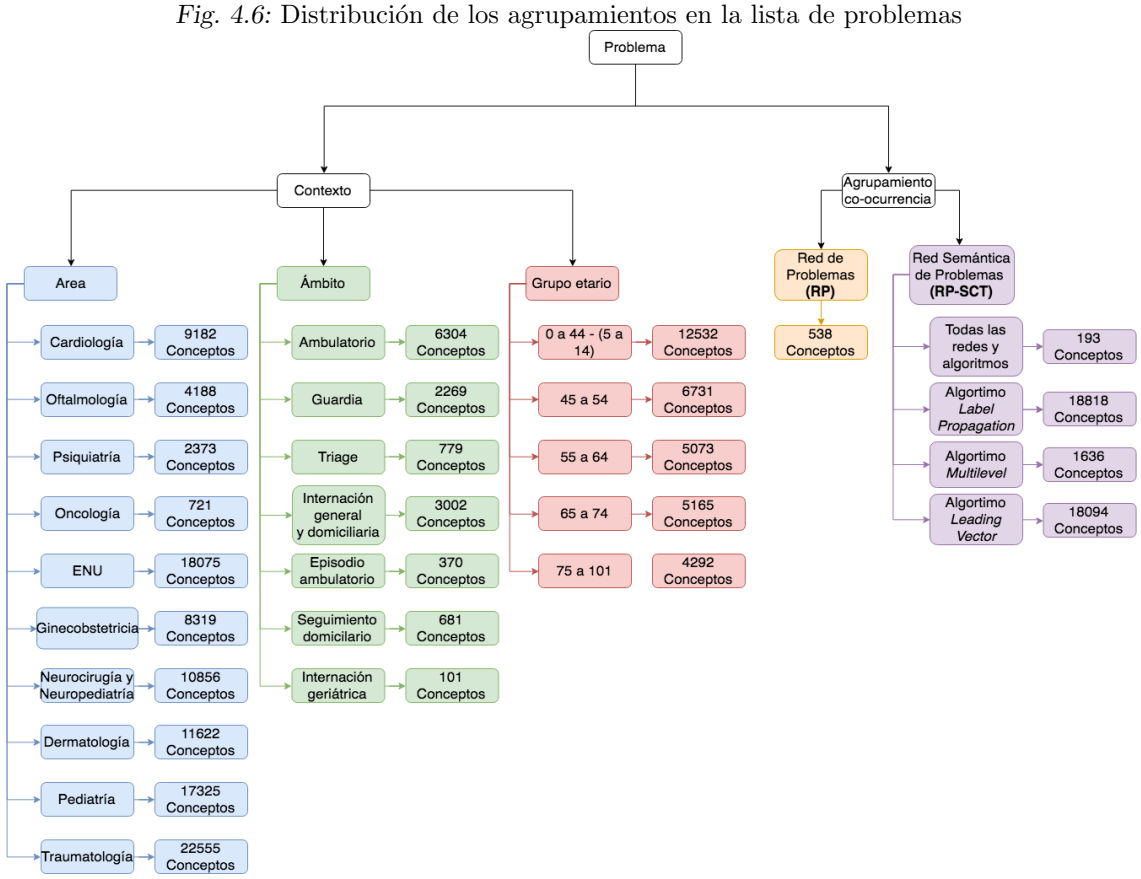
4.5. Discusión del capítulo

El objetivo de este capítulo fue realizar un análisis de *graph mining* a la lista de problemas del HIBA con el fin identificar patrones de red y aplicar algoritmos de agrupamiento que permitan encontrar comunidades.

En la identificación de patrones de red se evaluaron dos redes. La primera es una red formada sólo por los conceptos de la lista problemas y sus co-ocurrencias en los individuos **(RP)**. En la segunda se extiende **(RP)** con sus conexiones jerárquicas $|ES\_UN|$ a los concepto de Snomed CT **(RP-SCT)**. La RP tiene un mejor ajuste a la ley de potencias que la RP-SCT. Esto debido a que la RP-SCT tiene una cola mucho más pesada que la RP.

El siguiente aspecto evaluado fue los efectos de comunidad en los grafos. La diferencia de las métricas de transitividad y la transitividad local promedio, es que la primera es la medida de la transitividad local en toda la red, en la segunda se calcula por cada vértice la transitividad y los nodos con menos de dos vecinos son considerados como de transitividad cero (Watts y Strogatz, 1998). Las diferencias encontradas en estas dos métricas en las dos redes, indican una alta presencia de nodos que no forman tríadas. Aunque los valores

Fig. 4.6: Distribución de los agrupamientos en la lista de problemas



de la longitud media del camino mínimo entre nodos entre las dos redes es similar. Para un alto coeficiente de agrupamiento como es el caso de RP, el efecto es que la mayoría de los nodos que son homogéneos se encuentran en pocos saltos dentro de la red. Este efecto es similar al fenómeno de mundo pequeño en las redes sociales (Cook y Holder, 2006).

Se evidencia que la RP tiene estructuras más fuertes que la RP-SCT, además que los nodos están más conectados (promedio grados más altos y longitudes medias más pequeñas).

Para detectar comunidades se usaron los algoritmos de agrupamiento *leading vector*, *label propagation* y *multilevel*. Las características de estos algoritmos son diferentes y están descriptas en el capítulo 2 (Ver sección 2.2). La complejidad de estos algoritmos permite que puedan ser empleados en redes de larga escala y con grafos que no están conectados. Según los resultados obtenidos en la modularidad, *Leading vector* es un algoritmo divisivo cuya modularidad se ve afectado por el tamaño del grafo y *label propagation* va obteniendo mejores modularidades en la medida que crece el grafo. El algoritmo *multilevel* es el que mejores modularidades tiene y el número de grupos permanece constante sin importar que se modifique el tamaño del grafo.

Para evaluar la capacidad predictiva de los grupos obtenidos usé los datos de la lista de problemas en el año 2017. Lo que evalué era si en la lista de problemas del 2017 había coincidencia con los 10 problemas más cercanos por medidas de centralidad o por distancia semántica. Los problemas más cercanos debían coincidir en el mismo grupo de los que están en la lista de problemas del paciente. Las medidas estadísticas que usé fueron precisión y

exactitud.

Cuando las predicciones son exactas, los resultados con precisiones más altas se logran con 20 nodos excluidos en grados, cercanía y autovector. En el caso de intermediación aumenta en la medida que más se filtren datos, con un máximo local en 400 nodos. Esto es por la relación entre falsos positivos y verdaderos positivos, donde los filtros por grado, cercanía y autovector tienen mejores resultados porque tienen menos falsos positivos, y en el caso de la intermediación porque tienen más verdaderos positivos.

En el caso de las mediciones sin repeticiones se observa una exactitud similar en todos los experimentos y que hay una mejora no significativa en la precisión. A diferencia de los experimentos con repeticiones donde se obtiene una exactitud máxima de 0.397, en el caso de los experimentos sin repeticiones la exactitud máxima es de 0.255, esto se debe a que los verdaderos negativos cuentan con muchas repeticiones, en algunos casos reduciéndose a más de la mitad.

En el caso del ordenamiento por distancia semántica, los resultados mejoran en la precisión. Incluso sin realizar ningún filtro por medidas de centralidad, pero en cuanto a la medida de exactitud, se obtienen mejores resultados en el ordenamiento por centralidad. Por lo tanto, las medidas de centralidad permiten identificar mejor los verdaderos negativos que las distancias semánticas.

Cuando las predicciones no son exactas, sino que se toma como verdadero positivo si al menos hay una distancia de hasta dos nodos entre la respuesta y las predicciones, hay un incremento significativo de los verdaderos positivos. Por lo tanto, los valores de precisión y exactitud aumentan, presentándose los máximos en los casos en los que no se excluyen nodos. El mejor resultado se encuentra en el ordenamiento por distancia semántica de menor a mayor, con y sin repeticiones, los valores con repeticiones son P@10(0.995) y Acc@10(0.995) y sin repeticiones son P@10(0.997) y Acc@10(0.723). Aunque con las predicciones con distancias menores o iguales a dos nodos, se obtienen precisiones y exactitudes por encima del 0.900 se requiere de un análisis más detallado para evaluar si las predicciones son realmente verdaderos positivos.

El comportamiento diferente en los casos de intermediación se debe a que a diferencia del grado, cercanía y autovector, la intermediación no es una medida de cercanía, su valor está dado porque conecta otros pares de nodos. La interpretación es que al eliminar estos nodos entonces se eliminan los problemas que hacen parte de los síntomas subyacentes, pero no los problemas principales.

Las visualizaciones permiten explicar la razón de las mejoras en las métricas de los contextos de servicios de atención y del grupo etario. Se puede observar que a diferencia de los otros contextos, aquellos en los que se conserva complejidad (por el número de nodos y enlaces) y variedad de grupos (color en los nodos), tienen mejor capacidad predictiva que aquellos que se reducen mucho en tamaño o que tienen pocos grupos. Utilizando el contexto, en el caso con repeticiones la precisión máxima P@10 es del 0.942 y la exactitud máxima A@10 es de 0.943, y sin precisiones son P@10 de 0.948 y A@10 de 0.948. Estos valores fueron obtenidos en el contexto del grupo etario entre 75 y 101 años, son pacientes que por sus características tienen muchas enfermedades crónicas y varios registros en sus lista de problemas.

4.6. Conclusión del capítulo

En este capítulo se analizaron el grafo creados a partir de la lista de problemas y la co-ocurrencia en los pacientes (Red RP) y el grafo extendido con sus conexiones con Snomed CT (Red RP-SCT).

El primer análisis fue evaluar diferentes patrones de redes en estos grafos. Concluí que los grafos tienen un mejor ajuste a la ley de potencias que a otras distribuciones y que se evidencian efectos de comunidad. La utilidad de modelar la lista de problemas como un grafo radica en la posibilidad de detectar comunidades significativas que permitan realizar clasificaciones, detectar nodos influyentes y recuperar otros nodos del grupo usando algún criterio de proximidad o semejanza.

En la detección de comunidades significativas se aplicaron los algoritmos de agrupamiento *Leading vector*, *Multilevel* y *Label Propagation*. Esta detección se realizó en la la red RP y en la red RP-SCT conformada por problemas que tengan registros de al menos 5, 10, 100, 1000 y 10 000 pacientes con registros antes de diciembre de 2016. Esto tenía el propósito de encontrar comunidades que se repitieran a pesar de los enfoques diferentes de los algoritmos y de los tamaños de los grafos.

Los resultados permitieron encontrar comunidades, problemas que se detectaron como atípicos debido a su distancia semántica con los otros miembros del grupo, y asociaciones novedosas entre problemas semánticamente distantes. Los resultados también permitieron detectar grupos de problemas con un contexto similar en la red RP: Ej. Enfermedades generales (cluster_6), enfermedades relacionadas con la edad avanzada (cluster_27), enfermedades relacionada al embarazo (cluster_16), etc. Mientras que en la red RP-SCT se encontraron en su mayoría grupos pequeños pero que presentaban asociaciones entre pares de conceptos distantes semánticamente según su clasificación en la ontología de Snomed CT. Sin embargo al realizar una búsqueda bibliográfica encontré evidencias de la co-ocurrencia de estas enfermedades en pacientes.

El siguiente paso dentro de la metodología fue evaluar la capacidad predictiva de las comunidades encontradas. Para lo cual en el caso de los pacientes con registros en el 2017, tomé la lista de problemas hasta el 2016 para crear la *query*. La co-ocurrencia de los problemas de esta *query* dentro de las comunidades generó una lista de predicciones. La lista de predicciones es de tamaño 10 y su ordenamiento es por la medidas de centralidad o de distancias semánticas. También evalué filtrar de la lista, los nodos con mayores medidas de centralidad: grado, intermediación, cercanía y autovector. Las métricas evaluadas fueron precisión P@10 y exactitud Acc@10, tanto en predicciones exactas o si la distancia entre el problema del 2017 y la lista de predicciones es menor a dos nodos. Los resultados de estos experimentos dieron valores muy bajos, especialmente en el caso de las predicciones exactas donde la mayoría no supera en precisión el 0.100 y en exactitud el 0.300.

En la última sección, se filtraron las listas de sugerencias usando los refset de los contextos del área jerárquica, nivel asistencial y grupo etario. Las métricas mejoran significativamente en cada caso. En menor grado en el nivel asistencial donde los valores de precisión y exactitud se duplican, en promedio son P@10=0.150 y Acc@10=0.500 en las predicciones exactas. Y en mayor medida en el servicio de atención de salud y grupo etario. En estos últimos, algunos casos superan 0.900 en precisión y exactitud y la mayoría es ubica por encima de 0.800.

5. CONCLUSIONES Y FUTUROS PASOS

5.1. Conclusiones generales

En este trabajo he construido una clasificación de los conceptos de la lista de problemas, a partir de la co-ocurrencia de los problemas dentro de los pacientes y su registro asociado significativamente a diferentes contextos: servicio de atención de salud, nivel asistencial o ámbito y grupo etario. El objetivo de estos grupos o clasificaciones es contribuir al mejoramiento de la calidad de la lista de problemas de los pacientes en su completitud, actualización y granularidad, por medio de la recuperación contextual de información y el uso significativo de Snomed CT.

En la actualidad Snomed CT tiene un amplio cubrimiento de diferentes tópicos relacionados a la salud, es usado en muchos países y está traducido en diferentes idiomas. Uno de los aspectos claves en su uso significativo es la creación de refset o agrupamientos que puedan ser utilizados como vocabularios controlados dependientes de contextos, sin embargo Snomed CT no específica una metodología para la construcción de dichos agrupamientos.

En esta tesis, construí por medio de algoritmos de *graphmining* tres niveles diferentes de clasificación de conceptos de Snomed CT: Problemas asociados a niveles asistenciales, Problemas asociados a grupos etarios, Problemas asociados a servicios de atención de salud. Las relaciones entre los conceptos de estas clasificaciones están ponderados por su co-ocurrencia entre los pacientes del Hospital Italiano de Buenos Aires.

En el caso de los problemas asociados a servicios de atención de salud, realicé una validación de su cubrimiento con los refset publicados por el consorcio *Kaiser Permanente*. En el este capítulo repasaré las contribuciones de estas clasificaciones al proceso de la construcción de refset de Snomed CT, y presentaré algunas ideas sobre el trabajo futuro de esta línea de investigación.

Se aplicaron algoritmos de aprendizaje no supervisado al grafo construido con la red de problemas (RP) y sus conexiones semánticas son Snomed CT (RP-SCT). Los grupos que fueron consistentes en diferentes agrupamientos fueron evaluados. Se midió la precisión y exactitud de una lista de predicciones de 10 conceptos que intentaba predecir los problemas que fueron registrados a los pacientes en el año 2017, a partir de los registros de su lista de problemas previo al 31 diciembre de 2016.

Los refset de contextos permitieron acotar el universo de búsqueda, y mejoraron significativamente las métricas de precisión y exactitud.

El principal logro de esta tesis es la creación de estas agrupaciones, las cuales tienen múltiples aplicaciones entre las que se cuentan: la recuperación de información, el uso significativo de Snomed CT y la creación de refset con contexto. Espero que estas aplicaciones deriven en una mejora a la calidad de la lista de problemas.

5.2. Uso significativo de Snomed CT

El primer objetivo del uso significativo de Snomed CT es mantener actualizada la lista de problemas con diagnósticos actuales ya activos. Los recursos para lograr esto son los refset contextuales. Snomed CT presenta algunos públicos disponibles que sirvieron para evaluar el cubrimiento de los servicios de atención de salud.(IHTSDO, 2014)

La metodología para la construcción de los refset de Servicios de atención de salud y la comparación con otros experimentos realizados por nova scotia(Kuropatwa y Giannangelo, 2016) fueron presentados en el congreso Medical Informatics Europe 2018.(Ávila y cols., 2018)

Con el desarrollo de esta tesis, concluimos que la construcción de estos refset mejora significativamente la recuperación de información en la lista de problemas, que la precisión y exactitud mejoró significativamente con el uso de refset de contexto.

Los servicios del Hospital Italiano de Buenos Aires, son usados por historias clínicas electrónicas de Argentina, Uruguay y Chile. En trabajos futuros, sería muy beneficioso realizar un consenso entre el uso de las listas de problemas de todos estos centros médicos, para establecer una única lista de problemas controlada que además esté ponderada con la frecuencia de uso en las bases de datos clínicas. Un trabajo similar fue realizado por Kaiser Permanente. Como resultado Kaiser Permanente ha liberado una única lista de problemas desde el 2009 para ser usado en todas las clínicas de su consorcio y mantener la interoperabilidad(Dolin y cols., 2004). El trabajo futuro que se propone tendría la validez internacional de los países del cono sur.

5.3. Distancias semánticas

Estimar la distancia semántica entre términos es una de las herramientas más ampliamente usadas para el procesamiento y entendimiento de textos. En esta tesis fue usada para cuantificar la distancia entre los términos predichos y los términos de prueba, asumiendo que las distancias de tamaño dos eran lo máximo aceptable. La distancia fue medida con el camino más corto entre los dos nodos del grafo.

Una línea de investigación que se desprende de esta tesis es mejorar la manera como se calculan la distancias semánticas. Existen varias metodologías propuestas para medir la similaridad entre dos conceptos en una ontología. Específicamente los trabajos con Snomed CT se pueden dividir en dos categorías, enfoques basados en conocimiento y enfoques basados en corpus. (Mabotuwana, Lee, y Cohen-Solal, 2013; Ben Aouicha y Hadj Taieb, 2016; Sánchez y Batet, 2011; Harispe, Sánchez, Ranwez, Janaqi, y Montmain, 2014)

El enfoque basado en conocimiento explota principalmente la estructura jerárquica y las relaciones semánticas de la ontología para hacer inferencias sobre el conocimiento. Este enfoque mide la distancia entre conceptos usando técnicas como el camino más corto, conteo de aristas, profundidad ontológica y ancestro común más bajo *(Lowest Common Subsumer)*. La similaridad es determinada como el inverso de la distancia en su forma más simple, o alguna otra función matemática basada en la distancia ontológica.(Mabotuwana y cols., 2013)

El enfoque basado en corpus usa un gran corpus con textos específicos del dominio para determinar el valor de la información de cada concepto. Este valor se determina por la frecuencia del concepto en el corpus. Los menos frecuentes son vistos como más informativos que los más comunes.(Mabotuwana y cols., 2013)

No hay una métrica *gold estandar* para determinar la similaridad entre dos conceptos de una ontología. En un trabajo futuro se evaluarán los dos enfoques, dado que se cuenta con la ontología y el corpus de la lista de problemas. Esta línea de investigación tiene el objetivo de hallar una métrica con cierto nivel de confianza que determine cuándo dos conceptos son clínicamente similares.

5.4. Implementación en una Historia Clínica Electrónica

La implementación de los modelos en la historia clínica electrónica sugiere el completo desarrollo de un proyecto de software. Según los resultados teóricos obtenidos en esta tesis la línea futura tendrá el diseño de un experimento observacional de corte transversal con la implementación siguiendo los siguientes criterios de inclusión: áreas jerárquicas agrupadas por servicios de atención de salud: Cardiología, dermatología, endocrinología, nefrología, neurología, oftalmología, otorrinolaringología, pediatría, pisquiatría y traumatología, donde los valores de P@10 y Acc@10 superan el 0.800. También cuando el paciente pertenece a los grupos etarios: 15-24, 55-64 y 75-101, donde los valores de P@10 y Acc@10 superan el 0.900.

Para el desarrollo de proyectos de software, el departamento de informática del hospital sigue los fundamentos de gestión de proyectos. Se realiza un Project Charter donde se define el alcance, riesgos, la metodología y entregables. También una Estructura de Desglose del Trabajo (EDT) con el cronograma. Por otro lado, el hospital requiere que el Comité de Ética de Protocolos de Investigación (CEPI) avale el protocolo de investigación.

Una vez se han alcanzado los avales y el proyecto ingresa dentro del portfolio del departamento de informática, se da inicio al análisis, diseño, desarrollo y pruebas. Este ciclo de software se realiza iterativamente usando la metodología ágil scrum.

Apéndice

A. GRUPOS DE PROBLEMAS

Tab. A.1: Top de 10 conceptos de los agrupamientos con las mayores distancias semánticas en la red de problemas

Cluster(tamaño)	Grado		Cercanía		Intermedicación	
	Concepto	Valor	Concepto	Valor	Concepto	Valor
cluster_1(3)	QUIMIOTERAPIA	1.862	QUIMIOTERAPIA	0,0715	QUIMIOTERAPIA	42,33
	CARCINOMA ENDOMETRIAL	680	CARCINOMA ENDOMETRIAL	0,0713	CARCINOMA ENDOMETRIAL	18,63
	ADENOCARCINOMA DE ENDOMETRIO	598	ADENOCARCINOMA DE ENDOMETRIO	0,0713	ADENOCARCINOMA DE ENDOMETRIO	5,53
cluster_6(84)	FIEBRE	51.606	FIEBRE	0,0744	RESFRÍO COMÚN	2.307,62
	TOS	29.334	TOS	0,0737	FARINGITIS	1.395,75
	BRONCOESPASMO	24.608	VÓMITOS	0,0734	TOS CON FIEBRE	325,83
	VÓMITOS	22.430	DIARREA	0,0734	AMIGDALECTOMÍA	237,49
	DIARREA	21.892	IRRITACIÓN DEL OJO	0,0733	EXACERBACIÓN AGUDA DE ASMA	217,75
	IRRITACIÓN DEL OJO	19.910	FARINGITIS	0,0731	FORÚNCULO	149,61
	NEUMONÍA	18.860	ERUPCIÓN CUTÁNEA	0,0731	ENFERMEDAD HEPÁTICA INFLAMATORIA	124,90
	FARINGITIS	17.538	GASTROENTERITIS	0,0730	NEUMONITIS	106,20
	ERUPCIÓN CUTÁNEA	17.402	TRAUMA CEREBRAL	0,0730	INSUFICIENCIA DE LA PELÍCULA LAGRIMAL	94,53
	TRAUMA CEREBRAL	16.920	CERVICODINIA	0,0729	CONJUNTIVITIS	85,14
cluster_10(45)	MALESTAR GENERAL	46.096	MALESTAR GENERAL	0,0747	INMUNODEFICIENCIA COMBINADA SEVERA	455,20
	SÍNDROME DE INSUFICIENCIA RENAL CRÓNICA	15.020	APOYO NUTRICIONAL	0,0729	MALESTAR GENERAL	155,88
	APOYO NUTRICIONAL	14.356	SÍNDROME DE INSUFICIENCIA RENAL CRÓNICA	0,0726	ABDOMEN AGUDO	149,82
	ADMINISTRACIÓN DE ANTICOAGULANTE	8.552	ADMINISTRACIÓN DE ANTICOAGULANTE	0,0723	NEUROPATÍA DESMIELINIZANTE	85,62
	SÍNDROME DE INSUFICIENCIA RENAL	8.286	SÍNDROME DE INSUFICIENCIA RENAL	0,0722	INYECCIÓN DE GAMMAGLOBULINA	77,27
	ABDOMEN AGUDO	6.080	ABDOMEN AGUDO	0,0721	POLIRRADICULOPATÍA DESMIELINIZANTE INFLAMATORIA CRÓNICA	32,00
	HALLAZGO POSPROCEDIMIENTO	5.240	HALLAZGO POSPROCEDIMIENTO	0,0719	PERITONITIS	21,08
	TRATAMIENTO CON ANTIBIÓTICOS	4.142	TRATAMIENTO CON ANTIBIÓTICOS	0,0718	HALLAZGO POSPROCEDIMIENTO	20,57
	INSUFICIENCIA RENAL CRÓNICA REAGUDIZADA	3.704	DIABETES MELLITUS TIPO 1	0,0718	SÍNDROME DE INSUFICIENCIA RENAL CRÓNICA	18,50
	DIABETES MELLITUS TIPO 1	3.494	INSUFICIENCIA RENAL CRÓNICA REAGUDIZADA	0,0718	GLOMERULONEFRITIS	16,66
cluster_16(34)	PACIENTE ACTUALMENTE EMBARAZADA	17.324	MAREO	0,0730	AMENAZA DE TRABAJO DE PARTO PREMATURO (TRASTORNO)	283,58
	ANSIEDAD	14.658	ANSIEDAD	0,0728	INTOXICACIÓN POR FÁRMACO Y/O SUSTANCIA MEDICINAL	267,73
	MAREO	14.348	GRIPE	0,0726	SANGRADO VAGINAL	171,41
	PUERPERIO	10.694	PACIENTE ACTUALMENTE EMBARAZADA	0,0725	DOLOR SUPRAPÚBICO	142,86
	GRIPE	10.450	CAMBIO EN LOS SÍNTOMAS GINECOLÓGICOS	0,0722	TRASTORNO MENTAL	116,94
	TRASTORNO MENTAL	9.632	TRASTORNO MENTAL	0,0722	ÚTERO UNICORNE	114,16
	CAMBIO EN LOS SÍNTOMAS GINECOLÓGICOS	7.698	PUERPERIO	0,0722	INSOMNIO	85,10

Table A.1 Continúa de página anterior

Cluster(tamaño)	Grado		Cercanía		Intermedicación	
	Concepto	Valor	Concepto	Valor	Concepto	Valor
cluster_22(6)	CONTRACCIONES UTERINAS PRESENTES	7.618	CÓLICO RENAL	0,0722	CAMBIO EN LOS SÍNTOMAS GINECOLÓGICOS	55,21
	SANGRADO VAGINAL	7.576	SANGRADO VAGINAL	0,0721	CÓLICO RENAL	37,12
	CÓLICO RENAL	7.036	LIPOTIMIA	0,0721	PACIENTE ACTUALMENTE EMBARAZADA	35,84
	HEMOPTISIS	2.982	HEMOPTISIS	0,0717	CARCINOMA DE LARINGE	15,64
	DESNUTRICIÓN	2.722	DESNUTRICIÓN	0,0717	FIEBRE POSOPERATORIA	14,88
	DIABETES MELLITUS SECUNDARIA	1.762	DIABETES MELLITUS SECUNDARIA	0,0716	DESNUTRICIÓN	14,63
	FIEBRE POSOPERATORIA	1.420	FIEBRE POSOPERATORIA	0,0715	HEMOPTISIS	7,98
	TUMOR MALIGNO DE LA LARINGE	1.126	TUMOR MALIGNO DE LA LARINGE	0,0714	TUMOR MALIGNO DE LA LARINGE	4,80
	CARCINOMA DE LARINGE	734	CARCINOMA DE LARINGE	0,0713	DIABETES MELLITUS SECUNDARIA	3,83
cluster_27(59)	HIPERTENSIÓN ARTERIAL	30.020	HIPERTENSIÓN ARTERIAL	0,0740	COLANGITIS AGUDA	804,66
	LUMBALGIA	22.022	LUMBALGIA	0,0734	COGNICIÓN ALTERADA	386,24
	DISLIPIDEMIA	15.682	DISLIPIDEMIA	0,0732	DISLIPIDEMIA	182,20
	DOLOR DE HOMBRO	12.770	DOLOR DE HOMBRO	0,0729	CÁLCULO RENAL	171,41
	ACCIDENTE CEREBROVASCULAR	12.476	HIPOTIROIDISMO	0,0728	GOTA	155,88
	HIPOTIROIDISMO	11.264	TRASTORNO DE CONDUCCIÓN CARDÍACA	0,0726	DISFAGIA	145,24
	TRASTORNO DEPRESIVO	11.188	CAÍDA ACCIDENTAL	0,0726	LESIÓN TRAUMÁTICA DE LA PIERNA	128,45
	CAÍDA ACCIDENTAL	11.016	TRASTORNO DEPRESIVO	0,0726	ANGINA DE PECHO	114,08
	TROMBOSIS VENOSA PROFUNDA	11.014	ACCIDENTE CEREBROVASCULAR	0,0726	ANCIANO FRÁGIL	106,45
	TRASTORNO DE CONDUCCIÓN CARDÍACA	10.130	INDIGESTIÓN	0,0725	DOLOR DE HOMBRO	105,73
cluster_30(77)	SERVICIO DE MEDICINA TRANSFUSIONAL	27.586	PROCEDIMIENTO DE FISIOTERAPIA	0,0736	HIPOMAGNESEMIA	143,92
	PROCEDIMIENTO DE FISIOTERAPIA	25.532	SERVICIO DE MEDICINA TRANSFUSIONAL	0,0733	LEUCEMIA LINFOBLÁSTICA AGUDA COMÚN	121,80
	ENFERMEDAD INFECCIOSA DE LAS VÍAS URINARIAS INFERIORES	9.372	ENFERMEDAD INFECCIOSA DE LAS VÍAS URINARIAS INFERIORES	0,0725	SEPSIS	57,29
	SEPSIS	7.298	SEPSIS	0,0721	TUMOR MALIGNO DE TESTÍCULO	56,80
	NEUTROPENIA FEBRIL	6.320	TRASTORNO DE ADAPTACIÓN	0,0721	CARCINOMA DE MAMA	40,16
	TRASTORNO DE ADAPTACIÓN	5.836	HEMORRAGIA DIGESTIVA ALTA	0,0720	LEUCEMIA	32,11
	HEMORRAGIA DIGESTIVA ALTA	5.792	NEUTROPENIA FEBRIL	0,0719	REEMPLAZO TOTAL DE CADERA	30,02
	LINFOMA MALIGNO (CLÍNICO)	5.228	HALLAZGO RELACIONADO CON UN PROCEDIMIENTO	0,0719	NEUTROPENIA FEBRIL	27,73
	HALLAZGO RELACIONADO CON UN PROCEDIMIENTO	5.090	CIRROSIS HEPÁTICA	0,0718	TRASPLANTE DE HÍGADO	26,30
	CIRROSIS HEPÁTICA	4.632	LINFOMA MALIGNO (CLÍNICO)	0,0718	ENFERMEDAD INFECCIOSA DE LAS VÍAS URINARIAS INFERIORES	25,85
cluster_34(22)	DOLOR ABDOMINAL	35.502	DOLOR ABDOMINAL	0,0741	CÁLCULO VESICAL	121,93
	CEFALEA	20.358	CEFALEA	0,0733	DOLOR EPIGÁSTRICO	66,69
	LESIÓN TRAUMÁTICA DEL PIE	11.986	LESIÓN TRAUMÁTICA DEL PIE	0,0727	MASTITIS	45,56
	CONVULSIÓN	10.288	DOLOR EPIGÁSTRICO	0,0725	DOLOR ABDOMINAL	30,75
	DOLOR EPIGÁSTRICO	8.790	CONVULSIÓN	0,0724	FATIGA	21,56
	VÉRTIGO	5.754	VÉRTIGO	0,0722	HALLAZGO RELACIONADO CON EL VÓMITO	21,47
	EPILEPSIA	5.306	NÁUSEAS Y VÓMITOS	0,0720	CEFALEA	18,53

Table A.1 Continúa de página anterior

Cluster(tamaño)	Grado		Cercanía		Intermedicación	
	Concepto	Valor	Concepto	Valor	Concepto	Valor
cluster_36(25)	NÁUSEAS Y VÓMITOS	4.946	FATIGA	0,0720	MONONUCLEOSIS INFECCIOSA	17,00
	FATIGA	4.688	EPILEPSIA	0,0720	COLOCACIÓN DE UNA SONDA NASOGÁSTRICA	14,20
	MASTITIS	2.876	COLESTASIS	0,0717	HIPERAMONEMIA	5,94
	DISNEA	17.886	SÍNDROME DE DEPENDENCIA DEL TABACO	0,0732	ANEMIA FERROPÉNICA	575,25
	SÍNDROME DE DEPENDENCIA DEL TABACO	14.358	DISNEA	0,0730	AFASIA	46,02
	INSUFICIENCIA CARDÍACA CONGESTIVA	12.482	DORSALGIA (HALLAZGO)	0,0726	TUMOR DE KLATSKIN	30,90
	INSUFICIENCIA CARDÍACA	10.070	INSUFICIENCIA CARDÍACA CONGESTIVA	0,0724	EXACERBACIÓN AGUDA DE ENFERMEDAD PULMONAR OBSTRUCTIVA CRÓNICA	30,75
	DORSALGIA (HALLAZGO)	9.774	FIBRILACIÓN AURICULAR	0,0723	INSUFICIENCIA CARDÍACA CRÓNICA	18,50
	FIBRILACIÓN AURICULAR	8.682	INSUFICIENCIA CARDÍACA	0,0723	ARTERIOSCLEROSIS CORONARIA	17,81
	ARTERIOSCLEROSIS CORONARIA	6.636	ARTERIOSCLEROSIS CORONARIA	0,0721	AFASIA COMBINADA	16,63
	EXACERBACIÓN AGUDA DE ENFERMEDAD PULMONAR OBSTRUCTIVA CRÓNICA	6.146	ANEMIA FERROPÉNICA	0,0721	FIBRILACIÓN AURICULAR	13,77
	TROMBOEMBOLIA PULMONAR	6.104	TROMBOEMBOLIA PULMONAR	0,0720	TROMBOEMBOLIA PULMONAR	12,51
	DERRAME PLEURAL	5.812	DERRAME PLEURAL	0,0719	ÚLCERA POR DECÚBITO	11,37

Tab. A.2: Top de 10 conceptos de los agrupamientos con las mayores distancias semánticas en la red de problemas unida a la red semántica de SNOMED CT

Cluster(tamaño)	Grado		Cercanía		Intermediación	
	Concepto	Valor	Concepto	Valor	Concepto	Valor
cluster_0(3)	INCISIÓN DE LA TRÁQUEA	2.116	INCISIÓN DE LA TRÁQUEA	0,0716	AMIGDALECTOMÍA	237,49
	AMIGDALECTOMÍA	1.326	AMIGDALECTOMÍA	0,0715	INCISIÓN DE LA TRÁQUEA	11,79
	CIRUGÍA DE CATARATAS	1.208	CIRUGÍA DE CATARATAS	0,0715	CIRUGÍA DE CATARATAS	1,58
cluster_1(2)	DERRAME PLEURAL	5.812	DERRAME PLEURAL	0,0719	ENFERMEDAD PULMONAR OBSTRUCTIVA CRÓNICA, GRAVE	5,95
	ENFERMEDAD PULMONAR OBSTRUCTIVA CRÓNICA, GRAVE	1.060	ENFERMEDAD PULMONAR OBSTRUCTIVA CRÓNICA, GRAVE	0,0714	DERRAME PLEURAL	3,36
cluster_2(2)	FÍSTULA TRAQUEOESOFÁGICA	148	FÍSTULA TRAQUEOESOFÁGICA	0,0711	ÚTERO UNICORNE	114,16
	ÚTERO UNICORNE	36	ÚTERO UNICORNE	0,0706	FÍSTULA TRAQUEOESOFÁGICA	2,23
cluster_3(2)	FRACTURA DE FÉMUR	2.636	FRACTURA DE FÉMUR	0,0716	OSTEOGENIA IMPERFECTA	12,19
	OSTEOGENIA IMPERFECTA	50	OSTEOGENIA IMPERFECTA	0,0709	FRACTURA DE FÉMUR	2,83
cluster_6(2)	HIPERCORTISOLISMO	496	HIPERCORTISOLISMO	0,0713	TUMOR DE KLATSKIN	30,90
	TUMOR DE KLATSKIN	370	TUMOR DE KLATSKIN	0,0712	HIPERCORTISOLISMO	2,19
cluster_7(2)	PARASITISMO INTESTINAL	1.666	PARASITISMO INTESTINAL	0,0715	PARASITISMO INTESTINAL	31,32
	LITIASIS COLEDOCIANA	1.432	LITIASIS COLEDOCIANA	0,0715	LITIASIS COLEDOCIANA	7,59
cluster_8(2)	DIARREA CRÓNICA	3.406	DIARREA CRÓNICA	0,0718	FIBROSIS QUÍSTICA	36,17
	FIBROSIS QUÍSTICA	492	FIBROSIS QUÍSTICA	0,0712	DIARREA CRÓNICA	2,00
cluster_10(4)	GASTROENTERITIS	15.090	GASTROENTERITIS	0,0730	HIPOSPADIAS	7,70
	ESTREÑIMIENTO	11.590	ESTREÑIMIENTO	0,0726	ESTREÑIMIENTO	6,74
	HIPOSPADIAS	530	HIPOSPADIAS	0,0712	DISGENESIA GONADAL	2,28
	DISGENESIA GONADAL	58	DISGENESIA GONADAL	0,0709	GASTROENTERITIS	1,61
cluster_13(2)	VÓMITOS	22.430	VÓMITOS	0,0734	DIARREA	21,82
	DIARREA	21.892	DIARREA	0,0734	VÓMITOS	16,75
cluster_14(2)	ENFERMEDAD INFECCIOSA DE LAS VÍAS URINARIAS INFERIORES	9.372	ENFERMEDAD INFECCIOSA DE LAS VÍAS URINARIAS INFERIORES	0,0725	ENFERMEDAD INFECCIOSA DE LAS VÍAS URINARIAS INFERIORES	25,85
	HEMATOMA RETROPERITONEAL	542	HEMATOMA RETROPERITONEAL	0,0713	HEMATOMA RETROPERITONEAL	5,63
cluster_15(2)	RESFRÍO COMÚN	10.414	RESFRÍO COMÚN	0,0726	RESFRÍO COMÚN	2.307,62
	CONGESTIÓN NASAL	6.282	CONGESTIÓN NASAL	0,0722	CONGESTIÓN NASAL	3,96
cluster_17(2)	PUERPERIO	10.694	PUERPERIO	0,0722	EMBARAZO GEMELAR	8,27
	EMBARAZO GEMELAR	1.340	EMBARAZO GEMELAR	0,0714	PUERPERIO	3,29
cluster_19(3)	ACCIDENTE CEREBROVASCULAR	12.476	ACCIDENTE CEREBROVASCULAR	0,0726	DEMENCIA	20,17
	DEMENCIA	7.420	DEMENCIA	0,0721	ENFERMEDAD DE ALZHEIMER	8,17
	ENFERMEDAD DE ALZHEIMER	2.232	ENFERMEDAD DE ALZHEIMER	0,0716	ACCIDENTE CEREBROVASCULAR	5,21
cluster_22(2)	NEUTROPENIA FEBRIL	6.320	NEUTROPENIA FEBRIL	0,0719	HIPOMAGNESEMIA	143,92
	HIPOMAGNESEMIA	1.146	HIPOMAGNESEMIA	0,0715	NEUTROPENIA FEBRIL	27,73
cluster_26(3)	VULVOVAGINITIS	5.654	VULVOVAGINITIS	0,0721	VULVOVAGINITIS	32,83
	LEIOMIOMA DEL ÚTERO	2.934	LEIOMIOMA DEL ÚTERO	0,0717	BARTOLINITIS	6,25

Table A.2 Continúa de página anterior

Cluster(tamaño)	Grado		Cercanía		Intermediación	
	Concepto	Valor	Concepto	Valor	Concepto	Valor
cluster_29(2)	BARTOLINITIS	1.330	BARTOLINITIS	0,0715	LEIOMIOMA DEL ÚTERO	2,85
	INSUFICIENCIA CARDÍACA DIASTÓLICA	760	INSUFICIENCIA CARDÍACA DIASTÓLICA	0,0713	ARTERITIS DE TAKAYASU	9,25
	ARTERITIS DE TAKAYASU	58	ARTERITIS DE TAKAYASU	0,0709	INSUFICIENCIA CARDÍACA DIASTÓLICA	5,89
cluster_30(2)	TUMOR MALIGNO DE LA VESÍCULA BILIAR	320	NEOPLASIA DE LA AMPOLLA DE VATER	0,0713	TUMOR MALIGNO DE LA VESÍCULA BILIAR	9,04
	NEOPLASIA DE LA AMPOLLA DE VATER	312	TUMOR MALIGNO DE LA VESÍCULA BILIAR	0,0712	NEOPLASIA DE LA AMPOLLA DE VATER	1,52
cluster_33(5)	CONJUNTIVITIS AGUDA	5.982	CONJUNTIVITIS AGUDA	0,0722	ORZUELO	80,30
	ORZUELO	4.644	ORZUELO	0,0720	CONJUNTIVITIS AGUDA	60,15
	CUERPO EXTRAÑO EN LA CÓRNEA	1.688	CUERPO EXTRAÑO EN LA CÓRNEA	0,0716	CUERPO EXTRAÑO EN LA CÓRNEA	3,15
	OTITIS MEDIA PURULENTA	1.248	UVEÍTIS	0,0715	UVEÍTIS	2,75
	UVEÍTIS	902	OTITIS MEDIA PURULENTA	0,0714	OTITIS MEDIA PURULENTA	2,16
cluster_34(3)	APOYO NUTRICIONAL	14.356	APOYO NUTRICIONAL	0,0729	INYECCIÓN DE GAMMAGLOBULINA	77,27
	TRATAMIENTO CON ANTIBIÓTICOS INTRAVENOSOS	1.392	TRATAMIENTO CON ANTIBIÓTICOS INTRAVENOSOS	0,0715	TRATAMIENTO CON ANTIBIÓTICOS INTRAVENOSOS	10,36
	INYECCIÓN DE GAMMAGLOBULINA	1.348	INYECCIÓN DE GAMMAGLOBULINA	0,0715	APOYO NUTRICIONAL	5,11
cluster_35(2)	OPERACIÓN CESÁREA	3.106	OPERACIÓN CESÁREA	0,0717	OPERACIÓN CESÁREA	6,82
	LIGADURA TUBARIA BILATERAL	302	LIGADURA TUBARIA BILATERAL	0,0712	LIGADURA TUBARIA BILATERAL	6,00
cluster_36(4)	FRACTURA DE FÉMUR PROXIMAL	7.180	FRACTURA DE FÉMUR PROXIMAL	0,0720	DEBILIDAD DE EXTREMIDAD	33,47
	DEBILIDAD DE EXTREMIDAD	1.784	DEBILIDAD DE EXTREMIDAD	0,0716	PARESIA DE LA EXTREMIDAD INFERIOR	17,41
	PARESIA DE LA EXTREMIDAD INFERIOR	1.384	PARESIA DE LA EXTREMIDAD INFERIOR	0,0715	DISCOPATÍA	4,64
	DISCOPATÍA	740	DISCOPATÍA	0,0714	FRACTURA DE FÉMUR PROXIMAL	1,39
cluster_37(2)	CÁLCULO VESICAL	1.280	CÁLCULO VESICAL	0,0715	CÁLCULO VESICAL	121,93
	ESTENOSIS PILÓRICA	692	ESTENOSIS PILÓRICA	0,0713	ESTENOSIS PILÓRICA	3,41
cluster_38(2)	ANASTOMOSIS ARTERIOVENOSA QUIRÚRGICA PARA DIÁLISIS RENAL	780	ANASTOMOSIS ARTERIOVENOSA QUIRÚRGICA PARA DIÁLISIS RENAL	0,0714	RESECCIÓN DE RIÑÓN	15,92
	RESECCIÓN DE RIÑÓN	320	RESECCIÓN DE RIÑÓN	0,0712	ANASTOMOSIS ARTERIOVENOSA QUIRÚRGICA PARA DIÁLISIS RENAL	5,65
cluster_39(6)	ENFERMEDAD HEPÁTICA INFLAMATORIA	2.222	ENFERMEDAD HEPÁTICA INFLAMATORIA	0,0716	ENFERMEDAD HEPÁTICA INFLAMATORIA	124,90
	NEOPLASIA MALIGNA SECUNDARIA DE HÍGADO	1.944	NEOPLASIA MALIGNA SECUNDARIA DE HÍGADO	0,0715	COLECCIÓN INTRABDOMINAL	30,02
	HIPOFUNCIÓN SUPRARRENAL	1.424	HIPOFUNCIÓN SUPRARRENAL	0,0715	HIPOFUNCIÓN SUPRARRENAL	12,33
	COLECCIÓN INTRABDOMINAL	730	COLECCIÓN INTRABDOMINAL	0,0713	NEOPLASIA MALIGNA PRIMARIA DEL RETROPERITONEO	5,35

Table A.2 Continúa de página anterior

Cluster(tamaño)	Grado		Cercanía		Intermediación	
	Concepto	Valor	Concepto	Valor	Concepto	Valor
	CARCINOMA DE VEJIGA	538	CARCINOMA DE VEJIGA	0,0713	CARCINOMA DE VEJIGA	2,27
	NEOPLASIA MALIGNA PRIMARIA DEL RETROPERITONEO	32	NEOPLASIA MALIGNA PRIMARIA DEL RETROPERITONEO	0,0702	NEOPLASIA MALIGNA SECUNDARIA DE HÍGADO	2,06
	ERUPCIÓN CUTÁNEA	17.402	ERUPCIÓN CUTÁNEA	0,0731	ERUPCIÓN CUTÁNEA	71,98
cluster_40(3)	VARICELA	5.332	VARICELA	0,0719	ERUPCIÓN PRURIGINOSA	16,43
	ERUPCIÓN PRURIGINOSA	3.382	ERUPCIÓN PRURIGINOSA	0,0718	VARICELA	5,85
cluster_41(2)	OBSTRUCCIÓN INTESTINAL	6.422	OBSTRUCCIÓN INTESTINAL	0,0720	NEOPLASIA MALIGNA PRIMARIA DE OVARIO	51,19
	NEOPLASIA MALIGNA PRIMARIA DE OVARIO	2.010	NEOPLASIA MALIGNA PRIMARIA DE OVARIO	0,0715	OBSTRUCCIÓN INTESTINAL	9,63
cluster_45(2)	FRACTURA DE EXTREMO SUPERIOR DE TIBIA	210	FRACTURA DE EXTREMO SUPERIOR DE TIBIA	0,0711	FRACTURA DE EXTREMO SUPERIOR DE TIBIA	10,49
	FRACTURA DE FÉMUR DISTAL	110	FRACTURA DE FÉMUR DISTAL	0,0704	FRACTURA DE FÉMUR DISTAL	2,11
	ABDOMEN AGUDO	6.080	ABDOMEN AGUDO	0,0721	ABDOMEN AGUDO	149,82
cluster_47(5)	INSUFICIENCIA RENAL CRÓNICA REAGUDIZADA	3.704	INSUFICIENCIA RENAL CRÓNICA REAGUDIZADA	0,0718	PERITONITIS	21,08
	PERITONITIS	1.964	PERITONITIS	0,0716	PUBERTAD PRECOZ	6,02
	VÁLVULAS URETRALES POSTERIORES CONGÉNITAS	310	VÁLVULAS URETRALES POSTERIORES CONGÉNITAS	0,0711	INSUFICIENCIA RENAL CRÓNICA REAGUDIZADA	3,10
	PUBERTAD PRECOZ	298	PUBERTAD PRECOZ	0,0711	VÁLVULAS URETRALES POSTERIORES CONGÉNITAS	2,25
	NEOPLASIA DE LA PELVIS	1.322	NEOPLASIA DE LA PELVIS	0,0715	TUMOR MALIGNO DE TESTÍCULO	56,80
cluster_48(5)	TUMOR MALIGNO DEL CUELLO UTERINO	1.228	TUMOR MALIGNO DEL CUELLO UTERINO	0,0714	ENFERMEDAD DE CAROLI	6,47
	TUMOR MALIGNO DE TESTÍCULO	486	TUMOR MALIGNO DE TESTÍCULO	0,0712	LINFOMA GÁSTRICO	5,38
	LINFOMA GÁSTRICO	126	LINFOMA GÁSTRICO	0,0711	NEOPLASIA DE LA PELVIS	2,69
	ENFERMEDAD DE CAROLI	42	ENFERMEDAD DE CAROLI	0,0710	TUMOR MALIGNO DEL CUELLO UTERINO	2,25
cluster_49(2)	CARCINOMA DE PARÉNQUIMA PULMONAR	646	CARCINOMA DE PARÉNQUIMA PULMONAR	0,0713	NEUMONÍA ASPIRATIVA RECURRENTE	4,39
	NEUMONÍA ASPIRATIVA RECURRENTE	514	NEUMONÍA ASPIRATIVA RECURRENTE	0,0712	CARCINOMA DE PARÉNQUIMA PULMONAR	1,92
cluster_52(2)	DIABETES MELLITUS TIPO 2	11.354	DIABETES MELLITUS TIPO 2	0,0726	DIABETES MELLITUS TIPO 2	21,61
	HIPOGLUCEMIA	2.892	HIPOGLUCEMIA	0,0717	HIPOGLUCEMIA	3,05
cluster_53(2)	NEUMONITIS	5.126	NEUMONITIS	0,0719	NEUMONITIS	106,20
	ADENOCARCINOMA DEL PULMÓN	1.850	ADENOCARCINOMA DEL PULMÓN	0,0715	ADENOCARCINOMA DEL PULMÓN	3,60
	PROCTORRAGIA	9.166	INDIGESTIÓN	0,0725	CÁLCULO RENAL	171,41
	INDIGESTIÓN	8.366	PROCTORRAGIA	0,0724	DISFAGIA	145,24
	DISFAGIA	7.016	VEJIGA: INCONTINENTE	0,0723	PROCTORRAGIA	94,53
cluster_56(8)	VEJIGA: INCONTINENTE	6.504	DISFAGIA	0,0722	VEJIGA: INCONTINENTE	8,14
	CÁLCULO RENAL	5.112	CÁLCULO RENAL	0,0720	SÍNDROME DE ICTERICIA COLESTÁSICA	7,89
	HEMORRAGIA DIGESTIVA BAJA	4.132	HEMORRAGIA DIGESTIVA BAJA	0,0718	PIELONEFRITIS	5,51
	SÍNDROME DE ICTERICIA COLESTÁSICA	3.554	SÍNDROME DE ICTERICIA COLESTÁSICA	0,0717	HEMORRAGIA DIGESTIVA BAJA	3,72

Table A.2 Continúa de página anterior

Cluster(tamaño)	Grado		Cercanía		Intermediación	
	Concepto	Valor	Concepto	Valor	Concepto	Valor
cluster_57(9)	PIELONEFRITIS	2.494	PIELONEFRITIS	0,0717	INDIGESTIÓN	1,87
	HEMORRAGIA DIGESTIVA ALTA	5.792	HEMORRAGIA DIGESTIVA ALTA	0,0720	CIRROSIS INFECCIOSA	7,87
	CIRROSIS HEPÁTICA	4.632	CIRROSIS HEPÁTICA	0,0718	TUMOR DE CÉLULAS GERMINALES	5,88
	COLECISTITIS	3.570	COLECISTITIS	0,0718	HEMORRAGIA DIGESTIVA ALTA	5,09
	ABSCESO PERIANAL	2.204	ABSCESO PERIANAL	0,0716	PROCTITIS	4,53
	CIRROSIS INFECCIOSA	1.172	HEPATITIS AUTOINMUNITARIA	0,0714	HEPATITIS AUTOINMUNITARIA	3,99
	HEPATITIS AUTOINMUNITARIA	940	CIRROSIS INFECCIOSA	0,0714	COLECISTITIS	3,83
	PROCTITIS	562	PROCTITIS	0,0714	ABSCESO PERIANAL	3,23
	TUMOR DE CÉLULAS GERMINALES	390	TUMOR DE CÉLULAS GERMINALES	0,0712	ATRESIA DEL ESÓFAGO	3,20
	ATRESIA DEL ESÓFAGO	386	ATRESIA DEL ESÓFAGO	0,0711	CIRROSIS HEPÁTICA	2,35
cluster_60(5)	CONJUNTIVITIS	14.220	CONJUNTIVITIS	0,0729	CONJUNTIVITIS	85,14
	OTITIS MEDIA AGUDA	8.890	OTITIS MEDIA AGUDA	0,0723	OTITIS	47,27
	OTITIS	5.932	OTITIS	0,0721	QUERATITIS	11,81
	QUERATITIS	4.798	QUERATITIS	0,0720	OTITIS MEDIA AGUDA	11,80
	ÚLCERA CORNEAL	4.338	ÚLCERA CORNEAL	0,0719	ÚLCERA CORNEAL	8,32
cluster_61(2)	TRATAMIENTO CON ANTIBIÓTICOS	4.142	TRATAMIENTO CON ANTIBIÓTICOS	0,0718	PLASMAFÉRESIS	10,97
	PLASMAFÉRESIS	468	PLASMAFÉRESIS	0,0712	TRATAMIENTO CON ANTIBIÓTICOS	2,45
cluster_64(3)	TRASTORNO DE CONDUCCIÓN CARDÍACA	10.130	TRASTORNO DE CONDUCCIÓN CARDÍACA	0,0726	ANGINA DE PECHO	114,08
	ANGINA DE PECHO	6.282	ANGINA DE PECHO	0,0722	FIBRILACIÓN AURICULAR CRÓNICA	6,33
	FIBRILACIÓN AURICULAR CRÓNICA	4.548	FIBRILACIÓN AURICULAR CRÓNICA	0,0719	TRASTORNO DE CONDUCCIÓN CARDÍACA	1,90
cluster_65(6)	LINFOMA MALIGNO (CLÍNICO)	5.228	LINFOMA MALIGNO (CLÍNICO)	0,0718	LINFOMA DE CÉLULAS DEL MANTO	17,11
	MIELOMA MÚLTIPLE	3.220	MIELOMA MÚLTIPLE	0,0716	LINFOMA DE BURKITT	12,88
	LEUCEMIA MIELOIDE AGUDA	1.454	ENFERMEDAD DE HODGKIN	0,0714	ENFERMEDAD DE HODGKIN	12,79
	ENFERMEDAD DE HODGKIN	1.294	LEUCEMIA MIELOIDE AGUDA	0,0714	MIELOMA MÚLTIPLE	11,08
	LINFOMA DE CÉLULAS DEL MANTO	402	LINFOMA DE CÉLULAS DEL MANTO	0,0712	LEUCEMIA MIELOIDE AGUDA	8,53
	LINFOMA DE BURKITT	366	LINFOMA DE BURKITT	0,0711	LINFOMA MALIGNO (CLÍNICO)	2,50
cluster_66(8)	SÍNDROME DE INSUFICIENCIA RENAL CRÓNICA	15.020	SÍNDROME DE INSUFICIENCIA RENAL CRÓNICA	0,0726	SÍNDROME DE INSUFICIENCIA RENAL CRÓNICA	18,50
	SÍNDROME DE INSUFICIENCIA RENAL	8.286	SÍNDROME DE INSUFICIENCIA RENAL	0,0722	SÍNDROME NEFRÓTICO	6,85
	SÍNDROME NEFRÓTICO	2.402	SÍNDROME NEFRÓTICO	0,0716	SÍNDROME DE INSUFICIENCIA RENAL	5,95
	INSUFICIENCIA RENAL EN ESTADIO TERMINAL	1.522	INSUFICIENCIA RENAL EN ESTADIO TERMINAL	0,0715	GLOMERULOSCLEROSIS SEGMENTARIA FOCAL	3,83
	INSUFICIENCIA RENAL CRÓNICA PROGRESIVA	864	INSUFICIENCIA RENAL CRÓNICA PROGRESIVA	0,0714	SÍNDROME NEFRÓTICO CONGÉNITO	3,54
	GLOMERULOSCLEROSIS SEGMENTARIA FOCAL	434	GLOMERULOSCLEROSIS SEGMENTARIA FOCAL	0,0712	INSUFICIENCIA RENAL CRÓNICA PROGRESIVA	3,43
	PROTEINURIA NEFRÓTICA	392	PROTEINURIA NEFRÓTICA	0,0712	PROTEINURIA NEFRÓTICA	2,59
	SÍNDROME NEFRÓTICO CONGÉNITO	70	SÍNDROME NEFRÓTICO CONGÉNITO	0,0709	INSUFICIENCIA RENAL EN ESTADIO TERMINAL	2,33

Table A.2 Continúa de página anterior

Cluster(tamaño)	Grado		Cercanía		Intermediación	
	Concepto	Valor	Concepto	Valor	Concepto	Valor
cluster_67(2)	CONTRACCIONES UTERINAS PRESENTES	7.618	SANGRADO VAGINAL	0,0721	SANGRADO VAGINAL	171,41
	SANGRADO VAGINAL	7.576	CONTRACCIONES UTERINAS PRESENTES	0,0720	CONTRACCIONES UTERINAS PRESENTES	25,04

Tab. A.3: Grupos de dos problemas y evidencia científica

Red	ConceptId1	Problema1	Conceptid2	Problema2	Observación
	30989003	GONALGIA	197247001	FÍSTULA ENTEROCUTÁNEA	Sin evidencia en documentos científicos
	40917007	OBNUBILACIÓN MENTAL	48867003	BRADICARDIA	Efectos secundarios en uso de medicina para el delirio DOI:10.4088/PCC.09r00938yel
	233961000	ISQUEMIA DE EXTREMIDAD INFERIOR	363346000	NEOPLASIA MALIGNA	Sin evidencia en documentos científicos
	11483009	DEFECTO VENTILATORIO	301199001	PROBLEMA NASAL	Sin evidencia en documentos científicos
	45460008	TRANSFUSIÓN INTRAUTERINA	57325008	ISOINMUNIZACIÓN	DOI: 10.1016/j.transci.2017.10.007
	118944007	TRASTORNO DE HOMBRO	301810000	INFECCIÓN CATEGORIZADA POR LOCALIZACIÓN	Sin evidencia en documentos científicos
RP	40108008	TALASEMIA	75451007	TALASEMIA MAYOR	Especificación
	88032003	AMAUROSIS FUGAZ	400130008	ARTERITIS TEMPORAL	doi: 10.1016/j.pop.2010.07.005.
	81516001	NEFRECTOMÍA PARCIAL	126880001	NEOPLASIA DEL RIÑÓN	doi: 10.1016/j.juro.2015.09.099
	13212004	QUEMADURAS DE SEGUNDO GRADO DE SITIOS MÚLTIPLES	284196006	QUEMADURA DE PIEL	Sin evidencia en documentos científicos
	11466000	OPERACIÓN CESÁREA	287664005	LIGADURA TUBARIA BILATERAL	https://www.acog.org/Clinical-Guidance-and-Publications/Committee-Opinions/Committee-on-Coding-and-Nomenclature/Tubal-Ligation-with-Cesarean-Delivery
	13802001	ABSCESO DE LA AXILA	15296000	ESTERILIDAD	Sin evidencia en documentos científicos
	10191004	EFECTO ANTICOAGULANTE	309298003	HALLAZGO RELACIONADO CON TRATAMIENTO FARMACOLÓGICO	Sin evidencia en documentos científicos
	28626004	FÍSTULA VESICOCOLÓNICA	110352000	DETERIORO COGNITIVO MÍNIMO	Sin evidencia en documentos científicos
	126858004	NEOPLASIA DE LA AMPOLLA DE VATER	363353009	TUMOR MALIGNO DE LA VESÍCULA BILIAR	Sin evidencia en documentos científicos
	24079001	DERMATITIS ATÓPICA	106190000	ALERGIA (TRASTORNO)	doi: 10.1158/1055-9965.EPI-14-1243
	190905008	FIBROSIS QUÍSTICA	236071009	DIARREA CRÓNICA	https://www.ncbi.nlm.nih.gov/pubmed/24712288
	34095006	DESHIDRATACIÓN	236021006	HERNIA INGUINAL DERECHA	https://www.ncbi.nlm.nih.gov/pubmed/1124411
	37796009	MIGRAÑA	75934005	ENFERMEDAD METABÓLICA	doi: 10.3389/fneur.2012.00161
	60046008	DERRAME PLEURAL	313299006	ENFERMEDAD PULMONAR OBSTRUCTIVA CRÓNICA, GRAVE	doi: 10.5935/1678-9741.20140047.
	1372004	ÚTERO UNICORNE	95435007	FÍSTULA TRAQUEOESOFÁGICA	http://revcmhabana.sld.cu/index.php/rcmh/article/view/713/1164
	71620000	FRACTURA DE FÉMUR	78314001	OSTEOGENIA IMPERFECTA	https://www.ncbi.nlm.nih.gov/pubmed/1294087
	276975007	CARCINOMA DE LARINGE	363429002	TUMOR MALIGNO DE LA LARINGE	Especificación
	89444000	ESPECTRO DE HIPOGÉNESIS DE MIEMBRO - OROMANDIBULAR	276657008	SÍNDROME DE SUPERPOSICIÓN	Sin evidencia en documentos científicos
	47270006	HIPERCORTISOLISMO	253017000	TUMOR DE KLATSKIN	Sin evidencia en documentos científicos
	87282003	PARASITISMO INTESTINAL	307132003	LITIASIS COLEDOCIANA	http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S0034-74932000000100009
	190905008	FIBROSIS QUÍSTICA	236071009	DIARREA CRÓNICA	https://www.ncbi.nlm.nih.gov/pubmed/24712288
	12441001	EPISTAXIS	232354002	EPISTAXIS ANTERIOR	Especificación
	84229001	FATIGA	399153001	VÉRTIGO	https://symptomchecker.webmd.com/multiple-symptoms?symptoms=dizziness%7Cfatigue&symptomids=81%7C98&locations=66%7C66
	16932000	NÁUSEAS Y VÓMITOS	300359004	HALLAZGO RELACIONADO CON EL VÓMITO	Especificación
	62315008	DIARREA	422400008	VÓMITOS	https://symptomchecker.webmd.com/multiple-symptoms?symptoms=diarrhea|nausea-or-vomiting&symptomids=72|156&locations=24|22
RSP	4009004	ENFERMEDAD INFECCIOSA DE LAS VÍAS URINARIAS INFERIORES	236002003	HEMATOMA RETROPERITONEAL	https://encolombia.com/medicina/revistas-medicas/cirugia/vc-082/trauma-retroperitoneal/
	68235000	CONGESTIÓN NASAL	82272006	RESFRÍO COMÚN	https://www.healthline.com/health/common-cold-symptoms
	65147003	EMBARAZO GEMELAR	86569001	PUERPERIO	trivial

Table A.3 Continúa de página anterior

Red	ConceptId1	Problema1	Conceptid2	Problema2	Observación
	70536003	TRASPLANTE DE RIÑÓN	176136000	CISTOSTOMÍA E INSERCIÓN DE CATÉTER SUPRAPÚBICO	Sin evidencia en documentos científicos
	128926000	HALLAZGO POSPROCEDIMIENTO	367391008	MALESTAR GENERAL	trivial
	190855004	HIPOMAGNESEMIA	409089005	NEUTROPENIA FEBRIL	Efectos secundarios en uso de medicina para el hypomagnesemia doi: 10.3892/ol.2013.1301
	40108008	TALASEMIA	75451007	TALASEMIA MAYOR	Especificación
	13172003	PÚRPURA TROMBOCITOPÉNICA IDIOPÁTICA CRÓNICA	32273002	PÚRPURA TROMBOCITOPÉNICA IDIOPÁTICA	Especificación
	42343007	INSUFICIENCIA CARDÍACA CONGESTIVA	84114007	INSUFICIENCIA CARDÍACA	Especificación
	95214007	NEOPLASIA MALIGNA PRIMARIA DE HÍGADO	126851005	NEOPLASIA DEL HÍGADO	Especificación
	359789008	ARTERITIS DE TAKAYASU	418304008	INSUFICIENCIA CARDÍACA DIASTÓLICA	Sin evidencia en documentos científicos
	126858004	NEOPLASIA DE LA AMPOLLA DE VATER	363353009	TUMOR MALIGNO DE LA VESÍCULA BILIAR	Sin evidencia en documentos científicos
	198288003	ESTADO DE ANSIEDAD	225624000	ATAQUE DE PÁNICO	https://www.ncbi.nlm.nih.gov/pubmed/7962579
	11466000	OPERACIÓN CESÁREA	287664005	LIGADURA TUBARIA BILATERAL	https://www.acog.org/Clinical-Guidance-and-Publications/Committee-Opinions/Committee-on-Coding-and-Nomenclature/Tubal-Ligation-with-Cesarean-Delivery
	70650003	CÁLCULO VESICAL	367403001	ESTENOSIS PILÓRICA	Sin evidencia en documentos científicos
	79827002	ANASTOMOSIS ARTERIOVENOSA QUIRÚRGICA PARA DIÁLISIS RENAL	108022006	RESECCIÓN DE RIÑÓN	Sin evidencia en documentos científicos
	81060008	OBSTRUCCIÓN INTESTINAL	93934004	NEOPLASIA MALIGNA PRIMARIA DE OVARIO	Sin evidencia en documentos científicos
	371039008	TRASTORNO TROMBOEMBÓLICO	400047006	ENFERMEDAD VASCULAR PERIFÉRICA	Sin evidencia en documentos científicos
	20433007	FRACTURA DE EXTREMO SUPERIOR DE TIBIA	263232003	FRACTURA DE FÉMUR DISTAL	trivial
	24057006	AFASIA COMBINADA	87486003	AFASIA	Especificación
	254628007	CARCINOMA DE PARÉNQUIMA PULMONAR	430969000	NEUMONÍA ASPIRATIVA RECURRENTE	Sin evidencia en documentos científicos
	123845008	ADENOCARCINOMA DE ENDOMETRIO	254878006	CARCINOMA ENDOMETRIAL	Especificación
	44054006	DIABETES MELLITUS TIPO 2	302866003	HIPOGLUCEMIA	https://www.ncbi.nlm.nih.gov/pubmed/29496365
	205237003	NEUMONITIS	254626006	ADENOCARCINOMA DEL PULMÓN	Sin evidencia en documentos científicos
	236648008	RETENCIÓN AGUDA DE ORINA	267064002	RETENCIÓN DE ORINA	Especificación
	126713003	NEOPLASIA DE PULMÓN	309529002	MASA EN PULMÓN	Especificación
	33688009	COLESTASIS	59927004	INSUFICIENCIA HEPÁTICA	doi: 10.3978/j.issn.2305-5839.2015.AB083
	119249001	AGAMMAGLOBULINEMIA	119250001	HIPOGAMMAGLOBULINEMIA	https://www.ncbi.nlm.nih.gov/pubmed/3728553
	20720000	PLASMAFÉRESIS	281789004	TRATAMIENTO CON ANTIBIÓTICOS	https://www.ncbi.nlm.nih.gov/pubmed/26820918
	92814006	LEUCEMIA LINFOIDE CRÓNICA	109978004	LINFOMA DE LINFOCITOS T (CLÍNICO)	https://www.ncbi.nlm.nih.gov/pubmed/28698787
	3006004	ALTERACIÓN DE LA CONSCIENCIA	130987000	CONFUSIÓN AGUDA	https://www.ncbi.nlm.nih.gov/pubmed/22795469
	268471004	SANGRADO VAGINAL	289700000	CONTRACCIONES UTERINAS PRESENTES	Sin evidencia en documentos científicos

B. AGRUPAMIENTOS POR CONTEXTOS

Fig. B.1: Agrupamientos del grafo de la lista de problemas

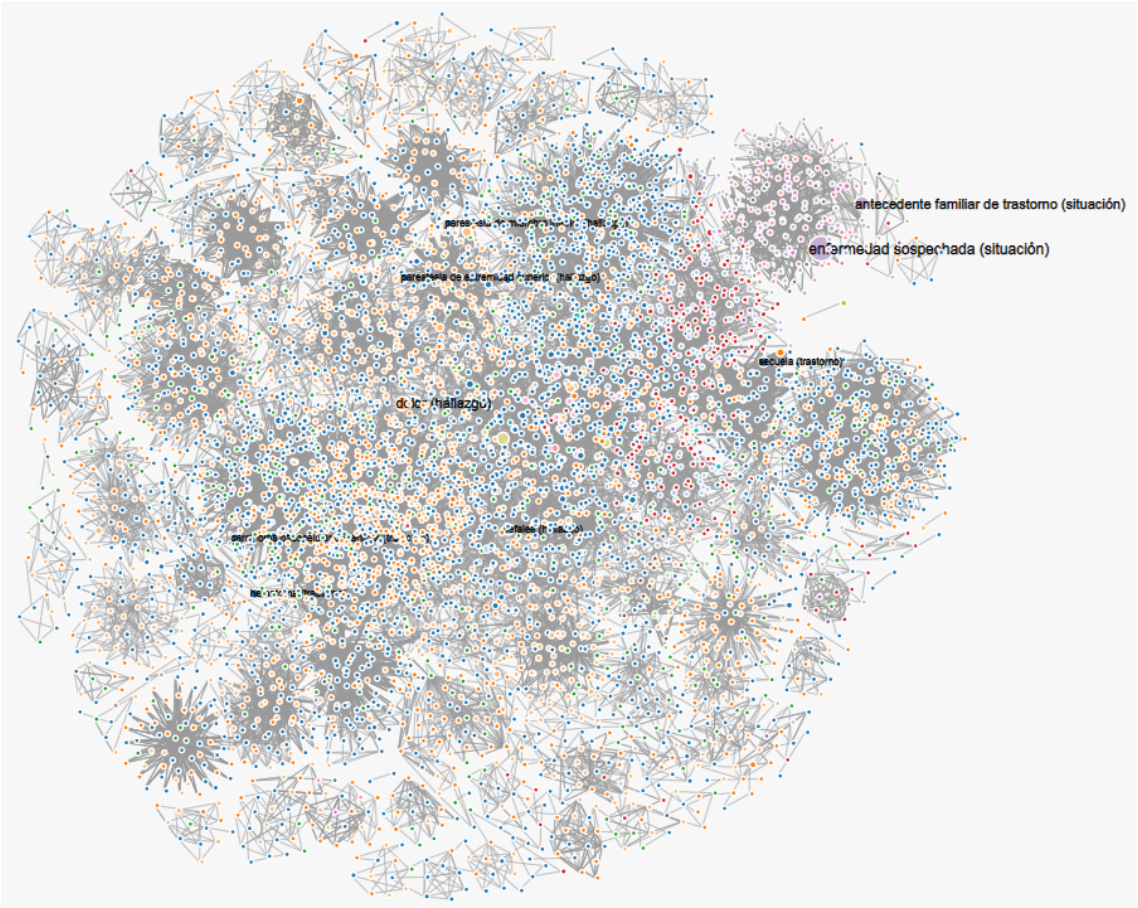


Fig. B.2: Agrupamientos del grafo de la lista de problemas en el contexto de los servicios de cardiología de adultos y pediátrica

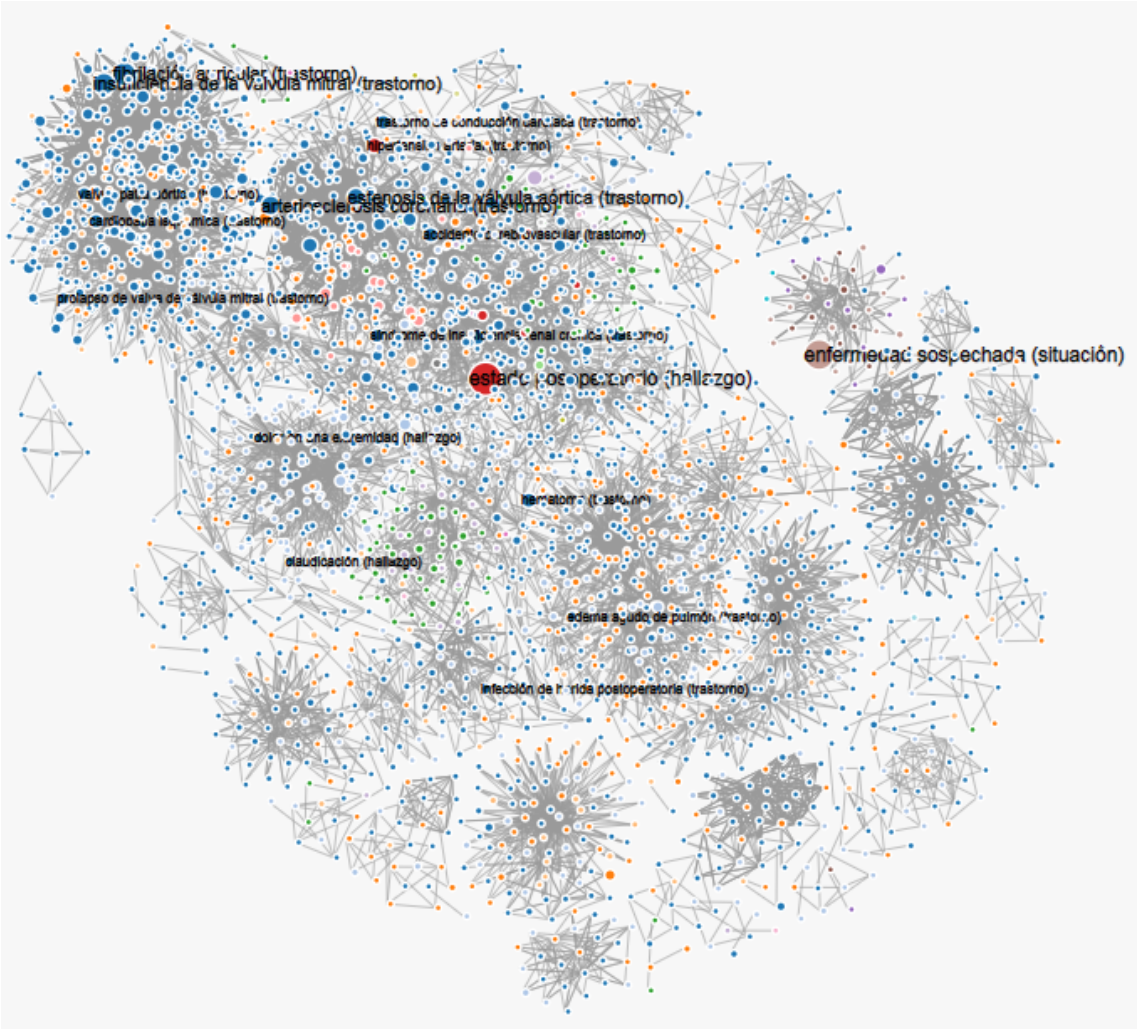


Fig. B.3: Agrupamientos del grafo de la lista de problemas en el contexto del servicio de dermatología

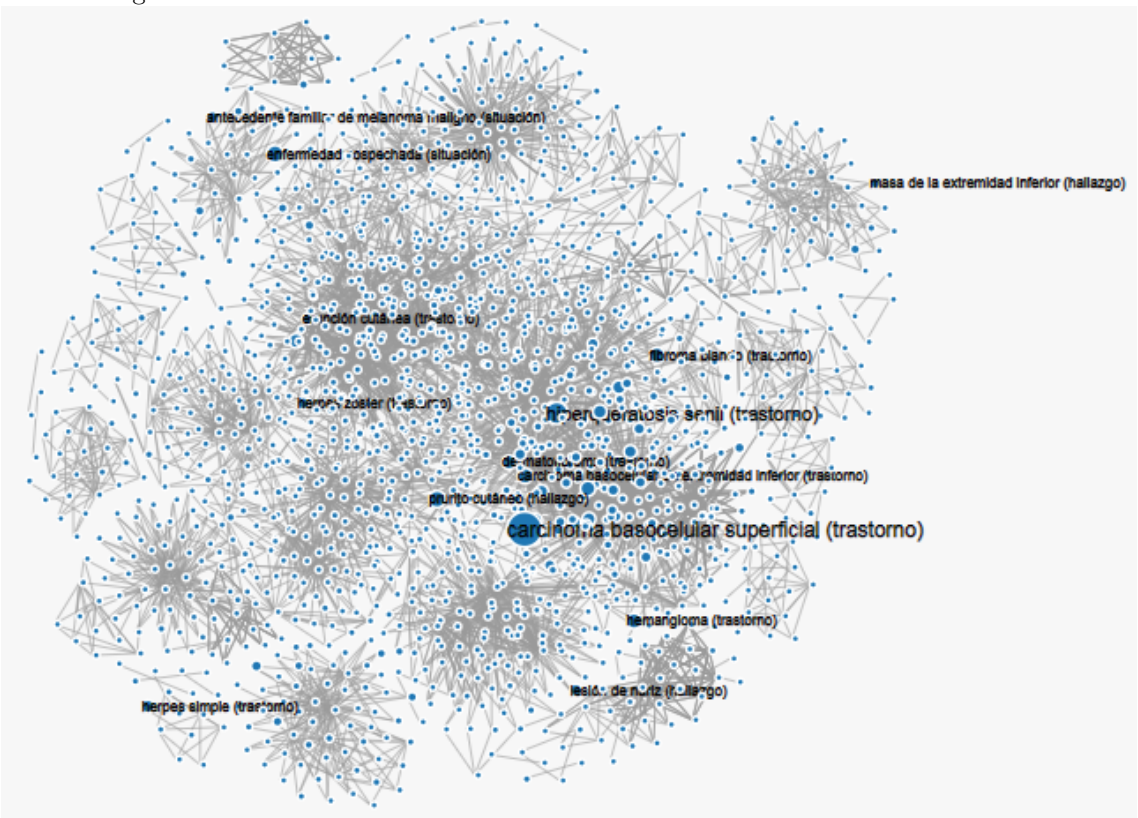


Fig. B.4: Agrupamientos del grafo de la lista de problemas en el contexto de los servicios de endocrinología, nefrología y urología

Fig. B.5: Agrupamientos del grafo de la lista de problemas en el contexto del servicio de neurología

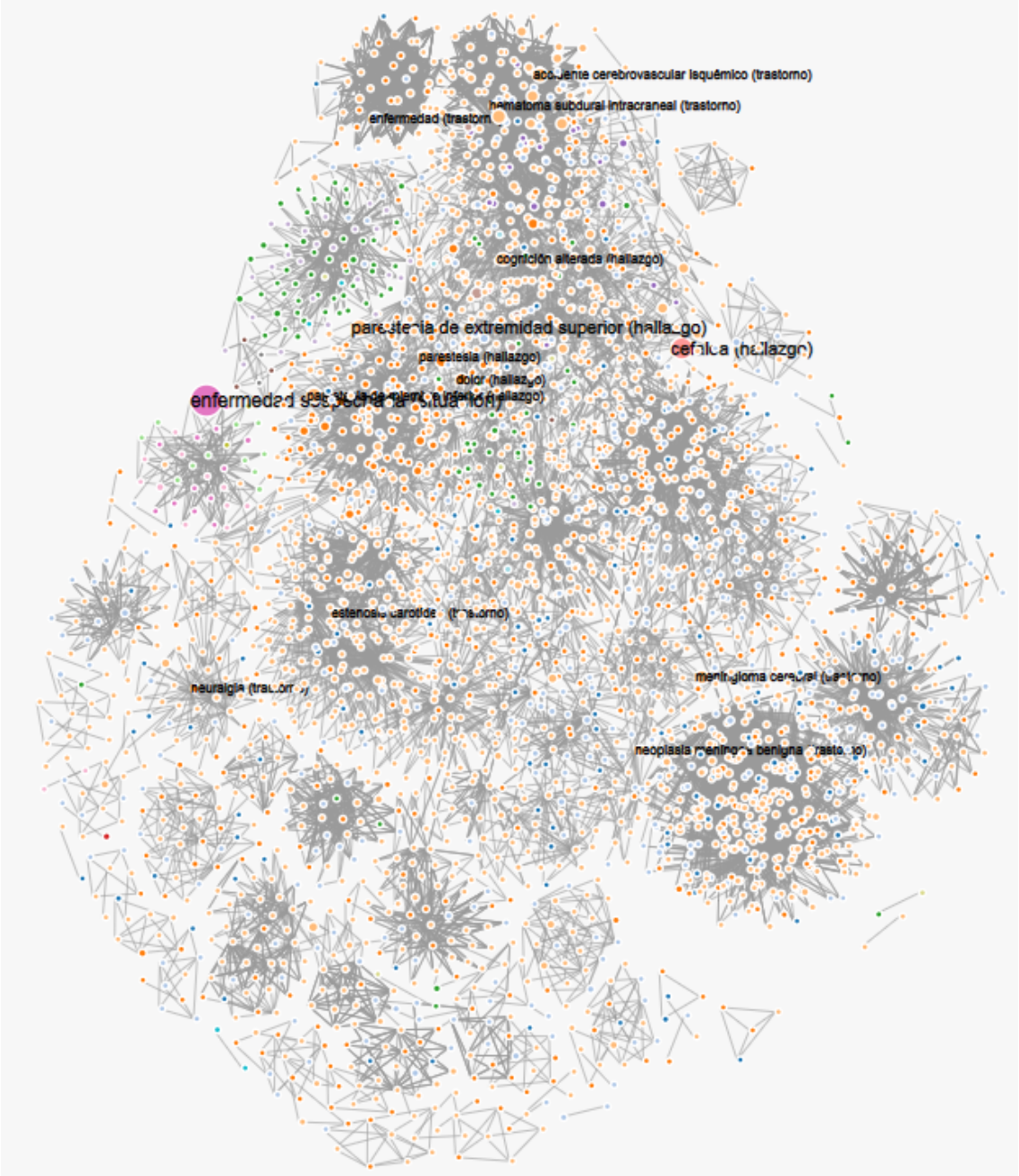


Fig. B.6: Agrupamientos del grafo de la lista de problemas en el contexto del servicio de pediatría

Fig. B.7: Agrupamientos del grafo de la lista de problemas en el contexto del nivel asistencial ambulatorio

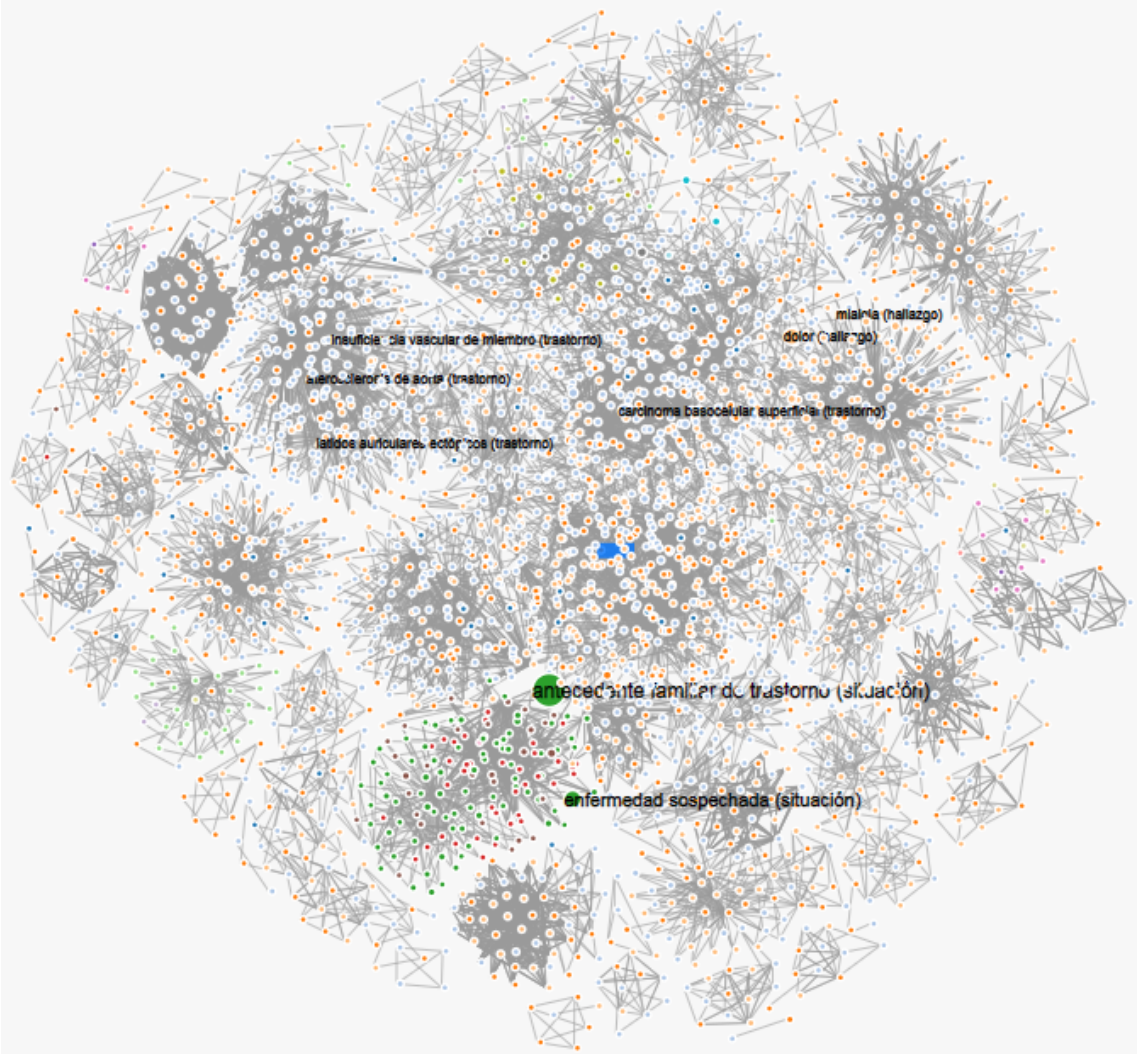


Fig. B.8: Agrupamientos del grafo de la lista de problemas en el contexto del nivel asistencial de episodio ambulatorio

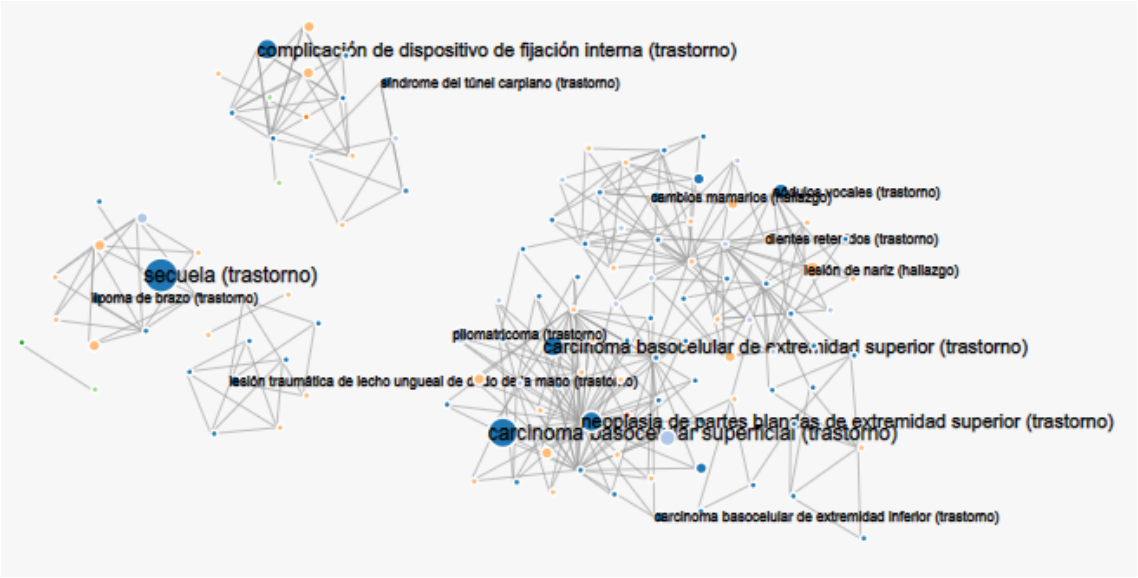


Fig. B.9: Agrupamientos del grafo de la lista de problemas en el contexto del nivel asistencial de internación domiciliaria

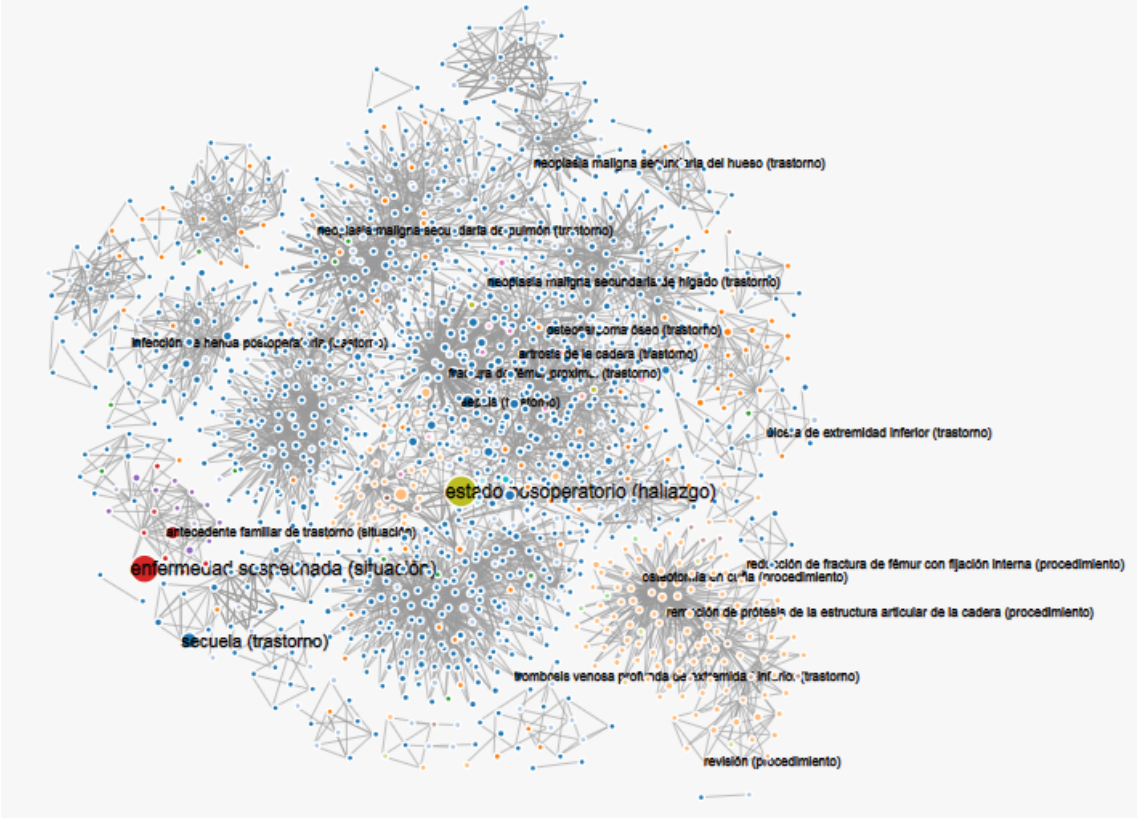


Fig. B.10: Agrupamientos del grafo de la lista de problemas en el contexto del nivel asistencial de internación general

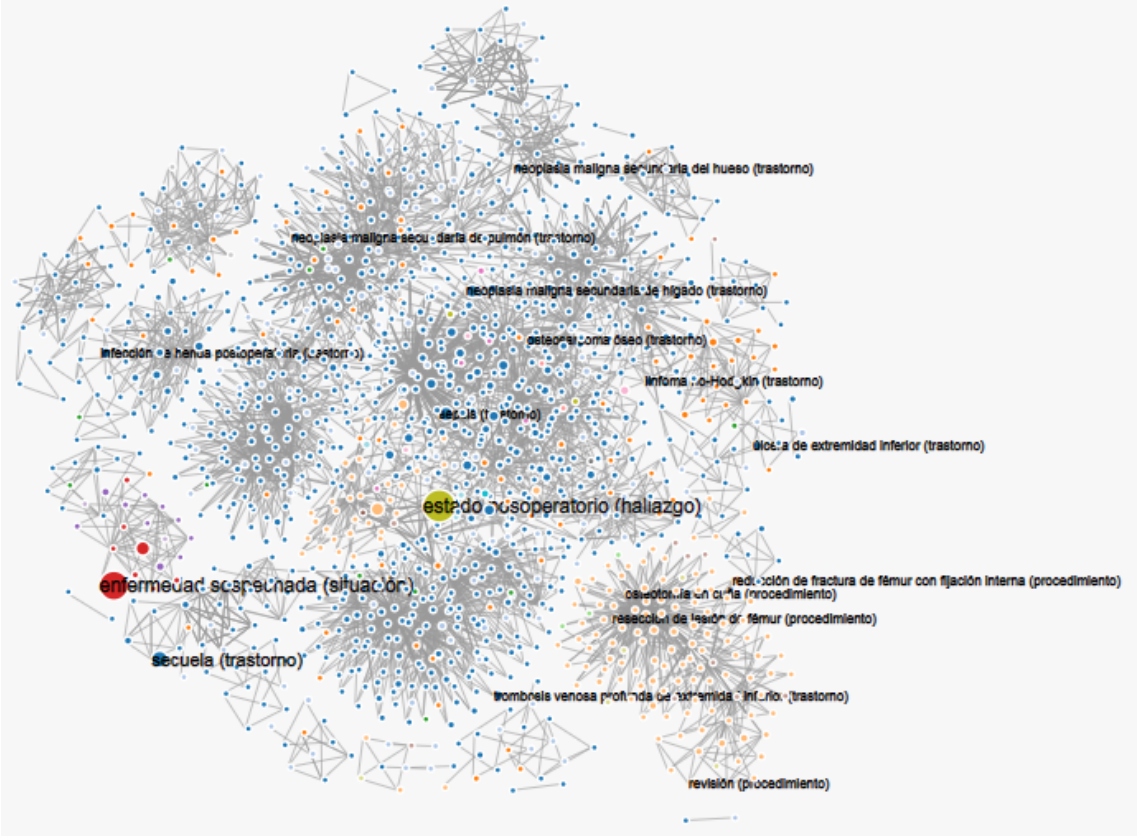


Fig. B.11: Agrupamientos del grafo de la lista de problemas en el contexto de grupo etario de 0 a 4, 15 a 24, 25 a 34 y 35 a 44 años

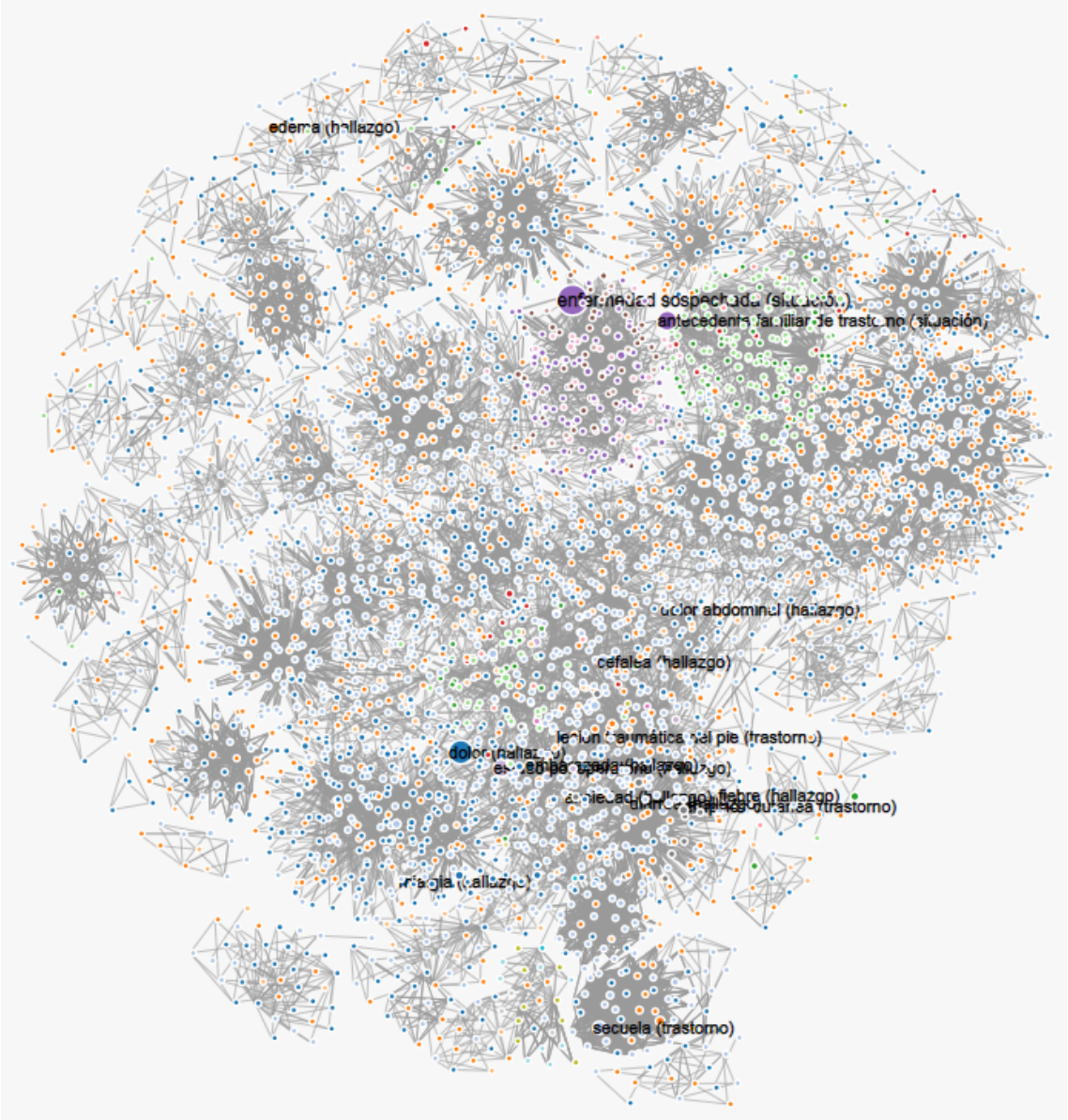
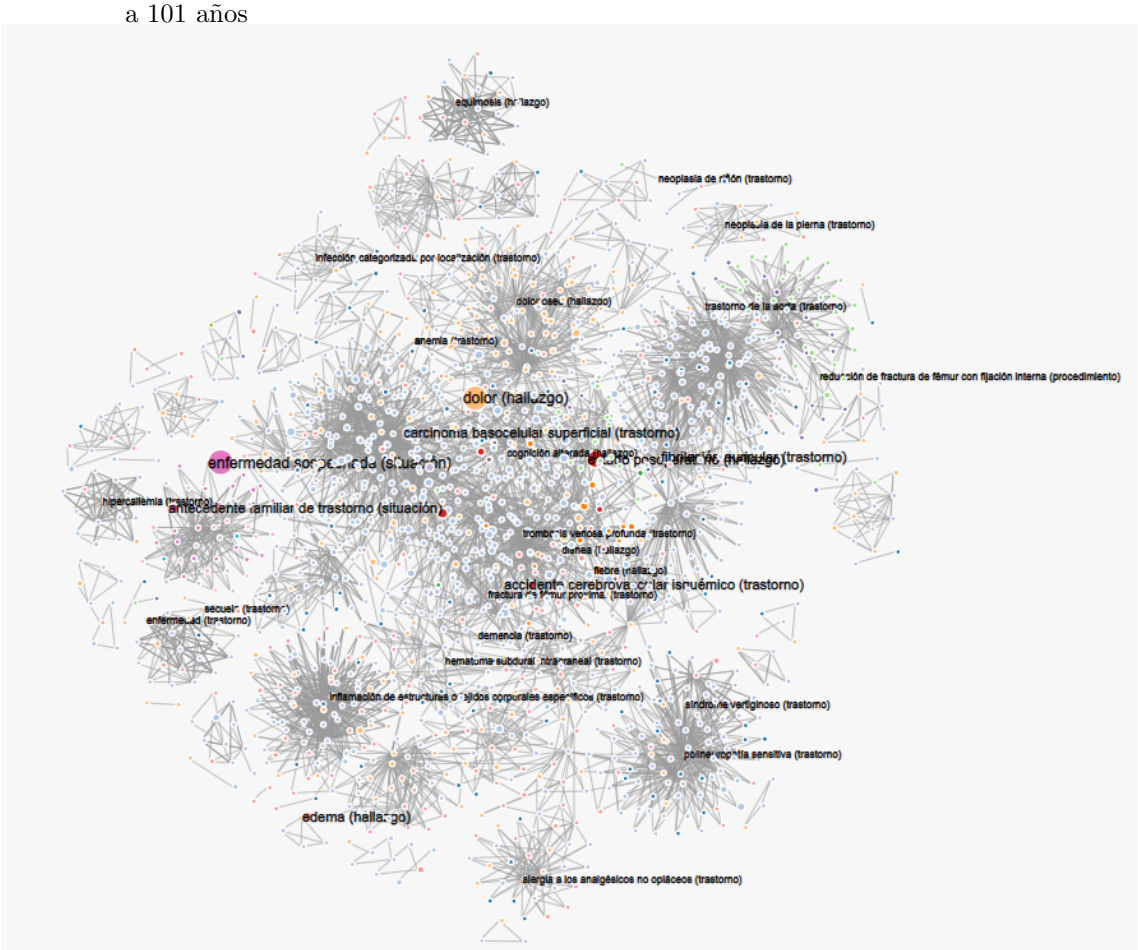


Fig. B.12: Agrupamientos del grafo de la lista de problemas en el contexto de grupo etario de 75 a 101 años



REFERENCIAS

Ávila, P., Castaño, J., Berinsky, H., Gambarte, L., Park, H., Pérez, D., ... Luna, D. (2018). Selection of Semantic Relevant Healthcare Services Subsets. *Studies in health technology and informatics*, *247*, 915–919. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/29678094

Balakrishnan, R., y Ranganathan, K. (2000). *A Textbook of Graph Theory*. New York, NY: Springer New York. Descargado de http://link.springer.com/10.1007/978-1-4419-8505-7 doi: 10.1007/978-1-4419-8505-7

Barabási, A.-L., Gulbahce, N., y Loscalzo, J. (2011, 1). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, *12*(1), 56–68. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/21164525http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3140052http://www.nature.com/doifinder/10.1038/nrg2918 doi: 10.1038/nrg2918

Ben Aouicha, M., y Hadj Taieb, M. A. (2016, 2). Computing semantic similarity between biomedical concepts using new information content approach. *Journal of Biomedical Informatics*, *59*, 258–275. Descargado de https://www.sciencedirect.com/science/article/pii/S1532046415002877 doi: 10.1016/J.JBI.2015.12.007

Bhattacharyya, S. B. (2016). SNOMED CT Basics. En *Introduction to snomed ct* (pp. 25–60). Singapore: Springer Singapore. Descargado de http://link.springer.com/10.1007/978-981-287-895-3_4 doi: 10.1007/978-981-287-895-3{_}4

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., y Lefebvre, E. (2008, 10). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. Descargado de http://arxiv.org/abs/0803.0476 doi: 10.1088/1742-5468/2008/10/P10008

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., y Hwang, D. U. (2006, 2). Complex networks: Structure and dynamics. *Physics Reports*, *424*(4-5), 175–308. Descargado de http://linkinghub.elsevier.com/retrieve/pii/S037015730500462X doi: 10.1016/j.physrep.2005.10.009

Brath, R., y Jonker, D. (2015). *Graph Analysis and Visualization: Discovering Business Opportunity in Linked Data*. John Wiley & Sons. Descargado de https://books.google.com.ar/books/about/Graph_Analysis_and_Visualization.html?id=pkPxBQAAQBAJ&pgis=1

Clauset, A., Newman, M. E. J., y Moore, C. (2004, 8). Finding community structure in very large networks. Descargado de http://arxiv.org/abs/cond-mat/0408187http://dx.doi.org/10.1103/PhysRevE.70.066111 doi: 10.1103/PhysRevE.70.066111

Cohen, R., y Havlin, S. (2010). *Complex Networks: Structure, Robustness and Function*. Cambridge University Press. Descargado de https://books.google.com/books?id=1ECLiFrKulIC&pgis=1

Cook, D. J., y Holder, L. B. (2006). *Mining Graph Data*. John Wiley & Sons. Descargado de http://dl.acm.org/citation.cfm?id=1050985

Csardi, G., y Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. Descargado de http://igraph.org

Diestel, R. (2005). *Graph theory*. Springer. Descargado de https://books.google.com.ar/books/about/Graph_Theory.html?id=aR2TMYQr2CMC&redir_esc=y

Dolin, R. H., Mattison, J. E., Cohn, S., Campbell, K. E., Wiesenthal, A. M., Hochhalter, B., ... Zingo, C. (2004). Kaiser Permanente's Convergent Medical Terminology. *Studies in health technology and informatics*, *107*(Pt 1), 346–50. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/15360832

Estrada, E., y Knight, P. (2015). *A first course in network theory*. Descargado de https://books.google.com.ar/books?id=4iB-BwAAQBAJ&pg=PA225&dq=network+is&hl=es&sa=X&ved=0ahUKEwiY39bNx6TfAhVEgJAKHQFgAtcQ6AEIMDAB#v=onepage&q=networkis&f=false

Fung, K. W., y Xu, J. (2015, 2). An exploration of the properties of the CORE problem list subset and how it facilitates the implementation of SNOMED CT. *Journal of the American Medical Informatics Association*, *22*(3), 649–658. Descargado de https://academic.oup.com/jamia/article-lookup/doi/10.1093/jamia/ocu022 doi: 10.1093/jamia/ocu022

Gan, M., Dou, X., y Jiang, R. (2013, 1). From ontology to semantic similarity: calculation of ontology-based semantic similarity. *TheScientificWorldJournal*, *2013*, 793091. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3603583&tool=pmcentrez&rendertype=abstract doi: 10.1155/2013/793091

Girvan, M., Newman, M. E. J., Cecconi, F., Loreto, V., y Parisi, D. (2002, 6). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(12), 7821–6. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/12060727http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC122977 doi: 10.1073/pnas.122653799

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., y Barabási, A.-L. (2007, 5). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(21), 8685–90. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/17502601http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1885563 doi: 10.1073/pnas.0701361104

Hagberg, A. A., Schult, D. A., y Swart, P. J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. En G. Varoquaux, T. Vaught, y J. Millman (Eds.), *Proceedings of the 7th python in science conference* (pp. 11–15). Pasadena, CA USA.

Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., y Montmain, J. (2014, 4). A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of Biomedical Informatics*, *48*, 38–53. Descargado de https://www.sciencedirect.com/science/article/pii/S1532046413001834?via%3Dihub doi: 10.1016/J.JBI.2013.11.006

Hersh, W., Buckley, C., Leone, T. J., y Hickam, D. (1994). OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. En *Sigir '94* (pp. 192–201). London: Springer London. Descargado de http://link.springer.com/10.1007/978-1-4471-2099-5_20 doi: 10.1007/978-1-4471-2099-5{_}20

Højen, A. R., Sundvall, E., y Gøeg, K. R. (2014). Methods and applications for visualization of SNOMED CT concept sets. *Applied clinical informatics*, *5*(1), 127–52. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/24734129http://

www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3974253 doi: 10.4338/ACI-2013-09-RA-0071

IHTSDO, I. H. T. S. D. O. (2014). *Snomed ct – supporting meaningful use.* Descargado de https://www.snomed.org/resource/resource/15 ([Online; accessed 01-September-2018])

IHTSDO, I. H. T. S. D. O. (2016a). *Editorial guide.*

IHTSDO, I. H. T. S. D. O. (2016b). *Snomed ct starter guide.*

Kaiser, M. (2008, 2). Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks. *New Journal of Physics*, *10*(8), 083042. Descargado de http://arxiv.org/abs/0802.2512http://dx.doi.org/10.1088/1367-2630/10/8/083042 doi: 10.1088/1367-2630/10/8/083042

Kuropatwa, O., y Giannangelo, K. (2016). *Snomed ct physician specialty subsets: From development to refinement with an eye on implementation and maintenance.* Descargado de https://confluence.ihtsdotools.org/pages/viewpage.action?pageId=29950370 ([Online; accessed 01-September-2018])

Lee, D., Cornet, R., Lau, F., y de Keizer, N. (2013, 2). A survey of SNOMED CT implementations. *Journal of biomedical informatics*, *46*(1), 87–96. Descargado de http://linkinghub.elsevier.com/retrieve/pii/S1532046412001530http://www.ncbi.nlm.nih.gov/pubmed/23041717 doi: 10.1016/j.jbi.2012.09.006

Lee, D., de Keizer, N., Lau, F., y Cornet, R. (2014, 2). Literature review of SNOMED CT use. *Journal of the American Medical Informatics Association*, *21*(e1), e11-e19. Descargado de https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2013-001636 doi: 10.1136/amiajnl-2013-001636

Lezcano, L., y Sicilia, M.-A. (2011). Connectivity and Semantic Patterns in Automatically Generated SNOMED-CT Subsets. En *Communications in computer and information science* (Vol. 240 CCIS, pp. 117–125). Descargado de http://link.springer.com/10.1007/978-3-642-24731-6_11 doi: 10.1007/978-3-642-24731-6{_}11

López Osornio, A., Gambarte, M. L., Otero, C., Gomez, A., Martinez, M., Soriano, E., ... others (2005). Desarrollo de un servidor de terminología clínico. En *8mo simposio de informática en salud-34 jaiio* (pp. 29–43).

López Osornio, A., Luna, D., y de Quiros, F. G. B. (2002). Creación de un sistema para la codificación automática de una lista de problemas. En *5to simposio de informática en salud-31 jaiio.*

López Osornio, A., Montenegro, S., García Martí, S., Toselli, L., Otero, C., Tavasci, I., ... de Quiros, F. (2004). Codificación múltiple de una lista de problemas utilizando la CIAP-2, CIE-10 y SNOMED CT. En *3er virtual congress of medical informatics-informedica.*

Luna, D., Franco, M., Plaza, C., Otero, C., Wassermann, S., Gambarte, M. L., ... González Bernaldo de Quirós, F. (2013, 1). Accuracy of an electronic problem list from primary care providers and specialists. *Studies in health technology and informatics*, *192*, 417–21. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/23920588

Luna, D., Otero, P., Gomez, A., Martinez, M., García Martí, S., Schpilberg, M., ... Quiros, F. (2003). Implementación de una Historia Clínica Electrónica Ambulatoria:"Proyecto ITALICA. En *6to simposio de informática en salud-32 jaiio* (Vol. 32).

Mabotuwana, T., Lee, M. C., y Cohen-Solal, E. V. (2013, 10). An ontology-based similarity measure for biomedical data – Application to radiology reports. *Journal of Biomedi-*

cal Informatics, *46*(5), 857–868. Descargado de https://www.sciencedirect.com/science/article/pii/S1532046413000889?via%3Dihub doi: 10.1016/J.JBI.2013.06.013

Manning, C. D., Raghavan, P., y Schuütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. Descargado de https://books.google.com.ar/books/about/Introduction_to_Information_Retrieval.html?id=GNvtngEACAAJ&source=kp_cover&redir_esc=y

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., y McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. En *Association for computational linguistics (acl) system demonstrations* (pp. 55–60). Descargado de http://www.aclweb.org/anthology/P/P14/P14-5010

Marbn, s., Mariscal, G., y Segovi, J. (2009). A Data Mining & Knowledge Discovery Process Model. En *Data mining and knowledge discovery in real life applications*. doi: 10.5772/6438

Newman, M. E. J. (2004, 6). Fast algorithm for detecting community structure in networks. *Physical Review E*, *69*(6), 066133. Descargado de https://link.aps.org/doi/10.1103/PhysRevE.69.066133 doi: 10.1103/PhysRevE.69.066133

Newman, M. E. J. (2006, 9). Finding community structure in networks using the eigenvectors of matrices. *Physical review. E, Statistical, nonlinear, and soft matter physics*, *74*(3 Pt 2), 036104. Descargado de http://arxiv.org/abs/physics/0605087 doi: 10.1103/PhysRevE.74.036104

Otero, C. (2014). *Improving the granularity of the electronic health record problem list* (Tesis Doctoral). Descargado de http://digitalcommons.ohsu.edu/etd/3532

Pedersen, T., Pakhomov, S. V., Patwardhan, S., y Chute, C. G. (2007, 6). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, *40*(3), 288–299. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/16875881http://linkinghub.elsevier.com/retrieve/pii/S1532046406000645 doi: 10.1016/j.jbi.2006.06.004

Pons, P., y Latapy, M. (2005, 10). Computing Communities in Large Networks Using Random Walks. En (pp. 284–293). Springer, Berlin, Heidelberg. Descargado de http://link.springer.com/10.1007/11569596_31 doi: 10.1007/11569596{_}31

Powers. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1).

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., y Parisi, D. (2004, 3). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, *101*(9), 2658–2663. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/14981240http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC365677http://www.pnas.org/cgi/doi/10.1073/pnas.0400054101 doi: 10.1073/pnas.0400054101

Raghavan, U. N., Albert, R., y Kumara, S. (2007, 9). Near linear time algorithm to detect community structures in large-scale networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, *76*(3 Pt 2), 036106. Descargado de http://arxiv.org/abs/0709.2938 doi: 10.1103/PhysRevE.76.036106

Rector, A. L., Brandt, S., y Schneider, T. (2011, 1). Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. *Journal of the American Medical Informatics Association : JAMIA*, *18*(4), 432–

40. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3128394&tool=pmcentrez&rendertype=abstract doi: 10.1136/amiajnl-2010-000045

Sánchez, D., y Batet, M. (2011, 10). Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of Biomedical Informatics*, *44*(5), 749–759. Descargado de https://www.sciencedirect.com/science/article/pii/S1532046411000645 doi: 10.1016/J.JBI.2011.03.013

Saramäki, J., Kivelä, M., Onnela, J. P., Kaski, K., y Kertész, J. (2007, 8). Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *75*(2), 027105. Descargado de http://arxiv.org/abs/cond-mat/0608670http://dx.doi.org/10.1103/PhysRevE.75.027105 doi: 10.1103/PhysRevE.75.027105

Scott, J., Wasserman, S., Faust, K., y Galaskiewicz, J. (1996). *Social Network Analysis: Methods and Applications* (Vol. 47) (n.º 2). Cambridge University Press. Descargado de http://www.jstor.org/stable/591741?origin=crossref doi: 10.2307/591741

Solid IT. (2019). *DB-Engines Ranking of Graph DBMS.* Descargado de https://db-engines.com/en/ranking/graph+dbms

Tang, L., y Liu, H. (2010). Graph Mining Applications to Social Network Analysis. En (pp. 487–513). Springer US. Descargado de http://link.springer.com/10.1007/978-1-4419-6045-0_16http://www.leitang.net/papers/graph_mining.pdf doi: 10.1007/978-1-4419-6045-0{_}16

Van Haaren, J., Davis, J., Lappenschaar, M., y Hommersom, A. (2013). Exploring disease interactions using Markov networks. *AAAI Workshop - Technical Report*, *WS-13-09*, 65–70. Descargado de https://lirias.kuleuven.be/bitstream/123456789/400315/1/hiai13-paper.pdf

Versi, E. (1992, 7). Gold standard is an appropriate term. *BMJ (Clinical research ed.)*, *305*(6846), 187. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/1515860http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1883235

Wang, J., Day, R., Visweswaran, S., y Hogan, W. (2010, 1). The use of semantic distance metrics to support ontology audit. *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, *2010*, 842–846. Descargado de http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041307&tool=pmcentrez&rendertype=abstract

Watts, D. J., y Strogatz, S. H. (1998, 6). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. Descargado de http://www.nature.com/articles/30918 doi: 10.1038/30918

Weed, L. L. (1968, 3). Medical records that guide and teach. *The New England journal of medicine*, *278*(11), 593–600. Descargado de http://www.ncbi.nlm.nih.gov/pubmed/5637758 doi: 10.1056/NEJM196803142781105

Zare, M., Zare, M., Pahl, C., Nilashi, M., Salim, N., y Ibrahim, O. (2015, 9). A Review of Semantic Similarity Measures in Biomedical Domain Using SNOMED-CT. *Journal of Soft Computing and Decision Support Systems*, *2*(6), 1–13. Descargado de http://jscdss.com/index.php/files/article/view/61