VNIVERSIDAD NACIONAL
AVTONOMA DE
MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERIA DE LA COMPUTACIÓN

EXTRACCIÓN DE TAXONOMÍAS DE TEXTOS Y MÉTRICAS DE CALIDAD DE LAS MISMAS UTILIZANDO CUBOS OLAP, UDFS Y SPS

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

MAESTRA EN CIENCIAS

(COMPUTACIÓN)

PRESENTA:

ANGELA VICTORIA INCLÁN JASSO

TUTOR: DR. JAVIER GARCÍA GARCÍA

FACULTAD DE CIENCIAS

UNAM

MÉXICO, D. F. MAYO 2014.

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Desde la aparición de la Web semántica (proyecto creado por World Wide Web Consortium) se ha pretendido desarrollar tecnologías que permitan publicar datos legibles por aplicaciones informáticas, es decir, se ha pretendido mejorar Internet por medio de la interoperabilidad entre los sistemas informáticos. Para esto se han creado y utilizado las ontologías, las cuales forman parte de la base en las tecnologías de la Web semántica y son similares a un diccionario o glosario, pero con mayor detalle y estructura, éstas tecnologías permiten a los equipos procesar su contenido.

Algunos autores consideran a las taxonomías como un subcaso de las ontologías [46, 47], por lo que el trabajo de esta tesis esta basado en [1], pero reduciéndolo al caso de las taxonomías.

En este trabajo se agregaron nuevos métodos para la obtención de las relaciones de los pares de términos que conforman las taxonomías, así como también se proponen métricas para evaluar el contenido de las taxonomías obtenidas. Todo esto implementándolo mediante expresiones de procesamiento analítico en línea (OLAP, por sus siglas en inglés), funciones definidas por el usuario (UDFs, por sus siglas en inglés) y procedimientos almacenados (SPs, por sus siglas en inglés).

Se realizan experimentos para obtener taxonomías con una versión de software desarrollada con procedimientos almacenados y otra con funciones definidas por el usuario, para posteriormente evaluar su eficiencia de acuerdo al tiempo de ejecución.

También se realizan experimentos para la obtención de la relación entre los pares de términos más frecuentes con métodos de correlación y peso de acuerdo a [1] y adicionalmente con otros métodos como lo son distancia y soporte y confianza. Se presenta también una modificación al método utilizado para obtener reglas de asociación (soporte y confianza), llamado método de soporte y confianza por intervalos.

AGRADECIMIENTOS

La culminación de este trabajo se debe a la contribución de varias personas, que aunque influyeron de diferente manera, me alentaron a terminar esta gran etapa que me trajo grandes y buenos cambios, por esta razón, quiero agradecer:

Al Conacyt, por brindarme la beca con la cual me fue posible llevar a cabo y concluir los estudios de la maestría, sin este apoyo ni siquiera hubiera pensado en la posibilidad de continuar estudiando.

A mis primos, que siempre están conmigo, apoyándome, distrayéndome, aconsejándome y escuchándome cuando los necesito.

A mi mamá y mi abuelita que me han brindado los sustentos necesarios para continuar estudiando, espero que estén muy orgullosas de mi.

A Carlitos quien me apoyo y sufrió conmigo por pasar las materias de la maestría, gracias chinito, sin ti me hubiera costado más trabajo terminar.

A mis sinodales de tesis, quienes aportaron con propuestas de mejoras para este trabajo, les agradezco todos sus comentarios, especialmente a la profesora Sofia.

A mi director de tesis, el Dr. Javier, quien me guió y alentó para terminar con este trabajo.

A Amalia y Lulú, quienes siempre me ayudaron a resolver las dudas de los tramites, me presionaron para terminar y escucharon mis quejas.

A mis nuevos amigos, Isaias, quien me escucho y apoyo en los momentos más difíciles que me tocaron vivir en esta etapa, a Rommel, quien me hizo abrir los ojos en varios aspectos de mi vida, a Onofre, quien escucho mis tonterías y me aconsejo respecto a ellas, a Ale, mi nueva mejor amiga, con quien he compartido varias experiencias y espero que aún falten muchas por vivir.

Un sueño más cumplido.

Índice general

Capítulo 1

Introducción

Desde la aparición de la Web semántica (proyecto creado por World Wide Web Consortium) se ha pretendido desarrollar tecnologías que permitan publicar datos legibles por aplicaciones informáticas, es decir, se ha pretendido mejorar Internet por medio de la interoperabilidad entre los sistemas informáticos.

Hace ya algunos años se han realizado investigaciones para lograr la creación de descripciones semánticas de los datos. Para esto se han creado y utilizado las ontologías, las cuales forman parte de la base en las tecnologías de la Web semántica y son similares a un diccionario o glosario, pero con mayor detalle y estructura, éstas tecnologías permiten a los equipos procesar su contenido. Las ontologías se componen de un conjunto de conceptos, axiomas y relaciones que describen un dominio de interés[1, 20].

Para la construcción de ontologías, es necesario resolver el problema de la extracción de términos y conceptos principales de un texto. Este problema se ha intentado resolver con distintas metodologías de lenguaje natural, con descomposición del valor singular y con estadística entre otros métodos. Cabe mencionar que algunos de estos métodos conllevan un gran tiempo de procesamiento o requieren de conocimiento previo. En este trabajo se pretende encontrar métodos y herramientas que permitan facilitar estos procesos[1].

Las ontologías pueden contener diferentes tipos de relaciones entre los términos que las conforman, por ejemplo la relación de jerarquía, la relación de pertenencia, entre otras; entre más relaciones y axiomas tengan más ricas se consideras la ontologías.

Algunos autores consideran a las taxonomías como un subcaso de las ontologías[46, 47], por lo que el trabajo de esta tesis esta basado en [1], pero reduciéndolo al caso de las taxonomías, esto se explica en el capítulo 2.

1.1. Objetivo

El presente trabajo se basa en [1], agregando nuevos métodos para la obtención de las relaciones de los pares de términos, así como también se proponen

métricas para evaluar el contenido de las taxonomías obtenidas. Todo implementándolo con procedimientos almacenados y funciones definidas por el usuario.

El objetivo de este trabajo de tesis es lograr la obtención eficiente de los conceptos principales de textos que se encuentran almacenados en una base de datos mediante expresiones de procesamiento analítico en línea (OLAP, por sus siglas en inglés), funciones definidas por el usuario (UDFs, por sus siglas en inglés) y procedimientos almacenados (SPs, por sus siglas en inglés) y con estos conceptos construir taxonomías con las cuales se podrá acceder a la información de una manera rápida y analizar la calidad del contenido de las taxonomías obtenidas mediante distintas métricas.

Los objetivos particulares de esta tesis son:

- Realizar experimentos para obtener taxonomías con una versión de software desarrollada con procedimientos almacenados y otra con funciones definidas por el usuario, para posteriormente evaluar su eficiencia de acuerdo al tiempo de ejecución (ambas versiones desarrolladas con SQL).
- Realizar experimentos para la obtención de la relación entre los pares de términos más frecuentes con métodos de correlación y peso de acuerdo a [1] y adicionalmente con otros métodos como lo son distancia y soporte y confianza. Se presenta también una modificación al método utilizado para obtener reglas de asociación (soporte y confianza), llamado método de soporte y confianza por intervalos.
- Medir la calidad de las taxonomías obtenidas al utilizar la métrica de palabras clave y la métrica de palabras contenidas en el título (ver capítulo 4), así mismo comparar los resultados obtenidos con ambas métricas de calidad.

1.2. Trabajos relacionados

Existen ya algunas herramientas que permiten obtener, modificar o dar mantenimiento a ontologías, de las cuales a continuación se da una breve explicación de algunas de ellas, por ejemplo:

- OntoBuilder[6]: El proyecto OntoBuilder soporta la extracción de ontologías de interfaces de búsqueda Web. Las ontologías de dominios similares son consolidadas en una única ontología con un dominio que puede ser consultado. Dado un ejemplo de formulario, llenado por el usuario y dado un nuevo formulario de otro sitio de Internet, OntoBuilder encuentra el mejor mapeo entre los dos formularios.

  El uso de ontologías, al contrario del esquema relacional o XML como un modelo de datos subyacente permite una representación flexible de los metadatos. OntoBuilder fue desarrollado con Java, y genera un diccionario de términos extrayendo las etiquetas y nombres de los campos de los formularios Web. Además reconoce una única relación entre términos y utiliza un algoritmo de empate.

En contraste con las ontologías obtenidas con el OntoBuilder las cuales provienen de formularios de páginas Web y son construidas con Java, el presente trabajo obtiene taxonomías de textos almacenados en una base de datos donde se utiliza solamente herramientas de SQL.

- OntoLT[2]: Esta herramienta soporta la extracción interactiva y/o la extensión de ontologías de textos. OntoLT provee un ambiente de la integración de análisis lingüístico en la ontología a través de la definición de reglas de mapeo que relacionan entidades lingüísticas en colecciones de textos.

  OntoLT provee una solución intermedia en el desarrollo de ontologías que permite a la ingeniería de la ontología inicializar en dominio específico de un corpus relevante de textos. Una secuencia de pasos automáticos e interactivos están involucrados en este proceso:

  - Análisis y anotación lingüística automática.
  - Preprocesamiento estadístico automático de extracción de entidades lingüísticas.
  - Definición interactiva de reglas de mapeo entre la extracción de entidades lingüísticas y las clases de Protégé.
  - Validación interactiva del usuario de las clases generadas por Protégé-OntoLT.
  - Integración automática de clases validadas en una existente o una nueva ontología.

  A diferencia del OntoLT, la construcción de las taxonomías realizada en este trabajo de tesis no realiza anotaciones lingüísticas a las palabras, ni se realizan reglas de mapeo, los términos se obtienen con la frecuencia, el peso y el método utilizado en cada experimento, ya sea la correlación, el soporte y confianza o la distancia.

- Altova SemanticWorks[10] es un editor visual RDF y OWL que genera automáticamente RDF /XML o nTriples basada en el diseño visual de la ontología. No hay versión de fuente abierta disponible.

  En comparación con lo realizado con la herramienta de Altova SemanticWorks la cual esta enfocada a la modificación de la ontología ya obtenida en alguno de los formatos RDF ú OWL, nosotros nos enfocamos en la obtención de las taxonomías y la medición de la calidad de éstas.

- TopBraid Compositer[11] es un entorno de modelado de clase empresarial para el desarrollo de ontologías de Web Semántica y la creación de aplicaciones semánticas.

- Knoodl[9]: esta herramienta facilita la orientación comunitaria de desarrollo de ontologías OWL basadas en RDF y bases de conocimiento. También

sirve como una plataforma de tecnología semántica, que ofrece un servicio basado en Java interface o una interfaz basada en SPARQL para que las comunidades puedan construir sus propias aplicaciones semánticas al utilizar sus ontologías y bases de conocimiento.

- ONTODUM[38]: En el artículo se introduce la técnica de GRAMO que sirve para la construcción del modelo de usuario y de dominio que pueden ser reutilizados en el desarrollo de aplicaciones multi-agente. ONTODUM representa el conocimiento de las técnicas para la especificación de los requisitos de una familia de sistemas multi-agente en un cierto dominio. Los modelos de dominio que se construyen con ONTODUM se utilizan en la definición de una técnica basada en patrones y ontologías para la construcción de marcos multi-agente.

  Como puede notarse, ONTODUM es diferente en comparación con el presente trabajo de tesis, principalmente por el uso de sistemas multi-agente que se utilizan en ONTODUM.

La mayor parte de herramientas existentes que obtienen, modifican o visualizan ontologías, están enfocadas en manejar datos de páginas o sistemas Web, e incluso la mayoría están desarrolladas con Java. Existen muy pocas herramientas desarrolladas con SQL, como es la que se presenta en este trabajo, con distintas métricas para obtener los términos y sus relaciones, completamente implementadas en SQL, con la premisa de contar con los documentos almacenados en la base de datos.

Por último se compara con el trabajo descrito en el artículo de ONTOCUBE[1], donde se obtienen ontologías con un solo tipo de relación (A tiene un B) mediante cubos OLAP (definido en la sección 3.4.1) por medio de las relaciones de los términos de acuerdo al valor de correlación existente entre dos pares de términos. La diferencia con este último trabajo mencionado es que en esta tesis se obtienen taxonomías, con las métricas de correlación, soporte y confianza, orden, distancia y soporte y confianza por intervalos (ver capítulo 4) y posteriormente se mide la calidad de éstas.

1.3. Estructura

Los capítulos de este trabajo se encuentran organizados de la siguiente manera:

En el capítulo 2 se da la definición de ontologías, así como de las taxonomías y cuales son algunos de los lenguajes que se utilizan para representarlas en el ámbito de Web Semántica. Por otro lado en el capítulo 3 se introduce el tema de las herramientas de SQL que se utilizaron para la obtención y procesamiento de las taxonomías, como lo son los cubos OLAP, las funciones definidas por el usuario y los procedimientos almacenados.

Posteriormente en el capítulo 4 se da una explicación de los métodos que fueron utilizados en los experimentos, así como de las métricas que se utilizaron

para medir la calidad de las taxonomías y la manera en que se desarrollaron los experimentos.

A continuación en el capítulo 5 se presenta la sección de los experimentos de cada corpus, cada uno también con sus correspondientes experimentos al utilizar las dos métricas de calidad, también en este capítulo se presenta el rendimiento de cada versión (SP y UDF).

En seguida en el capítulo 7 se encuentran las conclusiones del trabajo realizado, así como las mejoras propuestas.

A petición de una de las revisoras de esta trabajo se agrego una herramienta para lematización de las palabras contenidas en los documentos, aunque esta no sigue con los propósitos de este trabajo que es utilizar el lenguaje de SQL, como lo son los SPs y las UDFs, ésta se trata en el capítulo 6.

Capítulo 2

Ontologías y taxonomías

En este capítulo se define que son las ontologías y las taxonomías, y se describen algunos detalles de éstas. También se presentan los formatos en los que deben estar escritas y con los cuales las taxonomías desarrolladas en este trabajo podrían transformarse a ontologías.

2.1. Ontologías

Una ontología es la elaboración de un esquema conceptual de un dominio, y tiene como objetivo facilitar la comunicación y el intercambio de información entre diferentes sistemas y entidades. Su nombre fue tomado originalmente de "analogía"[21].

De acuerdo a Standard Upper Ontology Working Group (SUO)[20]: Una ontología es similar a un diccionario o glosario, pero con mayor detalle y estructura que permite a los equipos procesar su contenido. Una ontología se compone de un conjunto de conceptos, axiomas y relaciones que describen un dominio de interés.

Las ontologías se empiezan a utilizar a finales de los 80 en el campo de la inteligencia artificial como medio para la compartición y reuso de conocimiento. En la segunda mitad de los 90 se empiezan a aplicar a la Web para el reuso, la compartición y la integración de los datos entre diferentes páginas Web.

Algunos formatos utilizados en la Web para el manejo de ontologías son:

- OWL (Web Ontology Language)[24]: es un lenguaje de marcado para publicar y compartir datos por medio de ontologías en Internet. Esta construido sobre RDF y basado en XML. Se utiliza en el proyecto de Web semántica.
- RDF (Resource Description Framework)[25]: es un modelo estándar utilizado para representar los metadatos en Internet.

Existen diferentes tipos de ontologías, entre los que se distinguen:

- Ontologías de un dominio específico: son especificas para un proceso o una tarea específica de algún dominio concreto del conocimiento. Por ejemplo una ontología para del proceso de ventas.
- Ontologías genéricas: son las más abstractas, se utilizan para describir conceptos, como pueden ser el tiempo, el espacio, entre otros.
- Ontologías de representación: se utilizan para representar un objeto y sus características, estos objetos pueden ser procesos, funciones, componentes, entre otros.

En una ontología cada concepto consta de 3 componentes básicos, los cuales son: los términos, sus atributos (representan las características y permiten representar más a detalle el concepto) y las relaciones que existe entre éstos (representan correspondencias entre distintos conceptos y brindan estructura a la ontología).

2.1.1. Relaciones representadas en una ontología

Algunas de las relaciones que se representan en una ontología son[26]:

- Es un: se utiliza para la propiedad de inclusión, es decir el segundo concepto es más general que el primer concepto.
- Instancia de : representa objetos determinados de un concepto.
- Es parte de: esta relación se utiliza para indicar composición entre conceptos.

Por ejemplo, se tienen los conceptos carro, vehículo y motor. La relación entre carro y vehículo quedaría como “carro es un vehículo”, mientras que la relación entre carro y motor quedaría como “motor es parte de carro”, por otro lado para la relación “instancia de” definimos el objeto “carro-rojo” el cual es un carro en particular, de esta manera se tendría “carro-rojo es instancia de carro”.

Las ontologías representan una gran dificultad para su creación, descripción y anotación, pero uno de los beneficios más importantes es que por medio de éstas se obtiene una mejor recuperación de información.

2.1.2. Web Semántica

Es un proyecto creado por World Wide Web Consortium para desarrollar tecnologías que permitan publicar datos legibles por aplicaciones informáticas. Con estas tecnologías se pretende mejorar Internet con el objetivo de ampliar la interoperabilidad entre los sistemas informáticos.

La Web semántica es una Web extendida, es decir, se pueden obtener soluciones a problemas que se presentan cuando se realiza una búsqueda de información, debido a que se utiliza una infraestructura común, lo cual permite compartir, procesar y transferir información de manera sencilla. Para lograr esto utiliza lenguajes universales como son RDF y OWL.

A continuación se muestra en la Figura 2.1 un ejemplo del resultado que se obtiene en un buscador Web actualmente:

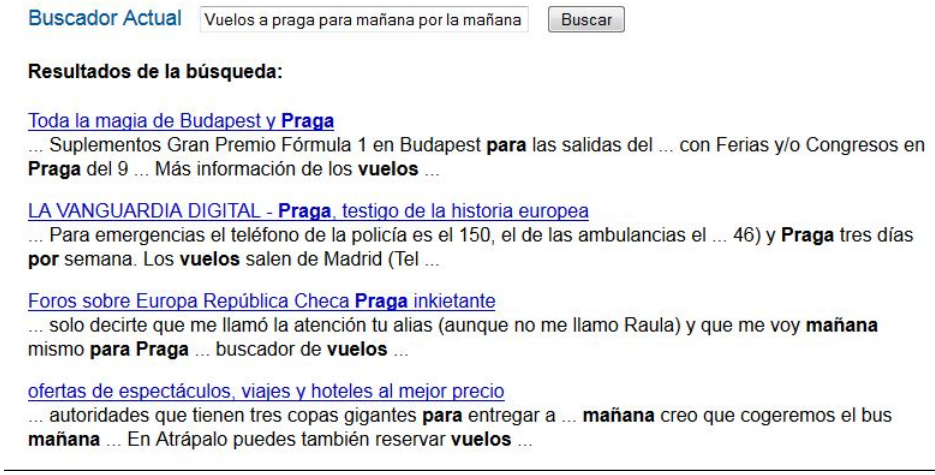

Figura 2.1: Ejemplo de búsqueda en Internet[27].

En la Figura 2.2 se muestra como se obtendría el resultado de la misma búsqueda, pero esta vez con un buscador semántico:

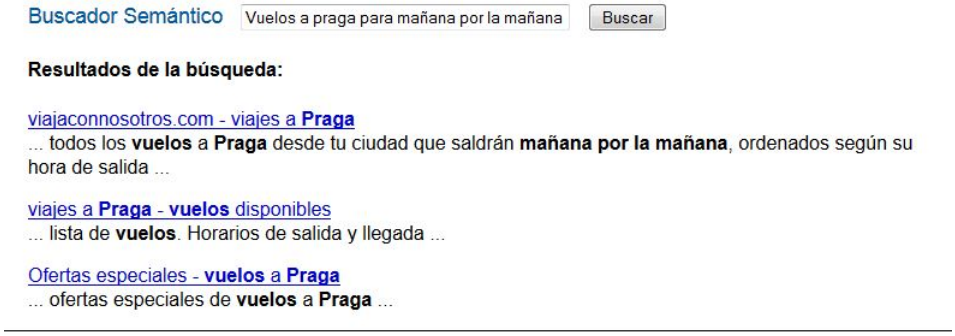

Figura 2.2: Ejemplo de búsqueda con un buscador semántico en Internet[27].

Como se observa en la Figura 2.2 con el buscador semántico se obtienen resultados más precisos que con el buscador común de Internet.

2.1.3. RDF

Es un lenguaje que permite representar la información de recursos que se encuentran en la Web, principalmente representa los metadatos de los recursos.

Sus características facilitan la mezcla de datos, sin necesidad de requerir que los datos tengan un esquema similar. El uso de este modelo simple, permite que los datos estructurados y semiestructurados se mezclen o se intercambien entre aplicaciones sin perder información[45].

RDF extiende la estructura de enlaces Web utilizado uris[1] para nombrar la relación entre dos cosas, describe los recursos como propiedades y sus valores, de esta manera se tiene una estructura con forma de un grafo dirigido y etiquetado, donde los bordes representan el enlace llamado entre dos recursos, representados por los nodos del grafo.

Un ejemplo de esto se muestra en la Figura 2.3, en la cual se representa una persona cuyo nombre es Eric Miller, el cual tiene el correo em@w3.org y su título correspondiente es Dr, se puede observar como se representa mediante uris que representan las propiedades de la persona.

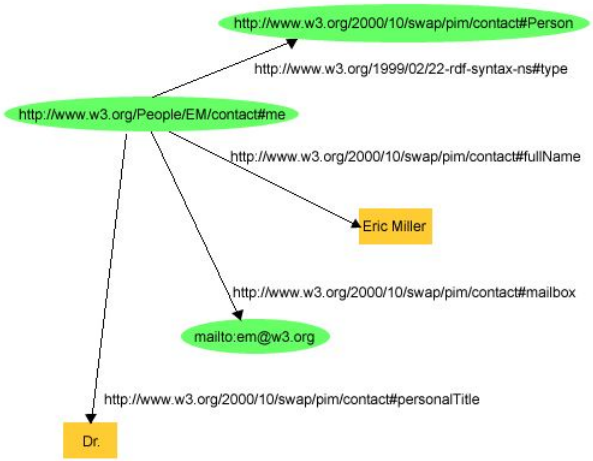

Figura 2.3: Representación de una persona en RDF[45].

Su sintaxis se basa en XML, como se puede observar en la Figura 2.4, la cual es la representación de la Figura 2.3 con RDF:

[1] Uris se refiere a identificador uniforme de recursos: es una cadena de caracteres que permite identificar de manera inequívoca un recurso, este puede ser un archivo, una página de Internet, un servicio, entre otros. Una URI es más completa que una URL porque permite incluir una subdirección.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">

  <contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:fullName>Eric Miller</contact:fullName>
    <contact:mailbox rdf:resource="mailto:em@w3.org"/>
    <contact:personalTitle>Dr.</contact:personalTitle>
  </contact:Person>

</rdf:RDF>
```

Figura 2.4: Ejemplo de RDF/XML, el cual describe una persona[45].

En la Figura 2.4 se muestra que se utilizan uris, las cuales tienen propiedades en este caso como *fullName*, *mailbox* y *personalTitle*. RDF define un recurso como cualquier cosa que pueda ser identificado por una referencia a una uri, por lo que al utilizar RDF se puede describir cualquier cosa y el estado de la relación entre las cosas.

2.1.4. OWL

Como se mencionó anteriormente, OWL es un lenguaje de marcado, que se utiliza cuando la información esta contenida en documentos y estos tienen que ser procesados, esta diseñado para procesar el contenido de la información, en lugar de solo presentarla a los usuarios. También se utiliza para representar términos y relaciones que existan entre estos (ontología).

Las herramientas existentes de OWL soportan el lenguaje RDF/XML, aunque los lenguajes turtle y Manchester syntax también son utilizados en algunos editores de ontologías.

Existen tres sub-lenguajes de OWL, que son utilizados dependiendo de las necesidades de expresión que se tienen, estos son[24]:

- OWL lite: este sub-lenguaje es suficiente para cuando solo se quiere tener relaciones de clasificación de jerarquía de conceptos y se tienen restricciones simples.
- OWL dl: con este sub-lenguaje se permite tener la máxima expresividad
- OWL full

Se cumplen las siguientes relaciones, pero las inversas no se cumplen[24]:

- Cada ontología creada con OWL lite es una ontología válida de OWL dl.
- Cada ontología creada con OWL dl es una ontología válida de OWL full.
- Cada conclusión válida de OWL lite es una conclusión válida de OWL dl.
- Cada conclusión válida de OWL dl es una conclusión valida de OWL full.

2.2. Taxonomías

La taxonomía de acuerdo a su significado etimológico es la ciencia que trata de los principios, métodos y fines de la clasificación y procura la organización jerarquizada[47].

Una taxonomía es una colección de términos de vocabulario controlado jerárquicamente, donde cada término de la taxonomía tiene una o más relaciones padre-hijo con otros términos de la misma taxonomía.

En particular, una ontología define los términos y conceptos organizados jerárquicamente, los cuales representan un área de conocimiento y una de sus utilidades es dar soporte a diversas aplicaciones que necesiten del conocimiento que ésta representa.

En ocasiones se utiliza la palabra ontología para referirse a los glosarios, diccionarios de datos, tesauros, taxonomías, esquemas, entre otros, pero una ontología formal debe estar expresado en un lenguaje de representación específico para ontologías, como lo es RDF u OWL.

Una ontología que contiene más relaciones que la relación clase-subclase es más rica (es decir representa mejor el conocimiento) que una taxonomía con solamente la relación clase-subclase[40], por lo que las taxonomías son una subclase de las ontologías.

Cabe mencionar que una de las mayores ventajas de las taxonomías es que permiten mejorar la navegación y el desarrollo de sistemas de búsqueda y en la recuperación de información.

En el presente trabajo, acordamos referirnos al producto obtenido mediante nuestras expresiones OLAP, UDFs y SPs como taxonomía más que ontología, aunque trabajos relacionados a este[1] han nombrado a las estructuras obtenidas como ontologías. Las taxonomías obtenidas en este trabajo son las estructuras con relación “es un”.

Capítulo 3

SPs, UDFs y Cubos OLAP en SQL

Como se mencionó en la introducción de este trabajo, se utilizan herramientas de SQL, desde la obtención de los términos hasta la obtención de los términos relacionados y la jerarquía que existe entre cada par de términos frecuentes, así como en la implementación de las métricas que se utilizan para medir la calidad de las taxonomías generadas, por esta razón se presenta a continuación una breve descripción de las herramientas utilizadas.

El Lenguaje Estructurado de Consulta (SQL, por sus siglas en inglés) se utiliza para crear las estructuras necesarias para almacenar los datos, así como actualizar y recuperar información que se encuentra almacenada en una base de datos[34].

Cada sistema manejador de bases de datos (DBMS, por sus siglas en inglés) tiene su propio lenguaje procedural, el cual es extendido de SQL. En este caso, se utilizó el sistema manejador SQL Server[53], por lo cual se explica que es y algunas características de T-SQL.

T-SQL (Transact-SQL) es una implementación de Microsoft y Sybase del SQL, esta basado en los estándares creados por ANSI[1], con la diferencia de que se agregaron varias mejoras a la funcionalidad, del cual se obtiene como resultado un lenguaje poderoso y versátil.

> "T-SQL fue diseñado con el exclusivo propósito de recuperar información y manipular la información."[29]

La cita anterior hace énfasis en que T-SQL será utilizado para la obtención de información y su manipulación por medio de las herramientas que a continuación se describen.

T-SQL contiene dos tipos de instrucciones:

[1] Es el Instituto Nacional Estadounidense de Estándares (American National Standars Institute) que supervisa el desarrollo de estándares para productos, servicios, procesos y sistemas en los Estados Unidos. http://www.ansi.org/ (30/12/2013 1:54 am).

- Lenguaje de definición de datos (DDL): es el subconjunto de T-SQL que se utiliza para crear, alterar o eliminar objetos en la base de datos.
- Lenguaje de manipulación de datos (DML): es el subconjunto de T-SQL con el cual se definen y gestionan todos los objetos que componen la base de datos.

Una de las principales características de T-SQL es que un conjunto debe ser considerado como una simple entidad, se debe considerar a alguna tabla entera como una entidad, en lugar de tomar cada tupla de la tabla como un objeto individual.

Dentro del DML se encuentran los procedimientos almacenados y las funciones definidas por el usuario, los cuales se presentan a continuación.

3.1. Procedimientos almacenados (SP)

Los procedimientos almacenados pueden ser definidos como un conjunto de instrucciones del DML. Para ejecutarse en SQL Server se realiza una llamada por medio de su nombre o con la instrucción EXECUTE ó EXEC[42].

Éstos se almacenan en la base de datos y son generalmente desarrollados en un lenguaje de bases de datos.

Estos procedimientos pueden recibir parámetros o no y devolver valores, pueden ser temporales (para la sesión local) ó global (para más sesiones del usuario). El código compilado del procedimiento almacenado se guarda en la memoria cache después de su primera ejecución con el objeto de que futuras ejecuciones del mismo sean más eficientes.

Los procedimientos almacenados pueden tener declaración y asignación de variables, e instrucciones como IF, CASE, WHILE y GOTO. También permiten instrucciones como SELECT, UPDATE, INSERT, DELETE y se puede asegurar la consistencia de los datos al utilizar dentro de éstos las transacciones, bloqueos y métodos para evitar y tratar las excepciones.

Los usos más comunes de procedimientos almacenados son[28]:

- Mejorar el rendimiento.
- Encadenamiento de instrucciones.
- Seguridad en la ejecución.
- Manipulación de datos del sistema.
- Implementación de las reglas del negocio.
- Tratamiento en cascada.
- Encapsular un proceso grande.

3.1.1. Sintaxis

La sintaxis para crear un procedimiento almacenado se muestra en la Figura 3.1, de donde se observan los elementos[30]:

- ;number el cual es un entero opcional utilizado para agrupar procedimientos con el mismo nombre.
- @parameter puede ser uno o más parámetros los que recibirá el procedimiento.
- VARYING especifica el conjunto de resultados admitidos como parámetros de salida.
- default indica que el procedimiento se puede ejecutar sin especificar el valor del parámetro de salida.
- OUTPUT indica que se trata de un parámetro de salida, el cual puede ser un marcador de posición de un cursor.
- RECOMPILE indica que el motor de la base de datos no debe almacenar en cache el plan de consulta del procedimiento.
- ENCRYPTION indica que el motor de la base de datos cambiará el texto del procedimiento en un formato confuso.

```
CREATE PROC [ EDURE ] [ owner. ] procedure_name [ ; number ]
  [ { @parameter data_type }
    [ VARYING ] [ = default ] [ OUTPUT ]
  ] [ ,...n ]

[ WITH
  { RECOMPILE | ENCRYPTION | RECOMPILE , ENCRYPTION } ]

[ FOR REPLICATION ]

AS sql_statement [ ...n ]
```

Figura 3.1: Sintaxis para crear un procedimiento almacenado[30].

Para la modificación de un procedimiento almacenado se utiliza el comando *ALTER PROCEDURE* que permite modificar el cuerpo del procedimiento y para la eliminación se utiliza *DROP PROCEDURE*.

A continuación se muestra como ejemplo un procedimiento almacenado en el cual se calcula el peso de los términos en el corpus[2]. Lo que se realiza en la función es que en las lineas 5 y 6 se obtiene el número de documentos que se tienen almacenados en la base de datos, mientras que en las lineas 7 a 10 se inserta en la tabla *term_weight_doc* los términos, su peso y el número de documentos en lo que se puede encontrar cada término.

[2]Corpus se refiere al conjunto de documentos de un tema específico que se utilizó para obtener las taxonomías.

Algoritmo 3.1 Ejemplo de un procedimiento almacenado

```
1  CREATE PROCEDURE tdidf_term
2  AS
3  BEGIN
4      DECLARE @num_docs INT
5      SELECT @num_docs=
6          (SELECT COUNT(*) from [document])
7      INSERT INTO term_weight_doc (term, idft, tdft)
8      SELECT term, log(convert(FLOAT,@num_docs)/COUNT(
           document_id)),
9          COUNT(document_id) FROM [term_doc]
10     GROUP BY term
11 END
```

3.1.2. Cursores

La instrucción de SQL Select devuelve un conjunto de tuplas, en el caso en que es necesario recorrerlos y realizar operaciones sobre ellos se utilizan los cursores. Un cursor es una herramienta que provee SQL para recorrer las tuplas obtenidas de una consulta.

La sintaxis para crear un cursor se muestra en la Figura 3.2:

```
DECLARE cursor_name [ INSENSITIVE ] [ SCROLL ] CURSOR
     FOR select_statement
     [ FOR { READ ONLY | UPDATE [ OF column_name [ ,...n ] ] } ]
[;]
Transact-SQL Extended Syntax
DECLARE cursor_name CURSOR [ LOCAL | GLOBAL ]
     [ FORWARD_ONLY | SCROLL ]
     [ STATIC | KEYSET | DYNAMIC | FAST_FORWARD ]
     [ READ_ONLY | SCROLL_LOCKS | OPTIMISTIC ]
     [ TYPE_WARNING ]
     FOR select_statement
     [ FOR UPDATE [ OF column_name [ ,...n ] ] ]
[;]
```

Figura 3.2: Sintaxis de cursores en SQL Server[44].

Como se observa en la Figura 3.2 los elementos para crear un cursor son:

DECLARE nombre CURSOR FOR consulta

Y contiene opciones que permiten declararlo solo para lectura, para actualizar los datos ó en la versión extendida se permite que se desplace de la primera fila a la última, entre otras opciones.

Posteriormente para utilizar el cursor, se abre con la sentencia:

OPEN nombre

Se le asigna la siguiente tupla sobre el cual se realizará el proceso con:

FETCH nombre INTO variable1, ... variableN

Se cierra con la instrucción:

CLOSE nombre

Y finalmente se elimina con la sentencia:

DEALLOCATE nombre

3.1.3. Características de los SPs

Algunas ventajas de utilizar procedimientos almacenados son:

- Permiten la programación modular.
- Mejoran la seguridad de la aplicación al incluir atributos como permisos o cadenas de propiedad.
- Simplifican la creación y mantenimiento de programas que realizan operaciones sobre una base de datos.

3.2. Funciones definidas por el usuario (UDF)

Una función definida por el usuario (UDF, por sus siglas en inglés) es una función personalizada. Las UDFs son una técnica de T-SQL para mejorar el proceso de desarrollo. La mayor restricción es que no permite ejecutar sentencias que cambien el estado de la base de datos.

A continuación se mencionan los tres diferentes tipos de funciones definidas por el usuario, los cuales se explican más adelante[42]:

- Funciones escalares.
- Funciones en linea de múltiples sentencias.
- Funciones de agregado.

3.2.1. Funciones escalares

Este tipo de funciones definidas por el usuario retornan un único valor escalar, aunque se aplica a todas las tuplas y pueden recibir o no parámetros. Se puede ejecutar dentro del cuerpo de estas funciones cualquier cantidad de instrucciones, desde muy sencillas hasta muy complejas.

La sintaxis para crear una función escalar se muestra en la Figura 3.3:

```
CREATE FUNCTION [ schema_name. ] function_name
( [ { @parameter_name [ AS ][ type_schema_name. ] parameter_data_type
    [ = default ] [ READONLY ] }
    [ ,...n ]
  ]
)
RETURNS return_data_type
    [ WITH <function_option> [ ,...n ] ]
    [ AS ]
    BEGIN
        function_body
        RETURN scalar_expression
    END
[ ; ]
```

Figura 3.3: Sintaxis de una función escalar.[35]

Las instrucciones que se ejecutan dentro de esta función escalar se realizan para cada tupla con las cuales es llamada la función. Es decir, una función escalar es llamada de la forma:

SELECT dbo.nom_funcion FROM tabla

Para entender mejor la Figura 3.3 se muestra el Algoritmo 3.2 con el código de una función escalar que se utilizó para obtener el peso de un término en el corpus correspondiente:

Algoritmo 3.2 Ejemplo de una función escalar

```
1  CREATE FUNCTION [dbo].[term_weigth]
2  ( @num_docs INT,
3    @doc_id_rep INT,
4    @txt varchar(MAX),
5    @doc INT )
6  RETURNS FLOAT AS
7   BEGIN
8     DECLARE @peso INT
9
10      SELECT @peso=log(convert(float,@num_docs)/
            @doc_id_rep)
11      FROM count_frec(@txt, @doc)
12      GROUP BY term
13    RETURN @peso
14  END
```

En la cual se muestra que internamente llama a otra función llamada *count_frec*, la cual es llamada en la parte del *from* y no en la parte del *select* de la consulta, por lo que esta función llamada internamente no es escalar, si no una función en linea de múltiples sentencias, que a continuación se explica. Esta función es equivalente al procedimiento almacenado mostrado en la sección 3.1.1.

3.2.2. Funciones en linea de múltiples sentencias

Las funciones en linea también pueden o no recibir parámetros, pero son diferentes a las funciones escalares. A diferencia de las funciones escalares, estas funciones devuelven un conjunto de resultados (un resultado por tupla a la que se le aplica la función) y no se puede utilizar la clausula ORDER BY.

En la Figura 3.4 se muestra la sintaxis de este tipo de UDF.

```
CREATE FUNCTION [ schema_name. ] function_name
( [ { @parameter_name [ AS ] [ type_schema_name. ] parameter_data_type
    [ = default ] [ READONLY ] }
    [ ,...n ]
  ]
)
RETURNS TABLE
    [ WITH <function_option> [ ,...n ] ]
    [ AS ]
    RETURN [ ( ] select_stmt [ ) ]
[ ; ]
```

Figura 3.4: Sintaxis de una función en linea de múltiples sentencias[35].

La manera para utilizar estas funciones es llamándolas en la parte del *FROM* de una consulta, como se muestra a continuación:

*SELECT * FROM dbo.nom_funcion*

3.2.3. Funciones de agregado

Este tipo de funciones realiza operaciones de cálculo en el conjunto de valores que reciben y devuelven un solo valor, como las funciones escalares. Se utiliza comúnmente con la clausula GROUP BY en una consulta de SQL.

Se muestra en la Figura 3.5 la sintaxis de una función de agregado.

```
CREATE FUNCTION [ schema_name. ] function_name
( [ { @parameter_name [ AS ] [ type_schema_name. ] parameter_data_type
    [ = default ] [READONLY] }
    [ ,...n ]
  ]
)
RETURNS @return_variable TABLE <table_type_definition>
    [ WITH <function_option> [ ,...n ] ]
    [ AS ]
    BEGIN
        function_body
        RETURN
    END
[ ; ]
```

Figura 3.5: Sintaxis de una función de agregado[35].

Las funciones de agregado se pueden llamar desde la clausula SELECT ó desde la clausula HAVING. Las funciones de agregado que se tienen ya definidas en T-SQL[35] son:

- AVG: devuelve el promedio de los valores de entrada.
- CHECKSUM_AGG: realiza el checksum de los valores en una agrupación.
- COUNT: cuenta el número de tuplas que pertenecen a un determinado grupo.
- MIN: regresa el valor mínimo del conjunto de valores de entrada.
- MAX: encuentra y devuelve el valor máximo del conjunto de valores de entrada.
- COUNT_BIG: realiza la misma operación que COUNT, con la diferencia que el valor que regresa es del tipo bigint.
- GROUPING: indica si una columna ha sido agrupada, la cual regresa el valor de 1 en caso verdadero y 0 en caso contrario.
- GROUPING_ID: calcula el nivel de agrupamiento, solo se puede utilizar en la parte del select, having o con el order by.
- SUM: suma todos los valores de entrada.
- STDEV: devuelve la desviación estándar de los valores de entrada.
- STDEVP: calcula la desviación estándar de la población en los valores especificados
- VAR: regresa la varianza de todos los conjuntos de valores seleccionados.
- VARP: retorna la varianza de la población de los valores especificados.

3.3. SP VS UDF

Algunas de las diferencias que existen entre los procedimientos almacenados y una función definida por el usuario se mencionan a continuación[28, 29, 42, 55]:

- Se tienen restricciones acerca de las sentencias de T-SQL (definido en 3) que se pueden usar en el cuerpo de una UDF, por ejemplo, no se permite ejecutar un INSERT a tablas que ya están creadas en la base de datos al utilizar UDFs.
- Las UDFs proporcionan mas flexibilidad que un procedimiento almacenado (SP), permitiendo que la salida de las UDF se utilice más fácilmente, ya sea en otras consultas o como parámetro de otra función.

- Se puede hacer referencia a un procedimiento almacenado dentro de una UDF.
- Una UDF puede ser rehusada en diferentes secciones de una consulta.
- Las UDFs permiten mayor mantenibilidad y claridad del código que los SPs.
- Las UDFs pierden portabilidad.
- Las UDFs escalares no utilizan CURSORES como los SP, ya que se ejecutan por cada tupla.
- Los planes de ejecución de las UDFs se almacenan de manera similar que para los SPs lo que permite al optimizador reutilizarlos en diferentes consultas.

En las siguientes secciones se muestran los planes de ejecución de una función con ambas versiones, esto para observar las diferencias entre cada una.

3.3.1. Plan de ejecución de una UDF

A continuación se muestra una serie de figuras con el plan de ejecución de la UDF que calcula el peso de los términos en el corpus, para utilizar esta función se realiza la siguiente consulta:

SELECT term, idft,tdft from onto_udf.dbo.tdidf_term('rank')

En la primera parte del plan (Figura 3.6) se puede observar que se realiza una consulta con la llamada a la función tdidf_term.

	StmtText
1	|--Sequence
2	|--Table-valued function(OBJECT:([onto_udf].[dbo].[tdidf_term]))
3	|--Table Scan(OBJECT:([onto_udf].[dbo].[tdidf_term]))

Figura 3.6: Primera parte del plan de ejecución.

En la segunda parte se muestra el código fuente de la función, esto se observa en la Figura 3.7

	StmtText
1	UDF: [onto_udf].[dbo].[tdidf_term]
2	CREATE function tdidf_term (@term varchar(100)) returns @term_weight_doc table (term varchar(100), idft float, tdft float) ...

Figura 3.7: Segunda parte del plan de ejecución.

En la tercera parte mostrada en la Figura 3.8 se realiza un escaneo sobre un índice, para buscar la llave primaria del documento, posteriormente se realiza un conteo para conocer el número de apariciones de los términos en cada documento.

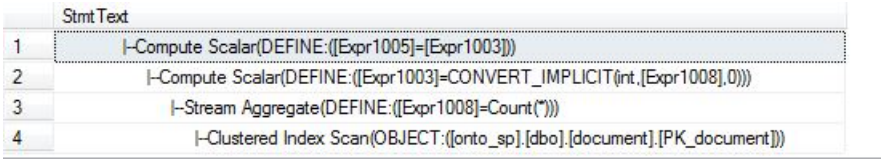

Figura 3.8: Tercera parte del plan de ejecución.

Posteriormente en la Figura 3.9 se tiene la consulta que inserta en una tabla de salida de la función los valores obtenidos en los pasos anteriores, estos se agrupan por término.

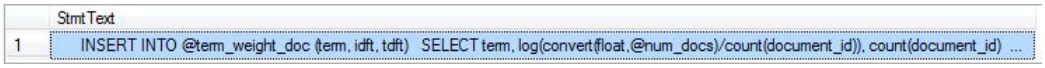

Figura 3.9: Cuarta parte del plan de ejecución.

Finalmente se tiene la Figura 3.10 con el plan de la consulta de la quinta parte, donde se realizan las operaciones de conversión a flotante, la división y el conteo de las apariciones de los términos.

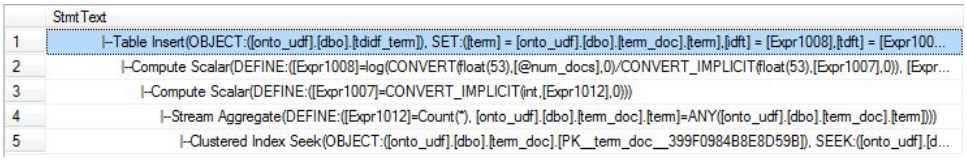

Figura 3.10: Quinta parte del plan de ejecución.

3.3.2. Plan de ejecución de un SP

Ahora se muestra el plan de ejecución para el mismo proceso, pero esta vez realizado con procedimientos almacenados. Para ejecutarlo se utiliza la siguiente consulta:

exec tdidf_term

En la primera parte mostrada en la Figura 3.11 del plan se observa la sentencia con la que es llamado el procedimiento almacenado y el código fuente.

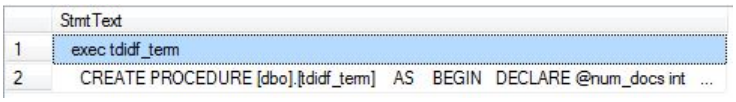

Figura 3.11: Primera parte del plan de ejecución.

En la segunda parte se realizan los mismos pasos que en la segunda parte de la UDF mencionada anteriormente, esto se observa en la Figura 3.12

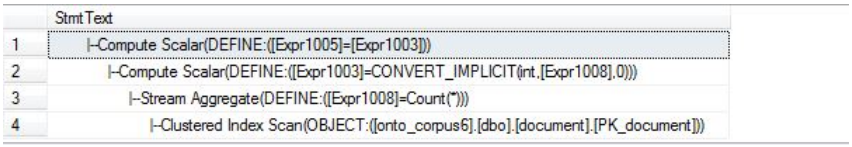

Figura 3.12: Segunda parte del plan de ejecución.

En la tercera parte representada en la Figura 3.13 del plan también puede observarse la sentencia que inserta los valores obtenidos en una tabla, pero esta vez con el procedimiento almacenado si se insertará en una tabla existente, en lugar de una tabla temporal de salida.

Figura 3.13: Tercera parte del plan de ejecución.

La última parte del plan mostrada en la Figura 3.14 se observa como se realizan las operaciones necesarias para obtener los valores que serán insertados en la tabla mencionada en la tercera parte.

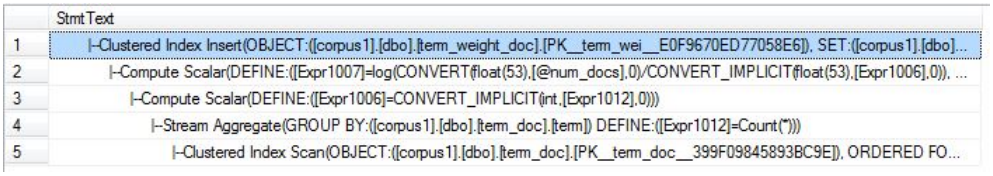

Figura 3.14: Cuarta parte del plan de ejecución.

Como puede observarse, en esencia tienen el mismo plan de ejecución, aunque la diferencia más notable es que el procedimiento almacenado si inserta los valores en una tabla creada en la base de datos y la función definida por el usuario no lo hace.

3.4. OLAP

El término de Procesamiento Analítico en Linea (por sus siglas en inglés, OLAP) fue definido como tal en un artículo de Arbor Software Corp, en 1993, donde se menciona que es “el proceso interactivo de crear, mantener, analizar y elaborar informes sobre datos”, donde los datos están percibidos como si estuvieran almacenados en un arreglo multidimensional[56].

Existen arquitecturas diferentes para los sistemas OLAP, estos pueden ser:

- ROLAP: significa Procesamiento Analítico Relacional en linea , es decir, se trata de sistemas y herramientas de Procesamiento Analítico en línea

(OLAP, por sus siglas en inglés) construidos sobre una base de datos relacional. No requieren, en principio, el almacenamiento de la información, ya que pueden acceder directamente a la fuente de dichos datos, las herramientas ROLAP acceden a los datos de una base de datos relacional y generan consultas SQL para calcular la información al nivel apropiado cuando un usuario final lo requiere.

- MOLAP: significa Procesamiento Analítico Multidimensional en linea, el cual usa una base de datos multidimensional para proporcionar el análisis y poder visualizar los datos multidimensionalmente. Se diferencia significativamente en que requiere un pre-procesamiento y almacenamiento de la información contenida en el cubo OLAP. MOLAP almacena estos datos en una matriz de almacenamiento multidimensional optimizada. Como resultado se obtiene un mejor desempeño, pero se requiere un mayor espacio de almacenamiento.

- HOLAP: Almacena los datos en una base de datos relacional mientras que las agregaciones de estos, se almacenan en una base de datos multidimensional.

El ROLAP es una alternativa a la tecnología MOLAP (Multidimensional OLAP) que se construye sobre bases de datos multidimensionales. Ambos tipos de herramientas, tanto ROLAP como MOLAP, están diseñadas para realizar análisis de datos a través del uso de modelos de datos multidimensionales, aunque en el caso de ROLAP estos modelos no se implementan sobre un sistema multidimensional, sino sobre un sistema relacional clásico. ROLAP requiere de menor almacenamiento que MOLAP, pero es más lento.

Las bases de datos MOLAP son más rápidas que sus semejantes ROLAP, aunque ambos tipos de herramientas están diseñadas para realizar análisis de datos a través de un modelo de datos multidimensional.

3.4.1. Cubo

Es una estructura multidimensional, en la que el almacenamiento físico de los datos se realiza en un vector multidimensional, con las siguientes características:

- Las dimensiones de los datos se colocan sobre los ejes.

- Los datos se almacenan en las celdas del cubo.

- Los cubos de datos pueden llegar a crecer hasta tener n número de dimensiones, en tal caso estos se convierten en hipercubos.

- Las dimensiones de un cubo se pueden resumir como una jerarquía, por ejemplo si en una dimensión se encuentra un día en particular, esta dimensión se puede clasificar en semana y esta a su vez en un mes y así sucesivamente puede clasificarse en una categoría de mayor jerarquía. Estas jerarquías permiten adentrarse en el cubo y observar con mayor detalle los datos o en caso contrario observarlos a grandes rasgos.

- Las dimensiones que permanecen constantes en un cubo se llaman rebanadas (slices), los valores de la dimensión se llaman miembros (members), mientras que el nivel de una posición es su posición en una jerarquía. Por ejemplo si se tiene una dimensión Fecha, sus niveles son año, semestre y mes, por lo que refiriéndose al semestre 1, se refiere al nivel 2 del cubo.

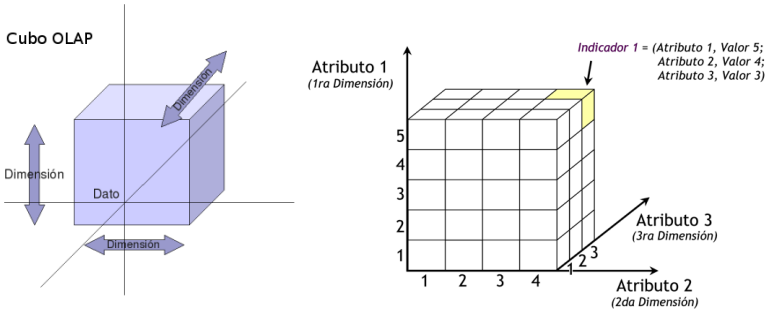

Figura 3.15: Ejemplo de un cubo de datos con dimensiones de fechas [58].

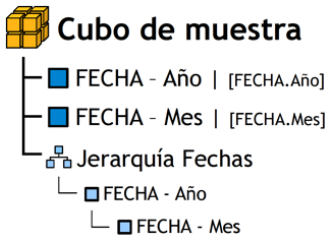

Figura 3.16: Ejemplo de la jerarquía del cubo con dimensiones de fechas[58].

3.4.2. Construcción y acceso a un cubo

Los cubos son estáticos, por lo que no están sujetos a cambios, es decir, si se requiere de un nuevo valor en alguna de las dimensiones de los cubos, estos deben ser rediseñados y generados nuevamente. El proceso de construcción de un cubo es una tarea compleja, ya que requiere un gran trabajo de diseño. Los parámetros por los cuales se analizarán los datos se conocen como las dimensiones del cubo y para accesar a ellos se indexan a partir de los valores de las dimensiones o ejes.

Una desventaja de esta forma de almacenamiento es que una vez poblada la base de datos no puede recibir cambios en su estructura. Para ello sería necesario rediseñar el cubo.

Para acelerar el tiempo de acceso a los datos del cubo, éste se mantiene en memoria. A esto se le llama cache de cubo.

Las principales operaciones que se pueden realizar en un cubo son[18]:

- Slice: es la acción de seleccionar un subconjunto rectangular de datos del cubo si se selecciona un único valor de una dimensión, del cual resulta un cubo de menor dimensión.
- Dice: Es la acción de formar un subcubo si se seleccionan diversos valores de múltiples dimensiones.
- Roll-up (Enrollar): esta acción permite navegar entre los niveles del cubo, que van desde lo más abstracto a lo más detallado.
- Drill-down (Profundizar): esta acción resume los datos a lo largo de una dimensión.
- Pivot (Rotar): Permite rotar el cubo para visualizar sus diferentes caras.

Capítulo 4

Metodología de extracción de taxonomías y medición de calidad

A continuación se presenta la metodología utilizada para la extracción de las taxonomías y la medición de calidad de las mismas, y se explica como se utilizó cada concepto en el contexto de este trabajo.

Para determinar la importancia de los conceptos principales se utilizaron métodos como la correlación, la distancia, el soporte y la confianza y una modificación a este método el cual indica la relevancia de la frecuencia de dos conceptos (términos) respecto a la colección completa. Y como medida adicional se utilizó el peso, el cual representa que tan seguido aparecen los conceptos en la colección.

Una vez obtenidos los conceptos principales se filtraron los pares de conceptos cuyo peso y correlación sean superiores a un umbral. De estos se obtendrá la relación "tiene un", la cual proporciona una jerarquía de los conceptos.

4.1. Operaciones realizadas

En los experimentos realizados se utilizaron varios métodos para encontrar la jerarquía de los términos que están relacionados para la definición de las ontologías y taxonomías de los corpus.

Una vez que se procesaron las palabras contenidas en los documentos de los corpora y se obtienen los k números más frecuentes estos son considerados como términos, es decir son los posibles conceptos que conformarán la taxonomía.

A continuación se explican las operaciones que se realizaron en cada paso del procesamiento, se explica para que se utilizaron y que significan.

4.1.1. Stemming

Se utilizó el algoritmo de Porter Stemmer, obtenido de [36], por lo general se refiere a un proceso heurístico crudo que corta el final de las palabras, con la esperanza de lograr este objetivo correctamente la mayoría de las veces, e incluye la eliminación de los afijos derivativos[59].

Cabe mencionar que este algoritmo es diferente para cada idioma, por lo que en el caso de esta tesis se utilizó el algoritmo para el idioma inglés, también en la versión de T-SQL.

Por ejemplo, a las palabras: "policies", "Policy", "Policies" y "policy" al aplicarles el algoritmo de Porter Stemmer se obtiene el término "polici" por lo que son tomadas como si fueran el mismo término, es decir, si en el documento i aparecen las palabras policies y policy una vez cada una, en total el término "polici" tendrá frecuencia=2 en el documento i.

Lo mismo pasa con los términos "scoring", "SCORE", "scores" y "Scores" que se agrupan en el término "score".

4.1.2. "Stop Words"

Las "stop words" son consideradas como las palabras más frecuentes que por lo general ocurren en todos los documentos con una frecuencia alta, esto quiere decir que no aportan información importante. Entre estas se consideran los pronombres, las preposiciones, las conjunciones y disyunciones. Para descartar estas palabras se utilizó una lista de "stop words" llamada "stoplist" que se tiene disponible en el sistema manejador de la base de datos. Para tener disponible la lista de "stop words" es necesario ejecutar la instrucción[54]:

CREATE FULLTEXT STOPLIST nombre_tabla

Esta instrucción genera una tabla con "stop words" de diferentes idiomas, pero para fines de este trabajo solo se utilizaron las palabras del idioma inglés.

	stopword	language	language_id
30	S	English	1033
31	T	English	1033
32	U	English	1033
33	V	English	1033
34	W	English	1033
35	X	English	1033
36	Y	English	1033
37	Z	English	1033
38	about	English	1033
39	after	English	1033
40	all	English	1033
41	also	English	1033
42	an	English	1033
43	and	English	1033
44	another	English	1033
45	any	English	1033
46	are	English	1033
47	as	English	1033
48	at	English	1033
49	be	English	1033
50	because	English	1033
51	been	English	1033

Figura 4.1: Tabla con ejemplos de “stop words”.

En la Figura 4.1 se puede observar una parte del contenido de la tabla de “stop words” y dos columnas las cuales indican el idioma correspondiente a las “stop words” y el *id* del idioma correspondiente.

4.1.3. Correlación

La correlación indica que tan frecuente es un término con respecto a otro término en los documentos de un corpus, se definen t y $t1$ términos del corpus, la correlación indica que tan seguido aparecen ambos términos en el mismo documento y que tan seguido aparece uno sin el otro[1]. Para obtener la correlación entre ambos términos se realiza el siguiente cálculo[1]:

$$\rho_{t,t1} = \frac{nQ_{t,t1} - L_t L_{t1}}{\sqrt{nQ_{t,t} - L_t^2}\sqrt{nQ_{t1,t1} - L_{t1}^2}}$$

De donde n es el número de documentos del corpus, y para obtener Q y L se utilizan las formulas [1, 14]:

$$L_t = \sum_i x_i \text{ y } Q_t = \sum_i x_i \star x_i^T$$

Donde x es la frecuencia del termino t en el documento i, por lo que L es la suma de las frecuencias del termino t y Q contiene la media cuadrada de las frecuencias del término t.

Dependiendo del valor de la correlación se puede concluir que mientras más se acerque el valor de correlación de los términos t y t1 significa que están muy relacionados y por el contrario entre más se aleje el valor de la correlación, los términos t y t1 están menos relacionados.

4.1.4. Peso

El peso de un término en un documento indica que tanta información relevante aporta respecto a éste. A continuación se muestra la función para calcular el peso de un término para todo el corpus[37]:

$$idf_t = log \frac{N}{df_t}$$

Donde df_t es el número de documentos en los que aparece el término t y N es el total de documentos en el corpus. Para asignarle un peso a un término en cada documento, se utiliza la función[37]:

$$tf\_idf_{t,i} = tf_{t,i} \times idft$$

Donde $tf\_\text{idf}_{t,i}$ es el número de ocurrencias del término t en el documento i. El resultado obtenido de $\text{tf\_idft}_{t,i}$ es [37]:

- Alto cuando el término t aparece muchas veces en un número pequeño de documentos o en muchos documentos
- Bajo cuando el término t aparece pocas veces en un documento.
- Más bajo cuando el término t aparece en todos los documentos del corpus.

4.1.5. Orden

El orden se utilizó para obtener la jerarquía de cada par de términos frecuentes, es decir, se utiliza para decidir si el término A es padre del término B o viceversa.

Para obtener el valor del orden de un par de términos (término1 y término2) se sigue el Algoritmo 4.1, donde posición se refiere a la ubicación de cada término de acuerdo a un documento, para ilustrar mejor este concepto se muestra la Figura 4.2 que contiene un fragmento de un documento almacenado en la base de datos:

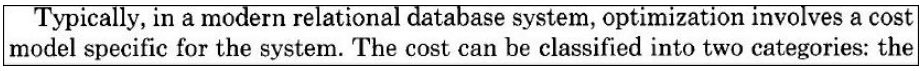
Typically, in a modern relational database system, optimization involves a cost model specific for the system. The cost can be classified into two categories: the

Figura 4.2: Ejemplo de definición de posición

La asignación de posiciones se realiza por cada término, al tomar el espacio en blanco como el separador entre términos, y se recuerda que los caracteres como son las comas, guiones y diagonales entre otros no se toman en cuenta, donde se obtiene la asignación de la siguiente manera.

Sea el texto mostrado en la Figura 4.2 entonces:

- “Typically” tiene asignada la posición 1
- “in” la posición 2

- "a" la posición 3
- "modern" la posición 4
- "relational" le corresponde la posición 5

Se obtiene cada término del ejemplo con un número consecutivo como valor de su posición.

Algoritmo 4.1 Algoritmo para determinar el orden de un par de términos

```
1  Obtener las posiciones correspondientes de los términos t y t1
     .
2    Para cada posición de t.
3      Para cada posición de t1.
4        Si posición de t < posición de t1.
5          sumar(+1)
6        Fin
7        Si posición de t > posición de t1.
8              sumar(-1)
9        Fin
10     Fin
11   Fin
12 Si suma >= 0
13   t es padre de t1
14 Fin
15 Si suma < 0
16   t1 es padre de t
17 Fin
```

Se tiene la Tabla 4.1 con datos de ejemplo, donde la primera columna contiene los términos, la segunda columna contiene el identificador del documento donde se encuentra dichos términos y por último la tercera columna que representa la posición del término en el documento:

Tabla 4.1: Ejemplo de posiciones de un par de términos frecuentes por documento.

término	**id_doc**	**posición1**	**posición2**	**posición3**
polici	1	3	12	50
cach	1	10	32	100
polici	2	12	23	
cach	2	11	41	

A continuación se muestra el proceso descrito en el Algoritmo 4.1 para obtener el valor del orden del ejemplo anterior.

Para el documento 1, si el término1 es considerado como “polici” y término2 es “cach” entonces el valor del orden en este documento es :

orden = 1+1+1= 3

Para el documento 2, considerando a los términos de la misma manera que en el documento 1, se obtiene el valor:

orden = -1+1=0

Por lo que al sumar los valores obtenidos en ambos documentos se obtiene el valor 3, por lo que se concluye que “polici” es padre de “cach”.

Ahora bien, si es el caso contrario, es decir, término1 es “cach” y término2 es “polici” se obtiene el valor:

orden= -1-1-1+1-1 = -3

Sin embargo en ambos casos se obtiene el mismo resultado de jerarquía: “polici” es padre de “cach”.

Si se presenta el caso mostrado en la Tabla 4.2 en el que un término tiene más ocurrencias que el otro término en un documento dados los pasos del algoritmo antes presentado se obtiene el valor:

Tabla 4.2: Ejemplo de posiciones de un par de términos frecuentes.

término	**id_doc**	**posición1**	**posición2**	**posición3**
polici	1	3	50	
cach	1	10	100	130

El valor obtenido es orden = 1 +1 + 1-1 +1+1 = 4, lo que significa que “polici” es padre de “cach”.

Ya que cada posición del primer término se compara contra cada posición del segundo término.

4.1.6. Distancia

El método se utiliza para medir el promedio de palabras que existen dados los términos *t* y *t1* y sus respectivas apariciones en los documentos. Esto se realiza para determinar si existe relación entre un par de términos y se ocupa en lugar del método de correlación y soporte y confianza en sus dos versiones que en se presentan.

Con este método se obtiene la distancia promedio de las apariciones de cada par de términos, donde se toma la definición de posición de igual manera que con el concepto de orden. El Algoritmo 4.2 muestra el procedimiento para obtener el valor de distancia promedio:

Algoritmo 4.2 Algoritmo para obtener la distancia promedio de un par de términos.

```
1  Obtener las posiciones correspondientes de los términos t
       y t1.
2    Para cada documento en el que aparezca t y t1
3          Para cada posición de t.
4             Para cada posición de t1.
5                   distancia_parcial += Distancia(t,t1)
6             fin
7          fin
8          distancia_doc_i= suma de las distancias parciales
               / (número de posiciones de t * número de
              posiciones de t1)
9    fin
10 distancia_promedio=suma distancia_doc_i / n
```

Donde n es el total de documentos del corpus, $posiciones_{doc_i}$ indica el número de posiciones del término t, $distancia_i$ corresponde a la distancia promedio que existe entre los términos t y $t1$ en el documento i.

Se muestra la Tabla 4.3 con un ejemplo de las posiciones de un par de términos.

Tabla 4.3: Ejemplo de los términos y sus correspondientes posiciones.

término	id_doc	posición1	posición2	posición3
polici	1	3	12	50
cach	1	10	32	100
polici	2	12	23	
cach	2	11	41	

Para el ejemplo anterior $distancia_{doc_1} = 25.20$ y $distancia_{doc_2} = 39.94$ por lo que la distancia promedio para los términos t y $t1$ sería 29.384.

Esto se realiza para cada par de términos frecuentes y posteriormente se filtran las combinaciones donde se obtienen solo los pares que tengan una distancia promedio menor a un cierto umbral.

4.1.7. Soporte y confianza

El soporte y confianza se utilizan muy comúnmente para definir reglas de asociación[32, 31], las cuales se utilizan entre otros en minería de datos[32]. En el

contexto de este trabajo se utilizan éstas reglas para conocer que par de términos aparecen frecuentemente en el mismo documento (donde cada documento se toma como una transacción), es decir, se utiliza para saber si dados los términos t y t1 tales que son los más frecuentes en el corpus de documentos; que tan probable es que se encuentren en los mismos documentos. Esta probabilidad indica si el par de términos deben o no estar relacionados en la taxonomía. Esto se puede conocer al calcular el soporte y la confianza, descritos a continuación.

Éstas reglas, nos permiten encontrar patrones que ocurren frecuentemente en los datos[31], los cuales indican que conjuntos de objetos ocurren frecuentemente en un conjunto de transacciones y están relacionados.

Una regla de asociación es una implicación t => t1 tal que:

- La intersección $t \cap t1$ es vacía.
- Se dice que t => t1 es verdadera si en el conjunto de transacciones D con confianza c, el c % de las transacciones contenidas en D que contienen a t contienen a t1.
- Para el caso del soporte, t => t1 tiene soporte s si en D el s % de las transacciones contienen $t \cup t1$.
- Si la confianza y el soporte de la regla de asociación esta por encima de un cierto umbral mínimo definido por el usuario esta se toma como válida.

Por ejemplo para conocer que tan frecuentemente los clientes de un supermercado que compran el producto *t* compran también el producto *t1*, y así crear estrategias de mercadotecnia.

Soporte

Sean n tuplas (en este caso los documentos se toman en como transacciones) en la base de datos y t un término frecuente del corpus de documentos, el soporte se refiere al número de documentos en los cuales aparece el término t ($docs_t$) sin importar la frecuencia que este tenga en cada documento, entre el número total de documentos del corpus. Es decir, es la fracción de documentos del corpus que contienen al término *t*.

$$Soporte_t = \frac{docs_t}{n}$$

Confianza

Sea s el soporte de t, y sea *t1* otro término frecuente del corpus, la confianza es la probabilidad de encontrar al término *t1* dados los documentos en los que se encuentra t (P($t1|t$)), es decir en que porción de los documentos se encuentra *t1* y t ($docs_{t1,t}$) dada la porción de documentos en los que se encuentra t ($docs_t$).

Dada la regla de asociación t ->*t1*, la confianza es la porción donde se puede encontrar la parte derecha de la implicación (*t1*) dada la parte izquierda (t).

$$Confianza_{t,t1} = \frac{docs_{t,t1}}{docs_t}$$

4.1.8. Soporte y confianza respecto a intervalos

El método que proponemos en este trabajo para medir la relación de los pares de términos más frecuentes es una modificación del método de soporte y confianza, en la cual cada documento es visto como un corpus y cada intervalo se toma como si fuera un documento. A continuación se describe:

Sea *ta* el tamaño del intervalo (por ejemplo *ta*=10, 100 ó 1000 palabras), sean *t* y *t1* términos frecuentes del corpus e *i*=1...n los documentos en el corpus.

El soporte de *t* es el promedio de los subsoportes obtenidos en cada documento.

El subsoporte de *t* en el documento *i* se define como el número de intervalos del documento *i* en los cuales el término *t* aparece, entre el tamaño del intervalo.

$$Subsoporte_t^{ta} = \frac{\#intervalos_t}{ta}$$

$$Soporte_{t,i}^{ta} = \frac{\sum subsoporte_t^{ta}}{ni}$$

Con *ni* igual al número total de intervalos en el documento (*m/ta* es el mínimo entero, con *m* el total de términos en el documento).

La confianza de la regla de asociación *t* -> *t1* es el promedio de las subconfianzas obtenidas de cada documento del corpus.

La subconfianza de la regla *t* -> *t1* en el documento *i*, se define como el número de intervalos de tamaño *ta* en los que aparecen los términos *t* y *t1* entre el número de intervalos en los que aparece el término *t*.

$$Subconfianza_{t,t1,i}^{ta} = \frac{\#intervalos_{t,t1}}{\#intervalos_t}$$

$$Confianza_{t,t1}^{ta} = \frac{\sum subconfianzas_{t,t1}^{ta}}{n}$$

Como proceso final, para obtener el soporte y la confianza de los términos *t* y *t1* en el corpus se calcula un promedio del soporte y confianza obtenidos en cada documento (*i*) como sigue: Sean *t* y *t1* términos del corpus y *n* el total de documentos del corpus.

$$Soporte_{t,t1} = \frac{\sum_i^n Soporte_{t,t1,i}^{ta}}{n}$$

$$Confianza_{t,t1} = \frac{\sum_i^n Confianza_{t,t1,i}^{ta}}{n}$$

4.2. Componentes del proceso de extracción de taxonomías

Para la implementación del proceso de extraer taxonomías de un conjunto de documentos almacenados en una base de datos se utilizaron los siguientes

componentes:

- Se creo una base de datos para cada corpus.
- Se construyeron cubos OLAP en cada base que contienen información acerca de los términos del corpus, como son su frecuencia en cada documento, su peso y su frecuencia total en el corpus.
- Se tienen procedimientos almacenados y funciones definidas por el usuario que realizan las siguientes tareas:
  - Separar los documentos en tokens.
  - Aplicar stemming a los tokens de los documentos.
  - Contar las frecuencias de las raíces (o lemmas en el caso de utilizar la lematización) y sus correspondientes pesos.
  - Obtener los k términos más frecuentes.
  - Aplicar el método para relacionar los pares de términos (ya sea correlación, soporte y confianza, distancia o soporte y confianza por intervalos).
  - Obtener la jerarquía de los pares de términos.

4.3. Metodología de construcción de taxonomías

Se muestra el Algoritmo 4.3 que describe a grandes rasgos la metodología que se siguió para la obtención de las taxonomías en los experimentos realizados:

Algoritmo 4.3 Algoritmo de obtención de taxonomías

```
1  Limpiar cada documento de caracteres no deseados.
2  Mientras haya documentos en el corpus.
3          Obtener la raíz de cada palabra del documento.
4          Mientras haya términos en el documento.
5                  Almacenar cada término con su posición e
                       id de documento.
6          fin
7          Calcular la frecuencia de cada palabra en el
               documento procesado.
8  fin
9  Calcular la frecuencia total de cada palabra en el corpus
       .
10 Calcular el peso de cada palabra para cada documento.
11 Calcular el peso de cada palabra para el corpus.
12 Obtener los k términos más frecuentes sin tomar en cuenta
        las stop words.
13 Para cada par de términos frecuentes que se  obtuvieron (
       t,t1).
14         Calcular el valor del método a utilizar (
               correlación, soporte y confianza, soporte y
               confianza respecto a intervalos ó distancia
               promedio).
15 Fin
16 Mientras valor_método(t,t1) > umbral_método.
17         Agregar a las relaciones (t,t1).
18 Fin
19 Para cada par de términos (t,t1) de las relaciones.
20         Calcular el valor del método (orden ó porcentaje
               de documentos) usado para obtener la jerarquía
               .          Fin
21 Crear la jerarquía de los términos.
22 FIN
```

Donde k es el número de términos frecuentes con los que el usuario desea construir la taxonomía.

Explicando el Algoritmo 4.3 para la construcción de la taxonomía se siguen los siguientes pasos:

- Dividir el texto de los documentos en términos, se utilizó el espacio blanco (“ “) como separador de los términos.
- Posteriormente se aplicó un algoritmo de stemming (Porter Stemming, Sección 4.1.1) a cada palabra de cada documento.
- Después se utilizó el concepto de “stop words” para filtrar los términos.

- Posteriormente se usaron métodos como correlación, soporte, confianza, distancia, peso, orden, soporte y confianza por intervalos para asociar cada par de términos.
- Finalmente se filtraron los pares de términos, donde se obtienen solamente en los que se obtuvo un valor del método mayor a un cierto umbral dado por el usuario.

4.3.1. Versiones de SPs y UDFs

Como se mencionó anteriormente, todos los métodos que se utilizaron para relacionar los pares de términos se implementaron con SPs y con UDFs. Para notar la diferencia entre ambas, a continuación se muestra el código de la parte con la que se calcula el método de soporte y confianza por intervalos, en ésta se calcula en que que intervalos aparecen los términos. Primero se muestra el Algoritmo 4.4 con la versión en UDF y en seguida el Algoritmo 4.5 con la versión con SP.

Algoritmo 4.4 Versión para calcular los intervalos con UDF.

```
1  CREATE FUNCTION intervalos_doc(
2          @tamanno INT,
3          @term1 VARCHAR(100)
4  ) RETURNS @intervalos TABLE (
5          num_intervalo INT,
6          doc_id INT,
7          tamanno INT,
8          term VARCHAR(100),
9          aparicion_term1 INT
10         PRIMARY KEY (num_intervalo, doc_id) )
11 AS BEGIN
12         INSERT INTO @intervalos
13         SELECT CEILING(CONVERT(FLOAT, posición)/@tamanno),
               doc_id, @tamanno, @term1,1
14         FROM onto_sp_intervalos.dbo.term_tmp
15         WHERE @term1 = onto_sp.dbo.fnPorterAlgorithm(term)
               ORDER BY posición
16         RETURN
17 END
```

Como puede observarse en el Algoritmo 4.4 es necesario crear una tabla dentro de la función, la cual es obtenida como el resultado de esta función, la cual contiene los datos de los intervalos en los que aparecen los términos y posteriormente se insertan en una tabla estos resultados, para su posterior procesamiento.

A diferencia de la versión SP, que se muestra en el Algoritmo 4.5, en el cual se insertan los datos directamente en una tabla auxiliar, pero en esencia es la

única diferencia entre ambas versiones, sin embargo en la Sección 5.12 se muestra el rendimiento de cada una.

Algoritmo 4.5 Versión para calcular los intervalos con SP.

```
1  CREATE PROCEDURE intervalos_doc_sp
2          @tamano INT,
3          @term1 VARCHAR(100)
4  AS BEGIN
5          INSERT INTO intervalos_sp
6          SELECT CEILING(CONVERT(FLOAT, posición)/@tamano),
               doc_id, @tamano, @term1, 1
7          FROM onto_sp_intervalos.dbo.term_tmp
8          WHERE @term1 = onto_sp.dbo.fnPorterAlgorithm(term)
9          ORDER BY posición
10 END
```

4.3.2. Evaluación: Métrica de palabras clave ("Keywords")

El proceso de desambiguación de un término es una tarea compleja, por lo que lo consideramos fuera del alcance de este trabajo, siendo también motivo para evaluar la calidad de los términos que componen las taxonomías construidas. Debido a ésto, se desarrolló una métrica que compara cuantos términos de la taxonomía corresponden a los términos definidos por el (los) autor(es) de cada articulo como palabras clave, de esta manera se toma en cuenta la parte del documento que tiene una supervisión por parte de un experto, esta métrica se definió de la siguiente manera:

- Se obtiene los términos que se utilizan en la taxonomía, estos se encuentran en la tabla de correlación (o la tabla correspondiente a la métrica utilizada para la obtención de la relación de cada par de términos) en donde se indican los pares de términos más frecuentes y la correlación que existe entre cada pareja.
- Se obtiene también las palabras clave de cada documento, las cuales se encuentran almacenadas en una tabla.
- Para empatar los términos de la tabla de la taxonomía y los términos de la tabla de "keywords" se le aplica el algoritmo de Porter a esta última, para obtener la raíz de las palabras clave, de lo contrario no se encontrarían ó se obtendría un falso positivo.

4.3.3. Evaluación: Métrica de palabras contenidas en el título

Como segunda medida de evaluación de calidad de los términos que componen las taxonomías obtenidas se desarrolló una métrica que compara cuantos términos de la taxonomía corresponden a los términos definidos por el(los) autor(es) de cada articulo en el título del documento, donde se toma de igual manera que en la métrica anterior la parte del documento que tiene supervisión por parte de un experto, la métrica se realizó como a continuación se describe:

- Se obtienen los términos que se utilizan en la taxonomía, estos se obtienen de la tabla de correlación o la tabla correspondiente a la métrica para obtener la relación de los pares de términos que se desea evaluar.
- Se obtiene también las palabras de los títulos de cada documento, las cuales se almacenan en una tabla.
- Para empatar los términos de la taxonomía y los términos de la tabla de títulos también se le aplica el algoritmo de Porter a esta última de la misma manera que con las palabras clave.

4.3.4. Evaluación: Ambas métricas

Como tercera medida de evaluación de calidad se compara cuantos términos de la taxonomía corresponden a los términos definidos por el(los) autor(es) de cada artículo en el título del documento y los términos definidos como palabras clave, es decir se toman los términos de las dos métricas de evaluación anteriormente mencionados.

Capítulo 5

Experimentación

En los experimentos que se muestra a continuación sólo se identifica la relación de la forma "tiene un" ("has a") como se realiza en[1], donde se explica que sean A y B términos de la taxonomía, si el término A es padre de B, entonces quiere decir que A tiene un B.

Por cada corpus se muestra una serie de experimentos, donde se construye una taxonomía con diferentes métodos para obtener la relación de los pares de términos y una serie de experimentos para medir la calidad de estas taxonomías, esto con dos diferentes métricas, (la métrica de palabras clave y la métrica de palabras contenidas en los títulos) mencionadas en la Sección 4.3.

Los diez corpus fueron obtenidos de las siguientes fuentes:

- **dl.acm.org** de donde se obtuvieron cinco corpora de diferentes temas de ciencias de la computación (los corpus 1,2,4,5,6,7).
- **www.ncbi.nlm.nih.gov/pubmed** del cual se obtuvo un corpus del área de medicina (el corpus 3).
- **http://oa-hermes.unam.mx/oa-hermes.html** de esta fuente se obtuvieron tres corpora correspondientes al área de medicina (corpus 8,9,10).

5.1. Arquitectura de referencia.

Todas las pruebas realizadas se llevaron a cabo en una computadora con las siguientes características:

- SQL Server 2012.
- Procesador Intel core i5 a 2.50 GHz.
- 6.0 GB de memoria RAM.
- Sistema operativo de Windows 7 de 64 bits.
- Java 6

5.2. Primer corpus

El primer corpus consta de 86 documentos obtenidos de la página **dl.acm.org** con la búsqueda "**query optimization**", de los cuales se procesaron 887,470 tokens, y se obtuvo la taxonomía mostrada en la Figura 5.1 con los 10 términos más frecuentes del corpus al utilizar los métodos correlación y peso de los términos.

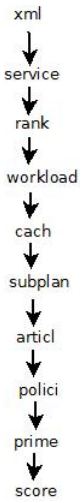

Figura 5.1: Jerarquía obtenida con los 10 términos más frecuentes.

Se hace énfasis en que solamente el término "articl" en la jerarquía de 10 términos no tiene relación con los demás términos y en la jerarquía obtenida con 5 términos no aparece, por lo que la primera jerarquía es considerada más consistente.

Se muestra en la Figura 5.2 el resultado de aplicar la operación de orden (como se mencionó en la Sección 4.1.5, si el término A aparece más veces antes que el término B en el corpus entonces A es padre de B y viceversa) en lugar del método usado en el caso anterior para determinar que término es el padre del otro[1], aún en el primer corpus.

[1]Sean los términos A y B, n1 y n2 el número de documentos en los que aparece el término A y B respectivamente. Si n1 - n2 $<$ 10 % del total de documentos del corpus entonces A es padre de B[1].

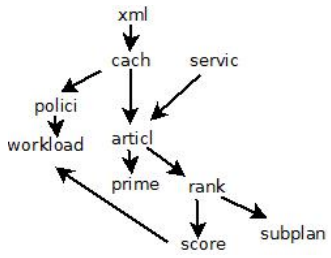

Figura 5.2: Jerarquía obtenida con los 10 términos más frecuentes.

Se muestra la Figura 5.3 con una porción de los resultados de aplicar el método de soporte y confianza a los k términos más frecuentes por pares (donde se toma a cada documento como transacción). Las columnas $term_t$ indican si es el término 1 ó 2 y las columnas $ndocs_t$ indican el número de documentos que contienen al término t.

	term1	term2	correlacion	peso	ndocs1	ndocs2	confianza
1	cach	polici	1.00	1.80	20	16	0.30
2	cach	workload	0.58	2.00	20	20	0.25
3	polici	workload	0.59	1.80	16	20	0.19
4	prime	articl	0.33	1.88	14	16	0.36
5	rank	score	0.66	1.52	21	11	0.48
6	rank	subplan	0.43	1.86	21	18	0.33
7	score	subplan	0.17	1.61	11	18	0.36
8	servic	articl	0.23	1.73	22	16	0.18
9	servic	prime	0.19	1.64	22	14	0.14
10	xml	articl	0.31	1.73	22	16	0.18

Figura 5.3: Tabla con los datos del método de soporte y confianza.

Como puede observarse en la tabla en algunos casos la correlación y la confianza varía por una cantidad pequeña, pero en otros casos es más notoria la diferencia, lo cual se puede observar en la Figura 5.4 la cual muestra la taxonomía obtenida con los datos anteriores.

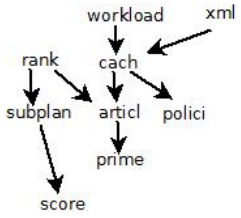

Figura 5.4: Taxonomía obtenida con el método de soporte y confianza.

Con la modificación al método de soporte y confianza (método de soporte y confianza por intervalos) se obtuvo la taxonomía mostrada en la Figura 5.5, si se

toman los intervalos con tamaño de 1000 palabras, se observa que se obtuvieron pocas relaciones entre los términos:

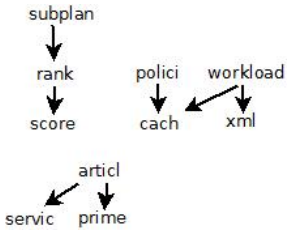

Figura 5.5: Taxonomía obtenida con el método de soporte y confianza por intervalos.

5.2.1. Métrica de palabras clave

Se muestra la Tabla 5.1 la cual contiene el porcentaje de coincidencias de palabras clave (keywords) de los documentos con los términos que conforman la taxonomía, este porcentaje depende del número de términos frecuentes y del método que se utilizó para determinar si un par de términos estaban asociados o no en la taxonomía.

Tabla 5.1: Coincidencias de términos de la taxonomía con las palabras clave.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	5.0	3.0	7.0	14.0
distancia	50.0	46.0	50.0	50.0
soporte y confianza	40.0	40.0	42.0	40.0
soporte y confianza por intervalos	40.0	43.0	42.0	38.0

Es decir, cuando se obtiene la taxonomía con los 20 términos más frecuentes y se utilizó el método de correlación; el 5 % de los términos que la conforman están definidos como “keywords” en los documentos del corpus; por otro lado cuando la taxonomía se elabora con el método de distancia: el 50 % de sus términos son “keywords”.

En la Tabla 5.1 se observa que el mayor número de coincidencias entre las palabras clave y la taxonomía en el caso donde se utilizan 20 términos frecuentes y el método de distancia.

5.2.2. Métrica de palabras contenidas en el título

En la métrica anteriormente reportada se tomaban las palabras clave definidas por el autor en cada articulo, esta vez se tomaron los términos de cada título para compararlos con los términos contenidos en la taxonomía, en la Tabla 5.2 se muestra el porcentaje de coincidencias de los términos de las taxonomías con las palabras de los títulos de los documentos.

Tabla 5.2: Correspondencia de términos del título con los términos de las taxonomías.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	15.0	10.0	10.0	12.0
distancia	50.0	43.0	45.0	46.0
soporte y confianza	40.0	36.0	37.0	36.0
soporte y confianza por intervalos	40.0	30.0	30.0	30.0

Y como puede observarse el mayor número de términos que contiene la taxonomía y que aparecen en los títulos de los artículos es cuando se utiliza el método de distancia con los 20 términos más frecuentes.

5.2.3. Resumen de evaluación primer corpus

Se muestra en la Tabla 5.3 una comparativa de la correspondencia de los términos usados en la taxonomía y las métricas de calidad utilizadas, las cuales son palabras clave, título y ambas.

Se puede observar que para este corpus se encontraron más coincidencias entre los términos de la taxonomía y los términos que contiene cada título de los documentos, esto para cada métrica utilizada para obtener las relaciones de los términos, como lo son correlación, distancia, soporte y confianza y soporte y confianza por intervalos.

5.2.4. Resultados del método de soporte y confianza por intervalos

La Tabla 5.4 muestra el tiempo que se requirió en la ejecución del proceso de obtener los intervalos de tamaño 100 en el corpus 1.

Tabla 5.3: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	**Ambas(%)**
correlación	20	5.0	15.0	15.0
distancia	20	50.0	50.0	60.0
soporte confianza	20	40.0	40.0	50.0
soporte y confianza intervalos	20	40.0	40.0	50.0

Tabla 5.4: Tiempo de ejecución del método de soporte y confianza por intervalos de tamaño 100.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
100	30	89.5
100	40	114.3
100	50	144.9

También se muestra la Tabla 5.5 con las medias de los resultados obtenidos de cambiar el valor del umbral de la confianza (con valores del umbral igual a 0.3, 0.4 y 0.5) y el número de términos utilizados para crear la taxonomía y compararlos con las palabras que el (los) autor (es) definieron como palabras clave y título del documento:

Tabla 5.5: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 100.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	100	17.0	16.0	21.0
40	100	15.6	15.3	20.0
30	100	9.0	9.0	11.6
20	100	5.6	6.3	7.6

Si se modifica el tamaño del intervalo, en esta ocasión a 500 palabras, se

muestra la Tabla 5.5 con los resultados correspondientes:

Tabla 5.6: Tiempo de ejecución del método de soporte y confianza por intervalos de tamaño 500.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
500	30	143.7
500	40	205.1
500	50	265.6

La Tabla 5.6 muestra el cambio en segundos del tiempo de ejecución de la obtención de los intervalos para el método de soporte y confianza con intervalos de tamaño 500.

Tabla 5.7: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 500.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	500	19.7	18.7	23.7
40	500	16.0	15.0	19.0
30	500	12.0	11.0	14.0
20	500	8.0	8.0	10.0

Al duplicar el tamaño del intervalo a 1000 palabras, se obtuvieron los resultados mostrados en la Tabla 5.8.

Tabla 5.8: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 1000.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	1000	19.7	18.7	23.7
40	1000	16.0	15.0	19.0
30	1000	12.0	11.0	14.0
20	1000	8.0	8.0	10.0

Así mismo se muestra en la Tabla 5.9 el tiempo de ejecución que requirió obtener los intervalos de tamaño 1000.

Tabla 5.9: Tiempo de ejecución del método de soporte y confianza por intervalos de tamaño 1000.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
1000	30	175.0
1000	40	246.1
1000	50	315.7

Si se observan las Tablas 5.7 y 5.8 se nota que no hubo diferencias en las palabras clave obtenidas al duplicar el tamaño del intervalo (1000 palabras).

5.3. Segundo corpus

El segundo corpus con el que se experimento son 80 documentos obtenidos nuevamente de la página **dl.acm.org** con la búsqueda **"semantic Web"**, esta vez, se procesaron 449,993 tokens y se obtuvo la taxonomía con los primeros 20 términos más frecuentes, con un umbral de 0.4 para filtrarlos de acuerdo a su correlación. La taxonomía obtenida es la correspondiente a la Figura 5.6:

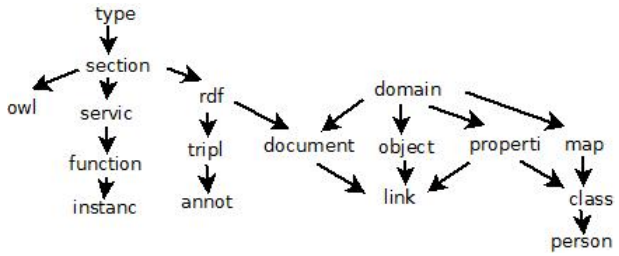

Figura 5.6: Taxonomía obtenida con los 20 términos más frecuentes del segundo corpus.

Si se cambia el criterio de clasificación para determinar cual de los términos en el conjunto de pares es padre del otro, es decir, se utiliza el criterio del orden de aparición de los términos en el corpus, (si el término A aparece más veces antes que el término B, entonces A es padre de B ó viceversa) se obtiene la taxonomía que se muestra en la Figura 5.7

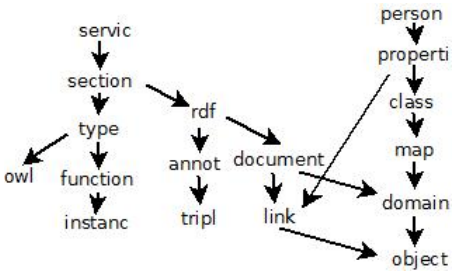

Figura 5.7: Taxonomía obtenida con la jerarquía de orden.

Si se toma en cuenta la distancia entre la aparición de cada uno de los términos como medida, donde se toman en cuenta las palabras y no los caracteres como la unidad mínima para las posiciones de los términos, se puede observas en la Figura 5.8 la taxonomía obtenida con el método de distancia:

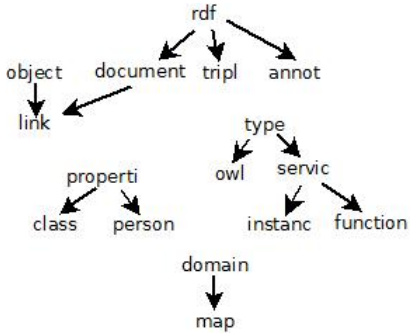

Figura 5.8: Taxonomía obtenida con el método de distancia.

De acuerdo a el método de soporte y confianza se calculó el valor correspondiente a cada par de términos del segundo corpus, y se obtuvo la taxonomía de la Figura 5.9:

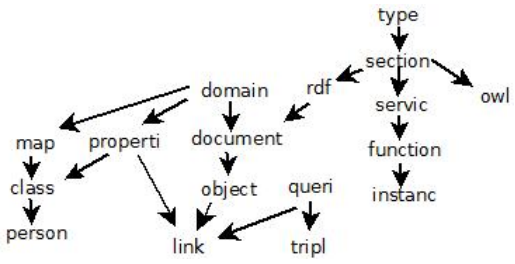

Figura 5.9: Taxonomía obtenida con el método de soporte y confianza.

Con la modificación mencionada anteriormente para el método de soporte

y confianza, pero esta vez con el segundo corpus, es decir, con el método de soporte y confianza por intervalos de tamaño 1000, se obtuvo la taxonomía que se observa en la Figura 5.10:

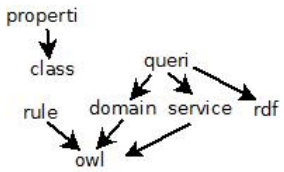

Figura 5.10: Taxonomía obtenida con el método de soporte y confianza por intervalos de tamaño 1000.

5.3.1. Métrica de palabras clave

Para observar la correspondencia de las palabras clave definidas en los documentos que conforman el segundo corpus se presenta la Tabla **??**.

Tabla 5.10: Coincidencias de términos de la taxonomía con las palabras clave.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	0.0	0.0	2.0	2.0
distancia	0.0	0.0	2.0	2.0
soporte y confianza	0.0	0.0	0.0	0.0
soporte y confianza por intervalos	2.0	4.0	4.0	

5.3.2. Métrica de palabras contenidas en el título

En la métrica anteriormente reportada se tomaban las palabras clave definidas por el autor en cada articulo, esta vez se tomaron los términos de cada título para compararlos con los términos contenidos en la taxonomía, en la Tabla 5.11 se muestra el porcentaje de coincidencias de los términos de las taxonomías con las palabras de los títulos de los documentos.

Tabla 5.11: Correspondencia de términos del título con los términos de las taxonomías.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	0.0	3.0	4.0	4.0
distancia	0.0	3.0	4.0	4.0
soporte y confianza	0.0	3.0	4.0	4.0
soporte y confianza por intervalos	0.0	3.0	5.0	12.0

En este caso se obtuvo un mayor número de coincidencias entre los términos de la taxonomía y los términos de los títulos, por lo que en algunos casos con la métrica de las palabras clave se obtiene un resultado mayor y en otros se obtiene un mejor resultado con la métrica del título.

5.3.3. Resumen de evaluación

Se muestra la Tabla 5.12 comparativa de la correspondencia de los términos usados en la taxonomía y las métricas de calidad utilizadas, las cuales son palabras clave, título y ambas.

Tabla 5.12: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	**Ambas(%)**
correlación	50	2.0	4.0	6.0
distancia	50	2.0	4.0	6.0
soporte confianza	50	0.0	4.0	4.0
soporte confianza intervalos	50	4.0	12.0	12.0

Se observa que para el segundo corpus las coincidencias entre los términos que conforman la taxonomía y las palabras clave como las palabras de los títulos son pocas con todas las métricas.

5.3.4. Resultados del método soporte y confianza por intervalos

Como se mostró con el primer corpus, se presenta la Tabla 5.13 con los tiempos de ejecución para calcular los intervalos de tamaño 100, donde se cambia el número de términos frecuentes con los que se construye la taxonomía y los resultados obtenidos de la correspondencia de palabras clave con los términos:

Tabla 5.13: Tiempo de ejecución del método de soporte y confianza por intervalos de tamaño 100.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
1000	30	10.7
1000	40	20.4
1000	50	26.5

En la Tabla 5.14 se muestran las medias de los resultados con el tamaño de intervalo igual a 100 y el umbral de la confianza con los valores 0.3, 0.4 y 0.5.

Tabla 5.14: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 100.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	100	1.7	8.2	9.2
40	100	1.7	5.7	6.2
30	100	1.0	5.0	5.0
20	100	1.0	3.0	3.0

La Tabla 5.15 contiene los tiempos de ejecución del método de soporte y confianza por intervalos de tamaño 500, mientras que la Tabla 5.16 contiene la media de los porcentajes de coincidencias de las palabras clave, las palabras contenidas en los títulos de los documentos y ambas con los términos que contiene la taxonomía.

Tabla 5.15: Tiempo de ejecución del método de soporte y confianza por intervalos de tamaño 500.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
1000	30	6.9
1000	40	12.7
1000	50	17.7

Tabla 5.16: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 500.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	500	2.0	10.7	11.7
40	500	2.0	8.7	9.7
30	500	1.0	7.7	7.7
20	500	1.0	4.0	4.0

La Tabla 5.17 muestra los tiempos de ejecución del proceso de obtener los intervalos de tamaño1000 para el segundo corpus.

Tabla 5.17: Tiempo de ejecución del método de soporte y confianza por intervalos de tamaño 1000.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
1000	30	7.2
1000	40	7.4
1000	50	12.6

Tabla 5.18: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 1000.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	1000	2.0	10.7	11.7
40	1000	2.0	8.7	9.7
30	1000	1.0	7.7	7.7
20	1000	1.0	5.0	5.0

5.4. Tercer corpus

El tercer corpus con el que se experimento cuenta con 80 abstract obtenidos de la página **www.ncbi.nlm.nih.gov/pubmed** con la búsqueda **"liver cancer**", estos abstract fueron obtenidos de la página antes mencionada en formato XML y posteriormente procesados para ser almacenados en la base de datos. Cabe mencionar que para este corpus se procesaron 13,449 tokens, esta cantidad de tokens es menor a los utilizados en los corpora anteriormente procesados.

A continuación se muestra la Figura 5.11 con los datos obtenidos al calcular la correlación entre cada par de términos, donde se puede observar que la columna "*term1*" corresponde al primer término, por lo que la columna "*term2*" corresponde al segundo término, por otro lado la columna "*docs*" corresponde al número de documentos en los que el par de términos aparecen juntos, con la correlación que indica la columna "*correlacion*" y el peso dado por la columna "*peso*".

	term1	term2	docs	correlacion	peso
1	group	control	13	0.67	1.87
2	control	compar	18	0.65	1.72
3	group	compar	18	0.53	1.83
4	metastas	resect	17	0.52	2.00
5	control	significantli	18	0.52	1.72
6	group	significantli	18	0.51	1.83
7	control	evalu	20	0.50	1.65
8	group	evalu	20	0.47	1.75
9	colorect	surviv	15	0.46	1.83
10	compar	significantli	18	0.45	2.00
11	metastas	colorect	18	0.44	1.94
12	diseas	compar	18	0.43	1.90
13	express	protein	12	0.43	1.71
14	resect	surviv	15	0.43	1.88
15	diseas	surviv	15	0.42	1.75
16	compar	evalu	20	0.40	1.90
17	metastas	surviv	15	0.40	1.88
18	resect	colorect	18	0.40	1.94
19	tumor	increas	20	0.39	1.91
20	express	tissu	19	0.38	1.89
21	model	activ	14	0.37	1.93
22	metastas	control	13	0.37	1.76
23	tumor	model	15	0.35	1.68

Figura 5.11: Datos de correlación del corpus de medicina.

Como puede observarse en la Figura 5.11 se obtienen términos como "*tumor*", "*surviv*", "*diseas*", "*alcohol*", entre otros, que a simple vista parecen términos que brindan información acerca del tema de este corpus.

Para este corpus del área de medicina se obtuvo la taxonomía mostrada en la Figura 5.12 con los primeros 20 términos más frecuentes:

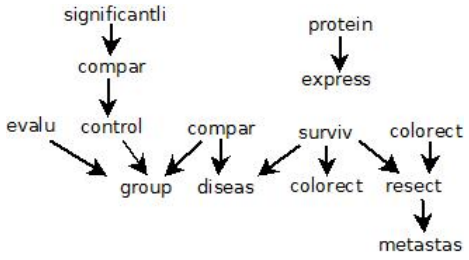

Figura 5.12: Taxonomía obtenida con el método de correlación.

5.4.1. Métrica de evaluación

En lo referente al tercer corpus no se pudo obtener la correspondencia de las palabras clave y la taxonomía debido a que el corpus solo contempla los abstracts de los artículos sin tener las palabras clave de cada uno.

5.5. Cuarto corpus

El cuarto corpus consta de 100 documentos obtenidos de la página **dl.acm.org** con la búsqueda "**data quality**", para este corpus se procesaron 635,299 tokens, de los cuales al obtener la taxonomía de los 20 términos más frecuentes, con el método de correlación para filtrar los términos relacionados entre ellos, se obtuvo la Figura 5.13:

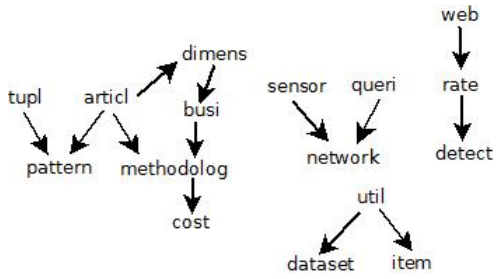

Figura 5.13: Taxonomía obtenida con los 20 términos más frecuentes y el método de correlación.

En contraste en la Figura 5.14 se puede observar que al cambiar el método utilizada para obtener la relación de jerarquía de los términos, en este caso refiriéndose a la jerarquía de orden, se obtiene una taxonomía diferente a la mostrada en la Figura 5.13.

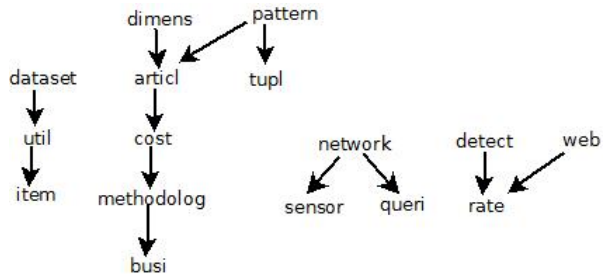

Figura 5.14: Taxonomía obtenida con los 20 términos más frecuentes y la jerarquía de orden.

En la Figura 5.15 se muestra la taxonomía obtenida al utilizar el método de soporte y confianza en lugar de la correlación para filtrar los términos relacionados, donde se obtuvieron diferentes términos, cabe mencionar que los umbrales del soporte y la confianza son 0.3 y 0.6 respectivamente, con lo cual se puede observar que para los términos en que se obtiene una correlación más alta, los valores de soporte y confianza no se mantienen de la misma manera.

	term1	term2	docs1	docs2	soporte	confianza
1	queri	web	53	59	0.53	0.69811320754717
2	dimens	softwar	57	59	0.57	0.631578947368421
3	dimens	busi	57	57	0.57	0.701754385964912
4	pattern	cost	47	58	0.47	0.659574468085106
5	pattern	rate	47	54	0.47	0.617021276595745
6	pattern	detect	47	55	0.47	0.638297872340426
7	class	cost	45	58	0.45	0.666666666666667
8	class	detect	45	55	0.45	0.6
9	articl	rate	36	54	0.36	0.638888888888889
10	articl	detect	36	55	0.36	0.638888888888889
11	cost	busi	58	57	0.58	0.672413793103448
12	cost	rate	58	54	0.58	0.637931034482759
13	cost	detect	58	55	0.58	0.655172413793104
14	busi	detect	57	55	0.57	0.614035087719298
15	util	rate	40	54	0.4	0.6
16	dataset	rate	38	54	0.38	0.657894736842105
17	dataset	detect	38	55	0.38	0.631578947368421
18	rate	detect	54	55	0.54	0.648148148148148
19	rate	item	54	47	0.54	0.62962962962963
20	detect	item	55	47	0.55	0.6

Figura 5.16: Soporte y confianza obtenida de los términos frecuentes.

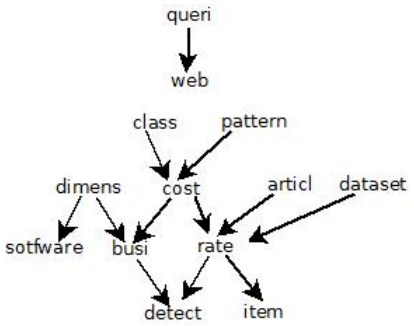

Figura 5.15: Taxonomía obtenida con el método de soporte y confianza.

Como se mencionó anteriormente, se muestra la Figura 5.16 con información del soporte y la confianza obtenidos por cada pareja de términos, donde las columnas *term1* y *term2* corresponden al primer y segundo término consecutivamente, mientras que las columnas docs1 y docs2 corresponden al número de documentos en los que aparece el primer y segundo término también respectivamente.

A continuación se muestra la Figura 5.17 con la taxonomía del corpus 4, obtenida al utilizar el método de soporte y confianza por intervalos, si se toma el tamaño de éstos como 100 palabras.

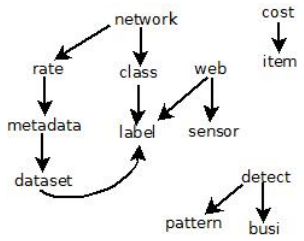

Figura 5.17: Taxonomía obtenida con el método de soporte y confianza por intervalos.

5.5.1. Métrica de palabras clave

Se muestra la Tabla 5.19 la cual contiene el porcentaje de coincidencias de palabras clave (keywords) de los documentos con los términos que conforman la taxonomía, este porcentaje depende del número de términos frecuentes y del método que se utilizó para determinar si un par de términos estaban asociados o no en la taxonomía.

Tabla 5.19: Coincidencias de términos de la taxonomía con las palabras clave.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	0.0	0.0	5.0	6.0
distancia	10.0	10.0	22.0	20.0
soporte y confianza	10.0	10.0	22.0	20.0
soporte y confianza por intervalos	10.0	10.0	17.0	16.0

5.5.2. Métrica de palabras contenidas en el título

En la métrica anteriormente reportada se tomaban las palabras clave definidas por el autor en cada articulo, esta vez se tomaron los términos de cada título para compararlos con los términos contenidos en la taxonomía, en la Tabla 5.20 se muestra el porcentaje de coincidencias de los términos de las taxonomías con las palabras de los títulos de los documentos.

Tabla 5.20: Correspondencia de términos del título con los términos de las taxonomías.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	10.0	16.0	17.0	18.0
distancia	40.0	40.0	42.0	40.0
soporte y confianza	25.0	30.0	35.0	34.0
soporte y confianza por intervalos	35.0	33.0	35.0	34.0

Se puede observar que para ambas métricas de obtención de las relaciones de los términos se encontraron más coincidencias al compararlas con las palabras que conforman los títulos de los documentos.

5.5.3. Resumen de evaluación

Se muestra la Tabla 5.21 comparativa de la correspondencia de los términos usados en la taxonomía y las métricas de calidad utilizadas, las cuales son palabras clave, título y ambas.

Tabla 5.21: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	**Ambas(%)**
correlación	50	6.0	18.0	18.0
distancia	50	20.0	40.0	44.0
soporte confianza	50	20.0	34.0	48.0
soporte confianza intervalos	50	22.0	50.0	54.0

Para el cuarto corpus puede observarse con el método de soporte y confianza por intervalos se encontró el mayor número de coincidencias con la taxonomía.

5.5.4. Resultados del método de soporte y confianza por intervalos

Se muestra a continuación una serie de tablas con los resultados de los experimentos realizados con la métrica de soporte y confianza por intervalos donde se cambian los valores del soporte, la confianza y el tamaño de los intervalos, así como la Tabla 5.22 con los tiempos de ejecución con el tamaño del intervalo igual a 100.

Tabla 5.22: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 100.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
100	30	395.4
100	40	506.3
100	50	637.1

Tabla 5.23: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 100.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	100	11.0	25.0	27.0
40	100	10.0	22.0	24.0
30	100	4.0	15.7	16.7
20	100	2.7	12.0	13.0

La Tabla 5.24 contiene los tiempos de ejecución de la métrica de soporte y confianza por intervalos de tamaño 500.

Tabla 5.24: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 500.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
500	30	387.4
500	40	523.8
500	50	657.3

Tabla 5.25: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 500.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	500	11.0	25.0	27.0
40	500	10.0	22.0	24.0
30	500	4.0	16.0	17.0
20	500	3.0	12.0	13.0

Tabla 5.26: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 1000.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
1000	30	388.2
1000	40	526.0
1000	50	673.5

Tabla 5.27: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 1000.

Términos	Intervalo	Palabras clave	Título	Ambas
50	1000	11.0	25.0	27.0
40	1000	10.0	22.0	24.0
30	1000	4.0	16.0	17.0
20	1000	3.0	12.0	13.0

5.6. Quinto corpus

El quinto corpus consta de 100 documentos obtenidos de la página **dl.acm.org** con la búsqueda "**streaming**", en el cual se tienen un total de 880,774 tokens. Y se generó la taxonomía mostrada en la Figura 5.18 al calcular la correlación entre los 20 términos más frecuentes:

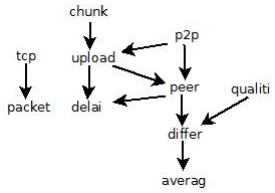

Figura 5.18: Taxonomía obtenida con el método de correlación.

Si se modifica la métrica para decidir si el término A es padre de B en una pareja de términos que se encuentran relacionados (es decir, se usa la métrica del orden) se obtiene la taxonomía correspondiente a la Figura 5.19, con el mismo umbral de correlación que en el caso anterior.

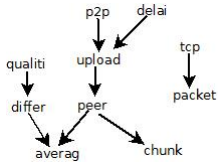

Figura 5.19: Taxonomía obtenida con los 20 términos más frecuentes y la jerarquía de orden.

En este caso se modificó la métrica con la que se mide el grado de relación en el cual se encuentran los términos, para la Figura 5.20 se utilizó la métrica

de soporte y confianza, de igual manera que con el corpus anterior se utilizaron los valores 0.3 y 0.6 como umbrales respectivamente.

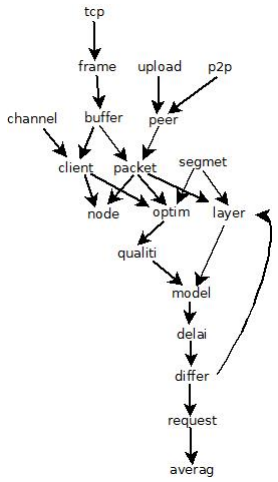

Figura 5.20: Taxonomía obtenida con el método de soporte y confianza.

La Figura 5.21 muestra la taxonomía que se obtuvo con la métrica de soporte y confianza por intervalos, en la que se nota que se obtiene un menos número de términos y relaciones que con la métrica de soporte y confianza.

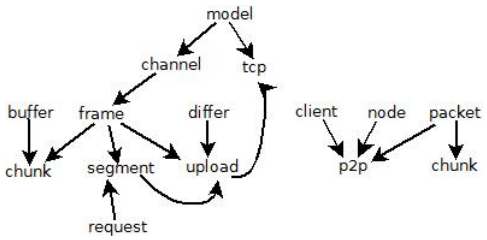

Figura 5.21: Taxonomía obtenida con el método de soporte y confianza por intervalos.

5.6.1. Métrica de palabras clave

Se muestra la Tabla 5.28 la cual contiene el porcentaje de coincidencias de palabras clave (keywords) de los documentos con los términos que conforman la taxonomía, este porcentaje depende del número de términos frecuentes y del método que se utilizó para determinar si un par de términos estaban asociados o no en la taxonomía.

Tabla 5.28: Coincidencias de términos de la taxonomía con las palabras clave.

Método	20 términos	30 términos	40 términos	50 términos
correlación	5.0	3.0	2.0	4.0
distancia	10.0	6.0	5.0	10.0
soporte y confianza	55.0	56.0	55.0	54.0
soporte y confianza por intervalos	50.0	53.0	50.0	48.0

5.6.2. Métrica de palabras contenidas en el título

En la métrica anteriormente reportada se tomaban las palabras clave definidas por el autor en cada articulo, esta vez se tomaron los términos de cada título para compararlos con los términos contenidos en la taxonomía, en la Tabla 5.29 se muestra el porcentaje de coincidencias de los términos de las taxonomías con las palabras de los títulos de los documentos.

Tabla 5.29: Correspondencia de términos del título con los términos de las taxonomías.

Método	20 términos	30 términos	40 términos	50 términos
correlación	5.0	3.0	5.0	4.0
distancia	10.0	6.0	7.0	10.0
soporte y confianza	60.0	60.0	62.0	54.0
soporte y confianza por intervalos	45.0	50.0	47.0	42.0

5.6.3. Resumen de evaluación

Se muestra la Tabla 5.30 comparativa de la correspondencia de los términos usados en la taxonomía y las métricas de calidad utilizadas, las cuales son palabras clave, título y ambas.

Tabla 5.30: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	jajajaj **Ambas(%)**
correlación	20	5.0	5.0	10.0
distancia	20	10.0	10.0	10.0
soporte confianza	20	55.0	60.0	70.0
soporte confianza intervalos	20	50.0	55.0	65.0

5.6.4. Resultados del método de soporte y confianza por intervalos

Se muestra a continuación la Tabla 5.32 con los resultados de los experimentos realizados con la métrica de soporte y confianza por intervalos donde se cambian los valores del soporte, la confianza, así como la Tabla 5.31 con los tiempos de ejecución con intervalos de tamaño 100.

Tabla 5.31: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 100.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
100	30	178.1
100	40	283.2
100	50	378.9

Tabla 5.32: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 100.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	100	23.0	23.0	30 .0
40	100	19.0	21.0	25.0
30	100	14.0	16.0	18.0
20	100	8.2	10.0	12.0

Se muestran las Tablas 5.33 y 5.34 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 500.

Tabla 5.33: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 500.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
500	30	265.5
500	40	470.5
500	50	631.8

Tabla 5.34: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 500.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	500	24.0	24.0	31.0
40	500	20.0	22.0	26.0
30	500	16.0	17.0	20.0
20	500	10.0	11.0	13.0

Se muestran las Tablas 5.35 y 5.36 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 500.

Tabla 5.35: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 1000.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
1000	30	319.1
1000	40	562.8
1000	50	734.2

Tabla 5.36: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 1000.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	1000	24.0	24.0	31.0
40	1000	20.0	22.0	26.0
30	1000	16.0	17.0	20.0
20	1000	10.0	11.0	13.0

5.7. Sexto corpus

El sexto corpus consta de 86 documentos obtenidos de la página **dl.acm.org** con la búsqueda "**face detection**", para este corpus se procesaron 392,566 tokens. De lo anterior se obtuvo la siguiente taxonomía con los primeros 20 términos más frecuentes y la métrica de correlación.

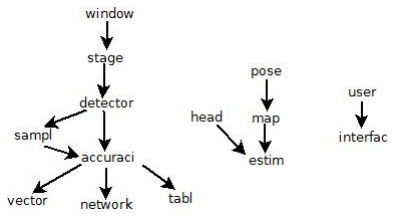

Figura 5.22: Taxonomía obtenida con los 20 términos más frecuentes y el método de correlación.

Al modificar la manera de decidir que término de una pareja es el padre del otro, es decir, al utilizar la jerarquía de orden con los datos obtenidos con la

correlación, se obtuvo la taxonomía de la Figura 5.23:

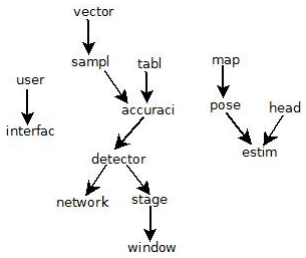

Figura 5.23: Taxonomía obtenida con los 20 términos más frecuentes y la jerarquía de orden.

En la cual se observa que en la mayoría de las parejas de términos no se mantuvo la relación donde A es padre de B, es decir, en la mayoría de los casos se encontró que B es padre de A, sin embargo pocas relaciones se mantuvieron, como es el caso de los términos "*user*" e "*interfac*".

En la Figura 5.24 se muestra una taxonomía en la que se puede observar un mayor número de relaciones entre los términos al utilizar la métrica de soporte y confianza.

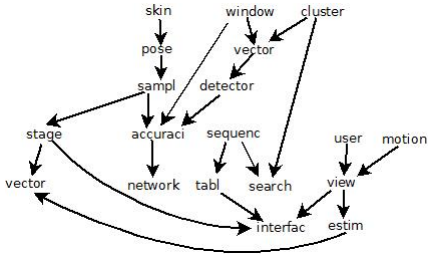

Figura 5.24: Taxonomía obtenida con los 20 términos más frecuentes y el método de soporte y confianza.

Por otro lado, se muestra en la Figura 5.25 la taxonomía que se obtuvo con la métrica de soporte y confianza por intervalos, en la cual se observan un número menor de relaciones entre los términos, así como éstos mismos.

5.7.1. Métrica de palabras clave.

Se muestra la Tabla 5.37 la cual contiene el porcentaje de coincidencias de palabras clave (keywords) de los documentos con los términos que conforman la taxonomía, este porcentaje depende del número de términos frecuentes y del método que se utilizó para determinar si un par de términos estaban asociados o no en la taxonomía.

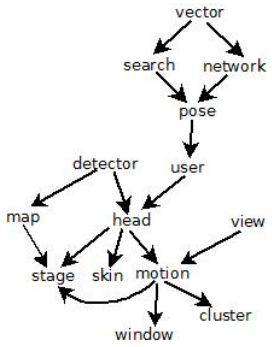

Figura 5.25: Taxonomía obtenida con el método de soporte y confianza por intervalos.

Tabla 5.37: Coincidencias de términos de la taxonomía con las palabras clave.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	5.0	3.0	10.0	12.0
distancia	10.0	6.0	5.0	10.0
soporte y confianza	30.0	26.0	27.0	26.0
soporte y confianza por intervalos	35.0	30.0	30.0	28.0

Del cuadro anterior se puede concluir que para el sexto corpus se obtuvo un mayor número de palabras clave que conforman la taxonomía con la métrica de correlación.

5.7.2. Métrica de palabras contenidas en el título

En la métrica anteriormente reportada se tomaban las palabras clave definidas por el autor en cada articulo, esta vez se tomaron los términos de cada título para compararlos con los términos contenidos en la taxonomía, en la Tabla 5.38 se muestra el porcentaje de coincidencias de los términos de las taxonomías con las palabras de los títulos de los documentos.

Tabla 5.38: Correspondencia de términos del título con los términos de las taxonomías.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	10.0	10.0	15.0	14.0
distancia	10.0	16.0	12.0	12.0
soporte y confianza	35.0	33.0	27.0	24.0
soporte y confianza por intervalos	45.0	40.0	37.0	36.0

5.7.3. Resumen de evaluación

Se muestra la Tabla 5.39 comparativa de la correspondencia de los términos usados en la taxonomía y las métricas de calidad utilizadas, éstas son palabras clave, título y ambas.

Tabla 5.39: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Título(%)**	**Palabras clave(%)**	**Ambas(%)**
correlación	50	12.0	14.0	20.0
distancia	50	18.0	12.0	22.0
soporte confianza	50	26.0	24.0	56.0
soporte confianza intervalos	50	28.0	38.0	52.0

Para el sexto corpus se encontró el mayor número de coincidencias con la métrica de soporte y confianza por intervalos y las palabras contenidas en los títulos de los documentos.

5.7.4. Resultados del método de soporte y confianza por intervalos

Se muestran las Tablas 5.40 y 5.41 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 100.

Tabla 5.40: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 100.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
100	30	21.5
100	40	36.5
100	50	45.9

Tabla 5.41: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 100.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	100	14.0	19.0	26.0
40	100	12.0	16.0	21.0
30	100	9.0	13.0	16.0
20	100	7.0	10.0	12.0

Se muestran las Tablas 5.42 y 5.43 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 500.

Tabla 5.42: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 500.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
500	30	41.0
500	40	71.0
500	50	93.9

Tabla 5.43: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 500.

Términos	Intervalo	Palabras clave	Título	Ambas
50	500	14.0	19.0	26.0
40	500	12.0	16.0	21.0
30	500	9.0	13.0	16.0
20	500	7.0	10.0	12.0

Se muestran las Tablas 5.44 y 5.45 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 1000.

Tabla 5.44: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 1000.

Tamaño de intervalo	Número de términos	Tiempo (sg)
1000	30	50.4
1000	40	86.5
1000	50	114.2

Tabla 5.45: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 1000.

Términos	Intervalo	Palabras clave	Título	Ambas
50	1000	14.0	19.0	26.0
40	1000	12.0	16.0	21.0
30	1000	9.0	13.0	16.0
20	1000	7.0	10.0	12.0

5.8. Séptimo corpus

Este corpus consta de 84 documentos obtenidos de la página **dl.acm.org** con la búsqueda **"data encryption"**, para el cual se procesaron 668,114 tokens.

La primera Figura de este corpus (Figura 5.26) corresponde a la taxonomía que se obtiene de los 20 términos más frecuentes y la métrica de correlación, se observa que en la mayoría de los términos solo están relacionados con otro término.

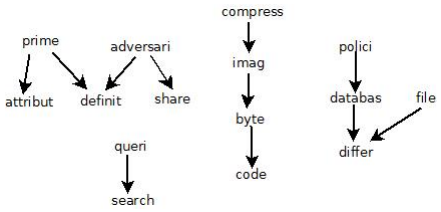

Figura 5.26: Taxonomía obtenida con la métrica de correlación.

En la Figura 5.27 se muestra la taxonomía que al igual que en el caso anterior, se utilizaron los 20 términos más frecuentes con la métrica de correlación, pero al utilizar la jerarquía de orden para obtener la jerarquía de cada par de términos, se puede observar que algunas relaciones se mantuvieron, mientras que otras resultaron de manera inversa.

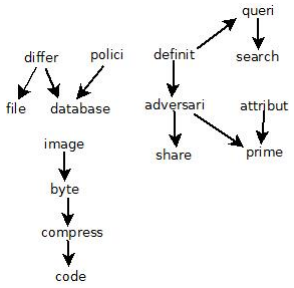

Figura 5.27: Taxonomía obtenida con la jerarquía de orden.

Ahora se muestra la Figura 5.28 con la taxonomía obtenida de aplicarle la métrica de soporte y confianza a los primeros 20 términos más frecuentes, se observa un mayor número de relaciones entre los términos que cuando se utilizó la métrica de correlación para este corpus.

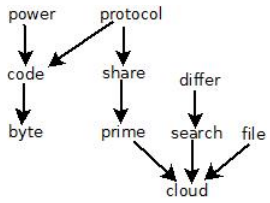

Figura 5.29: Taxonomía obtenida con la métrica de soporte y confianza por intervalos.

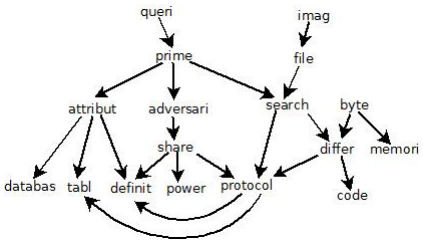

Figura 5.28: Taxonomía obtenida con la métrica de soporte y confianza.

Al utilizar la modificación a la métrica de soporte y confianza, es decir, si se usa el soporte y confianza por intervalos se obtuvo la taxonomía de la Figura 5.29 con un número de términos y relaciones menor que con las demás métricas.

5.8.1. Métrica de palabras clave

Se muestra la Tabla 5.46 la cual contiene el porcentaje de coincidencias de palabras clave (keywords) de los documentos con los términos que conforman la taxonomía, este porcentaje depende del número de términos frecuentes y del método que se utilizó para determinar si un par de términos estaban asociados o no en la taxonomía.

Tabla 5.46: Coincidencias de términos de la taxonomía con las palabras clave.

Método	20 términos	30 términos	40 términos	50 términos
correlación	5.0	6.0	7.0	12.0
distancia	50.0	43.0	42.0	46.0
soporte y confianza	65.0	43.0	40.0	42.0
soporte y confianza por intervalos	35.0	30.0	27.0	32.0

5.8.2. Métrica de palabras contenidas en el título

En la métrica anteriormente reportada se tomaban las palabras clave definidas por el autor en cada articulo, esta vez se tomaron los términos de cada título para compararlos con los términos contenidos en la taxonomía, en la Tabla 5.47 se muestra el porcentaje de coincidencias de los términos de las taxonomías con las palabras de los títulos de los documentos.

Tabla 5.47: Correspondencia de términos del título con los términos de las taxonomías.

Método	20 términos	30 términos	40 términos	50 términos
correlación	5.0	10.0	10.0	14.0
distancia	60.0	53.0	47.0	48.0
soporte y confianza	80.0	53.0	45.0	44.0
soporte y confianza por intervalos	35.0	33.0	30.0	32.0

5.8.3. Resumen de evaluación

Se muestra la Tabla 5.48 comparativa de la correspondencia de los términos usados en la taxonomía y las métricas de calidad utilizadas, las cuales son palabras clave, título y ambas.

Tabla 5.48: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Título(%)**	**Palabras clave(%)**	**Ambas(%)**
correlación	50	12.0	14.0	14.0
distancia	50	46.0	48.0	56.0
soporte confianza	50	42.0	44.0	58.0
soporte confianza intervalos	50	40.0	32.0	46.0

5.8.4. Resultados del método de soporte y confianza por intervalos

Se muestra a continuación una serie de tablas con los resultados de los experimentos realizados con la métrica de soporte y confianza por intervalos donde se cambian los valores del soporte, la confianza y el tamaño de los intervalos.

Tabla 5.49: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 100.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
100	30	25.9
100	40	40.5
100	50	64.1

Tabla 5.50: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 100.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	100	15.0	16.0	19.0
40	100	10.0	12.0	13.0
30	100	8.0	11.0	11.0
20	100	7.0	8.0	8.0

Se muestran las Tablas 5.51 y 5.52 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 500.

Tabla 5.51: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 500.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
500	30	37.4
500	40	59.8
500	50	103.4

Tabla 5.52: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 500.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	500	16.0	17.0	20.0
40	500	11.0	13.0	14.0
30	500	9.0	11.0	12.0
20	500	7.0	8.0	8.0

Se muestran las Tablas 5.53 y 5.54 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 1000.

Tabla 5.53: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 1000.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
1000	30	45.8
1000	40	72.7
1000	50	123.6

Tabla 5.54: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 1000.

Términos	Intervalo	Palabras clave	Título	Ambas
50	1000	16.0	17.0	20.0
40	1000	11.0	13.0	14.0
30	1000	9.0	11.0	12.0
20	1000	7.0	8.0	8.0

5.9. Octavo corpus

Este corpus consta de 76 documentos obtenidos de la página **http://oa-hermes.unam.mx/oa-hermes.html** con la búsqueda "**cancer breast**". Para este corpus se procesaron 512,791 tokens.

Al utilizar la métrica de correlación, con umbral de 0.4 para filtrar las relaciones de los términos, se obtuvo la taxonomía mostrada en la Figura 5.30

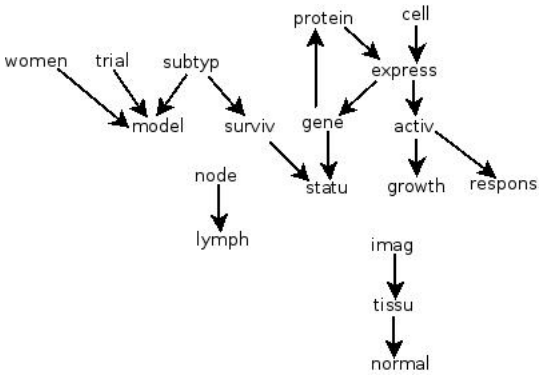

Figura 5.30: Taxonomía obtenida con el método de correlación.

La taxonomía de la Figura 5.31 se obtuvo al utilizar la métrica de correlación, como en el caso anterior, pero esta vez se utilizó también la jerarquía de orden (Sección 4.1.5) con lo cual se observa que las relaciones de jerarquía entre los pares de términos cambio en la mayoría de los casos.

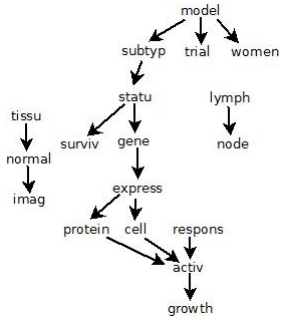

Figura 5.31: Taxonomía obtenida con el método de correlación y orden.

Esta vez al utilizar la métrica de soporte y confianza con sus respectivos umbrales para filtrar los términos, se obtuvo la taxonomía mostrada en la Figura 5.32, la cual se puede apreciar más disminuida a comparación de la taxonomía de la Figura 5.31.

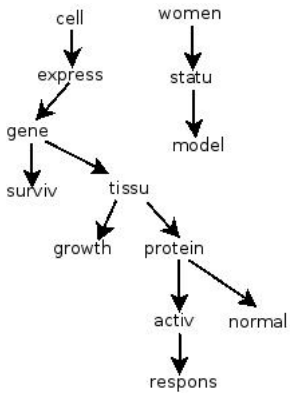

Figura 5.32: Taxonomía obtenida con el método de soporte y confianza.

Como en los corpus anteriores, al utilizar la métrica de soporte y confianza por intervalos, mostrada en la Figura 5.31 se obtiene una taxonomía con un número de términos notablemente menor que con las métricas mencionadas anteriormente.

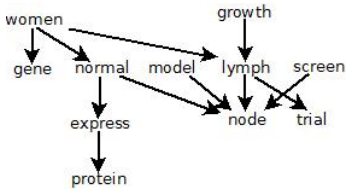

Figura 5.33: Taxonomía obtenida con el método de soporte y confianza por intervalos.

5.9.1. Métrica de palabras clave

Se muestra la Tabla 5.55 la cual contiene el porcentaje de coincidencias de palabras clave (keywords) de los documentos con los términos que conforman la taxonomía, este porcentaje depende del número de términos frecuentes y del método que se utilizó para determinar si un par de términos estaban asociados o no en la taxonomía.

Tabla 5.55: Coincidencias de términos de la taxonomía con las palabras clave.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	10.0	6.0	5.0	8.0
distancia	45.0	33.0	27.0	30.0
soporte y confianza	40.0	30.0	25.0	28.0
soporte y confianza por intervalos	45.0	33.0	25.0	26.0

Como puede observarse en la mayoría de los casos presentados por la métrica de soporte y confianza por intervalos se encontraron 9 coincidencias de palabras clave, las cuales corresponden al 45 % de la taxonomía.

5.9.2. Métrica de palabras contenidas en el título

En la métrica anteriormente reportada se tomaban las palabras clave definidas por el autor en cada articulo, esta vez se tomaron los términos de cada título para compararlos con los términos contenidos en la taxonomía, en la Tabla 5.56 se muestra el porcentaje de coincidencias de los términos de las taxonomías con las palabras de los títulos de los documentos.

Tabla 5.56: Correspondencia de términos del título con los términos de las taxonomías.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	5.0	3.0	10.0	12.0
distancia	55.0	50.0	50.0	52.0
soporte y confianza	50.0	46.0	45.0	48.0
soporte y confianza por intervalos	40.0	36.0	35.0	34.0

5.9.3. Resumen de evaluación

Se muestra la Tabla 5.57 comparativa de la correspondencia de los términos usados en la taxonomía y las métricas de calidad utilizadas, las cuales son palabras clave, título y ambas.

Tabla 5.57: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Título(%)**	**Palabras clave(%)**	**Ambas(%)**
correlación	20	10.0	5.0	10.0
distancia	20	45.0	55.0	65.0
soporte confianza	20	40.0	50.0	80.0
soporte confianza intervalos	20	45.0	50.0	60.0

5.9.4. Resultados del método de soporte y confianza por intervalos

Se muestra a continuación una serie de tablas con los resultados de los experimentos realizados con la métrica de soporte y confianza por intervalos donde se cambian los valores del soporte, la confianza y el tamaño de los intervalos.

Tabla 5.58: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 100.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
100	30	51.1
100	40	70.4
100	50	101.7

Tabla 5.59: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 100.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	100	11.7	18.7	22.7
40	100	9.2	15.2	18.2
30	100	9.7	12.7	15.7
20	100	8.2	8.7	10.7

Se muestran las Tablas 5.60 y 5.61 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 500.

Tabla 5.60: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 500.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
500	30	89.0
500	40	124.9
500	50	184.0

Tabla 5.61: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 500.

Términos	Intervalo	Palabras clave	Título	Ambas
50	500	13.0	20.0	24.0
40	500	10.0	16.2	19.2
30	500	10.0	13.2	16.2
20	500	9.0	10.0	12.0

Se muestran las Tablas 5.62 y 5.63 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 500.

Tabla 5.62: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 1000.

Tamaño de intervalo	Número de términos	Tiempo (sg)
1000	30	103.6
1000	40	149.0
1000	50	220.7

Tabla 5.63: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 1000.

Términos	Intervalo	Palabras clave	Título	Ambas
50	1000	13.0	20.0	24.0
40	1000	10.0	17.0	20.0
30	1000	10.0	14.0	17.0
20	1000	9.0	10.0	12.0

5.10. Noveno corpus

Este corpus consta de 100 documentos obtenidos de la página **http://oa-hermes.unam.mx/oa-hermes.html** con la búsqueda "**HIV**"; para este cor-

pus se procesaron 634,366 tokens.

La Figura 5.34 muestra la taxonomía obtenida con la métrica de correlación utilizada para obtener la relación entre los términos, con los 20 términos más frecuentes.

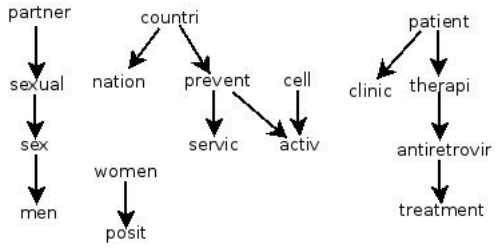

Figura 5.34: Taxonomía obtenida con el método de correlación.

En la Figura 5.35 se muestra que al cambiar la manera de obtener la jerarquía entre cada par de términos, es decir, al utilizar la jerarquía de orden, se obtiene una taxonomía muy diferente a la anterior, pero de igual manera que en otros corpus existen relaciones que se mantienen, aunque son minoría.

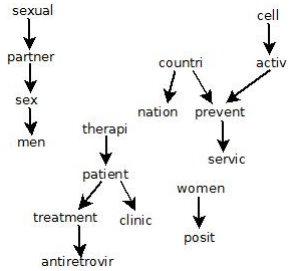

Figura 5.35: Taxonomía obtenida con la métrica jerarquía de orden.

La métrica de soporte y confianza aplicada al noveno corpus se muestra en la Figura 5.36, en la cual se puede observar que se encontraron más relaciones entre los términos que la conforman que al utilizar la métrica de correlación.

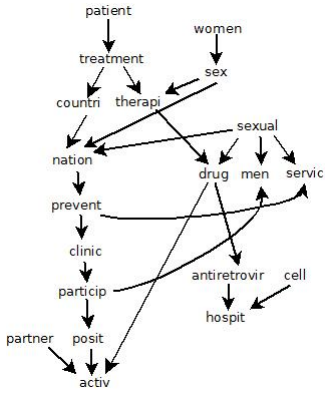

Figura 5.36: Taxonomía obtenida con el método de soporte y confianza.

Como en corpus anteriores se observó, al utilizar la métrica de soporte y confianza por intervalos se obtuvo una taxonomía con un número menor de términos, así como relaciones entre ellos, como se observa en la Figura 5.37.

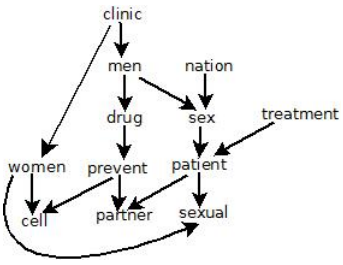

Figura 5.37: Taxonomía obtenida con el método de soporte y confianza por intervalos.

5.10.1. Métrica de palabras clave

Se muestra la Tabla 5.64 la cual contiene el porcentaje de coincidencias de palabras clave (keywords) de los documentos con los términos que conforman la taxonomía, este porcentaje depende del número de términos frecuentes y del método que se utilizó para determinar si un par de términos estaban asociados o no en la taxonomía.

Tabla 5.64: Coincidencias de términos de la taxonomía con las palabras clave.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	15.0	13.0	10.0	12.0
distancia	45.0	33.0	35.0	34.0
soporte y confianza	50.0	40.0	40.0	38.0
soporte y confianza por intervalos	35.0	26.0	27.0	30.0

Se puede observar en la Tabla 5.64 que el mayor número de coincidencias es en el caso del método de soporte y confianza con los 20 términos más frecuentes.

5.10.2. Métrica de palabras contenidas en el título

En la métrica anteriormente reportada se tomaban las palabras clave definidas por el autor en cada articulo, esta vez se tomaron los términos de cada título para compararlos con los términos contenidos en la taxonomía, en la Tabla 5.65 se muestra el porcentaje de coincidencias de los términos de las taxonomías con las palabras de los títulos de los documentos.

Tabla 5.65: Correspondencia de términos del título con los términos de las taxonomías.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	30.0	26.0	20.0	20.0
distancia	85.0	80.0	80.0	74.0
soporte y confianza	85.0	80.0	80.0	74.0
soporte y confianza por intervalos	45.0	43.0	47.0	48.0

5.10.3. Resumen de evaluación

Se muestra la Tabla 5.66 comparativa de la correspondencia de los términos usados en la taxonomía y las métricas de calidad utilizadas, las cuales son palabras clave, título y ambas.

Tabla 5.66: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Título(%)**	**Palabras clave(%)**	**Ambas(%)**
correlación	50	18.0	27.0	30.0
distancia	50	0.0	2.0	2.0
soporte confianza	50	34.0	74.0	82.0
soporte confianza intervalos	50	30.0	58.0	60.0

5.10.4. Resultados del método de soporte y confianza por intervalos

Se muestra a continuación una serie de tablas con los resultados de los experimentos realizados con la métrica de soporte y confianza por intervalos donde se cambian los valores del soporte, la confianza y el tamaño de los intervalos.

Tabla 5.67: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 100.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
100	30	52.8
100	40	87.8
100	50	195.2

Tabla 5.68: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 100.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	100	12.7	26.7	27.7
40	100	9.5	21.7	21.7
30	100	8.0	17.0	17.0
20	100	7.0	12.0	12.0

Se muestran las Tablas 5.69 y 5.70 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 500.

Tabla 5.69: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 500.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
500	30	84.5
500	40	151.2
500	50	359.2

Tabla 5.70: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 500.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	500	15.0	29.0	30.0
40	500	12.0	24.0	24.0
30	500	8.2	17.2	17.2
20	500	8.2	14.2	14.2

Se muestran las Tablas 5.71 y 5.72 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 1000.

Tabla 5.71: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 1000.

Tamaño de intervalo	Número de términos	Tiempo (sg)
1000	30	97.0
1000	40	184.8
1000	50	449.4

Tabla 5.72: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 1000.

Términos	Intervalo	Palabras clave	Título	Ambas
50	1000	15.0	29.0	30.0
40	1000	12.0	24.0	24.0
30	1000	9.0	18.0	18.0
20	1000	8.0	14.0	14.0

5.11. Décimo corpus

Este corpus consta de 96 documentos obtenidos de la página **http://oa-hermes.unam.mx/oa-hermes.html** con la búsqueda "**multiform gliobastoma**"; para este corpus se procesaron 669,939 tokens.

La Figura 5.38 corresponde a la taxonomía obtenida con los 20 términos principales, filtrándolos con la métrica de correlación.

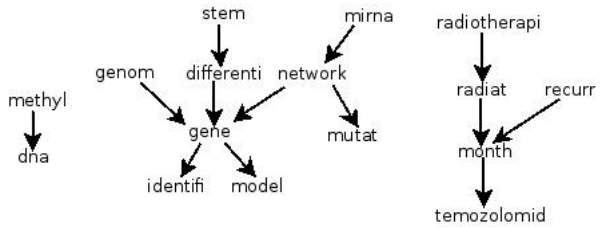

Figura 5.38: Taxonomía obtenida con el método de correlación.

La Figura 5.39 muestra la taxonomía obtenida al utilizar nuevamente la

métrica de correlación, pero esta vez se cambia la manera de obtener la jerarquía de los términos, en esta ocasión se utilizó la jerarquía de orden (Sección 4.1.5).

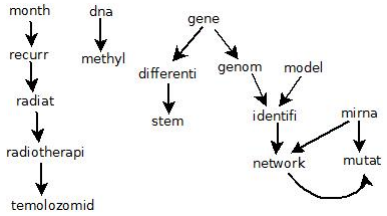

Figura 5.39: Taxonomía obtenida con el método de correlación y orden.

Como puede observarse, se obtienen los mismos términos que en el caso anterior, con la diferencia que en la mayoría de los casos no se mantuvo la relación A es padre de B, encontrándose que la mayoría cumple que B es padre de A.

Para la Figura 5.40 se muestra la taxonomía obtenida con la métrica de soporte y confianza, de igual manera se procesaron los 20 términos más frecuentes y con mayor peso en el corpus.

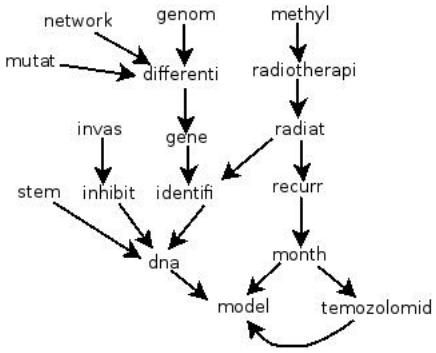

Figura 5.40: Taxonomía obtenida con el método de soporte y confianza.

Al utilizar la métrica de soporte y confianza por intervalos, con un umbral para la confianza de 0.4 y 100 palabras como el tamaño del intervalo, se obtiene la taxonomía mostrada en la Figura 5.41, en la cual se observa que es de menor tamaño que las taxonomías obtenidas con las métricas anteriores.

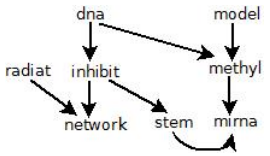

Figura 5.41: Taxonomía obtenida con el método de soporte y confianza por intervalos.

5.11.1. Métrica de palabras clave

Se muestra la Tabla 5.73 la cual contiene el porcentaje de coincidencias de palabras clave (keywords) de los documentos con los términos que conforman la taxonomía, este porcentaje depende del número de términos frecuentes y del método que se utilizó para determinar si un par de términos estaban asociados o no en la taxonomía.

Tabla 5.73: Coincidencias de términos de la taxonomía con las palabras clave.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	10.0	10.0	7.0	6.0
distancia	25.0	23.0	20.0	16.0
soporte y confianza	50.0	46.0	37.0	34.0
soporte y confianza por intervalos	10.0	12.0	13.0	14.0

5.11.2. Métrica de palabras contenidas en el título

Para el último corpus se presenta la Tabla 5.2, en ésta se muestra el porcentaje de coincidencias de los términos de las taxonomías con las palabras de los títulos de los documentos.

Tabla 5.74: Correspondencia de términos del título con los términos de las taxonomías.

Método	**20 términos**	**30 términos**	**40 términos**	**50 términos**
correlación	5.0	10.0	10.0	10.0
distancia	35.0	36.0	35.0	34.0
soporte y confianza	55.0	60.0	52.0	52.0
soporte y confianza por intervalos	30.0	26.0	27.0	30.0

5.11.3. Resumen de evaluación

Se muestra la Tabla 5.75 comparativa de la correspondencia de los términos usados en la taxonomía y las métricas de calidad utilizadas, las cuales son palabras clave, título y ambas.

Tabla 5.75: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Título(%)**	**Palabras clave(%)**	**Ambas(%)**
correlación	20	10.0	8.0	10.0
distancia	20	25.0	35.0	45.0
soporte confianza	20	50.0	55.0	75.0
soporte confianza intervalos	20	25.0	20.0	40.0

5.11.4. Resultados del método de soporte y confianza por intervalos

Se muestra a continuación una serie de tablas con los resultados de los experimentos realizados con la métrica de soporte y confianza por intervalos donde se cambian los valores del soporte, la confianza y el tamaño de los intervalos.

Tabla 5.76: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 100.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
100	30	16.6
100	40	27.7
100	50	53.6

Tabla 5.77: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 100.

Términos	**Intervalo**	**Palabras clave**	**Título**	**Ambas**
50	100	12.0	17.5	21.5
40	100	10.0	13.5	17.5
30	100	9.2	9.7	12.7
20	100	5.25	6.7	8.7

Se muestran las Tablas 5.78 y 5.79 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 500.

Tabla 5.78: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 100.

Tamaño de intervalo	**Número de términos**	**Tiempo (sg)**
500	30	27.8
500	40	51.3
500	50	91.2

Tabla 5.79: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 500.

Términos	Intervalo	Palabras clave	Título	Ambas
50	500	14.0	19.0	24.0
40	500	12.0	15.0	19.0
30	500	10.7	12.0	14.7
20	500	6.7	8.7	11

Se muestran las Tablas 5.80 y 5.81 con los tiempos de ejecución y los resultados obtenidos con los intervalos de tamaño 1000.

Tabla 5.80: Tiempo de ejecución de la métrica de soporte y confianza por intervalos de tamaño 100.

Tamaño de intervalo	Número de términos	Tiempo (sg)
500	30	43.4
500	40	74.0
500	50	115.8

Tabla 5.81: Correspondencia de keywords-título-taxonomía con intervalos de tamaño 1000.

Términos	Intervalo	Palabras clave	Título	Ambas
50	1000	14.0	20.0	25.0
40	1000	12.0	16.0	20.0
30	1000	11.0	13.0	16.0
20	1000	8.0	9.0	12.0

5.12. Discusión de resultados

En la presente sección se muestra un resumen de los resultados separándolos en dos partes.

- En la primera sección se resume el desempeño de ambas versiones (SP y UDF) con las que se extrajeron las taxonomías. Esto con la etapa de pre-procesamiento de los documentos contenidos en los corpus y la etapa en la que se utiliza cada método para la obtención de la relación de un par de términos.
- En la segunda sección se muestra un resumen acerca de la calidad de las taxonomías obtenidas, de acuerdo a las métricas de calidad aportadas en este trabajo (secciones 4.3.2 y 4.3.3) y los experimentos de los métodos utilizados para la obtención de las taxonomías.

Recordemos que los métodos presentados para la elaboración de la taxonomía son:

- Correlación: como fue definida en 4.1.3 indica que tan seguido aparecen dos términos en los documentos de un corpus y que tan seguido aparece uno sin el otro.
- Distancia: este método definido en 4.1.6 indica si los términos *t* y *t1* tienen una media de palabras que aparecen entre ambos términos menor a un cierto umbral, de ser así los términos *t* y *t1* se relacionan, en caso contrario no se toman en cuenta como una relación para la taxonomía.
- Soporte y confianza: la definición de este método se encuentra en 4.1.7 donde se menciona que se utiliza comúnmente para obtener reglas de asociación, particularmente en este trabajo, se utiliza para obtener el soporte y confianza de las reglas *t* -> *t1* donde *t, t1* son términos del corpus y la regla representa dado el término *t* que tan frecuente aparece el término *t1* en los documentos del corpus.
- Soporte y confianza por intervalos: la aportación de este método se presenta en la sección 4.1.8 la cual es una modificación al método de soporte y confianza, es decir, los documentos se parten en intervalos de un cierto tamaño (el cual puede variar para fines de la experimentación) y si se toman *t* y *t1* como términos del corpus se calcula el soporte y la confianza de la regla de asociación *t* -> *t1*, de cada intervalo de cada documento del corpus.

Por otro lado, si recordamos las métricas de evaluación de la calidad de las taxonomías donde se toman en cuenta la parte de los documentos que tienen supervisión humana, se presentaron:

- Métrica de palabras clave: la métrica se definió en 4.3.2 donde se menciona que se comparan las palabras clave definidas por los autores de los documentos como "keywords" con los términos contenidos en la taxonomía obtenida de cada corpus.
- Métrica de palabras contenidas en el título: esta métrica definida en 4.3.3 compara las palabras que aparecen en los títulos de cada documento con los términos que conforman la taxonomía obtenida de cada corpus.

- Ambas métricas: Sección 4.3.4, se toman las palabras definidas como palabras clave y las palabras contenidas en los títulos de los documentos y se compara con los términos que conforman las taxonomías definidas en cada corpus.

5.12.1. Desempeño de versiones.

En esta sección se presenta una serie de tablas con los tiempos de procesamiento que se requirió en las etapas necesarias para la obtención de las taxonomías de los corpus.

Se muestra en la Tabla 5.82 el tiempo necesario para el desarrollo de la primera etapa, que es el pre-procesamiento de los textos (separarlos en tokens de acuerdo a los espacios en blanco e insertar una tupla en una tabla previamente creada para almacenar el término, el documento en el que se encuentra y la posición de está y por último contar la frecuencia de cada uno) para cada una de las dos versiones, se observa que la versión de UDF requirió de un menor tiempo que la versión con SP:

Tabla 5.82: Tiempo de ejecución del pre-procesamiento.

Versión	**Tiempo (Seg)**
SP	574.7
UDF	410.2

Para la siguiente etapa de la extracción de las taxonomías se muestra la Tabla 5.83 en la cual se presenta la media (en segundos) del tiempo de ejecución de cada método utilizada para obtener las relaciones de cada par de términos frecuentes, con ambas versiones, SP y UDF. En esta tabla se observa que la diferencia de la media entre ambas versiones es de menos de 100 segundos respecto a SP contra UDF, se observa también que en algunos casos la versión de SP resulta más rápida que la versión de UDF, pero en otros casos ocurre lo contrario.

Tabla 5.83: Tabla de medias de tiempo de ejecución de métricas.

Métrica	**Versión SP (Seg)**	**Versión UDF (Seg)**
correlación	5.9	4.5
distancia	15.8	17.7
soporte y confianza	3.6	2.4
soporte y confianza por intervalos (100)	79.9	93.7
soporte y confianza por intervalos (1000)	179.3	167.1

Si se observa la Tabla 5.83 se concluye que el método aportado por nosotros en este trabajo, el método de soporte y confianza por intervalos, fue el que requirió del mayor tiempo mientras que el método de soporte y confianza fue más rápido si se habla en términos del tiempo de ejecución.

5.12.2. Calidad de las taxonomías.

Ya que se ha discutido acerca del desempeño de los métodos, es importante también tratar la calidad del contenido de las taxonomías obtenidas en este trabajo. Para lograr este objetivo, se presenta una evaluación de los métodos utilizados para la obtención de la relación de un par de términos, de acuerdo a las métricas de calidad que como ya se mencionó anteriormente son una parte de la aportación de este trabajo.

La interpretación de las métricas de calidad que definimos en la sección 4.3 es la siguiente: si la taxonomía no contiene ningún término del título o de las palabras clave entonces suponemos que la taxonomía fue mal construida, en este caso el valor de las métricas serían 0 o cercanas a 0, a medida que el número crece, es decir, si se encuentran más palabras del título o palabras clave en la taxonomía obtenida la calidad de ésta asciende, lo anterior debido a que estamos suponiendo que el autor utilizó términos importantes desde el punto de vista taxonómico tanto en el título como en las palabras clave.

Para esto se muestra la Tabla 5.84 donde se muestra el porcentaje promedio de coincidencias que se obtuvieron con cada métrica de calidad (presentadas en las secciones 4.3.2 y 4.3.3) y los distintos métodos de obtención de los términos.

Tabla 5.84: Tabla de medias de calidad obtenidas.

Método de obtención de términos	**Promedio (%) de palabras en títulos**	**Promedio (%) de palabras clave**	**Promedio (%) de palabras en título y palabras clave (Ambas métricas)**
correlación	11.1	9.1	13.3
distancia	28.4	24.0	34.4
soporte y confianza	42.7	34.1	59.5
soporte y confianza por intervalos	41.8	31.4	50.1

Se puede observar que el método de soporte y confianza es con el que se obtiene una media mayor a las demás, seguido por el método propuesto de soporte y confianza por intervalos, posteriormente sigue el método de distancia y por último se tiene el método de correlación. Con respecto al método de correlación, cabe mencionar que al cambiar valor del umbral a un valor menor (se tomó como umbral= 0.6) se podría encontrar un mayor porcentaje de coincidencias, lo cual hace al resultado dependiente del umbral de validación. También se observa que la métrica de calidad de las palabras contenidas en los títulos de los documentos, se obtuvo mayor porcentaje de coincidencias en comparación con la métrica de las palabras clave en la mayoría de los casos.

Nuestras métricas de calidad nos llevan a concluir que las taxonomías de más alta calidad son las obtenidas a través del método de soporte y confianza dado que en promedio el 50 % de los términos encontrados tanto en el título como en las palabras clave es arriba del 50 %.

Capítulo 6

Lematización

El proceso de lematización es obtener el lema de las palabras, por esta razón es muy parecido al stemming, pero con algunas diferencias, ya que lematización puede utilizar un diccionario para obtener el lema y en algunas ocasiones el resultado de éstos dos procesos resulta diferente aunque en algunas otras resulta el mismo.

Por ejemplo, para la palabra en inglés "are" el resultado de aplicarle stemming sería "are", por otro lado, si se le aplica lematización se obtendría la palabra "be". Por lo que las taxonomías obtenidas con lematización pueden ser diferentes a las obtenidas con stemming.

A petición de una de las revisoras de esta trabajo se utilizó una herramienta para lematización de las palabras de los documentos, aunque esta no sigue con los propósitos de este trabajo que es utilizar el lenguaje de SQL, como lo son los SPs y las UDFs.

La herramienta para lematización se obtuvo de [57] y esta desarrollada con Java. La cual permite realizar diferentes procesos en un archivo de texto, algunos de estos procesos son:

- tokenize: separa en tokens el contenido del documento.
- cleanxml: elimina las anotaciones de XML del documento.
- ssplit: separa una secuencia de tokens en oraciones.
- pos: le asigna una etiqueta gramatical a cada token.
- lemma: obtiene el lemma de las palabras del documento.
- ner: reconoce los nombre de personas, organizaciones o localidades.

A continuación se presentan las Figuras correspondientes a cada corpus y las taxonomías que se obtuvieron con los métodos; aplicando la lematización a las palabras de los documentos en el preprocesamiento.

Primer corpus

El primer corpus consta de 86 documentos obtenidos de la página **dl.acm.org** con la búsqueda "**query optimization**", de los cuales se procesaron 887,470 tokens, y se obtuvo la taxonomía mostrada en la Figura 6.1 con el método de correlación.

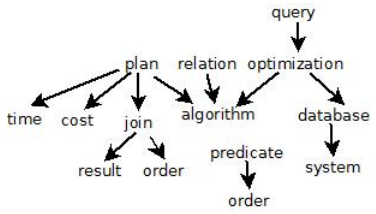

Figura 6.1: Taxonomía obtenida con el método de correlación.

Se muestra en la Figura 6.2 la taxonomía obtenida con el método de soporte y confianza.

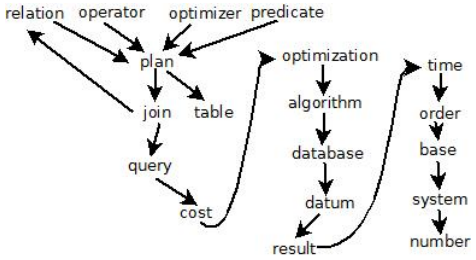

Figura 6.2: Taxonomía obtenida con el método de soporte y confianza.

Ahora se muestra la Tabla 6.3 con la taxonomía obtenida con el método de soporte y confianza por intervalos.

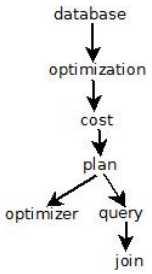

Figura 6.3: Taxonomía obtenida con el método de soporte y confianza por intervalos.

La Tabla 6.1 muestra las coincidencias de los términos que conforman la taxonomía con las métricas de calidad, las cuales son palabras clave, palabras contenidas en el título y ambas. Se puede observar en ésta que con todos los métodos se encontró mas del 50 % de coincidencias.

Tabla 6.1: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	**Ambas(%)**
correlación	20	60.0	60.0	60.0
distancia	20	60.0	85.0	90.0
soporte confianza	20	60.0	85.0	90.0
soporte y confianza intervalos	20	60.0	85.0	90.0

Segundo corpus

El segundo corpus con el que se experimento son 80 documentos obtenidos nuevamente de la página **dl.acm.org** con la búsqueda **"semantic Web**", esta vez, se procesaron 449,993 tokens.

Se muestra las Figuras 6.4 y 6.5 con las taxonomías obtenidas con los métodos de correlación y soporte y confianza respectivamente.

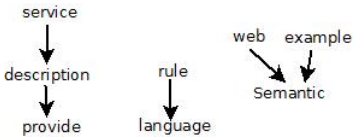

Figura 6.4: Taxonomía obtenida con el método de correlación.

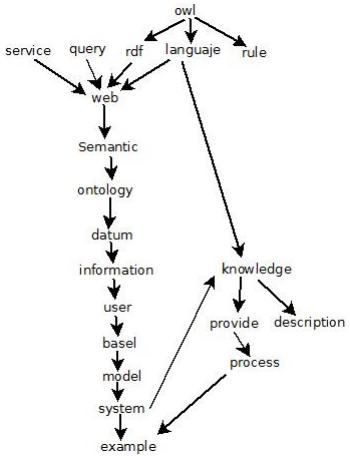

Figura 6.5: Taxonomía obtenida con el método de soporte y confianza.

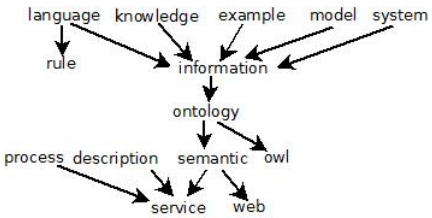

Figura 6.6: Taxonomía obtenida con el método de soporte y confianza por intervalos.

La Tabla 6.2 muestra las coincidencias de los términos que conforman la taxonomía con las métricas de calidad, las cuales son palabras clave, palabras contenidas en el título y ambas.

Tabla 6.2: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	**Ambas(%)**
correlación	20	25.0	30.0	30.0
distancia	20	0.0	0.0	0.0
soporte confianza	20	55.0	90.0	90.0
soporte y confianza intervalos	20	50.0	85.0	85.0

Cuarto corpus

El cuarto corpus consta de 100 documentos obtenidos de la página **dl.acm.org** con la búsqueda "**data quality**", para este corpus se procesaron 635,299 tokens.

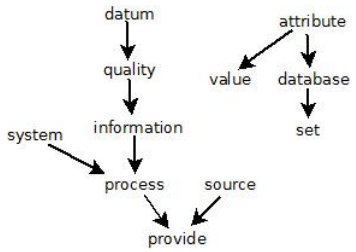

Figura 6.7: Taxonomía obtenida con el método de correlación.

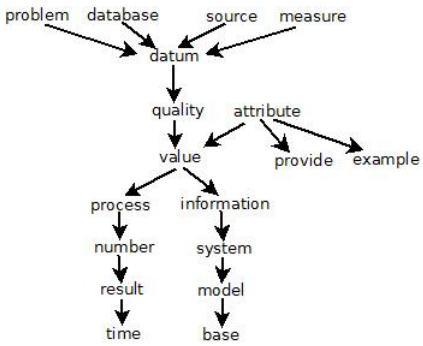

Figura 6.8: Taxonomía obtenida con el método de soporte y confianza.

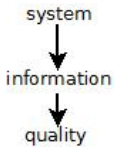

Figura 6.9: Taxonomía obtenida con el método de soporte y confianza por intervalos.

La Tabla 6.3 muestra las coincidencias de los términos que conforman la taxonomía con las métricas de calidad, las cuales son palabras clave, palabras contenidas en el título y ambas.

Tabla 6.3: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	**Ambas(%)**
correlación	20	40.0	50.0	55.0
distancia	20	40.0	50.0	60.0
soporte confianza	20	55.0	80.0	85.0
soporte y confianza intervalos	20	55.0	80.0	85.0

Quinto corpus

El quinto corpus consta de 100 documentos obtenidos de la página **dl.acm.org** con la búsqueda "**streaming**", en el cual se tienen un total de 880,774 tokens. Y se generó la taxonomía mostrada en la Figura 6.10 .

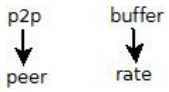

Figura 6.10: Taxonomía obtenida con el método de correlación.

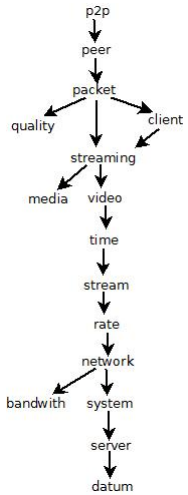

Figura 6.11: Taxonomía obtenida con el método de soporte y confianza.

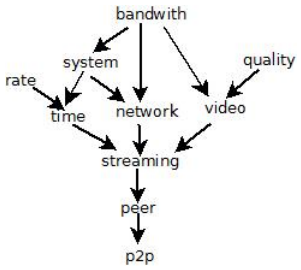

Figura 6.12: Taxonomía obtenida con el método de soporte y confianza por intervalos.

La Tabla 6.4 muestra las coincidencias de los términos que conforman la taxonomía con las métricas de calidad, las cuales son palabras clave, palabras contenidas en el título y ambas.

Tabla 6.4: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	**Ambas(%)**
correlación	20	20.0	20.0	15.0
distancia	20	10.0	10.0	10.0
soporte confianza	20	85.0	80.0	90.0
soporte y confianza intervalos	20	80.0	75.0	85.0

Sexto corpus

El sexto corpus consta de 86 documentos obtenidos de la página **dl.acm.org** con la búsqueda "**face detection**", para este corpus se procesaron 392,566 tokens.

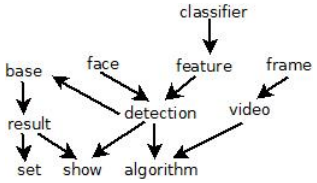

Figura 6.13: Taxonomía obtenida con el método de correlación.

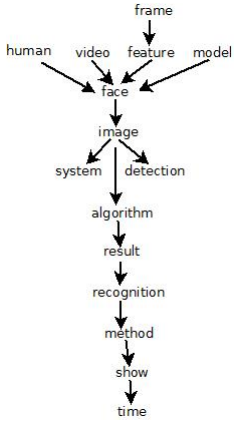

Figura 6.14: Taxonomía obtenida con el método de soporte y confianza.

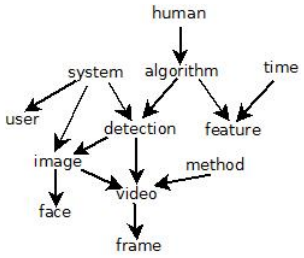

Figura 6.15: Taxonomía obtenida con el método de soporte y confianza por intervalos.

La Tabla 6.5 muestra las coincidencias de los términos que conforman la taxonomía con las métricas de calidad, las cuales son palabras clave, palabras contenidas en el título y ambas.

Tabla 6.5: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	**Ambas(%)**
correlación	20	60.0	40.0	60.0
distancia	20	0.0	0.0	0.0
soporte confianza	20	95.0	80.0	100.0
soporte y confianza intervalos	20	95.0	80.0	100.0

Séptimo corpus

Este corpus consta de 84 documentos obtenidos de la página **dl.acm.org** con la búsqueda **"data encryption"**, para el cual se procesaron 668,114 tokens.

La primer Figura de este corpus (Figura 6.16) corresponde a la taxonomía obtenida con el método de correlación.

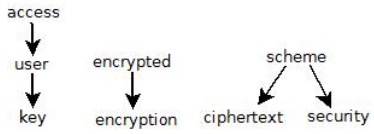

Figura 6.16: Taxonomía obtenida con el método de correlación.

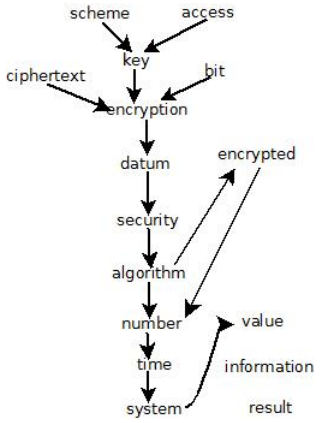

Figura 6.17: Taxonomía obtenida con el método de soporte y confianza.

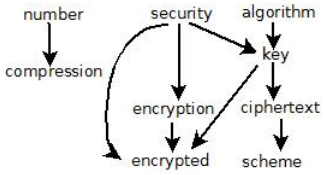

Figura 6.18: Taxonomía obtenida con el método de soporte y confianza por intervalos.

La Tabla 6.6 muestra las coincidencias de los términos que conforman la taxonomía con las métricas de calidad, las cuales son palabras clave, palabras contenidas en el título y ambas.

Tabla 6.6: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	**Ambas(%)**
correlación	20	40.0	30.0	40.0
distancia	20	0.0	0.0	0.0
soporte confianza	20	85.0	60.0	85.0
soporte y confianza intervalos	20	80.0	55.0	80.0

Octavo corpus

Este corpus consta de 76 documentos obtenidos de la página **http://oa-hermes.unam.mx/oa-hermes.html** con la búsqueda "**cancer breast**", para este corpus se procesaron 512,791 tokens.

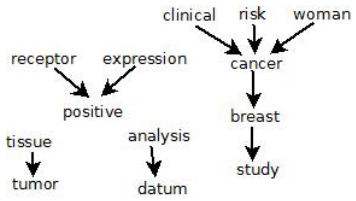

Figura 6.19: Taxonomía obtenida con el método de correlación.

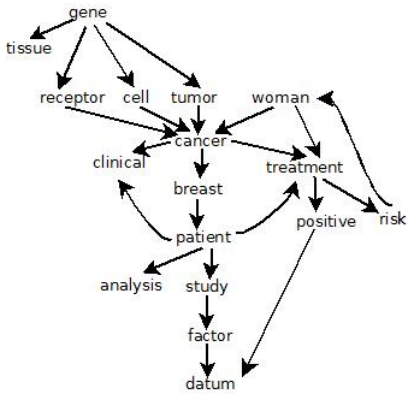

Figura 6.20: Taxonomía obtenida con el método de soporte y confianza.

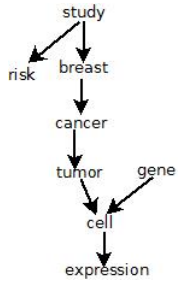

Figura 6.21: Taxonomía obtenida con el método de soporte y confianza por intervalos.

La Tabla 6.7 muestra las coincidencias de los términos que conforman la taxonomía con las métricas de calidad, las cuales son palabras clave, palabras contenidas en el título y ambas.

Tabla 6.7: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	**Ambas(%)**
correlación	20	30.0	55.0	55.0
distancia	20	50.0	85.0	85.0
soporte confianza	20	50.0	85.0	85.0
soporte y confianza intervalos	20	50.0	85.0	85.0

Noveno corpus

Este corpus consta de 100 documentos obtenidos de la página **http://oa-hermes.unam.mx/oa-hermes.html** con la búsqueda "**HIV**", para este corpus se procesaron 634,366 tokens.

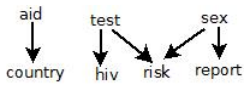

Figura 6.22: Taxonomía obtenida con el método de correlación.

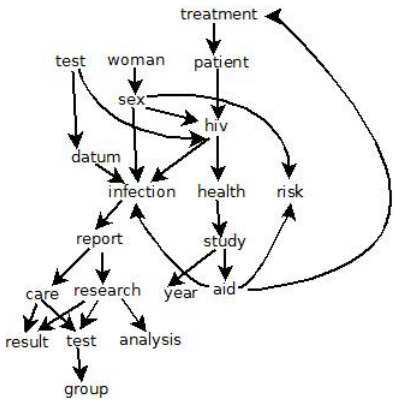

Figura 6.23: Taxonomía obtenida con el método de soporte y confianza.

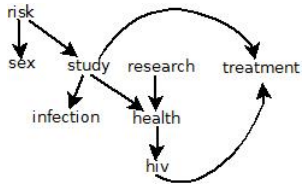

Figura 6.24: Taxonomía obtenida con el método de soporte y confianza por intervalos.

La Tabla 6.8 muestra las coincidencias de los términos que conforman la taxonomía con las métricas de calidad, las cuales son palabras clave, palabras contenidas en el título y ambas.

Tabla 6.8: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	**Ambas(%)**
correlación	20	30.0	30.0	10.0
distancia	20	0.0	0.0	0.0
soporte confianza	20	55.0	85.0	85.0
soporte y confianza intervalos	20	55.0	85.0	85.0

Décimo corpus

Este corpus consta de 96 documentos obtenidos de la página **http://oa-hermes.unam.mx/oa-hermes.html** con la búsqueda "**multiform gliobastoma**", para este corpus se procesaron 669,939 tokens.

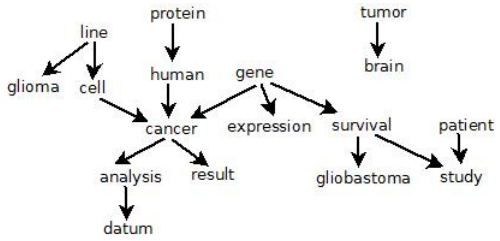

Figura 6.25: Taxonomía obtenida con el método de correlación.

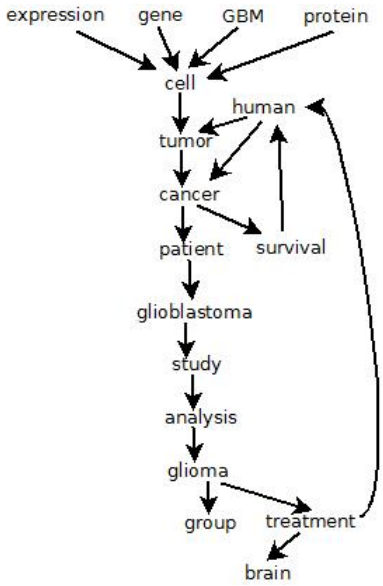

Figura 6.26: Taxonomía obtenida con el método de soporte y confianza.

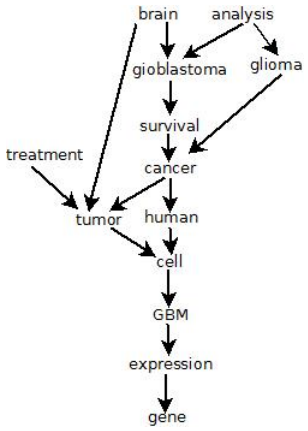

Figura 6.27: Taxonomía obtenida con el método de soporte y confianza por intervalos.

La Tabla 6.9 muestra las coincidencias de los términos que conforman la taxonomía con las métricas de calidad, las cuales son palabras clave, palabras contenidas en el título y ambas.

Tabla 6.9: Resumen de evaluación de las métricas.

Métrica	**No. de términos**	**Palabras clave(%)**	**Título(%)**	**Ambas(%)**
correlación	20	10.0	75.0	75.0
distancia	20	10.0	35.0	35.0
soporte confianza	20	10.0	80.0	80.0
soporte y confianza intervalos	20	10.0	75.0	75.0

6.1. Resultados de lematización

Para realizar la comparación del stemming con la lematización se muestra la Tabla 6.10 que se había mostrado previamente en la sección 5.12.2 donde se muestran las medias de las coincidencias encontradas entre los términos que conforman las taxonomías y los métodos de obtención de los términos, a los cuales se les aplico stemming en el preprocesamiento.

Tabla 6.10: Tabla de medias de calidad obtenidas utilizando stemming.

Método de obtención de términos	**Promedio (%) de palabras en títulos**	**Promedio (%) de palabras clave**	**Ambas métricas**
correlación	11.1	9.1	13.3
distancia	28.4	24.0	34.4
soporte y confianza	42.7	34.1	59.5
soporte y confianza por intervalos	41.8	31.4	50.1

Ahora se muestra la Tabla 6.11 con las medias de los porcentajes de las coincidencias de los términos de la taxonomía con las métricas de calidad del contenido de las taxonomías. Como puede observarse en ésta, se obtiene en general una media mayor que cuando se utiliza stemming, por lo que podemos concluir que las taxonomías generadas con lematización tienen mayor calidad que las generadas con stemming de acuerdo a las métricas de calidad definidas en este trabajo.

Tabla 6.11: Tabla de medias de calidad obtenidas utilizando lematización.

Método de obtención de términos	**Promedio (%) de palabras en títulos**	**Promedio (%) de palabras clave**	**Ambas métricas**
correlación	35.0	43.3	44.4
distancia	16.6	26.1	27.7
soporte y confianza	61.1	80.5	87.7
soporte y confianza por intervalos	53.8	78.3	85.5

Se observa en las tablas de resumen de cada corpus que en general se obtiene un número mayor de coincidencias de acuerdo a las métricas de calidad cuando se utiliza lematización, incluso en algunos casos se obtiene el 100 % de coincidencias.

Capítulo 7

Conclusiones y trabajo a futuro

A lo largo de este trabajo se presentaron distintos métodos (correlación, distancia, soporte y confianza por intervalos y soporte y confianza) para obtener taxonomías de corpus de documentos almacenados en la base de datos y distintas métricas de calidad (palabras del título y palabras clave), las cuales fueron utilizadas para determinar si son válidas o no las taxonomías obtenidas.

Con respecto a los métodos aportados en este trabajo; con el método definido como soporte y confianza por intervalos, se mostró que incrementar el tamaño de los intervalos es consistente con los resultados obtenidos con tamaños menores en este método, se observa también que se obtienen menos relaciones en comparación con los otros métodos presentados en este trabajo, por lo que se concluye que es más restrictivo, pero consistente, ya que también se obtuvo un porcentaje alto (en comparación con los demás métodos) de coincidencias de acuerdo a las métricas de calidad.

Como se mencionó en la sección 5.12 el método de soporte y confianza fue el más rápido con respecto al tiempo de ejecución y con el que se obtuvo las taxonomías de mayor calidad, ya que la media de coincidencias con las métricas de calidad es mayor al 50 %.

Mientras que respecto a las métricas de calidad, la métrica de palabras contenidas en los títulos de los documentos resultó encontrar un mayor número de coincidencias en comparación con la métrica de palabras clave. Cabe mencionar que al utilizar las palabras clave y las palabras contenidas en los títulos como una solo métrica de calidad el porcentaje de coincidencias más alto encontrado en los corpus fue de 80 %, lo cual consideramos que es un porcentaje muy alto, por lo que se concluye que las taxonomías obtenidas son de buena calidad.

Respecto al tiempo de ejecución de las versiones de procedimientos almacenados (SP) y funciones definidas por el usuario (UDF), como se muestra en la sección anterior, no hubo una gran diferencia en tiempo entre ambas versiones, pero cabe mencionar que la versión de UDF requirió de ejecutar un mayor

número de consultas a la base de datos, ya que no se puede insertar los datos obtenidos dentro del cuerpo de la función, en contraste a las SP que si lo permiten.

Por último de acuerdo a los experimentos realizados con lematización, se concluye que al obtener las taxonomías aplicándoles a los términos lematización en lugar de stemming se obtienen taxonomías con contenido (los términos que la conforman) de mejor calidad, ésto de acuerdo a las métricas de calidad definidas en este trabajo.

A continuación se mencionan los puntos a considerar para el trabajo futuro, los cuales son:

- Escribir las taxonomías obtenidas como ontologías formales expresadas en RDF u OWL.
- Definir otra métrica de calidad de las taxonomías utilizando el abstract de los documentos, ya que en esta parte del documento, el (los) autor (es) resumen el contenido del documento.
- Comparar las taxonomías obtenidas de los abstract de los documentos con las taxonomías del cuerpo de los documentos, para observar si son parecidas y de esta forma mejorar la interpretación de la calidad de éstas.

Bibliografía

[1] C. Garcia-Alvarado, Z. Chen, and C. Ordonez. "ONTOCUBE: Efficient Ontology Extraction using OLAP cubes", CIKM'11, 2011.

[2] P. Buitelaar, D. Olejnik, and M. Sintek. "OntoLT: A protégé plug-in for ontology extraction from text", In Proc. Of ISWC, 2003

[3] Z. Chen and C. Ordonez, "Efficient OLAP with UDFs", In Proc. ACM DOLAP Woskshop, pages 41-48, 2008.

[4] A. Elsayed, S. El-Beltagy, M. Rafea, and O. Hegazy, "Applying Data Mining for Ontology Building", In Proc. Of ISSR, 2007.

[5] N. Fukuta, T. Yamaguchi, T. Morita, and N. Izumi, "DODDLE-OWL: Interactive Domain Ontology Development with Open Source Software in Java", IEICE TOIS, 91(4):945-1356, 2010.

[6] A. Gal, G. Modica, and H. Jamil. "Ontobuilder: Fully automatic extraction and consolidation of ontologies from Web sources", In Proc. Of ICDE, page 853, 2004.

[7] C. Garcia-Alvarado, Z. Chen, and C. Ordonez, "OLAP-based query recommendation", In Proc. Of ACM CIKM, pages 1353-1356, 2010.

[8] Carlos Ordonez, Zhibo Chen, Javier García-García, "Interactive exploration and visualization of OLAP cubes", DOLAP 2011: 83-88, 2011.

[9] Knoodl knoodl.com/ (27/04/13 7:44 pm).

[10] SemanticWorks
http://www.altova.com/documents/SemanticWorksdatasheet.pdf
(18/12/2013 11:47 am).

[11] TopQuadrant Products
http://www.topquadrant.com/products/TB_Composer.html (11/06/2013 10:40 am).

[12] Javier García-García, Carlos Ordonez, "Repairing OLAP queries in databases with referential integrity errors", DOLAP págs. 61-66, 2010.

[13] Carlos Ordonez, Javier García-García, "Database systems research on data mining" SIGMOD Conference 2010: 1253-1254.

[14] C. Ordonez, "Statistical Model Computation with UDFs", IEEE TKDE, 2010.

[15] R. Navigli, and P. Velardi, "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites", Computational Linguistics (MIT Press Journals), Vol. 30, No. 2 , Pages 151-179, 2004

[16] M. Balakrishna, and M. Srikanth, "Automatic Ontology Creation from Text for National Intelligence Priorities Framework (NIPF)", Ontology for the intelligence community (George más on University), Pages 8-12, 2008.

[17] C. Xide, Lin, B. Ding, J. Han, F. Zhu, and B. Zhao. "Text Cube: Computing IR Measures for Multidimensional Text Database Analysis", In Proc. Of ICDM, Pages 905-910, 2008.

[18] P. Vassiliadis, "Modeling Multidimensional Databases, Cubes and Cube Operations", In Proc. Of SSDBM, Pages 53-62, 1998.

[19] T. Neimi, J. Nummenmaa, and P. Thanish, "Constructing OLAP Cubes Based on Queries", DOLAP'01, 2001.

[20] Standard Upper Ontology Working Group (SUO WG, http://suo.ieee.org/) (25/07/2013 5:15 pm).

[21] Guarino, N. "Formal Ontologies and Information Systems", Proceedings of FOIS'98, Italy, 1998.

[22] Mahesh, K., "Ontology Development for Machine Translation: Ideology and Methodology", NMSU. Computing Research Laboratory. Technical Report MCCS-96-292. New Mexico, 1996.

[23] Gruber, T. R., "A Translation Approach to Portable Ontologies", KnOW-Ledge Acquisition, Standard Upper Ontology Working Group, 5(2): 199-220, 1993.

[24] OWL http://www.w3.org/2001/sw/wiki/OWL (18/12/2013 12:13 pm).

[25] RDF http://www.w3.org/RDF/ (16/19/2013 10:01 pm).

[26] Ontologías y su representación jerárquica
http://catarina.udlap.mx/tales/documentos/mcc/sanchez_l_se/capitulo4.pdf
(26/12/2013 9:48 pm).

[27] Guía breve de Web Semántica
http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica
(26/12/2013 11:39 pm).

[28] Jérôme Gabillaud, "SQL Server 2008 SQL Transact SQL", Ediciones ENI, págs 402, 2009.

[29] Paul Turley, Dan Wood, "Beginning T-SQL with Microsoft SQL Server 2005 and 2008", John Wiley & Sons, 750 páginas.

[30] Create procedure http://msdn.microsoft.com/en-us/library/aa258259%28v=sql.80%29.aspx (08/06/2013 5:39 pm).

[31] Jiawei Han, Micheline Kamber, Jian Pei, "Data minig: Concepts and techniques", Data Minig Third edition, Elsevier, 697, 2006.

[32] John D. Holt, Soon M. Chung, "Multipass Algorithms for Mining Association Rules in Text Databases", KnOWLedge and Information Systems, Springer-Verlag London Ltd., 2001.

[33] Funciones de agregado (Transact-SQL) http://msdn.microsoft.com/es-es/library/ms173454.aspx (13/02/2013 5:32 pm).

[34] Chris Fehily, "SQL: Visual QuickStart Guide", Peachpit Press, 504 páginas, 2008.

[35] Create function. http://technet.microsoft.com/es-es/library/ms186755.aspx (01/10/2013 16:40 pm)

[36] Porter Stemming Algorithm http://tartarus.org/martin/PorterStemmer/ (22/04/13 11:18 am)

[37] D. Manning, Raghavan Prabhakar, Shótze Hinrich, "An introduction to information retrieval", Cambridge UP, 2007.

[38] Rosario Girardi, Carla Gomes de Faria, "ONTODUM", CLEI ELECTRONIC JOURNAL, 2004.

[39] Home - PubMed - NCBI http://www.ncbi.nlm.nih.gov/pubmed (11:56 am 21/06/2013)

[40] Samir Tatir, I. Budak Arpinar, Michael Moore, Amit P. Sheth, Boanerges Aleman-Meza, "OntoQA: Metric-Based Ontology Quality Analysis", University of Georgia, IEEE Workshop on KnOWLedge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and KnOWLedge Sources, 2005

[41] Kathi Kellenberger, Scoot Shaw, "Beginning T-SQL 2012", Apress, 456 páginas.

[42] Garth Wells, "Code Centric: T-SQL Programming with Stored Procedures and Triggers", Apress, 2001, 692 páginas.

[43] Enrique Paniagua Arís (coordinador); Belén López Ayuso, "La gestión tecnológica del conocimiento", EDITUM, 324 págs, 2007.

[44] Declare cursor http://technet.microsoft.com/es-es/library/ms180169.aspx (23/08/2013 01:07 pm)

[45] RDF Primer http://www.w3.org/TR/2004/REC-RDF-primer-20040210/ (30/09/2013 10:44 pm).

[46] Taxonomías
http://www.infor.uva.es/~sblanco/Tesis/Ontolog%C3%ADas.pdf
(12/12/2013 3:25 pm).

[47] Flor Nancy Díaz Piraquive, Luis Joyanes Aguilar, Víctor Hugo Medina García, "Taxonomía, ontología y folksonomía, ¿qué son y qué beneficios u oportunidades presentan para los usuariosde la Web?", Universidad & Empresa, Colombia, 2009.

[48] Ángel Velázquez Iturbide, "Tecnologías del software: Seminario de investigación e innovación en tecnologías del software", Librería-Editorial Dykinson, 258 págs, 2007.

[49] Amparo Alcina, Esperanza Valero, Elena Rambla, "Términología y sociedad del conocimiento", Peter Lang, 408 págs, 2009.

[50] Luis Bayardo Sanabria Rodríguez, David Macías Mora, "Formación de competencias docentes: diseñar y aprender con ambientes computacionales", U. Pedagógica Nacional, 140 págs, 2006.

[51] Javier Tuya, Isabel Ramos Román, José Javier Dolado Cosín, "Técnicas cuantitativas para la gestión en la ingeniería del software", Netbiblo, 373 págs, 2007.

[52] Ming-Syan Cheng, Philip S. Yu, Bing Liu, "Advances in Knowledge Discovery and Data Mining", Springer, 568 págs, 2002.

[53] http://www.microsoft.com/es-es/server-cloud/products/sql-server/ (27/04/2014 4:42 pm).

[54] Configure and manage Stopwords and Stoplist for Full-Text Search http://msdn.microsoft.com/en-us/library/ms142551.aspx (17/06/2014 2:31 pm).

[55] Michael Lee, Gentry Bieker, "Mastering SQL Server 2008", John Wiley & Sons, 792 págs, 2009.

[56] C. J. Date, Sergio Luis María Ruiz Faúdon, "Introducción a los sistemas de bases de datos", Pearson Education, 936 págs, 2001.

[57] The Stanford NLP nlp.stanford.edu/software/corenlp.shtml (19/06/2014 1:49 am)

[58] Data Warehouse
http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager (19/06/2014 2:59 am)

[59] Stemming and lemmatization http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html (27/06/2014 11:08 am)