Sistema para el apoyo al análisis de patentabilidad

Andrés Felipe Vásquez Castrillón
Cod. 201255724
andres.felipe.vasquez@correounivalle.edu.co

Alexander Escobar Luna
Cod. 201255669
alexander.escobar@correounivalle.edu.co

Director
Federico López Gómez, Ing.

Co-director
Victor Andrés Bucheli Guerrero, PhD.

Trabajo de grado para optar el título de Ingeniero de Sistemas

Universidad del Valle Sede Tuluá
Programa de Ingeniería de Sistemas
Octubre de 2016

Dedicatoria

A la persona que me ha dado todo.

Andrés

Al abuelo y a la Abuela por su incondicional apoyo a lo largo de éste camino, a mamá y a papá por estar ahí, a la reina de mi vida que siempre estuvo a pesar de las dificultades, al ingeniero y a la doctora por su gran apoyo, a todos los que estuvieron ahí dándome ánimo cada vez que lo necesité, a toda mi gran familia y a los tres fantásticos.

Alexander

Agradecimientos

Con agradecimiento a todas las personas que de alguna u otra forma hicieron posible la culminación de esta carrera.

Índice general

Índice de figuras

Índice de tablas

Resumen

Existe un interés por parte de las organizaciones en patentar sus resultados con el fin de proteger sus invenciones. Al realizar una solicitud de patente es necesario determinar si existen invenciones similares y qué tan novedosa es la invención presentada con respecto a otras.

En este trabajo se propone una herramienta de software que soporta el análisis de las solicitudes de patentes con el fin disminuir el tiempo dedicado a su evaluación. La aproximación propuesta utiliza técnicas de diferentes áreas de las Ciencias de la Computación tales como la Web Semántica, el Procesamiento del Lenguaje Natural y la Minería de Texto.

Palabras Clave: Procesamiento del Lenguaje Natural, Minería de Texto, Análisis de Patentes, Anotación Semántica.

Capítulo 1

Introducción

1.1. Descripción General del Problema

En un contexto en el que los mercados son cada vez más cambiantes y globales, la necesidad de proteger los productos o de afirmar y conservar la propiedad de los procesos e innovaciones es importante para mejorar la posición competitiva. Muy frecuentemente, empresas que han tardado años en desarrollar y madurar sus productos se dan cuenta de que están siendo copiados por sus competidores [1].

Los documentos de patente consisten en descripciones bien estructuradas de la innovación tecnológica y de los resultados de investigaciones, los cuales son típicamente escritos con una terminología en específica, que generalmente requiere de la experiencia de un experto de dominio para ser interpretada con el fin de llevar a cabo un análisis de patente. El rápido crecimiento del número de solicitudes de patentes, que según la WIPO[1] (World Intellectual Property Organization) en el 2014 fue de 4,5 por ciento en relación con el año 2013 y el incremento en los registros aumenta la dificultad de extraer y sintetizar el conocimiento contenido en las patentes. [2].

Los centros de investigación y desarrollo tales como las universidades, generalmente tienen dependencias que se encargan de asesorar a sus investigadores o inventores en el proceso de patentar sus invenciones o aportes. En este proceso de asesoría, es necesario determinar qué tan viable es solicitar una patente para un nuevo aporte o invención. A este proceso se le llama Análisis de Patentabilidad y busca determinar qué tan "patentable" es un aporte determinado. Para que una invención se pueda patentar, esta debe cumplir con los siguientes requisitos:

1. Debe ser novedoso, es decir, que no exista a nivel mundial.

[1] http://www.wipo.int/portal/en/index.html

2. Debe poseer un nivel inventivo, lo que equivale a decir, que no sea un desarrollo obvio para alguien experto en la materia que trata el invento.

3. Que lo inventado pueda ser utilizado o fabricado en cualquier industria, es decir, que debe tener una aplicación industrial.

Una forma de verificar si la invención cumple con estos requisitos es revisar bases de datos o repositorios de patentes con el objetivo de encontrar invenciones similares. Sin embargo, en muchos casos, el resultado de la búsqueda es demasiado grande, lo cual hace que una revisión manual, una a una, sea un proceso lento y complicado.

Teniendo en cuenta lo anterior, es evidente la necesidad de desarrollar una aplicación que haga uso de los diferentes conceptos de la ciencia de la computación y el uso de nuevas tecnologías con el fin de brindar apoyo al proceso de análisis de patentabilidad y de este modo agilizarlo.

1.1.1. Formulación del problema

¿Cómo facilitar el proceso de análisis de patentabilidad de solicitudes de patentes en un campo específico?

1.2. Objetivos

1.2.1. Objetivo general

Desarrollar una herramienta de software para apoyar al proceso de análisis de patentabilidad de solicitudes de patente en un campo específico.

1.2.2. Objetivos específicos

En la Tabla 1.1 se relacionan los objetivos específicos con las diferentes secciones exponen el resultado de los objetivos propuestos

Objetivo específico	**Sección**
Revisar la literatura de las diferentes técnicas de las Ciencias de la Computación aplicadas al proceso de análisis de patentabilidad.	2.2.2
Implementar un algoritmo que permita determinar qué tan similares son dos documentos.	3.3
Implementar un algoritmo que determine qué tan novedoso es un documento con respecto a un conjunto de documentos.	3.4.
Desarrollar una aplicación prototipo que integre los dos algoritmos propuestos	3.5.

Tabla 1.1: Relación de los objetivos específicos con el resultado esperado

1.3. Estructura del Documento

El resto del documento está organizado de la siguiente manera:

En el *Capítulo 2: Marco referencial*, se presenta el marco conceptual y los antecedentes sobre herramientas para el análisis de patentabilidad.

En el *Capítulo 3: Sistema para el apoyo al análisis de patentabilidad*, se presentan los resultados obtenidos en el proceso de desarrollo.

En el *Capítulo 4: Aspectos del desarrollo de software*, se muestran las diferentes fases cumplidas en el desarrollo y el proceso de ingeniería de software llevado a cabo para cumplir con los objetivos propuestos. Se recrean diferentes escenarios de pruebas para medir la efectividad del sistema teniendo en cuenta los objetivos y los requerimientos propuestos.

Este trabajo finaliza con el *Capítulo 5: Conclusiones y trabajos futuros.*

1.4. Publicación

Parte de este trabajo se ha publicado en:

A. Vásquez, A. Escobar, F. López y V. Bucheli. Herramienta de apoyo al análisis de patentabilidad. Congreso Colombiano de Computación - 11CCC, ISBN 978-1-5090-2965-5. Octubre 2016. Popayan, Colombia.

Capítulo 2

Marco Referencial

2.1. Marco Teórico

2.1.1. Patente

La Patente es un privilegio que le otorga el Estado al inventor como reconocimiento a la inversión y esfuerzos realizados por éste para lograr una solución técnica que le aporte beneficios a la humanidad. Dicho privilegio consiste en el derecho a explotar exclusivamente el invento por un tiempo determinado [3].

El titular de una patente puede decidir quién puede o no utilizar la invención patentada durante el periodo en el que está protegida, puede conceder autorización o una licencia a terceros para utilizar la invención con sujeción a las condiciones establecidas de común acuerdo y puede vender el derecho a la invención a un tercero, que se convertirá en el nuevo titular de la patente. A cambio de la protección que el Estado concede al inventor, éste debe revelar detalladamente la manera de producir y utilizar la invención.

Se cree que la primera patente fue concedida el año 1421, en Florencia, al ingeniero Filippo Bruneleshi, que había diseñado un mecanismo basado en engranajes para la carga de mármol en embarcaciones. La patente le concedía el monopolio en la manufactura de esta embarcación durante tres años. El primer decreto sobre patentes fue aprobado por el gobierno de Venecia en 1474 [1].

2.1.2. Análisis de patentabilidad

El análisis de patentabilidad consiste en determinar si una solicitud de patente es patentable o no. Para que una invención se pueda patentar, esta debe cumplir con los

siguientes requisitos [3]:

1. Debe ser novedoso, es decir, que no exista a nivel mundial.
2. Debe poseer un nivel inventivo, lo que equivale a decir, que no sea un desarrollo obvio para alguien experto en la materia que trata el invento.
3. Que lo inventado pueda ser utilizado o fabricado en cualquier industria, es decir, que debe tener una aplicación industrial.

No todos los productos o procedimientos se consideran invenciones y no todas las invenciones son patentables. En Colombia la legislación contiene una lista de lo que no se considera invención:

1. Los descubrimientos, las teorías científicas y los métodos matemáticos.
2. Los seres vivos o una parte de él tal como se encuentran en la naturaleza.
3. Los procesos biológicos naturales, el material biológico existente en la naturaleza o aquel que pueda ser aislado, inclusive genoma o germoplasma de cualquier ser vivo.
4. Las obras literarias y artísticas.
5. Los planes, reglas y métodos para el ejercicio de actividades intelectuales, juegos o actividades económico-comerciales.
6. Los programas de computador o el soporte lógico.
7. Las formas de presentación de información.

2.1.3. Anotación Semántica de Documentos

La anotación según la RAE[1] "constituye la acción y efecto de anotar", y anotar, se define como, "poner notas en un escrito, una cuenta o un libro". En lingüística computacional una anotación es una nota añadida a una parte específica de un recurso, que bien puede ser un texto, una imagen, un archivo de audio, etcétera.

La Semántica según el diccionario de la RAE es la disciplina que estudia el significado de las unidades lingüísticas y de sus combinaciones. Una de las ramas de la semántica es la semántica léxica que se define como: la rama de la semántica que estudia el significado de las palabras, así como las diversas relaciones de sentido que se establecen entre ellas.

[1] http://www.rae.es/

De acuerdo con las definiciones anteriores se puede decir que la Anotación semántica de documentos es darle significado a las palabras de un texto agrupándolas e integrándolas a un contexto determinado, de esta forma se consigue determinar el rol semántico que cumple una palabra en un documento de texto y se le enseña a la máquina el significado del documento.

Lo que se busca con estas anotaciones es conseguir que las máquinas entiendan el significado, la semántica, de los textos escritos [4], de esta forma se permite estructurar la información mediante su clasificación con base en conceptos semánticos. Este es el primer paso para permitir el procesamiento automático de la información, contenida en documentos, por parte de las máquinas.

2.1.4. *Clustering*

Clustering o agrupación es uno de los temas de investigación más importantes en las comunidades de *Machine Learning* y *Data Mining*. Su objetivo es agrupar las muestras en varios grupos, las muestras en el mismo grupo son similares, mientras que las muestras en diferentes grupos son diferentes [5].

La definición de clustering conduce directamente a la definición de un solo "clúster" (grupo). Muchas definiciones se han propuesto en los últimos años. Sin embargo, la mayoría de estas definiciones están basadas en términos vagamente definidos, tales como *similares* y *parecidos*. Este hecho revela la dificultad de tener una definición aceptable para el termino clúster [6].

En [7], los vectores son vistos como puntos en un espacio l-dimensional, y el cluster descrito como "regiones continuas de este espacio que continene una alta densidad de puntos, separado de otras regiones de alta densidad por regiones con una baja densidad de puntos" .

Así, teniendo en cuenta la definición de [7], en [6] intentan dar una definición para "clustering", que si bien no es universal, da una idea de los que es *Clustering*. Sea X un conjunto de datos, tal que:

$$X = \{x_1, x_2, ...x_N\}$$

Se define como un *m-clustering* de X, la partición de X en m conjuntos (*clúster*), C_1,...C_m, de manera que se cumplan las siguientes condiciones:

- $C_i \neq \emptyset,\ i = 1, ..., m$
- $\bigcup_{i=1}^{m} C_i = X$

- $C_i \bigcap C_j \neq \emptyset, i \neq j, i, j = 1, ..., m$

Los vectores contenidos en el cluster C_i son "más similares" entre sí, y son "menos similares" a los vectores de otros clusters.

2.1.4.1. Algoritmos de clustering

Toda la colección de clusters se conoce comúnmente como clustering [8]. Los algoritmos de clustering pueden ser vistos como esquemas que proporcionan agrupaciones sensibles considerando solamente una pequeña fracción del conjunto que contiene todas las posibles particiones de X. El resultado depende del algoritmo y los criterios usados [6]. Los algoritmos de clustering pueden ser divididos en las siguientes categorías:

- *Algoritmos secuenciales*
- *Algoritmos jerárquicos*
- *Algoritmos de clustering basados en la optimización de una función costo*

2.1.4.2. Algoritmo K-means

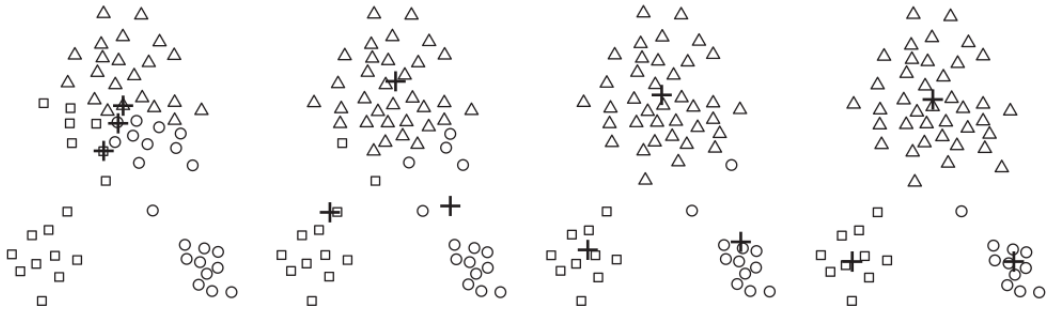
K-means pertenece a la categoría de algoritmos de clustering basado en optimización de una función. Es uno de las técnicas más populares dada su simplicidad [6]. Una descripción general se puede dar de la siguiente manera: en primer lugar se eligen los K centroides iniciales, donde K debe ser especificado por el usuario, es decir, el número de clusters deseados. A cada punto se le asigna el centroide más cercano, y cada colección de puntos asignados a un centroide es un cluster. A continuación el centroide de cada cluster es actualizado basado en los puntos asignados al cluster. Los pasos de asignar y actualizar se repiten hasta que los puntos del cluster no cambien, o hasta que los centroides sigan siendo los mismos [8].

La Figura 2.1 describe formalmente al algoritmo y la Figura 2.2 ilustra el uso de *K-means* en un conjunto específico de datos.

Algorithm 8.1 Basic K-means algorithm.

1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: **until** Centroids do not change.

FIGURA 2.1: Descripción de K-means (Tomada de [8])



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

FIGURA 2.2: Usando algoritmo *K-means* para encontrar tres clusters (Tomada de [8])

2.1.5. Medición de la novedad

La novedad es un concepto muy complejo. En [9] definen la novedad "como algo que era desconocido antes de un punto en particular en el tiempo, por lo tanto fue creado o descubierto en ese momento". Si la novedad se define como algo que había sido desconocido hasta ahora, entonces esto implica que la novedad también está directamente relacionado con lo que se conocía antes.

2.1.5.1. Técnicas para medir la novedad

Una patente representa una invención en un determinado campo de la tecnología y estudios previos consideran que parte de la información presentada en patentes es relativamente nueva [10]. Debido a las ventajas que presenta el contenido de una patente, ya se han establecido métodos para medir o determinar la novedad.

Uno de estos métodos es el de *analizar las citaciones de las patentes*, la cual es frecuentemente utilizada para el análisis de patentes y aumentar la comprensión en un campo tecnológico. Esta técnica tiene su origen en los estudios bibliométricos de publicaciones científicas y han sido aplicadas para analizar el flujo de conocimiento entre sectores tecnológicos y regiones geográficas, para descubrir patrones en la literatura de patentes y evaluar su valor. Sin embargo, se presentan varias desventajas: (i) el alcance de la información es limitado, ya que se centra exclusivamente en la información bibliográfica y no tiene en cuenta la sección de la descripción en la patente, (ii) la citación de patentes se limita a indicar enlaces individuales entre patentes y por lo tanto no es adecuado para un análisis de relación y(iii) las referencias de una patente están sujetas a decisiones estratégicas. Ésto puede influir en el número y la variedad de las referencias [9].

En [11], proponen una nueva medida para estimar la novedad en las invenciones usando las IPC (International Patent Classification) y evaluar dicha medida en comparación con la patentabilidad de las innovaciones. Las IPC se utilizan para la clasificación de la invenciones en las patentes y solicitudes de patentes según las diferentes áreas de la tecnología a la que pertenecen las invenciones.

También hay métodos basados en palabras claves para analizar las patentes y así hacer frente con las limitaciones que presenta el análisis de las citaciones de patentes. Este procedimiento implica una comparación de la ocurrencia de las palabras claves, por ejemplo medir la frecuencia de los términos o la extracción y clasificación de palabras. No obstante, la extracción y comparación de palabras son poco específicas para un análisis detallado [9].

2.2. Estado del arte o Antecedentes

2.2.1. Herramientas para el análisis de patentes

La gestión inteligente de las patentes es muy importante, ya que éstas constituyen un componente crucial del motor que impulsa la economía. Las patentes son de suma importancia para las empresas, las cuales proporcionan protección legal para las técnicas inventadas, procesos o productos [12]. Éstas mantienen información relevante sobre la propiedad intelectual de las invenciones [13]. Es así que se han desarrollado diversas herramientas que ayudan a hacer análisis de patentes, con el objetivo de que las organizaciones sean más competitivas y tomen mejores decisiones en la investigación y el desarrollo (I+D).

Entre las herramientas que permiten hacer un análisis de patentabilidad, están:

- **VantagePoint**[2]
  Es una herramienta de minería de texto que permite descubrir conocimiento en los resultados de búsqueda de bases de datos de patentes y de literatura.

  VantagePoint posibilita la agrupación de patentes por familias, los recuentos por frecuencia de número de patentes por año, por organización, por autor, hasta realizar sofisticados análisis estadísticos, y mostrar toda la información relacionada con cualquier término, organización, año, etc., que le interese al usuario, de manera interactiva y visual.

  Esta herramienta cuenta con las siguientes características:

[2] https://www.thevantagepoint.com/

- Importing: Obtener los datos brutos en VantagePoint y minar los datos en bruto para obtener más datos de él.
- Cleaning: Transformar los datos en un conjunto coherente, combinando lo que se quiere analizar como un grupo, y la fusión y normalización de datos de diversas fuentes.
- Analyzing: Mirar sus datos en una variedad de maneras.
- Reporting: Preparación para comunicar los resultados.
- Automating: Codificación de todo el proceso para que sea consistente y fácilmente repetible.

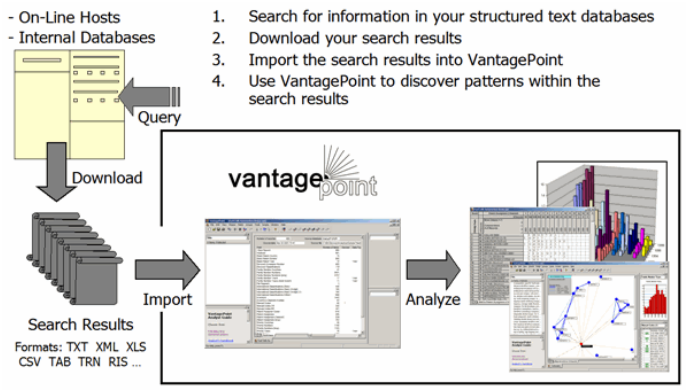


FIGURA 2.3: Proceso de VantagePoint

- **Wisdomain**[3]

  Se especializa en soluciones de minería de patentes, permitiendo a los profesionales de la IP (Propiedad Intelectual) buscar y analizar en los sectores de la tecnología de todo el mundo, con el objetivo de hacer más rápido y rentable la toma de decisiones sobre estrategias de I+D, posicionamiento competitivo, y las oportunidades de mercado.

  Su interfaz es sencilla y amigable, la cual permite crear rápidamente representaciones visuales con información de las patentes analizadas, ofreciendo la salida de estos datos en dos formas diferentes: tablas y gráficos (2D y 3D) e informes. Cuenta con herramientas inteligentes para aumentar la productividad de la búsqueda, con las cuales se puede hacer:

  - Asignar colores a las palabras clave y destacarlos en los documentos.
  - Revisar varios dibujos de patentes y bibliografía en modo de presentación.
  - Mostrar dos documentos de patentes lado a lado para facilitar la comparación.

[3]http://www.wisdomain.com/About/ko/CompanyEn

- Realizar el análisis de citas y crear gráficos directamente desde los resultados de búsqueda.
- Crear una presentación visual de la patente directamente del resultado de la búsqueda.
- Filtrar el resultado de la búsqueda, crear diagramas en 3D, descargar los diagramas en Excel.

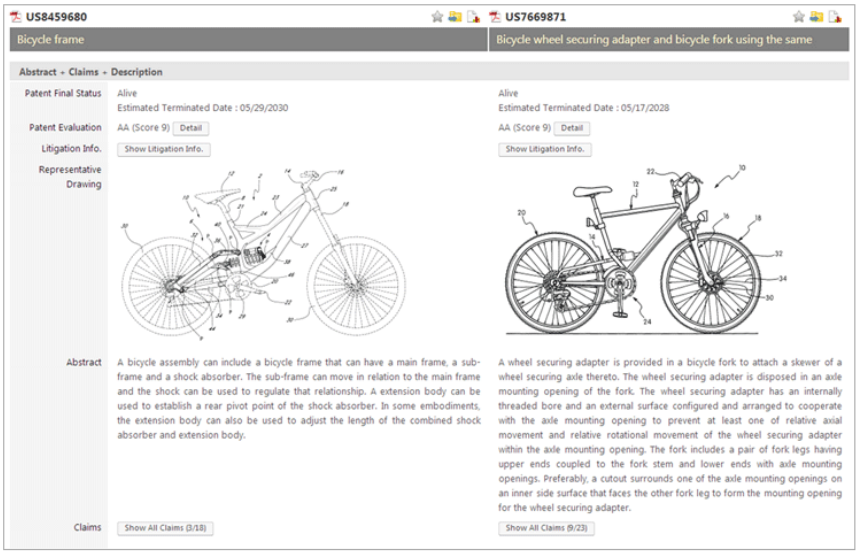


FIGURA 2.4: Comparación de dos patentes en Wisdomain

- **WordStat**[4]

  Es un software fácil y flexible para el análisis de texto, el cual hace uso de la minería de texto para la extracción rápida de temas y tendencias o medición cuidadosa y precisa con herramientas de vanguardia en análisis de contenido cuantitativo.

  Puede ser utilizada por cualquier persona que necesite extraer y analizar rápidamente información de grandes cantidades de documentos. Éste software de análisis de contenido y minería de texto se utiliza para:

  - Inteligencia de negocios y análisis competitivo de sitios web.
  - Extracción de información y descubrimiento de conocimiento desde reportes incidentales, quejas de clientes y mensajes.
  - Análisis de seguimiento informativo o de literatura científica.
  - Etiquetado automático y clasificación de documentos.
  - Detección de fraudes, atribución de autoría, análisis de patente.

[4]http://provalisresearch.com/es/products/software-de-analisis-de-contenido/

WordStat cuenta con la opción de clasificación jerárquica, la cual es una herramienta útil exploratoria para identificar rápidamente temas o grupos de documentos.

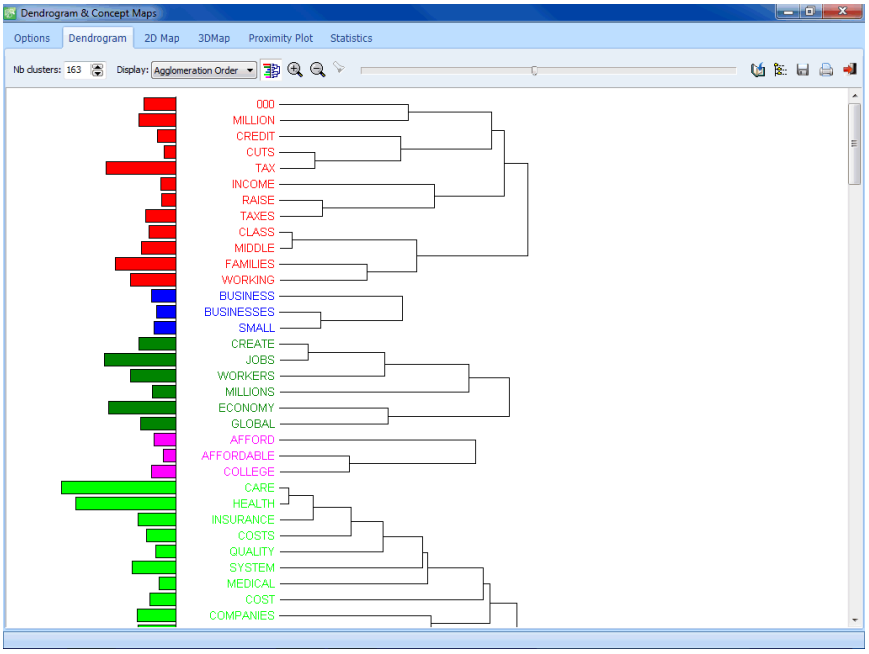


FIGURA 2.5: Ejemplo de clasificación en WordStat

- **MATHEO Patent**[5]

  Gracias a las dos principales fuentes de patentes del mundo (Espacenet[6] y USPTO[7]), Matheo Patent permite un rápido acceso a la información relativa a las patentes masivas.

  Ofrece herramientas de análisis de patentes "orientada al usuario final". Estas herramientas están diseñadas para ayudar a los expertos en su trabajo de análisis. Por lo tanto, puede crear sus propios grupos, añadir una puntuación correspondiente y añadir comentarios individuales sobre cada patente.

  Las patentes pueden ser analizados a través de selección automática de los diferentes tipos de información:

  - Los solicitantes, inventores.
  - Fechas.
  - Tecnologías.
  - Países.
  - Grupos definidos por el usuario.
  - Familia de patentes.

[5]http://www.matheo-software.com/matheo-patent/
[6]http://worldwide.espacenet.com/
[7]http://www.uspto.gov/

- Información legal.

FIGURA 2.6: Diagrama de operación de MATHEO Patent

Con base en los diferentes sistemas que se han descrito anteriormente, se puede evidenciar que existen diversos programas que permiten realizar análisis de patentes haciendo búsquedas por palabras claves, sin embargo, éstos todavía retornan una cantidad excesiva de resultados, además, no permiten realizar una búsqueda por archivo o por el texto (de la patente), el cual es el objetivo del presente trabajo de grado. Así, se ve la necesidad de construir un sistema que permita hacer más efectiva el análisis de patentes.

2.2.2. Técnicas utilizadas para el análisis de patentes

El análisis de patentes ha demostrado jugar un papel importante en la definición de la estrategia de negocio en diversas organizaciones [14], ya que permite identificar nuevas oportunidades tecnológicas. Las patentes contienen grandes volúmenes de datos estructurados y no estructurados, y debido a esto se requieren herramientas que sean lo suficientemente inteligentes como para llevar a cabo las tareas de análisis. No hay una clasificación evidente de técnicas y herramientas utilizadas en el análisis de patentes. Sin embargo, la gran cantidad de literatura sobre éste tema ha utilizado enfoques basados en la minería de texto y de visualización para analizar su contenido [10]. La figura 2.7 muestra las diferentes técnicas que se han utilizado para el análisis de patentes.

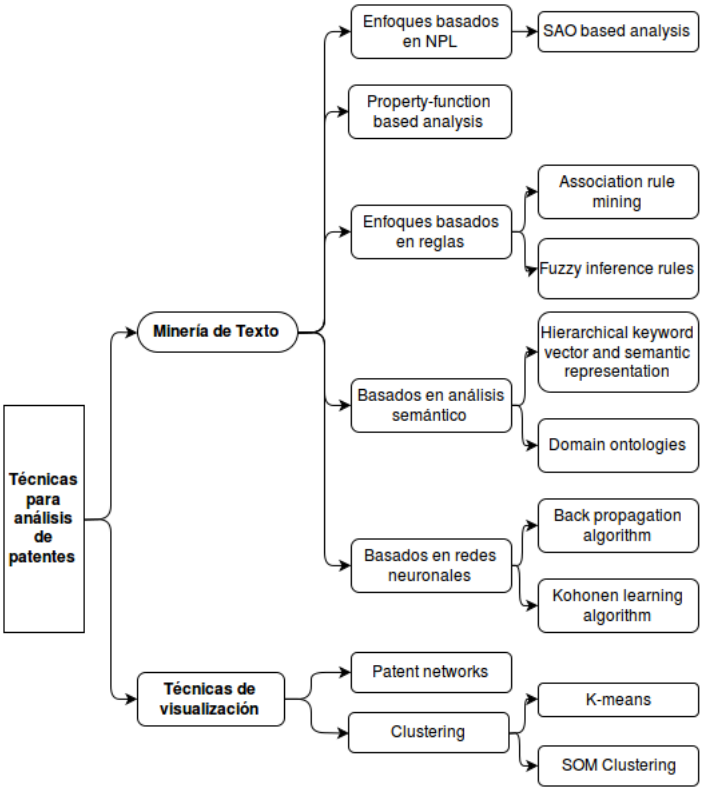


FIGURA 2.7: Técnicas para el análisis de patentes
Tomada de [10]

La investigación realizada en [15] se centra en la estrategia de seleccionar palabras clave para realizar minería de texto. Específicamente, se abordan cuatro factores; *(i)* qué sección de los documentos de patente adoptar para la selección de palabras clave, *(ii)* qué métodos utilizar para la selección de palabras clave, *(iii)* cuántas palabras clave seleccionar y *(iv)* cómo transformar los resultados de la selección de palabras clave en un formato de datos analizables. Los resultados fueron evaluados haciendo uso del algoritmo *K-means*, y se encontró que una selección entre 70 y 130 palabras de un resumen con *Tf-idf* y expresiones booleanas parecen representar la mejor opción para seleccionar las palabras clave, sin embargo, son pocos los estudios que han investigado si el método *Tf-idf* es el más eficaz. La figura 2.8 muestra el resultado de la investigación con las diferentes técnicas utilizadas y el número de palabras claves seleccionadas midiendo el índice de similitud entre 500 patentes.

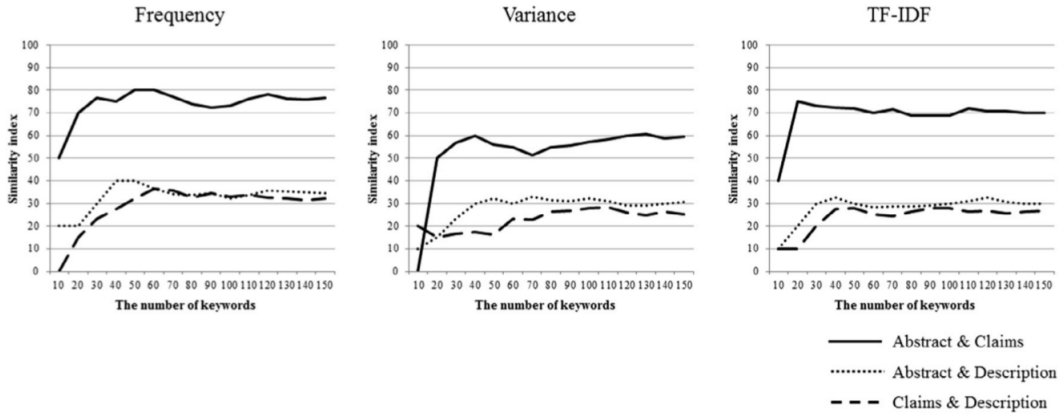


FIGURA 2.8: Similitud entre patentes con diferentes técnicas
Tomada de [15]

Las técnicas de minería de texto basados en la semántica se apoyan en dominios de conocimiento y crean relaciones entre conceptos específicos del dominio. Los tipos de técnicas son eficaces en identificar las similitudes entre las patentes y la determinación de las tendencias tecnológicas por un análisis lógicamente relacional de estructuras gramaticales [10]. Además, permiten determinar las invenciones de alta novedad en los datos de patentes.

En [9], presentan un enfoque basado en análisis semántico para identificar las invenciones de patentes que son altamente novedosas. Para esto, extraen las estructuras semánticas de los datos textuales de la patente. Después se identifican las estructuras semánticas y con análisis lingüístico en particular se relaciona el dominio y la situación de los elementos. Un último paso utiliza la medida de similitud para determinar la novedad de la patente. Este enfoque se avaluó a través de un caso de estudio y los autores afirman que es un método eficaz para identificar patentes altamente novedosas.

En [16] se realiza un estudio para medir la similitud semántica entre patentes y trabajos académicos con VSM *(Vector Space Model)* y LSA *(Latent Semantic Analysis)* usando artículos académicos reales y datos de patentes. Lo primero que hacen es mejorar la precisión en la indexación del documento usando algunas acciones previas básicas de procesamiento. En segundo lugar, los términos juegan un papel importante en los trabajos académicos y en las patentes, un método de reconocimiento de términos es aplicado para entender la semántica del texto. Finalmente, diferentes técnicas son usadas para generar vectores de documentos, como la TF-IDF *(Term frequency – Inverse document frequency)* y así, calcular la similitud entre cada documento y una patente. El estudio demostró que la precisión en la similitud es mejor en el VSM basado en términos y que éstos son más eficientes que los basados en palabras.

En [17], [18], [19] se proponen técnicas para el análisis de patentes basados en ontologías. En [17] se describe un Framework computacional que permite a los usuarios buscar y recopilar información a través de fuentes de información del gobierno y más específicamente en los sistemas de patentes. A través de un caso de estudio en el dominio biomédico, se ilustra cómo este Framework puede ser útil para la búsqueda y recuperación de información de una menera inteligente. En [18], se describe un Framework para expresar y extraer conocimiento de los documentos de patentes para acelerar el diseño de las conexiones de implantes dentales. Los resultados ayudan a los investigadores a comprender rápidamente el conocimiento clave de las conexiones. En [19], se presenta un nuevo enfoque de la búsqueda de patentes que combina el conocimiento semántico y ontologías utilizado para anotar las patentes procesadas con herramientas de procesamiento de lenguaje natural. La arquitectura utiliza reglas de lógica difusa para organizar las patentes anotadas y lograr una recuperación más precisa. En éste, se demostró que los resultados clasificados usando la anotación semántica son mejores que los resultados en función de la frecuencia de palabras en un texto.

La Tabla 2.1 muestra una comparación entre las diferentes técnicas utilizadas en el análisis de patentes, su finalidad y las características clave en la realización del análisis.

Técnica(s)	**Finalidad**	**Característica clave**
NPL, estructuras SAO	Identificación de vacíos y tendencias tecnológicas	Construcción de mapas dinámicos de patentes para extraer relaciones de texto estructurado y no estructurado
Minería de texto basada en SAO	Identificación de patentes prometedoras	Capacidad para manejar grandes datos de patentes y analizarlos automáticamente
Property–function based approach	Identificación de tendencias tecnológicas	No requiere un conjunto predefinido de palabras clave
Hierarchical keyword vectors, tree matching algorithms	Análisis de infracciones	Crear relaciones de dependencia entre los elementos de la patente (claim)
Continúa en la siguiente página		

Técnica(s)	**Finalidad**	**Característica clave**
K-nearest neighbor extraction, Ontology	Clasificar patentes según su relevancia	Clasificación de patentes basada en redes
Back-propagation algorithm	Determinación de la calidad de las patentes para las operaciones de I+D	Minimiza el tiempo para determinar la calidad de la patente de un dominio tecnológico
Clustering using K-means algorithm	Determinar el cambio de tendencias para tecnologías ubicuas	Hace una red semántica de palabras clave para determinar relaciones significativas
Patent Networks, bibliometric patent analysis	Identificación de tendencias tecnológicas	Utiliza gráficas y técnicas cuantitativas para construir redes

Tabla 2.1: Comparación de técnicas para el análisis de patentes

Capítulo 3

Herramienta de apoyo al análisis de patentabilidad

En este capítulo se propone una herramienta para apoyar el análisis de patentes, una aplicaión Web que facilita el tiempo de evaluación en una solicitud, determinando la similitud entre un conjunto de patentes y calculando la novedad de la solicitud presentada. La figura 3.1 muestra la arquitectura general de la aplicación.

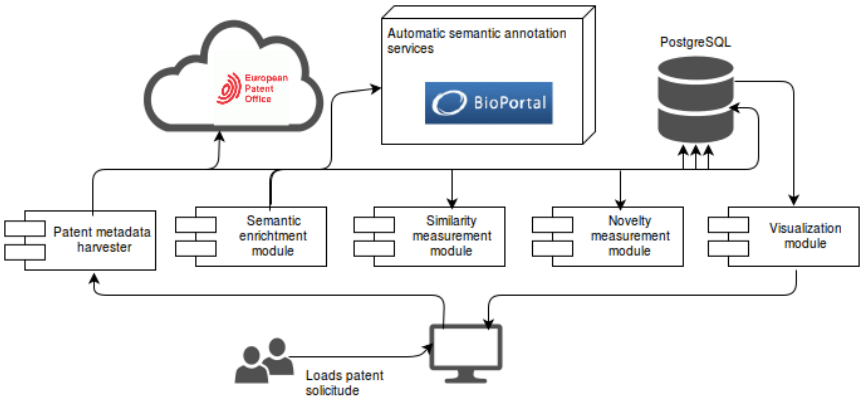


FIGURA 3.1: Arquitectura general de la aplicación

3.1. Obtención de patentes desde bases de datos

Para llevar a cabo el proceso de obtener los datos de las patentes, es necesario realizar peticiones o realizar consultas a bases de datos que permitan dicho proceso. Este es el caso de la *European Patent Office* (EPO[1], por sus siglas en inglés). La actividad principal de esta oficina es examinar las solicitudes de patentes y otorgar las patentes europeas.

[1]http://www.epo.org/index.html

También proporciona servicios de información sobre patentes y servicios de recuperación de datos.

Uno de los servicios Web que ofrece es el *Open Patent Services* (OPS[2]), el cual permite realizar consultas *machine-to-machine* proporcionando datos de patentes. Los datos son extraídos de la base de datos de la EPO. La figura 3.2 muestra el diagrama de flujo entre los servicios de la EPO y la OPS.

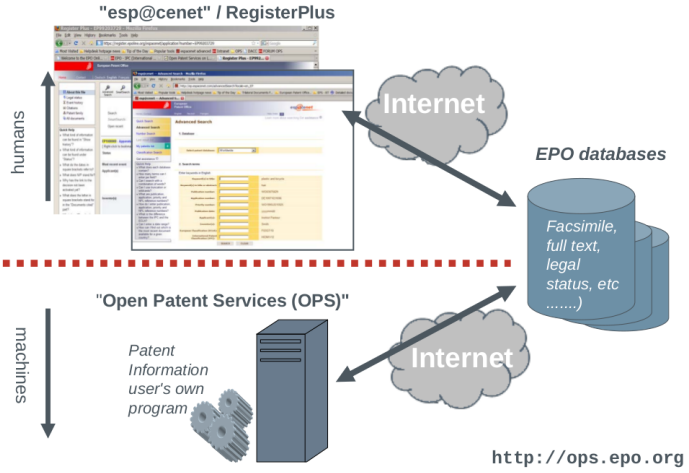


FIGURA 3.2: Diagrama de flujo entre la EPO y OPS

A través de diferentes endpoints (*puntos finales*) el servicio proporciona acceso a datos de patentes en todo el mundo, tales como los *claims, description, images, abstract y datos bibliográficos*. Además, permite hacer búsqueda de patentes por medio de palabras claves.

Todo el servicio está bajo la arquitectura REST, la cual es una tecnología para el intercambio de información entre el cliente y el servidor [20]. REST permite al cliente obtener información del servidor mediante un Web service. De esta manera, la estructura de la URI *(uniform resource identifier)*[3] de una solicitud al servicio OPS se puede construir de la siguiente manera:

```
protocol/authority/[version]/prefix/service/
reference-type/input-
format/input/[endpoint]/[constituent(s)]/output-format
```

FIGURA 3.3: Estructura general de una solicitud en OPS

Un ejemplo de solicitud al servicio se vería de esta forma:

[2]http://www.epo.org/searching-for-patents/technical/espacenet/ops.html
[3]https://www.w3.org/Addressing/

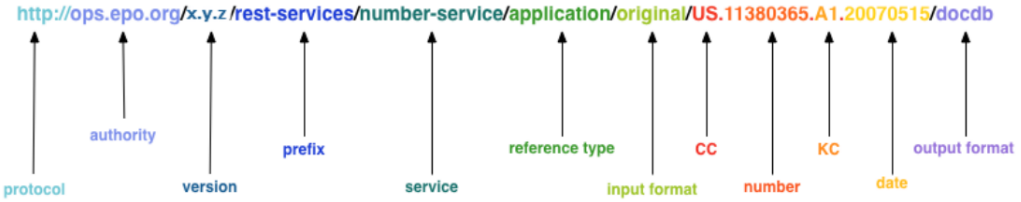


FIGURA 3.4: Petición al servicio de la OPS

La OPS también soporta el método HTTP Post. Esto es específicamente útil si se desea solicitar un gran número de publicaciones a través de recuperación de datos bibliográficos.

Al utilizar los servicios *RESTful* de la OPS es importante saber la estructura de los datos, de tal manera que la solicitud sea la adecuada y se recupere lo que se necesita. Por defecto casi todas las repuestas están en *XML*[4], sin embargo, también es posible obtener un formato *JSON (JavaScript Object Notation)*[5], el cual es un formato ligero, independiente del lenguaje y basado en texto para intercambiar datos.

La estructura genérica de una solicitud de datos publicado se muestra en la figura 3.5. Los parámetros en **negrita** son obligatorias:

```
GET http://ops.epo.org/rest-services/published-data/reference-
type publication or application or priority/input-format docdb or
epodoc/input e.g. EP1000000 or EP1000000.A1/[endpoint e.g.
biblio,abstract,equivalents,fulltext,claims,description,images]
and-or/[constituent e.g. biblio,full-cycle,abstract]?[parameter e.g.
Range=1]

Accept: application/exchange+xml
```

FIGURA 3.5: Estructura genérica de la petición a la OPS

La figura 3.6 muestra un ejemplo para la obtención de datos bibliográficos de una patente.

[4]https://www.w3.org/XML/
[5]http://www.json.org

```
GET http://ops.epo.org/rest-services/published-
data/publication/epodoc/EP1000000.A1/biblio

Accept: application/exchange+xml
```

```
POST http://ops.epo.org/rest-services/published-
data/publication/epodoc/biblio

Accept: application/exchange+xml
Request Body: EP1000000.A1
```

FIGURA 3.6: Estructura genérica de la petición a la OPS

Como se puede ver, la figura 3.6 muestra las dos formas de hacer la petición, una con el método *GET* y el otro con el método *POST*. Esta consulta retorna un archivo *XML*, con la información bibliográfica de la patente identificada con el número 1000000.

Para el caso de hacer una búsqueda de patentes sobre un tema en específico, la petición sería la misma a la que se muestra en la Figura 3.6, y lo único que cambiaría sería la *URI*. Por ejemplo, si se le pasa la siguiente URI: *http://ops.epo.org/3.1/rest-services/published-data/search/biblio?q=ta %3Dlung,cancer*, daría como resultado un archivo XML con las primeras 25 patentes encontradas en la búsqueda, relacionadas con las palabras *lung* y *cancer*. Si se desea obtener más de 25 patentes en una petición, lo que se hace es concatenar al final del enlace *&Range=1-100* y así, el resultado sería las primeras 100 patentes.

Sin embargo, el servicio de la OPS presenta ciertas limitaciones. Una de ellas es al momento de realizar una búsqueda de patentes por palabras clave, ya que si la respuesta a la petición tiene más de 2000 resultados, a partir de la patente en la posición 2001 en adelante no es posible acceder a ellas.

3.2. Anotación semántica con múltiples ontologías

Para realizar la anotación semántica se utiliza el servicio Web de *BioPortal*[6], el cual es un repositorio de ontologías biomédicas de libre acceso. Este repositorio cuenta con más de 300 ontologías que contienen alrededor de 4'867.799 términos del entorno biomédico [21].

BioPortal cuenta con unos componetes principales: la aplicación Web, los servicios API y miniaplicaciones que pueden ser instaladas en un sitio web. Entre los servicios que

[6]http://bioportal.bioontology.org

ofrecen estos componentes están: un buscador de ontologías, un buscador de términos, un recomendador de ontologías, anotación de texto, un buscador de recursos biomédicos, entre otros.

Para acceder a los servicios Web de *BioPortal* se hace uso de la *API REST*[7] que éste ofrece. Para esto, es necesario una *Api Key* para desarrolladores, la cual se obtiene al momento de crear una cuenta en la página de *BioPortal*.

Al hacer una petición al servicio Web, ésta nos retorna un *JSON* o un *XML* con las anotaciones y las ontologías del texto que se le pasa como parámetro. Estas ontologías son devueltas con una valoración numérica la cual indica la importancia de dicha ontología en el proceso de anotación.

Por otra parte, el servicio Web presenta algunas limitaciones o inconvenientes. Si se desea anotar un gran número de palabras o un texto demasiado extenso, el tiempo de respuesta de este servicio aumenta. Otro inconveniente es que si se realizan muchas peticiones al mismo tiempo, el servicio Web bloquea al usuario *(identificado con la Api Key)* por un determinado tiempo y en algunas ocasiones el servidor de BioPortal no está disponible, aunque esto ocurre por períodos de tiempo muy cortos.

La tabla 3.1 muestra los parámetros necesarios para realizar una petición:

Parámetro	**Descripción**
Apikey *(char)*	Una API Key es necesaria para acceder a cualquier llamado a la API
Input *(char)*	Texto que se desea anotar en el servicio Web.
mínimum_match_length *(int)*	Tamaño mínimo de una palabra que se quiere anotar.
exclude_numbers *(boolean)*	Permite excluir los números que hay en el texto para que no sean anotados.
longest_only *(boolean)*	Si se desea anotar sólo texto largo y no palabras sueltas.
Continúa en la siguiente página	

[7] http://data.bioontology.org/documentation

Parámetro	Descripción
exclude_synonyms *(boolean)*	Permite que se anote sólo una vez alguna palabra que tenga uno o más sinónimos dentro del texto que se va anotar, esto permite que no se hagan anotaciones repetidas por causa de sinónimos de las palabras que ya se anotaron

TABLA 3.1: Parámetros para una solicitud al servicio de *BioPortal*

3.3. Clustering y cálculo de la similitud entre patentes

Para el proceso de *Clustering* entre las patentes se usa el algoritmo *K-means*. En este caso, lo que se realiza es tomar las palabras anotadas que retorna el proceso de anotación semántica y con ellas construir una matriz de características *(o matriz de puntos)*. La codificación de los documentos en vectores se llama indexación [22]. La figura 3.7 muestra un esquema sobre el proceso de indexación.

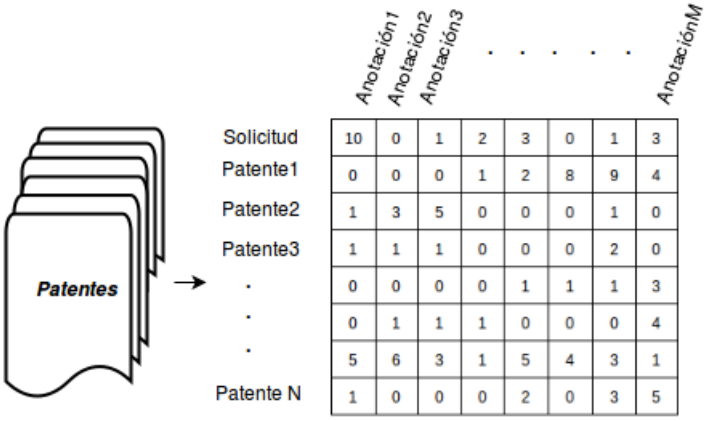


FIGURA 3.7: Proceso de indexación

La matriz de características se puede construir de diferentes maneras. Una de ellas es que cada una de las palabras anotadas se convierta en una columna de la matriz. Después se procede a identificar si una palabra correspondiente aparece en el documento de la patente o no. Al final queda una matriz de 1's y 0's, donde *1* significa que la palabra está en el documento y *0* en caso contrario.

Otra forma de construir la matriz es obteniendo la *frecuencia bruta* de cada palabra o término anotado t en el documento de la patente, es decir, el número de veces que

el término t aparece en la patente. Igual a la forma anterior, cada una de las palabras anotadas se convierte en una columna de la matriz y ésta es creada con los valores de la *frecuencia bruta*.

Finalmente, una tercera forma para crear la matriz de características es usar la medida Tf-idf *(frecuencia de término–frecuencia inversa de documento)*[8], la cual proporciona un modelo simple para evaluar la relevancia de las palabras clave dentro de un corpus o una colección de documentos. Aunque la medida Tf-idf es un factor de ponderación relativamente antigua, es ampliamente utilizado en la recuperación de información y minería de texto [23], es por este motivo que se decide hacer uso de la técnica *Tf-idf* para realizar parte del desarrollo de este proyecto.

Después de haber construido la matriz de características, ésta se le pasa como parámetro al algoritmo *K-means*, el cual también requiere otro parámetro que es el número de clusters k que se desea tener. Para determinar el valor de k lo que se hace es un análisis de silueta *(Silhouette Analysis)*, una medida que nos permite analizar la distancia de separación entre los clusters o agrupaciones resultantes. Esta medida tiene un rango entre [-1,1], donde un valor alto indica que la muestra está muy lejos de los grupos vecinos. Un valor de 0 indica que la muestra está sobre o muy cerca de la frontera de decisión entre dos grupos vecinos y los valores negativos indican que esas muestras podrían haber sido asignados al grupo incorrecto. Para la implementación del algoritmo *K-means* y para el análisis de silueta se hace uso de la librería *scikit-learn* [24] del lenguaje Python[9].

Para el caso de determinar la similitud entre patentes con un valor numérico se opta por determinar el *Cosine similarity*, la cual es una medida que permite calcular el valor de la similitud haciendo uso del vector de características; es comúnmente usado en la recuperación de información, y es una métrica de orientación y no de magnitud [25]. Esta medida proporciona un valor igual a 1 si el ángulo comprendido es cero. Cualquier ángulo existente entre los vectores, el coseno arrojaría un valor inferior a uno. Si los vectores fuesen ortogonales el coseno se anularía, y si apuntan en sentido contrario su valor sería -1. De esta forma, el valor de esta métrica se encuentra entre -1 y 1 [26]. La siguiente ecuación representa la fórmula para hallar el valor del coseno.

$$\cos(\theta) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \, \|\vec{b}\|}$$

[8] http://www.tfidf.com
[9] https://www.python.org

3.4. Medición de la novedad de una solicitud de patente frente a un conjunto de patentes similares

La medición de la novedad tal vez sea el proceso más complejo que se aborda en este trabajo de grado debido a que es considerado algo muy subjetivo. De este modo, no son muchas las investigaciones que se han llevado a cabo para medir la novedad, sin embargo, se pueden mencionar algunas propuestas como [9, 11, 27].

Así, para medir la novedad se selecciona parte del modelo propuesto en [9], la cual considera la novedad y la no-evidencia como un fenómeno conjunto, donde la *no-evidencia* podría ser útil como un efecto moderador de la relación entre la novedad de una patente y su impacto sobre otras variables como el valor de la patente. Esto es importante dado que se desea determinar patentes con un alto grado de innovación y no fenómenos similares de lo que ya existe o esté patentado.

El enfoque propuesto en [9] se puede resumir en los siguientes pasos:

1. Identificar y extraer las estructuras semánticas de las patentes.
2. Especificar el dominio de interés con un análisis lingüistico.
3. Medir la similitud basado en la importancia de las estructuras extraídas.
4. Calcular la novedad haciendo uso de la matriz de similitud.

El último paso es el proceso que determina la novedad de la patente sobre la base de una matriz de similitud y es la única parte del modelo que se selecciona e implementa en el presente trabajo de grado, debido a que la forma en que se calcula la similitud entre las patentes es diferente.

Dado que la novedad sólo puede determinarse en relación a las tecnologías establecidas o ya existentes, entonces el cálculo de la novedad se basa en la similitud entre una solicitud de patente y un conjunto de patentes. La máxima similitud de una patente con respecto a otras patentes anteriores puede ser considerada como su "vejez" *(oldness)*, esto quiere decir que se considera la novedad como una proporción particular de una patente que no se parece a patentes anteriores en términos de similitud. Por lo tanto, se calcula la novedad de una patente restando la máxima similitud de una patente anterior de uno. La siguiente ecuación representa la fórmula para determinar el valor de la novedad.

$$N_i = 1 - max(s_{i(n)}), \text{para todo n} < \text{i}$$

Donde N_i es la novedad de la patente i y $s_{i(n)}$ es la similitud de la patente i para cada patente n presentada antes de la patente i.

De esta manera, se realizan los siguientes pasos para el funcionamiento del sistema:

1. Se ingresan los datos de la patente a analizar y sus keywords.
2. El sistema guarda la patente y hace una búsqueda en la base de datos de la EPO usando como criterio de búsqueda las keywords ingresadas.
3. Una vez se obtienen las patentes desde la EPO, se guardan los metadatos y el abstract de cada una.
4. Se utilizan los abstracts obtenidos en el paso anterior para hacer anotaciones usando el servicio web RECOMMENDER de BIOPORTAL.
5. Con las anotaciones obtenidas se inicia el análisis de patentabilidad construyendo la matriz de características.
6. Con la matriz de características se aplica el algoritmo k-means y se determina la similitud entre las patentes.
7. Con el resultado de la similitud entre las patentes, se selecciona la mayor medida de similitud y se determina la novedad de la solicitud.
8. Finalmente se visualiza el gráfico del clúser obtenido y se presenta en una tabla la medida de similitud entre las patentes del mismo cluster de la patente analizada y su nivel de novedad.

3.5. Prototipo de software

Al ingresar a la aplicación se despliega la interfaz principal del sistema, la cual consta de un formulario que permite registrar los datos de la solicitud de patente que se desea analizar. La figura 3.8 muestra la forma de la interfaz principal.

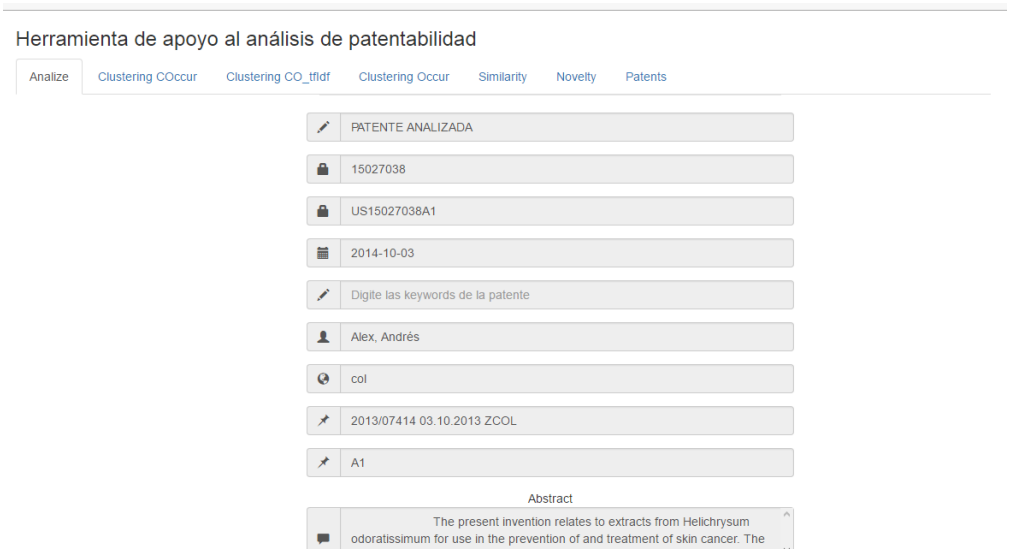


FIGURA 3.8: Interfaz principal del sistema

Después que el usuario haya ingresado los datos de la solicitud de patente, el sistema puede empezar a realizar el proceso de análisis. Al finalizar dicho proceso, el usuario podrá visualizar los diferentes resultados que retorna el sistema, entre los cuales están: la lista de patentes que se están analizando, la similitud entre el conjunto de patentes, el resultado de la novedad y el gráfico del clustering entre las patentes.

La Figura 3.9 muestra la interfaz con el resultado del análisis de la similitud.

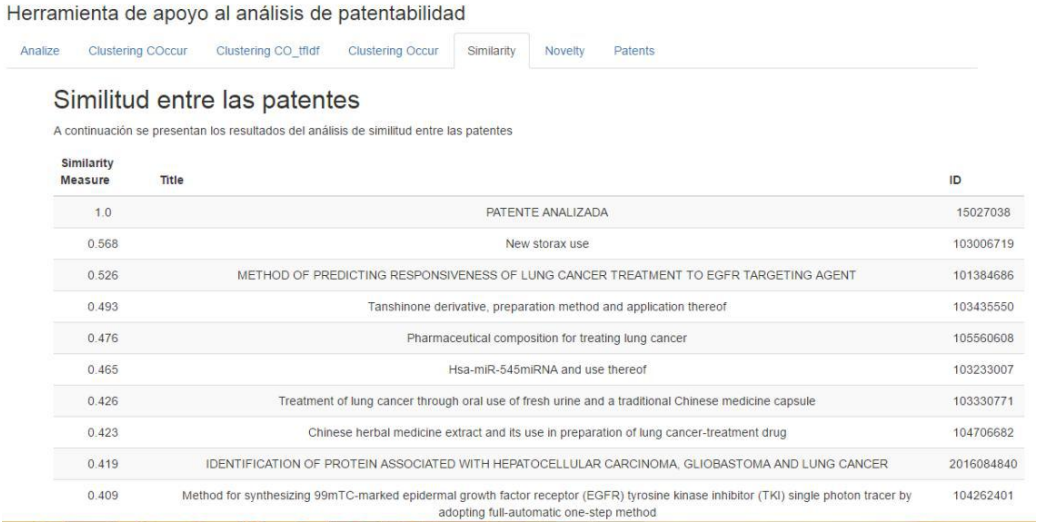


FIGURA 3.9: Interfaz de la similitud

La figura 3.10 muestra la forma en que se visualiza el resultado del clustering, en la cual fueron analizadas 661 patentes

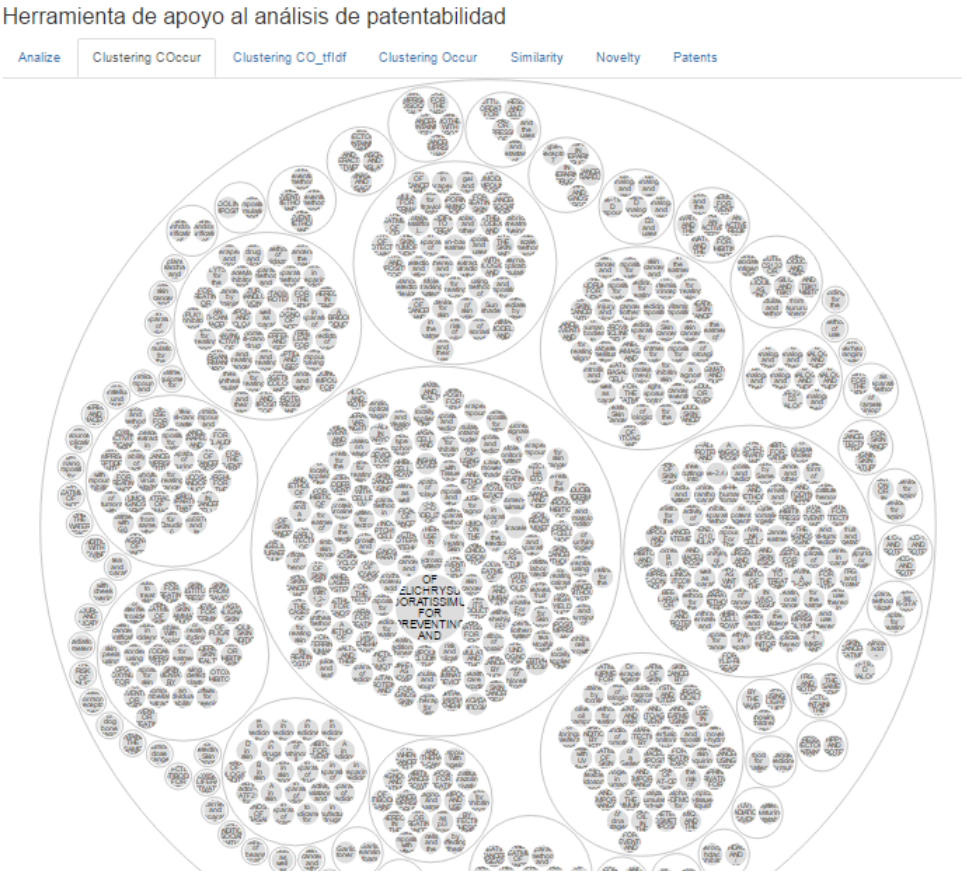


FIGURA 3.10: Interfaz del clustering

Capítulo 4

Aspectos del desarrollo de software

En este capítulo se expone el proceso de desarrollo y se hace la respectiva documentación. También se presentan las diferentes tecnologías usadas en la implementación de la herramienta.

La metodología de desarrollo seleccionada para llevar a cabo el presente trabajo de grado fue *XP (eXtreme Programming)*, la cual pertenece a las metodologías ágiles. *XP* enfatiza en el trabajo en equipo, el cual permite crear un entorno sencillo, pero eficaz, logrando un alto nivel de productividad y su autoorganización permite resolver problemas de forma eficiente. Además, no requiere un gran número de desarrolladores y propone el principio de "Programación en parejas".

4.1. Planeación

El primer paso para desarrollar un proyecto bajo la metodología *XP* es la planificación, donde los desarrolladores y las partes interesadas definen los requisitos y funcionalidades del proyecto, las cuales son detalladas en historias de usuario. Estas historias se plantearon de forma modular, donde cada módulo y su historia de usuario dan cumplimiento a los diferentes requisitos. Para un desarrollo de forma incremental los módulos se clasifican de básicos a complejos, así se logra una jerarquización en los módulos y se implementan los más simples y se realizan iteraciones para elaborar los módulos más complejos.

4.1.1. Especificación de requerimientos

Para determinar el alcance de la herramienta se define una serie de requerimientos, los cuales darán apoyo a las historias de usuario. La tabla 4.1 muestra los requerimientos funcionales del módulo de usuario.

Universidad del Valle	Especificación de requerimientos	Revisión: 001
Sistema para el apoyo al análisis de patentabilidad		Fecha
1. Módulo de usuario		
Requerimiento		Descripción
Requerimiento 1.1		El sistema en el módulo de usuario deberá contener un formulario en el cual se pueda ingresar los datos de la patente a analizar.
Requerimiento 1.2		El sistema en el módulo de usuario deberá contener un botón "Análizar" que iniciará el análisis de la patente.

Tabla 4.1: Requerimientos funcionles, módulo de usuario

El documento completo con los requerimientos están en el Anexo A.

4.1.2. Historias de usuario

Ahora que se tienen definidas las funcionalidades del sistema, se procede a redactar las historias de usuario. Dado que la herramienta propuesta busca optimizar el proceso de análisis de patentes, no son muchas las historias que se generan.

La tabla 4.2 muestra la historia de usuario que se generó a partir de los requerimientos.

	**Historia de usuario: Módulo Usuario**
Sistema para el apoyo al análisis de patentabilidad	**Fecha: 20/02/2016**
Título	Descripción
COMO usuario QUIERO analizar una solicitud de patente	El usuario ingresa a la aplicación y debe diligenciar los datos de la solicitud de patente (los campos a ingresar se pueden ver en el Anexo B). **Criterio de aceptación:** el sistema mostrará los metadatos de las patentes analizadas, el resultado del clustering de las patentes, la medición de la similitud y la de novedad.

Tabla 4.2: Historia de usuario: Módulo usuarios

4.1.3. Planeación de iteraciones

Con toda la documentación que se tiene, requerimientos e historias de usuario, se da inicio a la planeación de iteraciones de acuerdo a la metodología *XP*. Dado que *XP* permite a los desarrolladores adaptarse a las necesidades cambiantes que se pueden presentar en el trancurso del desarrollo, algunos de los requerimientos y sus historias de usuario fueron cambiadas para ajustar las funcionalidades del sistema. En total se realizaron 5 iteraciones. En la tabla 4.3 se puede observar la iteración 1.

	**Iteración 1. Revisión: 001**
Sistema para el apoyo al análisis de patentabilidad	**Fecha:**
Descripción	H-U/Requerimiento
En esta iteración se crea el modelo de la base de datos y se determinó el servicio Web para obtener los metadatos de las patentes, posteriormente se crea las consultas necesarias para acceder a dicho servicio	Se seleccionan los requerimientos 3.1 - 3.2

Tabla 4.3: Descripción de la iteración 1

En el anexo C se muestran las demás iteraciones que se llevaron a cabo.

4.2. Arquitectura

Para el desarrollo de la aplicación se implementa la arquitectura multicapa cliente-servidor, también conocida como Arquitectura en tres capas. Esta arquitectura está compuesta por la capa del cliente, es decir el equipo que solicita los recursos, equipado por una interfaz de usuario; la capa del servidor de aplicaciones, cuya tarea es proporcionar los recursos solicitados por el cliente y la capa de datos que se encarga de almacenar y mantener la integridad de la información. La figura 5.5 muestra la arquitectura sobre la cual se soporta la herramienta mediante el uso del diagrama de componentes. Dado que *XP* es una metodología ágil no se consideran necesarios más artefactos *UML*.

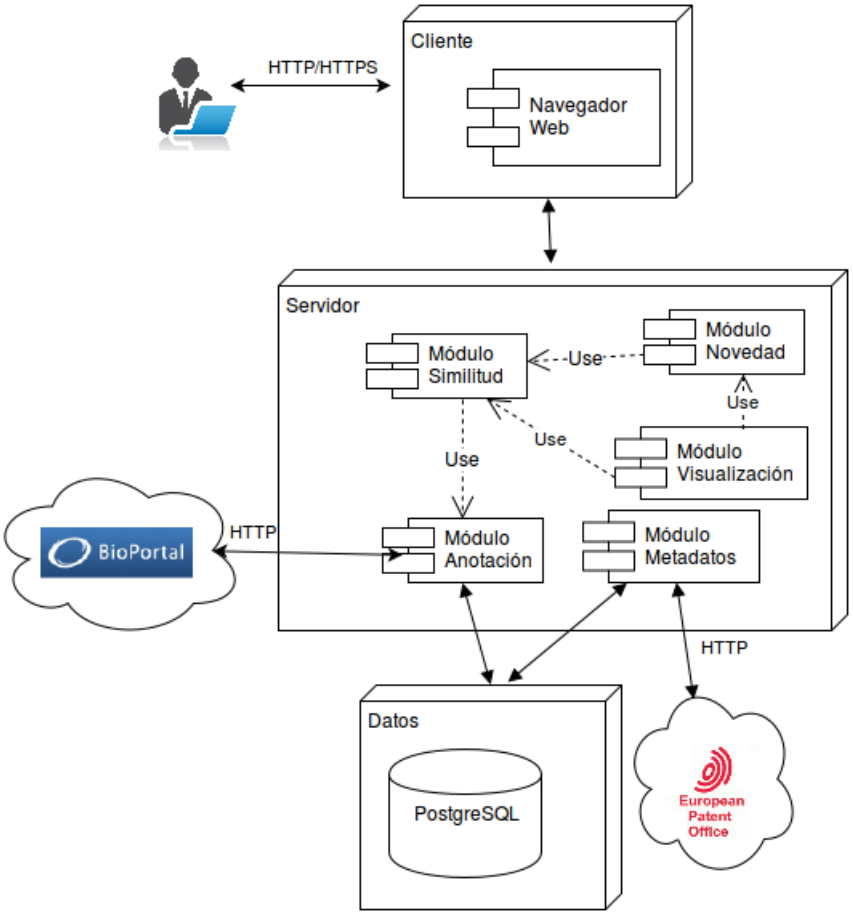


FIGURA 4.1: Diagrama de componentes

La capa de presentación *(cliente)* está construida con la librería D3 (*Data-Driven Documents*)[1] de *JavaScript*, la cual produce visualizaciones dinámicas en el navegador Web

[1]https://d3js.org

haciendo uso de HTML, SVG y CSS. La capa de presentación y la capa de servicios se comunican bajo la arquitectura *REST*, encargada de direccionar las peticiones de la capa de presentación a los diferentes módulos. Por su parte, la capa de datos tiene la responsabilidad de interactuar con la capa de servicios y mantener la información almacenada.

4.3. Codificación

La metodología *XP* propone que al final de cada iteración se genere una versión, módulo o componente de software.

La codificación de la aplicación (lado servidor) se hizo bajo el lenguaje de programación Python[2], cuya filosofía hace hincapié en una sintaxis que favorezca un código legible. También se hizo uso del editor de texto Sublime Text[3] en su versión gratuita, para la implementación del código. Para obtener las patentes se usa el servicio Web de la EPO, Las anotaciones se obtuvieron mediante el servicio de Bioportal. En el proceso de clustering y el de análisis de similitud se hizo mediante la librería scikit-learn, la cual cuenta con una gran variedad de clases y funciones para realizar data mining y análisis de datos. Esta librería es Open source y está bajo la licencia BSD *(Berkeley Software Distribution)*. Por otro lado, se hace uso del servidor propio de Flask como contenedor Web para probar y desplegar la aplicación Web.

Para el desarrollo de la capa de presentación, en la cual se implentaron las diferentes interfaces gráficas de usuario, se hizo uso del Microframework para Python Flask[4], basado en *Werkzeug*[5] y *Jinja2*[6]. También se hace uso de HTML[7] y *Bootstrap*[8], el cual es un framework para dar estilo a las aplicaciones Web. Para generar los distintos gráficos que se muestran después de realizar los procesos de miniería de texto, se usa la librería D3 *(Data-Driven Documents)* de JavaScript.

Además, para llevar a cabo esta etapa de codificación, se hizo necesario un sistema de control de versiones para la integración del código entre los diferentes desarrolladores. En este caso se elige el sistema *Bitbucket*[9], el cual permite la colaboración entre el equipo de desarrollo.

[2]https://www.python.org
[3]http://www.sublimetext.com
[4]http://flask.pocoo.org
[5]http://werkzeug.pocoo.org
[6]http://jinja.pocoo.org
[7]https://www.w3.org/html/
[8]http://getbootstrap.com
[9]https://bitbucket.org/product

El Anexo D contiene el instructivo para instalar la aplicación.

Capítulo 5

Pruebas

5.1. Selección de datos de prueba

Para el desarrollo de las pruebas se seleccionaron diversos datos con el objetivo de medir el comportamiento del sistema. Para ésto, se analizaron patentes con un tema en común y otra donde la relación entre las patentes era muy poca.

Después de realizar el proceso anterior, se hace un análisis con una patente cuyo estado legal es **"REFUSED"**. Esta patente se obtuvo del servicio *Patent Information Services for Experts*[1] ofrecido por la EPO. Es importante aclarar que dichas patentes no cuentan con un motivo por el cual haya sido rechazada.

5.2. Descripción del conjunto de datos

En primer lugar se decide probar el comportamiento de los algoritmos que determinan la similitud y la novedad. Para este caso, se realizan dos pruebas, comparando una solicitud de patente con una patente similar y otra comparación con una patente poco similar.

Como solicitud de patente a analizar se selecciona una patente de la base de datos de la *United States Patent and Trademark Office*: **Compositions and Methods to Inhibit EZH2 for the Treatment of Cardiovascular Diseases**, cuyo tema es el tratamiento de enfermedades cardiovasculares.

Las otras dos patentes con las cuales se hizo la comparación fueron: **Heat Dissipation Device and Control Method Thereof**, cuyo país de origen es Taiwan y **Use of**

[1] https://data.epo.org/expert-services/index-2-3-1.html

PCSK9 and LDL-R Activity For Treating Cardiovascular Risk, patente de los Estados Unidos. Éstas patentes fueron tomadas de la base de datos de la EPO.

A continuación se presenta el *abstract* de cada una de las patentes.

Compositions and Methods to Inhibit EZH2 for the Treatment of Cardiovascular Diseases *... "The present invention relates to compositions and methods for treatment and/or prevention of a cardiovascular disease. In one embodiment, the invention provides compositions and methods for decreasing one or more of the level, production, and activity of EZH2."*

Heat Dissipation Device and Control Method Thereof *... "A heat dissipation device including a driving module and a heat dissipation sheet is provided. The heat dissipation sheet has a connection portion and a swing portion and is connected to the driving module by the connection portion. When the driving module receives an input voltage to drive the connection portion, the connection portion drives the swing portion to swing back and forth so as to generate a heat dissipation airflow."*

Use of PCSK9 and LDL-R Activity For Treating Cardiovascular Risk *... "The present disclosure provides methods of assessing cardiovascular risk in a subject and/or of treating a subject having or at risk of developing a cardiovascular disease or disorder. In some embodiments, the method comprises determining an activity level of PCSK9 and/or a level of PCSK9 in a sample obtained from the subject, and initiating or modifying a treatment regimen."*

5.3. Ejecución de las pruebas

La siguiente tabla muestra el resultado del análisis entre las patentes similares. En ella se puede observar las palabras que fueron anotadas por el servicio de *Bioportal*, el número de ocurrencia de cada palabra en el texto y el valor de la similitud que se obtuvo.

Patentes similares: Similitud 0.887 usando el algoritmo tf-idf			
Compositions and Methods to Inhibit EZH2 for the Treatment of Cardiovascular Diseases		**USE OF PCSK9 AND LDL-R ACTIVITY FOR TREATING CARDIOVASCULAR RISK**	
Anotaciones	**Cantidad de ocurrencias**	**Anotaciones:**	**Cantidad de ocurrencias**
ACTIVITY	11	ACTIVITY	11
AND/OR	1	ACTIVITY LEVEL	1
CARDIOVASCULAR	1	AND/OR	2
CARDIOVASCULAR DISEASE	18	AT RISK	2
DECREASING	2	CARDIOVASCULAR	6
DISEASE	4	CARDIOVASCULAR DISEASE	16
EZH2	4	CARDIOVASCULAR RISK	1
INVENTION	2	DEVELOPING	1
LEVEL	6	DISCLOSURE	2
METHODS	6	DISEASE	2
PRESENT	7	DISEASE OR DISORDER	1
PREVENTION	6	DISORDER	4
PREVENTION OF	1	LEVEL	11
PRODUCTION	5	METHOD	6
TREATMENT	12	METHODS	3
		OBTAINED	1
		PCSK9	12
		PRESENT	8
		RISK	9
		RISK OF	1
		SAMPLE	3
		SAMPLE OBTAINED	1
		SUBJECT	6
		TREATING	2
		TREATMENT	8

FIGURA 5.1: Comparación entre patentes similares

En la siguiente tabla se puede ver el resultado del análisis entre las patestes diferentes.

Patentes distintas: Similitud: 0.0 con el algotritmo tf-idf			
Compositions and Methods to Inhibit EZH2 for the Treatment of Cardiovascular Diseases		**HEAT DISSIPATION DEVICE AND CONTROL METHOD THEREOF**	
Anotaciones	**Cantidad de ocurrencias**	**Anotaciones:**	**Cantidad de ocurrencias**
ACTIVITY	11	AIRFLOW	1
AND/OR	1	CONNECTION	12
CARDIOVASCULAR	1	DEVICE	7
CARDIOVASCULAR DISEASE	18	DISSIPATION	40
DECREASING	2	DRIVE	4
DISEASE	4	DRIVES	1
Continúa en la siguiente página			

FIGURA 5.2: Comparación entre patentes distintas

Patentes distintas: Similitud: 0.0 con el algotritmo tf-idf			
Compositions and Methods to Inhibit EZH2 for the Treatment of Cardiovascular Diseases		**HEAT DISSIPATION DEVICE AND CONTROL METHOD THEREOF**	
Anotaciones	**Cantidad de ocurrencias**	**Anotaciones:**	**Cantidad de ocurrencias**
EZH2	4	DRIVING	9
INVENTION	2	FORTH	2
LEVEL	6	GENERATE	2
METHODS	6	HEAT	60
PRESENT	7	HEAT DISSIPATION	32
PREVENTION	6	INCLUDING	1
PREVENTION OF	1	INPUT	11
PRODUCTION	5	MODULE	6
TREATMENT	12	PORTION	6
		PROVIDED	1
		SHEET	4
		VOLTAGE	3

FIGURA 5.3: Comparación entre patentes distintas

Se selecciona la patente **: Methods and compositions concerning poxviruses and cáncer** cuyo estado legal es "REFUSED". El resultado del análisis se puede ver a continuación:

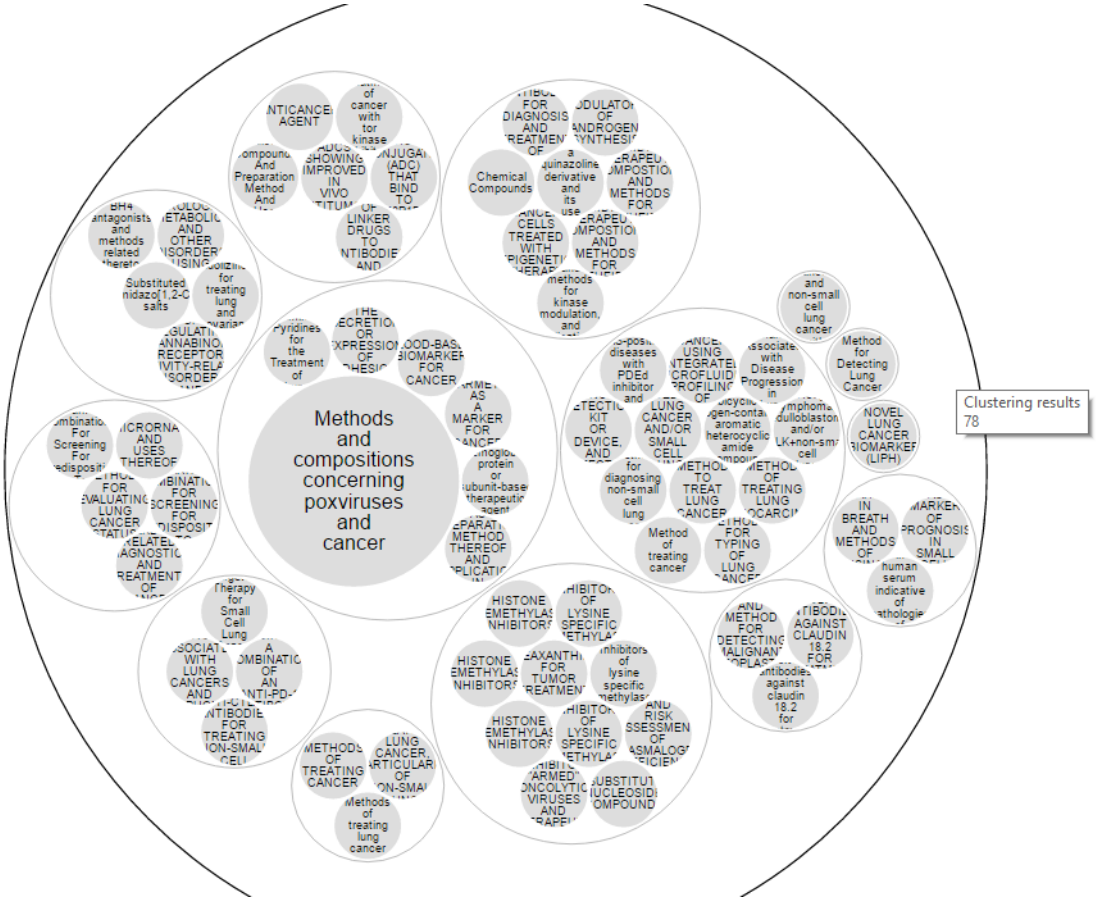


FIGURA 5.4: Análisis de patente rechazada

El resultado del análisis fue:

Similarity Measure	Title	Patent
0,84	Pharmaceutical composition comprising recombinant hemoglobin protein or subunit-based therapeutic agent for cancer targeting treatment	2015283394
0.83	METHODS AND COMPOUNDS FOR MODULATING THE SECRETION OR EXPRESSION OF ADHESION PROTEINS OR ANGIOPOIETINS OF CELLS	2016354397
0,76	ARMET AS A MARKER FOR CANCER	2016377625
0,75	BLOOD-BASED BIOMARKER FOR CANCER	2016370379
0.63	PERIPLANETA AMERICANA EXTRACT OR PERIPLANETA AMERICANA MEDICINAL POWDER AS WELL AS PREPARATION METHOD THEREOF AND APPLICATION IN PREPARATION FOR MEDICINE USED FOR PREVENTING AND TREATING RADIATION-INDUCED DAMAGES	2017000829
0,18	Fused Amino Pyridines for the Treatment of Lung Cancer	2016317508

FIGURA 5.5: Resultado de similitud en patente rechazada

Como se puede observar, el análisis realizado por el sistema determina que hay una patente muy similar, con un valor del 84 %. Dicha patente es *"Pharmaceutical composition comprising recombinant hemoglobin protein or subunit-based therapeutic agent for cancer targeting treatment"*.

Debido a que se encontró una patente muy similar, el nivel de novedad de la patente analizada es bajo.

Novelty Measure	**Title**	**Patent**
0,159	Pharmaceutical composition comprising recombinant hemoglobin protein or subunit-based therapeutic agent for cancer targeting treatment	2015283394

TABLA 5.1: Resultado de la novedad en patente rechazada

Capítulo 6

Conclusiones y trabajos futuros

En este capítulo se presentan las conclusiones del trabajo propuesto con respecto a los objetivos cumplidos y también al proceso de desarrollo del sistema. Finalmente se mencionan algunas líneas de trabajo futuro.

6.1. Conclusiones

En este trabajo se ha presentado una herramienta de software que soporta el análisis de solicitudes de patentes con el fin de disminuir el tiempo dedicado a su evaluación. Para lograr ésto, se hace uso de los diferentes mecanismos de las Ciencias de la Computación. La herramienta permite medir la similitud y la novedad de la solicitud entre las diferentes patentes encontradas en la bases de datos de la EPO *(European Patent Office)*. Además, realiza clustering sobre los datos de las patentes con el fin de facilitar el análisis de forma visual. Está basada en una arquitectura modular lo que hace posible extender las funcionalidades de la herramienta.

En el trabajo de grado se propusieron los siguientes objetivos específicos, de los cuales se concluye:

1. **Revisar la literatura de las diferentes técnicas de las Ciencias de la Computación aplicadas al proceso de análisis de patentabilidad.**

   El proceso de investigación sobre la literatura aplicada al análisis de patentes se llevó a cabo principalmente accediendo a la base de datos en línea de la Universidad del Valle[1].

[1] http://bd.univalle.edu.co/login

Gracias a la base de datos de la Universidad del Valle se logró obtener la información de importantes artículos científicos, lo que permitió evidenciar que han sido diversas las técnicas utilizadas para analizar documentos de patentes. Entre ellas están las que se enfocan en un análisis semántico y otras en seleccionar palabras clave. Las de análisis semántico utilizan ontologías de dominio para realizar el proceso de análisis de similitud. Las que utilizan palabras clave se basan en modelos matemáticos para determinar la relevancia de las palabras en el texto. Además, hay pocos ejemplos de aproximaciones en las que se propone el clustering de patentes usando anotaciones provenientes de ontologías de dominio.

2. **Implementar un algoritmo que permita determinar qué tan similares son dos documentos.**

   Existen múltiples formas de determinar la similitud entre documentos. En este trabajo se utilizó la distancia coseno entre los vectores de caracterísitcas construidos a partir de las anotaciones semánticas provenientes de *Bioportal* sobre los abstracts y los claims de las patentes analizadas. Esta aproximación se vio limitada por la disponibildiad de datos proveientes de la *EPO* y del servicio de anotación de *Bioportal*.

   Inicialmente se estaban anotando el *abstract* y los *claims* de las patentes, pero en el transcurso del desarrollo el servicio de la EPO presentó problemas para acceder a los *claims*, debido a esto sólo se pudo obtener el abstract de las patentes. Por otro lado, dado que el servicio de la EPO está en constante actualización se presentaba el caso de que algunas patentes no contaban con el abstract, de este modo, no era posible analizarla y por ende se descartaba.

   El servicio de *BioPortal*, utilizado para realizar las anotaciones también presentó algunas limitaciones, dado que este no permite anotar un texto que contenga más de 500 palabras; cuando ésto ocurria, lo que se hacía era dividir el texto de la patente. Además, el tiempo que tomaba anotar una patente era alrededor de 90 segundos, un tiempo bastante considerable teniendo en cuenta que el resultado de una búsqueda en la base de datos de patentes podría retonar más de 1000 patentes. Otra limitación es que el servicio sólo realiza anotaciones a textos en inglés. Así, las patentes que no contaban con el abstract en este idioma, no se tenían en cuenta.

3. **Implementar un algoritmo que determine qué tan novedoso es un documento con respecto a un conjunto de documentos.**

   La implementación de un algortimo que determine la novedad de una solicitud de patente entre un conjunto de patentes, fue un proceso que requirió una extensa revisión de la literatura existente, dado que son pocos los artículos científicos enfocados en este tema.

La función que se implementó está basada en parte a un modelo que se describe en la sección 3.4. Sin embargo, no podemos asegurar la efectividad de la misma, debido a que no se contó con la evaluación de un experto. Es por esto que más adelante, en la sección de trabajos futuros se propone probar la aplicación en ambientes reales relacionados con el ámbito de la propiedad intelectual.

Como se puede observar en el capítulo 5 de las pruebas, el nivel de novedad de la solicitud de patente depende de la similitud que se obtiene entre el conjunto de patentes. Es por esto que si se encuntra una patente muy similar a la solicitud analizada su nivel de novedad será bajo.

4. **Desarrollar una aplicación prototipo que integre los dos algoritmos propuestos.**

   El objetivo general de este trabajo de grado era el desarrollo de una herramienta para apoyar el proceso de análisis de patentes en un campo en específico. Así, teniendo en cuenta las ventajas que presentaba el servicio de Bioportal (ver sección 3.2), se opta por seleccionar el campo de la ***biomedicina***; un campo que abarca la medicina, la odontología y a las biociencias como: bioquímica, inmunología, química, biología, histología, genética, embriología, anatomía, fisiología, patología, ingeniería biomédica, zoología, botánica y microbiología.

   No fue posible probar la herramienta con algún experto en el área de la propiedad intelectual (PI), debido al poco tiempo con que se contaba.

   Sin embargo, las pruebas realizadas para medir el comportamiento de los algoritmos (ver capítulo 5), demuestra que el sistema podría ayudar de forma rápida a derterminar la similitud y el grado de novedad entre un conjunto de patentes y así lograr que el tiempo que toma analizar una patente disminuya.

   Se espera que la herramienta pueda ser utilizada con fines académicos e investigativos.

6.2. Trabajos futuros

Se identificaron diferentes líneas de trabajo futuros, entre los cuales se consideran:

- Probar la herramienta en ambientes reales con expertos para medir la efectividad de la propuesta, de este modo, se podría llevar a cabo el diseño y la ejecución de un esquema de pruebas con alguna entidad que realice análisis de patentabilidad, tal como la Oficina de Transferencia de Resultados de Investigación de la Universidad del Valle (OTRI)[2].

[2]http://viceinvestigaciones.univalle.edu.co/universidad-empresa-otri

- Debido a que la herramienta sólo trabaja con el abstract de las patentes, ésta se podría extender para que examine el texto completo, la cual podría ser una fuente valiosa para el análisis de patentes.
- Construir un grafo RDF a partir de los metadatos y anotaciones para hacer clustering y analizar los resultados obtenidos
- La implementación de un algoritmo de aprendizaje de máquina, el cual ayude a clasificar las patentes teniendo en cuenta la similitud y analizar su comportamiento.
- Extender la funcionalidad de la herramienta, que además de permitir el cálculo de la similitud entre un conjunto de patentes y el valor de la novedad, también se pueda realizar estadísticas de vigilancia tecnológica.

Referencias

[1] P. Escorsa and J. Valls. *Tecnología e innovación en la empresa*. Alfaomega., 2005.

[2] Charles V. Trappey Tong-Mei Wang A. J. C. Trappey, Hsin-Yi Peng. Ontology-based dental implant connection patent analysis. *in IEEE 17th International Conference, Computer Supported Cooperative Work in Design (CSCWD)*, pages pp. 257–262, 2013.

[3] Superintendencia de Industria y Comercio de Colombia. Patentes [online]. Available: http://www.sic.gov.co/drupal/patentes, 2016.

[4] Universidad de Valladolid. Anotaciones semánticas [online]. Available: http://www.infor.uva.es/ sblanco, 2006.

[5] Yuan Jiang, Yuliang Liao, and Guoxian Yu. Affinity Propagation Clustering Using Path Based Similarity. *Algorithms*, 9(3):46, jul 2016. ISSN 1999-4893. doi: 10.3390/a9030046. URL `http://www.mdpi.com/1999-4893/9/3/46`.

[6] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 2009.

[7] Leesse M. Everitt., Landau S. *Cluster Analysis*. 2001.

[8] V. Kumar Pang-Ning Tan, M. Steinbach. *Introduction to Data Mining*. Addison-Wesley, 2006.

[9] Jan M. Gerken and Martin G. Moehrle. A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. *Scientometrics*, 91(3):645–670, jan 2012. ISSN 0138-9130. URL `http://link.springer.com/10.1007/s11192-012-0635-7`.

[10] Assad Abbas, Limin Zhang, and Samee U. Khan. A literature review on the state-of-the-art in patent analysis. *World Patent Information*, 37:3–13, jun 2014. ISSN 01722190. URL `http://www.sciencedirect.com/science/article/pii/S0172219013001634`.

[11] T. Takahashi and T. Saiki. A measure to estimate the novelty of component combinations in technologies. In *2010 IEEE International Conference on Industrial Engineering and Engineering Management*, pages 2283–2286. IEEE, dec 2010. ISBN 978-1-4244-8501-7. doi: 10.1109/IEEM.2010.5674170. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5674170`.

[12] Y. Chen K. Cai R. Ma L. Zhang X. Wu X. Jin, S. Spangler and J. Han. Patent maintenance recommendation with patent information network model. *in ICDM'11*, pages pp. 280–289, 2011.

[13] Hedhili A. Khelif, K. and M. Collard. Semantic patent clustering for biomedical communities. *in International Conference on Web intelligence and Intelligent Agent Technology*, Vol. 1:pp. 419–422, 2008.

[14] Nisha Mattas and Deepti Mehrotra. Comparing Data Mining Techniques for Mining Patents. In *2015 Fifth International Conference on Advanced Computing & Communication Technologies*, pages 217–221. IEEE, feb 2015. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7079082`.

[15] Heeyong Noh, Yeongran Jo, and Sungjoo Lee. Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications*, 42(9):4348–4360, feb 2015. ISSN 09574174. URL `http://www.sciencedirect.com/science/article/pii/S0957417415000652`.

[16] Hongjiao Xu, Wen Zeng, Jie Gui, Peng Qu, Xiaohua Zhu, and Lijun Wang. Exploring similarity between academic paper and patent based on Latent Semantic Analysis and Vector Space Model. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 801–805. IEEE, aug 2015. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7382045`.

[17] S.; Lau G.T.; Kesan J.P. Law, K.H.;Taduri. An Ontology-Based Approach for Retrieving Information from Disparate Sectors in Government: The Patent System as an Exemplar. pages 2096–2105. IEEE, jan 2015. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7070064`.

[18] Charles V; Peng Hsin-Yi; Wang Tong-Mei Trappey, Amy J.C; Trappey. Ontology-based dental implant connection patent analysis. pages 257–262. IEEE, jun 2013. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6580972`.

[19] Daniela Boshnakoska, Ivan Chorbev, and Danco Davcev. Ontology supported patent search architecture with natural language analysis and fuzzy rules, 2013. ISSN 21945357. URL http://link.springer.com/10.1007/978-3-642-37169-1{_}27http://www.scopus.com/inward/record.url?eid=2-s2.0-84876277849{&}partnerID=40{&}md5=88b676f2b82493f1eece12c509ca2d36.

[20] Roy T. Fielding and Richard N. Taylor. Principled design of the modern web architecture. *Internet Technol*, (2):115–150, 2002.

[21] Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. BioPortal: Enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(SUPPL. 2):W541–5, jul 2011. ISSN 03051048. doi: 10.1093/nar/gkr469. URL http://www.ncbi.nlm.nih.gov/pubmed/21672956http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3125807.

[22] Tom Magerman, Bart van Looy, and Xiaoyan Song. Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, 82(2):289–306, feb 2010. ISSN 01389130. doi: 10.1007/s11192-009-0046-6. URL http://link.springer.com/10.1007/s11192-009-0046-6.

[23] Approximate TF-IDF based on topic extraction from massive message stream using the GPU. *Information Sciences*, 292:143–161, 2015. ISSN 00200255. doi: 10.1016/j.ins.2014.08.062.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[25] Zone Ching Lin, De Wei Wu, and Guo En Hong. Combination of improved cosine similarity and patent attribution probability method to judge the attribution of related patents of hydrolysis substrate fabrication process. *Advanced Engineering Informatics*, 30(1):26–38, 2016. ISSN 14740346. doi: 10.1016/j.aei.2015.11.003.

[26] Andres Felipe Rojas Hernandez and Nancy Yaneth Gelvez Garcia. Distributed processing using cosine similarity for mapping Big Data in Hadoop. *IEEE Latin America Transactions*, 14(6):2857–2861, jun 2016. ISSN 1548-0992. doi: 10.1109/TLA.2016.7555265. URL http://ieeexplore.ieee.org/document/7555265/.

[27] Daniel Kim, Daniel Burkhardt Cerigo, Hawoong Jeong, and Hyejin Youn. Technological novelty profile and invention's future impact. *EPJ Data Science*, 5 (1):8, dec 2016. ISSN 21931127. doi: 10.1140/epjds/s13688-016-0069-1. URL `http://www.epjdatascience.com/content/5/1/8`.

Anexo A: Especificación de requerimientos

El presente Anexo describe los requerimientos funcionales y no funcionales de la aplicación.

Requerimientos funcionales

Universidad del Valle	Especificación de requerimientos	Revisión: 001
Sistema para el apoyo al análisis de patentabilidad		Fecha: 15/02/2016
1. Módulo de usuario		
Requerimiento		Descripción
Requerimiento 1.1		El sistema en el módulo de usuario debe contener un formulario en el cual se pueda ingresar los datos de la patente a analizar.
Requerimiento 1.2		El sistema en el módulo de usuario debe contener un botón “Análizar” que iniciará el análisis de la patente.

Universidad del Valle	Especificación de requerimientos	Revisión: 001
Sistema para el apoyo al análisis de patentabilidad		Fecha: 15/02/2016
2. Módulo de visualización		
Requerimiento.		Descripción
Requerimiento 2.1		El sistema en el módulo de visualización deberá mostrar un gráfico en el que se muestren los clústers generados, resaltando el clúster al que pertenece la patente analizada.
Requerimiento 2.2		El sistema en el módulo de visualización deberá mostrar un gráfico con la valoración numérica de la similitud de una patente con respecto a la patente analizada, sólo se deben tener en cuenta las patentes que pertenecen al mismo clúster de la patente analizada.
Requerimiento 2.3		El sistema en el módulo de visualización deberá mostrar un gráfico con la valoración numérica de la novedad de la patente analizada.

Requerimientos no funcionales

Universidad del Valle	Especificación de requerimientos	Revisión: 001
Sistema para el apoyo al análisis de patentabilidad		Fecha: 15/02/2016
3. Módulo de análisis		
Requerimiento		Descripción
Requerimiento 3.1		En el momento que se inicia el análisis de la patente ingresada, el sistema debe hacer una búsqueda en los servicios web de las bases de datos de patentes, tomando como criterio de búsqueda las palabras clave ingresadas por el usuario.
Requerimiento 3.2		Luego de obtener las patentes, el sistema deberá guardar: identificador, país, kind, fecha, título, abstract, autores y clasificación de cada una de las patentes obtenidas.
Requerimiento 3.3		El sistema debe hacer la anotación semántica del abstract de cada patente y deberá guardar las anotaciones en la base de datos.

Anexo B: Historias de usuario

Los campos necesarios para iniciar el análisis de una patente se pueden observar en la siguiente Tabla:

Campo	**Tipo**	**Requerido**
Título de la solicitud	Alfanumérico	Sí
Identificador	Numérico	Sí
País	Alfabético	Sí
Fecha	Numerico	No
Abstract	Alfanumérico	Sí
Key-words	Alfabético	Sí
Autores	Alfabético	Sí
Clasificación	Alfanumérico, caracteres especiales	No
Kind	Alfanumérica	No

Anexo C: Documento final de iteraciones

	Iteración 1. **Revisión 002**
Sistema para el apoyo al análisis de patentabilidad	**Fecha: 22/05/2016**
Descripción	H-U/Requerimiento
En esta iteración se crea la interfaz donde el usuario ingresa los datos de la solicitud de la patente, el modelo de la base de datos y se determina el servicio Web para obtener los metadatos de las patentes, posteriormente se crean las consultas necesarias para acceder a dicho servicio	Se selecciona el requerimiento 1.1, 3.1, 3.2

Plan de trabajo		
Requerimiento	Prioridad	Fecha inicio
1.1	2	22/05/2016
3.1	4	28/05/2016
3.2	3	20/06/2016

	Iteración 2. **Revisión 002**
Sistema para el apoyo al análisis de patentabilidad	**Fecha: 02/07/2016**
Descripción	Requerimiento
Se desarrolla el módulo de anotación semántica por medio del servicio web de Bioportal. Se conecta el módulo con la base de datos y se crean los mecanismos necesarios para la interacción con el sistema.	Se selecciona el requerimiento 1.2, 3.3

Plan de trabajo		
Requerimiento	Prioridad	Fecha inicio
1.2	2	02/07/2016
3.3	4	07/07/2016

	Iteración 3. **Revisión 002**
Sistema para el apoyo al análisis de patentabilidad	**Fecha: 22/08/2016**
Descripción	Requerimiento
Se crea el módulo de análisis de patentabilidad para crear los diferentes clústers con las anotaciones generadas en la anterior iteración. También se implementa el mecanismo para determinar la similitud entre las patentes	Se selecciona el requerimiento 2.1, 2.2

Plan de trabajo		
Requerimiento	Prioridad	Fecha inicio
2.1	3	22/08/2016
2.2	4	01/09/2016
Novedades		
Tipo	Fecha	Descripción
Posponer		Parte del **requerimiento 2.1** fue postergado debido a que se debe evaluar la forma en que se debe mostrar el clustering. Se propone investigar diferentes técnicas de visualización e implementar la más adecuada.

	Iteración 4. **Revisión 003**
Sistema para el apoyo al análisis de patentabilidad	**Fecha: 25/09/2016**
Descripción	Requerimiento
Se implementa el módulo de la novedad y se evalua la mejor forma de visualizar el resultado.	Se selecciona el requerimiento 2.2, 2.3

Plan de trabajo		
Requerimiento	Prioridad	Fecha inicio
2.2	4	25/09/2016
2.3	4	05/10/2016
Novedades		
Tipo	Fecha	Descripción
Evaluación		Para la parte del requerimiento que fue postergada en la iteración 3. De acuerdo a lo realizado, se determinó hacer uso de la librearía d3js de Javascript para la visualización del clustering generado.

	Iteración 4. **Revisión 003**
Sistema para el apoyo al análisis de patentabilidad	**Fecha: 15/10/2016**
Descripción	H-U
Se integran todos los componentes del sistema.	Se selecciona la Historia de usuario 1

Plan de trabajo		
H-U	Prioridad	Fecha inicio
1	5	15/10/2016

Anexo D: Instalación

Para la instalación y puesta en marcha de la aplicación es necesario seguir los siguientes pasos:

1. Descargar desde la página oficial de Python `python.org/releases` la versión 'Windows X86-64 MSI Installer' (Última versión estable de Python 2 para Windows). Una vez descargado, hacemos doble-click en el instalador y aceptamos todas las condiciones del instalador. Al finalizar la instalación, ya deberíamos tener Python corriendo en nuestro sistema.

2. Son necesarias algunas librerías para que el sistema funcione, a continuación los scripts para hacer las instalaciones necesarias. Los scripts se deben ejecutar desde el Shell de windows:

   Nota: PATH es la ruta en la que se encuentra el archivo .whl, este archivo es necesario en caso de que el repositorio en el que se encuentra la librería no esté disponible.

   - python -m pip install numpy
   - python -m pip install PATH\numpy-1.11.1+mkl-cp27-cp27m-win32.whl
   - python -m pip install scikit-learn
   - python -m pip install scipy
   - python -m pip install PATH\scipy-0.18.1-cp27-cp27m-win32.whl
   - python -m pip install PATH\matplotlib-1.5.3-cp27-cp27m-win32.whl
   - python -m pip install virtualenv

3. Creamos una carpeta en la cual vamos a instalar la aplicación y dentro de ella ejecutamos los scrtipts.

   - C:\>mkdir project
   - C:\>cd project

4. Instalamos flask en el contenedor:
   - C:\project>python -m virtualenv flask
   - C:\project>flask\Scripts\pip install flask
5. Se crea el directorio app en el cual van a estar los archivos .py y .html
   - C:\project>mkdir app
   - C:\project>cd app
   - C:\project\app>mkdir static
   - C:\project\app>mkdir templates
   - C:\project\app>cd..
   - C:\project>cd flask
   - C:\project\flask>cd Scripts
6. Una vez ejecutados los anteriores comandos, sobre-escribimos la carpeta app con los archivos de código fuente y ejecutamos el servidor de flask con los siguientes comandos:
   - C:\project\flask\Scripts>activate
   - (flask) C:\project\flask\Scripts>cd
   - (flask) C:\>cd project
   - (flask) C:\project>cd app
7. Iniciamos la app con (flask) C:\project\app>python __init__.py

   Nota: La base de datos de la aplicación está en PostgreSQL. Es necesario crear el **usuario**: *admin*, **password**: *admin* y la base de datos annotation para que la aplicación funcione correctamente. Las tablas son creadas automáticamente desde la aplicación.

   Si lo anterior se hizo correctamente, ya se puede acceder a la aplicación en el navegador en la URL: `http://127.0.0.1:5000`