Técnicas de minería de texto aplicadas para la identificación de grupos de patentes afines en la industria del cacao

Paola Milena Rodríguez Millán

Trabajo de grado para optar al título de Ingeniera Industrial

Director:

Leonardo Hernán Talero Sarmiento

MSc. en Ingeniería Industrial

Codirectores:

Henry Lamos Díaz

Ph.D en Física-Matemática

Leidy Johanna Cárdenas Solano

MSc. en Ingeniería Industrial

Universidad Industrial de Santander

Facultad de Ingeniería Físico Mecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2020

AGRADECIMIENTOS

Gracias Dios mío por guiar mi camino, por la sabiduría y fortaleza para hoy alcanzar esta meta

Gracias a mi mamá por su amor y apoyo incondicional

Gracias a mi ángel que me cuida y por todas las enseñanzas que me dio en vida

Gracias al profesor Leonardo Talero Sarmiento por darme la oportunidad de este proyecto, por sus grandes conocimientos, por su apoyo, por su constante acompañamiento, por el tiempo dedicado y sobre todo por sus sabios consejos.

Gracias a mi codirectora Leidy Cárdenas Solano por confiar en mí, por su paciencia y su conocimiento compartido.

Gracias a mi codirector Henry Lamos Díaz por su orientación y conocimiento en este proyecto.

Gracias a mis amigos por su amistad, apoyo y ayuda en este largo camino.

Gracias a Juan Felipe Heredia por su conocimiento y apoyo en el proceso.

Gracias al grupo de investigación OPALO por las herramientas brindadas durante el desarrollo de este proyecto.

DEDICATORIA

A Dios, a mi mamá

Y a mi papá…

Ante el altar de tus férvidos recuerdos

Os rindo tierno mis cálidos hinojos

Viendo siempre tus divinos ojos

Que hace un año cerré con cándidos besos

Llego a ti en este hermoso día

Con la presea del deber cumplido,

Conservo de ti recuerdos infinitos

En la celda más íntima de mi alma,

Siendo ellos la fuerza de mi calma

Y para ti este grado que tanto anhelabas.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Lista de Apéndices

Ver apéndices adjuntos y pueden ser consultados en la base de datos de la Biblioteca UIS

Apéndice A. Artículo de investigación.

Resumen

Título: Técnicas de minería de texto aplicadas para la identificación de grupos de patentes afines en la industria del cacao[*]

Autores: Paola Milena Rodríguez Millán[**]

Palabras Claves: Minería de Texto, Análisis de Patentes, K-means, Cacao.

Descripción:

En la última década el número de patentes ha aumentado cada vez más rápido, y la cantidad de información técnica contenida en ellas dificulta su análisis. Los documentos de patentes permiten la generación de grandes cantidades de datos no estructurados, que pueden procesarse con la ayuda de diferentes técnicas de minería de textos. Las bases de datos de patentes disponibles, a través de su interfaz, permite la inclusión y exclusión de aplicaciones estándar en las ecuaciones de búsqueda, posibilitando la extracción efectiva de los datos requeridos para el análisis, dando paso a la aplicación de herramientas automatizadas con el fin de identificar relaciones y tendencias que pueden mejorar la ventaja competitiva. El presente trabajo de investigación tiene como objetivo la identificación de grupos de patentes afines en la industria del cacao por medio de una minería de texto, donde se aplicó una metodología de agrupamiento de documentos no supervisada, basada en datos no estructurados para los que se lleva a cabo un preprocesamiento de los datos, luego se genera un corpus como espacio vectorial de los documentos de patentes, una representación de TF-IDF para finalmente aplicar la técnica de agrupamiento k-means, que permitió identificar relaciones entre las patentes y los grupos generados.

---

[*] Trabajo de grado.

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Estudios Industriales y Empresariales. Director: MSc. Leonardo Hernán Talero Sarmiento, Codirectores: PhD. Henry Lamos Díaz, MSc. Leidy Johanna Cárdenas Solano.

Abstract

Title: Text mining techniques applied to identify groups of related patents in the cocoa industry[*]

Authors: Paola Milena Rodríguez Millán[**]

Keywords: Text Mining**,** Patent Analysis, K-means, Cocoa.

Description:

In the last decade the number of patents has increased faster and faster, and the amount of technical information contained in them makes their analysis difficult. Patent documents allow the generation of large amounts of unstructured data, which can be processed with the help of text mining techniques. The patent databases, through their interface, allows the inclusion and exclusion of standard applications in the search equations, enabling the effective extraction of the data required for the analysis, giving way to the application of automated tools in order to identify relationships and trends that can improve competitive advantage. The objective of this research work is to identify groups of related patents in the cocoa industry through text mining, where an unsupervised document grouping methodology was applied, based on unstructured data for which it is carried out. carry out a preprocessing of the data, then a corpus is generated as a vector space of the patent documents, a representation of TF-IDF to finally apply the k-means grouping technique, which seeks to identify relationships between the patents and the generated groups.

---

[*] Bachelor Degree.

[**] Faculty of Physicomechanical Engineering. Industrial and Business School. Board of Advisors: MSc. Leonardo Hernán Talero Sarmiento, Ph.D. Henry Lamos Díaz, MSc. Leidy Johanna Cárdenas Solano.

Introducción

La agricultura es un conjunto de actividades con gran importancia en la economía de Colombia. Estas actividades favorecen el desarrollo del país por su significativa participación en la producción interna y le generación de empleo, así como el aporte a la seguridad alimentaria, esencial sobre todo para los países menos industrializados (Nyberg, Saadat, & Zoraida, 2006). Durante las últimas décadas, han surgido diferentes acontecimientos mundiales, como la globalización, las cadenas de valor integradas, el acelerado crecimiento de innovación tecnológica e institucionales, el condicionamiento ambiental y el aumento del precio de bienes agrícolas, generando el reconocimiento de la capacidad que tiene la agricultura para ejercer múltiples funciones en el desarrollo de los países (Perfetti, Balcázar, Hernández, & Leibovich, 2013). En este sentido, en Colombia existe la necesidad de desarrollar estrategias para promover y fortalecer el desarrollo de la agricultura y de los territorios rurales para que los cultivos puedan contribuir al crecimiento económico, y reducir la pobreza y del hambre.

Uno de los cultivos que aportan en gran medida a la economía colombiana es el cultivo del cacao, dado que presenta una de las mayores participaciones en producción, con 54.926 toneladas y en área sembrada con un total de 134.575 hectáreas para el primer semestre del 2019 (Nacional Agropecuaria, 2020). Así mismo, el cacao ha sido diferenciado como uno de los productos agroindustriales con mayor potencial, debido a que es reconocida a nivel mundial la calidad de los genotipos que se cultivan en el país y, de este modo, poder posicionarse en el grupo de cacaos especiales. Sin embargo, “se han identificado problemas en la cadena de valor, como el bajo nivel de desarrollo tecnológico en el negocio de transformación de la postcosecha en las principales zonas productoras de Colombia, el descuido de los parámetros de calidad y de los requerimientos

del mercado internacional, y la relevancia, confianza e integración de los eslabones de la cadena del cacao" (Contreras, 2017).

El cacao en Colombia es un producto importante para la agricultura nacional y el desarrollo sostenible del país, dado que es un territorio privilegiado por su ubicación geográfica, variedad cultural, diversidad biológica, recursos naturales (Vargas, 2016) y condiciones agroecológicas favorables para la producción del cultivo (Ministerio de Agricultura y Desarrollo Rural, 2010)

Según el marco conceptual descrito anteriormente sobre el valor del cultivo de cacao y las ventajas del campo colombiano, se busca contribuir al mejoramiento del desarrollo tecnológico considerando las condiciones de cada unidad productiva (Contreras, 2017), por medio del fortalecimiento de la estructura y el mecanismo de investigación, teniendo en cuenta que la competitividad de un país o empresa debe basarse en la capacidad de su industria para innovar y mejorar continuamente sus productos, servicios y procesos.(Gavilanes, Rio, Cilleruelo, & Garechana, 2011).

El proceso de innovación se basa en la comprensión, transformación, y difusión del conocimiento, por lo que es necesario contar con herramientas que permitan una adecuada gestión y transformación de la información en conocimiento favorable para la toma de decisiones estratégicas. Tomando un mayor valor en la situación actual, donde la revolución del internet ha permitido el acceso a la información almacenadas en bases de datos y páginas web (Gavilanes, Rio, & Cilleruelo, 2010).

Con la fácil disponibilidad y acceso a las bases de datos de patentes, los investigadores aprovechan cada vez más estos recursos gratuitos de información, ya que son una muy buena fuente de información técnica y de innovación (Dou, 2004). Además, el análisis de patentes aporta información de gran valor a través de indicadores económicos que miden la relación entre el

desarrollo tecnológico y el crecimiento económico, su impacto en la productividad, el desempeño innovador a nivel mundial, entre otros (W. M. Wang & Cheung, 2011).

Según un informe de la OMPI, las solicitudes de patente presentadas a escala mundial han experimentado el crecimiento más rápido de los últimos 18 años, y asegura que, tras la crisis financiera de 2009, las solicitudes de títulos de propiedad intelectual (P.I.) presentadas a escala mundial y la producción económica mundial han seguido caminos divergentes, bajo esta premisa, el Director General de la OMPI, Francis Gurry, considera: "A pesar de que la recuperación económica producida desde la crisis de 2009 no ha sido homogénea y no ha logrado reducir los elevados niveles de desempleo que resultaban inaceptables, las solicitudes de títulos de P.I. presentadas han aumentado a un ritmo superior al existente antes de la crisis" (OMPI, 2013). Este crecimiento acelerado de patente ha exigido el desarrollo de sofisticadas herramientas de análisis de patentes. Por lo tanto, en la actualidad existen varias herramientas de aprendizaje automático y técnicas de minería de texto que son capaces de realizar una amplia gama de tareas, como analizar y pronosticar tendencias tecnológicas futuras, llevar a cabo una planificación tecnológica estratégica, detectar infracciones de patentes, determinar la calidad de las patentes y las patentes más prometedoras, e identificar puntos críticos tecnológicos y vacíos de patentes (Abbas, Zhang, & Khan, 2014).

Dicho lo anterior, el presente trabajo de investigación tiene como propósito la construcción de un modelo de aprendizaje automático no supervisado, para el análisis de documentos científicos relacionados con la industria del cacao, con el fin de identificar grupos de patentes, a partir del uso de la información obtenida en las bases de datos gratuitas y con el apoyo del software R-Studio, aplicando las técnicas de minería de texto.

Tabla 1.

Cumplimiento de objetivos del proyecto

Objetivos específicos	Cumplimiento
Seleccionar técnicas de minería de texto y su aplicación en el análisis de documentos científicos mediante una revisión de literatura.	Capítulo 3
Consolidar una base de datos estructurada de patentes relacionadas con la industria del cacao, a partir de una búsqueda en fuentes secundarias.	Numeral 6.1
Determinar grupos de patentes afines en la industria del cacao por medio de la aplicación de técnicas de minería de texto.	Numeral 6.4
Socializar los resultados investigativos mediante la construcción de un artículo de carácter publicable.	Apéndice B

1. Planteamiento del Problema

En Colombia, el cacao es uno de los principales productos de la cadena de producción agrícola nacional, el cual cuenta con condiciones agroecológicas favorables para su producción, que ha permitido alcanzar una calidad reconocida internacionalmente de cacao fino y de aroma (Ministerio de Agricultura y Desarrollo Rural, 2010). Sin embargo, debido al desconocimiento de los parámetros de calidad de los eslabones productivos y comerciales, se han identificado problemas en la cadena de valor, como el bajo nivel de desarrollo tecnológico del negocio de transformación de la postcosecha en las principales zonas productoras de Colombia, también existe el desconocimiento de los requerimientos del mercado internacional y los problemas de conexión, falta de confianza e integración de los eslabones de la cadena (Contreras, 2017), la falta de información para determinar tendencias mundiales de investigación y desarrollo, los limitados estudios de la dinámica comercial de los productos finales, y la ausencia de información disponible sobre las capacidades nacionales de esta cadena productiva (Ministerio de Agricultura y Desarrollo Rural, 2010).

Teniendo en cuenta lo anterior, resulta pertinente realizar una investigación que contribuya a identificar el desarrollo tecnológico en la cadena de valor, a partir de las patentes, como un indicador para medir el progreso tecnológico, ya que representan de manera concreta la creación y difusión de conocimiento en la actividad productiva (Casanova, 2019), dado que el sector cacaotero ha expuesto falencias tecnológicas que impiden que el cultivo alcance su potencial y no cumpla con las expectativas planteadas (Puentes, 2016). En relación con lo anterior, “Colombia es un país privilegiado por su ubicación geográfica, variedad cultural, climas diversos, flora, fauna, cuencas hidrográficas y recursos naturales. Estas fortalezas han hecho que la agricultura colombiana sea una valiosa fuente de ingresos para una parte de sus habitantes” (Vargas, 2016).

Sin embargo, la escasez de tierras idóneas para el cultivo del cacao y el bajo nivel de implementación de tecnología, han provocado que la producción del cacao se mantenga casi invariable en las últimas décadas (Puentes, 2016). Aun así, la producción nacional de cacao en grano ha presentado un aumento anual, alcanzando un máximo histórico con un total de 60.535 toneladas en el año 2017; no obstante, para el cierre del año 2018 la producción de este decayó en 6.45% al total registrado en el año anterior, debido al plan de renovación de plantaciones de la Federación Nacional de Cacaoteros. Esta reforma, contribuye a que, de las aproximadamente 80.000 hectáreas de cacao envejecidas en Colombia, 10.000 hayan sido renovadas. Esta renovación es una propuesta para los agricultores, quienes desde el inicio han tenido que cambiar, adaptarse y crear nuevos procesos para trabajar la tierra, lidiar con terrenos complejos y fenómenos climáticos extremos (Jose Graziano de Silva, 2013), teniendo en cuenta que el cacao es un cultivo de tardío rendimiento, por lo que los resultados se ven entre 3 o 4 años después de la siembra (SEMANA S.A, 2017).

Otras de las razones que dificulta la explotación agrícola, es que el 90% de la producción está centrado en granjas familiares, quienes producen más del 80% de los alimentos a nivel mundial, sin tener el conocimiento, las tecnologías y los recursos necesarios para desarrollar esta labor de una manera eficiente, lo que contribuye a la pobreza e inseguridad alimentaria que son dos de las principales problemáticas que padecen estas familias (FAO, 2014). Por tanto, se espera hacer un aporte en la consolidación de las innovaciones que han sido usadas, para referenciar a los agricultores y difundirlas, esperando generar un impacto en el mejoramiento de los procesos de agricultura y alimentación; alineado a solventar la problemática de alimentación para abastecer la demanda de una creciente población (FAO, 2014).

El desarrollo metodológico de este trabajo de investigación se centrará en un análisis de patentes enfocado en invenciones aplicables a los diferentes eslabones de la cadena de valor del cacao. Este análisis se describe como la ciencia de analizar grandes cantidades de información de propiedad intelectual, en relación con otras fuentes de datos, para descubrir relaciones y tendencias (Aristodemou & Tietze, 2018); y considera los documentos de patentes porque contienen importantes resultados de investigación, son largos y ricos en terminología técnica, por lo que se requieren muchos esfuerzos humanos para los análisis (Tseng, Lin, & Lin, 2007). Razón por la que se han aplicado diferentes herramientas automáticas para el análisis de patentes, dados los avances en técnicas de minería de texto que han logrado resultados eficientes para un público no especializado.

De hecho, es a partir de la minería de texto que se espera encontrar patrones implícitos, previamente desconocidos y potencialmente útiles para un conjunto de datos grandes. En la práctica, el proceso de minería de texto implica una serie de interacciones del usuario con las herramientas de minería de texto para explorar el conjunto de datos con el fin de encontrar dichos patrones. Después de complementarse con información adicional e interpretados por expertos experimentados, estos patrones pueden convertirse en inteligencia importante para la toma de decisiones (Tseng et al., 2007). Teniendo en cuenta que, a medida que se desarrollan herramientas más confiables para el análisis de texto, es posible capturar información de texto útil para un análisis que no estaba disponible en los enfoques de análisis de patentes convencionales. Por lo tanto, si las herramientas de minería de texto pueden extraer contenidos tecnológicos de manera efectiva, las bases de datos de patentes pueden proporcionar una fuente valiosa para el análisis tecnológico en profundidad (Noh, Jo, & Lee, 2015). En síntesis, este trabajo de investigación se centrará en identificar grupos de patentes afines en la industria del cacao, soportado por técnicas

de minería de textos, las cuales hacen parte del conjunto de herramientas de análisis de datos, con el fin de obtener información útil de cuerpos no estructurados.

2. Objetivos

2.1. Objetivo general

Identificar grupos de patentes afines en la industria del cacao a partir de técnicas de minería de texto.

2.2. Objetivos específicos

- Seleccionar técnicas de minería de texto y su aplicación en el análisis de documentos científicos mediante una revisión de literatura.
- Consolidar una base de datos estructurada de patentes relacionadas con la industria del cacao, a partir de una búsqueda en fuentes secundarias.
- Determinar grupos de patentes afines en la industria del cacao por medio de la aplicación de técnicas de minería de texto.
- Socializar los resultados investigativos mediante la construcción de un artículo de carácter publicable.

3. Revisión de la literatura

Los avances continuos en tecnologías digitales permiten el desarrollo de técnicas para procesar datos textuales de diferentes fuentes de información. Los datos son información concreta de gran valor para una economía competitiva impulsada por el desarrollo tecnológico. El aumento en la disponibilidad de datos brinda oportunidades para mejorar la toma de decisiones y la formulación de estrategias, para introducir innovaciones de próxima generación y tecnologías disruptivas. (Aristodemou & Tietze, 2018)

Durante la última década, se han producido avances fundamentales en el campo del análisis de patentes. Las patentes representan la actividad tecnológica o inventiva y la producción en diferentes campos. Por lo tanto, el análisis de patentes es útil para ayudar a centrar los esfuerzos en investigación, economía, y en actividades de toma de decisiones; sin embargo, la comprensión de los datos depende en gran medida de la clasificación exitosa de las patentes. Para comprender mejor el impacto del entorno tecnológico actual, se han desarrollado diferentes enfoques de selección y clasificación de patentes, aplicando técnicas de aprendizaje automático. A modo de ejemplo, Bass y Kurgan codificaron un grupo de patentes de nanotecnología utilizando un conjunto predefinido de características para desarrollar, probar, y analizar varios modelos de clasificación optimizados; concluyendo que los mejores modelos de clasificación son RIPPER y árboles de decisión C4.5(2010). A continuación, se presentan los documentos agrupados por metodologías de minería de texto para el análisis de patentes.

3.1. Método de espacio vectorial

En el artículo de Xiaoyu Zhang (2014) se desarrolla un algoritmo interactivo de clasificación de patentes basado en la fusión de múltiples clasificadores y el aprendizaje activo, que comprende la construcción y actualización del modelo de clasificación. Para la actualización del modelo, primero se adopta la interacción humano-computadora para equilibrar la efectividad con la eficiencia. Como segundo lugar, en el modelo de espacio vectorial (VSM) se adopta una máquina de vectores de soporte para la construcción de clasificadores con el fin de combinar los sub-clasificadores (super-kernel) para adquirir clasificadores mejorados. Por último, se introduce el aprendizaje activo para seleccionar los datos más informativos con el mayor potencial para el refinamiento del modelo.

Los autores J. L. Wu, Chang, Tsao y Fan (2016) hacen un estudio en el que emplean los enfoques de minería de datos para identificar y clasificar la calidad de la nueva patente en el tiempo. Se desarrolla un sistema automático de análisis y clasificación de la calidad de las patentes, llamado SOM-KPCA-SVM, de acuerdo con los indicadores y las características de la calidad de las patentes respectivamente. Primero, el enfoque del mapa autoorganizado (SOM) se utiliza para agrupar las patentes publicadas anteriormente en diferentes grupos de calidad de acuerdo con los indicadores de calidad de la patente y define el tipo de calidad del grupo. El enfoque del análisis de componentes principales del núcleo (KPCA) se utiliza para transformar el espacio de características no lineales con el fin de mejorar el rendimiento de la clasificación. Por último, la máquina de vectores de soporte (SVM) se utiliza para construir el modelo de clasificación de la calidad de la patente.

3.2. Análisis semántico

Magerman, van Looy y Song (2010) en su estudio, el cual tiene como objetivo evaluar la precisión de las técnicas de análisis de texto léxico basadas en el análisis semántico latente (LSA) para construir medidas de distancia que sean adecuadas para comprender las similitudes entre los documentos de texto de patentes y publicaciones, concluyen indicando que las diversas opciones utilizadas para derivar las medidas de similitud varían mucho en precisión. La diferencia es que algunas opciones de minería de texto reconocidas, como la reducción de dimensionalidad y LSA, pueden no producir los mejores resultados cuando se trata de conjuntos de documentos más pequeños.

Por otra parte, W. M. Wang & Cheung (2011) desarrollan un Sistema de Gestión de Propiedad Intelectual de Base Semántica (SIPMS) para apoyar la gestión de propiedad intelectual usando técnicas de análisis semántico y minería de textos para procesar y analizar los documentos de

patente. El método propuesto adopta bases de datos mundiales de patentes como fuentes de conocimiento y permite a los usuarios buscar documentos de patente existentes o documentos de propiedad intelectual relevantes que estén relacionados con una posible nueva invención, proporciona relaciones y patrones entre un grupo de documentos. Posteriormente, Niemann, Moehrle y Frischkorn (2017) proponen un nuevo método de visualización y extracción de textos de patentes para identificar patrones de patentes a lo largo del tiempo. Una característica básica del método es la aplicación de similitudes semánticas para desarrollar líneas de patentes.

En el siguiente año, An, Kim, Mortara y Lee (2018) despliegan un enfoque novedoso basado en la red de análisis semántico de preposición que supera las limitaciones del análisis de red existente basado en palabras clave y demuestra su potencial a través de una aplicación. Una preposición es una palabra que define la relación entre dos palabras vecinas y, en el caso de las patentes, las preposiciones ayudan a revelar las relaciones entre las palabras clave relacionadas con las tecnologías. Para el 2019, J. Wang & Chen (2019) realizan una investigación en la que se desarrolla un enfoque de minería de patentes de detección de novedad que integra el método de análisis semántico latente (LSA) y el método de detección de valores atípicos basado en ángulos (ABOD) para manejar los problemas de desajuste de vocabulario y la "maldición de la gran dimensionalidad" y la tecnología de ayuda. El método LSA se basa en el principio de que los términos utilizados en los mismos contextos tienden a tener significados similares, por lo tanto, el objetivo es identifican patentes atípicas para explorar posibles oportunidades tecnológicas.

Por otra parte y de manera similar, autores como Song, Kim y Lee (2017), Li, Xie, Jiang, Zhou y L. Huang (2019) y Moehrle y Caferoglu (2019) apuntaron sus esfuerzos en investigaciones que implementan enfoques novedosos de minería de texto para el análisis de patentes; aplicando algoritmos de clasificación basados en el análisis semántico con el propósito de identificar áreas,

tendencias y pronosticar tecnologías emergentes. Un esfuerzo afín entre investigaciones se evidencia en los estudios de Wang, Ren, Chen, Liu, Qiao, Huang (2019) y Wittfoth (2019), quienes coinciden en el uso de la metodología de análisis semántico para el desarrollo de un indicador. El primero presenta un indicador llamado DWSAO, para la medición de similitud entre patentes. DWSAO mide la importancia de las estructuras sujeto-acción-obejto (SAO) para caracterizar la tecnología de patentes al ponderar conceptos semánticos similares que no son comunes en un dominio. Mientras que, en el segundo estudio se proporciona un indicador tecnológico para evaluar el alcance de la patente

3.3. Métodos híbridos

Para el año 2010, C. H. Wu, Ken y Huang (2010). proponen un nuevo proceso de toma de decisiones de patentes asistido por inteligencia artificial (IA) mediante la integración de un enfoque de selección de expertos y una maquina de vectores híbrida basada en la genética (HGA-SVM). El modelo HGA-SVM propuesto puede clasificar de forma dinámica y automática los documentos de patentes al registrar y aprender los conocimientos y la lógica de los expertos.

Dada la evolución del análisis de patentes, se reconoce en dos etapas principales: Primera, evolución del análisis de patentes, y Segunda, evolución de la minería de patentes. Este tipo de análisis implica que los autores apliquen medidas, indicadores, y estadísticas para analizar datos y citas de bibliometría, luego realicen un análisis de red y de conglomerados para encontrar patrones más complicados en las bibliometría y citas, Por lo tanto, las medidas de similitud son herramientas fundamentales para identificar relaciones entre patentes, un ejemplo de esta metodología es lo desarrollado por Madani y Weber (2016). Por otra parte, Yi Zhang et al. en su trabajo construyen un método híbrido de medida de similitud basado en múltiples indicadores para analizar portafolios de patentes, para ello proponen dos modelos: Primero, similitud categórica, y Segundo, similitud

semántica. El modelo de similitud categórica enfatiza las clasificaciones internacionales de patentes (IPC), mientras que el modelo de similitud semántica enfatiza elementos textuales. (Yi Zhang et al., 2016)

Por su parte, Zhou, Huang, Zhang, y Yu (2019) hablan de un enfoque hibrido para detectar y agrupar tecnologías candidatas para una posible recombinación y un análisis manual adicional por parte de expertos. Teniendo en cuenta que, el objetivo de su investigación es mapear patrones e identificar tecnologías anómalas o innovadoras que muestren una promesa significativa para la resolución de problemas futuros y, por esta razón, la recombinación futura. En la metodología combinan análisis de red, minería de texto, indicadores cuantitativos, análisis de probabilidad, y juicios de expertos, como resultado de inteligencia técnica sobre recombinaciones tecnológicas, cooperaciones estratégicas y relaciones competitivas de la literatura actual y en el futuro.

3.4. Reglas de asociación

En el año 2015, se proponen algoritmos de reglas de asociación ponderados para el análisis de las relaciones entre las diferentes tecnologías. El enfoque propuesto se basa en tres algoritmos (BWARM, WARM y BOWARM) que determinan la diferencia y el impacto de tecnologías. De tal forma, este enfoque reconoce la importancia desigual de las patentes y la clase de tecnología en términos de su impacto tecnológico y su importancia comercial. (Altuntas, Dereli, & Kusiak, 2015)

Para el siguiente año, Seo, Yoon, Park, Coh, Lee y Kwon (2016) proponen un enfoque sistemático para identificar posibles oportunidades de productos, con el fin de explorar como identificar oportunidades de productos en función de las capacidades internas de una empresa. Primero extrajeron información de productos de los conjuntos masivos de patentes utilizando la

técnica de minería de texto y luego generaron reglas de asociación entre productos que emplean minería de reglas de asociación. Finalmente, estiman el valor potencial de las oportunidades de productos con un indicador de medición que permite identificar productos potenciales que se pueden realizar utilizando las capacidades de una empresa objetivo.

3.5. Clustering

(2017) G. Kim & Bae realizan un estudio en el que proponen un método novedoso para pronosticar tecnologías prometedoras mediante el análisis de patentes. El proceso general de la metodología propuesta consta de tres pasos. En primer lugar, para formar grupos de tecnología, agrupan los documentos de patente con una tecnología similar en un espacio bidimensional mediante la adopción de la técnica de k- medias, sobre la base de la clasificación cooperativa de patentes (CPC). Luego, con respecto al proceso de definición de clústeres de tecnología, examinan la combinación de CPC de cada clúster formado. Por último, los indicadores de patentes, como las citas anticipadas. Se analizan familias de patentes tríadicas y reivindicaciones independientes para evaluar si los grupos de tecnología son prometedores.

Posteriormente, Kyembambe, Cheng, Huang, He, Zhan (2017) W. Choi, Ahn, y Shin (2019) realizan estudios de minería de texto para el análisis de patentes empleando algoritmos no supervisados, como la técnica de agrupamiento k-means con el fin de identificar grupos de patentes de tecnologías emergentes. El agrupamiento encuentra una estructura a partir de una colección de datos sin etiquetar, y los objetos dentro de un grupo comparten características similares. Ese mismo año, Yi Zhang, Huang, Porter, Zhang, y Lu (2019) construyen un marco empírico que integra el aprendizaje automático y la bibliometría para investigar el paradigma Big Data. Específicamente, utilizando un perfil de I + D para revelar información sobre la dinámica estadística y la distribución geográfica de la investigación y, en particular, explorando las interacciones globales entre

organizaciones académicas en todo el mundo. El modelo introdujo un proceso de k-means para identificar los clústeres tecnológicos centrales de la investigación y detectar sus vías evolutivas.

3.6. Otras metodologías

Bass y Kurgan (2010) en su artículo, utilizan una serie de técnicas de aprendizaje automático para encontrar los factores más informativos que permiten diferenciar entre patentes de Alto y Bajo valor dentro del campo de la nanotecnología. Para ello codificaron un gran conjunto de patentes de utilizando un conjunto predefinido de características para desarrollar, probar, y analizar varios modelos de clasificación optimizados. Los seis clasificadores implementados son: los *árboles de decisión* C4.5, RIPPER, *Random Forest, MetaCost C4.5, MetaCost RIPPER y MetaCost Random Forest*, así como la regresión logística. Siendo árboles de decisiónC4.5 y RIPPER los mejores modelos de clasificación, lograron más del 20% de sensibilidad, más del 99% de especificidad y se caracterizaron por una relación TP/FP que muestra por cada tres patentes ALTAS correctamente clasificadas, solo se realizaría una clasificación incorrecta.

Posteriormente, Park, Ree y Kim (2013) proponen un nuevo método adoptando la minería de texto basada en sujeto-acción-objeto (SAO) con el fin de tratar grandes volúmenes de datos y analizarlos automáticamente, presentando un reflejo de los conceptos claves tecnológicos en una patente y comparando directamente con las bases de reglas definidas. Además, utilizan las tendencias de evolución de TRIZ y la extracción de textos para identificar futuras patentes prometedoras para la transferencia de tecnología. Las tendencias TRIZ, presentan patrones evolutivos de tecnologías o sistemas. Dado que los esquemas de clasificación de patentes que existen dependen de la tecnología o están basados en TRIZ. Los autores F. Zhu, Wang, D. Zhu y Liu, proponen un esquema de clasificación de patentes supervisado y orientado a los requisitos del usuario. En el artículo, el proceso automático del método basado en metadatos consta de un estudio

de caso sobre tecnologías *System on a chip* (SoC) en el que se aplican árboles de decisión, clasificador bayesiano ingenuo y máquinas de soporte vectorial para valida la efectividad del nuevo esquema de clasificación. Los experimentos se llevan a cabo con la técnica de ponderación de términos de TFIDF y la técnica de selección de características de IG (2015).

En ese mismo año, se aplicaron diversas metodologías para analizar los datos de patentes en la gestión de la tecnología, dados los avances en las técnicas de análisis de datos disponibles. En particular, implementando minería de texto para fines de análisis de patentes. Por lo tanto, en el artículo de Noh, Jo y Lee demuestran que los elementos del documento de la patente, los métodos de selección de palabras claves como el de frecuencia de término – frecuencia de documento inverso(TF-IDF) y el número total de palabras claves son estadísticamente insignificantes (2015).

Años después, Roh, Jeong yYoon (2017) hacen un estudio sobre minería de texto basada en palabras claves con el objetivo mejorar las limitaciones de las técnicas existentes, permitiendo derivar un nivel de información más detallado de las características del campo de la tecnología y las características de cada patente de acuerdo con el tipo de información tecnológica. Para esto, desarrollan una metodología para estructurar y estratificar la información tecnológica aplicando el proceso de programación neurolingüística (PNL).

A partir de los avances recientes en el aprendizaje automático y el procesamiento del lenguaje natural. Balsmeier, Assaf, Chesebro, Fierro, K. Johnson; S. Johnson, Li, Lueck, O'Reagan, Yeh, Zang y Fleming (2018) desarrollaron una serie de herramientas que ingieren, analizan, desambiguan, y construyen automáticamente una base de datos actualizada utilizando datos de patentes. Las herramientas identifican entidades únicas de inventor, cesionario y ubicación mencionada en cada patente. Es decir, una herramienta de mapeo de red de coinventor automatizada que visualiza tendencias en patentes en los últimos 40 años.

Para 2019, Woo, Yeom y Lee (2019) proponen un enfoque de aprendizaje automático para seleccionar ideas vinculando los contenidos en invenciones patentadas y el valor tecnológico de las mismas. En el centro del enfoque propuesto están la técnica de minería de texto para construir vectores de palabras clave a partir de patentes, y el algoritmo de k-vecinos más próximos, para capturar las relaciones entre los vectores de palabras clave y el número de citas directas de las patentes. Ese mismo año, J. M. Kim, Kim, Jung, y Jun (2019) hicieron un estudio tecnológico para proponer un nuevo método en el que se aplica técnicas de minería de texto para analizar patentes en función del tiempo, siendo un factor a tener en cuenta en el análisis tecnológico ya que la tecnología evoluciona con el tiempo. Como resultado del estudio, descubren que el modelo de regresión de obstáculos de Poisson es el más adecuado para encontrar relaciones entre las palabras claves.

Por su parte, Lima, Argenta, Zattar y Klein (2019) presenta un análisis de patentes, usando la minería de texto, a través de un algoritmo desarrollado en el lenguaje de programación R, con el propósito de identificar la etapa de desarrollo tecnológico de los paneles fotovoltaicos. Sus resultados principales muestran que los esfuerzos de empresas e investigadores se centran en el desarrollo de nuevos dispositivos de seguimiento, nuevas técnicas de fabricación, nuevos materiales, nuevos sistemas de ensamblaje y sistemas de autolimpieza. La aplicación de minería de texto ayudó a identificar los términos más frecuentes en los documentos. El algoritmo del lenguaje de programación R demostró ser muy preciso en el análisis, y el análisis del contenido, a través de los indicadores generados con la minería de texto, permitió identificar las tendencias presentes en los documentos. Por lo anterior, el descubrimiento de oportunidades tecnológicas (TOD) es una técnica importante para ayudar a las empresas a mantener la ventaja del mercado y el desarrollo sostenible a través del tiempo, donde se integra una serie de técnicas tales como *Latent*

Dirichlet Allocation (LDA), Multidimensional Scaling (MDS) y Local outlier factor (LOF). (Shi, Cai, & Song, 2019).

Por su parte, Huang y Chen (2019) proponen un nuevo algoritmo de clúster superpuesto (*overlapping clustering method*) para satisfacer la necesidad de métodos innovadores en la identificación de la estructura intelectual de diferentes temas de investigación en la industria de la computación en la nube. Por lo tanto, en este estudio se implementa un nuevo índice para maximizar la distancia general entre los centros de agrupación, y la aplicación de dos nuevos algoritmos. El primero de agrupación no exhaustiva y el segundo para el proceso de análisis de co-palabras. Para finales del 2019, los autores Choi, Ahn y Shin (2019) desarrollan un método para analizar visualizaciones de forma multidimensional a las patentes relacionadas con la industria de Big Data Geoespacial (GSBD). En el aspecto espacial, utilizaron los datos de citas de patentes. Mientras que, en el aspecto no espacial, analizaron la tendencia de las series temporales de actividades de innovación tecnológica de GSBD basadas en la clasificación industrial y las palabras claves tecnológicas.

Finalizando 2019, Xiaoyu Wang, Zhai, Lin Y Wang (2019) presentan un método semi-supervisado para extraer información tecnológica en capas basada en la combinación de análisis semántico, coincidencia de reglas y aprendizaje semi-supervisado de documentos científicos con el fin de extender el alcance de la minería tecnológica. En el estudio más reciente de esta revisión, se construye una red semántica tecnológica (TechNet) que cubre los conceptos elementales en todos los dominios de la tecnología y sus asociaciones semánticas. Para derivar el TechNet, se utilizaron técnicas de procesamiento de lenguaje natural para extraer términos de textos de patentes masivos y se emplearon algoritmos recientes de incrustación de palabras para vectorizar dichos términos y establecer sus relaciones semánticas. (Sarica, Luo, & Wood, 2020).

3.7. Síntesis de la revisión.

El cambio tecnológico juega un papel crítico en el desarrollo industrial y social. Este cambio se manifiesta en nuevos productos, procesos o materiales y, como tale, los rastros tangibles de productos, procesos o materiales en forma de patentes o publicaciones. Estas publicaciones de nuevo conocimiento pueden usarse para identificar, medir y monitorear los cambios tecnológicos (Parraguez, Škec, e Carmo, & Maier, 2020). Particularmente, los volúmenes de documentos científicos permiten realizar análisis detallados de la información, de modo que se puedan implementar técnicas de minería de texto y el uso de algoritmos de clasificación no supervisada (clustering). Así mismo, se resalta el uso del método de agrupación de k-means para encontrar relaciones, tendencias y mejorar la ventaja competitiva en aspectos sustanciales al momento de tomar decisiones, algoritmo que se usará en el desarrollo de esta investigación.

4. Marco Teórico

4.1. Inteligencia artificial

La inteligencia artificial (IA) es una rama de la informática, es decir, inteligencia ejecutada por máquinas, la cual se encarga de estudiar modelos informáticos que pueden realizar actividades principales del ser humanos con base a dos de sus características típicas: el razonamiento y el comportamiento (Takeyas, 2007). El objetivo de la IA es la investigación teórica y el desarrollo de sistemas informáticos que puedan realizar tareas de inteligencia biológica o humana, con funciones como reconocimiento, toma de decisiones, control y con capacidad de aprender a partir de ejemplos. Por lo tanto, se puede observar que la IA puede considerarse como la simulación de la inteligencia cerebral (Fan, Fang, Wu, Guo, & Dai, 2020). Los sistemas de IA son autónomos, lo que implica que pueden funcionar sin intervención humana y pueden aprender e identificar

patrones para tomar decisiones y llegar a diferentes conclusiones basadas en el análisis de diversas situaciones (Čerka, Grigienė, & Sirbikytė, 2017).

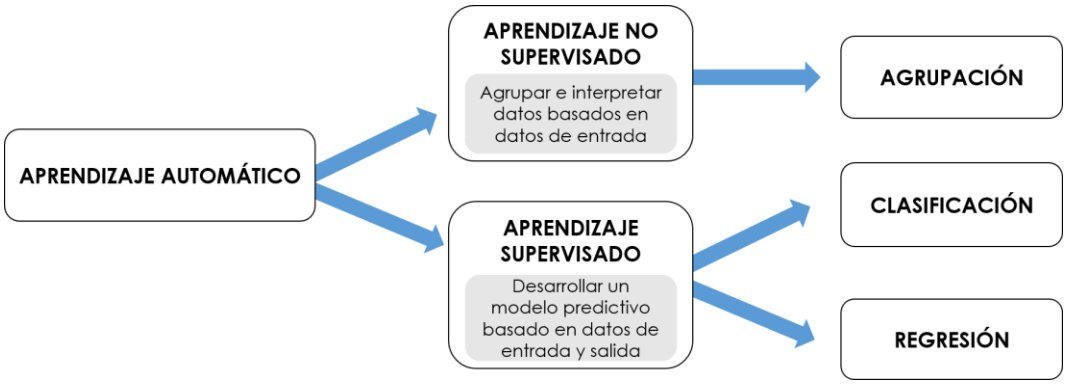
4.2. Aprendizaje automático (*Machine Learning, ML*)

Es una rama de la inteligencia artificial que tiene como objetivo el desarrollo de modelos o técnicas con la capacidad de generalizar comportamientos a partir de entradas de información en forma de ejemplos. Según Mitchell (1997), ML es el estudio de algoritmos informáticos que mejoran automáticamente por medio de la experiencia. De manera más amplia lo define como "un programa de computadora aprende de la experiencia E con respecto a alguna clase de tareas T y la medida de desempeño P, si su desempeño en tareas en T, medido por P, mejora con la experiencia E". Las técnicas de aprendizaje automático permiten el diseño y desarrollo de algoritmos que aprenden de la experiencia, es decir, son capaces de generar conocimiento de forma automática a partir de datos. Con lo anterior, se refiere a que un sistema puede adquirir e integrar conocimiento a través de observaciones a gran escala y mejora mientras aprende nuevos conocimientos en lugar de ser programado con ese conocimiento (Shapiro, 1992).

A continuación, se muestran los diferentes enfoques del aprendizaje automático.

Figura 1

Métodos de aprendizaje automático



Nota: Adaptado de *Introducing Machine Learning.* Math Works (2016, p.4)

4.2.1. *Aprendizaje supervisado*

Es una técnica del aprendizaje automático que permite obtener variables de resultado a partir de datos de entrenamiento, los cuales consisten de pares de objetos que se denominan vectores, con el propósito de crear una función que pueda predecir el valor correspondiente a cualquier dato de entrada válido. "El algoritmo de aprendizaje supervisado predice el valor de un atributo en el conjunto de datos (otros atributos conocidos) a partir de los datos cuya etiqueta se conoce e introduce una relación entre la etiqueta y otro conjunto de atributos. Estas relaciones se utilizan para predecir datos con etiqueta desconocida" (S. Y. Rodríguez & Díaz, 2009).

> "El objetivo del aprendizaje automático supervisado es construir un modelo que haga predicciones basadas en evidencia en presencia de incertidumbre. Un algoritmo de aprendizaje supervisado toma un conjunto conocido de datos de entrada y respuestas conocidas a los datos (salida) y entrena un modelo para generar predicciones razonables para la respuesta a nuevos datos" ("Introducing Machine Learning," 2016, p. 5).

4.2.2. *Aprendizaje no supervisado*

El aprendizaje no supervisado consiste en un método de aprendizaje automático en el que un modelo es ajustado a un conjunto de datos que no tienen ninguna etiqueta. Este aprendizaje se caracteriza porque no hay conocimiento a priori, y trata los datos de entrada como un conjunto de variables aleatorias (Espino, A.I.L., Mur, R.A. y de Miguel, 2004). "El método de aprendizaje no supervisado se utiliza para extraer inferencias de conjuntos de datos que consisten en datos de entrada sin respuestas etiquetadas. Por lo tanto, la agrupación (clustering) es la técnica más común del aprendizaje no supervisado. Se utiliza para el análisis de datos exploratorio para encontrar patrones o grupos en los datos" ("Introducing Machine Learning," 2016).

4.3. Agrupación (*Clustering*)

Es una técnica que se basa en el aprendizaje no supervisado, que consiste en dividir los datos en grupos de objetos similares. Para medir la similitud entre objetos se suelen utilizar diferentes formas de distancia: distancia euclidiana, distancia de Manhattan, distancia de Mahalanobis, etc. La representación de los datos a través de una serie de clústeres, conlleva la pérdida de detalles, pero se obtiene la simplificación de los mismos.

> "El clustering juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos espaciales (como GIS o datos procedentes de astronomía), aplicaciones web, marketing, diagnóstico médico, análisis de ADN en biología computacional, entre otras" (Garre, M., Cuadraro, J., Sicilia, M. A., Rodríguez, D., & Rejas, R, 2007, p. 9).

Esta técnica también puede ser utilizada para detectar anomalías o valores atípicos que no encajan en los clústeres al momento de realizar la segmentación de los grupos. Además, el método de agrupación es un procedimiento estadístico multivariante de clasificación automática de documentos, que busca reorganizar la información de un conjunto de datos sobre una muestra de entidades en grupos relativamente homogéneos a los que se denomina como clústeres. El análisis de agrupación se diferencia por la poca información que se conoce sobre la estructura de las categorías. Por lo tanto, su objetivo es ordenar en grupos de tal manera que el grado de asociación es alto entre los miembros del mismo grupo y bajo entre miembros de grupos diferentes.

4.3.1. *Agrupación de textos* (*Text Clustering, TC*)

La agrupación de documentos de texto en diversos grupos de categorías es un paso fundamental para indexar, buscar, gestionar y extraer grandes cantidades de datos de texto en la web o en los

sistemas de información corporativos. Además, el agrupamiento de texto se puede describir intuitivamente como hallazgos, dado un conjunto de vectores de datos en un espacio multidimensional. TC permite mejores resultados en la información al navegar y organizar documentos en jerarquías de agrupación significativas, proporcionando un complemento útil para los motores de búsqueda de texto tradicionales basados en palabras claves. Teniendo en cuenta que, este método, descubre automáticamente la estructura implícita de una colección original de documentos no estructurados, identificando los temas más frecuentes y distribuyendo los documentos en varios clústeres. La característica de esta distribución es que maximiza la similitud entre los elementos de un mismo clúster y maximiza la diferencia entre los clústeres (Jing, 2008, p. 1).

4.4. Minería de datos (*Data Mining*)

La minería de datos se define como el análisis y extracción de conocimiento a partir de datos (S. Y. Rodríguez & Díaz, 2009). Esta técnica ha surgido de la capacidad de analizar grandes volúmenes de información, mediante métodos automáticos o semiautomáticos, y con el fin de obtener resúmenes que faciliten la toma de decisiones. "Data mining combina técnicas semiautomáticas de inteligencia artificial, análisis estadístico, bases de datos, y visualización gráfica, para la obtención de información que no esté representada explícitamente en los datos. Así mismo, con el propósito de sustentar los procesos de toma de decisiones con un mayor conocimiento, descubre relaciones, tendencias, desviaciones comportamientos atípicos y patrones, y así se puede ubicar esta disciplina en el nivel más alto de los procesos tecnológicos de análisis de datos" (Beltrán, 2014).

> "La minería de datos hace uso de todas las técnicas que puedan aportar información útil, desde un sencillo análisis gráfico, pasando por métodos estadísticos más o menos complejos,

> complementados con métodos y algoritmos del campo de la inteligencia artificial y el aprendizaje automático que resuelven problemas típicos de agrupamiento automático, clasificación, predicción de valores, detección de patrones, asociación de atributos" (S. Y. Rodríguez & Díaz, 2009, p. 74).

"Esta disciplina está conformada por una serie de técnicas y herramientas adecuadas para el proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y comprensible de grandes conjuntos de datos con el fin de predecir automáticamente tendencias y comportamientos, además, describe modelos previamente desconocidos. Las tareas de la minería de datos se pueden dividir en dos categorías: minería de datos descriptiva y minería de datos predicativa" (Perichinsky et al., 2003).

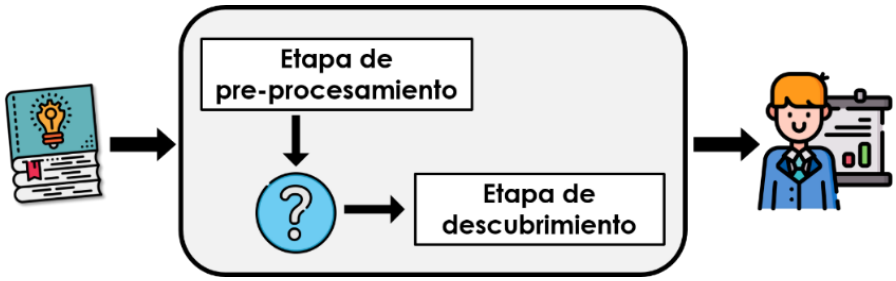
4.4.1. *Minería de texto*

La minería de texto es una reciente área de investigación del procesamiento de textos. La cual "se define como el proceso de descubrimiento de patrones interesantes y nuevos conocimientos en una colección de textos, es decir, la minería de texto es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección, pero que surgen de relacionar el contenido de varios de ellos" (Hearst, 1999; Kodratoff, 1999; Montes-y-gómez, 2001).

"Este proceso consiste de dos etapas principales: una etapa de preprocesamiento y una etapa de descubrimiento. En la primera etapa, los textos se transforman a algún tipo de representación estructurada o semiestructurada que facilite su posterior análisis, mientras que en la segunda etapa las representaciones intermedias se analizan con el objetivo de descubrir en ellas algunos patrones interesantes o nuevos conocimientos" (Tan, 1999). La **¡Error! No se encuentra el origen de la referencia.** ilustra este proceso.

Figura 2

Proceso de minería de texto.



Nota: Adaptado de *Montes-y-gómez* (2001, p. 4).
https://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf

Dependiendo del tipo de métodos usados en la etapa de preprocesamiento es el tipo de representación del contenido de los textos construida; y dependiendo de esta representación, es el tipo de patrones descubiertos. La **¡Error! No se encuentra el origen de la referencia.** muestra los tres tipos de estrategias empleadas en los actuales sistemas de minería de texto (Montes-y-gómez, 2001).

Figura 3

Estado del arte de la minería de texto

Etapa de pre-procesamiento	Tipo de representación	Tipo de descubrimientos
Categorización	Vector de temas	Nivel temático
Full- text	Secuencia de palabras	Patrones de lenguaje
Extracción de información	Tabla de datos	Relaciones entre entidades

Nota: Adaptado de *Montes-y-gómez* (2001, p. 4)

https://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf

4.5. Bolsa de palabras (*Bag Of Words)*

La bolsa de palabras es uno de los métodos de representación más populares para la categorización de objetos. La idea principal de este método es cuantificar cada punto clave extraído en un punto de palabras visuales y después representar cada imagen en un histograma de palabras visuales. Para esto, generalmente es usado un algoritmo de clustering que lograr generar las palabras visuales. Aunque varios estudios han mostrado resultados alentadores de la representación de la bolsa de palabras para la categorización de objetos, estudios teóricos sobre las propiedades del método de la bolsa de palabras están casi intactas, posiblemente debido a la dificultad que representa el uso de un proceso de clustering. Teniendo en cuenta los buenos resultados que se obtienen al realizar una categorización de texto, la representación por medio de un modelo de bolsa de palabras se convierte en uno de los métodos más populares para representar el contenido de una imagen y ha sido exitosamente aplicado a la categorización de objetos. (Yin Zhang, Jin, & Zhou, 2010).

4.5.1. *Preprocesamiento de texto*

El preprocesamiento de texto tiene como objetivo hacer la entrada de documentos más consistente para facilitar la representación del texto, lo cual es necesario para la mayoría de las tareas de análisis de texto. Los métodos tradicionales de preprocesamiento de texto incluyen la eliminación de palabras de detención y la derivación. La eliminación de palabras de detención elimina las palabras mediante una lista de palabras que se consideran generales o sin sentido; la derivación reduce las palabras flexionadas (o algunas veces derivadas) a su forma de tallo, base o raíz, por ejemplo, “mirar”, “mirando”, “miró” son representadas por la palabra “mirar”, por lo que las palabras con formas variantes pueden considerarse como la misma característica.

Los métodos de preprocesamiento dependen de la aplicación específica. En muchas aplicaciones, como *Opinion Mining* o NLP, necesitan analizar el mensaje desde un punto de vista sintáctico, lo que requiere que el método conserve la estructura original de la oración. Sin esta información, es difícil distinguir "¿De qué universidad se graduó el presidente?" y "¿Qué presidente se graduó de la Universidad de Harvard?", que tienen vocabularios superpuestos. En este caso, se debe evitar eliminar las palabras que contienen sintaxis (Aggarwal & Zhai, 2013).

4.6. K-means

Dentro de los principales algoritmos de agrupación en clústeres iterativos y de escalada de colinas es el algoritmo K-means. El agrupamiento de K-means es un método comúnmente utilizado para dividir automáticamente un conjunto de datos en k grupos (MacQueen, 1967). “El nombre de K-means viene porque representa cada uno de los clústeres por la media (o media ponderada) de sus puntos, es decir, por su centroide. La ventaja de la representación del centroide es que tiene un significado gráfico y estadístico inmediato. Por lo tanto, cada clúster se caracteriza por su centro o centroide, que se ubica en el centro o el medio de los elementos que componen el cluster. K-means es traducido como K-medias y se realiza en 4 etapas”.(Cambronero & Moreno, 2006, p. 7)

Figura 4

Etapas del algoritmo K-means

Etapa 1: Elegir aleatoriamente K objetos que forman así los K clusters iniciales. Para cada cluster k, el valor inicial del centro es = xi, con la xi únicos objetos de Dn pertenecientes al cluster.

Etapa 2: Reasigna los objetos del cluster. Para cada objeto x, el prototipo que se le asigna es el que es más próximo al objeto, según una medida de distancia, (habitualmente la medida euclidiana).

Etapa 3: Una vez que todos los objetos son colocados, recalcular los centros de K cluster. (los baricentros).

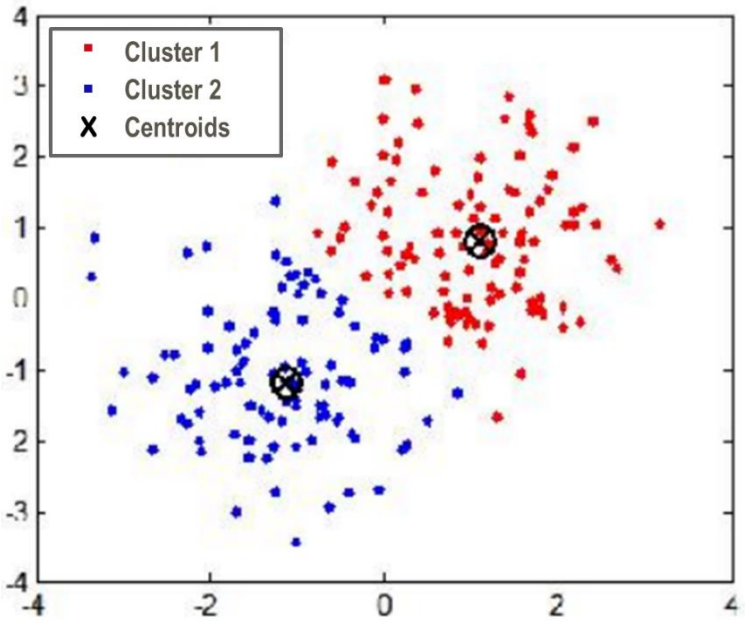
Etapa 4: Repetir las etapas 2 y 3 hasta que no se hagan más reasignaciones. Aunque el algoritmo termina siempre, no se garantiza el obtener la solución óptima.

Nota:. Adaptado de *Cambronero, C.G, & Moreno*, I. G., 2006.

El algoritmo creado por MacQueen es un algoritmo de clustering basado en un modelo estadístico llamado *finite mixtures* (mezcla de distribuciones). Una mezcla es un conjunto de k distribuciones, representando k clusters (Cambronero & Moreno, 2006). En la Figura 5, se observa como cada elemento del conjunto se va asignando al grupo con centroide más cercano.

Figura 5

Representación del algoritmo K-means



Nota: Adaptado de *Cambronero, C.G, & Moreno*, I. G., 2006.

4.6.1. *Algoritmo Lloyd*

Es un método de aprendizaje no supervisado porque permite obtener información sobre objetos sin proporcionar explícitamente un conocimiento previo sobre los mismos (Medina-Veloz, Luna-Rosas, Tavarez-Avendaño, & Narvaez-murillo, 2016). "El algoritmo de Lloyd es una técnica de agrupamiento que permite unificar n objetos en k clases. Al agrupar los objetos en distintas clases se concreta un método de predicción que se aplica iterativamente tras encontrar un nuevo elemento. Así mismo se puede establecer en una de las clases conocidas, esperando que sus características sean similares a las del resto de elementos de dicha clase. Este algoritmo comienza dividiendo los elementos de entrada en las distintas clases iniciales, ya sea mediante el uso de datos heurísticos o mediante el azar, para proceder posteriormente con el cálculo del punto medio o centro de gravedad de cada agrupación. De esta forma se construye una nueva partición, asociando cada elemento con el centro de gravedad más cercano y recalculando de nuevo los centroides" (Colomé Abril, 2012).

4.6.2. *Algoritmo de Forgy*

El algoritmo de Forgy se propuso en 1965 y es un modelo simple de centroide por lotes. Un centroide es el centro geométrico de un objeto convexo y se puede considerar como una generalización de la media (Morissette & Chartier, 2013). El algoritmo de Forgy consiste en la siguiente secuencia de pasos:

> ***Paso 1:*** Comenzar con cualquier configuración inicial
>
> Ir al Paso 2 si se comienza por un conjunto de k centroides
>
> Ir al Paso 3 si se comienza por una partición del conjunto de objetos en k grupos
>
> ***Paso 2:*** Asignar cada objeto a clasificar al centroide más próximo. Los centroides permanecen fijos en este paso
>
> ***Paso 3:*** Calcular los nuevos centroides como los baricentros de los k conglomerados obtenidos
>
> ***Paso 4:*** Alternar los Pasos 2 y 3 hasta que se alcance un determinado criterio de convergencia. (Medina-Veloz et al., 2016, p. 4)

4.6.3. *Algoritmo de Hartigan.*

Este algoritmo busca la partición del espacio de datos con la suma de cuadrados del error (SSE) dentro del clúster localmente óptimo. Significa que "puede asignar un objeto a otro subespacio, incluso si actualmente pertenece al subespacio del centroide más cercano, si al hacerlo se minimiza la suma de cuadrados total dentro del grupo. Los centros de los clústeres se inicializan de la misma forma que en el algoritmo de Forgy / Lloyd. Luego, los objetos se asignan al centroide más cercano a ellos y los centroides se calculan como la media de los puntos de datos asignados. Finalmente, de manera iterativa se busca que ningún objeto cambie el clúster, es decir, hasta que un cambio

haga que los clústeres sean más variables internamente o más similares externamente.*"*(Morissette & Chartier, 2013)

4.6.4. *Algoritmo de MacQueen.*

Es un algoritmo iterativo (también llamado en línea o incremental). La principal diferencia con el algoritmo de Forgy / Lloyd es que los centroides se vuelven a calcular cada vez que un objeto cambia de subespacio y también después de cada pasada por todos los objetos. "Los centroides se inicializan de la misma manera que en el algoritmo de Forgy / Lloyd y las iteraciones son las siguientes. Para cada caso, a su vez, si el centroide del subespacio al que pertenece actualmente es el más cercano, no se realiza ningún cambio. Si otro centroide es el más cercano, el caso se reasigna al otro centroide y los centroides tanto para el subespacio antiguo como para el nuevo se recalculan como la media de los casos correspondientes. El algoritmo es más eficiente ya que actualiza los centroides con más frecuencia y, por lo general, necesita realizar una pasada completa a través de los casos para converger en una solución" (Morissette & Chartier, 2013).

4.7. Medidas de distancia

Se utiliza una medida de distancia para calcular el grado de similaridad entre dos puntos de datos, cuanto mayor sea la distancia, más diferentes son los objetos y menor la probabilidad de que los métodos de clasificación los asigne en el mismo grupo o clúster (Malki, Rizk, El-Shorbagy, & Mousa, 2016).

4.7.1. *Distancia ecludiana*

La distancia euclidiana $D(i, j)$ entre coordenadas de dos puntos de datos u objetos se calcula como la raíz de la diferencia cuadrática.

$$D\ (i,j) = \sqrt{\sum_{k=1}^{n} \left(X_{ik} - X_{jk}\right)^2} \quad (1)$$

4.7.2. *Distancia de Mahalanobis*

La distancia Mahalanobis entre un par de puntos con la misma distribución de probabilidad se define por la expresión:

$$D\ (i,j) = \sqrt{(x_i - x_j)' \sum^{-1} (x_i - x_j)} \quad (2)$$

Donde la matriz asociada a la forma cuadrática $\sum^{-1}$ es la inversa de la matriz de varianzas.

4.7.3. *Distancia Manhattan.*

La longitud Manhattan entre dos puntos se calcula como la suma de las diferencias absolutas de sus coordenadas.

$$D\ (i,j) = \sum \left|x_{ik} - x_{jk}\right| \quad (3)$$

4.7.4. *Distancia de Chebyshev*

Es la distancia máxima de las diferencias entre sus dimensiones. Esta distancia se define por la siguiente expresión:

$$D\ (i,j) = \max \left|x_{ki} - x_{jk}\right| \quad (4)$$

4.7.5. *Distancia de Minkowski*

Es la distancia generalizada de la distancia euclidiana y Manhattan en un espacio vectorial normalizado. La distancia Minkowski de orden p entre dos puntos se define como :

$$D\,(i\,,j) = \left(\sum_{k=1}^{d} \left|x_{ik} - x_{jk}\right|^{P}\right)^{1/P} \quad (5)$$

Cuando $p = 1$, la expresión anterior coincide con la distancia Manhattan, y para $p = 2$, coincide con la distancia euclidiana. En el caso que p tome un valor infinito, obtenemos la distancia de Chebyshev.

4.8. Método del codo (Elbow Method)

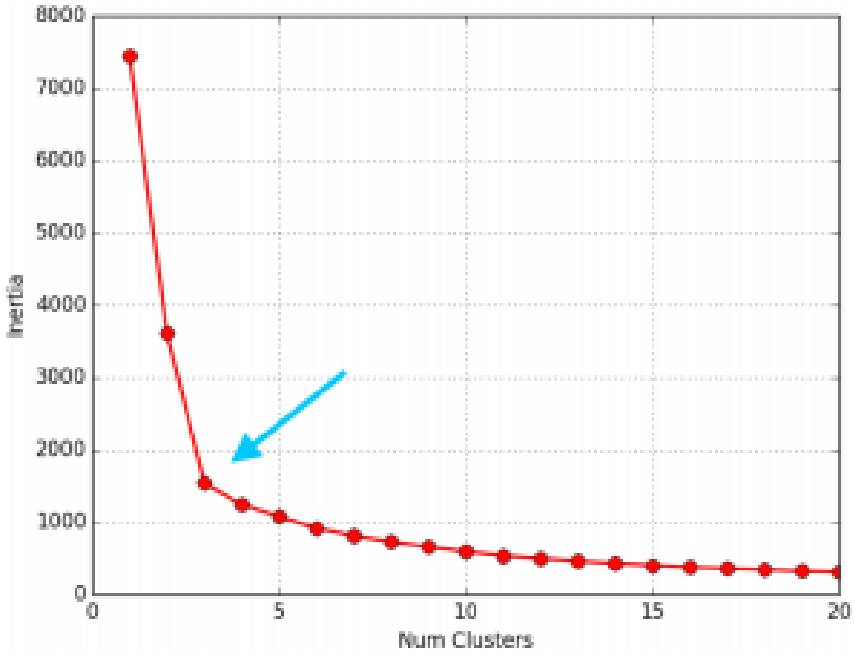
Este método utiliza los valores de la inercia obtenidos después de aplicar K-means a diferentes números de Clústeres (desde 1 a N Clústeres), siendo la inercia la suma de las distancias al cuadrado entre cada objeto del Cluster y su centroide:

$$Inercia = \sum_{i=0}^{N} \|x_i - \mu\|^2 \quad (6)$$

"Después de aplicar K-means de 1 a N Clústeres para obtener el valor de la inercia, representamos en una gráfica lineal la inercia relativa al número de Clústeres. En esta gráfica se debe apreciar un cambio brusco en la evolución de la inercia, teniendo en cuenta la forma representada por una línea similar a la forma de un brazo y su codo. El punto en el que se observa ese cambio brusco en la inercia indica un número recomendado de Clústeres a seleccionar para ese conjunto de datos; dicho de otra manera: el punto que representaría al codo del brazo será el número de Clústeres a seleccionar para ese conjunto de datos" (Moya, 2016).

Figura 6

Representación gráfica del método del codo



Nota: Modificado de *Ricardo Moya* (2016, p. 62)

4.9. Vigilancia tecnológica

"La vigilancia tecnológica (VT) es un esfuerzo sistemático y organizado de una empresa para observar, capturar, analizar, difundir con precisión, y recuperar información sobre hechos relacionados con el entorno económico, tecnológico, social o empresarial, porque puede implicar una oportunidad o amenaza. Requiere una atención personal o una actitud vigilante, y la suma organizada de estas actitudes conduce a la función de vigilancia de la empresa. En última instancia, la vigilancia filtra, interpreta y valora la información para que los usuarios puedan tomar decisiones y actuar de forma más eficaz" (González, Sánchez, & Caira, 2013)

La vigilancia tecnológica es una de las funciones que, siguiendo a Morin (1985), "requiere la gestión de la tecnología. El autor francés la asocia con las expectativas que genera y el grado de libertad que permite para la gestión. La vigilancia está estrechamente relacionada con la gestión

de la innovación y la estrategia de la empresa. Sin un pensamiento estratégico previo, es difícil considerar esfuerzos de vigilancia coordinados". (Palop & Vicente, 1999).

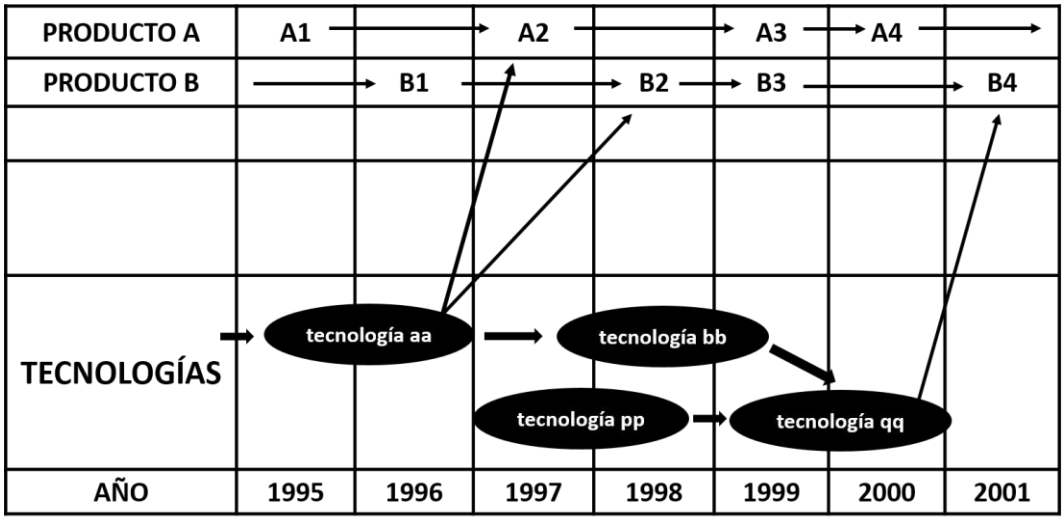
La vigilancia se proyecta sobre la toma de decisiones empresarial alertando sobre posibles amenazas y oportunidades, aportando nuevos elementos y enfoques, y reduciendo el riesgo (Palop & Vicente, 1999). "La vigilancia es un concepto inherente a la gestión de tecnología (GT), la cual involucra procesos de planeación, dirección, control y coordinación del desarrollo e implementación de la información para entender y anticiparse a los cambios tecnológicos, haciendo una detección temprana de eventos que representan oportunidades o amenazas potenciales" (León, Castellanos, & Vargas, 2006, p. 93).

4.9.1. *Mapeo tecnológico (Roadmapping)*

El roadmapping es un proceso que contribuye a la integración de negocios y tecnología, y a la definición de estrategia tecnológica al mostrar la interacción entre productos y tecnologías a lo largo del tiempo; teniendo en cuenta tanto el producto a corto como a largo plazo y sus aspectos tecnológicos. El principio de una hoja de ruta de la tecnología de producto se ilustra en la Figura 7. Los productos A y B, y las tecnologías aa y pp, requeridas para desarrollar y producir estos productos son mostrados por aproximadamente cinco años por delante. Productos A3 y A4, y las tecnologías bb y qq evolucionar desde el principio los productos A1 y A2, y las tecnologías aa y pp, respectivamente. Tenga en cuenta que la tecnología qq "mata" la tecnología bb.

Figura 7

Hoja de ruta para los productos y tecnologías



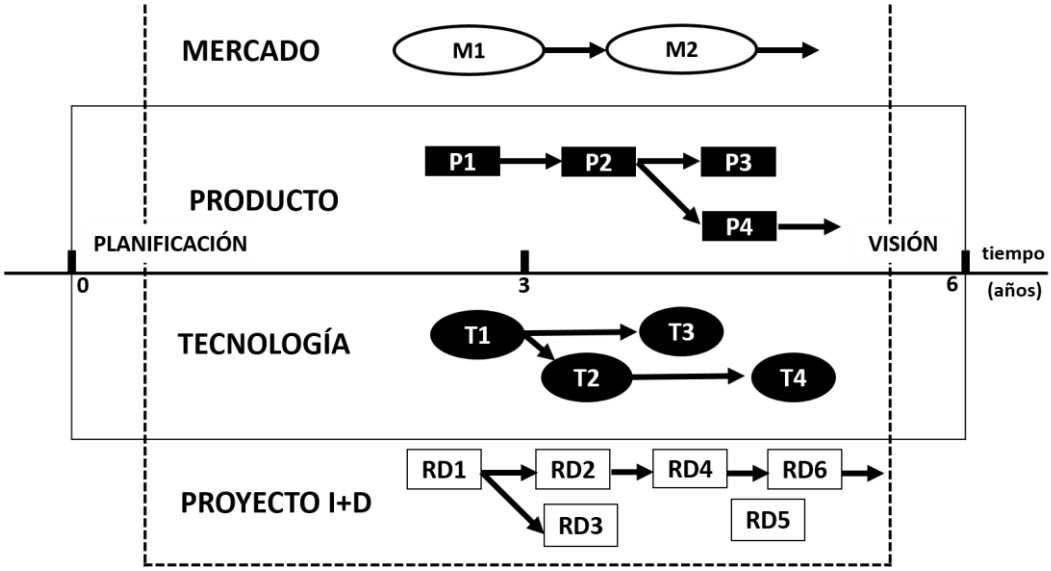
Nota: Adaptado de *Groenveld* (2007, p. 50)

Las hojas de ruta de tecnología de producto requieren una buena comprensión de los mercados y las aplicaciones para definir los productos en términos de los requisitos del cliente. Desde estos requisitos, las funciones técnicas del producto son determinadas, para luego determinar las tecnologías necesarias para realizar estas funciones. Todas estas actividades ayudan a mejorar, en particular, la parte frontal del proceso de creación del producto (el concepto y la fase de ideas) al proporcionar una mejor información. Los proyectos de I + D ayudan a construir las capacidades tecnológicas necesarias. La Figura 8 muestra cómo estas áreas están relacionadas.

El intervalo de tiempo que se mostrará en una hoja de ruta depende del tipo de productos y del nivel de agregación considerado. Las hojas de ruta que describen productos específicos con ciclos de vida cortos (p. ej., portátiles, productos de audio) a menudo no cubrirán más de tres o cuatro años, mientras que las hojas de ruta que presentan información acerca de una categoría de producto genérico (por ejemplo, almacenamiento óptico) pueden extenderse a diez años.

Figura 8.

Las hojas de ruta de la tecnología de un producto



Nota: La construcción de hojas de ruta exige una estrecha cooperación entre las funciones responsables de estas áreas. Adaptado de *Groenveld* (2007, p. 51)

Los 2–3 años iniciales son la llamada fase de planificación. Esta fase está impulsada por programas de productos predeterminados y recursos para el desarrollo que fueron asignados. La fase de visión se encuentra un poco más adelante (3–6 años); el proceso de mapeo depende fuertemente de la construcción de la visión. Las hojas de ruta pueden ser aplicadas a sistemas, a un producto gama, a productos específicos, a componentes o a procesos relacionados con la producción (Groenveld, 2007).

4.10. Análisis de patentes.

El análisis de patentes describe la ciencia de analizar grandes cantidades de información de propiedad intelectual en relación con otras fuentes de datos, para descubrir relaciones y pronosticar tendencias tecnológicas futuras con el fin de llevar a cabo una planificación tecnológica estratégica e identificar *hotspots* tecnológicos y vacíos de patentes (Aristodemou, Tietze, Athanassopoulou, & Minshall, 2017).

El análisis de documentos de patente a menudo se ha empleado para generar indicadores económicos que miden el vínculo entre el desarrollo tecnológico y el crecimiento económico, estiman los flujos de conocimiento tecnológico y su impacto en productividad, comparan el desempeño innovador en el contexto internacional, evalúan la competitividad de las empresas, desarrollan planes tecnológicos, priorizar la inversión en I + D, o monitorear el cambio tecnológico en las empresas (W. M. Wang & Cheung, 2011). Los indicadores de patentes proporcionan una herramienta de pronóstico muy útil para los tomadores de decisiones en los sectores público y privado (Campbell, 1983).

El análisis de patentes consta de tres etapas principales: preprocesamiento, procesamiento y procesamiento posterior. Estas etapas implican una serie de pasos que incluyen, extraer patentes de bases de datos de patentes, extraer la información de las patentes y analizar la información extraída para inferir las conclusiones lógicas (Abbas et al., 2014).

4.10.1. *Patente*

Una patente es un derecho exclusivo que se concede sobre una invención, es decir, una patente es el derecho legal de un inventor de excluir a otros de hacer o usar una invención particular. Este derecho a veces se denomina "derecho de propiedad intelectual" y se considera un incentivo para la innovación. Además, este derecho está habitualmente limitado en el tiempo a 20 años desde la fecha de presentación de la solicitud, en la mayoría de los países. El principio detrás de la patente moderna es que un inventor tiene un tiempo limitado para excluir a otros de suministrar o usar una invención para alentar la actividad inventiva al evitar la imitación inmediata. A cambio, se requiere que el inventor haga pública la descripción y la implementación de la invención en lugar de mantenerla en secreto, permitiendo a otros desarrollar más fácilmente el conocimiento contenido en su invención (Hall, 2007).

4.10.1.1. Claims

Las reclamaciones o reivindicaciones (claims) son la parte de la patente que define la extensión, es decir, el alcance de la protección legal que se busca para la invención. La descripción y los dibujos se utilizan para interpretar las reivindicaciones.

Las reivindicaciones de una patente son independientes y dependientes. Las independientes constan de dos partes: la parte de la técnica y la parte de caracterización. Las reivindicaciones independientes contienen las principales características de la invención y las reivindicaciones dependientes son las que incluyen todas las características técnicas de manera más concreta dentro de un marco establecido por la reivindicación independiente de la cual depende.(Oficina Española de Patentes y Marcas - OEPM, 2006)

4.11. Cacao.

El cacao es el fruto del árbol *Theobroma cacao* (nativo de la región amazónica de América del Sur). Según exponen Morales, García y Mñendez: "el árbol de cacao es una planta perenne de la familia de las Esterculiáceas que rinde varias cosechas al año. Alcanza una altura media de 6 m y tiene hojas lustrosas de hasta 30 cm de longitud y pequeñas flores rosas que se forman en el tronco y en las ramas más viejas. Sólo el 30% de las 6.000 flores que se abren durante el año llegan a formar semillas, éstas, llamadas comúnmente habas de cacao están encerradas en una mazorca o piña de color pardo rojizo de 28 cm en promedio de longitud. Las semillas de cacao, tienen un sabor amargo, son de color púrpura o blancuzco y se parecen a las almendras. La cáscara (testa) representa 10-14% del peso seco de la semilla de cacao, mientras que el núcleo o cotiledón se compone de la mayor parte del restante 86-90%. El cotiledón confiere sabores y aromas característicos de chocolate" (2012, p. 80).

> "El grano de cacao está compuesto principalmente de grasa (alrededor de un 55% después de fermentado, tostado, y secado). Un 60% de la grasa del cacao es saturada, rica en ácidos grasos como el esteárico (34%) o el palmítico (28%). Pero también contiene ácidos grasos insaturados como el oleico (35%) que juega un papel importante en la protección vascular al disminuir el colesterol y las LDL (Lipoproteínas de Baja Densidad) y aumentar las HDL (Lipoproteínas de Alta Densidad o colesterol bueno). El siguiente ingrediente más importante es la proteína o los elementos nitrogenados, incluidos la *theobromina* y la cafeína, que existen en pequeñas cantidades en el grano. Almidones y azúcares forman del 20 al 25 % del peso del grano" (Medina & Vargas, 2009; Morales et al., 2012, p. 80).

El cacao (aproximadamente del 80% al 90%) proviene de pequeñas fincas familiares, a través de cinco a seis millones de productores de cacao. La productividad de los granos de cacao es de 500 a 600 kg por hectárea en América. Por lo tanto, el rendimiento por hectárea varía según el país y el tipo de cacao cultivado allí. Según la World Cocoa Foundation (WCF), hay un total de 5 a 6 millones de productores de cacao en todo el mundo que comprenden alrededor de 40 millones a 50 millones de personas que dependen del cultivo del cacao para sus ingresos, lo que contribuye a la producción anual de cacao en todo el mundo en 4,2 millones de toneladas valoradas en $ 11,8 mil millones. creciendo a una tasa del 3% anual desde la última década (Beg, Ahmad, Jan, & Bashir, 2017).

4.11.1. *Procesamiento del cacao*

El procesamiento del cacao es el mismo de los últimos 150 años que pasa por una serie de equipos para transformarse en un producto.(Beg, Ahmad, Jan, & Bashir, 2017)

4.11.1.1. Cosecha. *Theobroma cacao* normalmente comienza a dar frutos después de 3 años y el rendimiento alcanza su máximo después de 8 a 9 años. Se pueden encontrar mazorcas maduras en los árboles de cacao durante todo el año, pero algunos países tienen dos temporadas de alta producción por año. Los cambios en las condiciones ambientales pueden afectar la cosecha, la productividad y el rendimiento del cultivo. La recolección se realiza manualmente a través de un alfanje o un cuchillo largo para obtener una vaina de los árboles con una herramienta. Estas vainas se abren para sacar los frijoles de la vaina. Una mazorca consta de 20 a 50 granos, según la variedad de cacao (Beg et al., 2017)

4.11.1.2. Fermentación. Los frijoles frescos obtenidos de las mazorcas se pueden empaquetar en canastas, cajas o amontonar en pilas que se pueden cubrir con hojas de plátano para iniciar la fermentación anaeróbica. Este proceso dura de 3 a 7 días para cumplir tres propósitos principales a saber, licuefacción, eliminación de pulpa mucilaginosa y desarrollo de aroma, color y sabor. La etapa de fermentación determina la calidad del cacao en polvo. (Beg et al., 2017)

4.11.1.3. Secado Los granos de cacao se pueden secar mediante secado solar al aire libre o mediante el horno secador de aire caliente para evitar el deterioro por bacterias.(Beg et al., 2017)

4.11.1.4. Embalaje y transporte. Los frijoles secos se empaquetan en sacos para almacenarlos en almacenes y exportarlos a diferentes países. La empresa exportadora almacena el cacao en bolsas de plástico o sisal y arpillera. A veces se requiere un secado adicional en este punto.(Beg et al., 2017)

4.11.2. *Especies*

"Existen aproximadamente 22 especies de Theobroma, y cerca de 15 son utilizados por su pulpa comestible o semillas. El cacao es la especie más importante. *Theobroma grandiflorum* (cupuaçu), *Theobroma gileri* (cacao de montaña), *Theobroma bicolor* (Macambo) y subincanum *Theobroma* (cacao silvestre) son otras especies utilizadas por su pulpa dulce, comestibles y semillas comestibles".(Medina & Vargas, 2009, p. 12)

4.11.3. *Variedades de cacao*.

Los tipos de cacao se clasifican en tres grupos principales: criollo, forastero y trinitario.(Medina & Vargas, 2009)

4.11.3.1. Criollo "El cacao criollo o cacao nativo es el desarrollado en el norte de América del Sur y América Central, son frutos de finas paredes, de color rojo o amarillo, algunos pueden ser verdes o blancos. Las semillas son grandes, redondas, de color blanco o púrpura pálido, no astringente. En general, los granos de cacao criollo se consideran que tienen un sabor más fino que el de otras variedades de cacao. Los árboles criollos no son muy resistentes a las enfermedades, y por lo tanto son problemáticos para los agricultores para hacerlos crecer y mantenerlos sanos".(Medina & Vargas, 2009, p. 20)

4.11.3.2. Forastero “El cacao tipo forastero son de la cuenca del Amazonas, y tienen una pared gruesa, fruta suave, generalmente de color amarillo. Las semillas son aplanadas y de color púrpura. El tipo de cacao forastero es muy productivo y es el que domina la producción de cacao en el mundo debido a su resistencia a las enfermedades”. El grano forastero tiene un sabor más pronunciado a "chocolate" y por eso se considera "granos a granel".(Medina & Vargas, 2009, p. 21)

4.11.3.3. Trinitario. “El tipo de cacao trinitario surgió en la isla de Trinidad, como un híbrido de los tipos criollo y forastero. Son muy variables, y se considera de alta calidad para la producción de chocolate. Al igual que con el tipo forastero, las vainas de cacao trinitario no están normalmente en punta, y la piel de las vainas es relativamente suave (en comparación con la de las vainas de los criollos). Los granos del cacao también son planos y de color púrpura cuando se cortan por la mitad y se ha extendido por todo el mundo como un cultivo muy importante de cacao” (Medina & Vargas, 2009, p. 23).

5. Metodología

En este capítulo se presenta de forma detallada las etapas de la metodología aplicada para identificar grupos de patentes afines a la industria del cacao por medio de técnicas de minería de texto.

5.1. Construcción del conjunto de datos

Para el desarrollo de este trabajo de investigación se realizó una búsqueda y recopilación de datos de patentes afines a la industria del cacao a través de la base de datos Orbit intelligence[1]. Orbit es una base de datos de patentes comercial, su sitio web disponible por Questel proporciona cobertura de texto completo de las colecciones de patentes de *Patent Cooperation Treaty* (PCT), asociadas a las bases de datos de China, Europa, Japón, Estados Unidos, y otras colecciones. De acuerdo con el esquema propietario de familias de patentes FamPat, los documentos de patentes se agrupan en registros de bases de datos. Por lo tanto, se puede acceder a los diferentes registros de una invención utilizando los datos que contiene cualquier documento de patente perteneciente a la familia de patentes FamPat asociada (WIPO, 2012). Se realizaron búsquedas avanzadas con las palabras claves "cocoa" y "chocolate", como se muestra en la Tabla 2, en todos los campos de las patentes que se han concedido.

En primera instancia, para el presente trabajo se llevó a cabo la descarga de los documentos de patentes que corresponden a los resultados recuperados de la búsqueda del 20/08/2020 según la Tabla 2, la cual no se tuvo restricción en ningún campo, lo que significa que se tenían en cuenta todas las patentes que existen en la base de datos usada con relación a la palabra clave "cocoa". Debido al volumen de documentos y campos de las patentes descargadas, no fue posible el procesamiento de los datos ya que el archivo total en formato csv pesaba aproximadamente 34 GB, por lo que se hizo necesario la aplicación de filtros o acotamientos por medio de la interface de la base de datos Orbit, para la ecuación de búsqueda de las patentes.

[1] https://orbit.com

Tabla 2.

Búsqueda de patentes

Base de datos	Ecuación de búsqueda	Número de resultados recuperados	Fecha
Orbit intelligence	(cocoa OR chocolate)/TI/AB/CLMS/DESC /ODES/OBJ/ADB/ICLM/KEYW/TX	224.768	19/06/2020
Orbit intelligence	(cocoa)/TI/AB/CLMS/DESC/ODES/OBJ /ADB/ICLM/KEYW/TX	424.423	20/08/2020
Orbit intelligence	(cocoa)/TI/AB/DESC/ODES/ICLM AND (STATE/ACT=ALIVE)	71.094	07/09/2020

Por lo anterior, los documentos de patentes comprenden una serie de elementos para analizar la información asociada a registros de patentes se puede agrupar en dos categorías; datos estructurados y no estructurados. Los datos estructurados de una patente es la información que se presenta en la primera página de un documento de patente (número de patente, fecha de publicación, inventores, etc.). En cambio, los datos no estructurados contienen diferentes estructuras y estilos, como los párrafos de texto (título, resumen, descripción y reivindicaciones). El análisis de datos no estructurados es un método novedoso de análisis de patentes, que se centra en datos de patentes desde un punto de vista científico y estratégico para la toma de decisiones competitivas(Y. Choi & Hong, 2020).

Como esta investigación se enfoca en un análisis cualitativo soportado por técnicas cuantitativas a los datos no estructurados de las patentes. La selección de documentos fue delimitada a aquellos campos como título, resumen, descripción, y reivindicaciones independientes, lo anterior aplicado a documentos activos o vigentes como lo muestra la ecuación de búsqueda de la fecha 07/09/2020 en la Tabla 2. Mediante esta estrategia se recuperaron 71.094 patentes de la base de datos, donde se descargaron los siguientes campos: título, resumen, descripción, reivindicaciones (dependiente

e independientes), para cada documento de patente. Como la plataforma Orbit permitía la descarga de 1000 registros de manera simultánea, se realizaron 72 descargas, estos documentos fueron unidos mediante el código de programación descrito en Figura 9.

Figura 9

Código de programación en Python usado para unir los 72 documentos

```
import OS
import pandas as pd
cwd = os.path.abspath('')
fils = os.listdir(cwd)

#MÉTODO PARA CONCATENAR LAS PATENTES DESCARGADAS EN 72 DOCUMENTOS DE EXCEL
df = pd.DataFrame()
for file in file:
  if file.endswith('.xlsx'):
    df=df.append(pd.read_excel(file, usecols= "A:F"), ignore_index= False)")
df.head()
df.to_excel('PATENTES_TOTAL.csv')
```

5.2. Preprocesamiento de datos

En la práctica, se ha encontrado que la preparación y la limpieza de los datos corresponde aproximadamente al 80% de la ingeniería total que se aplica a los datos. Lo anterior debido a que, una gran parte del trabajo en la minería de datos está basado en la existencia de datos de calidad, ya que estos constituyen la entrada de los algoritmos de la minería de datos, y estos algoritmos, asumen que los datos están bien distribuidos y que no contienen valores con error o valores faltantes. (S. Zhang, Zhang, & Yang, 2003)

Por esta razón, luego de obtener las 71.094 patentes con sus respectivos campos, se procede a realizar el preprocesamiento de los datos. Para ello, se construye un *notebook* de Jupyter en el cual se ejecutan funciones del lenguaje de programación R, buscando así retirar y sustituir términos con

el fin de eliminar valores con errores a la par de encontrando raíces de términos. Este proceso se llevó a cabo para garantizar que los datos no cuenten con patrones que puedan ser inútiles al momento de su procesamiento ya que pueden estar incompletos o pueden ser inconsistentes; este preprocesamiento genera datos de calidad, lo que lleva a generar patrones de calidad que permiten purificar la información, corregir errores, eliminar valores atípicos y reducir la ambigüedad de la información. Del mismo modo, como se menciona anteriormente, este proceso genera un dataset más pequeño que el original lo que puede mejorar significativamente la eficiencia de la minería de datos (S. Zhang et al., 2003). En la Figura 10 se adjunta el código utilizado durante esta etapa, el proceso consiste de los siguientes pasos:

- Retirar las *stopwords* (conectores, pronombres y artículos) de la lengua inglesa puesto que las patentes están en este idioma.
- Retirar la puntuación de los textos dentro de las patentes.
- Retirar números.
- Retirar espacios en blanco excesivos.
- Sustituir todos los términos a su contraparte en minúscula.
- Retirar términos relacionados a links de páginas web como lo pueden ser: “http://”, “https://”, etc.
- Extraer solo las raíces de los términos para evitar redundancias.

Figura 10

Código de programación en R usado durante la etapa de preprocesamiento

```
#PREPROCESAMIENTO DE DATOS
# Eliminar palabras vacias
DATA <- removeWords(DATA, stopwords("english"))
#Nos deshacemos de la puntuación
DATA <- removePunctuation(DATA)
#Remover los números
DATA <- removeNumbers(DATA)
#Eliminar los espacios vacios
DATA <- stripWhitespace(DATA)
#Texto en minúscula
DATA <- tolower(DATA)

PALABRAS <- read.csv("stopwords.csv")
DATA <- removeWords(DATA, PALABRAS$i)
DATA <- gsub("http://", "", DATA) #links
DATA <- gsub("https://", "", DATA) #links
DATA <- gsub("www.", "", DATA) #links
DATA <- gsub("#", "", DATA) #NÚmerales
DATA <- gsub("%", "", DATA) #Porcentaje
DATA <- gsub("_", "", DATA) #Guión al piso
DATA <- gsub("*", "", DATA) #Asterisco
DATA <- gsub("()", "", DATA) #Paréntesis
DATA <- gsub("the", "", DATA) #Palabra
DATA <- gsub(read.csv("stopwords.csv"), "", DATA) #Stopwords
DATA <- stemDocument(DATA, "english") #Raíces de palabras
```

5.3. Minería de textos

La metodología para determinar los grupos de patentes afines se basa en la aplicación del algoritmo k-means a una matriz de documentos y términos transformada. Lo anterior implica que es necesario transformar el conjunto de patentes una vez depurada en una estructura matricial afín, esto se logra mediante tres grandes etapas: Construcción del Corpus, estructuración de la Matriz TF-IDF, y agrupamiento de documentos.

5.3.1. *Construcción del Corpus*

Después de aplicar la limpieza de texto a los datos de las 71.094 patentes recuperadas, se procede a convertir en formato de un vector a todos los documentos dado que hasta el momento hacen parte de una lista y finalmente a un *corpus.* Un Corpus es una colección de documentos que contienen texto en lenguaje natural (Perry, 2017). En la Figura 11 se describe el Código implementado en R para la construcción del corpus.

El corpus está diseñado como la "librería" original de documentos, que se han convertido a texto sin formato codificado en UTF-8 y se han guardado a nivel de corpus y de documento junto con los metadatos. En términos de procesamiento y análisis, el corpus está diseñado como un contenedor de texto más o menos estático. Esto significa que el texto en el corpus no está diseñado para ser modificado internamente, por ejemplo, limpiando o procesando previamente. Lo más importante es que el texto se puede extraer del corpus durante el procesamiento y asignarlo a nuevos objetos, pero la idea es mantener el corpus como una copia de la referencia original para otros análisis, como la puntuación e indexación de análisis en el mismo corpus. (Benoit et al., 2018)

Figura 11

Código de programación en R usado para durante la construcción del corpus

```
#CREAR CORPUS
corpus <- Corpus(VectorSource(DATA))
corpus
```

5.3.2. *Estructuración de la Matriz TF-IDF*

Una vez construido el vector de unidades de lenguaje, se procede a generar una matriz de término-frecuencia. Esta matriz recibe el nombre de TF-IDF por sus siglas en inglés *(Term Frequency-*

Inverse Document Frequency). Se aplica esta métrica debido a que es una de las más reconocidas en cuánto a precisión al momento de clasificar los datos, puesto que considera la importancia de cada palabra usando dos características: La frecuencia del término y en cuántos archivos puede ser encontrado el término (Hakim, Erwin, Eng, Galinium, & Muliady, 2014).

TD-IDF es una técnica que clasifica a un término por su importancia y está basado en documentos y vectores de términos que representan la frecuencia del término y su presencia en un documento. Esta técnica se rige por la siguiente ecuación:

$$d^{(i)} = TF(w_i, d).IDF(w_i) \quad (7)$$

Donde,

$w_i = Término\ muestra.$

$d = Documento.$

$d^{(i)} = TFIDF\ del\ término\ wi\ en\ el\ documento\ d.$

$TF(wi, d) = Frecuencia\ del\ término\ wi\ en\ el\ documento\ d.$

$IDF(wi) = Frecuencia\ inversa\ del\ documento.$

La aplicación del algoritmo TF-IDF de un término (wi) en un documento (d) puede ser procesada usando ecuación (7) . La frecuencia del término TF(wi,d) es el conteo del término wi en un documento d. Un valor grande de la frecuencia del término indica que dicho término es prominente en cierto documento. La presencia de términos en múltiples documentos es suprimida debido a que tiende a asociarse con palabras vacías. Este proceso de eliminación se lleva a cabo con el segundo componente (IDF).

$$IDF(w_i) = \log\left(\frac{|D|}{DF(w_i)}\right) \quad (8)$$

Donde,

$IDF(w_i) = Frecuencia\ Inversa\ del\ documento.$

$w_i = Término\ muestra.$

$|D| = Conteo\ total\ de\ documentos.$

$DF(w_i) = Conteo\ de\ documentos\ que\ contienen\ el\ término\ wi.$

Si un término está presente en todos los documentos el numerador es igual al denominador en la ecuación (8) y esto da como resultado que IDF(wi)= Log(1) lo que es igual a 0. Pero si un término está presente en un número menor de documentos relativamente implicaría que DF(wi) < |D|, y esto da como resultado que IDF(wi)= Log(>1) y esto genera un número positivo. El vector de la presencia del término fue usado para calcular IDF. Mientras que TFI-DF, identifica la importancia de los términos en cierto número de documentos (Morgan, 2019).

Al ejecutar este algoritmo se obtiene una matriz de término-frecuencia. Los valores de esta matriz hacen referencia a la frecuencia en que un término aparece en dicho documento de la siguiente manera:

Tabla 3.

Ejemplo matriz TF-IDF

	Documento 1	**Documento 2**	**Documento 3**	**Documento 4**
Method	0.601	0.623	0.601	0.601

Composit	0.248	0.301	0	0.301
Food	0.124	0	0.248	0
Chocolate	0	0.265	0	0

En la anterior tabla vemos como ejemplo, el término “Food” aparece 0.124 veces en el documento 1 mientras que este mismo término no aparece en el documento 2.

Es importante resaltar que, si el conjunto de términos en el total de los documentos analizados es amplio, la matriz TF-IDF contendrá una gran cantidad de vectores con múltiples valores de 0, máxime, cuando un término es usado en un único documento. La poca presencia de términos en los documentos se denomina *sparsity* (escasez), esta es una medida que calcula la cantidad de elementos que contienen 0 en la matriz de término-frecuencia (McCullagh & Polson, 2018). Cuando una matriz tiene altos valores de sparsity (normalmente medido de manera porcentual) implica que hay una gran cantidad de palabras poco frecuentes, una matriz con estas características puede ser simplificada y reducida de tamaño, puesto que, al eliminar términos poco frecuentes o huérfanos, la pérdida de información es baja. Durante el presente proyecto se propone remover los términos con un *sparsity* del 99% (es decir, que son cero en el 99% de los documentos); al remover estos términos se logra reducir la cantidad de términos pasando de 1'640.120 a 579 términos, reduciendo considerablemente la carga computacional o consumo de recursos. Es de importancia resaltar que, la matriz de 579 términos y 71.904 documentos precisa de alrededor de 40 GB de memoria RAM para su procesamiento mediante el algoritmo k-means, En la Figura 12 se registra la estimación de los recursos computacionales en un log de Google Colab.

Figura 12

Estimación de los requerimientos de procesamiento de la matriz TF-IDF

```
Error: cannot allocate vector of size 40.0 Gb
Traceback:
```

5.3.3. *Selección del algoritmo y distancia*

Con el fin de seleccionar el algoritmo idóneo para el agrupamiento, así como su distancia, y teniendo bajo consideración las restricciones en cuanto a requerimientos de cómputo, se propone realizar una comparación tipo benchmarking a una muestra de documentos, en total se analizaron 30902 documentos, tamaño seleccionado a partir del cálculo de una muestra aleatoria con un margen de error del 2% y un nivel de confianza del 99%.

5.3.4. *Determinación de la cantidad de clústeres*

Teniendo en cuenta que el presente trabajo se enfoca en la aplicación de un algoritmo de aprendizaje automático no supervisado, se propone seleccionar la cantidad de grupos mediante la estrategia denominada Método del codo, esta metodología consiste en utilizar los valores de la inercia obtenidos después de aplicar la técnica de K-means a los diferentes Clústeres (desde 1 a N Clústeres), siendo la inercia la suma de las distancias al cuadrado entre cada objeto del Clúster y su centroide (Moya, 2016). Se resalta nuevamente que, debido a las restricciones en cuanto a requerimientos de cómputo y que esta estrategia implica el cálculo recursivo de clústeres (lo cual consume recursos de manera considerable), esta metodología es aplicada la misma muestra aleatoria con la cual se realiza la selección del algoritmo y de la distancia.

5.3.5. *Agrupamiento de documentos*

Con el fin de extraer información relevante de los documentos de patentes, se propone el uso de la técnica de k-means. No obstante, debido a limitantes computacionales mencionadas previamente,

en este proyecto no fue posible agrupar los 71.904 documentos de manera simultánea haciendo uso de un computador doméstico, e incluso de servidores en la nube como Google Colab (Intel(R) Xeon(R) CPU @ 2.20GHz, 12.753184 GB de RAM, Tesla K80 con 12 GB de GDDR5 VRAM). Por consiguiente, se propone sortear esta limitación dividiendo el conjunto total de patentes (71.904) en 18 subgrupos (ya que en una etapa anterior fue posible hacer cálculos con grupos de aproximadamente 4000 patentes) y aplicándole el algoritmo k-means a cada uno de los subgrupos, para posteriormente unificar los resultados.

Es importante resaltar que, fue necesario congelar los generadores de números pseudoaleatorios con el fin de que cada subconjunto de patentes iniciara con la misma cantidad de centroides en la misma ubicación espacial. Si bien este proceso no es exactamente igual a aplicar un agrupamiento a todos los documentos de manera simultánea y puede existir pérdida de información, todas las patentes distribuidas en todos los subgrupos comparten el mismo sistema de coordenadas y comparten los mismos centroides iniciales.

6. Resultados

A continuación, se presentan los resultados obtenidos del procesamiento de datos de los documentos de patentes divididos en las siguientes categorías: Conjunto de datos, Selección del algoritmo y distancia, Cantidad de clústeres, e Identificación de los grupos de patentes afines en la industria del cacao. Este trabajo se realizó conjuntamente utilizando una maquina local con las siguientes especificaciones: Procesador: Intel® Core ™ i7-8750H CPU @ 2.20GHz 2.21GHz, Memoria instalada (RAM): 16.0 GB, Sistema operativo: Windows 10 Pro.

6.1. Conjunto de datos

Como resultado de la construcción del conjunto de datos de este trabajo de investigación, después de concatenar los 72 paquetes descargados en la base de datos de patentes como formato de Excel, se obtiene un archivo como base de datos en formato csv (*comma-separated values*) que pesa 2.47 GB. Esta base de datos consta de 6 campos o columnas, cada columna representa una variable y cada fila corresponde al documento de patente (registro único). En promedio el conjunto de datos cuenta con 5963 caracteres en total el cual fue publicado en Kaggle[2] de manera pública para el acceso a todas las personas. En la Tabla 4 se describen las diferentes variables de la base de datos de patentes.

Tabla 4.

Variables de la base de datos de patentes

Variable	Descripción
Título	Nombre de la patente
Abstract	Describe la invención
Description	Antecedentes de la invención y retrata el estado de la técnica hasta ese momento
Claims	Define el alcance de la protección buscada por el solicitante
Independent claims	Características técnicas que definen el concepto inventivo y el objetivo de ella
Dependent claims	Características técnicas o limitaciones adicionales de la que depende

6.2. Selección del algoritmo y distancia

Con el fin de seleccionar el algoritmo de agrupación y técnica de cálculo de distancia más adecuados, se procede a evaluar el rendimiento de cada uno en una muestra aleatoria de patentes

[2]https://www.kaggle.com/pmrodriguez3/patentes-de-cacao

de 30902 documentos. Obteniendo como resultado que el algoritmo de agrupamiento y la técnica de cálculo de distancia más adecuados para este problema fueron “Hartigan” y “Manhattan” respectivamente. Esto, porque al evaluar el rendimiento del algoritmos y distancia presentan métricas significativas las cuales se comparan en la Tabla 5 y la

Tabla ***6***. Como no se encontraron mayores diferencias entre los algoritmos, se selecciona el algoritmo el algoritmo de Hartigan el cual se caracteriza por su eficiencia y se encuentra documentado en anteriores trabajos realizados dentro del grupo de investigación OPALO.

Tabla 5.

Métrica suma de cuadrados entre clúster

Algoritmo	Euclidiana	Manhattan	Minkowski (p=3)
Hartigan	24494.9	401710.3	17735.5
Lloyd	24494.9	401710.3	17735.5
Forgy	24494.9	401710.3	17735.5
MacQueen	24494.9	401710.3	17735.5

Tabla 6.

Métrica de variabilidad

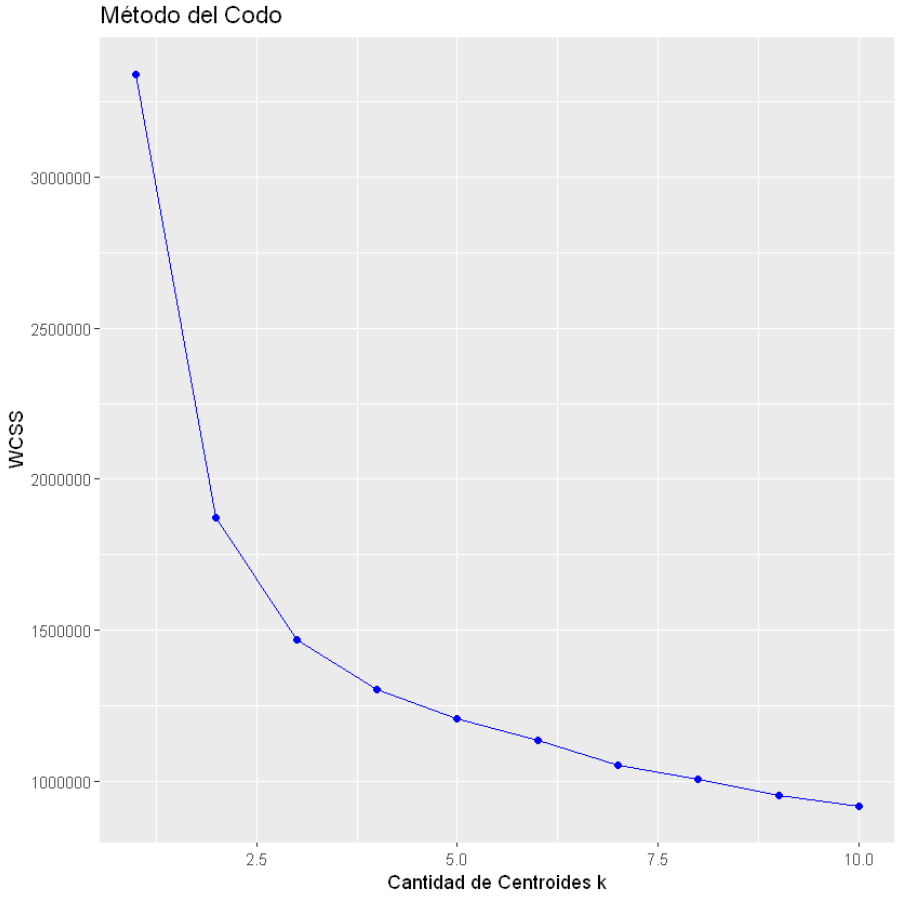
Algoritmo	Euclidiana	Manhattan	Minkowski (p=3)
Hartigan	87.71%	95.24%	82.9%
Lloyd	87.71%	95.24%	82.9%
Forgy	87.71%	95.24%	82.9%
MacQueen	87.71%	95.24%	82.9%

6.3. Cantidad de clústeres

Una vez se aplica la técnica de manera iterativa el algoritmo k-means a la muestra aleatoria, se determina que el un número de centroides a escoger es 4, número el cual fue seleccionado luego de aplicar el método del codo, lo anterior teniendo en cuenta que disminuye la variabilidad al aumentar el número de clústeres a seleccionar no es tan drástica (ver Figura 13).

Figura 13

Cantidad óptima de clústeres



6.4. Identificación de los grupos de patentes afines en la industria del cacao

Como primer resultado, se presentan 4 clústeres de documentos. En la Tabla 7. se indica la cantidad de patentes pertenecientes a cada uno de los grupos identificados. Se encuentra que, para cada uno de los clústeres realizados a los 18 subconjuntos, Asociada a esta tabla se obtuvo un valor medio de variabilidad de 83.6%, lo que indica que la pérdida de información en cada uno de los 18 subgrupos fue aproximadamente inferior al 20%.

Tabla 7.

Distribución de los documentos

Clústeres	Cantidad
1	18.917
2	16.358
3	20.088
4	15.731

Teniendo en cuenta que los documentos asociados a cada uno de los cuatro clústeres comparten un uso de términos similar, se procede a leer los documentos de la siguiente manera. (i) se realiza una lectura y análisis de los títulos de cada uno de los 71.904 documentos, (ii) se realiza una lectura del resumen y los reclamos a una muestra aleatoria dentro de cada uno de los cuatro clústeres. La muestra aleatoria se selecciona con un margen de error del 5% y nivel de confianza del 99% para una cantidad de patentes total de 639 (Clúster 1), 635 (Clúster 2), 640 (Clúster 3), y 634 (Clúster 4). A continuación, se describen cada uno de los cuatros clústeres luego de la estrategia de lectura.

6.4.1. *Clúster 1*

Métodos y compuestos para el desarrollo de tratamientos de enfermedades

Esta agrupación se caracterizado por la presencia de patentes relacionadas con el tratamiento de enfermedades crónicas y severas como lo puede ser el cáncer, así mismo, se mencionan métodos

y sistemas que se utilizaron para la invención de estos tratamientos, teniendo en cuenta su aplicación en diferentes escenarios, es decir, métodos para probar y dosificar componentes activos asociados al cacao. La utilización de diferentes derivados y sustancias son descritas según el tipo de padecimiento a tratar, por ejemplo, cáncer, problemas neurológicos, o enfermedades cutáneas. Del mismo modo, no solo se mencionan tratamientos, también se destacan los beneficios que se obtienen a partir del aprovechamiento de las capacidades inhibidoras del cacao y sus compuestos.

En la Figura 14 se presenta una nube de palabras con los términos más comunes en los documentos de este grupo, de manera similar, en la Figura 15 se indican los términos más frecuentes, por su parte, en la Tabla 8 se indica un breve análisis de 10 patentes seleccionadas de manera aleatoria y que perteneces a este clúster. Para finalizar con el ejemplo, a continuación, se muestra un título, resumen y reclamo de una patente perteneciente a este grupo.

Figura 14

Nube de palabras del clúster 1

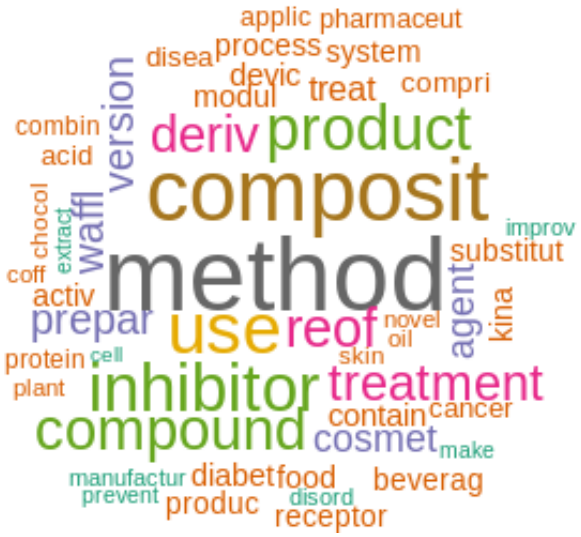


Figura 15

Gráfica de frecuencia de términos del clúster 1

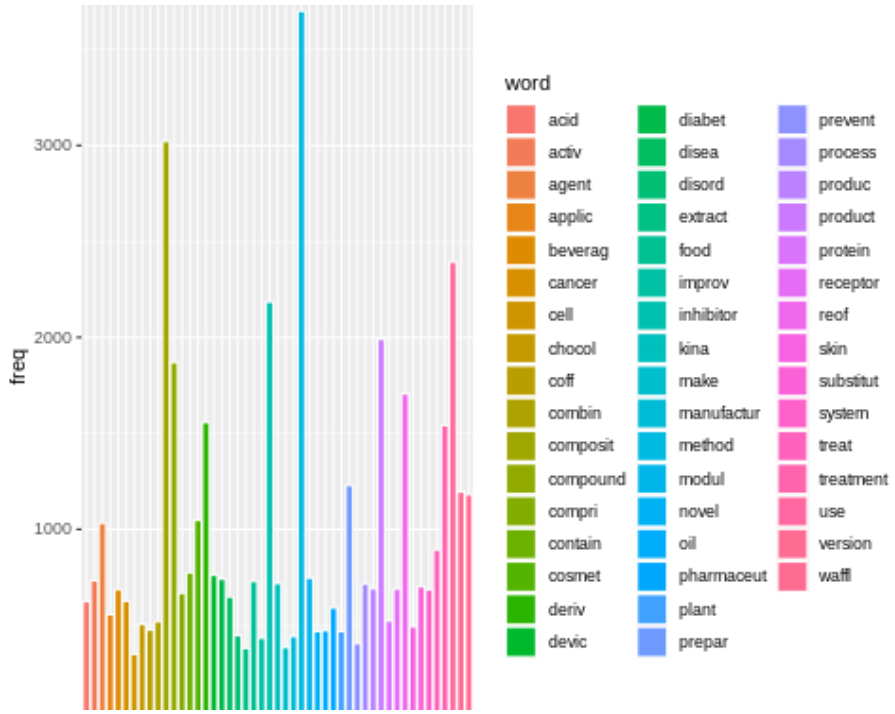


Figura 16

Ejemplo de patente del clúster 1

Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2479555 C 2013/02/05

(11)(21) **2 479 555**

(12) **BREVET CANADIEN**
CANADIAN PATENT

(13) **C**

(86) Date de dépôt PCT/PCT Filing Date: 2003/03/21
(87) Date publication PCT/PCT Publication Date: 2003/10/02
(45) Date de délivrance/Issue Date: 2013/02/05
(85) Entrée phase nationale/National Entry: 2004/09/16
(86) N° demande PCT/PCT Application No.: US 2003/008944
(87) N° publication PCT/PCT Publication No.: 2003/079998
(30) Priorité/Priority: 2002/03/21 (US60/366,363)

(51) Cl.Int./Int.Cl. *A61K 31/353* (2006.01), *A23G 1/00* (2006.01), *A23L 1/30* (2006.01), *A61K 31/352* (2006.01), *A61K 9/00* (2006.01), *A61P 9/06* (2006.01)

(72) Inventeur/Inventor:
SIES, HELMUT, DE

(73) Propriétaire/Owner:
MARS, INCORPORATED, US

(74) Agent: KIRBY EADES GALE BAKER

(54) Titre : UTILISATION DE FLAVANOLS DE CACAO ET D'OLIGOMERES DE FLAVANOLS DE CACAO POUR LE TRAITEMENT DU DYSFONCTIONNEMENT COGNITIF ET POUR AMELIORER LA FONCTION COGNITIVE CHEZ LES PATIENTS PRESENTANT UNE ATTEINTE NEUROLOGIQUE

(54) Title: USE OF COCOA FLAVANOLS AND OLIGOMERS THEREOF TO TREAT COGNITIVE DYSFUNCTION AND IMPROVE COGNITIVE FUNCTION IN NEURO-COMPROMISED PATIENTS

Nota: Adaptado de *Orbit.*

Title: *"Use of cocoa flavanols and oligomers thereof to treat cognitive dysfunction and improve cognitive function in neuro-compromised patients"* (Sies, 2003)

Abstract: ***"****This invention relates to compositions containing polyphenols, for example, cocoa polyphenols such as flavanols and their related oligomers, and methods for treating abnormalities in gap junctional communication of cells,, such as cancer, heart arrhythimia, neuro-degenerative diseases and cognitive dysfunction."* (Sies, 2003)

Claims*. " 1. Use, in the manufacture of a composition for treating cognitive dysfunction in a human or a veterinary animal, of a cocoa extract, chocolate liquor, cocoa cake, cocoa powder or cocoa nib, each comprising at least the following compounds: (a) a flavanol having the structure: and (b) a dimer composed of two units of the above flavanol connected via interflavan linkages 4.fwdarw.6 and/or 4.fwdarw.8; wherein the composition comprises at least 100 mg of the cocoa polyphenol compounds per unit of the composition."* (Sies, 2003)

2. "Use according to claim 1, wherein the composition is a pharmaceutical composition." (Sies, 2003)

3. "Use according to claim 1, wherein the composition is a food." (Sies, 2003)

4. "Use according to claim 1, wherein the cocoa polyphenol compounds further comprise at least one oligomer selected from trimers to pentamers composed of three to five units, respectively, of the above flavanol connected via interflavan linkages 4.fwdarw.6 and/or 4.fwdarw.8." (Sies, 2003)

5. "Use according to claim 1, wherein the cocoa polyphenol compounds further comprise at least one oligomer selected from trimers to decamers composed of three to ten units, respectively, of the above flavanol connected via interflavan linkages 4.fwdarw.6 and/or 4-8." (Sies, 2003)

6. "Use according to claim 1, wherein the composition is a confectionery." (Sies, 2003)

7. "Use according to claim 1, wherein the composition is a chocolate." (Sies, 2003)

8. "Use according to claim 1, wherein the composition is a beverage." (Sies, 2003)

9. "Use according to claim 1, wherein the composition is a dietary supplement." (Sies, 2003)

10. "Use according to claim 1, wherein the composition is a pet food." (Sies, 2003)

11 "Use, in the manufacture of a composition for improving a cognitive function in a subject suffering from a neurodegenerative disease, of a cocoa extract, chocolate liquor, cocoa cake, cocoa powder or cocoa nib, each comprising at least the following compounds: (a) a flavanol having the structure: and (b) a dimer composed of two units of the above flavanol connected via interflavan linkages 4.fwdarw.6 and/or 4.fwdarw.8; wherein the composition comprises at least 100 mg of the cocoa polyphenol compounds per unit of the composition; wherein the composition is formulated for administration in accordance with a regimen; and wherein the subject is a human or veterinary animal." (Sies, 2003)

12. "Use according to claim 11, wherein the composition is a pharmaceutical composition." (Sies, 2003)

13. "Use according to claim 11, wherein the composition is a food". (Sies, 2003)

14. "Use according to claim 11, wherein the cocoa polyphenol compounds further comprise at least one oligomer selected from trimers to pentamers composed of three to five units, respectively, of the above flavanol connected via interflavan linkages 4.fwdarw.6 and/or 4.fwdarw.8". (Sies, 2003)

15. "Use according to claim 11, wherein the cocoa polyphenol compounds further comprise at least one oligomer selected from trimers to decamers composed of three to ten units, respectively, of the above flavanol connected via interflavan linkages 4.fwdarw.6 and/or 4.fwdarw.8". (Sies, 2003)

16. "Use according to claim 11, wherein the composition is a confectionery." (Sies, 2003)

17. "Use according to claim 11, wherein the composition is a chocolate." (Sies, 2003)

18. "Use according to claim 11, wherein the composition is a beverage." (Sies, 2003)

19. "Use according to claim 11, wherein the composition is a dietary supplement." (Sies, 2003)

20. "Use according to claim 11, wherein the neurodegenerative disease is Alzheimer's disease." (Sies, 2003)

21. "Use according to claim 11, wherein the neurodegenerative disease is Parkinson's disease." (Sies, 2003)

22. "Use for the treatment of cognitive dysfunction in a human or a veterinary animal, of a cocoa extract, chocolate liquor, cocoa cake, cocoa powder or cocoa nib, each comprising at least the following compounds:(a) a flavanol having the structure:and (b) a dimer composed of two units of the above flavanol connected via interflavan linkages 4.fwdarw.6 and/or 4.fwdarw.8; wherein the composition comprises at least 100 mg of the cocoa polyphenol compounds per unit of the composition." (Sies, 2003)

23. "Use according to claim 22, wherein the compounds are in the form of a pharmaceutical composition." (Sies, 2003)

24. "Use according to claim 22, wherein the compounds are in the form of a food." (Sies, 2003)

25. "Use according to claim 22, wherein the cocoa polyphenol compounds further comprise at least one oligomer selected from trimers to pentamers composed of three to five units, respectively, of the above flavanol connected via interflavan linkages 4.fwdarw.6 and/or 4.fwdarw.8." (Sies, 2003)

26. "Use according to claim 22, wherein the cocoa polyphenol compounds further comprise at least one oligomer selected from timers to decamers composed of three to ten units, respectively, of the above flavanol connected via interflavan linkages 4.fwdarw.6 and/or 4.fwdarw.8." (Sies, 2003)

27. "Use according to claim 22, wherein the compounds are in the form of a confectionery." (Sies, 2003)

28. "Use according to claim 22, wherein the compounds are in the form of a chocolate." (Sies, 2003)

29. "Use according to claim 22, wherein the compounds are in the form of a beverage." (Sies, 2003)

30. "Use according to claim 22, wherein the compounds are in the form of a dietary supplement." (Sies, 2003)

31. "Use according to claim 22, wherein the compounds are in the form of a pet food." (Sies, 2003)

32. "Use for the improvement of cognitive function in a subject suffering from a neurodegenerative disease, of a cocoa extract, chocolate liquor, cocoa cake, cocoa powder or cocoa nib, each comprising at least the following compounds:(a) a flavanol having the structure:and (b) a dimer composed of two units of the above flavanol connected via interflavan linkages 4.fwdarw.6 and/or 4.fwdarw.8; wherein the composition comprises at least 100 mg of the cocoa polyphenol compounds per unit of the composition; wherein the composition is formulated for administration in accordance with a regimen; and wherein the subject is a human or veterinary animal". (Sies, 2003)

33. "Use according to claim 32, wherein the compounds are in the form of a pharmaceutical composition." (Sies, 2003)

34. "Use according to claim 32, wherein the compounds are in the form of a food." (Sies, 2003)

35. "Use according to claim 32, wherein the cocoa polyphenol compounds further comprise at least one oligomer selected from trimers to pentamers composed of three to five units, respectively, of the above flavanol connected via interflavan linkages 4.fwdarw.6 and/or 4.fwdarw.8." (Sies, 2003)

36. "Use according to claim 32, wherein the cocoa polyphenol compounds further comprise at least one oligomer selected from trimers to decamers composed of three to ten units, respectively, of the above flavanol connected via interflavan linkages 4.fwdarw.6 and/or 4.fwdarw.8." (Sies, 2003)

37. "Use according to claim 32, wherein the compounds are in the form of a confectionery." (Sies, 2003)

38. "Use according to claim 32, wherein the compounds are in the form of a chocolate." (Sies, 2003)

39. "Use according to claim 32, wherein the compounds are in the form of a beverage." (Sies, 2003)

40. "Use according to claim 32, wherein the compounds are in the form of a dietary supplement." (Sies, 2003)

41. "Use according to claim 32, wherein the neurodegenerative disease is Alzheimer's disease." (Sies, 2003)

42. "Use according to claim 32, wherein the neurodegenerative disease is Parkinson's disease." (Sies, 2003)

Tabla 8.

Ejemplos de patentes clúster 1

Título	Descripción
(WO2006079731) Use of cacao polyphenols for treating a prostate hyperplasia, a especific cacao extracts and applications	Esta patente habla sobre el uso de polifenoles de cacao para la prevención y el tratamiento de la hiperplasia de próstata, por medio de la compresión con distintos lípidos y xantinas, del mismo modo, como este compuesto es usado en la composición de productos farmacéuticos para el tratamiento de esta enfermedad y distintos desordenes cognitivos.(TROPLIN & BERNAERT, 2006)
(ES2331724) Compositions and methods to improve vascular health	Esta patente habla sobre una composición farmacéutica teniendo como compuesto base la procianidina de cacao y un compuesto reductor del nivel del colesterol basado en esterol o estanol, para la prevención y el tratamiento de diferentes afectaciones de la salud vascular, como la aterosclerosis y la enfermedad cardiovascular, este tratamiento está pensado para su aplicación en un mamífero, como puede ser un ser humano. (Kati A. Chevaux, Amy, Schmitz, & H., 2010)
(AU2015201692) Method to determine responsiveness of cancer to epidermal growth factor receptor targeting treatments	Esta intervención habla sobre un método para determinar una mayor probabilidad de respuesta a un tratamiento contra el cáncer, basado en un receptor del factor de crecimiento epidérmico (EGFR) por medio del aprovechamiento de las propiedades de la Xantina. Este tratamiento se centra específicamente en un individuo afectado con cáncer de pulmón y pretende determinar la reacción que este presenta en la cadena de polimerasa y la variación del gen receptor del factor de crecimiento epidérmico en el ADN. (*method to determine responsiveness of cancer to epidermal growth factor receptor targeting treatments*, 2006)
(US9629789) Rosacea treatments using polymetal complexes	Esta patente describe el mecanismo de acción de un método para tratar la rosacea, una afectación de la piel. Este tratamiento propone un régimen dividido en una etapa matutina y en una nocturna. La etapa de la mañana propone la limpieza del área afectada con un limpiador antimicrobiano o antibiótico, después, realizar la aplicación

	de una solución hidratante a base de Cu y Zn, esto ayudará al enrojecimiento y la limpieza definitiva, y por último sugiere complementar la protección con un filtro solar. A diferencia de esta sesión, la sesión nocturna no comprende la aplicación de una protección complementaria con un filtro solar.(Faryniarz, Ramirez, & Ounian, 2013)
(US20160143927) Compounds, Methods, and Treatments for Abnormal Signaling Pathways for Prenatal and Postnatal Development	Un compuesto eficaz para moderar la actividad de PKM2 a través de cascadas de señalización, para regular las vías metabólicas en tejidos de rápido crecimiento relacionados con enfermedades como el cáncer, la obesidad y la diabetes, incluido cualquier D-Chiro inositol. Del mismo modo, se refiere a métodos y composiciones para regular, antagonizar o estimular diferentes vías de transducción de señales, que pueden converger durante el desarrollo prenatal, el desarrollo postnatal y el desarrollo de organismos adultos para regular eventos de patrón y expresión génica.(Jennings, 2015)
(WO2000062631) Cocoa extract containing dietary fiber	Esta invención describe el desarrollo de un tratamiento contra la diabetes basado en un extracto de cacao que contiene fibra dietética, este extracto se obtiene por medio de un proceso de extracción aplicado a la cáscara del grano de cacao, la cual presenta una serie de características que logran mejorar distintos transtornos intestinales y metabólicos además de la diabetes. (Lee, Lee, & Kwon, 2000)
(US20200237871) Methods for mitigating liver injury and promoting liver hypertrophy, regeneration and cell engraftment in conjunction with radiation and/or radiomimetic treatments	Esta patente describe distintos métodos y kits para reducir el daño hepático y promover la regeneración y el trasplante de hígado en casos donde ha sido tratada con radioterapia dirigida. De este modo, habla de la utilización de imitadores de trombopoyetina, como RWJ-800088 o romiplostim y además de esto, la radiación realiza una gran parte del trabajo ya que reduce la enfermedad hepática y promueve efectos beneficiosos por medio de la regeneración y los trasplantes hepáticos.(EICHENBAUM & GUHA, 2020)
(US7265138) Vanilloid receptor ligands and their use in treatments	Esta patente habla sobre la implementación de compuestos llamados Valinoides para el tratamiento de afectaciones de diferentes tipos, como el dolor agudo, inflamatorio y neuropático, problemas vasculares, problemas óseos, diabetes, problemas gastrointestinales, alergias, problemas cutáneos, trastornos bronquiales y problemas gástricos. (Doherty et al., 2004)
(WO2018232448) c	Esta patente habla de un método que se centra en el tratamiento de distintos trastornos del sueño, habla sobre el desarrollo del compuesto y la utilización de distintos cannabioniodes y extractos naturales como aceite de cacao y su reacción a la hora de intervenir en el mejoramiento de afectaciones o trastornos del sueño. Esta propuesta surge de la necesidad continua de desarrollar nuevos tratamientos para tratar este tipo de afectaciones, pero con la particularidad de que se deriven de una fuente natural, ya que hay una creciente de pacientes que padecen este tipo de trastornos y esta se convierte en una alternativa sostenible.(GORDON, SMITH, WASHER, WASHER, & KARELIS, 2018)

(US8563066) Sustained release of nutrients in vivo	Esta patente habla sobre una serie de compuestos nutricionales basados en la epicatequina, molécula presente en el cacao, producidos con el fin de optimizar el rendimiento atlético por medio de dosis controladas suministradas en el acto, este tratamiento busca mejorar la coordinación del sujeto y lograr una mayor concentración.(Sexton, Krishnan, & Vendra, 2007)

6.4.2. *Clúster 2*

Métodos y compuestos para la fabricación de cosméticos

Este grupo hace referencia a la utilización de distintos componentes derivados del cacao, para el uso externo en el área de los cosméticos, así como en el cuidado general de la piel, resaltando las consecuencias de su utilización, sus beneficios, su reacción a agentes externos como la luz solar y su efecto a largo plazo. Del mismo modo, las patentes abordan los diferentes modos en los que estos componentes derivados pueden ser usados, es decir, su presentación al público, como exfoliantes, emulsiones, aceites y los cosméticos tradicionalmente conocidos.

En la Figura 17 se presenta una nube de palabras con los términos más comunes en los documentos de este grupo, de manera similar, en la Figura 18 se indican los términos más frecuentes, por su parte, en la Tabla 9 se indica un breve análisis de 10 patentes seleccionadas de manera aleatoria y que perteneces a este clúster. Para finalizar con el ejemplo, a continuación, se muestra un título, resumen y reclamo de una patente perteneciente a este grupo.

Figura 17

Nube de palabras del clúster 2

Figura 18

Gráfica de frecuencia de términos de clúster 2

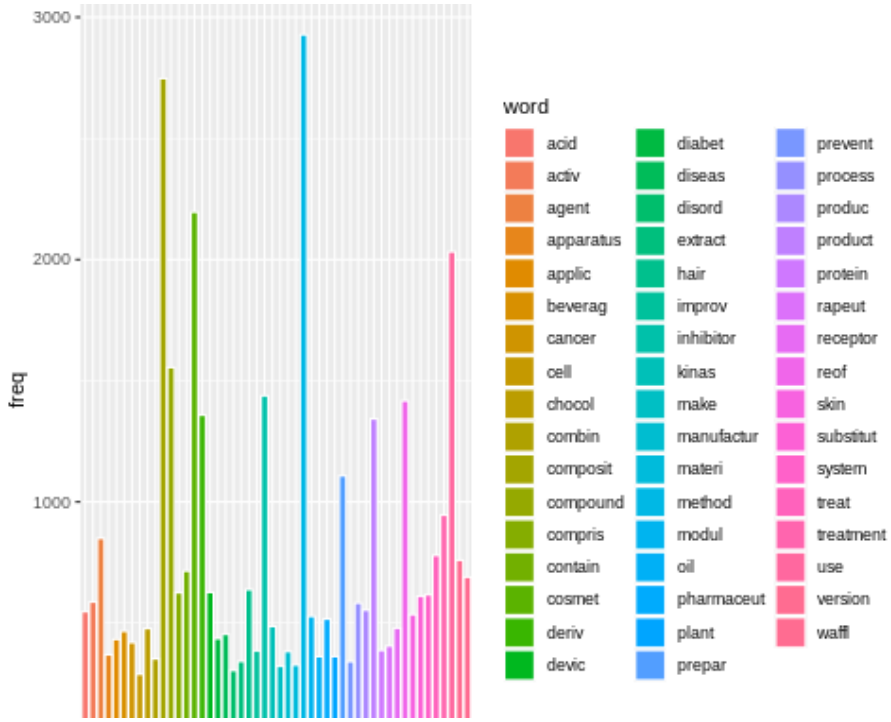
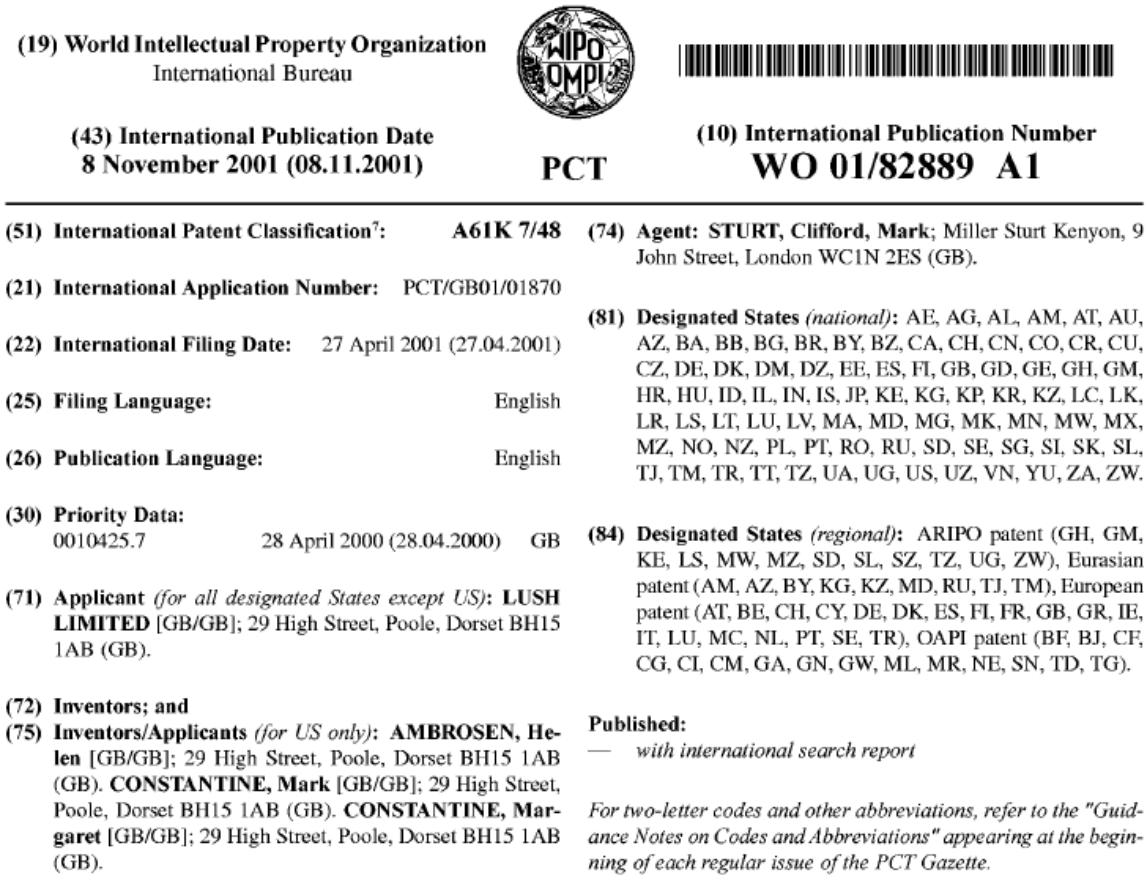


Figura 19

Ejemplo de patente del clúster 2



(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 November 2001 (08.11.2001)

PCT

(10) International Publication Number
WO 01/82889 A1

(51) International Patent Classification[7]: A61K 7/48

(21) International Application Number: PCT/GB01/01870

(22) International Filing Date: 27 April 2001 (27.04.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0010425.7 28 April 2000 (28.04.2000) GB

(71) Applicant *(for all designated States except US)*: **LUSH LIMITED** [GB/GB]; 29 High Street, Poole, Dorset BH15 1AB (GB).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **AMBROSEN, Helen** [GB/GB]; 29 High Street, Poole, Dorset BH15 1AB (GB). **CONSTANTINE, Mark** [GB/GB]; 29 High Street, Poole, Dorset BH15 1AB (GB). **CONSTANTINE, Margaret** [GB/GB]; 29 High Street, Poole, Dorset BH15 1AB (GB).

(74) Agent: **STURT, Clifford, Mark**; Miller Sturt Kenyon, 9 John Street, London WC1N 2ES (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Nota: Adaptado de Orbit.

Title: *"Cosmetic lotions comprising cocoa butter"* (Ambrosen, Constantine, & Constantine, 2001)

Abstract: *"A cosmetic lotion comprising 16 % to 76 % by weight cocoa butter and having a solid rather than liquid form. Also a method of manufacturing a cosmetic lotion having an oil based component and a water based component characterised by selecting the oil based component to comprise cocoa butter so that the cocoa butter constitutes 16 % to 76 % by weight of the final form of the lotion; heating the cocoa butter to a temperature in the range 55° to 70 °C; cooling the cocoa butter to a temperature in the range 35° to 25 °C*

and at a temperature in that range adding the water based component." (Ambrosen et al., 2001)

Claims. *1. "A cosmetic lotion comprising 16% to 76% by weight cocoa butter and having a solid rather than liquid form."* (Ambrosen et al., 2001)

2. "A cosmetic lotion as claimed in claim 1, comprising 30% to 75% by weight cocoa butter." (Ambrosen et al., 2001)

3. "A cosmetic lotion as claimed in claim 1, comprising 50% to 75 % by weight cocoa butter." (Ambrosen et al., 2001)

4. "A cosmetic lotion as claimed in claim 1, comprising 60% to 75% by weight cocoa butter." (Ambrosen et al., 2001)

5. "A cosmetic lotion as claimed in any preceding claim, wherein the lotion contains an oil based ingredient in addition to the cocoa butter." (Ambrosen et al., 2001)

6. "A cosmetic lotion as claimed in claim 5, wherein the said oil based ingredient is almond oil." (Ambrosen et al., 2001)

7. "A cosmetic lotion as claimed in any preceding claim, wherein the lotion contains a water based ingredient-comprising part of a fruit." (Ambrosen et al., 2001)

8. "A cosmetic lotion as claimed in any preceding claim, wherein the lotion contains a water based ingredient-comprising part of a vegetable." (Ambrosen et al., 2001)

9. "A cosmetic lotion as claimed in claim 1." (Ambrosen et al., 2001)

10. "A cosmetic lotion as claimed in claim 1." (Ambrosen et al., 2001)

11. "A method of manufacturing a cosmetic lotion having an oil based component and a water based component characterised by selecting the oil based component to comprise cocoa butter so that the cocoa butter constitutes 16% to 76% by weight of the final form of the lotion; heating the cocoa butter to a temperature in the range 55° to 70° C; cooling the cocoa butter to a temperature in the range 35° to 25° C and at a temperature in that range adding the water based component." (Ambrosen et al., 2001)

12. "A method as claimed in claim 11, further comprising the step of: at a temperature in the range 30° to 20°C pouring the mixed oil and water based components in to one or more moulds." (Ambrosen et al., 2001)

13. "A method as claimed in claim 12, further comprising the step subsequent to said step of pouring of chilling the lotion to a temperature in the range 20° to 10°C."

14. "A method as claimed in claim 11, comprising the steps of: heating the Cocoa Butter to approximately 60°C, emulsifying the water based ingredient at approximately 30°C, cooling the lotion to approximately 25°C, pouring the lotion in to one or more moulds and chilling the lotion at 16°C." (Ambrosen et al., 2001)

Tabla 9.

Ejemplos de patentes clúster 2

Título	Descripción
(WO2017157998) Peptide and saccharide hydrolysate of cocoa beans, cosmetic compositions containing same, and cosmetic uses of same	Esta patente habla de un método de preparación de composiciones cosméticas para el cuidado y la protección de la piel con respecto a los rayos UV y para el fortalecimiento de la piel ante la presencia de los signos de envejecimiento, por medio de un hidrolizsdo péptido de granos de cacao. (Coquet et al., 2017)
(WO2016201506) A skin-care composition and kit for and method of producing it	Esta patente habla del desarrollo de una crema nutritiva para el uso terapéutico en la piel, formada a base de aceite de macadamia, aceite de almendras, aceite de aguacate, aceite de coco, cacao, y otras sustancias de origen natural, logrando así, un método eficaz para nutrir la piel. (RIDDLE, 2016)
(CN106389161) Preparation method of cosmetic component with moisturizing efficacy	Esta patente habla del desarrollo de un componente cosmético teniendo como base el aloe y los granos de cacao, con la intención de integrar propiedades humectantes sobresalientes. Del mismo modo, describe detalladamente el proceso de desarrollo teniendo en cuenta las sustancias necesarias y los procesos que se deben llevar a cabo para su fabricación. (*Preparation method of cosmetic component with moisturizing efficacy*, 2016)
(JP2005232058) Skin care preparation for external use and fiber each having lipid decomposition promoting effect	Esta patente se da para resolver problemas cutáneos por medio de un producto de uso externo para el cuidado de la piel, que aumente la capacidad de absorción y humectación a través de distintos compuestos basados en los derivados de la xantina, alcohol isoestearílico, sebacato de diisopropilo y un dineopentanoato de alquilen poliglicol, todo esto, por medio de una fibra que contenga cada uno de estos compuestos de manera externa y por separado. (Abe, Koga, Okamoto, Oki, & Takeoka, 2004)

(JP2019123674) Oil-in-water emulsified sunscreen cosmetics	Esta patente surge por la necesidad del desarrollo de nuevos cosméticos que contengan mejores propiedades con respecto a la protección solar y que proporcionen una mejor respuesta en cuánto a su contacto con el agua, por esto, se creó un cosmético emulsionado de agua con aceite contando con la presencia de diferentes sustancias naturales, como el cacao y distintos aceites de origen vegetal, del mismo modo, estos cosméticos son a base de polímeros, un bloqueador de rayos UV compuesto de partículas inorgánicas hidrófobas y diferentes compuestos activos a diferentes temperaturas.
(CN109793672) Application of fish oligopeptide in cosmetics	Esta patente describe el proceso de desarrollo de cosméticos a base de oligopéptido de pescado, por parte de derivados de animales, y de oligopéptidos naturales sacados de los granos de cacao y la almendra logrando con esta aplicación mejores y nuevas propiedades con respecto a los cosméticos tradicionales. De este modo, se evidencia que al usarse este oligopéptido la capacidad de absorción y de humectación aumenta, del mismo modo, esta sustancia ayuda al proceso de regeneración y metabolismo de las células lo cual proporciona una ventaja al hablar de afectaciones por piel muerta o afectaciones por sustancias nocivas externas. (JIAHENG, CHUNCHEN, JINGBO, JUMAO, & YIHONG, 2018)
(JP2018162241) Milky lotion-like skin cosmetics	Esta patente se centra en el desarrollo de un producto cosmético similar a una loción humectante de tipo cremoso a base de aceites y grasas vegetales, en combinación con frutos naturales como el cacao, el eucalipto y distintos aceites de origen vegetal que, al ser mezclado con compuestos activos, alcoholes y óxidos, dan como resultado un producto cosmético de una textura uniforma y de finas partículas de emulsión para el cuidado externo de la piel.
(WO2019211470A1) Extract of shells of theobroma cacao beans for controlling rosacea and skin redness	Esta patente relata el desarrollo de una composición cosmética o dermatológica, para la prevención de distintas afectaciones de la piel, o para el tratamiento de condiciones como la rosácea, siendo así una aplicación en el campo dermatológico, esta composición se caracteriza por el aprovechamiento principal del extracto de cáscaras de granos de cacao. (ARIES & POIGNY, 2018)
(US20170100314) Thickened skin care product	Esta patente presenta la composición cosmética de productos para el cuidado de la piel sensible y delgada. Este producto está compuesto por un emulsionante que ayuda a mejorar la estabilidad de la piel y a regular la temperatura de las composiciones cosméticas por medio del aprovechamiento de las propiedades de composiciones de origen natural como el aceite de cacao y el aceite de almendras, logrando así un mejor tratado en la piel sensible, como es el caso del rostro.(Dickhof, Waldmann-Laue, Heinen, & Hartmann, 2015)
(US9877900) Use of cosmetics against infrared radiation	Esta patente presenta un método para el desarrollo de un compuesto cosmético con mejores propiedades en cuánto a la protección contra la radiación IR, esta composición se logra con la aplicación de extractos vegetales de semilla de café verde como principal compuesto natural, lo cual mezclado con las vitaminas E y C y distintos derivados, además de esto, se refuerza con el uso

de cosméticos auxiliares, conformando así un compuesto cosmético con propiedades sobresalientes al hablar de protección contra la radiación IR.(DOUCET, OLIVIER, PUJOS, MURIEL, ROBERT, CECILE, BERNINI, DOROTHEE, PISSAVINI, 2018)

6.4.3. *Clúster 3*

Métodos y compuestos para la fabricación de medicamentos

Este grupo se relaciona con la fabricación de distintos productos a base de compuestos, derivados, extractos y enzimas para el tratamiento de distintas enfermedades, desordenes y condiciones, así como también la fabricación de agentes de control en la búsqueda de la reducción del daño y el avance de estas enfermedades. Relata las condiciones que se presentan al usar estos compuestos, su proceso de fabricación, los intermediarios utilizados para esta fabricación y los métodos de aplicación contemplados para su uso. Convirtiendo así, estas aplicaciones en un factor importante en la industria farmacéutica partiendo del aprovechamiento de las propiedades, los compuestos y los derivados del cacao en la fabricación de medicamentos.

En la Figura 20 se presenta una nube de palabras con los términos más comunes en los documentos de este grupo, de manera similar, en la Figura 21 se indican los términos más frecuentes, por su parte, en la Tabla 10 se indica un breve análisis de 10 patentes seleccionadas de manera aleatoria y que perteneces a este clúster. Para finalizar con el ejemplo, a continuación, se muestra un título, resumen y reclamo de una patente perteneciente a este grupo.

Figura 20

Nube de palabras clúster 3

Figura 21

Gráfica de frecuencia de términos del clúster 3.

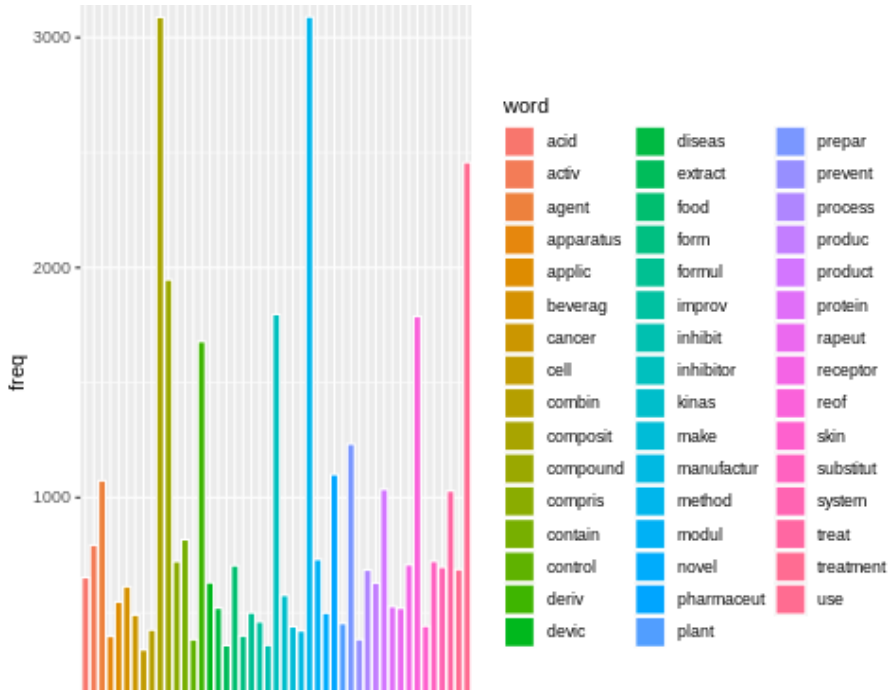
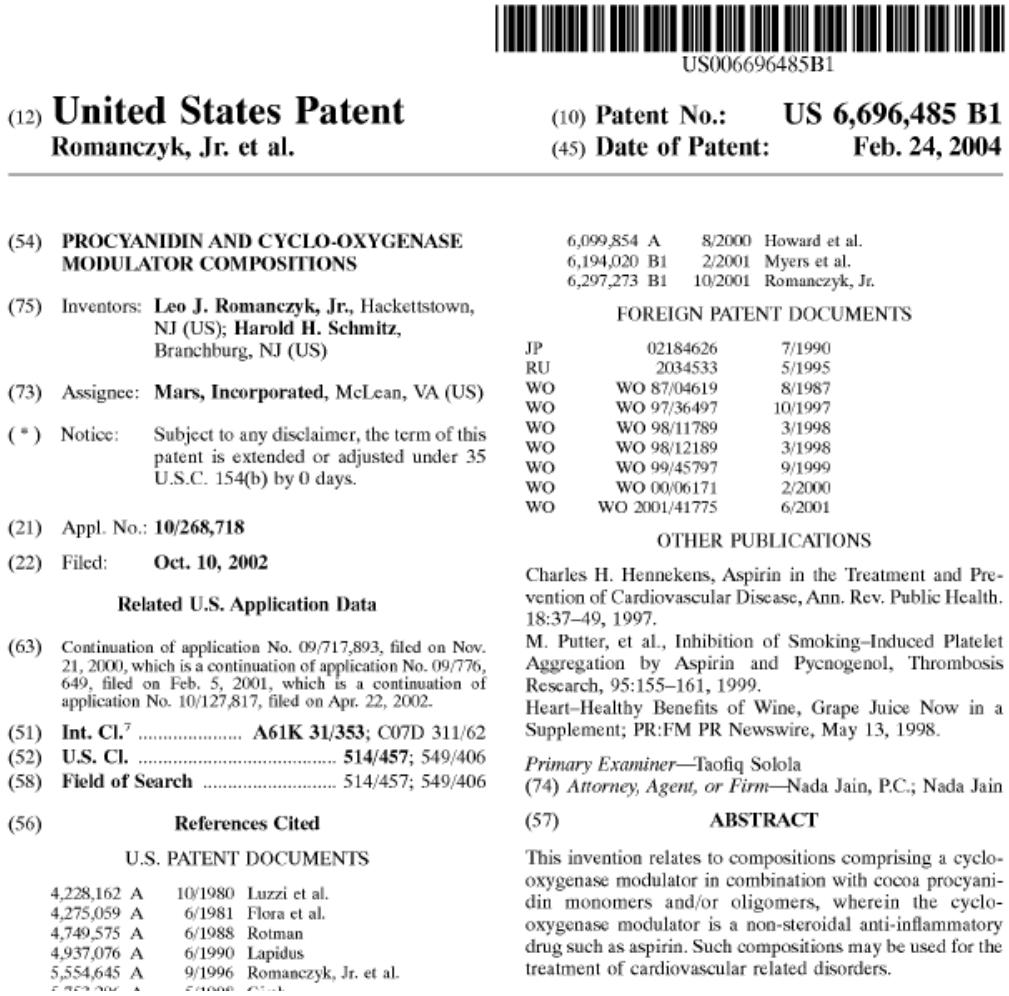


Figura 22

Ejemplo de patente del clúster 3



US006696485B1

(12) **United States Patent**
Romanczyk, Jr. et al.

(10) **Patent No.: US 6,696,485 B1**
(45) **Date of Patent: Feb. 24, 2004**

(54) **PROCYANIDIN AND CYCLO-OXYGENASE MODULATOR COMPOSITIONS**

(75) Inventors: **Leo J. Romanczyk, Jr.**, Hackettstown, NJ (US); **Harold H. Schmitz**, Branchburg, NJ (US)

(73) Assignee: **Mars, Incorporated**, McLean, VA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **10/268,718**

(22) Filed: **Oct. 10, 2002**

Related U.S. Application Data

(63) Continuation of application No. 09/717,893, filed on Nov. 21, 2000, which is a continuation of application No. 09/776,649, filed on Feb. 5, 2001, which is a continuation of application No. 10/127,817, filed on Apr. 22, 2002.

(51) **Int. Cl.**[7] **A61K 31/353**; C07D 311/62
(52) **U.S. Cl.** **514/457**; 549/406
(58) **Field of Search** 514/457; 549/406

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,228,162	A	10/1980	Luzzi et al.
4,275,059	A	6/1981	Flora et al.
4,749,575	A	6/1988	Rotman
4,937,076	A	6/1990	Lapidus
5,554,645	A	9/1996	Romanczyk, Jr. et al.
6,099,854	A	8/2000	Howard et al.
6,194,020	B1	2/2001	Myers et al.
6,297,273	B1	10/2001	Romanczyk, Jr.

FOREIGN PATENT DOCUMENTS

JP	02184626	7/1990
RU	2034533	5/1995
WO	WO 87/04619	8/1987
WO	WO 97/36497	10/1997
WO	WO 98/11789	3/1998
WO	WO 98/12189	3/1998
WO	WO 99/45797	9/1999
WO	WO 00/06171	2/2000
WO	WO 2001/41775	6/2001

OTHER PUBLICATIONS

Charles H. Hennekens, Aspirin in the Treatment and Prevention of Cardiovascular Disease, Ann. Rev. Public Health. 18:37–49, 1997.
M. Putter, et al., Inhibition of Smoking–Induced Platelet Aggregation by Aspirin and Pycnogenol, Thrombosis Research, 95:155–161, 1999.
Heart–Healthy Benefits of Wine, Grape Juice Now in a Supplement; PR:FM PR Newswire, May 13, 1998.

Primary Examiner—Taofiq Solola
(74) *Attorney, Agent, or Firm*—Nada Jain, P.C.; Nada Jain

(57) **ABSTRACT**

This invention relates to compositions comprising a cyclo-oxygenase modulator in combination with cocoa procyanidin monomers and/or oligomers, wherein the cyclo-oxygenase modulator is a non-steroidal anti-inflammatory drug such as aspirin. Such compositions may be used for the treatment of cardiovascular related disorders.

Nota: Adaptado de *Orbit.*

Title: *"Procyanidin and cyclo-oxygenase modulator compositions"*(Romanczyk & Schmitz, 2002)

Abstract: *"This invention relates to compositions comprising a cyclo-oxygenase modulator in combination with cocoa procyanidin monomers and/or oligomers, wherein the cyclo-oxygenase modulator is a non-steroidal anti-inflammatory drug such as aspirin. Such compositions may be used for the treatment of cardiovascular related disorders."* (Romanczyk & Schmitz, 2002)

Claims (14). "*What is claim is: 1.A composition comprising an effective amount of cocoa procyanidin monomer and/or oligomer in admixture with a cyclo-oxygenase modulator."* (Romanczyk & Schmitz, 2002)

2. "The composition of claim 1, wherein the cyclo-oxygenase modulator is a non-steroidal anti-inflammatory drug." (Romanczyk & Schmitz, 2002)

3. "The composition of claim 1, wherein the non-steroidal anti-inflammatory drug is an aspirin." (Romanczyk & Schmitz, 2002)

4. "The composition of claim 1, wherein the cocoa procyanidin is a dimer." (Romanczyk & Schmitz, 2002)

5. "The composition of claim 1, wherein the cocoa monomer and/or oligomer is in the form of a cocoa extract or cocoa procyanidin-containing fraction thereof." (Romanczyk & Schmitz, 2002)

6. "The composition of claim 1, wherein the monomer comprises epicatechin and the oligomer comprises an epicatechin-containing oligomer." (Romanczyk & Schmitz, 2002)

7. "A composition comprising an effective amount of a polymeric compound of the formula an or a pharmaceutically acceptable salt or derivative thereof." (Romanczyk & Schmitz, 2002)

8. "The composition of claim 7, wherein the cyclo-oxygenase modulator is a non-steroidal anti-inflammatory drug." (Romanczyk & Schmitz, 2002)

9. "The composition of claim 7, wherein the non-steroidal anti-inflammatory drug is an aspirin." (Romanczyk & Schmitz, 2002)

10. "The composition of claim 7, wherein n is 2." (Romanczyk & Schmitz, 2002)

11. "The composition of claim 9, wherein n is 2." (Romanczyk & Schmitz, 2002)

12. "The composition of claim 7, wherein n is 3-12." (Romanczyk & Schmitz, 2002)

13. "The composition of claim 9, wherein n is 3-12." (Romanczyk & Schmitz, 2002)

14. "The composition of claim 7, wherein A is epicatechin" (Romanczyk & Schmitz, 2002)

Tabla 10.

Ejemplos de patentes clúster 3

Título	Descripción
(CN107375079) Extraction method of cocoa seed extract and application of cocoa seed extract in hand cream	Esta patente se basa en la descripción del proceso de extracción del contenido de la semilla de cacao, comprendiendo los distintos procesos por los que tiene que pasar la semilla, para lograr como producto una crema para el cuidado de las manos, reparadora, antiflamatoria y antiarrugas. (*A kind of extracting method of cacao seed extract and its application in hand frost*, 2017)
(US8603547) Use of cocoa extract	Esta invención se centra en el control del peso de una persona por medio de la utilización del cacao mediante la extracción y fermentación para así lograr un contenido de polifenoles en estos, superior al 25%. (Bernaert & Allegaert, 2007)
(US8557867) Inhibitors of NCCa-ATP channels for therapy	La patente habla del desarrollo de métodos y composiciones químicas para el tratamiento de la hemorragia intraventricular o necrosis hemorrágica progresiva (NPH), causada por una afectación en la médula espinal. El proceso está basado en el uso de inhibidores de un canal de NCca-ATP, como los inhibidores de SUR1 O TRPM4, del mismo modo, estos compuesto buscan ayudar a tratar problemas del mismo tipo pero en bebés, tratando específicamente sus células cerebrales.(Simard, 2008)
(US20200113902) Compositions and methods for treating cancers with covalent inhibitors of cyclin-dependent kinase 7 (cdk7)	Esta patente se centra en el uso de inhibidores covalentes para lograr identificar sujetos que padecen diferentes tipos de cáncer y sus diferentes respuestas a un tratamiento usando este tipo de inhibidores, así mismo, se les hace seguimiento a estas composiciones, pero con la presencia o la ausencia de biomarcadores presentes en distintas clases de terapias usadas contra el cáncer.
(WO200880810) Composition comprising cocoa fibre	Esta invención habla sobre una composición de uso farmacéutico creada a base de fibra de cacao y compuesta por fibra de cacao y oligofructosa, que funciona de manera determinante al momento de tratar afectaciones estomacales como estreñimiento y tránsito de colón lento.
(US20190008824) Cocoa polyphenols and soluble dietary fiber for use in the treatment or prevention disorders associated with an above-normal number of granulocytes in a tissue	Esta patente habla sobre la implementación de una composición basada en al menos un polifenol de cacao unido a una fibra dietética soluble, como un posible tratamiento para una afectación de tejidos que tienen sus granulocitos por encima de lo normal.(Blanchard, Holvoet, Ran-Ressler, & Kuslys, 2019)

(US10500183) Acetylcholinesterase inhibitors for treatment of dermatological conditions	Esta patente se basa en la intervención de afecciones cutáneas por medio del inhibidor acetilcolinesterasa, usándolo de forma tópica u oral. Del mismo modo, habla de cómo el producto ataca directamente las áreas de afectación más aguadas por los ácaros, así como también los sitios donde puede haber presencia de estos, este tratamiento logra una remisión más completa de los signos y las consecuencias de estas afectaciones, logrando también tratar otros tipos de afectaciones cutáneas como la dermatitis, las erupciones, el acné y las irritaciones generales. (SPALLITTA, 2017)
(WO2009102121A2) Solid lipid nanoparticles for drug delivery, a production method therefor, and an injectable preparation comprising the nanoparticles	La presente invención se centra en la administración de fármacos por medio de una inyección conformada por nanopartículas no tóxicas y sin efectos secundarios en el cuerpo a intervenir, formando una cáscara hecha de un poloxámero, un núcleo de nanolípidos y una matriz liídica de una mezcla entre cacao en polvo y un fármaco general.(Na et al., 2009)
(WO2013141267A1) Polyphenol stabilizer, composition containing the stabilizer and processed product	Esta invención se basa en la creación de un producto farmacéutico, con la características de que posee un estabilizador de polifenoles derivados del cacao y catequinas sacadas del té por medio de procesos de temperatura y extracción, que suprime la reducción de procianidinas, lo cual cumple la función de estabilizar el organismo y resulta ser un gran beneficio para la salud en general.(*Polyphenol stabilizer, and composition and processed goods containing said stabilizer*, 2013)
(US8841449) Compounds useful as inhibitors of ATR kinase	Esta patente se centra en la fabricación de compuestos basados en la proteína quinasa ATR utilizadas como inhibidores y como farmacéuticamente se han intervenido estos compuestos para conformar métodos de tratamiento para diversas enfermedades, trastornos y afecciones. Del mismo modo relata los distintos procesos utilizados para fabricar los compuestos y como estas quinasas ejercen su función en su aplicación in vitro.(Charrier, Durrant, & Knegtel, 2014)

6.4.4. *Clúster 4*

Métodos y compuestos para la producción de chocolate

En este clúster que hace referencia a productos alimenticios que se pueden obtener a partir del cacao y sus diferentes presentaciones, como helados, galletas y pasteles. Así mismo, habla de la composición nutricional de los productos, de su proceso de preparación, de la relación del cacao como ingrediente principal y su combinación con diferentes productos para así lograr distintos resultados sacando el mayor provecho a las distintas propiedades que brinda el cacao, como su sabor y su textura.

En la Figura 23 se presenta una nube de palabras con los términos más comunes en los documentos de este grupo, de manera similar, en la Figura 24 se indican los términos más frecuentes, por su parte, en la Tabla 11 se indica un breve análisis de 10 patentes seleccionadas de manera aleatoria y que perteneces a este clúster. Para finalizar con el ejemplo, a continuación, se muestra un título, resumen y reclamo de una patente perteneciente a este grupo.

Figura 23

Nube de palabras del clúster 4

Figura 24

Gráfica de frecuencia de términos de clúster 4

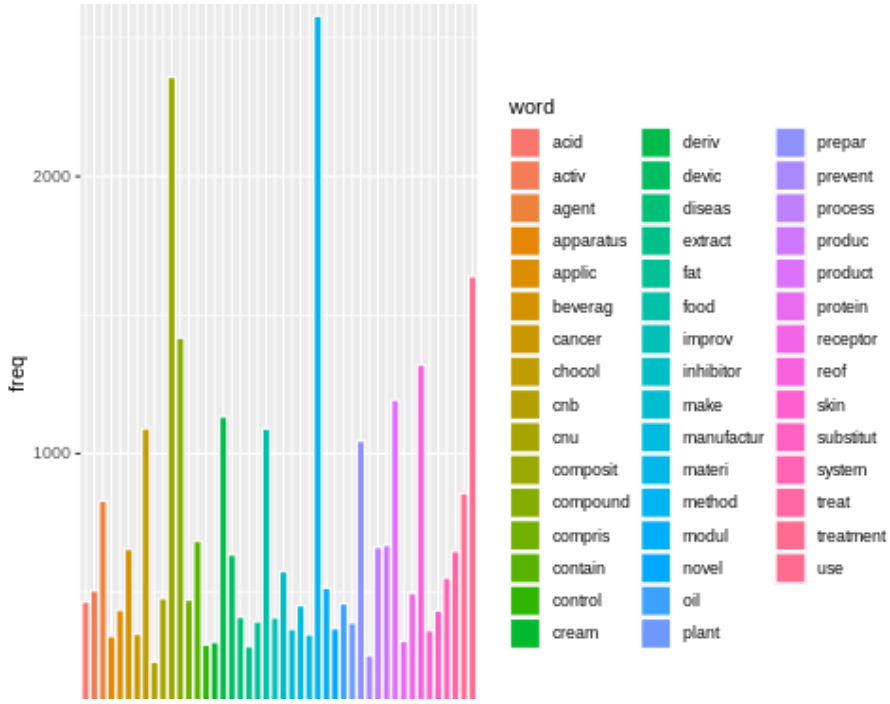
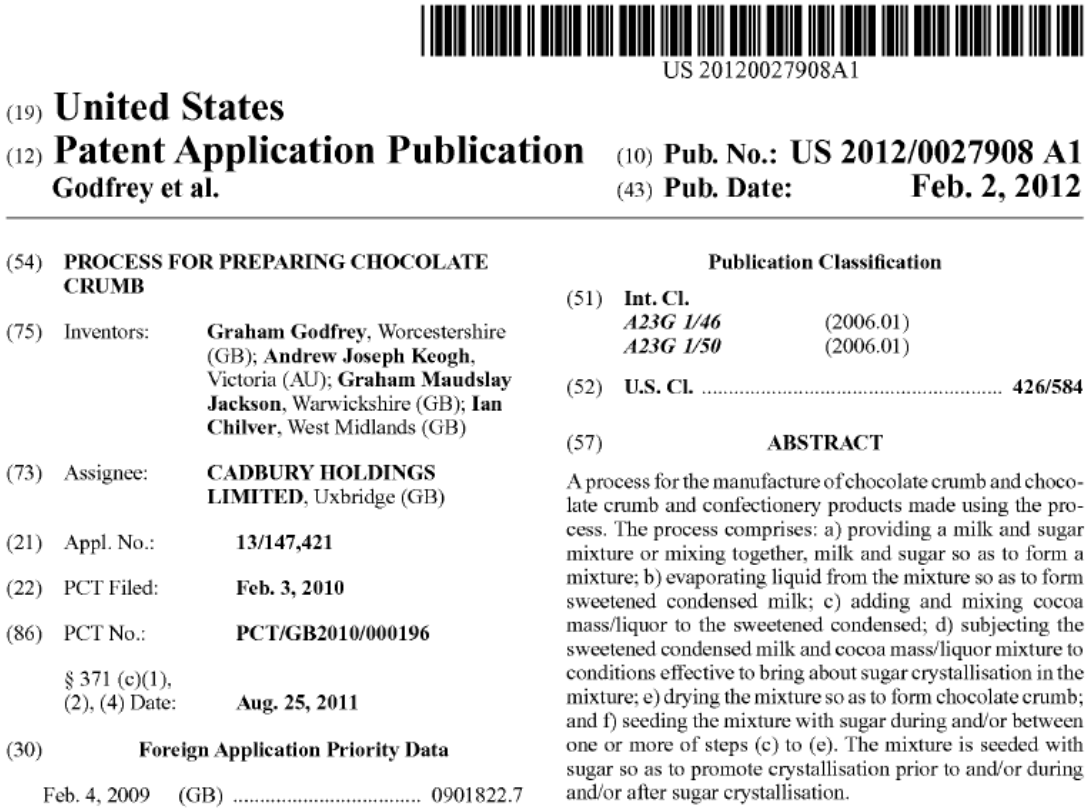


Figura 25

Ejemplo de patente del clúster 4



US 20120027908A1

(19) **United States**

(12) **Patent Application Publication**

Godfrey et al.

(10) **Pub. No.: US 2012/0027908 A1**

(43) **Pub. Date: Feb. 2, 2012**

(54) **PROCESS FOR PREPARING CHOCOLATE CRUMB**

(75) Inventors: **Graham Godfrey**, Worcestershire (GB); **Andrew Joseph Keogh**, Victoria (AU); **Graham Maudslay Jackson**, Warwickshire (GB); **Ian Chilver**, West Midlands (GB)

(73) Assignee: **CADBURY HOLDINGS LIMITED**, Uxbridge (GB)

(21) Appl. No.: **13/147,421**

(22) PCT Filed: **Feb. 3, 2010**

(86) PCT No.: **PCT/GB2010/000196**

§ 371 (c)(1), (2), (4) Date: **Aug. 25, 2011**

(30) **Foreign Application Priority Data**

Feb. 4, 2009 (GB) 0901822.7

Publication Classification

(51) **Int. Cl.**
A23G 1/46 (2006.01)
A23G 1/50 (2006.01)

(52) **U.S. Cl.** **426/584**

(57) **ABSTRACT**

A process for the manufacture of chocolate crumb and chocolate crumb and confectionery products made using the process. The process comprises: a) providing a milk and sugar mixture or mixing together, milk and sugar so as to form a mixture; b) evaporating liquid from the mixture so as to form sweetened condensed milk; c) adding and mixing cocoa mass/liquor to the sweetened condensed; d) subjecting the sweetened condensed milk and cocoa mass/liquor mixture to conditions effective to bring about sugar crystallisation in the mixture; e) drying the mixture so as to form chocolate crumb; and f) seeding the mixture with sugar during and/or between one or more of steps (c) to (e). The mixture is seeded with sugar so as to promote crystallisation prior to and/or during and/or after sugar crystallisation.

Nota: Adaptado de *Orbit.*

Title: *"Process for preparing chocolate crumb"*(Godfrey, Keogh, Jackson, & Chilver, 2012)

Abstract: *"A process for the manufacture of chocolate crumb and chocolate crumb and confectionery products made using the process. The process comprises: a) providing a milk and sugar mixture or mixing together, milk and sugar so as to form a mixture; b) evaporating liquid from the mixture so as to form sweetened condensed milk; c) adding and mixing cocoa mass/liquor to the sweetened condensed; d) subjecting the sweetened condensed milk and cocoa mass/liquor mixture to conditions effective to bring about sugar crystallisation in the mixture; e) drying the mixture so as to form chocolate crumb; and f) seeding the mixture with sugar during and/or between one or more of steps (c) to (e). The mixture is seeded with sugar*

so as to promote crystallisation prior to and/or during and/or after sugar crystallization." (Godfrey et al., 2012)

Claims. *"1. A process for preparing chocolate crumb comprising: a) providing a milk and sugar mixture or mixing together, milk and sugar so as to form a mixture; b) evaporating liquid from the mixture so as to form sweetened condensed milk; c) adding and mixing cocoa mass/liquor to the sweetened condensed; d) subjecting the sweetened condensed milk and cocoa mass/liquor mixture to conditions effective to bring about sugar crystallisation in the mixture; e) drying the mixture so as to form chocolate crumb; and f) seeding the mixture with sugar during and/or between one or more of steps (c) to (e)."* (Godfrey et al., 2012)

2. "A process as claimed in claim 1, wherein step (c) takes place after step (a) and/or after or during step (b)." (Godfrey et al., 2012)

3. "A process as claimed in claim 1, wherein step (f) comprises seeding the mixture with up to 25% of the total sugar in the chocolate crumb." (Godfrey et al., 2012)

4. "A process as claimed in claim 1, wherein the mixture is seeded with sucrose." (Godfrey et al., 2012)

5. "A process as claimed in claim 1, wherein step (b) comprises subjecting the mixture to heat." (Godfrey et al., 2012)

6. "A process as claimed in claim 5, wherein step (b) further comprises subjecting the mixture to a lowered pressure." (Godfrey et al., 2012)

7. "A process as claimed in claim 1, wherein the milk is formed from powdered milk and water." (Godfrey et al., 2012)

8. "A process as claimed in claim 1, wherein step (a) further comprises the addition of water." (Godfrey et al., 2012)

9. "A process as claimed in claim 1, wherein the milk comprises liquid milk." (Godfrey et al., 2012)

Tabla 11

Ejemplos de patentes clúster 4

Título	Descripción
(US8372456) Method for producing a soluble cocoa product from cocoa powder	Esta patente relata el desarrollo de un método para producir un producto de cacao soluble partiendo de una composición basada en el cacao en polvo, de este modo, describe las distintas etapas necesarias para la producción de diferentes productos obtenidos de este modo.(Bernaert, Blondee, & Clercq, 2013)
(US20040071848) Process for producing cocoa butter and cocoa powder by liquefied gas extraction	Esta patente describe el proceso de tratamiento que se le da a la masa de cacao para la obtención de productos de tipo manteca y sólidos, por medio de su disolución con hidrocarburos saturados para lograr una mezcla suspendida, del mismo modo, relata la afectación que puede llegar a sufrir esta masa en diferentes ambientes, como cambios de temperatura y su tipo de mezclado.(W., F., & C., 2004)
(BR102015021203) Procedure for obtaining cocoa, chocolates and other cocoa-based food products with a higher functional and sensory properties through fermentation of cocoa seeds in natural fruit juices with added microbial agents	La presente patente describe el proceso de obtención de distintos productos derivados del cacao, como licores, chocolates y otros productos alimenticios, potenciando sus propiedades funcionales por medio de la fermentación de la semilla en jugos de frutas naturales con agentes microbianos añadidos. Del mismo modo, describe otros procesos que le proporcionan a la semilla propiedades antioxidantes, por medio de la intervención en su maduración. (*Procedure for obtaining cocoa, chocolates and other cocoa-based food products with a higher functional and sensory properties through fermentation of cocoa seeds in natural fruit juices with added microbial agents*, 2016)
(CN107410625) Preparation method of cocoa butter chocolate products	La patente describe diferentes tecnologías utilizadas para la fabricación de alimentos y productos de chocolate a base de manteca de cacao. Del mismo modo, describe los distintos métodos de preparación para lograr determinada característica en el producto, como la *crocancia*, la suavidad al masticar, su fragancia y su calidad en general. Menciona distintas sustancias externas como la leche condensada y el aceite vegetal que son utilizados en estas mezclas para brindar propiedades nutritivas y convertirlas en fuentes de energía y calor.
(CN104187539) Method for preparing chocolate essence from cocoa bean husk	Esta invención habla sobre un método para la fabricación de esencia de chocolate teniendo como base la cáscara de la semilla de cacao. Describe de manera detallada el proceso por el que pasa la cáscara de la semilla de cacao para poder lograr una reacción que permita derivar de esta la esencia de chocolate.(*Method for preparing chocolate essence from cocoa bean husk*, 2014)
(CN107518139) Cocoa butter chocolates	Esta patente describe las generalidades de chocolates a base de manteca de cacao. De manera detallada, nombra los componentes necesarios para su fabricación y su respectiva cantidad en peso para su preparación. Todo esto, para lograr un producto final crujiente, rico en nutrientes y de buena calidad.(HUAXING, 2017)

(CN210184414U) Cocoa bean refiner for chocolate production	Esta patente se basa en la descripción del funcionamiento de una máquina refinadora de granos de cacao usada para la producción de chocolate. Logra detallar cada una de las secciones de este proceso por medio de esta máquina, y las distintas alteraciones que sufre el grano de cacao para convertirse finalmente en una materia prima garantizada.
(CN110679705) Preparation method of chocolate candy	La patente describe el desarrollo de un método de preparación de bombones de chocolate, que comprende desde su inicio, el conocimiento de los ingredientes necesarios, su proporción para ser aplicado en la mezcla y los procesos que sufren estos ingredientes a diferentes temperaturas, para lograr un producto almacenado y de gran calidad.(*Preparation method of chocolate candy*, 2017)
(CN101874538B) Method for preparing elastic chocolate sauce	La patente relata el proceso de preparación de una salsa de chocolate elástica, caracterizada por estar hecha a base de aceite, productos lácteos, y carbohidratos. Seguido a esto, describe el proceso de preparación y los distintos procesos a los que se ve sometido este compuesto.(*Method for preparing elastic chocolate sauce*, 2009)
(WO2020152112) Vegan chocolate	Esta patente el proceso para la obtención de chocolate vegano, teniendo como base una constituyente de grano de cacao, como la manteca de cacao o la masa de cacao, que al ser mezclada con edulcorantes y al usar harina de avena hidrolizada logra un compuesto capaz de generar el producto antes mencionado.(BACHMÜLLER & BOEKHAUS, 2020)

Discusión

El propósito del presente trabajo de investigación se centró en identificar grupos de patentes afines a la industria del cacao por medio del uso de técnicas de minería de textos, las cuales forman parte de la variedad de herramientas de análisis de datos y la inteligencia artificial, particularmente, de la rama del aprendizaje no supervisado. Este proyecto opta por este tipo de técnicas ya que permiten de manera matemática identificar similitud entre documentos, lo cual permite obtener información útil de información de datos no estructurados.

En primera instancia, para el desarrollo del presente trabajo se realizaron búsquedas avanzadas con las palabras claves “cocoa” y “chocolate” en todos los campos de las patentes que se han concedido, teniendo en cuenta el enfoque del proyecto fue necesario el uso de solo la primera

palabra clave mencionada anteriormente, la cual generó un total de 424.423 resultados que fueron descargados en 425 archivos de Excel pero debido al peso del archivo total en formato csv de aproximadamente 34GB, no fue posible el procesamiento de los datos para esta base de datos. Por esta razón, se hizo una segunda descarga de un total de 71.094 documentos de patentes, obtenidos de la aplicación de acotamientos en la ecuación de búsqueda de las patentes. La descarga de los documentos de patentes para el análisis se hace a través de la licencia de la Universidad Industrial de Santander (UIS) en una base de datos comercial de patentes que se conoce como Orbit Intelligence.

A partir de la revisión de literatura de los documentos científicos relacionados al análisis de patentes aplicando minería de texto, se encuentra que es común el uso de técnicas híbridas de agrupación y clasificación de datos, donde un gran componente tiene que ver la con lectura e interpretación semántico de los documentos bajo análisis, además, se encuentra que debido a la cantidad y variedad de documentos, los análisis de patentes de este estilo suelen enfocarse a industrias específicas en un horizonte determinado. En este caso en particular, se optó por no limitar el horizonte de búsqueda de las patentes y seleccionar datos no estructurados como almacenados en párrafos de texto (título, resumen, descripción y reivindicaciones). El análisis de datos no estructurados es un método novedoso de análisis de patentes, que se centra en datos de patentes desde un punto de vista científico y estratégico para la toma de decisiones competitivas. (Y. Choi & Hong, 2020). Como esta investigación se enfoca en un análisis cualitativo soportado por técnicas cuantitativas a los datos no estructurados de las patentes. La selección de documentos fue delimitada a aquellos campos específicos que se nombran anteriormente, lo anterior aplicado a documentos activos o vigentes. Por esta estrategia aplicada se lograron recuperar las 71.094 patentes mencionadas anteriormente.

Ahora bien, respecto a los aspectos cuantitativos, en la revisión de literatura se resalta el uso de algoritmos vectoriales y técnicas de agrupamiento basadas en aglomeración. Por lo tanto, en este proyecto se escogió la técnica conocida como k-means, debido a su capacidad de extraer información relevante de un gran volumen de datos técnicos a un bajo costo computacional. En este orden de ideas, de los antecedentes literarios se logran destacar los autores como Kim, Bae (2017), Kyebambe, Cheng, Huang, He, Zhang (2017), Choi, Ah, Shin (2019) Yi Zhang, Huang, Porter, Zhang, y Lu (2019) por sus estudios de minería de texto para el análisis de patentes empleando algoritmos no supervisados en el que llevan a cabo agrupaciones de documentos científicos por medio de la técnica k-means para identificar clústeres que comparten datos similares con el fin conocer las tendencias tecnológicas.

En cuanto a la generación de clústeres, este trabajo se enfrentó a las limitaciones computacionales para resolver el problema de agrupamiento, incluso considerando el uso de potentes herramientas de aprendizaje automático como Google Colab. Por ello, en el presente trabajo se tomaron dos grandes decisiones. La primera, eliminar términos poco frecuentes mediante la identificación de los *sparcity vectors*, con esto se redujo la cantidad de términos cuya composición mayoritariamente está definida por valores 0 (McCullagh & Polson, 2018). Por lo anterior, significa que cuando una matriz tiene un alto porcentaje de sparsity implica que existe una gran cantidad de palabras poco frecuentes, una matriz con estas características puede ser simplificada y reducida de tamaño, dado que, al eliminar términos poco frecuentes, la pérdida de información es relativamente baja. El segundo, proponiendo dividir el dataset en 18 subconjuntos y generando la agrupación de manera independiente para luego unirla. En este segundo caso, esta propuesta se realiza teniendo en cuenta que todos los documentos en sus respectivos subconjuntos comparten el mismo espacio vectorial donde se generan en la misma ubicación los cuatro puntos

iniciales, es decir, los cuatro centroides iniciales del algoritmo k-means no son aleatorios sino fijos. Esta estrategia si bien no es exactamente igual a hacer el agrupamiento con el conjunto de datos completo, se presenta como una alternativa viable en términos de eficiencia computacional, buscando un equilibrio entre eficiencia y eficacia.

En cuanto a los cuatro grupos identificados, se determina el uso de lecturas aplicadas a muestras aleatorias bajo el concepto de que todos los individuos dentro del mismo grupo son similares en su contenido basado en la frecuencia de los términos para cada clúster, estrategias similares fueron identificadas en la literatura por Lima, Argenta, Zattar y Klein (2019) encontrando los términos más frecuentes en los documentos y comprobando que el algoritmo del lenguaje de programación R es preciso en el análisis, y que el análisis del contenido generado con la minería de texto permitió una agrupación acertada. En cuanto a la información dentro de los grupos, se determina que todos los clústeres son afines a las siguientes características o palabras como: métodos para la fabricación, producción y preparación de los diferentes derivados del cacao, compuestos y propiedades que cuenta el fruto del cacao, inhibidor en componente químicos y medicamentos, como también las palabras productos y usos. No obstante, se lograron identificar las diferencias entre los grupos.

Conclusiones

A manera de conclusión, durante esta investigación se pudo apreciar como las diferentes técnicas de minería de texto han sido aplicadas para estudios de datos estructurados y no estructurados a lo largo de los años, por lo que ha sido de gran importancia en el campo de la investigación, dado que son herramientas precisas para el análisis de grandes volúmenes de datos, brindando información significativa acerca de tendencias, patrones y relaciones para el análisis del desarrollo tecnológicos, con el fin de lograr pronósticos y mejoramiento de estrategias competitivas en las

diferentes áreas. Para este proyecto se encontró de manera precisa que el cacao es un producto que está en constante innovación por sus diversos beneficios que aporta a la salud, dado a las propiedades curativas que tienen sus nutrientes y enzimas naturales. De esta misma manera, existe la mayor cantidad de patentes, es decir, la mayor cantidad de inventivas registradas pertenecen al clúster denominado: Métodos y compuestos para la fabricación de medicamentos

Es preciso resaltar que, a manera general las patentes expresan tendencias tecnológicas en los productos y procesos, y mediante su análisis se puede obtener un diagnóstico del desarrollo de un producto o técnica. Las patentes suponen un gran beneficio económico y científico para las empresas, dado que existen una gran cantidad de documentos de patentes las cuales van creciendo de manera acelerada. Por consiguiente, trabajos como el acá propuesto donde se mezclan metodologías cuantitativas y cualitativas permite un rápido análisis a grandes volúmenes de datos basándose en la similitud estructural de los términos que conforman las patentes. No obstante, para ello es necesario sortear el reto de trabajar con datos no estructurados para así generar las categorías. Debido a los resultados obtenidos, con los cuales fueron posible identificar información subyacente en las bases de datos de patentes consultadas, se determina que fue posible identificar grupos de patentes afines en la industria del cacao a partir de técnicas de minería de texto. Dando cumplimiento al objetivo general del presente proyecto.

De la metodología se concluye que el preprocesamiento de datos son grandes herramientas para reducir la capacidad computacional necesaria para trabajar un volumen de datos como este. De esta etapa se destaca que de 1'640.120 términos se pudo resaltar la importancia de un número significativamente menor de 579 términos, a los cuales se les aplica una matriz termino-frecuencia de tamaño considerablemente menor. También se resalta la facilidad de uso del lenguaje de

programación R y su gran variedad de paquetes y librerías, lo cual permite la visualización de los resultados de una manera muy didáctica.

Resaltando que en este trabajo se logró agrupar los documentos de patentes en cuatro grandes grupos denominados (i) Métodos y compuestos para el desarrollo de tratamientos de enfermedades, (ii) Métodos y compuestos para la fabricación de cosméticos, (iii) Métodos y compuestos para la fabricación de medicamentos y, por último, (iv) Métodos y compuestos para la producción de chocolate. Esta categorización se logra a partir de la identificación de semejanzas en los cuatros grupos propuestos mediante una lectura de documentos, lo que implica que estas metodologías cuantitativas (como minería de texto) apoya los procesos cualitativos (como búsqueda de contenido), disminuyendo el tiempo de revisión. Se resalta finalmente que, a pesar de identificar diferencias entre los cuatro grupos descritos, como es el área a la que se dedica según la fabricación de productos (como puede ser de cosméticos, de comidas y de salud), todas las patentes consultadas comparten similitudes, principalmente desde la descripción de propiedades químicas.

Recomendaciones

Sentando las bases para nuevas investigaciones se establecen una serie de recomendaciones con el fin de transmitir los aprendizajes adquiridos después de esta investigación. En primera instancia se sugiere la exploración de distintos algoritmos y sobre todo técnicas de análisis y/o minería de textos distintas a las usadas en esta tesis, con el fin de evaluar su desempeño en el estudio de patentes, entre ellas se recomiendan: máquinas de soporte vectorial, como lo recomendó la literatura; del mismo modo, K-means difuso, para adaptar más el algoritmo a inconsistencias en los datos; y el uso de una técnica hibrida, para extraer las ventajas de cada estrategia.

En segundo lugar, para disminuir la carga computacional sobre los dispositivos de cómputo y agilizar el proceso de análisis, se recomienda, la exploración del paradigma de programación en paralelo, mayormente, así como en este proyecto, se segmenta el conjunto de datos. Del mismo, modo se sugiere la exploración de otros lenguajes de programación tales como: Python y MATLAB.

En tercer lugar, a manera de trabajo futuro, se propone el desarrollo de una herramienta que además de crear los clústeres, estime la perdida de información en un conjunto de datos al aplicar técnicas como las aplicadas en este trabajo de grado, esto con el fin de dar a conocer la precisión de futuros trabajos afines.

En cuarta instancia, se recomienda aplicar los conocimientos aprendidos en esta tesis para otras familias de patentes, con la finalidad de evaluar la efectividad de estas técnicas en el análisis de patentes y al mismo tiempo, encontrar patrones en otros productos con patentes. Asimismo, el extender el alcance de futuros trabajos, usando no solo datos no estructurados sino también datos estructurados.

A manera de experimento, por último, se sugiere contrastar esta metodología con una de tipo cualitativa, la cual podría abrir nuevas puertas al análisis de patentes del mismo modo que en este trabajo de grado, y a la final comparar que concluye un análisis cuantitativo como el mostrado acá, con uno totalmente distinto usando análisis cualitativo.

Referencias Bibliográficas

A kind of extracting method of cacao seed extract and its application in hand frost. (2017).

Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information*, *37*, 3–13. https://doi.org/10.1016/j.wpi.2013.12.006

Abe, K., Koga, N., Okamoto, T., Oki, M., & Takeoka, E. (2004). *Skin care preparation for external use and fiber each having lipid decomposition promoting effect*.

Aggarwal, C. C., & Zhai, C. X. (2013). Mining text data. *Mining Text Data*, *9781461432*, 1–522. https://doi.org/10.1007/978-1-4614-3223-4

Altuntas, S., Dereli, T., & Kusiak, A. (2015). Analysis of patent documents with weighted association rules. *Technological Forecasting and Social Change*, *92*, 249–262. https://doi.org/10.1016/j.techfore.2014.09.012

Ambrosen, H., Constantine, M., & Constantine, M. (2001). *Cosmetic lotions comprising cocoa butter*.

An, J., Kim, K., Mortara, L., & Lee, S. (2018). Deriving technology intelligence from patents: Preposition-based semantic analysis. *Journal of Informetrics*, *12*(1), 217–236. https://doi.org/10.1016/j.joi.2018.01.001

ARIES, M. F., & POIGNY, S. (2018). *Extract of shells of theobroma cacao beans for controlling rosacea and skin redness*.

Aristodemou, L., & Tietze, F. (2018). The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. *World Patent Information*, *55*(February), 37–51. https://doi.org/10.1016/j.wpi.2018.07.002

Aristodemou, L., Tietze, F., Athanassopoulou, N., & Minshall, T. (2017). Exploring the Future of Patent Analytics: A Technology Roadmapping Approach. *R&D Management Conference 2017, Leuven, Belgium, November*(5), 1–9. https://doi.org/10.17863/CAM.13967

BACHMÜLLER, T., & BOEKHAUS, D. (2020). *Vegan chocolate*.

Balsmeier, B., Assaf, M., Chesebro, T., Fierro, G., Johnson, K., Johnson, S., … Fleming, L. (2018).

Machine learning and natural language processing on the patent corpus: Data, tools, and new measures. *Journal of Economics and Management Strategy*, *27*(3), 535–553. https://doi.org/10.1111/jems.12259

Bass, S. D., & Kurgan, L. A. (2010). Discovery of factors influencing patent value based on machine learning in patents in the field of nanotechnology. *Scientometrics*, *82*(2), 217–241. https://doi.org/10.1007/s11192-009-0008-z

Beg, M. S., Ahmad, S., Jan, K., & Bashir, K. (2017). Status, supply chain and processing of cocoa-A review. https://doi.org/10.1016/j.tifs.2017.06.007

Beltrán, B. (2014). *Minería de datos*. Retrieved from http://bbeltran.cs.buap.mx/NotasMD.pdf

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., … Lowe, W. (2018). Guía de Inicio Rápido • quanteda. Retrieved October 9, 2020, from European Research Council. website: https://quanteda.io/articles/pkgdown/quickstart_es.html

Bernaert, H., & Allegaert, L. (2007). *Use of cocoa extract*.

Bernaert, H., Blondee, I., & Clercq, lDirk De. (2013). *Method for producing a soluble cocoa product from cocoa powder*.

Blanchard, C., Holvoet, S., Ran-Ressler, R., & Kuslys, M. (2019). *Cocoa polyphenols and soluble dietary fiber for use in the treatment or prevention disorders associated with an above-normal number of granulocytes in a tissue*.

Cambronero, C. G., & Moreno, I. G. (2006). Algoritmos de aprendizaje: knn & kmeans. *Inteligencia En Redes de Comunicación, Universidad Carlos III de Madrid*.

Campbell, R. S. (1983). Patent trends as a technological forecasting tool. *World Patent Information*, *5*(3), 137–143. https://doi.org/10.1016/0172-2190(83)90134-5

Casanova, H. (2019). Las patentes como indicadores de innovación tecnológica | CAF. Retrieved January 23, 2020, from CAF website: https://www.caf.com/es/conocimiento/visiones/2019/08/las-patentes-como-indicadores-de-innovacion-tecnologica/

Čerka, P., Grigienė, J., & Sirbikytė, G. (2017). Is it possible to grant legal personality to artificial intelligence software systems? *Computer Law and Security Review*, *33*(5), 685–699.

https://doi.org/10.1016/j.clsr.2017.03.022

Charrier, J.-D., Durrant, S. J., & Knegtel, R. (2014). *Compounds useful as inhibitors of ATR kinase*.

Choi, W., Ahn, J., & Shin, D. (2019). Text mining geo-visualization of patent documents on geo-spatial big-data industry. *Spatial Information Research*, *27*(1), 109–120. https://doi.org/10.1007/s41324-018-0201-3

Choi, Y., & Hong, S. (2020). Qualitative and quantitative analysis of patent data in nanomedicine for bridging the gap between research activities and practical applications. *World Patent Information*, *60*, 101943. https://doi.org/10.1016/j.wpi.2019.101943

Colomé Abril, X. (2012). *Aproximación al reajuste automático de centroides mediante la heurística de Lloyd para resolver el problema de las K-Medias*. Retrieved from www.uoc.edu

Contreras, C. (2017). Análisis de la cadena de valor del cacao en Colombia: generación de estrategias tecnológicas en operaciones de cosecha y poscosecha, organizativas, de capacidad instalada y de mercado. *Universidad Nacional de Colombia, Facultad de Ingeniería.*, 221. Retrieved from http://www.bdigital.unal.edu.co/59141/1/1032373448-2017.pdf

Coquet, C., Gondran, C., Imbert, I., MANTELIN, J., Domloge, N., Garnier, S., … CICCHETI, E. (2017). *Peptide and saccharide hydrolysate of cocoa beans, cosmetic compositions containing same, and cosmetic uses of same*.

Dickhof, S., Waldmann-Laue, M., Heinen, S., & Hartmann, I. (2015). *Thickened skin care product*.

Doherty, E. M., Fotsch, C. H., Hungate, N. H. W., Liu, Q., Norman, M. H., Xi, N., & Xu, S. (2004). *Vanilloid receptor ligands and their use in treatments*.

Dou, H. J. M. (2004). Benchmarking R&D and companies through patent analysis using free databases and special software: A tool to improve innovative thinking. *World Patent Information*, *26*(4), 297–309. https://doi.org/10.1016/j.wpi.2004.03.001

DOUCET, OLIVIER, PUJOS, MURIEL, ROBERT, CECILE, BERNINI, DOROTHEE, PISSAVINI, M. (2018). *Use of cosmetics against infrared radiation*.

EICHENBAUM, G., & GUHA, C. (2020). *Methods for mitigating liver injury and promoting liver hypertrophy, regeneration and cell engraftment in conjunction with radiation and/or radiomimetic treatments*.

Espino, A.I.L., Mur, R.A. y de Miguel, M. A. S. (2004). *Aprendizaje Automatico En Conjuntos De Clasificadores Heteroge Neos Y Modelado De Agentes*. 150.

Fan, J., Fang, L., Wu, J., Guo, Y., & Dai, Q. (2020, January 14). From Brain Science to Artificial Intelligence. *Engineering*. https://doi.org/10.1016/j.eng.2019.11.012

FAO. (2014). Organizaciones de Naciones Unidas para la Alimentacion y la Agricultura. Retrieved November 14, 2019, from http://www.fao.org/fao-stories/article/es/c/1170985/

Faryniarz, J. R., Ramirez, J. E., & Ounian, H. (2013). *Rosacea treatments using polymetal complexes*.

Garre, M., Cuadraro, J., Sicilia, M. A., Rodríguez, D., & Rejas, R. (2007). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. *Revista Española de Innovación, Calidad e Ingeniería Del Software*, (January 2016).

Gavilanes, J., Rio, R. M., & Cilleruelo, E. (2010). *Aproximación al estudio de patentes. Indicadores utilizados en la minería de textos*.

Gavilanes, J., Rio, R. M., Cilleruelo, E., & Garechana, G. (2011). *Aplicación de la minería de textos. Análisis de patentes*.

Godfrey, G., Keogh, A. J., Jackson, G. M., & Chilver, I. (2012). *Process for preparing chocolate crumb*.

González, K., Sánchez, J., & Caira, N. (2013). Herramientas Informativas para la Vigilancia Tecnológica en Diseños Curriculares de Universidades Públicas. *GECONTEC: Revista Internacional de Gestión Del Conocimiento y La Tecnología*, *1*(2), 2–13.

GORDON, M., SMITH, S., WASHER, S., WASHER, P., & KARELIS, H. (2018). *Elizabeth M. DohertyChristopher H. FotschNianhe HanRandall W. HungateQingyian LiuMark H. NormanNing XiShimin Xu*.

Groenveld, P. (2007). Roadmapping integrates business and technology. *Research Technology Management*, *50*(6), 49–58. https://doi.org/10.1080/08956308.2007.11657472

Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W. (2014). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. *Proceedings - 2014 6th International Conference on*

Information Technology and Electrical Engineering: Leveraging Research and Technology Through University-Industry Collaboration, ICITEE 2014, 0–3. https://doi.org/10.1109/ICITEED.2014.7007894

Hall, B. H. (2007). Patents and patent policy. *Oxford Review of Economic Policy*, *23*(4), 568–587. https://doi.org/10.1093/icb/grm037

Hearst, M. A. (1999). Untangling Text Data Mining. *School of Information Management & Systems. University of California, Berkeley*.

Huang, J. Y., & Chen, R. C. (2019). Exploring the intellectual structure of cloud patents using non-exhaustive overlaps. In *Scientometrics* (Vol. 121). https://doi.org/10.1007/s11192-019-03219-4

HUAXING, G. (2017). *Cocoa butter chocolates*.

Introducing Machine Learning. (2016). *Perspectives on Ontology Learning*, (January 2014).

Jennings, B. B. (2015). *Compounds, Methods, and Treatments for Abnormal Signaling Pathways for Prenatal and Postnatal Development*.

JIAHENG, Z., CHUNCHEN, L., JINGBO, Z., JUMAO, Y., & YIHONG, L. (2018). *Application of fish oligopeptide in cosmetics*.

Jing, L. (2008). Survey of Text Clustering. *The University of Hong Kong, HongKong, China*.

Jose Graziano de Silva. (2013). *El estado mundial de la agricultura y la alimentacion. FAO Organizacion de las naciones unidas para la alimentacion y la agricultura*. Retrieved from http://www.fao.org/fao-stories/article/es/c/1170985/

Kati A. Chevaux, Amy, D., Schmitz, K., & H., J. H. (2010). *Compositions and methods to improve vascular health*.

Kim, G., & Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change*, *117*, 228–237. https://doi.org/10.1016/j.techfore.2016.11.023

Kim, J. M., Kim, N. K., Jung, Y., & Jun, S. (2019). Patent data analysis using functional count data model. *Soft Computing*, *23*(18), 8815–8826. https://doi.org/10.1007/s00500-018-3481-6

Kodratoff, Y. (1999). Knowledge discovery in texts: A definition, and applications. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1609*, 16–29. https://doi.org/10.1007/BFb0095087

Kyebambe, M. N., Cheng, G., Huang, Y., He, C., & Zhang, Z. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis. *Technological Forecasting and Social Change*, *125*(August), 236–244. https://doi.org/10.1016/j.techfore.2017.08.002

Lee, S.-Y., Lee, J.-S., & Kwon, I. (2000). *Cacao extract including dietary fiber*.

León, A. M., Castellanos, O. F., & Vargas, F. A. (2006). Valoración, Selección y pertinencia de herramientas de software utilizadas en vigilancia tecnológica. *Ingeniería e Investigación*, *26*(1), 92–102. Retrieved from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-56092006000100012&lng=en&nrm=iso

Li, X., Xie, Q., Jiang, J., Zhou, Y., & Huang, L. (2019). Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. *Technological Forecasting and Social Change*, *146*(June 2018), 687–705. https://doi.org/10.1016/j.techfore.2018.06.004

Lima, A. I. De, Argenta, A. B., Zattar, I. C., & Kleina, M. (2019). Applying Text Mining to Identify Photovoltaic Technologies. *IEEE Latin America Transactions*, *17*(5), 727–733. https://doi.org/10.1109/TLA.2019.8891940

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–296.

Madani, F., & Weber, C. (2016). The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis. *World Patent Information*, *46*, 32–48. https://doi.org/10.1016/j.wpi.2016.05.008

Magerman, T., van Looy, B., & Song, X. (2010). 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.

https://doi.org/10.1007/s11192-009-0046-6

Malki, A. Al, Rizk, M. M., El-Shorbagy, M. A., & Mousa, A. A. (2016). Hybrid Genetic Algorithm with K-Means for Clustering Problems. *Open Journal of Optimization*, *05*(02), 71–83. https://doi.org/10.4236/ojop.2016.52009

McCullagh, P., & Polson, N. G. (2018). Statistical sparsity. *Biometrika*, *105*(4), 797–814. https://doi.org/10.1093/biomet/asy051

Medina-Veloz, G., Luna-Rosas, F. J., Tavarez-Avendaño, J. F., & Narvaez-murillo, R. U. (2016). Calibración y selección del modelo de aprendizaje no supervisado K-Medias , de una encuesta sobre factores de riesgo en el consumo de drogas entre estudiantes. *Revista de Análisis Cuantitativo y Estadístico*, *3*(7), 1–9. Retrieved from www.ecorfan.org/bolivia

Medina, J. D. L. C., & Vargas, O. (2009). CACAO: Operaciones Poscosecha. Retrieved from http://www.fao.org/3/a-au995s.pdf

Method for preparing chocolate essence from cocoa bean husk. (2014).

Method for preparing elastic chocolate sauce. (2009).

METHOD TO DETERMINE RESPONSIVENESS OF CANCER TO EPIDERMAL GROWTH FACTOR RECEPTOR TARGETING TREATMENTS. (2006).

Ministerio de Agricultura y Desarrollo Rural. (2010). *Ministerio de Agricultura y Desarrollo Rural Países productores exclusivos de cacao fino de aroma Países productores mixtos de cacao fino de aroma*. 20.

Mitchell, T. M. (1997). *Machine Learning*. (April), 414. Retrieved from https://books.google.ca/books?id=EoYBngEACAAJ&dq=mitchell+machine+learning+1997&hl=en&sa=X&ved=0ahUKEwiomdqfj8TkAhWGslkKHRCbAtoQ6AEIKjAA

Moehrle, M. G., & Caferoglu, H. (2019). Technological speciation as a source for emerging technologies. Using semantic patent analysis for the case of camera technology. *Technological Forecasting and Social Change*, *146*(July 2018), 776–784. https://doi.org/10.1016/j.techfore.2018.07.049

Montes-y-gómez, M. (2001). Minería de Textos: Un nuevo reto computacional. *3rd International Workshop on Data Mining MINDAT2001*. Retrieved from

http://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf

Morales, J. de J., García, A., & Méndez, E. (2012). ¿Qué sabe usted acerca Cacao? *Revista Mexicana de Ciencias Farmaceuticas*, Vol. 43, pp. 79–81. Retrieved from http://www.redalyc.org/articulo.oa?id=57928311010

Morgan. (2019). IJACSA. In *Journal of Chemical Information and Modeling* (Vol. 53, pp. 1689–1699). https://doi.org/10.1017/CBO9781107415324.004

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 15–24. https://doi.org/10.20982/tqmp.09.1.p015

Moya, R. (2016). Machine learning. *Machine Learning*, *45*(13), 3–5. Retrieved from https://khasathan.in.th/archives/1217/machine-learning-4-การนำ-machine-learning-มาใช้งานต้องทำอะไรบ้าง

Na, K., Jin, S.-W., Lee, D.-H., Hahm, K.-B., Shinn, H.-C., & Chung, G.-J. (2009). *Solid lipid nanoparticles for drug delivery, a production method therefor, and an injectable preparation comprising the nanoparticles*.

Nacional Agropecuaria, E. (2020). *Boletín Técnico Encuesta Nacional Agropecuaria (ENA)*. Retrieved from https://www.dane.gov.co/files/investigaciones/agropecuario/enda/ena/2019/boletin_ena_2019-I.pdf

Niemann, H., Moehrle, M. G., & Frischkorn, J. (2017). Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application. *Technological Forecasting and Social Change*, *115*, 210–220. https://doi.org/10.1016/j.techfore.2016.10.004

Noh, H., Jo, Y., & Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications*, *42*(9), 4348–4360. https://doi.org/10.1016/j.eswa.2015.01.050

Nyberg, J., Saadat, S. O., & Zoraida, G. (2006). Agricultura, expansión del comercio y equidad de género. *Organizacion de Las Naciones Unidas Para La Agricultura y La Alimentacion*, 1–

59. Retrieved from http://www.fao.org/3/a0493s/a0493s02.htm#bm2

Oficina Española de Patentes y Marcas - OEPM. (2006). Claims. Retrieved October 1, 2020, from https://es.espacenet.com/help?locale=es_ES&method=handleHelpTopic&topic=claims

OMPI. (2013). Las solicitudes de patente presentadas a escala mundial han experimentado el crecimiento más rápido de los últimos 18 años. Retrieved May 9, 2020, from https://www.wipo.int/pressroom/es/articles/2013/article_0028.html

Palop, F., & Vicente, J. M. (1999). Vigilancia Tecnológica E Inteligencia Competitiva. Su Potencial Para La Empresa Española. *Revista Electrónica Gestión de Las Personas y Tecnología*, *5*, 116. Retrieved from http://scholar.googleusercontent.com/scholar?q=cache:Ei24Mz8j1yMJ:scholar.google.com/+Vigilancia+Tecnológica+e+Inteligencia+Competitiva:+Una+Contribución+al+Desarolo+de+Regiones+o+Teritorios+Inteligentes&hl=es&as_sdt=0,5%5Cnhttp://www.delfos.co.cu

Park, H., Ree, J. J., & Kim, K. (2013). Identification of promising patents for technology transfers using TRIZ evolution trends. *Expert Systems with Applications*, *40*(2), 736–743. https://doi.org/10.1016/j.eswa.2012.08.008

Parraguez, P., Škec, S., e Carmo, D. O., & Maier, A. (2020). Quantifying technological change as a combinatorial process. *Technological Forecasting and Social Change*, *151*(November 2019), 119803. https://doi.org/10.1016/j.techfore.2019.119803

Perfetti, J. J., Balcázar, Á., Hernández, A., & Leibovich, J. (2013). Políticas para el desarrollo de la agricultura en Colombia. In *Sociedad de Agricultores de Colombia, Fedesarrollo.* Retrieved from www.bancoagrario.gov.co

Perichinsky, G., Servente, M., Servetto, A., García-Martínez, R., Orellana, R., & Plastino, A. (2003). *Taxonomic Evidence and Robustness of the Classification Appying Intelligent Data Mining*. 1797–1808. Retrieved from https://www.researchgate.net/publication/228640451_TAXONOMIC_EVIDENCE_AND_ROBUSTNESS_OF_THE_CLASSIFICATION_APPLYING_INTELLIGENT_DATA_MINING

Perry, P. O. (2017). *corpus: Text Corpus Analysis*. Retrieved from http://corpustext.com

Polyphenol stabilizer, and composition and processed goods containing said stabilizer. (2013).

Preparation method of chocolate candy. (2017).

Preparation method of cosmetic component with moisturizing efficacy. (2016).

Procedure for obtaining cocoa, chocolates and other cocoa-based food products with a higher functional and sensory properties through fermentation of cocoa seeds in natural fruit juices with added microbial agents. (2016).

Puentes, D. B. (2016). Cacao (theobroma cacao L.) en el departamento del Huila en Colombia. Limitantes y oportunidades para el sector cacaotero. *Revistas.Sena.Edu.Co*, 8–9. Retrieved from http://revistas.sena.edu.co/index.php/riag/article/viewFile/1434/1563

RIDDLE, D. (2016). *A skin-care composition and kit for and method of producing it*.

Rodríguez, P. (2020). Patentes de Cacao | Kaggle. Retrieved October 8, 2020, from https://www.kaggle.com/pmrodriguez3/patentes-de-cacao

Rodríguez, S. Y., & Díaz, A. A. (2009). *Herramientas de Minería de Datos*.

Roh, T., Jeong, Y., & Yoon, B. (2017). Developing a methodology of structuring and layering technological information in patent documents through natural language processing. *Sustainability (Switzerland)*, *9*(11). https://doi.org/10.3390/su9112117

Romanczyk, L. J., & Schmitz, J. H. H. (2002). *Procyanidin and cyclo-oxygenase modulator compositions*.

Sarica, S., Luo, J., & Wood, K. L. (2020). TechNet: Technology semantic network based on patent data. *Expert Systems with Applications*, *142*. https://doi.org/10.1016/j.eswa.2019.112995

SEMANA S.A. (2017). Industria del cacao colombiano produce récord. Retrieved April 18, 2020, from Dinero website: https://www.dinero.com/edicion-impresa/pais/articulo/industria-del-cacao-colombiano-produce-record/251611

Seo, W., Yoon, J., Park, H., Coh, B. youl, Lee, J. M., & Kwon, O. J. (2016). Product opportunity identification based on internal capabilities using text mining and association rule mining. *Technological Forecasting and Social Change*, *105*, 94–104. https://doi.org/10.1016/j.techfore.2016.01.011

Sexton, F. A., Krishnan, S., & Vendra, V. K. (2007). *Sustained release of nutrients in vivo*.

Shapiro, S. C. (1992). *Encyclopedia fo Artificial Inteligence* (pp. 641–663). pp. 641–663. Retrieved from http://www.cse.yorku.ca/~tsotsos/Homepage of John K_files/teai-92.PDF

Shi, X., Cai, L., & Song, H. (2019). Discovering potential technology opportunities for fuel cell vehicle firms: A multi-level patent portfolio-based approach. *Sustainability (Switzerland)*, *11*(22). https://doi.org/10.3390/su11226381

Sies, H. (2003). *Use of cocoa flavanols and oligomers thereof to treat cognitive dysfunction and improve cognitive function in neuro-compromised patients*.

Simard, J. M. (2008). *Inhibitors of NCCa-ATP channels for therapy*.

Song, K., Kim, K. S., & Lee, S. (2017). Discovering new technology opportunities based on patents: Text-mining and F-term analysis. *Technovation*, *60–61*(March), 1–14. https://doi.org/10.1016/j.technovation.2017.03.001

SPALLITTA, F. A. (2017). *Acetylcholinesterase inhibitors for treatment of dermatological conditions*.

Takeyas, B. L. (2007). Introducción a la inteligencia artificial. *23*, *12*(32), 1.

Tan, A. (1999). Text Mining : The state of the art and the challenges. *Kent Ridge Digital Labs*.

TROPLIN, P., & BERNAERT, H. (2006). *USE OF CACAO POLYPHENOLS FOR TREATING A PROSTATE HYPERPLASIA, A SPECIFIC CACAO EXTRACT AND APPLICATIONS*.

Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information Processing and Management*, *43*(5), 1216–1247. https://doi.org/10.1016/j.ipm.2006.11.011

Vargas, K. (2016). La agricultura colombiana en el contexto de la globalización. *Elcampesino.Com*. Retrieved from https://www.elcampesino.co/la-agricultura-colombiana-en-el-contexto-de-la-globalizacion/

W., G. T., F., J. B., & C., P. I. (2004). *Process for producing cocoa butter and cocoa powder by liquefied gas extraction*.

Wang, J., & Chen, Y. J. (2019). A novelty detection patent mining approach for analyzing technological opportunities. *Advanced Engineering Informatics*, *42*(August 2018), 100941. https://doi.org/10.1016/j.aei.2019.100941

Wang, W. M., & Cheung, C. F. (2011). A Semantic-based Intellectual Property Management System (SIPMS) for supporting patent analysis. *Engineering Applications of Artificial Intelligence*, *24*(8), 1510–1520. https://doi.org/10.1016/j.engappai.2011.05.009

Wang, Xiaoyu, Zhai, Y., Lin, Y., & Wang, F. (2019). Mining layered technological information in scientific papers: A semi-supervised method. *Journal of Information Science*, *45*(6), 779–793. https://doi.org/10.1177/0165551518816941

Wang, Xuefeng, Ren, H., Chen, Y., Liu, Y., Qiao, Y., & Huang, Y. (2019). Measuring patent similarity with SAO semantic analysis. *Scientometrics*, *121*(1), 1–23. https://doi.org/10.1007/s11192-019-03191-z

WIPO. (2012). *Guía Para Bases De Datos Tecnológicas*. Retrieved from http://www.wipo.int/edocs/pubdocs/es/patents/434/wipo_pub_l434_11.pdf

Wittfoth, S. (2019). Measuring technological patent scope by semantic analysis of patent claims – An indicator for valuating patents. *World Patent Information*, *58*(July), 101906. https://doi.org/10.1016/j.wpi.2019.101906

Woo, H. G., Yeom, J., & Lee, C. (2019). Screening early stage ideas in technology development processes: a text mining and k-nearest neighbours approach using patent information. *Technology Analysis and Strategic Management*, *31*(5), 532–545. https://doi.org/10.1080/09537325.2018.1523386

Wu, C. H., Ken, Y., & Huang, T. (2010). Patent classification system using a new hybrid genetic algorithm support vector machine. *Applied Soft Computing Journal*, *10*(4), 1164–1177. https://doi.org/10.1016/j.asoc.2009.11.033

Wu, J. L., Chang, P. C., Tsao, C. C., & Fan, C. Y. (2016). A patent quality analysis and classification system using self-organizing maps with support vector machine. *Applied Soft Computing Journal*, *41*, 305–316. https://doi.org/10.1016/j.asoc.2016.01.020

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381. https://doi.org/10.1080/713827180

Zhang, X. (2014). Interactive patent classification based on multi-classifier fusion and active learning. *Neurocomputing*, *127*, 200–205. https://doi.org/10.1016/j.neucom.2013.08.013

Zhang, Yi, Huang, Y., Porter, A. L., Zhang, G., & Lu, J. (2019). Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study. *Technological Forecasting and Social Change*, *146*(April 2018), 795–807. https://doi.org/10.1016/j.techfore.2018.06.007

Zhang, Yi, Shang, L., Huang, L., Porter, A. L., Zhang, G., Lu, J., & Zhu, D. (2016). A hybrid similarity measure method for patent portfolio analysis. *Journal of Informetrics*, *10*(4), 1108–1130. https://doi.org/10.1016/j.joi.2016.09.006

Zhang, Yin, Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1–4), 43–52. https://doi.org/10.1007/s13042-010-0001-0

Zhou, X., Huang, L., Zhang, Y., & Yu, M. (2019). A hybrid approach to detecting technological recombination based on text mining and patent network analysis. In *Scientometrics* (Vol. 121). https://doi.org/10.1007/s11192-019-03218-5

Zhu, F., Wang, X., Zhu, D., & Liu, Y. (2015). A Supervised Requirement-oriented Patent Classification Scheme Based on the Combination of Metadata and Citation Information. *International Journal of Computational Intelligence Systems*, *8*(3), 502–516. https://doi.org/10.1080/18756891.2015.1023588