TRABAJO DE GRADO

ANÁLISIS TEXTUAL DE LOS TÍTULOS, RESÚMENES Y PALABRAS CLAVES DE LA REVISTA DE CIENCIAS E INGENIERÍA EN EL PERIODO 2000-2016 PUBLICADAS EN SABER ULA

Por

Br. Hilianna La Cruz

Tutor: Prof. Rafael Borges

DEDICATORIA

Llegar a esta etapa de mi vida ha sido unos de los momentos más felices, es el resultado de un largo recorrido lleno de momentos muy especiales y de mucha sabiduría, donde pude conocer personas maravillosas que me aportaron su granito de arena y me ayudaron a dar muchos de los pasos que me trajeron hasta el final de este capítulo de mi vida, que no es más que el comienzo de uno nuevo, una nueva etapa con nuevas pruebas, oportunidades y conocimiento. Es por ello que quiero agradecer:

A Dios Todopoderoso y a la Virgen Santísima, por acompañarme en todo momento e iluminarme en este largo recorrido lleno de sabiduría, entendimiento, paciencia, amor y sobre todo por darme la fuerza para continuar luchando a pesar de todas las dificultades.

A mis padres Elsy María Montilla (+) e Hilario José La Cruz, gracias a ustedes hoy mi sueño es una realidad, gracias por darme la educación para convertirme en lo que hoy soy y por creer en mí. Siempre estuvieron a mi lado dándome su apoyo en los momentos difíciles, esto se lo debo a ustedes. Te amo enorme papá y a ti mamita, te amaré y extrañaré por siempre, gracias por guiar mis pasos desde el cielo.

A mis hermanos Marianna y Domingo, por acompañarme en todo momento y por existir. Fueron mi inspiración y mis ganas de seguir adelante para lograr mi meta y convertirme en un ingeniero como ustedes. Los Amo.

A mis tíos, que son como unos segundos padres, Zulma, Mare, Blanca, Cristina, Diego, Eugenio, Carlos, José R (Cuchi), Ricardito, Diana, Luzmariel, Omaira, Blanca La Cruz, Humberto, José G, Milagros (+). Este pequeño triunfo es de ustedes. Los Amo.

A mi abuela Juana, por acompañarme en todo momento y por apoyarme, te amo viejita este triunfo te lo dedico a ti.

A mis primos, porque han estado en el momento indicado para apoyarme, corregirme y darme palabras de aliento cuando más lo he necesitado, los quiero mucho.

A mi segunda familia adoptiva, Andrith, Sophia, Adriana, Sra Rita, Sr Adolfo, Gustavito, Edwin, Ediwson, porque siempre han estado allí presentes en todo momento, que Dios les pague por su bondad y cariño hacia mí. Los Amo.

A mis segundas madres, Nancy, Ninoska y Carolina, gracias por brindarme su apoyo en todo momento, por ser las que a veces me ponen los pies sobre la tierra, por sus palabras de aliento y por estar cuando más las he necesitado. Las Amo.

A mis sobrinos y ahijados, María Victoria, Mariangel Andrea, José Andrés, Ángel Domingo, Victoria Estefanía, Alejandra, Sebastian, José Ignacio, Jesús Alejandro, Juan Diego por ser quienes sacan de mí la mejor parte, regalándome sus sonrisas y travesuras. Los amo inmenso mis pequeños.

A mis Hijos adoptivos, José Alejandro y Nicolás que con su cariño y amor me han sacado más de una sonrisa en los momentos que más lo he necesitado, gracias mis niños. Los Amo.

A mis queridos amigos, Daniel, Javier, Rhona, Yoana, Leonel, Joanelly, Daniela, Emilly, Mariangely, Arantxa, Miguel, Dianita, Estefani, Sra Lucy, Gabriel. Ustedes le han dado a mi vida un toque especial y me han llenado de alegría en muchos momentos de tristeza, mil gracias porque han hecho que el recorrido sea menos pesado. Los Amo.

AGRADECIMIENTO

A la ilustre Universidad de Los Andes, mi máxima casa de estudios, a la cual debo mi superación profesional y a cada uno de los profesores que estuvieron involucrados en dicha formación.

Vaya mi especial agradecimiento a mi tutor, profesor Rafael Borges, quien supo guiar con sabiduría cada una de las asesorías, las cuales me impulsaron para consumar este gran logro. Bendiciones para usted profe.

ÍNDICE GENERAL

Índice de contenido

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ANEXOS

Anexo

ANÁLISIS TEXTUAL DE LOS TÍTULOS, RESÚMENES Y PALABRAS CLAVES DE LA REVISTA DE CIENCIAS E INGENIERÍA EN EL PERIODO 2000-2016 PUBLICADAS EN SABER ULA

Hilianna La Cruz

Universidad de Los Andes. Facultad de Ingeniería. Escuela de Ingeniería de Sistemas**.**

RESUMEN

El presente trabajo especial de grado, se realizó un análisis de texto a la revista de Ciencia e Ingeniería publicada en el repositorio institucional SABER ULA, perteneciente a la Universidad de Los Andes para el periodo comprendido entre los años 2000 y 2016, el cual tuvo como objetivo general Analizar textualmente los títulos, resúmenes y palabras claves de la revista Ciencias e Ingeniería en el periodo 2000 – 2016, publicadas en Saber ULA. Mediante la utilización de la herramienta R a fin de realizar el análisis textual (minería de texto) a los títulos, resúmenes y palabras claves, con el uso de diversas librerías para efectuar dicha minería, lo que facilitó la compresión de nueva información y, de alguna manera, permitió determinar la temática de una forma más sencilla. En relación al análisis textual, no se siguió ninguna metodología en específico. Se crearon nubes de palabras para cada año en estudio, contentivas de las palabras claves, resúmenes, títulos y la nube de comparación; para la realización de este análisis de texto y de las nubes fue necesario crear una base de datos en archivos txt, donde se almacenaron para cada año los títulos, resúmenes, palabras claves y autores. También fue necesario la depuración de los artículos publicados con respecto a los caracteres especiales, signos de puntuación, números, acentuaciones, entre otros. En cambio, para realizar la recurrencia de los autores de dicha revista, esto si se realizó mediante la aplicación de la metodología de análisis planteada por Meeker y Escobar (1998) y Nelson (2003). Se estimaron modelos no paramétricos para el promedio acumulado del número de publicaciones recurrentes en la revista en función de la edad (meses de observación) para las publicaciones de los autores aquellos que tuvieron 2 o más artículos publicados. Con los modelos paramétricos estimados se obtuvieron los pronósticos del número promedio de artículos recurrentes de los autores.

Palabras Claves: Análisis textual, Temática, Recurrencia de autores, Evolución temporal, Nube de palabras

INTRODUCCIÓN

En las últimas cuatro décadas del siglo XX, con el desarrollo internacional de las tecnologías de la información y la comunicación (TIC), se ha producido un cambio importante en el formato de las publicaciones académicas que ha pasado de su versión en papel, al formato digital y al uso libre cada vez más generalizado en internet.

Motivado a ello, se ha dado paso a la creación de lo que es llamado un repositorio institucional (RI) entendido como una nueva tendencia de preservación intelectual asumida por muchas universidades a nivel mundial al facilitar la gestión, difusión y el fácil acceso desde internet a documentos electrónicos originados en estas instituciones y que reflejan su producción intelectual y gestión institucional.

En el caso de la Universidad de Los Andes el proyecto institucional se lleva a cabo a través de SABER ULA, cuyo objetivo principal es difundir y preservar la producción intelectual de los miembros de la comunidad universitaria (profesores e investigadores de todas las dependencias y de las unidades de investigación que posee esta institución académica.

Partiendo de estas consideraciones surge la necesidad de analizar las revistas que se encuentran en dicho repositorio, específicamente en el área de Ciencia e Ingeniería en el periodo comprendido entre los años 2000 y 2016, con la finalidad de observar aquellos temas que se repiten en torno a una investigación, es decir, evaluar la temática mayormente empleada en un determinado periodo (evolución en el tiempo) así como también la recurrenci1a de autores.

Para ello se ha efectuado una revisión exhaustiva de los documentos de interés relacionados con la investigación y, posteriormente, con los datos extraídos se

crearon nubes de palabras que facilitaron la búsqueda de la información correspondiente.

En este contexto la investigación se estructura en cinco capítulos. En el primero, se exponen los aspectos preliminares como: antecedentes, planteamiento del problema, objetivos, justificación y alcance de la investigación. El Capítulo II contiene el marco teórico de la investigación, en el cual se amplía la visión del tema dando a conocer conceptos claves para el desarrollo del proyecto. En el tercer capítulo se desarrolla la metodología del estudio, la cual se adhiere a lo planteado por Meeker y Escobar (1998), así como también se describe la base de datos utilizada. En el Capítulo IV, se presentan los resultados obtenidos como producto del análisis textual y de recurrencias. En el Capítulo V, se encuentran las conclusiones y recomendaciones del estudio. Finalmente, se presentan las referencias bibliográficas y los anexos respectivos de la investigación.

CAPÍTULO I

ASPECTOS PRELIMINARES

1.1 Antecedentes

Los antecedentes de otros estudios constituyen precedentes que sirven de consulta permanente para el desarrollo de nuevos trabajos que tengan algún tipo de vinculación con el que se pretenda realizar. De acuerdo a lo señalado por Ramírez (2010), "Consiste en dar al lector toda la información posible acerca de las investigaciones que se han realizado, tanto a nivel nacional como internacional sobre el problema que se pretende investigar" (p. 40). Según esta cita, los antecedentes permiten guiar nuevos estudios para tener una visión más clara de cómo orientar la investigación a emprender. En el presente estudio, para el análisis textual de títulos, resúmenes y palabras claves de los artículos, se han seleccionado los siguientes antecedentes.

Cui et al. (2010), realizaron su trabajo en el "Contexto que conserva la visualización dinámica de nube de palabra" (título original en inglés *Context Preserving Dynamic Word Cloud Visualization*) en la cual presentan dentro de su artículo que, introduciendo un método de visualización que combina el gráfico de tendencias con las nubes de palabras para ilustrar las evoluciones de los contenidos temporales en un conjunto de documentos. En concreto, utilizaron un gráfico de tendencias para codificar la evolución semántica general del contenido de los documentos a lo largo del tiempo.

En este trabajo, la evolución semántica de una colección de documentos se modela por la significación variada del contenido del documento, representado por un conjunto de palabras clave representativas, en diferentes momentos. En cada punto de tiempo, también usaron una nube de palabras para representar las palabras clave

representativas. Dado que las palabras en una nube de palabras pueden variar unas de otras a lo largo del tiempo (por ejemplo, palabras con mayor importancia), utilizamos mallas de geometría y un modelo adaptativo de fuerza dirigida para establecer nubes de palabras para resaltar las diferencias de palabra entre dos nubes de palabras subsiguientes.

Dicho método también asegura la coherencia semántica y la estabilidad espacial de las nubes de palabras a través del tiempo. Asimismo, el trabajo se encarna en un sistema de análisis visual interactivo que ayuda a los usuarios a realizar análisis de texto y obtener ideas de una gran colección de documentos. La evaluación preliminar demuestra la utilidad y utilidad de nuestro trabajo.

Cabe señalar, que dicho antecedente fue tomado como referencia pues justifica las teorías relacionadas con la variable nubes de palabras en la evolución de las mismas; a través del tiempo, también, sirve como punto de partida si se toma el método para utilizarse en un contexto espacial.

Wu *et. al*., (2011), trabajaron en "Preservar - Semántica de las nubes de la palabra mediante el tallado de la costura" (título original en inglés *Semantic - PreservingWord Clouds by Seam Carving*). Donde exponen que las nubes de palabras están proliferando en Internet y han recibido mucha atención en el análisis visual. Aunque las nubes de palabras pueden ayudar a los usuarios a comprender rápidamente el contenido principal de una colección de documentos, su capacidad de comparar visualmente documentos es limitada.

Este artículo presenta un nuevo método para crear nubes de palabras que preservan la semántica mediante el aprovechamiento de la talla de costura a medida, un operador bien establecido de cambio de tamaño de imagen con contenido. El método puede optimizar un diseño de nube de palabras quitando una costura de izquierda a derecha o superior a inferior de forma iterativa y graciosa del diseño. Cada costura es una trayectoria conectada de regiones de baja energía determinadas por una función de energía basada en Gauss.

Asimismo, con la talla de la costura, podemos empaquetar la nube de la palabra de forma compacta y eficaz, preservando su estructura semántica general. Además, diseñamos un conjunto de técnicas de visualización interactiva para las nubes de palabras creadas para facilitar el análisis visual de texto y la comparación. Se realizan estudios de casos para demostrar la efectividad y utilidad de nuestras técnicas.

Lo anterior expuesto, sirve como base a la actual investigación al coincidir en el estudio de temáticas relacionadas con la variable nubes de palabras, técnicas de visualización interactivas, las cuales servirán en la presente investigación como propuestas para facilitar el análisis y producción de nuevas conclusiones a través de la visualización de artículos, mediante la creación de las nubes de palabras. Por su parte, para el análisis de eventos recurrentes se revisaros una serie de trabajos, entre los cuales destacaron los siguientes:

Gómez (2007), efectuó un "Análisis de eventos recurrentes para las fallas de las celdas de reducción electrolítica de la CVG-Venalum". Donde expone que realizo un análisis de datos de eveno recurrentes, a los datos de fallas y costo de fallas de las celdas de reducción electrolíticas de la corporación venezolana de Guayana de Venalum (C.V.G-Venalum) para el periodo comprendido entre 1979 y 2004, mediante la aplicación de la metodología de análisis planteada por Meeker y Escobar (1998) y Nelson (2003). Se estimaron modelos no paramétricos para el promedio acumulado del número y costo de fallas recurrentes de la celdas de reducción electrolíticas en función de la edad (en días de operación) de dichas celdas, para las líneas de producción con tecnología Reynolds e Hydro Aluminium y considenrando los distintos modos de fallas y modos de fallas general.

Con los modelos no paramétricos estimados, se realizaron comparaciones entre los modos de fallas por línea de producción y de fallas generales entre las líneas de producción. Además se estimaron modelos paramétricos básicos de procesos de conteo para la tasa de recurrencia del número de fallas y la tasa de costos de fallas recurrentes de las celdas de reducción electrolítica, para las líneas de producción de los tipos de tecnología nombrados anteriormente y considerando también los distintos modos de fallas y modos de fallas generales. Con los modelos paramétricos

estimados se realizaron pronósticos del número y costo promedio de fallas recurrentes de las celdas de reducción para un año.

En el antecedente descripto anteriormente, sirve como base metodológica del estudio realizado, además de sustentar las bases teóricas pues analiza dentro de la variable las fallas recurrentes en las celdas electrolíticas, coincidiendo con las recurrencias que se presentan en las publicaciones de diferentes artículos y autores relacionados con la variable en estudio.

Por su parte, Lopera y Manotas (2011), realizaron una investigación titulada "Aplicación del análisis de datos recurrentes sobre interruptores FL245 en Interconexión Eléctrica S.A." Donde exponen que los datos recurrentes surgen cuando una unidad (o un grupo de tales unidades) es monitoreada a través del tiempo y un evento particular (o grupo de eventos) ocurre en varios puntos del periodo de observación, por ejemplo, los tiempos de episodios recurrentes de una enfermedad en pacientes o los tiempos de reparación de un producto manufacturado. Muchos sistemas, subsistemas y componentes (que genéricamente son denominadas "unidades") tienen asociadas más de una causa o modo de falla. En algunas aplicaciones, y para ciertos propósitos, es importante distinguir entre las causas o modos de falla. Para mejorar la confiabilidad, es esencial identificar la causa de falla hasta el nivel de componente, y en muchas aplicaciones, hasta la causa física real de una falla. En este trabajo, se presenta una aplicación del análisis de datos recurrentes realizado sobre interruptores tipo FL245 (unidades reparables en Interconexión Eléctrica S.A., ISA), que incluye el uso de métodos estadísticos no paramétricos y paramétricos considerando varios modos de falla.

Lo anterior expuesto, en el antecedente fue tomado como referencia dado que en sus basamentos teóricos y conclusiones refleja similitud con la investigación desarrollada en cuanto a las dimensiones relacionadas métodos estadísticos empleados para las recurrencias de las distintas publicaciones de autores. Igualmente, puede apreciarse de los antecedentes antes señalados, todos ellos se vinculan con el presente estudio, bien sea en el contexto de nubes de palabras o en lo relativo al

análisis de eventos recurrentes; así como también vale señalar la metodología empleada.

1.2 Planteamiento del Problema

La comunidad académica consiste en un grupo jerárquico que fomenta la investigación a través de las publicaciones dentro de las diferentes especialidades, entendiéndose por publicación científica cualquier forma de divulgar entre varias personas los resultados de alguna experiencia científica o los hallazgos de algún fenómeno o entes nuevos para la ciencia. Su divulgación puede tomar diversas formas, tales como orales, escritas o electrónicas; estas últimas generalmente se realizan en forma periódicas a través de elementos electrónicos. Sus características principales son la rápida difusión, el ahorre de coste y la fiabilidad para su uso, donde el principal papel es el acceso a la información de manera rápida y económica.

Ahora bien, cuando se publican artículos de interés o de carácter científico, se realizan principalmente con la finalidad de contribuir en el desarrollo de temas que son importantes tanto para la comunidad científica como para la sociedad. En el caso de las revistas, estas forman parte de la infraestructura de comunicación y del propio acervo intelectual de las comunidades académicas que las producen y consumen. Operan en medios específicos a través de una trama de autores, lectores y estructuras de distribución, y en esa medida constituyen una expresión de las características e intereses, los alcances y limitaciones de las redes académicas y temas que las nutren. En la proporción en que lo hacen, las revistas se convierten en referentes fundamentales para los investigadores profesionales y, más en general, para los académicos especializados (Contreras, 2011).

Estas revistas presentan como condición básica el hecho de ser instrumentos de las comunidades científicas, regidos por normas académicas, y valorados por la relevancia de los materiales publicados dentro de campos especializados del saber,

teniendo una importante relevancia disciplinaria en la cual se define el valor de la producción académica. Estos saberes están dirigidos a un público de expertos constituido, principalmente, por miembros de agrupaciones científicas, académicos y profesionales en proceso de formación. Al respecto, tanto por el lenguaje utilizado como los supuestos epistémicos, puede inferirse que no se trata de un público abierto quienes hacen uso de estas revistas académicas.

Ante este referente, vale señalar el caso de Saber ULA, Repositorio Institucional en el cual se gestiona la publicación, preservación y acceso libre, a texto completo, de documentos derivados de la producción intelectual e institucional de la Universidad de Los Andes, de los miembros de esta comunidad universitaria (profesores e investigadores de todas las dependencias y unidades de investigación que posee la ULA; además de servir de fuente de almacenamiento e intercambio común de la información producida en la Universidad en todos los campos del conocimiento, registrando en forma sistemática la información derivada de la gestión institucional universitaria.

De manera específica, es oportuno mencionar el caso de los artículos publicados en la revista Ciencia e Ingeniería y que son gestionados en el portal Saber ULA. Esta revista es una publicación cuatrimestral, patrocinada por la Facultad de Ingeniería de la Universidad de Los Andes, Mérida, Venezuela, en la cual se publican artículos científicos, relacionados con todas las áreas de la ingeniería y las ciencias aplicadas a estas. Todos los artículos son sometidos a un proceso de revisión, poniéndose a la disposición de docentes, investigadores y profesionales, de las ramas básicas y aplicadas de la Ingeniería, un medio de promoción y difusión que les brinda la oportunidad de dar a conocer el fruto de sus trabajos y les permite expresar sus opiniones respecto a cualquier actividad fundamental en sus áreas de experticia.

Cabe agregar que, las publicaciones de una diversidad de autores han generado una red amplia de académicos y un público determinado, quienes se muestran

interesados en los artículos; no obstante, esa afluencia de publicaciones ha concebido un exceso de materiales, lo que aunado a una diversidad de enfoques, metodologías y matices temáticos de los artículos publicados; constituyen un proceso que puede resultar engorroso cuando se desea conocer la producción intelectual de un tema en particular.

Asimismo, en relación con la consistencia temática que refleja la convergencia de redes académicas activas en la revista Ciencias e Ingeniería ha hecho de estas publicaciones, importantes espacios de expresión académica; por ello, el número de lectores se ha visto incrementado.

El contexto antes descrito, ha conllevado a la investigadora a centrar su estudio en la determinación de la temática que se emplea y la recurrencia de los autores, con la finalidad de dar a conocer a cualquier investigador la exploración en ambos aspectos, pues la penetración de la edición electrónica en el entorno universitario introduce importantes cambios en el proceso de investigación, resultados, edición y difusión en la revista en cuestión. Además de tomar en consideración la recurrencia con la cual los autores publican, la variación de los temas, o si las investigaciones tienen que ver con un área o estudio específico.

1.3 Justificación

La justificación del presente estudio se desarrolla desde varios puntos de vista: a nivel teórico, esta investigación ampliará la visión sobre la importancia de las revistas electrónicas para la comunidad científica y para la persona que realiza una publicación, pues puede acceder a tener la información y el conocimiento de manera rápida y económica, a través de una nube de palabras como recurso visual que se utiliza para representar las palabras más destacadas que componen un determinado texto, mostradas de manera abstracta, en las que son representadas en un mayor tamaño aquellas palabras que aparecen con mayor frecuencia o son más importantes.

Igualmente, esta investigación se justifica desde el punto de vista práctico debido a que aportará información sobre análisis textual (minería de texto) de los títulos, resúmenes y palabras claves de la revista Ciencia e Ingeniería a través de la creación de nubes de palabras por medio del software estadístico R studio y, del uso de diversas librerías para realizar dicha minería, lo que facilitará la compresión de nueva información y de alguna manera determinar la temática de una forma más sencilla.

En relación al análisis textual, no se siguió una metodología en específico, sino que se utilizaron varios pasos de algunos artículos publicados en R bloggers, como "*Intro to Text Analysis with R"* que realizan minería de texto, pero sin seguir ninguna metodología en particular.

Desde un ámbito metodológico, debe señalarse que se efectuó un análisis de eventos recurrentes siguiendo la metodología propuesta por Meeker y Escobar (1998) y Nelson (2003), para la fidelidad de publicación de los autores, definiéndose está como la recurrencia de publicación en la revista, la cual consiste en la identificación de un modelo que explica el comportamiento de la Función Acumulada Promedio (FAP) mediante un algoritmo de decisión propuesto por estos autores. Este análisis se hace mediante dos subtipos de análisis: no paramétrico y el paramétrico.

1.4 Objetivos de la investigación

1.4.1 Objetivo general

Analizar textualmente los títulos, resúmenes y palabras claves de la revista Ciencias e Ingeniería en el periodo 2000 – 2016, publicadas en Saber ULA.

1.4.2 Objetivos específicos

- ✓ Identificar los principales temas publicados en la revista Ciencias e Ingeniería de la ULA, mediante técnicas de análisis textuales.
- ✓ Analizar textualmente los datos extraídos de la mencionada revista, discriminados por año y por quinquenio.
- ✓ Determinar la recurrencia de autores discriminando por año y quinquenio.
- ✓ Elaborar una nube de palabras a partir de los datos extraídos de las revistas de interés.

1.5 Alcance

En el presenta trabajo pretende analizar textualmente los títulos, resúmenes y palabras claves de la revista Ciencia e Ingeniería en el periodo 2000 – 2016, publicadas en Saber ULA, lo que permitirá conocer año a año, por temática y a cuánto asciende la recurrencia de publicaciones.

CAPÍTULO II

MARCO TEÓRICO

En este capítulo se dan a conocer las bases conceptuales o conceptos teóricos que sustentan la investigación, los cuales sirven para introducir al lector en el tema que se desarrolla. Arias (2012), argumenta que "El marco teórico es el producto de la revisión documental-bibliográfica, y consiste en una recopilación de ideas, posturas de autores, conceptos y definiciones, que sirven de base a la investigación por realizar" (p. 106). En este sentido, a continuación, se presentan una serie de aspectos que tienen relación con el presente trabajo, tales como análisis textual de las revistas electrónicas, la temática, la recurrencia de autores, evolución temporal, así como las nubes de palabras.

2.1 Saber ULA

De acuerdo a lo expuesto por el Repositorio Institucional de la Universidad de Los Andes, "Saber ULA es un repositorio institucional de acceso libre, que gestiona la publicación y preservación a texto completo de documentos derivados de la producción intelectual e institucional de la Universidad de Los Andes". Los profesores e investigadores de la Universidad de Los Andes, grupos de investigación, editores de revistas universitarias y la comunidad universitaria, pueden publicar a través de Saber ULA documentos del quehacer académico y de investigación.

Este proyecto institucional es impulsado por el Consejo de Computación Académica de la ULA, con la participación de diferentes dependencias, tales como: CCA (Consejo Computacional Académico), CDCHT (Consejo de Desarrollo Científico, Humanístico y Tecnológico de la Universidad de Los Andes), CEP

(Consejo de Estudiantes de Posgrado). Además, es desarrollado y administrado por el Parque Tecnológico de Mérida, a través del Centro de Teleinformación (CTI).

Por otra parte, se tiene que un repositorio institucional (RI) es una nueva tendencia de preservación intelectual asumida por muchas universidades en el mundo para gestionar, difundir y facilitar el acceso a través de internet a aquellos documentos electrónicos originados en estas instituciones, y que reflejan su producción intelectual y gestión institucional. Son sistemas que almacenan y mantienen la información digital de la producción académica y científica de la universidad. Suelen incluir tesis doctorales, artículos de carácter científico, ponencias o comunicaciones a congresos, revistas electrónicas editadas por la institución, materiales docentes, entre otros.

Los RI constituyen un avance importante en la preservación de la información universitaria, propiciando el acceso libre al conocimiento y optimizando el uso de internet a partir de las posibilidades que ofrece. Adicionalmente, aumentan la visibilidad de investigadores y universidades a nivel mundial, a través de internet, lo cual constituye un aspecto importante para el intercambio académico (Repositorio Institucional de la Universidad de Los Andes, s/f).

En lo que respecta a los objetivos de Saber ULA, el Repositorio Institucional de la Universidad de Los Andes menciona los siguientes:

- ✓ Difundir y preservar la producción intelectual de los miembros de la comunidad universitaria (profesores e investigadores de todas las dependencias y unidades de investigación que posee la ULA).
- ✓ Servir de fuente de almacenamiento e intercambio común de la información producida en la Universidad de Los Andes en todos los campos del conocimiento.

- ✓ Registrar en forma sistemática la información derivada de la gestión institucional universitaria.
- ✓ Reforzar la presencia en internet de todos los centros, grupos, institutos, laboratorios y postgrados de la ULA.
- ✓ Fomentar la creación de publicaciones electrónicas.
- ✓ Contribuir con el libre acceso al conocimiento generado en la universidad.

Asimismo, Saber ULA otorga los servicios de publicación de documentos a textos completo de la comunidad universitaria: artículos o "*papers*", tesis, guías de estudio, presentaciones, estadísticas y otros datos; publicación sistematizada de datos de investigadores (perfil, contacto, currículo) y su lista de publicaciones con enlaces a los contenidos de texto completo o referencias almacenadas en la base de datos; creación y publicación en internet de revistas electrónicas pertenecientes a unidades de investigación o dependencias de la institución; publicación periódica de eventos, además del acceso a través de un portal web a todos los contenidos y datos de la base de datos de Saber ULA.

2.2 Revista Ciencia e Ingeniería

El portal de la revista Ciencia e Ingeniería de la Universidad de Los Andes, la define como "una revista multidisciplinaria arbitrada, cuya finalidad es publicar trabajos científicos de investigación básica y aplicada a las diferentes ramas de la Ingeniería, en sus secciones de: artículos invitados, revisiones, artículos de investigación, notas técnicas y comentarios". En ella se publican artículos científicos en los idiomas español e inglés, relacionados con todas las áreas de la ingeniería y las ciencias aplicadas a estas.

Es patrocinada por la Facultad de Ingeniería de esta casa de estudios, con miras a estimular los esfuerzos científicos, tecnológicos, docentes y de extensión de los miembros de la Comunidad Científica Nacional e Internacional. Al respecto, la

revista Ciencia e Ingeniería de la Universidad de Los Andes, en su página web, explica lo siguiente:

> Con esta publicación se pone a disposición de docentes, investigadores y profesionales, de las ramas básicas y aplicadas de la ingeniería, un medio de promoción y difusión que les brinde la oportunidad de dar a conocer el fruto de sus trabajos y les permita expresar sus opiniones respecto a cualquier actividad fundamental en sus áreas de experticia.

En cuanto al alcance, la revista provee investigaciones que aportan soluciones a las diferentes necesidades presentes en la sociedad, relacionados con procesos de análisis, síntesis y evaluación de métodos en las diferentes áreas de la ciencia, como energía, medio ambiente, minerales y metalúrgica, materiales y recursos naturales, biotecnología e ingeniería computarizada (revista Ciencia e Ingeniería de la Universidad de Los Andes, s/f).

Su política constante se basa en divulgar información sobre nuevos proyectos económicos y sociales que buscan ofrecer mejores alternativas en el desarrollo y adaptabilidad de procesos, métodos analíticos de laboratorios, nuevas técnicas experimentales, entre otros.

De acuerdo al portal de la revista Ciencia e Ingeniería de la Universidad de Los Andes, para publicar en la revista se deben seguir algunos procedimientos como son:

1. Acuse de recibo dirigida al autor(es) en físico y/o correo electrónico.
2. Evaluación preliminar del trabajo o artículo por parte del comité editorial para verificar si cumple con las normas y demás requisitos establecidos. Si las cumple, el mismo se envía a los árbitros, quienes emiten un veredicto sobre la publicación o no del trabajo o artículo. Si el trabajo o artículo no cumple con las normas y requisitos se le notificará por escrito la decisión al autor(es).

El arbitraje de los artículos enviados a la revista Ciencia e Ingeniería se cumple en la modalidad doble ciego, tal como lo expresa dicha revista en su página web, al establecer que “todos los artículos son sometidos a un proceso de revisión por pares

independientes bajo la modalidad de doble ciego". En esta modalidad, a los tres árbitros es enviado el artículo y una planilla de evaluación elaborada por el comité editor, en la que se contemplan las especificaciones requeridas por la revista para la publicación de los artículos, ensayos y reseñas. Los árbitros las devolverán debidamente llenas, con las observaciones correspondientes.

Entre los aspectos a evaluar se consideran la originalidad, pertinencia del tema, solidez de las argumentaciones, estructura del trabajo, organización interna del artículo, solidez de las conclusiones, y el resumen en caso de artículo.

Es importante destacar que la Universidad de Los Andes ha ocupado puestos importantes de visibilidad en la web, gracias a los contenidos difundidos a través de su repositorio institucional Saber ULA. Los ejemplares de la revista se publican cada tres meses por año y no muestra un número determinado de artículos. Esta revista provee acceso libre inmediato a su contenido bajo el principio de que hacer disponible gratuitamente investigación al público apoya a un mayor intercambio de conocimiento global.

2.3 Tipos de tecnología

DSpace es uno de los softwares de código abierto preferido por las instituciones académicas para gestionar repositorio de ficheros, al preservar los documentos digitales a largo plazo. Al respecto, Tello, Méndez, García y Zambrano (s/f) explican:

> DSpace es un sistema de información con arquitectura de repositorio digital que captura, almacena, ordena, preserva y distribuye material de investigación digital con el propósito de garantizar que se preserve y distribuya toda la producción intelectual generada al interior de las instituciones que hacen uso de éste (p. 5).

En este sentido, DSpace resulta la solución más ajustada para configurar el repositorio de objetos de aprendizaje, puesto que facilita la incorporación de diferentes tipos de recursos educativos y de tipo documental, disponiendo de una variedad de perfiles de usuarios gracias a su versatilidad.

El software DSpace está desarrollado en plataforma "*opensource*", es gratuito y se puede personalizar según las necesidades. Asimismo, Rodríguez y Sulé (2008), explican que dicho software fue "diseñado por el "*Massachusetts Institute of Technology*" (MIT) y los laboratorios de HP para gestionar repositorios de ficheros (textuales, audio, vídeo), facilitando su depósito, organizándolos en comunidades, asignándoles metadatos y permitiendo su difusión a recolectores o agregadores".

Su objetivo inicial fue crear un sistema escalable y sostenible, capaz de acoger las más de 100.000 unidades de contenido digital producidas cada año por los profesores e investigadores del MIT: artículos, informes, comunicaciones, bases de datos, programas de ordenador, grabaciones de video, presentaciones utilizadas en las clases, entre otros.

Actualmente, es la plataforma informática usada por el repositorio institucional de la Universidad de Los Andes, Saber ULA. Está constituido por un conjunto de herramientas, para gestionar contenidos digitales de acuerdo con el modelo OAIS ("*Reference Model for an Open Archival Information System"*). La revista Ciencia e Ingeniería utiliza el "*Open Journal Systems"* 2.4.2.0, un programa de publicación de código abierto para la gestión de revistas que desarrolla, financia y distribuye gratuitamente el "*Public Knowledge Project*" bajo la Licencia Pública general GNU, cuya estructura se muestra en la figura 1.

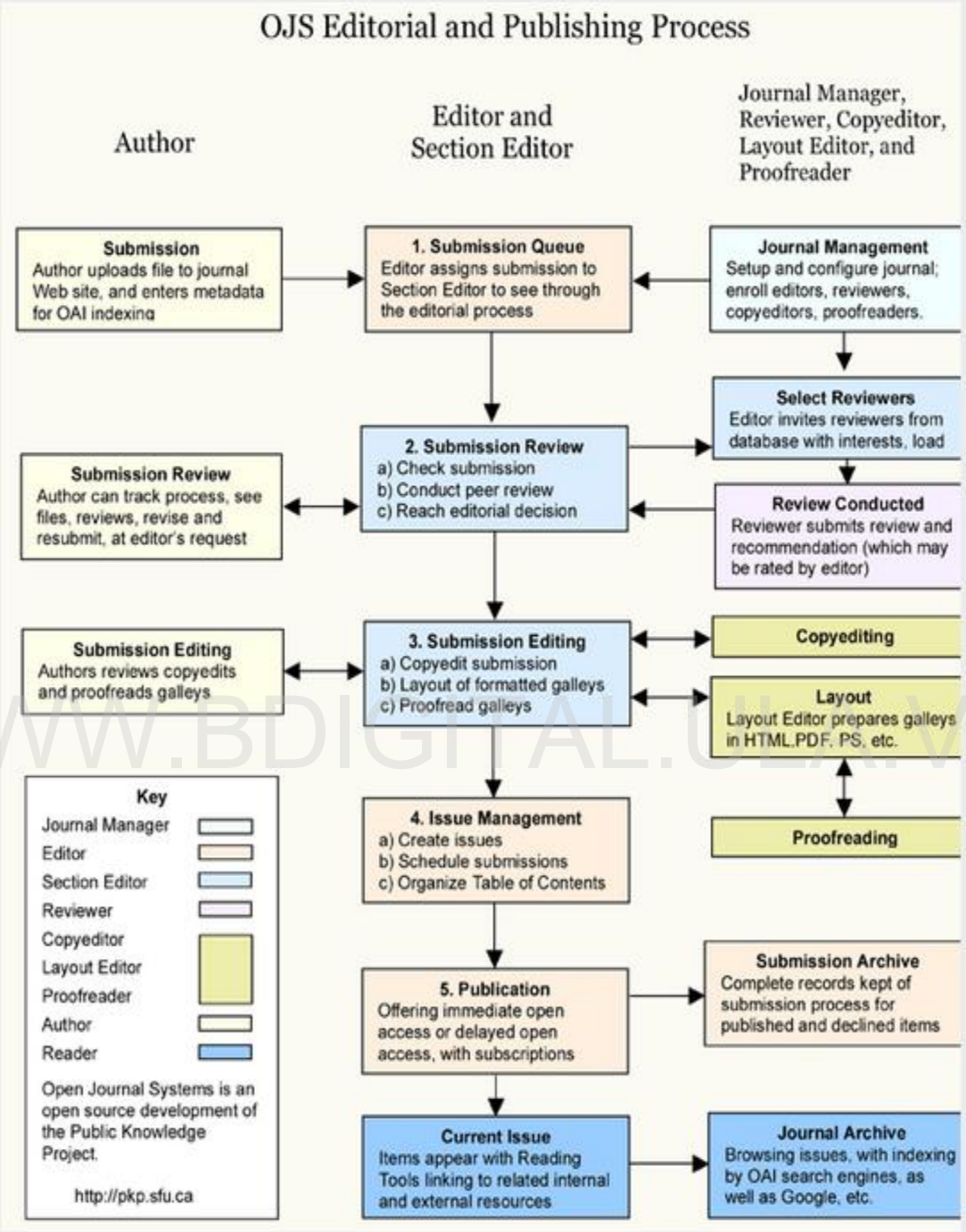

Figura 1 Estructural Open Journal Systems

Fuente: Willinsky, Stranack, Smecher y MacGregor (2010).

2.4 Software utilizado

Software estadístico R

R es un conjunto integrado de servicios de software para la manipulación de datos, cálculo y representación gráfica. En este sentido, The R Project for Statistical Computing lo define como:

> R es un lenguaje y entorno de computación y gráficos estadísticos. Es un proyecto GNU, que es similar al lenguaje S y el medio ambiente, fue desarrollado en los Laboratorios Bell (antes de AT & T, ahora Lucent Technologies) por John Chambers y sus colegas. R se puede considerar como una implementación diferente de S. Existen algunas diferencias importantes, pero mucho código escrito para S se ejecuta inalterado bajo R.

Incluye un manejo eficaz de los datos y la instalación de almacenamiento, un conjunto de operadores para los cálculos de matrices, en las matrices particulares una gran coherencia, colección integrada de herramientas intermedias para el análisis de datos, instalaciones gráficas para el análisis y visualización de datos, ya sea en pantalla o en copia impresa, simple y eficaz lenguaje de programación bien desarrollado que incluye condicionales, bucles, funciones recursivas definidos por el usuario y las instalaciones de entrada y salida.

La facilidad en el uso de los modelos matemáticos y en el diseño de calidad de los reportes sobre algún estudio, y el control general sobre los reportes por el usuario, hacen de R una herramienta valiosa en el estudio de la estadística. Al respecto, Zambrano (2012) sostiene lo siguiente:

> Dentro de sus paquetes estadísticos, se pueden observar variedad de modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series temporales, clasificación, "*clustering*", el uso de gráficos como son los histogramas, los diagramas de caja, los polígonos de frecuencia y

> otros. La importancia que tiene R es que facilita tu código abierto permitiendo que se aporte al desarrollo de su estructura (p. 45).

Este software ofrece una amplia variedad de técnicas gráficas estadísticas y es altamente extensible. El lenguaje S es, a menudo, el vehículo de elección para la investigación en metodología estadística, y R proporciona una ruta de código abierto para la participación en esa actividad.

Uno de los puntos fuertes de R es la facilidad con la que se han diseñados parcelas con calidad de publicación pueden ser producidos, incluidos los símbolos y fórmulas matemáticas cuando sea necesario. CRAN ("*Comprehensive R Archive Network"*) se ha hecho cargo de los valores predeterminados para las opciones de diseño de menor importancia en los gráficos, pero el usuario mantiene el control total.

Es relevante señalar que está disponible como software libre, bajo los términos de la "*Free Software Foundation's*" Licencia Pública General de GNU en forma de código fuente. Se compila y se ejecuta en una amplia variedad de plataformas UNIX y sistemas similares (incluyendo FreeBSD y Linux), Windows y MacOS (Zambrano, 2012).

R como S está diseñado en torno a un cierto lenguaje de programación, y permite a los usuarios añadir funcionalidad adicional mediante la definición de nuevas funciones. Gran parte del sistema está a su vez escrito en el dialecto R de S, lo que hace que sea fácil para los usuarios seguir las elecciones hechas algorítmicas. Para las tareas computacionalmente intensivas, C, C++ y Fortran se pueden vincular y llamados en tiempo de ejecución. Los usuarios avanzados pueden escribir código C para manipular objetos R directamente.

Muchos usuarios prefieren pensar en R como un entorno en el que se implementan técnicas estadísticas. R puede ser extendido (fácilmente) a través de paquetes. Hay cerca de ocho paquetes suministrados con la distribución de R y

muchos más están disponibles a través de la familia CRAN, de sitios de Internet que cubren una gama muy amplia de la estadística moderna. R tiene su propio formato de documentación similar al látex, que se utiliza para suministrar una amplia documentación, tanto en línea en varios formatos y en papel.

Software estadístico S Plus

S-PLUS es una implementación distribuida comercialmente del lenguaje de programación S. Se desarrolla, distribuye y está apoyado por TIBCO ("*The Information Bus Company"*) Software Inc. S-PLUS se escribe y se ejecuta en el TIBCO Spotfire S entorno de programación estadística. Los estadísticos e investigadores de una amplia gama de industrias utilizan S-PLUS para llevar a cabo el análisis estadístico avanzado en grandes conjuntos de datos. Entre los estadísticos de cálculo habituales incluye: pruebas de hipótesis y construcción de intervalos de confianza, análisis de varianza, análisis exploratorio de datos, entre otros.

Este programa se inicia a finales de 1970 en los Laboratorios Bell. Fue diseñado para ser un lenguaje de computación estadístico, que haría más fácil crear software de análisis estadístico. S-PLUS es una versión propietaria de la lengua S, que fue desarrollado por Ciencias Estadísticas en 1988 para su uso dentro de sus paquetes de software de análisis estadístico. Ciencias Estadísticas, junto con todos sus productos y el lenguaje S-PLUS, fueron adquiridas por MathSoft en 1993.

La versión que se va a utilizar para la investigación es la 6.1, la cual incluye dentro de sus principales características: análisis multivariado de datos, análisis de sobrevivencia, escalamiento multidimensional y regresión no paramétrica. Entre sus características se pueden mencionar entorno flexible para el análisis de datos, una colección extensa y coherente de herramientas estadísticas para análisis de datos, un lenguaje para expresar modelos estadísticos y herramientas para utilizar modelos estadísticos lineales y no lineales, facilidades para el análisis de datos y su

presentación tanto en la computadora como en papel, un lenguaje de programación orientado a objetos que puede ser fácilmente extendido.

En este sentido, S-PLUS y R son las dos implementaciones modernas del lenguaje de programación S. En la práctica, hay algunas diferencias en términos de ámbito léxico, modelos y una variedad de diferencias sintácticas menores. Sin embargo, los tres idiomas son muy similares y una gran cantidad de código se puede ejecutar igualmente bien en los tres entornos.

2.5 Análisis textual

El análisis textual de datos del presente proyecto se realizará a través de la minería textual que, según Eíto y Senso (2004), "es una aplicación de la lingüística computacional y del procesamiento de textos que pretende facilitar la identificación y extracción de nuevo conocimiento a partir de colecciones de documentos" (p. 11). De esta manera, la minería de texto recoge distintas técnicas formuladas en el ámbito de la recuperación textual y la lingüística computacional. Asimismo, Sullivan citado por ibídem (2004), señala que:

> La minería de texto es el proceso de compilar, organizar y analizar grandes colecciones de documentos para apoyar en la distribución de información a los analistas y a las personas encargadas de tomar decisiones, y para descubrir relaciones entre hechos relacionados que se reparten entre distintos dominios de investigación (p. 13).

Así pues, la minería de texto tiene como objetivo procesar y presentar la información disponible en grandes colecciones de documentos en un formato que facilite su comprensión y análisis. También permite crear la distribución de información para descubrir las relaciones entre hechos interconectados que pertenecen a distintos dominios de investigación, y así apoyar la toma de decisiones.

En otro sentido, un aspecto importante a favor de la minería de texto es que con ella se están materializando numerosas técnicas desarrolladas en el marco de la recuperación de información. Las herramientas de minería de texto ofrecen una ayuda importante en el proceso de acceso e interpretación de la información disponible en los documentos.

Una de las herramientas más utilizadas para realizar minería de texto es a través de la librería Tm (paquete de minería de textos) y "*Wordcloud"* (paquete generador de nube de palabras), los cuales permiten analizar textos y visualizar rápidamente las palabras clave como una nube de palabras (también referidas a nube de texto o nube de etiquetas), siendo así un formato que facilita su análisis y la deducción de nuevas conclusiones.

2.6 Nubes de palabras

Una nube de palabras o nube de etiquetas (a las que también se alude como nube de "*tags*", al proceder del inglés "*tag cloud*") es una representación visual de palabras o tags generadas automáticamente. En este sentido, Guallar, Orduña y Olea (2014) señalan que:

> Las tag clouds (nubes de etiquetas) representan las categorías en forma de nube de palabras, donde el color y el tamaño de la palabra codifican si hay o no muchos documentos recuperados bajo esa categoría. Normalmente las etiquetas que aparecen en la nube suelen ser hipervínculos que llevan al listado de los documentos que han sido recuperados bajo esa etiqueta (p. 303).

En este sentido, la nube de palabras visualiza gráficamente el peso específico de cierto tema y permite acceder directamente a todas las palabras publicadas sobre el mismo. Proporcionan un medio para que los usuarios se formen una impresión general del conjunto de los contenidos y la esencia de lo que trata.

Uno de sus principales usos es la visualización de las etiquetas de un sitio web de modo que los temas más frecuentemente tratados por dicho sitio se muestren con mayor relevancia. Por lo general, se muestran en orden alfabético y ponderado visualmente dependiendo de su popularidad. El tamaño de las etiquetas se corresponde con su frecuencia de uso.

En este tipo de visualización las palabras por sí solas pueden dar información, pero la manera en que son representadas le agregan valor adicional. Aunque todas las palabras que aparecen en la representación visual son intereses de los usuarios, se pueden observar que algunos intereses predominan más que otros. Además, pueden evolucionar a medida que los datos asociados cambian a través del tiempo.

Las nubes de etiquetas, adquieren relevancia al comenzar el siglo XXI como una característica de los primeros sitios webs y blogs de la denominada Web 2.0, para mostrar la distribución y frecuencia de palabras clave que describen el contenido del sitio, siendo una herramienta que ayuda a la navegación al disponer de información más rápida.

Existen distintas aplicaciones para la creación de nubes de palabras que permiten implementar después los resultados en otras herramientas o utilizarlas para analizar los términos que se utilizan en determinados textos, siendo Wordle y Tagxedo aplicaciones en línea ampliamente utilizadas para la creación de dichas nubes. Al respecto, Gómez (2013), explica:

> Wordle es una aplicación gratuita disponible en línea sin necesidad de descargar ni instalación, que permite la generación de nubes de etiquetas a partir de un texto, de una dirección URL o del perfil de Delicious. Las nubes se pueden presentar con distintos formatos.
> Incluye una serie de opciones para eliminar los números de la nube, para mantener las palabras tal como se han escrito, para convertir todas las palabras en minúscula, o en mayúsculas. Asimismo, también se pueden eliminar las palabras comunes de un idioma, o mantenerlas; esta última

es la opción que permite dejar palabras muy frecuentes como artículos, preposiciones…
Para modificar la nube podemos cambiar el tipo de fuente y el color, pero además podemos crear de manera aleatoria otros formatos de nube. También se puede fijar el máximo de palabras que conforman la nube. Otra de las opciones es presentar la nube con las palabras ordenadas alfabéticamente, lo que facilitará la localización de las palabras. Otras funciones son el tipo de borde en la nube: redondeado o recto; la orientación de todas las palabras en horizontal o de la mayoría, y lo mismo para la orientación vertical; que utiliza todas las orientaciones incluida la inclinada.
Por su parte, Tagxedo se trata de una herramienta puesta a disposición de manera gratuita en la Web, por lo que no es necesaria la instalación de ningún programa específico, aunque sí requiere tener instalado el *plugging* Microsoft Silverlight.
Crea nubes de etiquetas a partir de una dirección URL, una identidad de Twitter, de Delicious, de noticias, de búsquedas o de canales RSS, o con cualquier texto, ya sea cargando el fichero o introduciéndolo en la caja de búsquedas. También se puede cargar un fichero en formato XAP (formato propietario de aplicaciones generadas con Silverlight) haciendo clic sobre *browse* y buscándolo a través del explorador de archivos.
El programa permite hacer nubes de palabras donde se puede personalizar la fuente utilizada, el color, la orientación y, lo que supone una novedad frente a Wordle, elegir de la forma de la nube y que las palabras rellenen la imagen, o sea, solo el contorno. Una vez que se ha creado la nube, es posible guardarla, imprimirla, compartirla e incluso guardar la imagen en miniatura (p. 42).

2.7 Nubes de etiquetas con texto en línea

Se refiere a un párrafo (bloque) hecho exclusivamente de Inline HTML elementos, tales asspan, font, em, b, i, fuerte, y el Hno. Una etiqueta, incluso uno con espacios, debe permanecer en una sola línea. El espacio en blanco fuera de las etiquetas está en una determinada fuente y tamaño de fuente predeterminado. Cualquier área fuera de la etiqueta, pero dentro de la nube de etiquetas será referida como "blancos", irre- perspectiva del color de fondo.

Las fuentes y los tamaños de las fuentes correspondientes a diferentes etiquetas se imponen mediante el inline- ements con, por ejemplo, el atributo de estilo HTML.

El ancho disponible para la nube de etiquetas también está determinado dependiendo del diseño de la página, pero la altura de la nube de etiquetas se supone un parámetro libre. Naturalmente, las fuentes y tamaños de fuente, así como la etiqueta-nube ancho se determinan mediante el navegador Web, así como por el contenido de la página.

Mientras que las propiedades CSS "*letter-spacing*" y "*word-spacing*" permiten cambiar el ancho de las frases, hay limitaciones específicas de aplicación. La principal vista tiene el ancho y el alto de cada etiqueta fija. Asimismo, el espacio horizontal entre las etiquetas debe ser al menos tan grande como el espacio normal en la fuente predeterminada. Por lo tanto, no se va a incluir una pena para exprimir las etiquetas o espacios. Esto está en consonancia con la actual raza de motores de diseño del explorador Web.

De igual manera, las etiquetas son comúnmente ordenadas alfabéticamente en las nubes; sin embargo, no se encuentra ninguna evidencia de que realmente los usuarios naveguen nubes de etiquetas alfabéticamente. Para grandes nubes, un sencillo cuadro de búsqueda ECMAScript destacando "*tags"* comenzando con algún texto, puede resultar conveniente buscar etiquetas específicas.

2.8 Recurrencia

Los datos de tiempos de eventos son importantes en muchos campos de aplicación y surgen en diversas áreas del conocimiento, particularmente cuando un sujeto o unidad es monitoreado en el tiempo y durante un evento. Al respecto, Lopera y Manotas (2011), señalan lo siguiente:

> Los datos recurrentes surgen cuando una unidad (o un grupo de tales unidades) es monitoreada a través del tiempo y un evento particular (o grupo de eventos) ocurre en varios puntos del periodo de observación, por ejemplo, los tiempos de episodios recurrentes de una enfermedad en pacientes o los tiempos de reparación de un producto manufacturado.

> Muchos sistemas, subsistemas y componentes (que genéricamente son denominadas “unidades”) tienen asociadas más de una causa o modo de falla (p. 249).

Es así como en algunas aplicaciones, y para ciertos propósitos, es importante distinguir entre las causas o modos de falla. Para mejorar la confiabilidad es esencial identificar la causa de falla hasta el nivel de componente y, en muchas aplicaciones, hasta la causa física real de una falla. En el presente trabajo, se presenta una aplicación del análisis de datos recurrentes en las publicaciones de los autores que incluye el uso de métodos estadísticos no paramétricos y paramétricos.

Para Meeker y Escobar (1998), los procesos recurrentes son vistos como una sucesión de tiempo en los cuales ocurre el evento. Por ejemplo, $T_1, T_2, \ldots,$ puede ser una reparación, una falla, un reemplazamiento, entre otros. Generalmente, son observados por el investigador en intervalos de tiempo previamente fijados.

Ahora bien, en situaciones donde la unidad es estudiada en un periodo de tiempo que excluye los registros de ocurrencias previas al inicio del periodo de observación, se dice que la historia del evento de la unidad es censurada por la izquierda (edad censurada por la izquierda).

Asimismo, las situaciones en las cuales durante el periodo de observación hubo intervalos de tiempo en los que no se realizaron los respectivos registros de las recurrencias del evento, se dice que la historia de eventos de la unidad fue censurada por intervalos. Por consiguiente, la censura es una característica presentada por la muestra la cual indica la falta de información de registro de ocurrencias del evento.

Finalmente, los datos de eventos recurrentes se pueden referir tanto a una única unidad como a una colección de unidades. Para una colección de unidades, los datos son combinados en un conjunto de datos que forman un único proceso. El proceso

formado por una colección de unidades se conoce como superposición de varios procesos o proceso superimpuesto (Meeker y Escobar, 1998).

2.9 Función de conteo

Los eventos aleatorios, por ejemplo, las averías de equipos industriales, pueden ser representados mediante una función de conteo $N(t)$, definida para todo $t > 0$, que representa el número de sucesos que han ocurrido durante el periodo de tiempo transcurrido desde 0 a t. El instante 0 indica el tiempo en el que se comienza a observar el suceso aleatorio determinado. Para cada instante t, el valor $N(t)$ es un valor observado de una variable aleatoria. Para cada variable aleatoria $N(t)$, los únicos valores posibles son los enteros 0,1,2, ...

2.10 Procesos de conteo

De acuerdo a lo expuesto por Parzen (1972) (citado en Gomez (2007, p.13)), es un proceso de valores enteros $\{N(t), t \geq 0\}$ que cuenta el número de puntos aparecidos en un intervalo. Estos puntos están distribuidos por un mecanismo estocástico determinado. Un caso típico, es que los puntos representan los instantes $\tau_1, \tau_2, \ldots$en los que han ocurrido los eventos de un carácter especificado (en el presente contexto, son los llamados eventos recurrentes), donde $0 < \tau_1 < \tau_2 < \cdots$.

2.11 Función acumulada promedio (FAP)

La función acumulada promedio (FAP) es la media de la correspondiente distribución del número o costo de eventos recurrentes en alguna edad t. La función acumulada promedio para el número de eventos recurrentes en alguna edad t, $M(t)$, se denota por:

$$M(t) = E[N(t)]$$

La función acumulada promedio para el costo de eventos recurrentes en alguna edad t, se denota por:

$$M_c(t) = E[N_c(t)]$$

Donde $N_c(t) = kN(t)$ y k es una constante. $\{N_c(t), t \geq 0\}$ es llamado proceso de costo. Tanto $M(t)$, como $M_c(t)$, se consideran como una curva media; es decir, el punto medio de toda la población de curvas que pasa a través de la línea vertical en cada edad t. En el caso de eventos discretos, esta curva media es una función de paso con varios pasos pequeños, uno para cada evento de la población.

2.12 Tasa de recurrencia instantánea o función de intensidad

La tasa de recurrencia instantánea o función de intensidad es la derivada de la FAP del número de eventos recurrentes con respecto al tiempo, y se denota por $m(t)$; es decir:

$$M'(t) = m(t)$$

Se le llama instantánea porque depende de la edad del tiempo en el tiempo t. Se interpreta como el número de eventos por unidad de tiempo o de espacio por unidad poblacional (Baena y Salazar, 2006; Nelson, 2003).

2.13 Tasa de costo de fallas recurrentes

De igual manera que la tasa de recurrencia instantánea, la tasa de costo de fallas, es la derivada de la FAP del costo de eventos recurrentes con respecto al tiempo, y se denota por $m_c(t)$, es decir:

$$M_c' = m_c(t)$$

Se interpreta como el costo en unidades monetarias o cualquier unidad que defina al costo del evento recurrente por unidad de tiempo por unidad poblacional.

2.14 Modelo no paramétrico

El modelo no paramétrico para una población de unidades, es la población de las funciones de historia acumulada de todas las unidades. Por otra parte, se entiende como función de historia acumulada para el caso de eventos discretos (eventos que ocurren en un punto del tiempo), al número o costo acumulado de recurrencias del evento discreto como una función del número de unidades de tiempo en estudio; es decir, $N(t)$ y $N_c(t)$, respectivamente. Por tanto, para una única unidad, los datos de recurrencias pueden ser expresados como el número o costo acumulado de recurrencias el intervalo de edad de la unidad, estos es $N_i(t)$ o $N_{c_i}(t)$ en $(0,t)$.

Se asume que un modelo poblacional para el número acumulado de eventos recurrentes discretos es el proceso de conteo, el cual es denotado como $\{N(t), t \geq 0\}$. Este modelo es usado para describir una población de unidades, basado en la FAP, $M(t)$, en la edad t. Además, asumiendo que los modelos no paramétricos y paramétricos para la FAP del número acumulado de eventos recurrentes discretos son aplicables a la FAP del costo acumulado de eventos recurrentes discretos.

A continuación, se presentan las bases teóricas para la estimación no paramétrica, métodos de comparación y el modelo paramétrico para la FAP del número de acumulado de eventos recurrentes discretos.

2.14.1 Estimación no paramétrica de la FAP

La estimación no paramétrica puntual de la FAP puede ser obtenida como se muestra a continuación: sea $N_i(t)$ el número acumulado de recurrencias para la unidad i antes del tiempo t, en donde $i = 1, \ldots, n$, siendo n el conjunto de unidades bajo observación y sea $t_{ij}, j = 1, \ldots, m_i$ los tiempos de recurrencias para la unidad i. Un simple estimador de la FAP poblacional en el tiempo $t, M(t)$, debería ser la media muestral de los $N_i(t)$ valores de las unidades que aún están operando en el tiempo t. En este sentido, Nelson (2003) muestra en detalle diferentes estimadores que son empleados según los tipos de censura que se presenten en datos de eventos recurrentes.

Para la estimación no paramétrica de la FAP, $M(t)$, de datos de edad exacta con censura por la derecha (edad exacta, porque se está considerando el caso de eventos discretos), se deben seguir los siguientes supuestos:

- ✓ La población objetivo está claramente especificada y muestreada.
- ✓ Las unidades de la muestra son una muestra aleatoria simple de la población objetivo.
- ✓ La censura por la derecha de la muestra de historias es aleatoria. Esto equivale a que las historias son estadísticamente independientes de sus edades censuradas.
- ✓ La función de historia de cada unidad en la población se extiende a través de cada unidad del rango de la muestra de datos.
- ✓ La media poblacional, $M(t)$, es finita sobre el rango de edad.

Todas las edades recurrentes son distintas entre sí y distintas de las edades censuradas.

2.14.2 Estimación de la varianza y límites confidenciales para la FAP

Para una muestra aleatoria de $n \geq 2$ funciones acumuladas; es decir, más de dos unidades de una población de unidades finita o infinitas, existe la verdadera $Var[\bar{M}(t_j)]$. Sea $d(t_k)$ el número aleatorio de recurrencias en t_k para una función acumulada seleccionada de forma aleatoria de la población de funciones acumuladas. Entonces, la verdadera varianza de $\bar{M}(t_j)$ para una población de funciones acumuladas (Nelson, 1995) (citado en Gómez (2007, p.15)) es:

$$Var[\bar{M}(t_j)] = \sum_{k=1}^{j} \frac{Var[d(t_k)]}{\delta.(t_k)} + 2\sum_{k=1}^{j-1}\sum_{v=k+1}^{j} \frac{Cov[d(t_k), d(t_v)]}{\delta.(t_k)} \quad (1)$$

Para estimar $Var[d(t_k)]$, se usa el supuesto de que $d_i(t_k)$, $i = 1, \ldots, n$, es una muestra aleatoria de la población de valores $d(t_k)$; es decir, el número de recurrencias en el tiempo t_k para la función acumulada de la i-ésima unidad es una muestra aleatoria del número de recurrencias en t_k de la función acumulada de la i-ésima unidad en la población.

Cuando el número de funciones acumuladas muestreadas es mayor que el 5% o 10% de la población, se sustituyen las ecuaciones (2) y (3) en (1):

$$\hat{Var}[d(t_k)] = \left[1 - \frac{\delta.(t_k)}{N}\right] \sum_{i=1}^{n} \frac{\delta_i(t_k)}{\delta.(t_k)} \left[d_i(t_k) - \acute{d}(t_k)\right]^2 \quad (2)$$

$$\hat{Cov}[d(t_k), d(t_v)] = \left[1 - \frac{\delta.(t_v)}{N}\right] \sum_{i=1}^{n} \frac{\delta_i(t_v)}{\delta.(t_v)} \left[d_i(t_v) - \acute{d}(t_v)\right] d_i(t_k) \quad (3)$$

En las ecuaciones (2) y (3), N es el número total de funciones acumuladas en la población de interés, $\acute{d}(t_k)$ y $\acute{d}(t_v)$ son los cocientes $d(t_v)/\delta.(t_v)$ y $d(t_k)/\delta.(t_k)$ respectivamente.

Por otro lado, los intervalos de confianza del $100(1-\alpha)$ para la FAP en un tiempo específico t basados en: $Z_{\hat{M}(t)} = \frac{[\hat{M}(t)-M(t)]}{\hat{se}_{\hat{M}(t)}}\ NOR(0,1)$ es:

$$\hat{M}(t) \pm z_{(1-\alpha/2)}\hat{se}_{\hat{M}(t)} \qquad (4)$$

donde: $\hat{se}_{\hat{M}(t)} = \sqrt{\hat{Var}[\hat{M}(t)]}$.

Los supuestos a seguir para obtener los límites confidenciales (4) son:

- ✓ La distribución muestral de $\hat{M}(t)$ es cercana a la normal.
- ✓ Todas las varianzas y covarianzas poblacionales en la ecuación (1) existen y son finitas.
- ✓ La población tiene un número infinito de unidades (o historias).

Además de los supuestos anteriores, se deben considerar los supuestos para estimar la FAP poblacional expuestos en la sección II.12

2.14.3 Comparación no paramétrica de dos FAP muéstrales

Sea la diferencia entre las funciones acumuladas promedio en el tiempo t, denotadas por $\Delta_M(t) = M_1(t) - M_2(t)$ de dos poblaciones o procesos con funciones acumuladas promedio $M_1(t)$ y $M_2(t)$, respectivamente. Asumiendo que las dos muestras son estadísticamente independientes, se tiene que:

Un estimador paramétrico para $\Delta_M(t)$ es: $\hat{\Delta}_M(t) = \hat{M}_1(t) - \hat{M}_2(t)$. Ahora, un estimador para la varianza de la diferencia, es decir, $V[\Delta_M(t)]$ es: $\hat{V}\left[\hat{\Delta}_M(t)\right] = \hat{V}\left[\hat{M}_1(t)\right] + \hat{V}\left[\hat{M}_2(t)\right]$.

Además, si se asume que la distribución muestral acumulada de la diferencia observada $\left[\hat{M}_1(t) - \hat{M}_2(t)\right]$ es cercana a la normal, un intervalo de confianza del $100(1-\alpha)$, para $\Delta_M(t)$, basado en $Z_{\hat{\Delta}_M} = \frac{[\hat{\Delta}_M(t) - \Delta_M(t)]}{\hat{se}_{\hat{\Delta}_M}}$ $NOR(0,1)$ es:

$$\hat{\Delta}_M \pm z_{(1-\alpha/2)}\hat{se}_{\hat{\Delta}_M}.$$

Cabe destacar que diferentes tipos de eventos no necesitan ser estadísticamente independientes (Nelson, 2003). Por tanto, no se supone independencia entre los distintos modos de fallas.

2.15 Modelo paramétrico

El proceso de Poisson es el modelo paramétrico más simple de procesos de conteo para el número de recurrencias de un evento en un tiempo observado t. Un proceso de valor entero $\{N(t), t \geq 0\}$ es un proceso de Poisson con tasa de recurrencia $m(t)$ si se satisfacen las siguientes condiciones (Baena y Salazar, 2006; Meeker y Escobar, 1998; Parzen, 1972):

- ✓ El número de recurrencias acumuladas en el tiempo cero, es cero (denotado como N(0) = 0).
- ✓ El número de recurrencias en intervalos de tiempos disjuntos son estadísticamente independientes. Un proceso con esta propiedad, se dice que tiene "incrementos independientes".
- ✓ La tasa de recurrencia m(t) es positiva y M(a,b) = E[N(a,b)]= $\int_a^b m(u)du < \infty$, cuando 0≤ a < b < ∞.

Con esto se sigue, que para un proceso de Poisson, la N(a, b), tiene una distribución de Poisson con función de densidad de probabilidad:

Pr [N (a, b) = d] = $\frac{[M(a,b)]^d}{d!}$exp [−M (a, b)] con d = 0, 1, 2, . . .

2.15.1 Proceso Poisson homogéneo (PPH)

Según Meeker y Escobar (1998), un proceso Poisson homogéneo (HPP, por sus siglas en inglés) es un proceso de Poisson con tasa de recurrencia constante; es decir, m(t) = 1/Θ, si se cumplen las siguientes condiciones:

- ✓ N(a, b) tiene distribución de Poisson con parámetro M(a, b) = (b − a)/Θ.
- ✓ El número esperado de recurrencias en (a, b] es M(a, b). Equivalentemente, el número esperado de recurrencias por unidad de tiempo sobre (a, b] es constante e igual a 1/Θ. Esta propiedad se conoce como incrementos estacionarios.
- ✓ Los tiempos entre recurrencias, τ_j = Tj − Tj−1, son independientes e idénticamente distribuidos (iid), con distribución exponencial con parámetros Θ, es decir, exp (Θ).
- ✓ El tiempo Tk = $\tau_1 + \cdots + \tau_k$ para k-ésima ocurrencia tiene una distribución GAM (Θ, k).

2.15.2 Proceso Poisson no homogéneo (PPNH)

Según Meeker y Escobar (1998), un proceso de Poisson no homogéneo (PPNH) es un proceso de Poisson con una tasa de recurrencia no constante m(t); por tanto, los tiempos de ocurrencias no son idd.

Con un PPNH, el numero esperado de recurrencia por unidad sobre el tiempo (a,b] es:

$$\frac{M(a,b)}{b-a} = \frac{1}{b-a}\int_a^b m(t)dt.$$

A menudo, un modelo PPNH es especificado en términos de la tasa de recurrencia m(t), usando m(t)m(t, Θ), la cual es una función del vector Θ de parámetros desconocidos.

El modelo PPNH regla de potencia para la tasa de recurrencia es:

$$m(t; \beta,\eta)=\frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} \text{ Con } \beta > 0, \eta > 0$$

En donde el correspondiente promedio del número acumulado de recurrencias sobre (0,t]; es decir FAP, viene dada por: $M(t; \beta,\eta)= \left(\frac{t}{\eta}\right)^{\beta}$. Por lo que este modelo se reduce al modelo proceso de Poisson homogéneo cuando β=1.

Por otro lado, un modelo PPNH Loglineal para la tasa de recurrencia es:

$$m(t{:}\gamma_0,\gamma_1) = exp(\gamma_0 + \gamma_1 t)$$

En donde, el promedio del número acumulado de recurrencias sobre el tiempo (0,t], es decir la FAP, viene dada por: $M(t,\gamma_0,\gamma_1 = \frac{[exp(\gamma_0)][exp(\gamma_1 t)-1]}{\gamma_1}$. Por consiguiente, este método se reduce al modelo proceso de Poisson homogéneo cuando $\gamma_1 = 0$.

2.15.3 Pronóstico de recurrencias futuras con un proceso Poisson

Según Mekeer y Escobar (1998), el número esperado de recurrencias en un intervalo [a, b] es $\int_a^b m(t,\theta)dt$. El estimador de máxima verosimilitud de una predicción puntual es $\int_a^b m(t,\hat{\theta})dt$. Una predicción puntual para el número de recurrencias usando el modelo PPNH potencial es:

$$\int_a^b m(t,\hat{\theta})dt = \left(\frac{1}{\hat{\eta}}\right)^{\hat{\beta}} \left(b^{\hat{\beta}} - a^{\hat{\beta}}\right)$$

Similarmente, una predicción puntual para el número futuro de recurrencias utilizando el modelo PPNH loglineal es:

$$\int_a^b m(t,\hat{\theta})dt = \frac{[exp(\hat{\gamma}_0)]}{\hat{\gamma}_1}[exp(\hat{\gamma}_1 b) - exp(\hat{\gamma}_1 a)]$$

2.15.4 Procesos de Renovación

Una secuencia de recurrencias en los tiempos T_1, T_2… es un proceso de renovación si los tiempos entre recurrencias $T_j = T_j - T_{j-1}$, con $j = 1,2, ...,$ y $T_0 = 0$, son independiente e idénticamente distribuidos (iid). La MCF para un proceso de renovación se conoce también como la función de renovación. Es de notar que un HPP es un proceso de renovación (cuyos tiempos entre recurrencias se distribuyen exponencial con T_j= 1/m(t)) y que el NHPP, no lo es (Meeker y Escobar, 1998).

Antes de usar un modelo de proceso de renovación, es importante evaluar los supuestos del modelo tales como la tendencia y la no-independencia de los tiempos entre recurrencias (es importante destacar que, en general, incrementos independientes no son lo mismo que tiempos entre recurrencias independientes). De esta manera, cuando se prueba que un modelo es un proceso de renovación, los pronósticos se pueden hacer utilizando las distribuciones teóricas conocidas; es decir, haciendo inferencia con estas (ob. cit.).

En este sentido, los autores anteriormente citados sostienen que las características típicas de interés de un proceso de renovación, incluyen:

- ✓ La distribución de los valores T_j.
- ✓ La distribución del tiempo hasta la k-ésima recurrencia, con k = 1, 2, ...
- ✓ La tasa de recurrencia (o renovación en este caso).
- ✓ El número de recurrencias que serán observadas en un intervalo de tiempo futuro dado.

Cuando la curva de la FAP estimada no corresponde a una línea recta, se dice que el proceso es no homogéneo y, por ende, el tiempo entre recurrencias no se distribuye exponencial. Pero cuando esto no ocurre; es decir, que la curva es aparentemente una línea recta, se sospecha que se puede estar tratando de un PPH o un proceso de renovación y es necesario entonces utilizar ciertas herramientas para evaluar los respectivos supuestos. En el capítulo 16 de Meeker y Escobar (1998), se presentan dichas herramientas. Es decir, si la curva de la FAP no es una línea recta, HPP y proceso de renovación se descartan. Pero si la curva es, aparentemente una línea recta, entonces se deben probar los supuestos.

De todas formas, SPLIDA permite evaluar si un proceso es HPP o NHPP utilizando el modelo regla de la potencia (donde se estiman β y η y el modelo log lineal (donde se estiman γ_0 y γ_1)). Antes de elegir un NHPP es importante descartar si se trata de un proceso de renovación, debido a que puede suceder que los datos no correspondan a un HPP, pero si a un proceso de renovación. La versión de SPLIDA utilizada en el presente trabajo, no tiene disponible el módulo para evaluar si se trata de un proceso de renovación. Sin embargo, es posible descartar el proceso de renovación con una prueba de bondad de ajuste de los tiempos entre recurrencia, pues estos deben ser iid y su distribución debe ser la exponencial.

CAPÍTULO III

MARCO METODOLÓGICO

3.1 Análisis Textual (Minería de texto)

El análisis textual de datos en el presente trabajo se realizó a través de la minería textual, que es una aplicación de la lingüística computacional y del procesamiento de textos, que consiste en compilar, organizar y analizar grandes colecciones de documentos para facilitar la identificación y extracción de nuevo conocimiento. Asimismo, la minería de texto también permite crear la distribución de información para descubrir las relaciones entre hechos interconectados que pertenecen a distintos dominios de investigación, y así apoyar la toma de decisiones.

Una de las herramientas más utilizadas para realizar minería de texto fue a través de la librería Tm (paquete de minería de textos) y Wordcloud (paquete generador de nube de palabras), las cuales permitieron analizar textos y visualizar rápidamente las palabras clave como una nube de palabras, también referidas a nube de texto o nube de etiquetas, siendo así un formato que facilita su análisis y la deducción de nuevas conclusiones.

Inicialmente, se elaboró una base de datos extrayendo los títulos, resúmenes y palabras claves de la revista Ciencia e Ingeniería, seguidamente se guardaron en formatos TXT ordenados por cada año, desde el 2000 al 2016, para proceder a realizar el análisis de los datos, para el cual se realizó la instalación de las siguientes librerías del paquete estadístico R Studio versión 3.3.1:

install.packages ("tm") paquete de minería de texto
install.packages ("SnowballC") paquete de texto derivado

install.packages ("wordcloud") paquete generador de nube de palabras
install.packages ("RColorBrewer") paquete generador de paletas de colores

Una vez instaladas dichas librerías, se cargaron como se muestra a continuación:

Library ("tm")
Library ("SnowballC")
Library ("wordcloud")
Library ("RColorBrewer")

Ahora bien, para el análisis de texto se procedió a realizar los siguientes pasos:

1. Se cargaron los datos que se encuentran guardados en texto planos (TXT). Estos archivos contenían los títulos, resúmenes y palabras claves.
2. Se creó la función para depurar los textos. Con esta función se eliminaron los caracteres especiales, signos de puntuación, números, acentuaciones, entre otros. A continuación, se muestra un fragmento de la función depurar:

```
depurar<-function(direccion){

  dat<-read.delim(direccion, header=FALSE)
  df<-as.data.frame.data.frame(dat)
  df<-df$V1
  df<-gsub("[[:punct:]]","",df[])
  df<-gsub("[[:digit:]]","",df[])
  return(df[])
}
```

3. Luego se creó un método para buscar los archivos cargados y guardar los archivos depurados.

```
setwd("~/")
dr="~/articulosdep/"
dr1<-matrix(0,1,17) #crea vector que almacena las direcciones de los archivos
    for(i in 0:16){

      dr1[1,i+1]=paste0(dr,2000+i,"/") #asigna la dirección
      setwd(dr1[1,1+i])# se mueve a la dirección
      resumen<-depurar(paste0(dr1[1,1+i],"R",2000+i,".txt"))#guarda los datos de
resumen depurados en el año i
      titulo<-depurar(paste0(dr1[1,1+i],"T",2000+i,".txt"))#guarda los datos de titulo
depurados en el año i
      palc<-depurar(paste0(dr1[1,1+i],"PC",2000+i,".txt"))#guarda los datos de
palabras claves depurados en el año i

  setwd("~/articulosdep/resdep") #cambia de directorio
  write(resumen,paste0("R",2000+i,".txt"), sep='\t')# crea el archivo.txt depurado por
año

  setwd("~/articulosdep/restit") #cambia de directorio
  write(titulo,paste0("T",2000+i,".txt"), sep='\t')

  setwd("~/articulosdep/respc") #cambia de directorio
  write(palc,paste0("PC",2000+i,".txt"), sep='\t')
    }
```

4. Se creó una carpeta para concatenar los archivos por cada quinquenio, por ejemplo, “quin1”, “quin2”, “quin3”. Estas carpetas recogieron toda la información de cada 5 años.
5. Después de concatenar los archivos, se realizó la minería de texto de la siguente manera:

- ✓ Se cargaron todos los archivos como un corpus:

```
mi.corpus<-Corpus(DirSource(my.path),readerControl =
list(read=readPlain,language="spanish"))
```

- ✓ Se limpiaron de nuevo los archivos utilizando la función (Tm_map). Esta función convertía las palabras en minúscula, quitaba las palabras comunes y los espacios en blancos extras.

```
doc <- tm_map(mi.corpus[t],content_transformer(tolower))
# remueve numeros
doc <- tm_map(doc[], removeNumbers)
# Elimina palabras comunes en español
doc <- tm_map(doc[], removeWords, stopwords("spanish"))
doc <- tm_map(doc[], removeWords, sw1)
doc <- tm_map(doc[], removeWords, sw)
#Quitar palabra personalizadas
#Especifique sus palabras personalizadas como un vector de caracteres
#doc <- tm_map(doc[], removeWords, c("propiedad", "alto","flujo"))
# quita puntuaciones
doc <- tm_map(doc[], removePunctuation)
# Elimina espacios en blancos extra
doc <- tm_map(doc[], stripWhitespace)
```

- ✓ Se construyó una matriz de término-documento. Esta es una tabla que contiene la frecuencia de las palabras mediante la función TermDocumentMatrix:

```
#Crea la matriz termino-documeto
dtm <- TermDocumentMatrix(doc)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
```


- ✓ Luego se crearon dos nubes de palabra: una que mostraba las palabras frecuentes, y otra que muestra la comparación de los datos por medio del WordCloud.

```
wordcloud(names(v),freq =
v,max.words=k,random.order=F,rot.per=0,colors=brewer.pal(7,"Dark2"))
```

```
comparison.cloud(m,max.words=10,random.order=F,rot.per=0,colors=brewer.pal(7,"Dark2")
```

6. Todos los pasos para la minería de texto se llevaron a dos funciones: una para generar las nubes de frecuencias de las palabras (función **nube**), y otra para generar la nube de comparación de las palabras en todos los archivos en

estudio (Función **nube2**) con la finalidad de agilizar la creación de las nubes para cada año.

Función nube

```
nube<-function(my.path,k=15){
mi.corpus<-Corpus(DirSource(my.path),readerControl =
list(read=readPlain,language="spanish"))
doc <- tm_map(mi.corpus[],content_transformer(tolower))
doc <- tm_map(doc[], removeNumbers)
doc <- tm_map(doc[], removeWords, stopwords("spanish"))
doc <- tm_map(doc[], removeWords, sw1)
doc <- tm_map(doc[], removeWords, sw)
doc <- tm_map(doc[], removePunctuation)
doc <- tm_map(doc[], stripWhitespace)
dtm <- TermDocumentMatrix(doc)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
wordcloud(names(v),freq =
v,max.words=k,random.order=F,rot.per=0,colors=brewer.pal(7,"Dark2"))
}
```


Función nube2

```
nube2<-function(my.path){
  mi.corpus<-Corpus(DirSource(my.path),readerControl =
list(read=readPlain,language="spanish"))
  doc <- tm_map(mi.corpus[],content_transformer(tolower))
  doc <- tm_map(doc[], removeNumbers)
  doc <- tm_map(doc[], removeWords, stopwords("spanish"))
  doc <- tm_map(doc[], removeWords, sw1)
  doc <- tm_map(doc[], removeWords, sw)
  doc <- tm_map(doc[], removePunctuation)
  doc <- tm_map(doc[], stripWhitespace)
  dtm <- TermDocumentMatrix(doc)
  m <- as.matrix(dtm)
  v <- sort(rowSums(m),decreasing=TRUE)
  d <- data.frame(word = names(v),freq=v)
  head(d, 10)

  comparison.cloud(m,max.words=10,random.order=F,rot.per=0,colors=brewer.pal(7,"Dark2")
)
  }
```

7. Luego, se hizo el llamado de las funciones pasándoles como parámetros el año a evaluar, para así generar las nubes de palabras para palabras claves, resúmenes, títulos y la de comparación. Este procedimiento se realizó desde el año 2000 al 2016.

```
#Año 2000
##nube de las palabras claves

```{r}
nube("AT2000",1,k=15)
```


##nube de los resumenes

```{r}
nube("AT2000",2,k=15)
```

##nube de los titulos

```{r}
nube("AT2000",3,k=15)
```

##nube comparativa

```{r}
nube2("AT2000")
```
```

8. Por último, se crearon las nubes de palabras por quinquenio utilizando las dos funciones creadas anteriormente y las carpetas generadas en pasos previos a las funciones.

```
##nube desde 2007 al 2011

```{r, fig.height=7,fig.align='center', fig.width=9,}

nube("quin2",30)
nube2("quin2")
```
```

3.2 Análisis de datos de eventos recurrentes

La estadística paramétrica es una rama de la estadística inferencial que comprende los procedimientos estadísticos y de decisión, que están basados en las distribuciones de los datos reales. La mayoría de procedimientos paramétricos

requiere conocer la forma de distribución para las mediciones resultantes de la población estudiada. Para la inferencia paramétrica, es requerida como mínimo una escala de intervalo; esto quiere decir que los datos deben tener un orden y una numeración del intervalo.

Se elaboró una base de datos extrayendo de la revista Ciencia e Ingeniería los nombres de los autores de los artículos, así como el año en que fueron publicados y el volumen de edición de la revista. A estos datos se les realizó un filtro con respecto a las publicaciones de los autores, donde solo se consideraron aquellos que realizaron dos o más publicaciones. Seguidamente se hizo una transformación de la variable año a meses para así determinar las publicaciones en periodos de un mes.

Ahora bien, para realizar el análisis de recurrencia se siguió la metodología de Meeker y Escobar (1998), y de Nelson (2003), la cual consiste en la validación de supuestos, estableciendo el modelo más adecuado para la tasa de recurrencia, la cual se hizo siguiendo los siguientes pasos que se presentan a continuación:

3.2.1 Análisis no paramétrico

1. Se graficó la estimación puntual no paramétrica de la función acumulada promedio (FAP) con intervalo de confianza de 95% para la estimación de cada publicación recurrente de los autores.
2. Análisis e interpretación de los gráficos obtenidos en el paso anterior.

Para el cálculo de los límites confidenciales para la estimación no paramétrica de la FAP, se requiere que la distribución de la FAP sea cercana a la normal. Por tanto, la evaluación del supuesto de normalidad se realizó previo al paso 1.

3.2.2 Análisis paramétricos

1. Evaluación de los modelos procesos de Poisson homogéneo (PPH) y el proceso de Poisson no homogéneo (PPHN).

El proceso de Poisson es un modelo relevante tanto en las aplicaciones como en la teoría general de los procesos estocásticos de tiempo continuo, que consiste en contar eventos que ocurren a lo largo del tiempo. El tiempo entre cada par de eventos consecutivos tiene una distribución exponencial con el parámetro λ, y cada uno de estos tiempos entre llegadas se supone que es independiente de otros tiempos entre llegadas.

2. Se evaluó el resultado del paso anterior, para decidir qué modelo se lleva a cabo para estimar las recurrencias de la publicación de los autores.
3. Se realizó un pronóstico para predecir la futura recurrencia de publicaciones utilizando el modelo más adecuado.

Las pruebas para evaluar el tipo de proceso que mejor se ajusta a los datos de la Revista de Ciencia e Ingeniera fueron:

Prueba 1. Pruebas para la evaluación de un proceso de Poisson homogéneo.

Consisten en la estimación de los parámetros η y β del modelo PPNH Regla de Potencia y de los parámetros γ_o y γ_1 del modelo PPNH Log-lineal. Con los resultados de esta estimación se evaluó si $\hat{\beta}$ es significativamente igual a 1 y si $\hat{\gamma}_1$ es significativamente igual a 0. Según la teoría, si $\beta = 1$ y $\gamma_1 = 0$ entonces el proceso equivale a un proceso de Poisson homogéneo (PPH).

Prueba 2. Prueba para la evaluación de un proceso de renovación.

En el caso que se descarte un PPH, se procede a realizar esta prueba, debido a que si el proceso de la tasa de recurrencia no corresponde a un PPH puede ser que corresponda a un proceso de renovación. Para ello se puede realizar una prueba de bondad de ajuste de los tiempos entre recurrencias para ver si estos se distribuyen en forma exponencial.

Según la teoría, no se está en presencia de un proceso de Poisson homogéneo; por lo tanto, se realizó las pruebas de bondad siguiendo la metodología aplicada por Baena y Salazar (2006), de acuerdo al teorema de Drenick (1960), para saber si se encontraba ante un proceso de renovación. Es importante destacar que la versión de SPLIDA que se utilizó no tenía disponible el módulo para evaluar si se trataba de un proceso de renovación. Sin embargo, es posible descartar el proceso de renovación con una prueba de bondad de ajuste de los tiempos entre recurrencias, pues estos deben ser iid y su distribución debe ser la exponencial.

Prueba 3. Prueba para la evaluación de un proceso de Poisson no homogéneo.

En caso de llegarse a descartar los procesos de Poisson y renovación, se procede a graficar el ajuste de los modelos PPNH Regla de Potencia y PPNH Log-lineal, para realizar las comparaciones entre los modelos y determinar cuál modelo se ajusta mejor a los datos de autores recurrentes de la revista.

En la figura 2, se presenta un esquema que describe la metodología de análisis de datos de eventos recurrentes empleados en el presente trabajo.

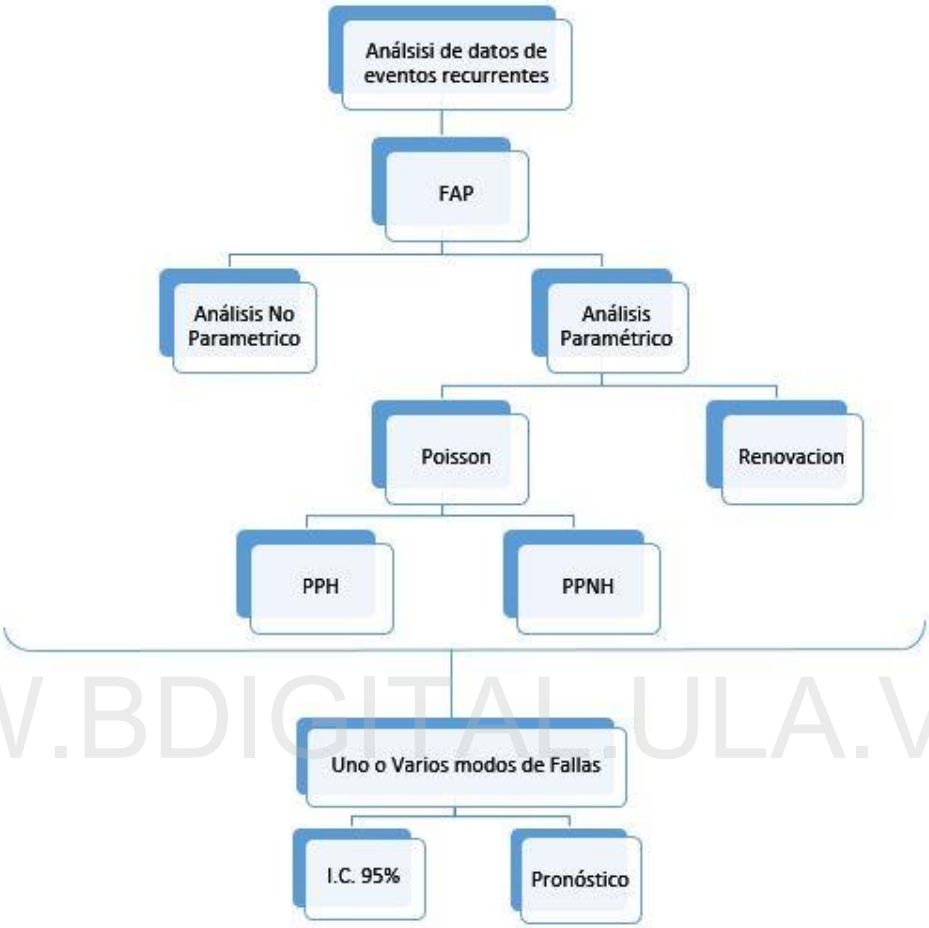

Figura 2. Metodología de análisis de datos de eventos recurrentes empleados.

Fuente: Baena y Salazar (2006).

3.3 Recopilación de datos para el análisis textual

Para la recopilación de los datos, fue necesario extraer de las Revistas de Ciencia e Ingeniería los títulos, resúmenes y palabras claves de cada artículo. Luego, se guardaron en archivos de textos planos (.txt) en carpetas separadas por año desde el 2000 al 2016, donde se observaron un total de 400 artículos distribuidos de la siguiente manera:

- ✓ Año 2000: 13 artículos.
- ✓ Año 2001: 18 artículos.
- ✓ Año 2002: 21 artículos.
- ✓ Año 2003: 22 artículos.
- ✓ Año 2004: 20 artículos.
- ✓ Año 2005: 16 artículos.
- ✓ Año 2006: 16 artículos.
- ✓ Año 2007: 25 artículos.
- ✓ Año 2008: 34 artículos.
- ✓ Año 2009: 29 artículos.
- ✓ Año 2010: 22 artículos.
- ✓ Año 2011: 53 artículos.
- ✓ Año 2012: 16 artículos.
- ✓ Año 2013: 19 artículos.
- ✓ Año 2014: 18 artículos.
- ✓ Año 2015: 18 artículos.
- ✓ Año 2016: 20 artículos.

Se observó que existen diferencias en la cantidad de artículos publicados por año, destacando mayor cantidad de artículos publicados en el año 2011, debido a los dos volúmenes de la edición especial.

3.4 Base de datos para análisis de datos de eventos recurrentes

Los artículos publicados en la revista Ciencia e Ingeniería, comprendidos en el periodo del 2000 al 2016, fueron escritos por 925 autores (entre ellos se cuentan los coautores). Sin embargo, para los eventos recurrentes son de interés únicamente los autores con dos o más publicaciones. Por lo tanto, los datos para el análisis de eventos recurrentes comprenden la cantidad de doscientos quince (215) autores.

Como la revista publica tres volúmenes por año en promedio cada cuatro meses, el identificador para describir el tiempo de publicación de los autores está expresado en meses. Así, el mes cuatro indica el primer volumen del año 2000, mientras que el mes dieciséis (16) indica el primer volumen del año 2001, hasta llegar al mes doscientos cuatro (204), indicando el tercer volumen del año 2016.

A continuación, se presenta la descripción de las variables utilizadas en la base de datos:

Variables	**Descripción**
An	Identificador de los autores
Mes	Identificador de las variables tiempo
Evento	Identificador del evento sobre la publicación del autor

Tabla 1. Descripción de las variables utilizadas en la base de datos.

Estas variables se crearon para la aplicación de la metodología donde:

- ✓ An: es el identificador que se le asignó a cada autor desde 1 hasta n = 216, para así facilitar el análisis en las gráficas resultantes.
- ✓ Mes: es el identificador para describir el tiempo de publicación de los autores, la cual también puede ser expresada en tiempo calendario.
- ✓ Evento: es una variable indicadora con dos variables posible: "*Star"* cuando se inició la observación del autor y "*End"*, cuando se deja de observar dicho autor.

3.5 Software utilizado para el análisis textual

Los resultados y salidas graficas (nubes de palabras) de los títulos, resúmenes y palabras claves se obtuvieron a través del paquete estadístico R Studio versión 3.3.1 (Copyright (C) 2016).

3.6 Software utilizado para el análisis de datos de eventos recurrentes

Los resultados y salidas gráficas de las estimaciones puntuales y por intervalo para la FAP, las estimaciones paramétricas y las gráficas de los modelos ajustados a los datos se obtuvieron a través del programa SPLIDA (Meeker y Escobar, 2004), el cual funciona en el paquete estadístico S-Plus, versión 6.1 (Insightful Corporation, 2001).

CAPÍTULO IV

RESULTADOS OBTENIDOS

4.1 Análisis textual de la revista Ciencia e Ingeniería mediante la nube de palabras

Los resultados y salidas gráficas (nubes de palabras) de los títulos, resúmenes y palabras claves se obtuvieron a través del paquete estadístico R versión 3.3.1 (Copyright (C) 2016). Previamente, se realizó la depuración de los artículos publicados para cada año en estudio (2000-2016) con respecto a los caracteres especiales, signos de puntuación, números, acentuación, entre otros.

Seguidamente, se presentaron los resultados (nubes de palabras) para Palabras Claves (PC), Resúmenes, Títulos y la nube de comparación de los ítems mencionados, con objeto de identificar los posibles temas abordados para cada año. En el proceso de investigación se pudo percibir que, para la compresión de la temática de las publicaciones de la revista para cada año, se necesitaron más que las palabras con mayores frecuencias mostradas en las nubes, siendo necesaria la generación de mayor cantidad de detalles, tales como la generación de etiquetas que agrupen a los autores por campos de investigación, mayor conocimiento de los procesos históricos en relación a la ciencia y tecnología, revisión exhaustiva de los artículos publicados, entre otros.

Vale destacar que la investigación se realizó con el apoyo de herramientas tecnológicas de análisis estadístico, además del programa Microsoft Excel (Ver cálculos en anexo pag 83). De esta manera, se describió un conjunto de palabras claves con base a su recurrencia en los artículos científicos publicados en la revista dentro del repositorio institucional Saber ULA durante el periodo 2010-2016. Se

asociaron las palabras de acuerdo a la familiaridad que guardaban con las diferentes disciplinas académicas de la Facultad de Ciencias en las carreras de Ingeniería, de igual manera que se asociaron las palabras con los campos de investigación y aplicación.

Las palabras de mayor dimensión o tamaño se encontraron directamente relacionadas con su aparición en la variedad de artículos, así como también su posicionamiento en el centro de la nube de palabras, de acuerdo a la identificación de palabras recurrentes en los siguientes aspectos: PC, Resúmenes, Título y Comparación. De igual manera, la disminución en el dimensionamiento de la palabra y su alejamiento del centro de la nube obedece a la menor cantidad de veces o recurrencia que aparece dicha palabra en los artículos científicos. A continuación, se describe la nube de palabras correspondiente a cada año estudiado:

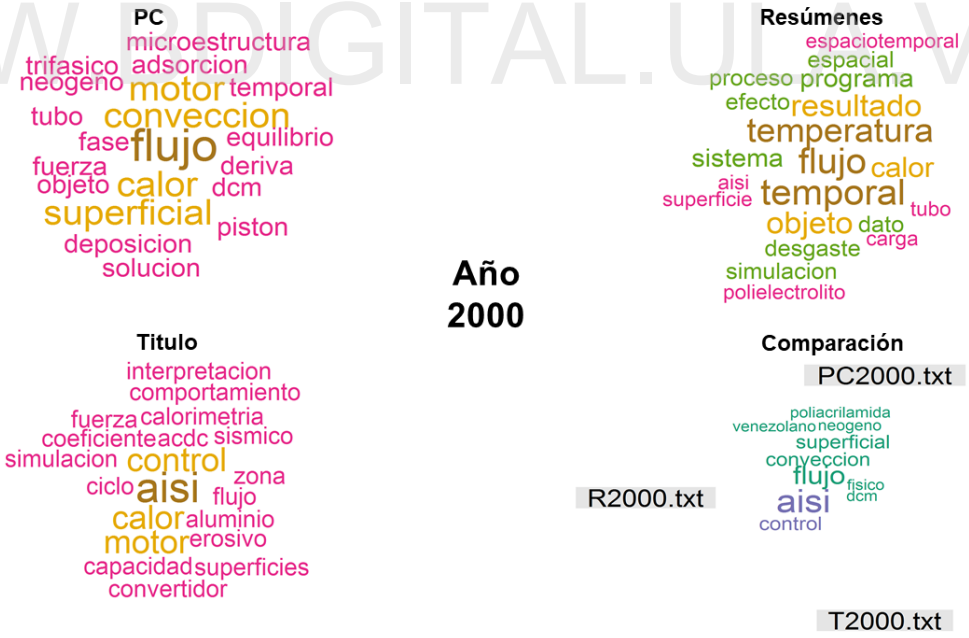

Figura 3. Nubes de palabras año 2000.

Se debe acotar que se observaron dos elementos que facilitaron la identificación de los temas: el primero, se debe a que los artículos, por lo general, relacionados con

las ciencias se fundamentaban principalmente en la investigación y, el segundo, por la parte de ingeniería, que dichos artículos se basaban generalmente en la aplicación, así que la temática estuvo accionada por la investigación y aplicación. Los temas fueron asociados a dichas disciplinas con palabras comunes entre ellas, por ejemplo, al mencionar la palabra "calor", la misma estaría relacionada con la disciplinas física y química si el tema es asociado a la investigación, mientras que, si esta palabra se fundamenta en la aplicación, estaría directamente vinculada con las ramas afines de la ingeniería (eléctrica, química, mecánica y sistemas).

De acuerdo a lo observado en la figura 3, se determinó que para el año 2000 los temas estaban enfocados en las ciencias aplicadas en áreas de ingeniería, al generarse mayor frecuencia de palabras como *temperatura*, *calor*, *motor*, *control*, *adsorción*, *solución*, *aluminio*, *aisi*, *Polieelectrolito*. De esta manera, se estableció un enfoque dirigido a las áreas de flujo de control de temperatura y motores.

Figura 4. Nubes de palabras año 2001

En la figura 4 se evidenció que, para el año 2001, los temas se encontraban enfocados en las áreas de sistemas, observándose mayor recurrencia a palabras como *experimental, procesos, cálculos, control, programa*. En este sentido, se encontraron vocablos referentes al diseño de sistemas de control.

Figura 5. Nubes de palabras año 2002.

Por su parte, en la figura 5 se determinó que los temas se enfocaron en áreas relacionadas a la mecánica para el año en estudio, al encontrarse mayor inclinación a palabras como *máquina, presión, válvula, motor, dosado, acero, inyección*, entre otras. Se corroboró que estos vocablos están enfocados al estudio sobre motores.

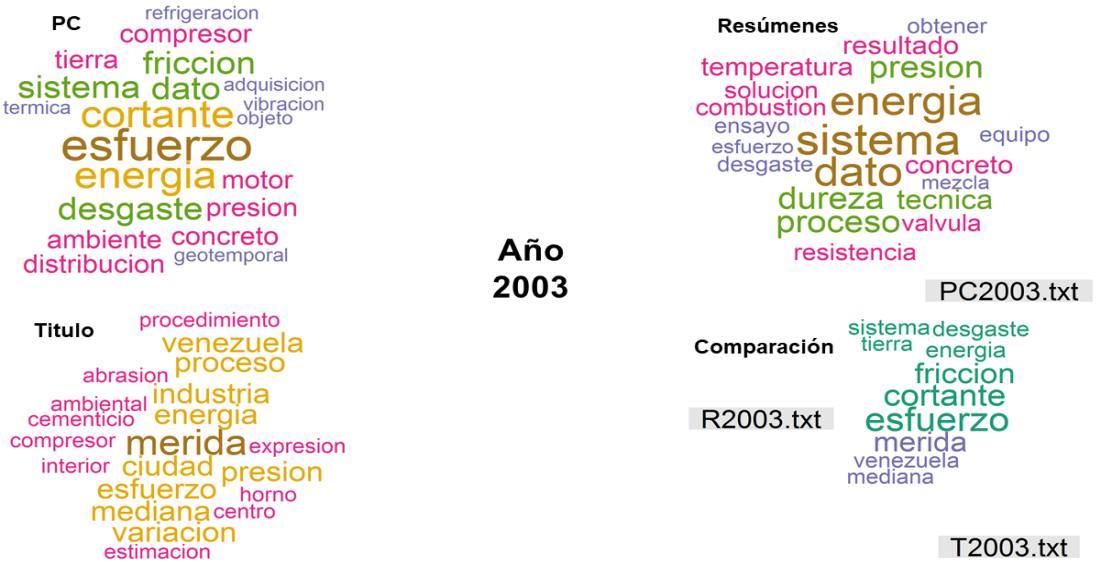

Figura 6. Nubes de palabras año 2003

Según lo observado en la figura 6, se puntualizó que los temas estaban orientados en áreas de civil durante el año 2003, debido a que se percibió gran recurrencia a palabras como *mezcla, concreto, resistencia, ensayo, presión, variación, compresor*, entre otras; estableciéndose que las mismas están enfocadas al estudio, diseño y pruebas realizadas al concreto.

Figura 7. Nubes de palabras año 2004

Se evidenció que para el año 2004, los temas estuvieron enfocados en áreas de sistemas, al aparecer palabras como *automatización, web, programa,* entre otras (ver figura 7). Se observó además que existía una tendencia a los métodos de control y automatización, suponiéndose que los vocablos estaban dirigidos al diseño de sistemas de control.

Figura 8. Nubes de palabras año 2005

Por su parte, en la figura 8 se estableció que para el año en estudio los temas estaban enfocados en áreas de sistema utilizando métodos de control y simulación, puesto que se apreció una significativa inclinación a palabras como *automatización, simulación, control, robusto, digital, programación*, lo que determinó que se direccionaban hacia la programación de sistemas de control lineales.

Figura 9. Nubes de palabras año 2006

De acuerdo a lo observado en la figura 9, se determinó que para el año 2006 los temas estaban enfocados en áreas de química, debido a que hubo mayor tendencia a palabras como *calor, presión, quemada, encendido, chispa, combustible, concentración, gasolina*. Se generó un vector enfocado a los procesos de combustión.

Figura 10. Nubes de palabras año 2007

En la figura 10 se puntualizó que, en el año 2007, los temas se centraron en áreas de sistemas, al encontrarse mayor recurrencia a palabras como *herramienta*,

programa, *simulación*, *interfaz*, *usuario*, *agente*, *software*, determinándose un enfoque en simulación de sistemas.

Figura 11. Nubes de palabras año 2008

Para el año 2008, los temas se encaminaron al área de sistema, puesto se percibió mayor uso de palabras como *sistemas*, *simulación*, *control*, *diseño*, *programa* (ver figura 11). Además, se notó una tendencia hacia métodos de control y simulación, lo que determinó que estaban dirigidas a sistemas de control.

Figura 12. Nubes de palabras año 2009

Por otra parte, se puntualizó que para el año 2009 los temas se orientaron al área de sistemas, al existir mayor inclinación a palabras como *dato, sistema, web, programa, simulación, software, modelado* (ver figura 12), estableciéndose de esta manera que se encontraban enfocadas en el modelado de sistemas.

Figura 13.Nubes de palabras año 2010

De acuerdo a la figura 13, para el año 2010 los temas se localizaron en el área de sistemas, debido a la frecuencia de palabras tales como *sistema*, *modelado*, *control*, *algoritmo*, *red*, determinándose que estaban orientadas a la aplicación de los sistemas.

Figura 14. Nubes de palabras año 2011

En la figura 14 para el año en consideración, los temas se determinaron dentro del área de sistemas, debido a la aparición de palabras como *simulación, sistema, herramienta, software, red, web, agente*. Así, se pudo evidenciar que las mismas se encontraban orientadas a la simulación de sistemas (de aprendizaje en ingeniería).

Figura 15. Nubes de palabras año 2012

De acuerdo a la figura 15, para el año 2012 los temas estaban enfocados en áreas de sistemas, debido a la presencia de palabras como *simulación*, *sistema*, *diseño*, *programa*, *software*, estableciéndose que estaban canalizadas hacia la simulación de sistemas.

Figura 16. Nubes de palabras año 2013

Según lo observado la figura 16, se determinó que para el año 2013 los temas destacaron aspectos vinculados al área de sistemas y métodos, al encontrarse gran inclinación a palabras como *Venezuela, sistemas, resistencia, método, emulsión y surfactante*, lo que estableció su orientación hacia sistemas y métodos en Venezuela.

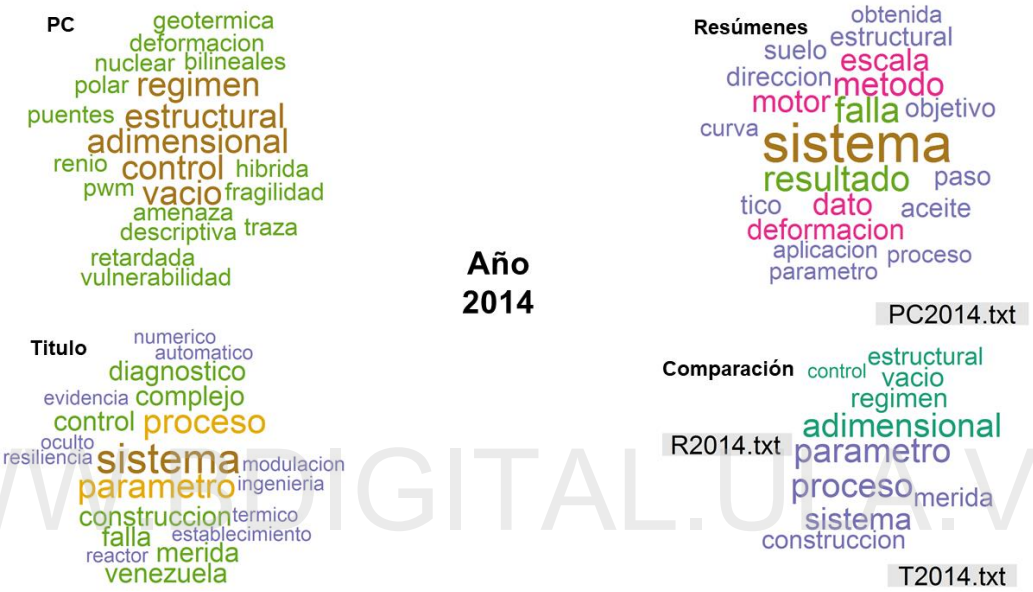

Figura 17. Nubes de palabras año 2014

En la figura 17 se evidenció que, en el año 2014, los temas desarrollados destacaron las áreas de sistemas y civil, encontrándose mayor frecuencia de palabras como *control, construcción, estructural, aplicación, puentes,* lo que permitió determinar que estos vocablos se encontraban relacionados con los sistemas de control en ingeniería civil.

Figura 18. Nubes de palabras año 2015

Para el año 2015, la temática estuvo enfocada en áreas de sistemas (figura 18), al encontrarse mayor inclinación a palabras como *estabilidad sistema, red, aplicación, proceso,* por lo que se intuye que las mismas se orientaron a la estabilidad de los sistemas.

Figura 19. Nubes de palabras año 2016

Por su parte, en la figura 19 se puntualizó que para el año 2016 los temas estaban enfocados en áreas de sistemas y control, al encontrarse mayor frecuencia a palabras como *sistemas, control, flujo, diseño, carreteras, construcción, edificación, procesos, método,* lo que establece un enfoque en sistemas de control de flujo.

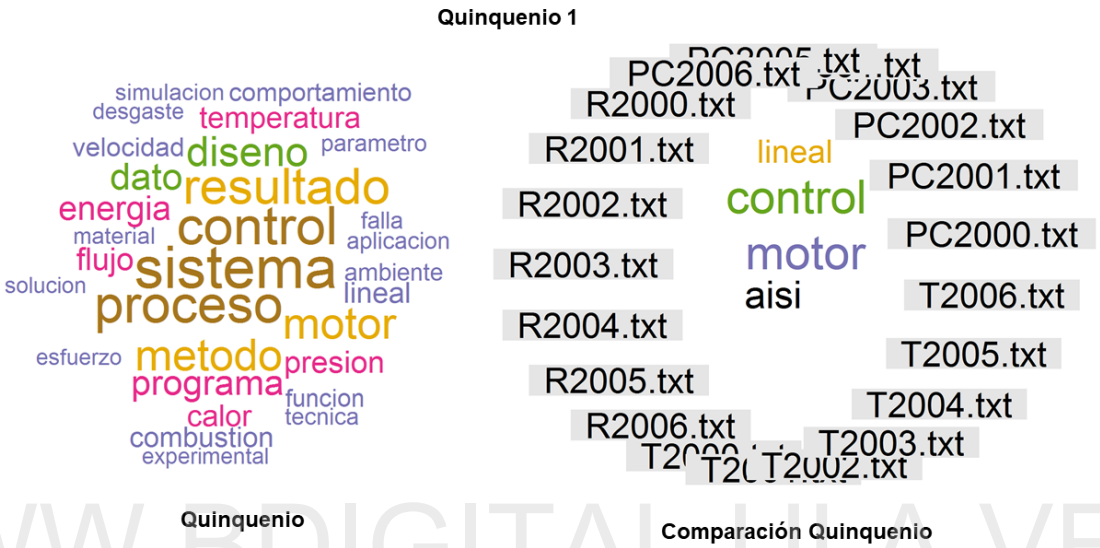

Figura 20. Quinquenio 1 y comparación

En la figura 20 se estableció que, para el primer quinquenio, los temas estaban enfocados en áreas de sistemas y mecánica. Se encontró mayor recurrencia a vocablos como *simulación, diseño, motor, control, programa*. De esta manera, se evidenció que en estas palabras predominó lo referente a sistema de control en motores.

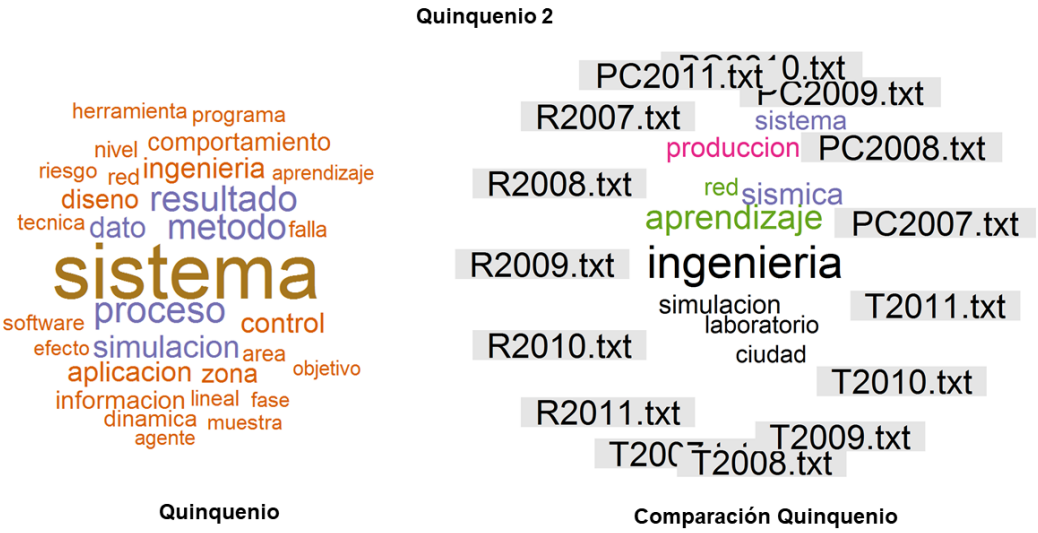

Figura 21. Quinquenio 2 y comparación

En la figura 21 se puntualizó que, para el segundo quinquenio, los temas estaban enfocados en áreas de sistemas, al aparecer palabras como *sistema, simulación, red, software, diseño, control, programa,* determinando que las mismas destacaron la parte de ingeniería de sistemas.

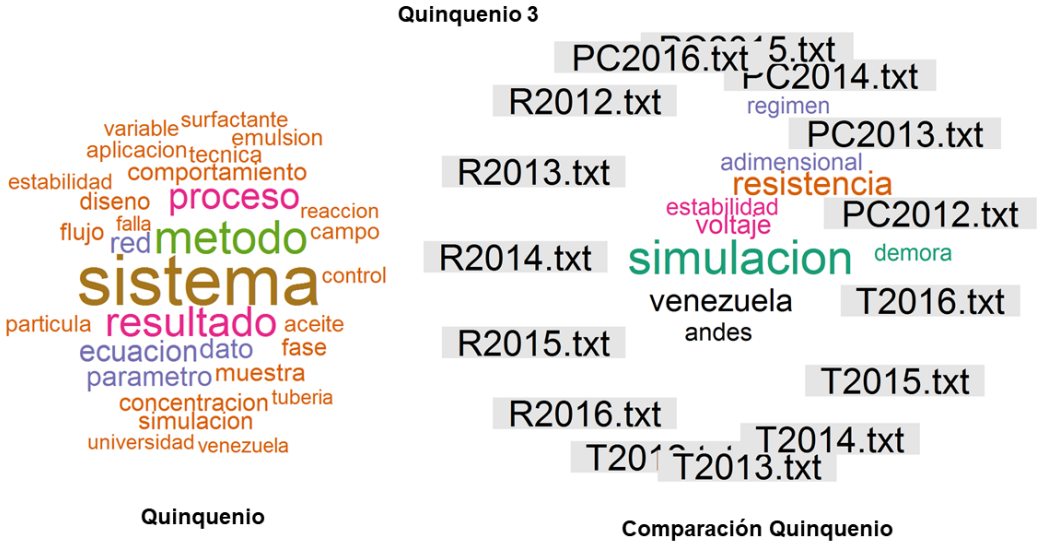

Figura 22. Quinquenio 3 y comparación

De acuerdo a lo observado en la figura 22 se determinó que, para el tercer quinquenio, los temas estaban enfocados en áreas de sistemas. Se encontró mayor recurrencia en palabras como *sistema, simulación, red, diseño*, evidenciando así que se orientaron a la simulación de sistemas.

Con respecto a los tres quinquenios, se percibió una tendencia al decrecimiento de la ingeniería tradicional (eléctrica, química, civil), lo que evidenció un crecimiento a las ingenierías emergentes como ingeniería de sistemas (simulación, control y computación).

4.2 Análisis de los eventos recurrentes de los autores de la revista de Ciencia e Ingeniería

Los análisis realizados en el presente trabajo fueron obtenidos a través del paquete estadístico S-PLUS 6.0, usando el módulo SPLIDA, el cual pueden ser utilizado gratuitamente para fines académicos. En este caso, se hizo bajo la autorización otorgada por Rafael Borges la licencia de la Facultad de Ciencias Estadísticas.

Previamente, se realizó un filtro con respecto a las publicaciones de los autores donde solo se tomaron en cuenta aquellos que realizaron dos o más publicaciones. A continuación, se transformó la variable año a meses para así determinar las publicaciones en periodos de mes.

Seguidamente, se presentaron los resultados para las estimaciones puntuales, y de intervalo de la FAP para la recurrencia de los autores, para proceder a elegir el modelo apropiado y de acuerdo a ello, se realizó el pronóstico de las futuras publicaciones.

Análisis de recurrencia de los autores

Este análisis se inició con una descripción general, exploratoria y descriptiva de los datos. Para los eventos recurrentes, se representaron a menudo mediante el gráfico de evento-tiempo, según la metodología de Nelson (2003).

Los gráficos evento-tiempo son una representación de las publicaciones de los autores en una unidad de tiempo, donde cada autor se observa mediante una línea recta que tiene como longitud la edad censurada o tiempo máximo de observación, en la que se inicia en la edad cero (0) o en la edad del autor en el momento en que se inició su observación (en este caso inició en 8). Cada publicación se denota como una X en la recta.

En la tabla 2 se presenta una descripción de la base de datos recurrentes de la publicación de los autores:

Summary of recu data data
Number of rows in data matrix= 677
Number of unique units or group IDs in the recurrence data object = 215
Number of observation windows = 215
Number of recurrences = 462
Sum of costs/counts: 462
Number of unique recurrence times = 54
Time units: MES
Recurrence time minimum: 8 MES
Recurrence time maximum: 200 MES
Endpoint time maximum: 204 MES

Tabla 2. Base de datos recurrentes de la publicación de los autores.

Se puede observar que se tienen seiscientos setenta y siete (677) datos, de los cuales cuatrocientos sesenta y dos (462) corresponden a las recurrencias, debido a que doscientos quince (215) datos pertenecen a las censuras a derecha u observaciones de sus autores. Se observa también que la unidad de medida es Mes, y que la primera publicación se observó a los ocho meses de edad en alguno de los autores.

Por otra parte, la figura 23 ilustra la distribución de los eventos recurrentes de los autores desde su primera publicación; cada "X" representa una recurrencia.

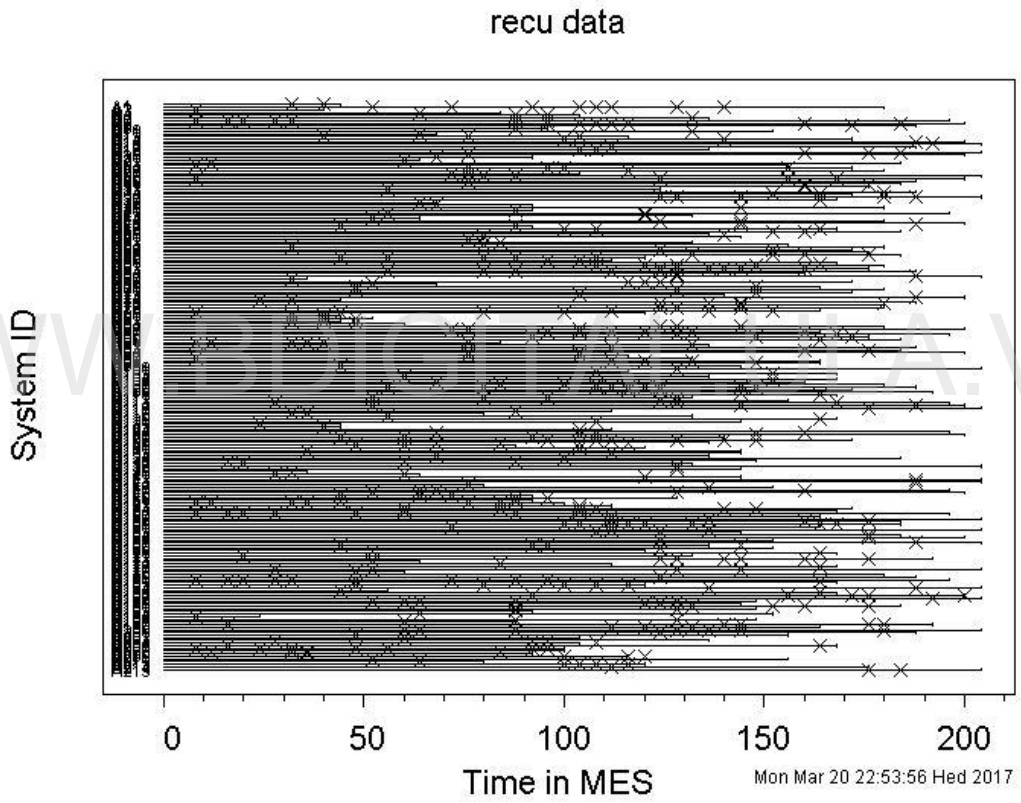

Figura 23. Gráfico de eventos de publicación de los autores recurrente.

Además, se puede apreciar que los autores tienen diferentes edades, y que solo se presentaron censuras a la derecha. También, se puede observar que se trata de autores que publican frecuentemente. La distribución de los eventos de publicación en el tiempo que se muestra en esta figura, indica que la tasa de recurrencia m(t) de los doscientos dieciséis (216) autores, no es constante en el tiempo.

Los análisis no-paramétricos y paramétricos, que se presentan en los siguientes apartados, ayudan a confirmar o descartar esta sospecha.

Análisis no paramétrico

La estimación puntual y por intervalo (I.C. 95%) de la FAP, por el método no-paramétrico, se presenta en la figura 24.

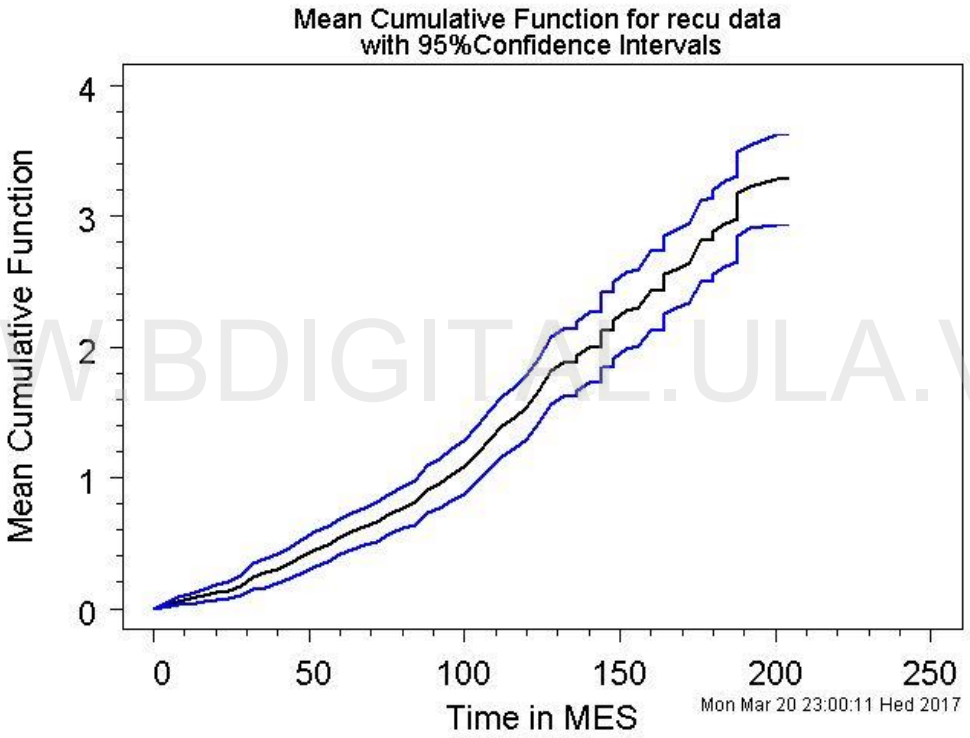

Figura 24. FAP estimada en SPLIDA para las publicaciones de los autores con intervalos de confianza del 95%

Se puede apreciar que la tasa de recurrencia no es constante, pues la tendencia de la FAP no sigue una línea recta. Igualmente, se observa que la recurrencia de publicación es menor en los primeros meses. La gráfica de los FAP presenta una tendencia creciente para la tasa de recurrencia donde se advierte que el promedio de recurrencias acumuladas para los autores, durante un poco más de doscientos (200)

meses de publicación del autor, se encuentra entre los tres y cuatro artículos, aproximadamente.

Análisis paramétrico

A continuación, se presenta el procedimiento que se siguió en el análisis paramétrico. Se comenzó haciendo la validación de supuestos y, con base en los resultados, se estableció el modelo más apropiado. A partir de ese punto, fue posible establecer el procedimiento para realizar las respectivas predicciones del número de eventos futuros en un intervalo de tiempo de interés, tomando en consideración que las mismas dependían del modelo seleccionado.

Prueba para la tendencia de un proceso de Poisson homogéneo (PPH) vs un proceso de Poisson no homogéneo (PPNH)

Para evaluar PPH contra PPNH, se estimaron los parámetros β y γ_1 de los modelos PPNH Regla de la Potencia y Loglineal, respectivamente. Según la teoría, si β = 1 o γ_1 = 0, entonces el proceso equivale a un modelo PPH.

Estimaciones de los parámetros del Proceso de Poisson No Homogéneo (PPNH) con el modelo de Regla de la Potencia

En la tabla 3 se observan las estimaciones por máxima verosimilitud de los parámetros β y γ_1 del modelo PPNH utilizando la Regla de la Potencia obtenidos en el módulo SPLIDA.

recu data

NHPP Power Rule model

Response units: MES

Maximum likelihood estimation results

Appears to have converged; relative function convergence

Log likelihood at maximum point: -2377

	MLE	Std.Err.	95% Lower	95% Upper
eta	92.299	3.62508	85.194	99.404
beta	1.527	0.06612	1.397	1.656

Tabla 3. Estimación por máxima verosimilitud de los parámetros β y γ_1 del modelo Regla de la Potencia.

En la tabla presentada anteriormente se observa, además, que los intervalos de confianza para el parámetro β es mayor que 1 (el intervalo de confianza al 95% para β es: [1.397, 1.656]), con el modelo de PPNH Regla de Potencia para la tasa de recurrencia de publicaciones de los autores. Por lo tanto, se descarta PPH en la prueba de HPP Vs. NHPP con el modelo de regla de la potencia.

Estimaciones de los parámetros del proceso de Poisson no homogéneo (PPNH) con el modelo log-lineal

En la tabla 4 se presentan las estimaciones por máxima verosimilitud de los parámetros $\gamma 0$ y $\gamma 1$ del modelo PPNH utilizando la Regla Log-Lineal, obtenidas en el módulo SPLIDA.

recu data

NHPP Log Linear model

Response units: MES

Maximum likelihood estimation results

Appears to have converged; relative function convergence

Log likelihood at maximum point: -2385

	MLE	Std.Err.	95% Lower	95% Upper
gamma0	-4.886621	0.1004883	-5.083574	-4.689667
gamma1	0.007144	0.0008788	0.005421	0.008866

Tabla 4. . Estimación por máxima verosimilitud de los parámetros γ0 y γ1 del modelo PPNH Loglineal

Se puede visualizar que la estimación de $\gamma 1$ es mayor que 0 (el intervalo de confianza del 95% para $\gamma 1$ es: [0.005421, 0.008866]); por tanto, se descarta un PPH en la prueba de HPP vs. NHPP con el modelo log-lineal.

Prueba para evaluar proceso de renovación

Debido a que con los modelos PPNH de Regla de Potencia y Log-Lineal no resultó un PPH para las tasas de recurrencia de publicaciones de los autores, fue necesario probar si se trata de un proceso de renovación o no. Para ello, se realizó una prueba de bondad de ajuste de los tiempos entre recurrencias, para ver si estos se distribuyen en forma exponencial. En la tabla 5 se presenta el resultado de varios estadísticos de bondad de ajuste calculados en R Studio.

```
        One-sample Kolmogorov-Smirnov test

data:  Tiempo
D = 0.24732, p-value < 2.2e-16
alternative hypothesis: two-sided


        Anderson-Darling test of goodness-of-fit
        Null hypothesis: exponential distribution
        with parameter rate = 0.00860228716645489

data:  Tiempo
An = 82.294, p-value = 8.863e-07


      Cramer-von Mises test of goodness-of-fit
      Null hypothesis: exponential distribution
      with parameter rate = 0.00860228716645489

data:  Tiempo
omega2 = 15.974, p-value = 5.452e-09
```


Tabla 5. Resultados de varios estadísticos de bondad de ajuste calculados en R Studio.

En la tabla 5 se puede apreciar que, a un nivel de significancia incluso más pequeño que $\alpha = 0.05$, los tiempos entre recurrencias de las publicaciones de los autores no se distribuyen exponencialmente. Por tanto, el proceso de recurrencia tampoco corresponde a un proceso de renovación. Luego de las tres pruebas realizadas, se determinó que los datos de recurrencia correspondían a un PPNH. En consecuencia, el siguiente paso fue elegir el modelo PPNH que mejor se ajustara a los datos, el modelo de Regla de Potencia o el Log-Lineal.

Elección del modelo paramétrico del proceso de Poisson no homogéneo (PPNH)

La elección del modelo paramétrico PPNH que mejor se ajustara a los datos de recurrencia, se realizó de acuerdo con las estimaciones obtenidas junto con los gráficos. De las figuras 25 y 26 se pueden observar que los modelos PPNH Regla de Potencia y Log-Lineal se ajustaban muy bien a los datos de publicación de los autores, lo que dificulta elegir por medio de la gráfica el modelo que mejor se adapta. Para ello, se observa de las tablas 3 y 4 el valor de la función de log-verosimilitud para el modelo de Regla de Potencia (-2377), y el valor de la función Log-verosimilitud para el modelo Log-Lineal (-2387), determinando que el modelo Log-Lineal sugiere un mejor ajuste por ser mayor al modelo de Regla de Potencia.

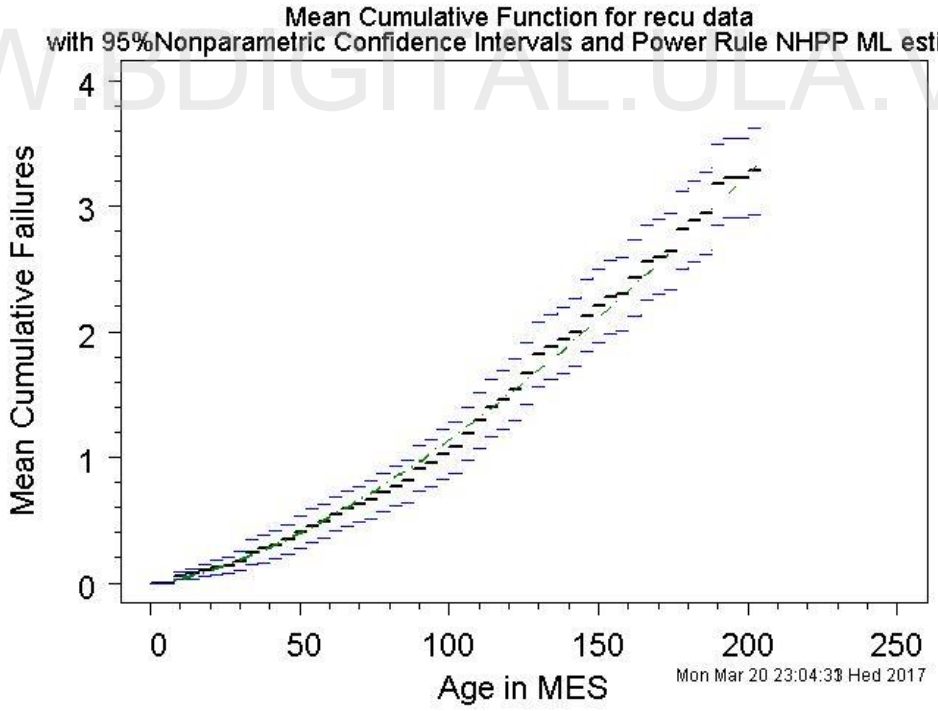

Figura 25. Regla de la potencia.

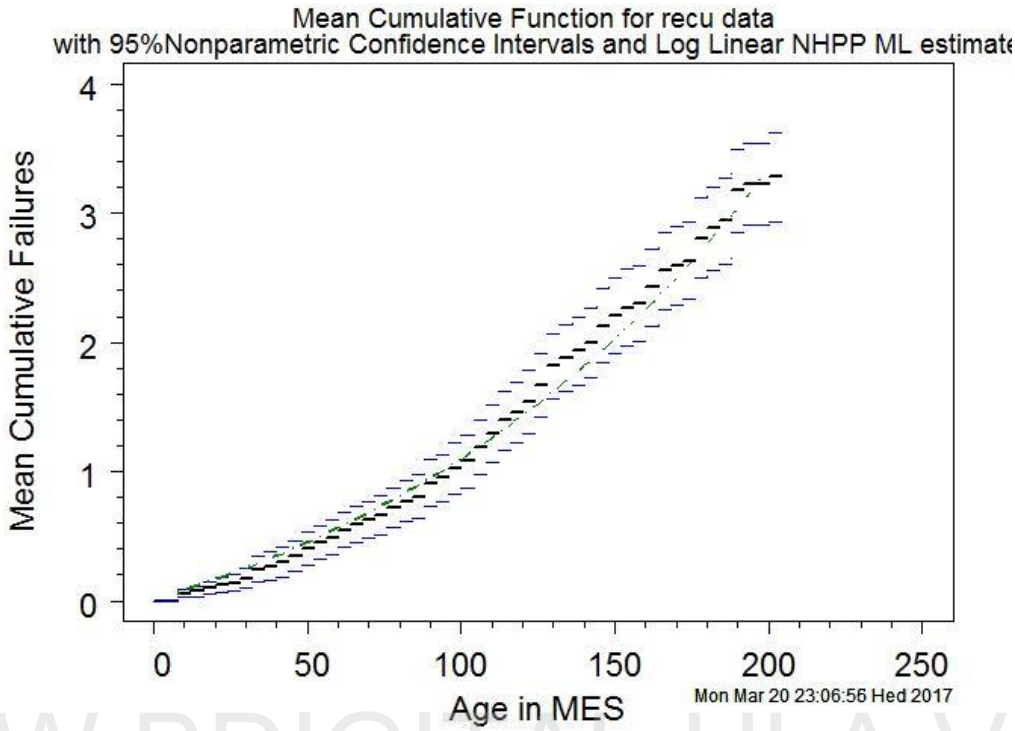

Figura 26. Log-Lineal

Si se selecciona el modelo de PPNH, se tendría la siguiente tasa de recurrencia: $m(t{:}\,\gamma_0, \gamma_1) = \exp(\gamma_0 + \gamma_1 t)$.

$$m(t{:}\,-4.886221, 0.007144) = \exp(-4.86621 + 0.007144t)$$

Considerando t en meses.

Pronóstico de recurrencia futura con un proceso de Poisson no homogéneo

Log-Lineal

$$\int_a^b m(t,\hat{\theta})dt = \frac{\exp(\hat{\gamma}_0)}{\hat{\gamma}_1}[\exp(\hat{\gamma}_1 b) - \exp(\hat{\gamma}_1 a)]$$

$$\int_{210}^{222} m(t,\hat{\theta})dt = \frac{\exp(-4.886221)}{0.007144}[\exp(0.007144 * 222) - \exp(-4.886221 * 210)]$$

$$\int_{210}^{222} m(t,\hat{\theta})dt \cong 5.161511$$

Se estima que la próxima publicación de la revista se realice aproximadamente en 5 meses, según los datos obtenidos de los pasos anteriores.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

Se logró analizar textualmente los títulos, resúmenes y palabras claves de la revista Ciencias e Ingeniería en el periodo 2000 – 2016, para ello, se realizó una depuración de los artículos, con respecto a los caracteres especiales, signos de puntuación, números, acentuaciones, entre otros. A fin de desarrollar este objetivo se realizó lo siguiente:

Se identificaron los principales temas publicados en la revista Ciencias e Ingeniería de la Universidad de Los Andes, a través de las nubes de palabras que se obtuvieron con base a los resultados del software R, mediante un algoritmo para realizar la minería de texto, determinándose las palabras usadas con mayor frecuencia dentro de los artículos de los títulos, resúmenes y palabras claves por cada año en estudio. La identificación de estos temas se realizó de manera muy subjetiva por falta de información. Adicionalmente, se realizó una búsqueda usando el programa Microsoft Excel para sustentar la identificación de los temas, creándose dos libros, el primero llamado (2000) que contenía los resúmenes, autores y títulos. El segundo libro se denominó (2000a), mediante el cual se realizó la búsqueda de las palabras claves de la nube de resúmenes, obteniendo como resultado el autor y título del artículo que atendía a la palabra buscada para el periodo del mismo año.

Se logró determinar la recurrencia de los autores, mediante la utilización del software estadístico S-Plus y el módulo Splida, utilizando la metodología de Meeker y Escobar (1998) y Nelson (2003), el cual consistió en la identificación de un modelo que explica el comportamiento de la Función Acumulada Promedio (FAP), mediante

el algoritmo de decisión propuesto por estos autores. Este estudio se hizo mediante dos subtipos de análisis: el no paramétrico y el paramétrico. La recurrencia recayó en 215 de los 716 autores que realizaron sus publicaciones. Obteniendo que el mejor modelo que se adaptó fue el modelo de Proceso de Poisson No Homogéneo (PPNH) Log-Lineal, descartando el Proceso de Poisson Homogéneo (PPH) y el Proceso de Renovación, por último, se logró determinar que la futura publicación se realizaría en cinco (5) meses aproximadamente.

Vale destacar que se elaboró la nube de palabras, mediante el análisis de texto por año y quinquenio, empleando el software R y el uso de paquetes como ***Wordcloud, Tm,*** entre otros. Las nubes que se realizaron se obtuvieron para cada año, constando de 4 nubes, una para Palabras Claves (PC), Resúmenes, Títulos, comparación y las nubes de los quinquenios.

5.2 Recomendaciones

- ✓ Utilizar la información que se obtuvo de manera empírica para que se formule modelo que permita verificar la relación y la aparición de las palabras existentes a partir de las nubes de palabras creadas.
- ✓ Incentivar a la comunidad de investigación a que continúen el análisis de las nubes de palabras creadas en el presente proyecto.
- ✓ Enmarcar los artículos de los autores por campos de investigación de las diferentes carreras.
- ✓ Promover ampliamente la inclusión de trabajos de carácter completamente prácticos (El Hospital Universitario de Los Andes, la cárcel, los servicios públicos, el transporte, etc…)
- ✓ Incentivar a los autores con premios o reconocimientos para aumentar las publicaciones por cada volumen que se genera al año.

REFERENCIAS

Arias, F. (2012). **El proyecto de investigación. Introducción a la metodología científica.** (7ª ed.). Caracas, Venezuela: Editorial Episteme.

Baena, A. y Salazar, J. (2006). **Análisis de recurrencia de falla aplicado a una empacadora de líquidos en la Cooperativa Lechera Colanta**. Medellín, Colombia. Universidad Nacional de Colombia. Estadística Industrial XVI Simposio de Estadística.

Contreras, O. (2011). **La comunidad académica y sus medios: la consolidación de una revista de ciencias sociales**. [Artículo en línea]. Disponible: http://www.scielo.org.mx/scielo.php [Consulta: abril 22, 2017].

Cui, W., Wu, Y., Liu, S., Wei, F. Zhou, M. y Qu, H. (2010). **Context-Preserving Dynamic Word Cloud Visualization**. IEEE Computer Graphics and Applications, 30(6): 42-53.

Easy Guides. (2017). **Text mining and word cloud fundamentals in R: 5 simple steps you should know**. [Documento en línea]. Disponible: https://www.r-bloggers.com/. [Consulta: marzo, 26, 2017].

Eíto, R. y Senso, J. (2004). **Minería textual**. [Revista en línea]. Disponible: http://www.elprofesionaldelainformacion.com/. [Consulta: abril, 6, 2017].

Galili, T. (2016). **Intro to Text Analysis with R**. [Documento en línea]. Disponible: https://www.r-bloggers.com/. [Consulta: marzo, 25, 2017].

Gómez, A. (2007). **Análisis de las Fallas Recurrentes de las Celdas de Reducción Electrolítica de la Corporación Venezolana de Guayana VENALUM en el**

período (1997-2004). Trabajo de grado para optar al título de Licenciado en Estadística de la Universidad de Los Andes, Mérida-Venezuela.

Gómez, R. (2013). **Etiquetar en la web social**. Barcelona, España: Editorial UOC.

Guallar, J., Orduña, E. y Olea, I. (2014). **Anuario ThinkEPI 2014: Análisis de tendencias en información y documentación**. (8ª ed.). España: Editorial UOC.

Lopera, C. y Manotas, E. (2011). **Aplicación del análisis de datos recurrentes sobre interruptores FL245 en Interconexión Eléctrica S.A.** Medellín, Colombia. Revista Colombiana de Estadística. Volumen 34, no. 2, pp. 249 a 266.

Meeker, W. y Escobar, L. (1998). **Statistical Methods for Reliability Data**. Estados Unidos de América: John Wiley & Sons, Inc.

Nelson, W. (2003). **Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications**. Nueva York, Estados Unidos de América: American Statistical Association and Society for Industrial and Applied Mathematics.

Ramsden, A. y Bate, A. (2008). **Using Word Clouds in Teaching and Learning**. [Documento en línea]. Disponible: http://opus.bath.ac.uk/. [Consulta: marzo, 7, 2017].

Ramírez, T. (2010). **Cómo hacer un proyecto de investigación.** (3ª ed). Venezuela: Editorial Panapo.

Repositorio Institucional de la Universidad de Los Andes. [Página web en línea]. Disponible: http://www.saber.ula.ve/. [Consulta: febrero, 19, 2017].

Revista Ciencia e Ingeniería de la Universidad de Los Andes. [Página web en línea]. Disponible: http://erevistas.saber.ula.ve/cienciaeingenieria. [Consulta: febrero, 22, 2017].

Rodríguez, J. y Sulé, A. (2008). **DSpace: un manual específico para gestores de la información y la documentación**. [Documento en línea]. Disponible: http://bid.ub.edu/. [Consulta: febrero, 18, 2017].

Tello, A., Méndez, F., García, J. y Carrillo, E. (s/f). **Repositorios educativos digitales sobre computación Grid**. [Artículo en línea]. Disponible: http://rlcu.org.ar/. [Consulta: febrero, 3, 2017].

The R Project for Statistical Computing. [Página web en línea]. Disponible: www.R-project.org. [Consulta: marzo, 2, 2017].

Willinsky, J., Stranack, K., Smecher, A. y MacGregor, J. (2010). **Open Journal Systems: A Complete Guide to Online Publishing**. [Documento en línea]. Disponible: http://pkp.sfu.ca/ojs. [Consulta: febrero, 18, 2017].

Wu, Y., Provan, T., Wei, F., Shixia, L. y Ma, K. (2011). **Semantic-Preserving Word Clouds by Seam Carving**. IEEE Symposium on Visualization. Volumen 30, Número 3.

Zambrano, M. (2012). **Curso virtual de estadística en Moodle con apoyo de R para una universidad colombiana**. [Documento en línea]. Trabajo de grado para optar el grado de Master en Estadística Aplicada de la Universidad de Granada. Granada, España. Disponible: http://masteres.ugr.es/. [Consulta: febrero, 6, 2017]

ANEXOS

Anexo 1. Prueba de normalidad, realizadas en SPSS

Parámetros de distribución estimados

		MES
Distribución normal	Ubicación	116,25
	Escala	52,371
Los casos están sin ponderar.		

MES

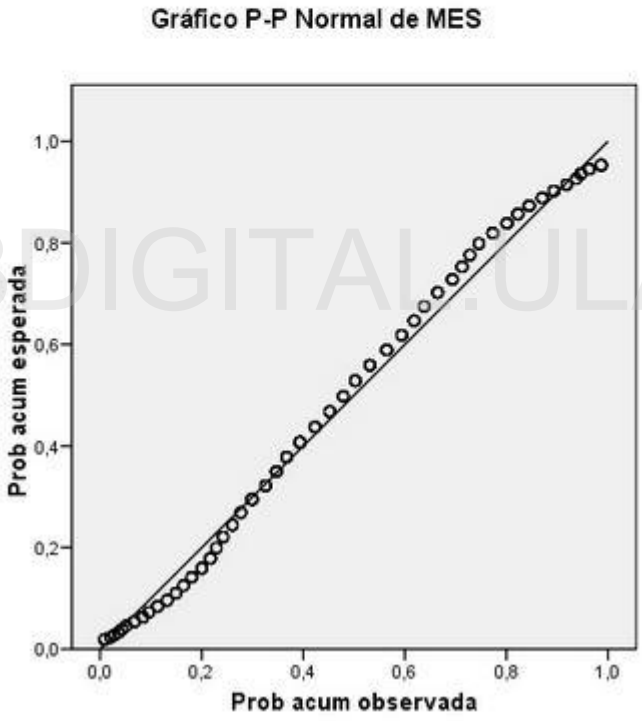

Figura 27. Test de Normalidad en SPSS.

Anexo 2. Utilización del Programa Microsoft Excel

Utilización del Programa Microsoft Excel, para ayudar a fundamentar la identificación de los temas, en base a la búsqueda de palabras claves extraídas de la nube de palabras de Resúmenes.

Año 2001	Resumen	Autores	Titulos
	1. En el pasado se ha estudiado el resalto hidráulico circular libre producido por la desviación axisimétrica de un chorro circular que incide normalmente sobre una superficie plana y lisa. Se ha encontrado la relación de profundidades secuentes, en función del numero de Froude del flujo deflectado y de la altura de una pantalla de retención que determina la formación del resalto. También se conocen las relaciones funcionales que permiten obtener la disipación de energía. En este trabajo, se presenta un estudio teórico y experimental sobre un disipador hidráulico circular imperfecto que se genera para alturas elevadas de la pantalla de retención y pequeños radios de ubicación, con los cuales el resalto pierde las características de flujo uniforme en las secciones de aguas arriba y de aguas abajo. Se desarrolla una ecuación general que permite predecir la profundidad adimensional de sumersión en función de cinco parámetros. También se obtiene una ecuación para calcular la disipación de energía relativa q volumen de hidrocarburo extraído y la ubicación de los máximos desplazamientos a lo largo del tiempo, cuando se varían los módulos de elasticidad y la superficie inicial de cedencia de los materiales involucrados. Finalmente, estos estudios demuestran que este tipo de análisis puede ser de gran provecho en la industria petrolera para estudiar el comportamiento de pozos y diseñar los proyec	1. Alix Teresa Moncada Moreno, Julián Aguirre Pe, María Luisa Olivero	1. Disipador radial imperfecto

Figura 28. Libro llamado 2001, que contiene Resumen, Autores y Títulos

Flujo			Año 2001	
Resultado		Palabra Cl	Autor	Nombre del Articulo
Sistema		Flujo	1. Alix Teresa Moncada Moreno, Julián Aguirre Pe, María Luisa Olivero	1. Disipador radial imperfecto
Metodo			NA	NA
Ecuación			NA	NA
Linea			NA	NA
Solución			NA	NA
Diseño			NA	NA
Programa			NA	NA
Dato			NA	NA
Control			9. H. Juárez S., P. Peykov, T. Diaz B., A. Vivaldo V., G. Romero P.	9. La estructura MOS como sensor de flujo de fotones
Temperatura			10. Julián Aguirre Pe, María Luisa Olivero, Alix Teresa Moncada Moreno	10. Efecto del número de Froude densimétrico en el transporte de sedimentos
Intercambiador			11. Luis M. Sarache B.	11. Analisis multifactorial y optimizacion de Normal Butano (NC4) en un proces
Experimental			NA	NA
Velocidad			NA	NA
Variable			NA	NA
Transmisión			15. L. Patiño, H. Espinoza, O. Velásquez	15. Estudio numérico de la transferencia de calor en la cerveza en el proceso de
Selección			NA	NA
Transporte			NA	NA
Parámetro			NA	NA
			NA	NA

Figura 29. Libro llamado 2001a, que arroja como resultados los Autores y Títulos al hacer la búsqueda de la Palabra Clave

	Año 2006	Resumen	Autores	Titulos
		1. El yogur es uno de los productos lácteos mas consumidos. Para su producción es necesario aumentar la concentración de sólidos en la leche y existen varios métodos para lograrlo: evaporación, adición de proteína láctea y concentración utilizando membranas. En este estudio se compararon dos métodos, la adición de proteína láctea y la concentración con membranas de ultrafiltración. Con estos métodos se obtuvieron yogures de diferentes concentraciones de sólidos: 11,5, 14,8 y 20%. Los yogures obtenidos fueron sometidos a pruebas de viscosidad, sinéresis y penetración. También fueron probados por un panel de jueces no entrenados. Los resultados indican que los yogures con mejores características son aquellos obtenidos por medio de la ultrafiltración, siendo el mejor el que tiene un 20% de sólidos.	1. S. Tolosa, Johnny J. Bullón T., Antonio Luis Cárdenas Rodulfo, Carmen Borregales	1. Producción de yogur utilizando membranas cerámicas para incrementar el porcentaje de sólidos en la leche
		2. Las redes de autómatas estocásticamente acoplados constituyen un ejemplo sencillo de sistemas auto-organizados: bajo ciertas condiciones aparece sincronización espontánea en este tipo de sistemas. Una motivación para estudiar la aparición de la sincronización espontánea es su posible aplicación al problema de la sincronización de símbolos en sistemas de comunicaciones. En este trabajo se emplea un algoritmo genético para optimizar la operación síncrona espontánea en una red de autómatas de Winfree estocásticamente acoplados. En el caso estudiado se desprecian los posibles retardos asociados al acople entre elementos así como la influencia del ruido y de las interferencias.	2. José Manuel Albornoz M.	2. Optimización de la sincronización de una red de autómatas de Winfree acoplados estocásticamente por un algoritmo genético

Figura 30. Libro llamado 2006, que contiene Resumen, Autores y Títulos

Sistema		Año 2006	
Método	Palabra Clave	Autor	Nombre del Artículo
Combustión	Sistema	1. S. Tolosa, Johnny J. Bullón T., Antonio Luis Cárdenas Rodulfo, Carmen Borregales	1. Producción de yogur utilizando membranas cerámicas para incrementar el porcentaje de s
Proceso		2. José Manuel Albornoz M.	2. Optimización de la sincronización de una red de autómatas de Winfree acoplados estocás
Resultado		3. Manuel A. Rodríguez H., Elsa Mora Gallardo, Christian Cavé	3. Síntesis de voz en el dialecto venezolano por medio de la concatenación de difonos
Diseño		4. María de Jesus Martín Valera, Jesús Omar Araque Maldonado, Carlos Gilberto Villamar Linares	4. Características del proceso de combustión en un motor CFR de gasolina
Dato		5. T. Mochizuki, Leonardo Rennola Alarcón	5. Programa multimedia para la enseñanza de combustión y hornos de proceso
Nivel		6. Jean Francois Dulhoste Vivien, Carlos Javier Jerez Rico, D. Georges, G. Besançon	6. Técnica de control de nivel de agua no lineal para un sistema de canales abiertos de dos t
Mezcla		NA	NA
Aplicación		8. Pedro Cerda, Pedro José Rivero Rivero, William Lobo Quintero	8. Evaluación del factor de respuesta R en estructuras de concreto armado con pisos blandos
Piso		NA	NA
Calor		10. José L. Paredes Q., Juan Marcos Ramírez Rondón, Giorgio Alessandro Bianchi Donayre	10. Métodos robustos de normalización de microarreglos de ADNc basados en la mediana
Aleta		NA	NA
Motor		NA	NA
Modulo		13. Emilio López	13. Efecto del corrugado de las aletas helicoidales sobre el flujo de calor
Presión		NA	NA
Control		NA	NA
Variación		NA	NA
Combustible		NA	
Comportamiento			

Figura 31. Libro llamado 2006a, que arroja como Resultados los Autores y Títulos al hacer la búsqueda de la Palabra Clave