TRABAJO FINAL PRESENTADO EN OPCIÓN AL TÍTULO DE MÁSTER EN INFORMÁTICA AVANZADA.

PROCEDIMIENTO PARA LA ESTRUCTURACIÓN Y ALMACENAMIENTO DE DOCUMENTOS EN EL SISTEMA DE RECUPERACIÓN DE INFORMACIÓN ORIÓN

Autora: Ing. Yennifer Delgado Mesa

Tutores:

DrC. Juan Pedro Febles Rodríguez

DrC. Arturo Orellana García

La Habana

2018

Declaración de autoría

Declaro por este medio que yo: Yennifer Delgado Mesa, con carnet de identidad 92101332095, soy la única autora del trabajo final de maestría: **"Procedimiento para la estructuración y almacenamiento de documentos en el Sistema de Recuperación de Información Orión"**, desarrollada como parte de la Maestría en Informática Avanzada y autorizo a la Universidad de las Ciencias Informáticas a hacer uso de la misma en su beneficio, así como los derechos patrimoniales con carácter exclusivo.

Y para que así conste, firmo la presente declaración jurada de autoría en La Habana a los_____ días del mes de_____del año 2018.

_____

Ing. Yennifer Delgado Mesa

Autor

_____ _____

DrC. Juan Pedro Feble Rodríguez

Tutor

DrC. Arturo Orellana García

Tutor

A mi familia que tanto me ayudó a lograr esta meta.

A mi hijo y mi esposo que son los dos motivos para día a día salir adelante.

Resumen

Con el objetivo de recuperar y visualizar la información alojada en la web cubana se desarrolló en el país una herramienta de recuperación y visualización llamada Orión. Como deficiencias fundamentales de este buscador se encuentran la definición de sus metadatos para estructurar de forma correcta los archivos digitales que se rastrean, la carencia de un espacio para el rastreo y almacenamiento de videos y el diseño de la arquitectura base posee deficiencias que no permiten ejecutar con calidad los mecanismos de rastreo y la indexación provocando sobrecarga en lo servidores. Para darle solución a la problemática anterior se diseña un procedimiento que se encarga de estructurar y almacenar los documentos para el sistema de recuperación de información Orión, combinando técnicas de rastreo e indexación, con el objetivo de mejorar la eficacia del almacenamiento y la calidad de los resultados brindados a los usuarios. Los aportes fundamentales de la investigación se centran en la estructuración de los documentos a partir de la definición de sus metadatos y la aplicación del modelo vectorial, la definición de las arquitecturas de hardware para el despliegue de los componentes de rastreo e indexación y los roles y habilidades necesarias para ejecutar correctamente el procedimiento. Para validar la capacidad del procedimiento de mejorar la eficacia del almacenamiento y la calidad de los resultados brindados a los usuarios se aplicaron técnicas cuantitativas y cualitativas. Los resultados obtenidos por estas técnicas fueron contrastados a través de una triangulación metodológica que permitió validar el cumplimiento del objetivo de la investigación.

Palabras clave: almacenamiento, estructura, metadato, sistema de recuperación de información.

Abstract

In order to recover and visualize the information hosted on the Cuban website, a recovery and visualization tool called Orion was developed in the country. As fundamental deficiencies of this search engine are the definition of its metadata to correctly structure the digital files that are tracked, the lack of a space for the tracking and storage of videos and the design of the base architecture has deficiencies that do not allow to execute with quality tracking and indexing mechanisms causing overload in the servers. To solve the above problems, a procedure is designed to structure and store the documents for the Orion information retrieval system, combining tracking and indexing techniques, with the aim of improving the storage efficiency and the quality of the results provided to users. The fundamental contributions of the research focus on the structuring of documents from the definition of their metadata and the application of the vector model, the definition of hardware architectures for the deployment of the tracking and indexing components and the roles and skills necessary to correctly execute the procedure. To validate the capacity of the procedure to improve the storage efficiency and the quality of the results provided to the users, quantitative and qualitative techniques were applied. The results obtained by these techniques were contrasted through a methodological triangulation that allowed to validate the fulfillment of the research objective.

Keywords: storage, structure, metadata, information retrieval system.

Índice de tablas

Índice de figuras

INTRODUCCIÓN

En los últimos años el crecimiento de Internet ha sido exponencial. Existen incontables datos sobre cualquier tópico, lo que hace de ésta la fuente más completa de información. Sin embargo, hay más cantidad de artículos de los que las personas manejan efectivamente (Velez y Santos, 2014). Según Internet Live Stats, en la red mundial de hoy, existen más de mil millones de sitios web, por lo que se crea una sobrecarga de información hacia los usuarios (Internetlivestats.com, 2017).

Este cúmulo de información, que se encuentra en diferentes formatos (texto, audio, imágenes, videos), ocasiona que los usuarios de internet, que buscan respuestas a sus interrogantes sobre la información publicada, no encuentren o se les dificulte encontrar respuestas que satisfagan sus necesidades (Umagandhi y Kumar, 2017). La Recuperación de Información, en lo adelante RI, según Baeza (1999), es la parte de la informática que estudia RI (no datos) de una colección de documentos escritos. En la actualidad adquiere un rol más importante debido al valor que tiene la información para la generación de conocimiento.

Se puede plantear que disponer o no de la información justa en tiempo y forma puede resultar en el éxito o fracaso de una operación (Tolosa y Bordignon, 2008). Las herramientas más utilizadas globalmente para acceder a la información dispersa en la web son los Sistemas de Recuperación de Información, en lo adelante SRI. Seroubian (2013) define estos sistemas como agentes de software, programas que residen en una computadora y tienen la misión de registrar, clasificar, indizar y almacenar en bases de datos los documentos de los más diversos sitios de forma automatizada; además, presentan la posibilidad de acceder a esas bases de datos para su consulta.

En el caso del desarrollo del dominio .cu continúa su ritmo creciente. Según el sitio web oficial del Centro Cubano de Información de Red (Cubanic, 2018) hasta el 27 de octubre del 2018 se contaba aproximadamente con unos 7552 dominios registrados bajo .cu y con ello una amplia gama de información disponible, ya sea de deporte, medio ambiente, ciencia, cultura, tecnología, noticias; de ahí la importancia de realizar una búsqueda eficaz de los datos que se desea analizar. En relación con este objetivo la Universidad de las Ciencias Informáticas (UCI), cuenta con una infraestructura tecnológica privilegiada, conectando en su red informática interna una gran cantidad de computadoras.

Producto del desarrollo tecnológico que posee la UCI y los centros de investigación y desarrollo de software que radican dentro de la misma, se ejecutan diversas actividades de investigación y eventos de carácter docente y científico. Como parte de los proyectos que desarrolla la universidad se encuentra una herramienta de recuperación y visualización de la información alojada en la web cubana llamada Orión. Actualmente este buscador está en uso por la red universitaria y ha sido meritorio de diversos premios y avales por la importancia e impacto que brinda para todos los usuarios; convirtiéndolo en una herramienta eficiente para la búsqueda de información alojada en la web cubana y permitiendo a los usuarios un rápido acceso a los recursos publicados en la intranet

de la red universitaria (Leyva, 2016).

Para ejecutar los procesos de rastreo, almacenamiento y visualización de información, los SRI basan su funcionamiento en modelos de RI. Los modelos de representación de documentos clásicos se basan generalmente en el modelo booleano o en el modelo vectorial (Jaimes y Vega, 2005; Tolosa y Feuerstein, 2014; Gudivada, Baeza y Raghavan, 2015). Estos modelos de representación de documentos son adecuados para documentos de texto, que pueden ser páginas web u otros objetos (como elementos multimedia) que estén descritos de forma textual. En los sistemas de recuperación de información no se suele trabajar directamente con los documentos de texto sino con representaciones más estructuradas de los mismos (Martínez, 2006).

En el caso específico de Orión, éste cuenta con tres componentes principales que permiten acceder a los documentos publicados, (entiéndase por documentos texto, imágenes, videos): interfaz web de consulta para realizar las búsquedas, mecanismo de indexación de contenidos y el mecanismo de rastreo. Durante el proceso de rastreo en el buscador Orión, se recepcionan diversos documentos en los cuales se repiten metadatos como el título, contenido, formato y url. Esto ocasiona la duplicidad de documentos y que se dificulte el proceso de recuperar información válida; brindándoles a los usuarios más información de la necesaria para responder a sus interrogantes de búsquedas.

La sobrecarga de los servidores de almacenamiento con información duplicada, provoca que el tiempo de respuesta a los usuarios no sea óptimo. Esto se produce debido a que el proceso de cálculo de similitud entre la consulta insertada por el usuario y los documentos almacenados se ejecuta en un mayor tiempo, además de provocar un uso ineficiente del almacenamiento.

En el proceso de estructuración de documentos el sistema no es capaz de verificar que estos tengan todos los metadatos definidos; lo cual ocasiona que se almacenen documentos incompletos que no aportan ningún valor a la colección, brindando resultados que no contienen campos como el título, resumen o url, perjudicando la homogeneidad de los resultados brindados en la interfaz de visualización. En la arquitectura diseñada para este buscador no se definió un espacio para el procesamiento y almacenamiento de videos por lo que el proceso de recuperación de esos archivos se ve limitado y los usuarios no tienen acceso a material de este tipo que responda a sus interrogantes de búsqueda.

Partiendo de esta situación se plantea el siguiente **problema de investigación:** ¿Cómo estructurar y almacenar los documentos que se gestionan en el buscador Orión, de manera que mejore la eficacia del almacenamiento y la calidad de los resultados brindados a los usuarios?

Para solucionar el problema planteado, se determinó que el **objeto de estudio** se centra en el proceso de estructuración y almacenamiento de documentos en sistemas recuperación de información enmarcado en el **campo de acción** estructuración y almacenamiento de documentos en el sistema de recuperación de información Orión.

Se define como **objetivo general**: Desarrollar un procedimiento para la estructuración y almacenamiento de documentos en el sistema de recuperación de información Orión, que integre técnicas de rastreo e indexación y mejore la eficacia del almacenamiento y la calidad de los resultados brindados a los usuarios.

Para dar cumplimiento al objetivo general, se han trazado los siguientes **objetivos específicos**:

1. Elaborar el marco teórico referencial de la investigación relacionada con la estructuración y almacenamiento de documentos para el sistema de recuperación de información Orión.
2. Definir los metadatos de los documentos en el sistema de recuperación de información Orión para estructurarlos de forma correcta.
3. Diseñar una arquitectura de hardware distribuida para los componentes de rastreo e indexación del buscador Orión.
4. Implementar un procedimiento para la estructuración y almacenamiento de documentos en el sistema de recuperación de información.
5. Validar la propuesta de solución a partir de los métodos científicos definidos.

Para dar respuesta a la problemática se fundamenta la siguiente **hipótesis de investigación:** La aplicación de un procedimiento para la estructuración y almacenamiento de documentos para el SRI Orión, que integre técnicas de rastreo e indexación, mejora la eficacia del almacenamiento y la calidad de los resultados brindados a los usuarios.

Operacionalización:

Variable independiente: Procedimiento para la estructuración y almacenamiento de documentos en el SRI.

Variable dependiente 1: Eficacia de la información almacenada.

Dimensiones para medir la eficacia de la información almacenada:

- **Completitud de los documentos:** Se entiende por un documento completo cuando posee todos los metadatos definidos en la propuesta de solución (Jaramillo et al., 2014).
- **Eficacia del almacenamiento:** Se medirá por el número de bytes que se precisan para almacenar los datos (Sequera, 2010).

Variable dependiente 2: Calidad de los resultados brindados a los usuarios.

Dimensiones para medir la calidad de los resultados devueltos a los usuarios:

- **Precisión**: Evitar devolver resultados que no corresponden con la consulta realizada (Avedaño et al., 2013; Roa et al., 2013; Jaramillo et al., 2014; Romá, 2014; García, 2015; Quiñones, 2015).
- **Exhaustividad**: Obtener tantos documentos relevantes como sea posible (Avedaño et al., 2013; Roa et al., 2013; Jaramillo et al., 2014; Romá, 2014; García, 2015; Quiñones, 2015).
- **Eficiencia**: Velocidad de respuesta a las peticiones de los usuarios (Sequera, 2010).
- **Aceptación de usuario**

Para dar cumplimiento a los objetivos propuestos, se utilizan diferentes métodos científicos. Los principales son (Sampieri, Fernández y Baptista., 2014):

Métodos teóricos:

Método analítico-sintético: empleado en la descomposición del problema científico en elementos por separado y la profundización en el estudio de cada uno de ellos, para luego sintetizarlos en la propuesta de solución.

Método histórico-lógico: se empleó para el análisis de los principales sistemas de recuperación de información, su surgimiento, su evolución y su estado actual, en función de comprender mejor el objeto de estudio de la investigación.

Método hipotético-deductivo: como guía de la investigación se hace uso de una hipótesis científica. A partir de la observación y el análisis del fenómeno en cuestión, se formuló una hipótesis que será comprobada en el proceso de validación.

Métodos Empíricos

Análisis documental: se realizaron consultas a documentación científica para el estudio de los referentes teóricos.

Encuesta: mediante su aplicación se obtuvo mediciones cuantitativas de los elementos cualitativos y cuantitativos abordados en la investigación.

Medición: permite medir los resultados al comparar el SRI antes de integrarle el procedimiento y después del mismo.

Experimentación: mediante los experimentos se evaluó la capacidad del procedimiento para mejorar la calidad de los resultados brindados a los usuarios y la calidad de la información almacenada.

Criterio de expertos empleando el escalamiento de Likert: a partir de su aplicación a expertos se evaluaron los elementos teóricos que fundamentan la investigación.

Técnica Iadov: se aplicó para evaluar el nivel de satisfacción de potenciales usuarios con respecto al procedimiento propuesto.

La **contribución** principal de la investigación se expresa en los siguientes aportes:

Aporte práctico:

- Procedimiento para la estructuración y almacenamiento de documentos en el SRI Orión.

Aporte social:

- La aplicación del procedimiento contribuye a la mejora del proceso de RI, convirtiéndose en un apoyo fundamental en el proceso investigativo y de búsqueda de información alojada en la web cubana.

El presente documento está estructurado en tres capítulos, introducción, conclusiones, recomendaciones y anexos.

Capítulo 1: Marco teórico referencial de la investigación relacionado con el proceso de estructuración y almacenamiento de documentos.

Se abordan los conceptos fundamentales asociados al dominio del problema expuesto. Se realiza una revisión bibliográfica de los principales sistemas y modelos de recuperación de información; cómo estos estructuran los documentos que almacenan, los estándares de datos más utilizados, la arquitectura de hardware que utilizan y los modelos clásicos de recuperación de información.

Capítulo 2: Procedimiento para la estructuración y almacenamiento de documentos en el sistema de recuperación de información Orión.

Se plantea la propuesta del procedimiento para la estructuración y almacenamiento de documentos en el sistema de recuperación de información Orión. Se describe cada una de las etapas del procedimiento, la estructura que debe poseer cada documento y su arquitectura de hardware para un correcto almacenamiento.

Capítulo 3: Validación del procedimiento para la estructuración y almacenamiento de documentos en el sistema de recuperación de información Orión.

Se presentan los resultados de las pruebas realizadas al procedimiento para la estructuración y almacenamiento de documentos en el sistema de recuperación de información Orión. Se valida la calidad de la información almacenada y de los resultados brindados a los usuarios a través de dos experimentos, la técnica de Iadov y el escalamiento de Likert. Luego se hace uso de la triangulación metodológica demostrando la confiabilidad de dichos resultados.

Finalmente se presentan las conclusiones, las recomendaciones, las referencias bibliográficas y los anexos.

CAPÍTULO 1: MARCO TEÓRICO REFERENCIAL DE LA INVESTIGACIÓN

En este capítulo se analizan y presentan los conceptos fundamentales asociados al dominio del problema expuesto. Además, se realiza un análisis de los aspectos principales relacionados con los SRI, la estructuración y almacenamiento de documentos, estándares de metadatos, arquitectura de hardware y modelos de la RI para una mejor comprensión del problema.

1.1 La recuperación de información

La RI surge a finales de la década de 1950. Sin embargo, en la actualidad adquiere un rol más importante debido al valor que tiene la información. Blázquez (2013) considera la RI como el proceso por el cual las demandas informativas y documentales del usuario son resueltas en un sistema de información, compuesto por un corpus documental de volumen variable; cuyo tratamiento de indexación y almacenamiento hacen posible su estructuración, interrogación y representación por medio del empleo de algoritmos matemáticos, estadísticos y semánticos. Según Korfhage (1997), es la localización y presentación a un usuario de información relevante a una necesidad de información expresada como una pregunta.

Luego de analizados los conceptos anteriormente abordados, se resume la RI como el proceso encargado del procesamiento, estructuración y almacenamiento de documentos que permite brindar mediante algoritmos matemáticos, respuestas a las interrogantes de los usuarios relativas a la información publicada en la web. Según Baeza (1999) el problema de la RI, figura 1, puede ser estudiado desde dos puntos de vista: el computacional y el humano. El primer caso tiene que ver con la construcción de estructuras de datos y algoritmos eficientes que mejoren la calidad de las respuestas. El segundo caso corresponde al estudio del comportamiento y de las necesidades de los usuarios.

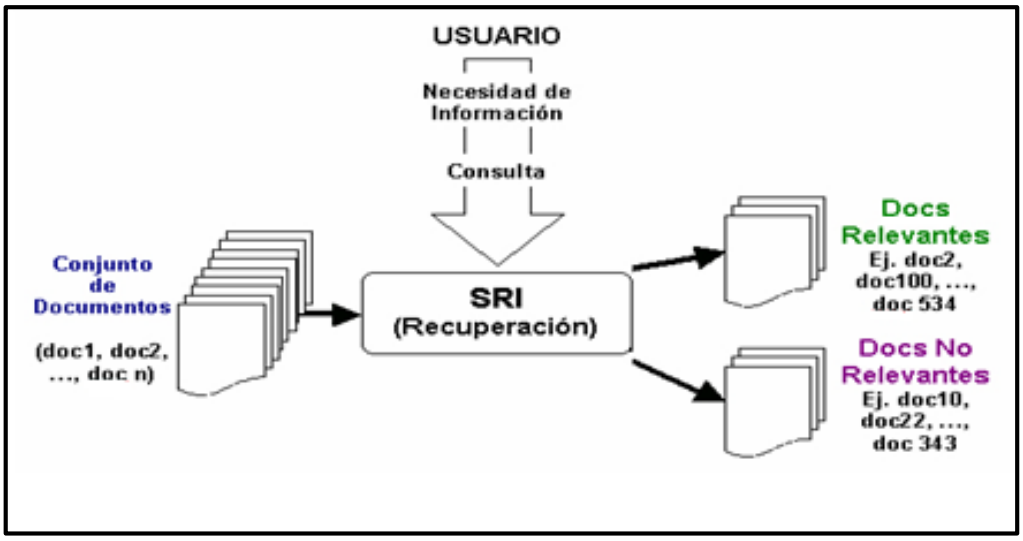


Figura 1. Problema de la RI. Fuente: (Tolosa y Bordignon, 2008).

La tarea de recuperar información puede plantearse de diversas formas en relación a la manera en que el usuario interactúa con el sistema, o bien a partir de las facilidades que este le presenta (Méndez, 2001; Ríssola y Tolosa, 2015):

Recuperación inmediata: El usuario define su necesidad de información y posteriormente, obtiene referencias a los documentos que el sistema determina como relevantes.

- Recuperación ad-hoc: en general la necesidad de información que el usuario presenta se traduce a una consulta en texto libre que posteriormente el sistema procesa y evalúa.
- Navegación o browsing: el sistema ofrece una interfaz con tópicos por los cuales el usuario navega obteniendo referencias a documentos relacionados. En este caso no se expresa una consulta explícita, hecho que facilita la búsqueda a usuarios que no poseen una necesidad clara.

Recuperación Diferida: El usuario detalla sus necesidades definiendo un perfil de modo tal, que el sistema entrega de forma continua aquellos documentos que se incorporen a la colección. Esta práctica recibe el nombre de filtrado y ruteo.

En la presente investigación se asume el uso de la variante ad-hoc de la recuperación inmediata por la relevancia que aporta brindar a los usuarios resultados afines a sus necesidades de búsqueda expresadas en una consulta. De forma general la problemática fundamental de la RI establece que:

- En la web existe un gran número de documentos que conforman una gran colección con diversas categorías.
- Los usuarios con necesidad de respuestas a sus interrogantes, acceden a los SRI e insertan su consulta o criterio de búsqueda, esperando una respuesta que satisfaga sus interrogantes.
- El SRI debe ser capaz de devolver a los usuarios los resultados más relevantes.

1.2 Los Sistemas de Recuperación de Información

Un SRI es una herramienta que posee un conjunto de componentes que permiten el rastreo, indexación y visualización de la información recolectada, como parte de las respuestas ofrecidas a los usuarios después que ejecutan una consulta (Lafferty y Zhai, 2017). El problema principal de un SRI es brindar a los usuarios exactamente los documentos que satisfagan sus necesidades de búsqueda. Para ello se valen de una arquitectura básica organizada en los componentes que se ilustran en la figura 2.

Estos componentes responden a cuatro procesos fundamentales (Shen et al., 2014; Ríssola y Tolosa, 2015; Mora, 2016; Lafferty y Zhai, 2017; Ensias, 2017):

- Proceso de indexación
- Proceso de consulta
- Proceso de evaluación
- Proceso de retroalimentación del usuario

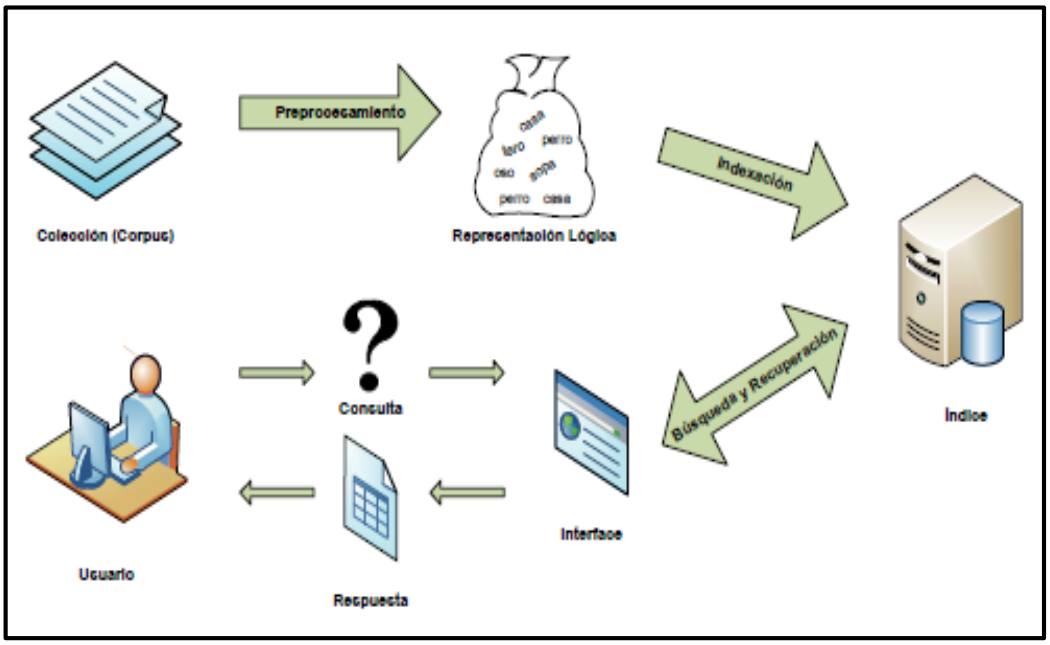


Figura 2. Arquitectura básica de un SRI. Fuente: (Ríssola y Tolosa, 2015).

Esta clasificación no está cerrada, sino que dependiendo del SRI y de sus mecanismos se pueden definir nuevos procesos. La autora de la presente investigación se basará en el proceso de indexación y añade como proceso fundamental en el funcionamiento de un buscador, el rastreo de la información alojada en la web como base para la actualización e incremento constante de la información almacenada.

1.2.1 Proceso de rastreo

Un rastreador web es un programa o script programado que navega por la World Wide Web de una manera sistemática y automatizada. La estructura de la WWW opta una forma gráfica, es decir, los enlaces presentados en una página web pueden utilizarse para abrir otras páginas web. Internet puede ser representada como un grafo dirigido, donde una página web es un nodo y un hipervínculo un borde, por lo tanto, la operación de búsqueda puede ser resumida como el proceso de recorrer un gráfico dirigido. Siguiendo la estructura enlazada de la Web, el rastreador puede recorrer varias páginas nuevas a partir de una página inicial. Un rastreador web se mueve de página a página, mediante el uso de la estructura gráfica de las páginas web (Bandagale, Sawantdesai, Paradkar y Shirodkar, 2017).

Estos programas también se conocen como robots, arañas o gusanos. Los rastreadores Web están diseñados para recuperar páginas Web e insertarlas en el repositorio local (Kausar, Dhaka y Singh, 2013, Udapure et al., 2014; Gupta y Goyal, 2015; Agre y Mahajan, 2015). Según Cambazoglu y Baeza (2015) la función principal del rastreador es localizar las páginas en la Web y descargar su contenido, que se almacena en el disco para su posterior procesamiento. Este proceso normalmente implica varias tareas de análisis, extracción y clasificación.

Para llevar a cabo las tareas de búsqueda, Orión utiliza Nutch como herramienta de rastreo (Orión_Universidad, 2017). Este es un programa de código libre diseñado especialmente para rastrear la web recuperando las páginas que componen la red a través de la estructura de enlaces existentes entre ellas. Nutch crea una base de datos con todos los enlaces encontrados, al tiempo que guarda una copia de todas las páginas localizadas y el resultado del análisis de su contenido, pues incorpora parsers para muchos formatos, no solamente HTML (Cafarella, 2004; Khare et al., 2004).

1.2.2 Estándares en la web

Una de las formas más utilizadas a nivel internacional para representar los datos de la web son los estándares de metadatos. Los metadatos son datos sobre datos, información estructurada que describe a otra información y que permite encontrarla, gestionarla, controlarla, entenderla y preservarla en el tiempo (Rodríguez, 2016). A continuación se realiza un estudio de los estándares de metadatos más utilizados en el mundo.

Learning Object Metadata (LOM): es un modelo de datos, usualmente codificado en XML, utilizado para describir un objeto de aprendizaje y otros recursos digitales similares empleados en el apoyo al aprendizaje. Su propósito es ayudar a la reutilización de objetos de aprendizaje y facilitar su internacionalidad, usualmente en el contexto de sistemas de aprendizaje en línea (Sandoval et al., 2015).

Text Encoding Initiative (TEI): Se encarga de desarrollar un estándar para representar textos en formato digital, basado en el metalenguaje XML. Este estándar ofrece un esquema que incluye y describe un conjunto de elementos que permiten marcar las principales características estructurales, interpretativas y conceptuales de diferentes tipologías de textos (literarios, periodísticos, científicos, etc.), con el objetivo de poder ser procesadas posteriormente (Baena et al., 2014). Es utilizado específicamente en bibliotecas o en colecciones de textos digitales. A pesar de que el estándar TEI hace casi tres décadas que se aplica, la evolución tecnológica ha renovado el interés en él (Zhang et al., 2015) y aunque es muy utilizado, presenta varias desventajas vinculadas a su rigidez: no se pueden crear elementos ni atributos, algunos elementos son obligatorios en cada texto y se debe respetar una jerarquía para cada elemento.

Metadata Object Description Schema: Define un esquema para un grupo de elementos bibliográficos que pueden ser usados con una variedad de propósitos, particularmente para aplicaciones bibliotecarias. A criterio de la autora de esta investigación este estándar al estar enfocado en aplicaciones bibliotecarias no reúne entre sus metadatos los necesarios para describir un documento web.

Resource Description Framework (RDF): es una base para procesar metadatos. Proporciona interoperabilidad entre aplicaciones que intercambian información legible por máquina en la Web. RDF destaca por la facilidad para habilitar el procesamiento automatizado de los recursos Web. A pesar de ser un estándar robusto a criterio de la autora de esta investigación, añade complejidad al proceso de estandarización de documentos.

OWL: El Lenguaje de Ontologías Web, está diseñado para ser usado en aplicaciones que necesitan procesar el contenido de la información en lugar de únicamente representar información para los humanos. OWL facilita un mejor mecanismo de interpretabilidad de contenido Web que los mecanismos admitidos por XML, RDF y esquema RDF (RDF-S) proporcionando vocabulario adicional junto con una semántica formal. A pesar de las ventajas que ofrece OWL, complejiza el proceso de estandarización ya que no solo ejecuta proceso de estructuración sino que también procesa el contenido de la información que estos contienen.

Simple Knowledge Organization System (SKOS): es un vocabulario RDF para la representación de sistemas de organización del conocimiento semiformales tales como: tesauros, taxonomías, esquemas de clasificación y listas de encabezamiento de materias. SKOS está basado en RDF por lo que dichas representaciones pueden ser legibles por máquinas e intercambiarse entre aplicaciones de software, así como publicarse en la World Wide Web. SKOS es un estándar robusto, pero no está enfocado en la estructuración de documentos web.

Online Information Exchange (ONIX): es el estándar internacional para la representación y comunicación de información sobre los productos de la industria del libro en formato electrónico. La autora de la presente investigación lo considera un estándar práctico en la representación de libros, pero no se adecua a la generalidad de documentos en la web.

Esquema de Metadatos Dublin Core (DC): Es el esquema de meta-información más utilizado a nivel mundial para describir los metadatos de los recursos digitales. Para maximizar las posibilidades de interoperar con otras colecciones de datos, se utiliza a DC como base de un esquema de metadatos (Piedra et al., 2015). Su objetivo es elaborar normas interoperables sobre metadatos y desarrollar vocabularios especializados en metadatos para la descripción de recursos que permitan sistemas de recuperación más inteligentes (Sandoval et al., 2015). La misión de DC es conseguir que sea más fácil encontrar recursos usando internet (Dublin Core Metadata Initiative, 2017). Entre las ventajas más significativas de este estándar es que describe cualquier recurso digital y que cada uno de los campos es opcional por lo que lo vuelve flexible y simple. Este estándar define 15 descriptores, entre ellos

Título: Se refiere al título que lleva por nombre el documento.

Descripción: En este campo se hace un breve resumen sobre el contenido del objeto digital.

Lenguaje: En este campo se establecen las siglas correspondientes al idioma en que se presenta la publicación.

Fecha: Se anota la fecha de elaboración del registro.

Formato: En este campo se registra el tipo de extensión con que se presenta el objeto digital, ya sea HTML, JPG, GIFF, PDF, AVI.

Identificador: Se refiere a la dirección electrónica de origen a la que está adscrito el material. Para ello se utilizan las siglas URL.

Después de analizados los estándares de metadatos más utilizados, esta autora decide basar la propuesta de estructura de documentos a utilizar, en el estándar DC; debido a que este estándar

recoge los descriptores más representativos de un documento web y su amplia utilización a nivel internacional facilita los procesos de integración con otros sistemas.

1.2.3 Proceso de indexación

Según Blázquez (2013), Büttcher et al. (2016), Fagan (2017) y Lafferty y Zhai (2017), indexar es la acción de construir un fichero inverso de forma automática o manual. Este proceso es necesario para localizar y recuperar rápidamente cada uno de los términos del texto de un documento. Esto significa que a cada palabra se le asigna un identificador del documento en el que aparece, un indicador de la posición que ocupa en el texto (párrafo, línea, número de caracter de inicio) y un número de identificación para ese término propiamente dicho (único e irrepetible). De esta forma se conoce la posición exacta de cada término en los documentos de la colección y posibilita el posterior análisis de frecuencias.

Un correcto proceso de indexación debe funcionar de forma dinámica. Esto se traduce en que se deben realizar constantes modificaciones a los contenidos y se debe incrementar su número. Para lograr un mejor tiempo de ejecución de las consultas se debe disminuir el tamaño de los documentos, por lo que antes de proceder a su almacenamiento definitivo se deben ejecutar una serie de mecanismos de depuración de páginas web para la extracción y procesamiento de textos.

Depuración y supresión de código fuente: Para efectuar un proceso de depuración y limpieza del contenido, es preciso identificar qué bloque de HTML contiene la información central del documento. En esencia el web crawler debe ser capaz de determinar qué elementos corresponden a la interfaz visual, la navegación hipertextual, los menús, la publicidad, los banners; así como cualquier otro elemento accesorio que no forma parte del contenido (Blázquez, 2013). Este proceso permite eliminar el contenido no representativo del documento que se analiza, reduciendo así su tamaño y mejorando su representación.

Tokenización: Es el proceso que descompone el texto de un documento en mínimas unidades. Los elementos resultantes son denominados tokens. La lista de ítems que conforman los tokens son utilizados posteriormente para el procesamiento del lenguaje natural, estadístico, lingüístico, almacenamiento y posterior recuperación de información. Los tokens a su vez pueden ser identificados mediante una codificación ASCII o en su defecto Hexadecimal, con el objeto de facilitar la identificación uno a uno de cada caracter que compone la palabra. Este proceso permite la identificación de cadenas de caracteres de forma unívoca, de cara a posteriores tratamientos de depuración, eliminación de signos de puntuación o la reducción morfológica (Blázquez, 2013).

Conversión a minúsculas, eliminación de signos de puntuación y acentos: En este proceso se convierte todo el texto a minúscula, se eliminan los signos de puntuación y los acentos. Según Blázquez (2013) esta actividad facilita el proceso de identificar la similitud entre la consulta y los documentos y aunque el usuario introduzca una consulta incorrecta la información puede ser recuperada. De esta forma el sistema de recuperación identifica mediante codificación hexadecimal qué caracteres debe reemplazar, modificar o en su defecto eliminar. Es habitual que tales

programas integren la referencia completa de caracteres, denominadas también como tablas de equivalencias entre caracteres.

Transliteración y reemplazo de caracteres especiales: En muchos casos, la conversión de un texto a minúsculas, la supresión de acentos y signos de puntuación no es suficiente. En muchos casos el idioma en el que está escrito el documento o las particularidades del texto implican el uso de caracteres especiales que requieren transliteración o reemplazo por otro caracter más sencillo y equivalente en el teclado estándar. Estos procesos resultan complejos, puesto que en todo momento se debe asegurar la identificación del caracter original (Blázquez, 2013).

Eliminación de palabras vacías: Las palabras vacías o stop words son aquellas que no aportan significado al documento sino que se utilizan para seguir las reglas del idioma (Nava, Domínguez y González, 2015). En este proceso se eliminan las palabras vacías del documento, por ejemplo: pronombres, partículas interrogativas y ciertas preposiciones (Durán, Ramírez y Juganaru, 2014). La cantidad de ocurrencias de una palabra en el texto determina si es o no un stop word, dado que cuantas más ocurrencias existan menos relevancia tiene en el texto. Dentro de este grupo se encuentran los artículos, los pronombres, las preposiciones, y las conjunciones. Esta técnica permite reducir el tamaño del texto para analizar, eliminando aproximadamente el 30% o 40% de dichas palabras (Ramos y Velez, 2016). Este proceso de selección pasa por determinar la importancia de un término en el documento, de tal forma que, si es lo suficientemente importante, se escogerá para ser incluido en el conjunto de términos final.

Reducción morfológica: Luego de depurar los datos, el siguiente paso en el procesamiento de los documentos es la reducción morfológica. Este proceso depura todos los términos de un texto, reduciendo su número de caracteres, simplificando su forma original, género, número, desinencia, prefijo o sufijo, en una forma de palabra más común o normalizada; debido a que la mayor parte de ellas tienen la misma significación semántica. Este proceso reduce el tamaño de los términos, del diccionario, fichero inverso y mejora la exhaustividad de los resultados en la recuperación de información. Uno de los métodos más conocidos para llevar a efecto la reducción morfológica es el algoritmo de Martin Porter, diseñado para eliminar las palabras más comunes del inglés inicialmente y posteriormente aplicado para terceros idiomas (Blázquez, 2013).

Stemming: Es uno de los procesos que se ejecutan en la reducción morfológica. En este proceso se reducen las palabras a sus unidades mínimas con significado, logrando así identificar y acotar la raíz de cada una de ellas. Además permite la recuperación de los documentos que tienen variaciones sintácticas de los términos que se están recuperando (Ramírez, 2007; Uysal y Gunal, 2014; Balakrishnan y Lloyd-Yemoh, 2014). Estos algoritmos se basan en un conjunto sencillo de reglas que truncan las palabras hasta obtener una raíz común (Matías et al., 2016).

Lematización: Consiste en convertir una palabra en su forma sin flexionar (lema). Por ejemplo, para la palabra “límite” su raíz o stem sería “limit” y su lema (puede no coincidir el stem y la raíz léxica). Para ello se necesita, en general, un diccionario morfosintáctico. Con el uso de este tipo de técnicas se consigue, por un lado, reducir el tamaño del índice a construir y por otro, aumentar la

cobertura del sistema de recuperación (aunque a la vez suele perderse precisión). El motivo es que muchas palabras distintas serán tratadas por el sistema como si fueran la misma (Granados, 2013; Bolaños, 2015).

La eliminación de documentos duplicados: se estima que en internet existen muchas páginas web duplicadas, aproximadamente el 30 %. La eliminación de documentos duplicado permite mejorar el rendimiento y reducir el espacio de almacenamiento. Pero la tarea de identificar documentos similares no es trivial, ya que pueden darse diferentes razones que compliquen esta labor, como por ejemplo el formato del documento, dos documentos pueden ser idénticos en contenido pero estar en diferentes formatos (Mora, 2016).

Selección de términos: Según Cabeza (2014) el cálculo de la importancia de cada término se conoce como ponderación del término. La ponderación de los términos es el proceso que tiene como finalidad conocer la importancia de los términos para representar un documento y permitir su posterior recuperación. Esto implica que se debe determinar el poder de resolución de los términos de la colección, o lo que es lo mismo, la capacidad de los términos para representar el contenido de los documentos en la colección, que permitan identificar cuáles son relevantes o no ante la consulta del usuario. Al valor e índice que es capaz de determinar este extremo se le denomina "peso del término" o "ponderación del término" y su cálculo implica determinar la "Frecuencia de aparición del término TF" y la "Frecuencia inversa del documento para un término IDF" (Berry, 2004; Blázquez, 2013).

Factor TF: El factor TF es la suma de todas las ocurrencias o el número de veces que aparece un término en un documento, lo que permite determinar su capacidad de representación. A este tipo de frecuencia de aparición también se la denomina "Frecuencia de aparición relativa" porque atañe a un documento en concreto y no a toda la colección. Si su frecuencia de aparición TF es baja la representatividad es elevada y si es alta el documento tiene muy baja representatividad. Su cálculo se efectúa una vez el texto del documento ha sido normalizado, según los procesos de depuración. Posteriormente se lleva a cabo el conteo de las veces que el término aparece presente en el documento, figura 3 (Blázquez, 2013; Cabeza, 2014), a través de la fórmula 1.

$$tf(n) = \sum_{(n)} D1 \tag{1}$$

Vector	T1	T2	T3	T4	T5	T6
D1	2	1	0	0	1	1
D2	1	0	0	0	1	0
D3	0	1	1	1	1	1
D4	0	0	0	0	0	1

Figura 3. Cálculo del TF de un término. Fuente: (Blázquez, 2013).

Factor IDF: El factor IDF de un término es inversamente proporcional al número de documentos en los que aparece dicho término. Esto significa que cuanto menor sea la cantidad de documentos, así como la frecuencia absoluta de aparición del término, mayor será su factor y cuanto mayor sea la frecuencia absoluta relativa a una alta presencia en todos los documentos de la colección, menor será su factor discriminatorio (Blázquez, 2013).

El factor IDF es único para cada término de la colección. Esto significa que su cálculo, fórmula 2, el IDF de un término dado (n) se realiza aplicando el logaritmo en base 10 de N (Número total de documentos de la colección) dividido entre la frecuencia de documentos para un término (n) en la colección (o lo que es lo mismo el número de documentos de la colección en los que aparece el término (n) dado), figura 4.

$$\mathrm{IDF(n)} = \log_{10} \frac{N}{DF(n)} \qquad (2)$$

Idf(T1) = log (4/2) = log 2 = 0.301
Idf(T2) = log (4/2) = log 2 = 0.301
Idf(T3) = log (4/1) = log 4 = 0.602
Idf(T4) = log (4/1) = log 4 = 0.602
Idf(T5) = log (4/3) = log 1.33 = 0.124
Idf(T6) = log (4/3) = log 1.33 = 0.124

Figura 4. Cálculo del IDF de un término. Fuente: (Blázquez, 2013).

Ponderación TF-IDF: El peso de un término en un documento es el producto de su frecuencia de aparición en dicho documento (TF) y su frecuencia inversa de documento (IDF), fórmula 3. Esto significa que el peso o ponderación se calcula para cada término en cada documento (Blázquez, 2013), figura 5.

$$TF\text{-}IDF(n,d) = TF(n,d)\ x\ IDF(n,d) \qquad (3)$$

Vector	T1	T2	T3	T4	T5	T6
D1	2*0.301=0.602	0.301	0	0	0.124	0.124
D2	0.301	0	0	0	0.124	0
D3	0	0.301	0.602	0.602	0.124	0.124
D4	0	0	0	0	0	0.124

Figura 5. Peso TD-IDF para un término en un documento. Fuente: (Blázquez, 2013).

Como herramienta de indexación, el buscador Orión utiliza Solr. Este funciona como un servidor de búsquedas utilizado para la indexación en tiempo real. Es escalable, permitiendo realizar búsquedas distribuidas y replicación de índices. Según la página principal de Solr, las características más destacables de este servidor de búsquedas son (Smiley et al., 2015):

- Capacidades avanzadas de búsqueda de texto completo.
- Optimizado para un tráfico Web elevado.
- Interfaces abiertas basadas en estándares XML, JSON y HTTP.
- Completas interfaces Web de administración.
- Estadísticas del servidor expuestas mediante JMX para su monitorización.
- Escalable linealmente, replicación automática del índice, recuperación automática.
- Indexación cuasi-inmediata.
- Arquitectura extensible mediante plugins.

1.3 Modelos de recuperación de información

A continuación se realiza un estudio de varios modelos computacionales analizando 3 factores que a criterio de esta autora, son necesarios para definir un procedimiento para la estructuración y almacenamiento de documentos en un SRI:

- Metadatos para documentos, imágenes y videos
- Estructuración de documentos
- Arquitectura de hardware

1.3.1 Modelos clásicos de recuperación de información

A pesar de que los modelos clásicos de recuperación de información sólo definen la representación de las consultas insertadas por los usuarios y los documentos almacenados en el SRI, además de las expresiones matemáticas para el cálculo de similitud entre ambos, son ampliamente utilizados en el campo de la RI, como guía para el desarrollo de SRI. A continuación se abordan los tres modelos más reconocidos en la bibliografía estudiada.

Modelo booleano: está basado en la teoría de conjuntos y el álgebra booleana. Cada documento es representado por un conjunto de palabras clave (Arora y Bhardwaj, 2014). Velasco (2014) constata que la relevancia de una palabra clave en un documento se mide únicamente considerando su presencia o ausencia dentro del mismo, lo que lo convierte en un sistema binario puro. Los términos son extraídos de los documentos y representan el contenido de los mismos. Se utilizan operadores lógicos: AND, OR y NOT, y los resultados son referencias a documentos, donde la representación de la consulta satisface las restricciones lógicas de la expresión de búsqueda. Gómez (2014) argumenta que en el modelo original no hay orden de relevancia sobre el conjunto de respuestas a la consulta, todos los documentos poseen la misma relevancia**.** Por las carencias de la lógica, este modelo no soporta la relevancia de documentos**.**

Se definen como principales ventajas y desventajas del modelo booleano las siguientes (Blázquez, 2013; Monsalve, 2015):

Ventajas:

- Es eficiente y simple.
- Utiliza el álgebra booleana.
- Útil para realizar experimentos de distintas naturalezas.
- El modelo booleano permite procesar colecciones muy grandes rápidamente.
- Es un modelo flexible ya que permite el empleo de distintas conectivas para precisar la consulta del usuario. Permite aproximar bastante las consultas por frase exacta y resulta válido para recuperar por medio de vocabulario controlado.

Desventajas:

- En muchos casos, las necesidades de información son complejas y ello entraña cierta dificultad a la hora de expresar las consultas mediante fórmulas lógicas que pueden incluso llegar a concatenarse.
- A veces el usuario puede imponer una lógica semántica que no se corresponda con la lógica algebraica de Boole, implicando un uso incorrecto de los operadores.
- El volumen de resultados no se puede controlar, ya que la consulta plantea una resolución absoluta para toda la colección en la que se aplica. Esto significa que el resultado puede ser excesivamente grande o pequeño.
- Los resultados obtenidos pueden ser perfectamente relevantes o absolutamente irrelevantes. No hay gradación o término medio, ya que el funcionamiento del modelo booleano se basa en equiparación exacta.
- No analiza la frecuencia del término y con todos los operadores OR se obtienen demasiados resultados, mientras que con todos AND se recuperan muy pocos.

Este modelo a pesar de brindar algunas ventajas significativas como la eficiencia y la simplicidad no satisface las necesidades definidas en esta investigación. Al no analizar la frecuencia de los términos en un documento puede devolver a los usuarios resultados que no satisfagan las necesidades de búsquedas de los mismos.

Modelo probabilístico: Fue propuesto por Robertson y Jones (Robertson y Jones, 1976). El objetivo del modelo probabilístico, es calcular la probabilidad de relevancia dados una consulta y un documento (Fernández, 2013; Machado, 2015; Amador, 2015). La idea de este modelo es la siguiente: dada una consulta, existe un conjunto de documentos que contiene los documentos relevantes y no otros. Si se tuviera una descripción adecuada de este conjunto, no habría problema para encontrar los documentos, pero no se tiene. Por lo que el modelo presupone que existe ese conjunto de documentos relevantes **(R)**. Los documentos que no aparecen en este conjunto se consideran no relevantes **(R')**. Teniendo en cuenta que la probabilidad de que el documento d sea relevante se representa **P(R|d)** y que la probabilidad de que el documento d no sea relevante se representa **P(R'|d)**, un documento será relevante si: **P(R|d) > P(R'|d).**

Las principales ventajas y desventajas de este modelo son abordadas a continuación (Blázquez, 2013; Monsalve, 2015):

Ventajas:

- Ordena los resultados por relevancia.
- Sigue un razonamiento matemático basado en probabilidades, lo que permite que tenga extensiones populares como las redes bayesianas.
- Es considerado uno de los mejores modelos dados sus buenos resultados con colecciones reales y corpus de entrenamiento.
- Su método de recuperación es mediante equiparación parcial, superando al método de equiparación exacta del modelo booleano.
- Retroalimentación por relevancia, acepta feedback.
- Asume la independencia de los términos de la consulta.
- Asigna pesos a los términos, permitiendo recuperar los documentos que probablemente sean relevantes.

Desventajas:

- Mantiene el modelo binario de recuperación de información, sin analizar todos los términos del documento como ocurriría en el modelo vectorial.
- Requiere alta capacidad de computación, resultando complejo de implementar.
- Necesita efectuar una hipótesis inicial que no siempre resulta acertada.
- No analiza la frecuencia de aparición de cada término en el documento, tal como lo haría un modelo vectorial.

La asignación de pesos a los términos de este modelo hace de él una fuente teórica a analizar en esta investigación. Valorando las desventajas que posee, analizando su complejidad de implementación y la carencia del manejo de la frecuencia de aparición de los términos en el documento.

Modelo vectorial: El modelo vectorial fue presentado por Salton en 1975 y posteriormente asentado en 1983 junto a Mc Gill. Parte del fundamento de que se puede representar los documentos como vectores de términos; entonces los documentos pueden situarse en un espacio vectorial de n dimensiones, es decir, con tantas dimensiones como elementos tenga el vector (Lizcano Bueno y Pérez, 2016; Agredo et al., 2013; Singh y Singh, 2015).Este modelo utiliza el enfoque lingüístico léxico, el cual se refiere al análisis concerniente a palabras individuales; y se basa en dos elementos fundamentales: un esquema de pesos y una medida de similitud (Torres y Arco, 2016), de una consulta dada por el usuario con respecto a los documentos de la colección cuyos términos fueron ponderados mediante TF-IDF (Blázquez, 2013; Lal, Qamar, y Shiwani, 2018). Este modelo se basa en tres principios esenciales (Martínez, 2006):

- La equiparación parcial es la capacidad del sistema para ordenar los resultados de una búsqueda, basado en el grado de similitud entre cada documento de la colección y la consulta.

- La ponderación de los términos en los documentos, no limitándose a señalar la presencia o ausencia de los mismos, sino adscribiendo a cada término en cada documento un número real que refleje su importancia en el documento.
- La ponderación de los términos en la consulta, de manera que el usuario puede asignar pesos a los términos de la consulta que reflejen la importancia de los mismos en relación a su necesidad informativa.

El modelo de espacio vectorial emplea el peso de los términos para cada documento. Además refleja la relevancia de los términos del documento de cara a su representatividad en la colección, adquiriendo una forma como la que se muestra en la figura 6.

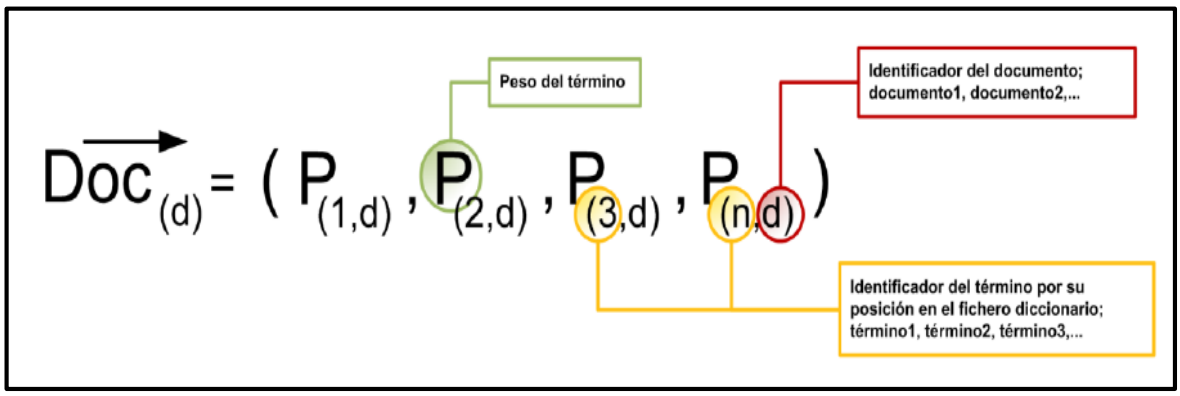


Figura 6. Representación del vector de un documento. Fuente: (Blázquez, 2013).

A este conjunto de números reales, que son los pesos que representan al documento, se les denomina vector del documento; permitiendo su representación en el espacio vectorial y en consecuencia, su tratamiento matemático. Posteriormente la colección sigue lo que se denomina un Proceso de Vectorización, por el que todos los documentos son representados mediante pesos TF-IDF, la consulta del usuario también requiere de dicho tratamiento. Lo que significa que se tiene que ponderar la importancia de los términos de la consulta para poder generar el Vector de la consulta del usuario (Blázquez, 2013).

El siguiente paso es aplicar el proceso de equiparación para determinar la similitud entre la consulta y los documentos. Se pueden utilizar diferentes funciones para realizar el cálculo del coeficiente de similitud, entre las que destacan: producto escalar, coseno del ángulo entre los dos vectores (las más utilizadas), coeficiente de Dice y coeficiente de Jaccard (Monsalve, 2015).

De forma general el modelo vectorial posee una serie de ventajas e inconvenientes que son abordados a continuación (Blázquez, 2013; Aguilar y Mosquera, 2015; Monsalve, 2015):

Ventajas:

- El modelo vectorial propone diferentes técnicas para el cálculo de similitud entre un documento y una consulta.
- Analiza los pesos TF-IDF para determinar la representatividad de los documentos de la colección.
- Utiliza poco espacio de almacenamiento al representar los documentos como vectores.

- Mejores resultados en experimentos, sobre todo en grandes colecciones.

Desventajas:

- Al ser un modelo estadístico-matemático, no analiza la estructura sintáctico-semántica del lenguaje natural.

Teniendo en cuenta el estudio anterior y basado en el estudio de Balamurugan y Iyswarya (2017), tabla 1, se decide basar el proceso de estructuración y representación de documentos en el modelo vectorial debido a sus ventajas, siendo las más importantes a criterio de esta autora, su forma de estructurar y representar los documentos en forma de vectores.

Tabla 1. Comparación de los modelos clásicos. Fuente: (Balamurugan y Iyswarya, 2017).

Modelos de RI / Atributos	**Booleano**	**Probabilístico**	**Vectorial**
Concepto	Basado en la teoría de conjuntos y el álgebra booleana.	Basado en el principio de clasificación de probabilidad.	Basado en el concepto de vectores.
Representación	Los documentos están representados por los términos de índice extraídos de documentos, y las consultas son términos en expresiones booleanas.	Documentos y consultas están representados por vectores binarios.	Representado en forma de vectores de términos ponderados.
Tipo de información	No considera información semántica.	Considera información semántica.	Considera información semántica.
Ocurrencia de palabras	El número de ocurrencias no se menciona.	Ocurrencia basada en la relevancia de la probabilidad.	Maneja el número de ocurrencia.
Ventajas	Fácil de implementar.	Rangos por probabilidades.	Modelo simple, los pesos no son binarios.
Desventajas	No clasifica documentos, recupera demasiados o muy pocos.	Pesos binarios, frecuencias ignoradas y supuestos de independencia.	No maneja sinónimos.

1.3.2 Definición de metadatos

SRI Google

El motor de búsqueda Google, persigue como objetivo que sus usuarios encuentren la información que necesitan y consigan hacerlo de la forma más sencilla y rápida posible. Ofrece servicios como búsqueda de imágenes, libros, noticias, videos, documentos, entre otros. La búsqueda avanzada en Google, incluye aspectos comunes con el resto de las búsquedas personalizadas que se ofrecen; como lo son, los criterios para encontrar todas, cualquiera o ninguna de las palabras de una frase, o incluso expresiones exactas. Dichos criterios, también se pueden incluir en el cuadro de búsqueda mediante el uso de operadores especiales definidos por el propio buscador (Argomedo et al., 2014). Específicamente para la búsqueda avanzada de documentos se ofrecen filtros que permiten obtener resultados específicos por idioma, región, última actualización, sitio o dominio donde se encuentra el contenido, términos en que aparece como por ejemplo título o en su dirección de web, tipo de archivo donde cada página busca el formato que prefiera, derechos de uso (López et al., 2015). Google brinda además la posibilidad de buscar imágenes a partir de diferentes criterios de búsqueda. Algunos de estos criterios son:

- Mostrar imágenes dada una o varias palabras: muestra imágenes asociadas a todas las palabras del criterio de búsqueda.
- Mostrar imágenes dada una frase exacta: muestra imágenes que contengan exactamente el criterio de búsqueda.
- Mostrar imágenes asociadas a cualquiera de varias palabras: agrega el operador lógico disyuntivo “OR” entre las palabras del criterio de búsqueda.
- Mostrar imágenes que no estén asociadas a ninguna de las palabras: agrega el operador de negación “-” delante de la palabra que no se quiere incluir.
- Mostrar imágenes por tamaño: muestra imágenes según el tamaño deseado.
- Mostrar imágenes por color: muestra imágenes donde predomine un color en particular o varios colores seleccionados previamente por el usuario.
- Mostrar imágenes por tipo: muestra imágenes según el tipo de imagen deseada por el usuario. Pueden ser rostros, fotografías, dibujos en líneas, animadas entre otros.

Para la búsqueda avanzada de videos se ofrecen filtros que permiten obtener resultados específicos por idiomas, duración, fecha de publicación, calidad de los videos, sitio o dominio donde se encuentra el contenido, videos subtitulados e incluso, se incluye un filtros para evitar contenido para adultos. Google, considera un video de corta duración a aquellos con menos de 4 minutos, de duración media a los que tienen entre 4 y 20 minutos y de larga duración, a los de más de 20 minutos, incluso se permite hacer búsquedas para cualquier duración, constituyendo este, el comportamiento por defecto. El filtro para las fechas de publicación incluye rangos de fechas para la última hora, día, semana o mes, además como forma predeterminada busca videos publicado en cualquier momento. Por otra parte, se permiten buscar videos de cualquier calidad o solamente aquellos de alta calidad.

SRI Bing

Bing es un motor de búsqueda de Microsoft. Utilizar Bing es muy parecido a utilizar otros buscadores, en el que puedes ingresar una o más palabras clave en la barra de búsqueda para encontrar páginas web que mencionen las palabras o frases que se ingresasen. En Bing, se puede llevar a cabo una búsqueda de artículos, imágenes, videos, noticias, sitios de compras y más. También se puede ingresar comandos específicos con las palabras claves para reducir los resultados de búsqueda.

Para una búsqueda avanzada en el buscador Bing, se cuentan con filtros que limitan los resultados según el sitio, tipos de archivos, título, idioma, región, búsqueda segura, ubicación (Rautenstrauch, 2010). Algunas de las opciones de búsqueda avanzada de imágenes que brinda Bing son las siguientes:

- Búsqueda de imágenes por tamaño: esta opción al igual que en Google muestra imágenes según el tamaño deseado ya sea mediano, pequeño, papel tapiz, grande, etc.
- Buscar imágenes por color: muestra imágenes donde predomine un color en particular o varios colores seleccionados previamente por el usuario.
- Por tipo de imagen: muestra imágenes según el tipo de imagen deseada por el usuario.
- Buscar imágenes por diseño: muestra imágenes cuadradas, rectangulares entre otros diseños.

Para una búsqueda avanzada de videos en el buscador Bing, se cuenta con filtros que limitan los resultados según la duración del contenido multimedia, la fecha en la que fue publicado en la web, resolución con la que cuenta, fuente o lugar de origen y una búsqueda segura para excluir el contenido para adultos de los resultados. Bing considera un video corto a aquellos con menos de 5 minutos de duración y largo a los de más de 20 minutos, mientras que un video de mediana duración es aquel comprendido entre 5 y 20 minutos. Como comportamiento por defecto, Bing ofrece videos de cualquier duración dentro de sus resultados.

Similar a Google, el filtro para las fechas de publicación de Bing, incluyen rangos de fechas para la última hora, día, semana o mes, y busca videos publicados en cualquier momento de forma predeterminada. Por otra parte, el filtro de resoluciones limita los resultados de las búsquedas según las dimensiones de las imágenes y de los videos. Para esto, cuenta con valores de 360, 480, 720, 1080 píxeles o mayor. Como comportamiento predeterminado muestra videos de cualquier resolución.

Otro de los filtros que ofrece este buscador, está relacionado con la fuente o el origen de los videos. Para ello definen algunos sitios web populares como lo son: YouTube, MySpace, Dailymotion y Metacafe.

La búsqueda segura puede ser desactivada para que no filtre el contenido para adultos o se puede realizar de forma estricta o moderada. En el primer caso se filtra textos, imágenes y videos para adultos y en el segundo, solamente las imágenes y los videos, siendo este, la configuración predeterminada.

SRI Yahoo

Yahoo es un buscador muy popular y con una de las mejores bases de datos. Su misión es ser el servicio global de Internet más esencial para consumidores y negocios. Este permite a los usuarios obtener respuestas a interrogantes planteadas diariamente en un lenguaje natural. Para el caso de la búsqueda de documentos, el sistema incluye fecha, dominio, idioma, país, título y tipo (Vargas, 2016).

Los SRI internacionales no proponen un mecanismo de estandarización de metadatos, ni define cuales son los más utilizados, pero las funcionalidades que brindan permitió identificar metadatos necesarios que ayudan al usuario a realizar búsquedas más certeras (Vargas, 2016).

Yahoo también brinda una búsqueda avanzada de imágenes filtrando por:

- Búsqueda de imágenes por tamaño: esta opción al igual que en los demás buscadores, muestra imágenes según el tamaño deseado por el usuario, aunque esta se diferencia de los demás por permitir buscar imágenes con una proporción determinada para fondos de pantalla.
- Buscar imágenes por color: muestra imágenes donde predomine un color en particular o varios colores seleccionados por el usuario.

Para el caso de la búsqueda de videos, el sistema incluye el criterio de relevancia para filtrar los resultados. Mediante este criterio se pueden buscar los videos más relevantes (valor por defecto para este filtro), populares, vistos o los más recientes. Además, se pueden buscar videos que hayan sido publicados en el día, la semana o el mes en curso, teniendo este filtro como comportamiento predeterminado logrando mostrar los resultados independientemente de la fecha en la que haya sido publicado el contenido.

SRI Duck duck go

Es un motor de búsqueda establecido en Pensilvania, Estados Unidos. Este motor utiliza la información de sitios de origen público con el objetivo de aumentar los resultados tradicionales y mejorar la relevancia. Su filosofía hace hincapié en la privacidad y en no registrar la información del usuario.

Este buscador trabaja en asociación con Yahoo, esta asociación le permitió incorporar un filtrado de búsqueda sobre la base de fechas y además, en los resultados de un sitio mostrar enlaces que facilitan acceder a sus subsecciones. Además contiene filtros para realizar una búsqueda por regiones y por fecha. En el caso de las imágenes se le agregan los filtros de tamaño, diseño, color y en los videos resolución, duración y licencia (Singh et al., 2013).

SRI Aol Search

AOL Inc., anteriormente conocida como America Online, es una empresa de servicios de internet y medios con sede en Nueva York. El sistema muestra muy poca información de los documentos que devuelve siendo los filtros más utilizados, color y tamaño (en imágenes) y resolución (en videos) (Sullivan, 2004).

SRI Yandex

Es una corporación multinacional rusa especializada en productos y servicios relacionados con Internet, incluidos servicios de búsqueda e información, comercio electrónico, transporte, navegación, aplicaciones móviles y publicidad en línea. Es la compañía de tecnología más grande de Rusia, el mayor motor de búsqueda en Internet en Rusia y es el quinto motor de búsqueda más grande del mundo. Alguna de las opciones que utiliza este buscador para el filtrado son las fechas, la categoría de la información (cultura, deporte, salud, naturaleza) y en el caso de los videos le añade la cantidad de visitas que obtuvo (Savenkov et al., 2011).

El estudio de estos SRI permitió definir un conjunto de metadatos que los buscadores utilizan para brindar de forma organizada la información que se le muestra al usuario, tabla 2 y con ello definir como metadatos a utilizar los siguientes, al ser estos los más empleados por estos buscadores:

Documentos: título, url, resumen, formato, fecha, palabras claves, lenguaje, sitio, autor, categoría.

Imágenes: título, url, resumen, fecha, tamaño, formato, lenguaje, región, autor, categoría.

Videos: título, url, resumen, fecha, duración, formato, subtitulado, tamaño, director, categoría.

Tabla 2. Metadatos utilizados por los SRI.

SRI	Google	Bing	Yahoo	Duck duck go	Aol Search	Yandex
Metadatos de documentos	Título	Título	Título	Título	Título	Título
	Url	Url	Url	Url	Url	Url
	Resumen	Resumen	Resumen	Resumen	Resumen	Resumen
	Formato	Formato	Formato	Fecha	Fecha	Fecha
	Fecha	Fecha	Fecha	Región	Idioma	Formato
	Palabras claves	Palabras claves	Palabras claves			Categoría
	Autor	Autor	Autor			
	Idioma	Idioma	Región			
	Región					
Metadatos de imágenes	Título	Título	Título	Título	Título	Título
	Url	Url	Url	Url	Url	Url
	Resumen	Resumen	Resumen	Resumen	Resumen	Resumen
	Formato	Formato	Formato	Formato	Formato	Formato
	Fecha	Fecha	Fecha	Tamaño	Fecha	Categoría
	Tamaño	Tamaño	Tamaño	Color	Tipo	
	Lenguaje	Color	Licencia	Tipo	Color	
	Autor	Tipo	Región	Diseño		
	Idioma	Diseño	Autor			
	Región	Gente				

		Licencia				
Metadatos de videos	Título	Título	Título	Título	Título	Título
	Url	Url	Url	Url	Url	Url
	Resumen	Resumen	Resumen	Resumen	Resumen	Resumen
	Formato	Formato	Formato	Fecha	Formato	Fecha
	Fecha	Fecha	Fecha	Resolución	Fecha	Categoría
	Duración	Duración	Duración	Licencia	Tamaño	Cantidad de visitas
	Director	Resolución	Director		Resolución	
	Idioma	Fuente	Subtitulado			
	Región	Precio				
	Subtitulado					

1.3.3 Arquitectura de hardware para SRI

Cambazoglu y Baeza (2015)

Proponen una arquitectura basada en una clusterización de varios nodos para rastreo, indexación y procesamiento de consultas, figura 7. A criterio de la autora de esta investigación es importante analizar esta propuesta en el diseño de la arquitectura que da soporte de hardware al procedimiento que se propone ya que la escalabilidad es uno de sus principios.

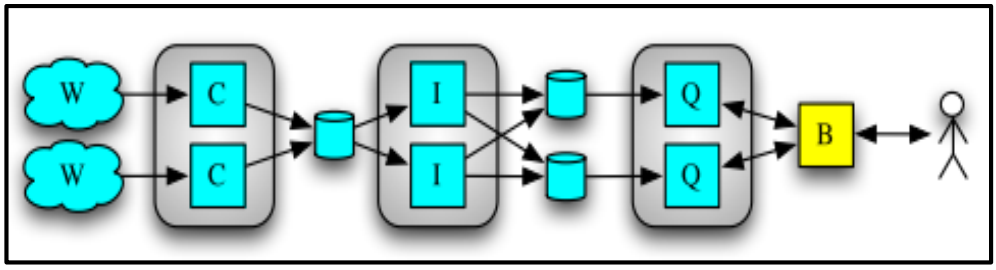


Figura 7. Arquitectura multinodo. Fuente: (Cambozaglu y Baeza, 2015).

Modelo de recuperación Google

Datos bibliográficos brindados por Benavides (2015) confirman que el tamaño exacto de los centros de datos que Google utiliza es desconocido y las cifras oficiales se mantienen poco precisas intencionadamente. Según una estimación del año 2000, la granja de servidores de Google estaba compuesta por 6000 procesadores, 12.000 discos duros IDE (dos por máquina) en cuatro centros físicos: dos en Silicon Valley y dos en Virginia. Cada centro tenía una conexión de fibra óptica de 2488 Mbit/s y otra de 622 Mbit/s. Los servidores ejecutan un software llamado Google Web Server. En la actualidad se calcula que este SRI posee desplegado en los centros de datos que tiene por todo el mundo, más de 900 000 servidores (Unocero, 2017; Rich, 2011). Sin embargo no se encontraron datos exactos en la bibliografía que evidencien la arquitectura de hardware que organiza el flujo de procesamiento de datos en todos estos servidores. Brin y Page (2012) publicaron un esquema sobre la arquitectura base de Google, figura 8, sin embargo, esta es muy genérica y

no ofrece todos los datos necesarios que sirvan de guía para el desarrollo de una arquitectura robusta para un SRI y todos sus componentes.

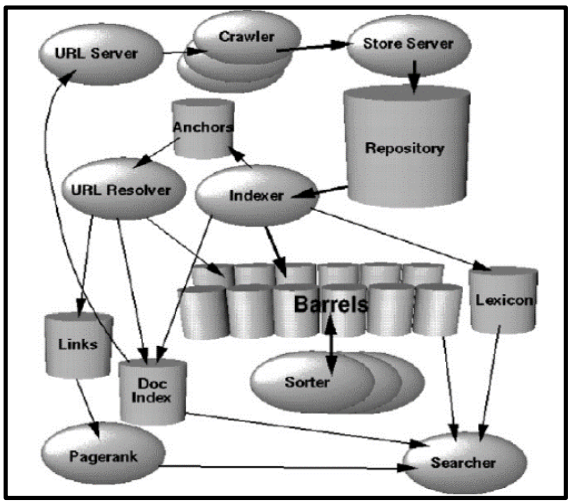


Figura 8. Arquitectura de Google. Fuente: (Brin y Page, 2012).

Pino (2014)

Propone el diseño de una arquitectura distribuible, que aunque sólo define los componentes básicos de un SRI, a criterio de esta autora está muy bien concebida. Esta investigación integra conceptos como el balanceo de carga y alta disponibilidad, figura 9.

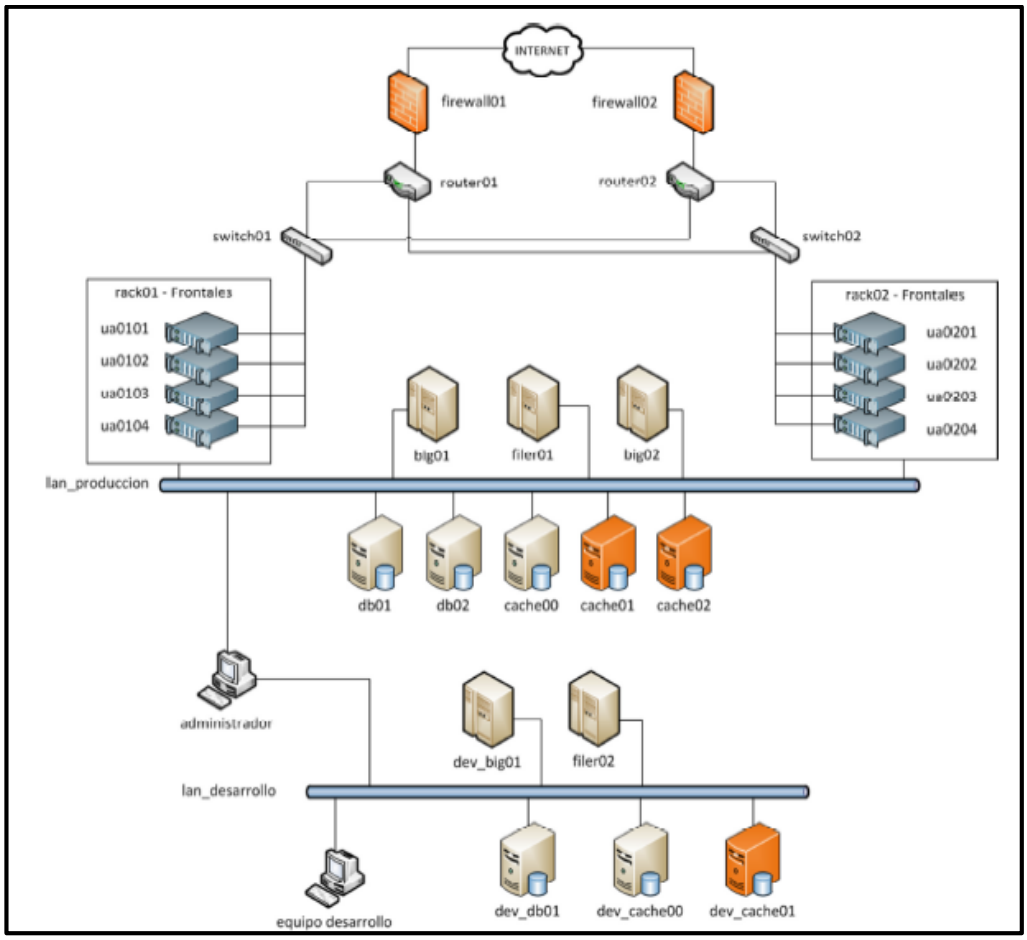


Figura 9. Arquitectura de un crawler semántico. Fuente: (Pino, 2014).

Abudaqqa y Patel, 2015

Proponen una arquitectura genérica y distribuida para SRI, figura 10, que entre sus ventajas permite a los usuarios hacer búsquedas de temas específicos que son respondidas por buscadores especializados en distintos temas, lo que permite aumentar el rendimiento del SRI al no tener que buscar las respuestas en la totalidad de los documentos indexados. Esta característica de la arquitectura propuesta, es un aspecto importante a tener en cuenta en esta investigación debido a que uno de los objetivos fundamentales del diseño de arquitecturas para sistemas, es lograr un rendimiento que permita brindar resultados en un tiempo aceptable para los usuarios.

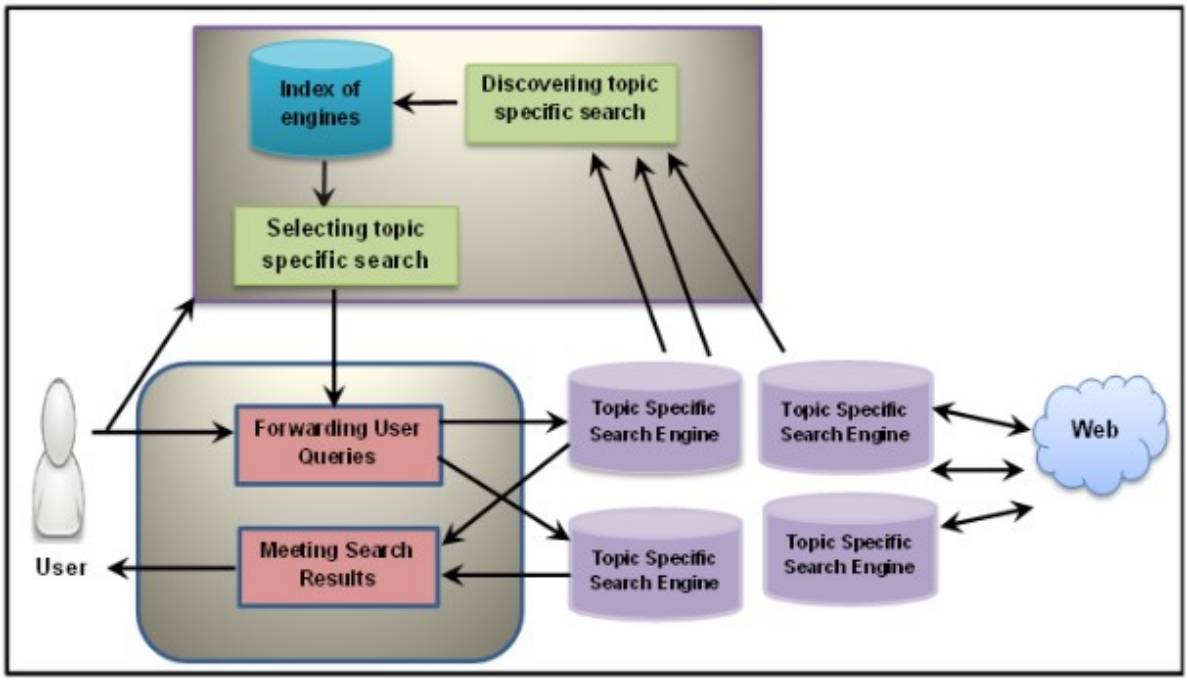


Figura 10. Arquitectura distribuida para SRI. Fuente: (Abudaqqa y Patel, 2015).

Verma y Kochar, 2016

Proponen una arquitectura multi agente para SRI basada en el procesamiento semántico que permite brindar resultados a los usuarios relacionados con sus preferencias. La arquitectura está basada en dos módulos fundamentales, el módulo agente que interviene con los usuarios y refina los resultados y el módulo descriptor semántico que convierte los textos en bloques de datos. La arquitectura multi agente complejiza el nivel de configuraciones necesarias para definir una arquitectura de hardware en un SRI y esta se especializa en el procesamiento de las preferencias de los usuarios, figura 11.

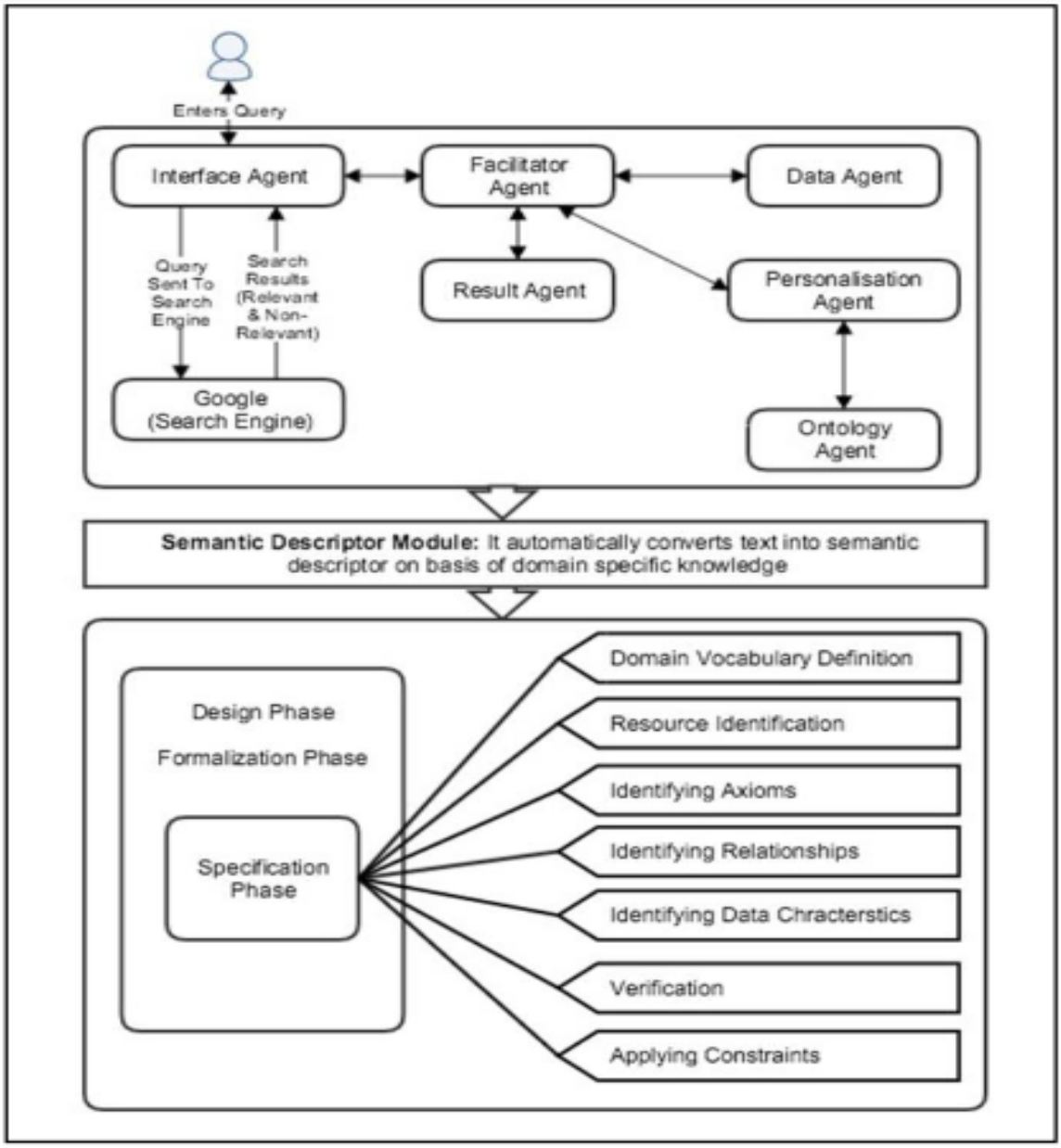


Figura 11. Arquitectura multi agente para SRI. Fuente: (Verma y Kochar, 2016).

Zhang, Yan-hong, Wei-jun y Zhong-xian, 2013

La arquitectura está compuesta por 5 componentes. El primero, recopilación distribuida de información, propone un esquema distribuido de rastreo de la información estructurado en nodos esclavos y maestros que responden a distintas tareas de rastreo de la información en la web. El componente de indexación distribuida de información, permite procesar la información rastreada, calcular el valor de PageRank de cada página, establecer el índice invertido y posteriormente almacenar los documentos de forma distribuida. El componente de recuperación distribuida de información, crea un vector de preferencias para cada usuario que permite personalizar los resultaos de las búsquedas. El componente base de datos de intereses de los usuarios, almacena toda la información relacionada con las preferencias de los usuarios. El componente interfaz de usuario, sirve de capa de abstracción entre los usuarios y los servidores brindando un interfaz que no sólo devuelve resultados sino que permite crear un mecanismo de retroalimentación en el cual los

usuarios pueden evaluar la calidad de los resultados, figura 12. Esta arquitectura resume en gran medida el funcionamiento general de rastreo, indexación y visualización de un SRI, como elemento importante para esta investigación se detecta el uso de arquitecturas distribuidas en el rastreo usando la filosofía nodos esclavos nodos maestros y aunque no brinda datos específicos sobre una arquitectura de hardware si permite hacer aproximaciones sobre la misma.

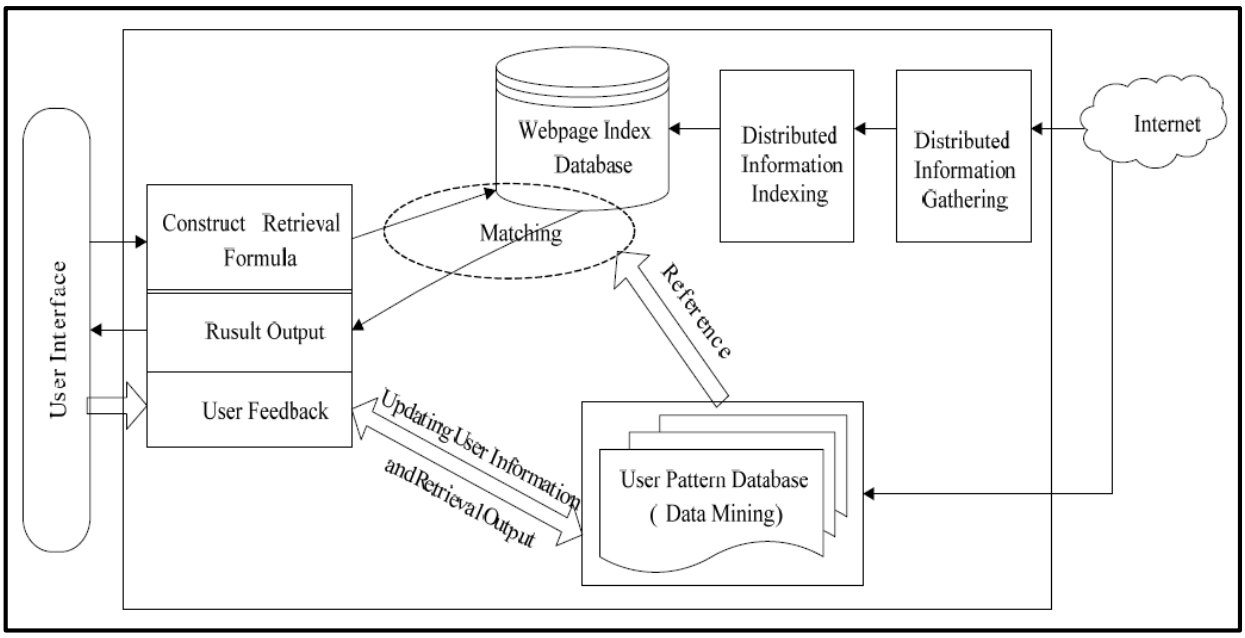


Figura 12. Arquitectura distribuida con componentes de retroalimentación del usuario. Fuente: (Zhang, Yan-hong, Wei-jun y Zhong-xian 2013).

El estudio realizado permitió detectar como las investigaciones se centran en estructuras muy generales. Existe poca evidencia de propuestas de arquitecturas de hardware adaptables para los componentes de un SRI y son de dominios muy específicos, lo que dificulta su generalización al desarrollo de buscadores. El procedimiento propuesto en esta investigación estará basado en una arquitectura distribuida sustentada por las investigaciones de Zhang et al. (2013) y Pino (2014) que permita ejecutar varios procesos de rastreo de forma paralela e incorpora mecanismos de rastreo centrado para dedicar rastreadores a contenidos específicos y así poder mejorar periódicamente los contenidos almacenados, (Pino, 2014; Zhang et al., 2013).

Conclusiones del capítulo

A pesar de que en la bibliografía no se define un procedimiento a seguir en los SRI, todos los modelos estudiados aportaron conocimientos acerca de cómo lograr una adecuada estructuración y almacenamiento de los documentos a través del preprocesamiento del texto, la utilización del modelo clásico vectorial y una arquitectura de hardware distribuida.

Los fundamentos teóricos de la RI confirman la importancia de establecer mecanismos de rastreo e indexación de la información alojada en la Web, que permitan una constante actualización de los datos recolectados; además de una correcta estructura de los documentos almacenados.

El estudio sobre las arquitecturas demostró que aunque éstos no describen una arquitectura de hardware adaptable a cualquier SRI, su análisis contribuyó a determinar que la escalabilidad, la disponibilidad y el balanceo de carga son los pilares fundamentales que debe cumplir.

El estudio de los modelos clásicos permitió definir el proceso de estructuración y representación de documentos en el modelo vectorial, ya que utiliza poco espacio de almacenamiento debido a su forma de estructurar y representar los documentos en forma de vectores.

El estudio bibliográfico de los SRI y de los estándares de metadatos permitió definir los metadatos asociados a cada tipo de contenido para lograr devolver al usuario documentos que satisfagan sus interrogantes de búsqueda.

CAPÍTULO 2: PROCEDIMIENTO PARA LA ESTRUCTURACIÓN Y ALMACENAMIENTO DE DOCUMENTOS EN EL SISTEMA DE RECUPERACIÓN DE INFORMACIÓN ORIÓN.

En el presente capítulo se presentan las etapas por las que transita el procedimiento para la estructuración y almacenamiento de documentos en el sistema de recuperación de información Orión. Se describe el flujo del procedimiento a seguir en el sistema de recuperación de información, la estructura que debe tener cada documento y la arquitectura de hardware apropiada para su correcto almacenamiento.

2.1 Estructura general del procedimiento

El procedimiento que se propone describe un proceso de RI, figura 13, que inicia con el funcionamiento del mecanismo de rastreo. Los datos de entrada de este procedimiento son las url de partida organizada en un semillero. Este semillero se actualiza con los nuevos enlaces encontrados al procesar la información rastreada, guardando en una base de datos temporales los link y un identificador hash que se le asigna a cada documento para evitar volverlos a procesar. El próximo paso es indexar el contenido para su estructuración, a través de vectores, para lograr un correcto almacenamiento de los documentos. Finalmente se almacenan todos los documentos indexados en sus respectivos núcleos. Como base del procedimiento se diseña una arquitectura de hardware distribuida para los componentes de rastreo e indexación con el fin de evitar la sobrecarga en estos servidores.

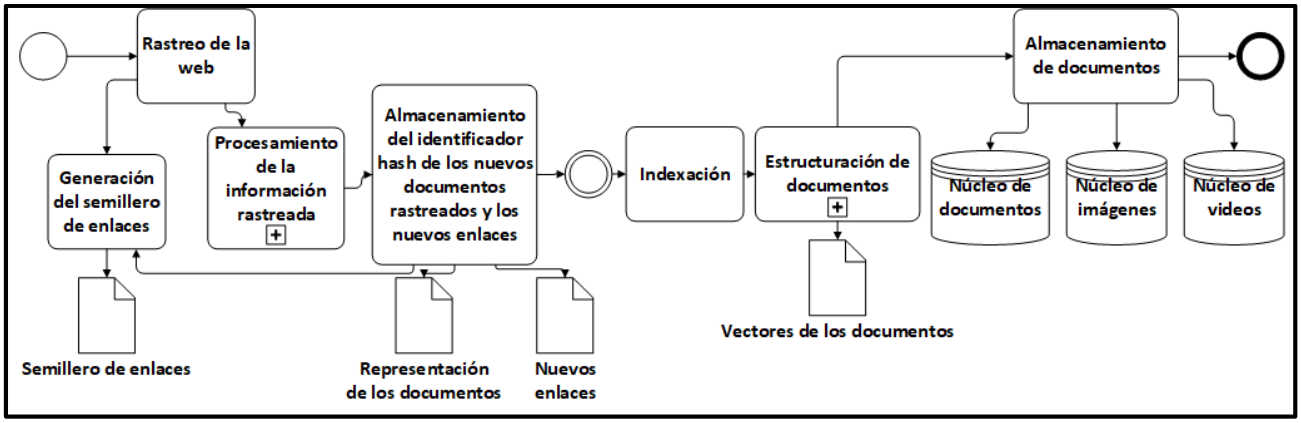


Figura 13. Flujo del procedimiento para la estructuración y almacenamiento de documentos.

En la tabla 3 se representan las 4 etapas fundamentales que integran el procedimiento de estructuración y almacenamiento de documentos, así como los roles que participan y las entradas y las salidas de cada etapa. Las dos primeras etapas responden al mecanismo de rastreo y las siguientes al mecanismo de indexación.

Tabla 3. Representación del procedimiento de estructuración y almacenamiento de documentos.

Procedimiento para la estructuración y almacenamiento de documentos			
Roles	Entradas	Etapas	Salidas
		Inicio	
Especialista en rastreo	Semillero	Rastreo de la web	Documentos rastreados
Especialista en rastreo	Documentos rastreados	Procesamiento de la información rastreada	Documentos estandarizados
Especialista en indexación	Documentos estandarizados	Estructuración de documentos	Documentos estructurados
Especialista en indexación	Documentos estructurados	Almacenamiento de documentos	
		Fin	

2.2 Mecanismo de rastreo

2.2.1 Rastreo de la web

Se utiliza como mecanismo de recolección del motor de búsqueda cubano, Nutch. Este permite recolectar información en formato PPT, DOC, PDF, HTML, XML, TXT, RTF, GIF, JPG, PNG, AVI, MP4, MKV, MPG, entre otros. Antes de comenzar el rastreo es necesario configurar cada rastreador con los que se cuenta, figura 14. Se propone utilizar tres rastreadores, donde cada uno rastrea solo un tipo de contenido (documentos, imágenes o videos).

Figura 14. Configuración de los rastreadores.

Para comenzar el rastreo se genera el semillero utilizando el comando **bin/nutch inject crawl/crawldb urls**. Este proceso cuenta con las siguientes fases:

- Inyección: Se obtienen las URLs contenidas en el semillero inicial de Nutch para el comienzo del rastreo, figura 15.

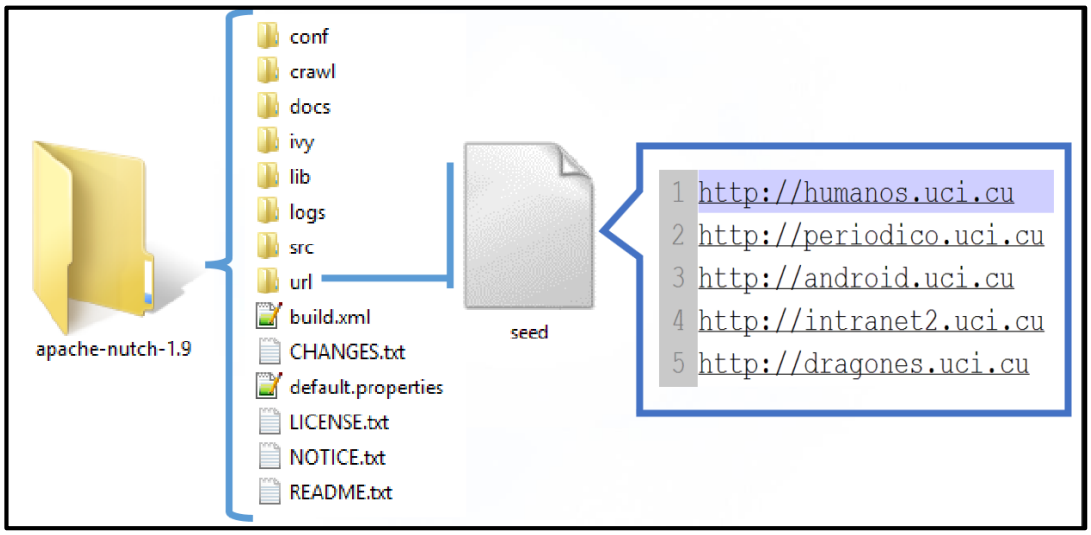


Figura 15. Fase de Inyección.

- Generación: Se genera una nueva lista de selección de URLs. En el primer rastreo solo se toman los links del semillero inicial, y en los demás se agregan los nuevos enlaces encontradas en los rastreos anteriores, figura 16.

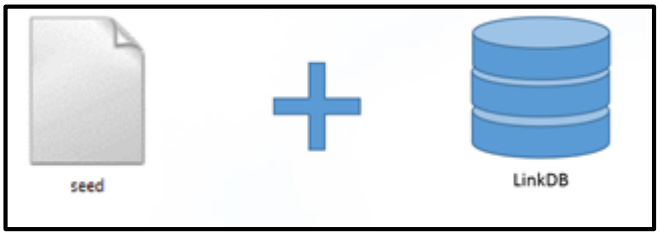


Figura 16. Fase de Generación.

- Selección: Se selecciona una de las URLs de la lista generada para establecer conexión con esa página para su descarga. Si está activa la almacena en el semillero para ser rastreada. En caso de que el sitio no esté activo se pasa esa url a una lista negra para chequearlas y cuando estén restablecidas se incorporan al conjunto de url del semillero, figura 17.

```
Public void Activas (ArrayList<String> enlaces) {

List<String> listaactiva = new ArrayList<String>();
List<String> listaactiva = new ArrayList<String>();

for ( int i= 0, i > enlaces.lenght , i++ ){

if(enlaces[i].activa == false)
Listanegra.adicionar(i,enlaces[i]);
}
else
Listaactiva. adicionar(i,enlaces[i]);
}
```

Figura 17. Fase de Selección.

2.2.2 Procesamiento de la información rastreada

Una vez descargada la página, se procede al análisis de su contenido (texto y metadatos). En esta fase se realizan las siguientes tareas, ellas son:

1. Procesamiento de tokens para reconocer la estructura, figura 18. Se descompone el texto de un documento en mínimas unidades, permitiendo la identificación de cadenas de caracteres de forma unívoca con el objetivo de facilitar la identificación de cada carácter que compone la palabra. Los elementos resultantes son utilizados posteriormente para su procesamiento.

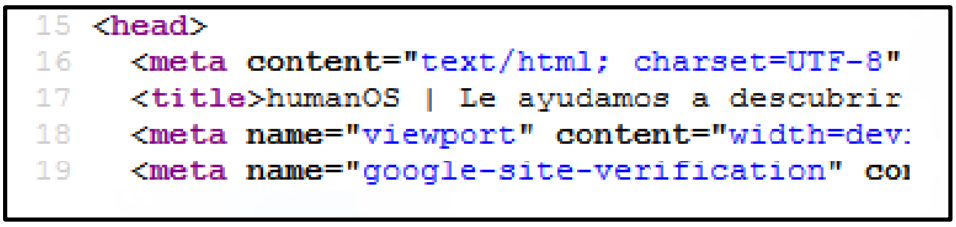


Figura 18. Fase de procesamiento, reconocimiento de estructura.

2. Conversión de los documentos a texto plano.
3. Identificación de metadatos asociados a cada tipo de documentos, figura 19.

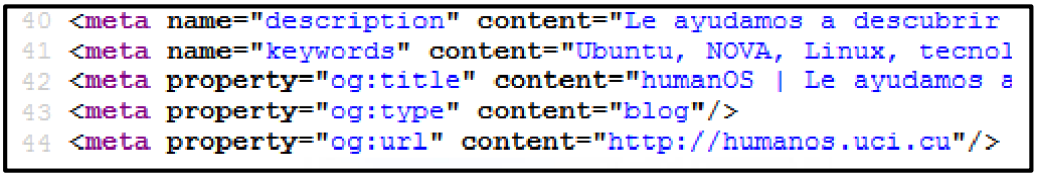


Figura 19. Fase de procesamiento, identificación de metadatos.

En el rastreo se aplica un proceso de estandarización, basado en la propuesta del estándar Dublin Core, que permite definir una estructura correcta para cada uno de los documentos y evita almacenar información que carezca de utilidad. En esta propuesta se definen un grupo de metadatos necesarios a partir del estudio realizado a los SRI a nivel internacional y otro grupo de metadatos adicionales que proporcionan información más abundante del recurso digital que se rastrea, tabla 4:

Tabla 4. Metadatos necesarios para cada tipo de documento.

Documentos	**Imágenes**	**Videos**
Imprescindibles		
Título	Título	Título
Url	Url	Url
Resumen	Resumen	Resumen
Fecha	Fecha	Fecha
Formato	Tamaño	Duración
	Formato	Formato
Adicionales		

Palabras claves	Lenguaje	Subtitulado
Lenguaje	Región	Tamaño
Sitio	Autor	Director
Autor	Categoría	Categoría
Categoría		

Título: Se refiere al título que lleva por nombre el documento, la imagen o los videos.

Url: Se refiere a la dirección electrónica de origen a la que está adscrito el material

Resumen: En este campo se hace un breve resumen sobre el contenido del objeto digital.

Fecha: Se anota la fecha de elaboración del registro.

Formato: En este campo se registra el tipo de extensión con que se presenta el objeto digital, ya sea HTML, JPG, GIFF, PDF, AVI.

Palabras claves: Las palabras que mayor relevancia tienen en el documento.

Lenguaje: En este campo se establecen las siglas correspondientes al idioma en que se presenta la publicación.

Sitio: El sitio de donde proviene el documento.

Autor: Se escribe al autor intelectual del documento digital.

Tamaño: En este campo se especifica el tamaño de la imagen.

Región: Se define el lugar de procedencia de la imagen.

Duración: En este campo se especifica la duración del video.

Subtitulado: Se puntualiza si el video está o no subtitulado.

Director: Director del material.

Categoría: En este campo se define la categoría a la que pertenece el archivo digital (cultura, deporte, naturaleza, entre otras).

Se pueden añadir más metadatos dependiendo de los tipos de recursos que indexe el SRI, pero es indispensable garantizar que cada documento tenga en su estructura los metadatos propuestos en el proceso de estandarización. Si se rastrean documentos que no cuentan con los metadatos definidos como imprescindibles, el sistema no los tomará en cuenta para evitar devolverle al usuario documentos incompletos que no aporten ningún valor a su búsqueda.

4. Identificación de nuevos enlaces a visitar, figura 20.

```
<a href="http://humanos.uci.cu/2015/07/como-se-hace-usando-facebook-messenger-en
<a href="/tag/messenger/" rel="tag">Messenger</a>,
<a href="/tag/pidgin/" rel="tag">Pidgin</a>
```

Figura 20. Fase de procesamiento, identificación de enlaces.

Al terminar el procesamiento, Nutch actualiza el semillero con los nuevos enlaces encontrados y elabora un identificador hash de los nuevos documentos rastreados procesando el título, resumen y el contenido de los mismos y finalmente envía toda la información a la aplicación que se encargará

de indizar todo el contenido, para su posterior consulta.

Cualidades del rastreo:

Las cualidades necesarias que posee el mecanismo de rastreo son las siguientes:

Rendimiento: el rastreador deber hacer uso eficiente de los recursos de hardware disponibles, organizando el proceso de rastreo en un grupo de servidores.

Escalabilidad: el rastreador debe estar diseñado de tal manera que permita aumentar la velocidad de rastreo, añadiendo servidores extras y ancho de banda.

Calidad: el rastreador debe ser capaz de identificar páginas correctamente estructuradas favoreciendo su indexación y su continua actualización; además debe poseer la habilidad de identificar contenidos duplicados.

Actualización: se debe garantizar que el SRI tenga almacenado las versiones más actuales de los documentos.

Extensibilidad: debe tener un diseño modular que permita la adición de nuevos componentes y funcionalidades.

2.3 Mecanismo de indexación

El mecanismo de indexación tiene como base la herramienta Apache Solr. A través del comando **bin/crawl urls crawl http://dirección:8080/solr**, Nucth le envía toda la información rastreada a Solr. Previamente a indexar documentos en Solr es necesario definir los campos que conforman los documentos que se indexarán y especificar el tipo de dato de cada campo y su identificador. Esta configuración se define en el archivo schema.xml, figura 21.

Figura 21. Configuración del servidor de indexación.

2.3.1 Estructuración de documentos

Con el objetivo de estructurar los documentos para su posterior almacenamiento se definen dos etapas principales:

1. **Preprocesamiento del texto**

Cuando se procesa el texto en lenguaje natural no todas las palabras aportan información válida para la representación de un documento. Por esta causa, el sistema de RI procesa el texto de los documentos y determina la mayoría de los términos importantes, siendo estos los de mayor poder discriminativo debido a sus raras ocurrencias en los documentos de la colección. En la figura 22 se esboza la fase de preprocesamiento tomando como entrada un documento.

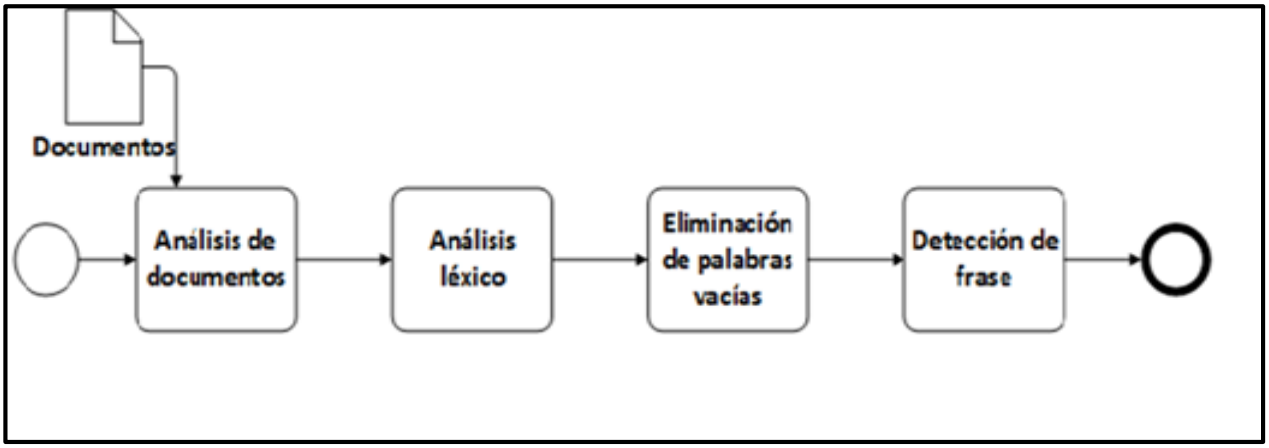


Figura 22. Procesamiento de texto en un sistema de RI. Fuente: (Ceri et al., 2013)

- **Análisis de documentos.**

El análisis de documentos se ocupa del reconocimiento y la descomposición de la estructura del documento en componentes individuales (tokens). Este proceso permite la identificación de cadenas de caracteres de forma unívoca, de cara a posteriores tratamientos de depuración. Se pueden utilizar diferentes filtros para lograr la descomposición del documento como son:

<filter class= “solr. **StandardTokenizer**Factory”/>: Parsea de forma inteligente el texto generando tokens en espacios y signos de puntuación.

<filter class= “solr. **WhiteSpaceTokenizer**Factory”/>: Genera tokens a partir de los espacios presentes.

- **Análisis léxico.**

El análisis léxico se encarga de la identificación correcta de acentos, abreviaturas, fechas y casos. La dificultad de esta operación depende mucho del idioma en cuestión: por ejemplo, el idioma inglés no tiene diacríticos ni casos, el francés tiene signos diacríticos pero ningún caso, el alemán tiene diacríticos y casos. Se encarga además, de la conversión a minúsculas, eliminación de signos y acentos y la transliteración, remplazo de caracteres especiales y normalización a minúsculas. Algunos de los filtros utilizados son:

Normalización a minúsculas: <filter class= “solr. **LowerCase**FilterFactory”/>

Sustitución a sinónimos: <filter class= “solr. **Synonym**FilterFactory”/>

Remplazo de caracteres: <filter class= “solr. **PatternRemplace**FilterFactory”/>

- **Eliminación de palabras vacías**

Una etapa posterior opcionalmente aplicada a los resultados del análisis léxico es la eliminación de la palabra de terminación, es decir, la eliminación de las palabras de alta frecuencia (pronombres,

partículas interrogativas y ciertas preposiciones), utilizando el filtro <filter class= “solr. **Stop**FilterFactory”/>.

- **Detección de Frase**

Este proceso depura todos los términos de un texto, reduciendo su número de caracteres, simplificando su forma original, género, número, desinencia en una forma de palabra más común o normalizada; debido a que la mayor parte de ellas tienen el mismo significado semántico. Además se encarga de eliminar los prefijos o sufijos de palabras para normalizarla. Este es un proceso que normalmente usa diccionarios y análisis morfológicos de palabras para devolver la forma básica y de ese modo colapsa sus formas. Uno de los filtros más utilizados con este objetivo es: <filter class= “solr. **PorterStem**FilterFactory”/>

2. Vectorización

En este momento se tienen todos los términos candidatos a formar parte de la base de datos documental. El siguiente paso consistirá en determinar la importancia de cada término, de tal forma que si es lo suficientemente importante se escogerá para ser incluido en el conjunto de términos finales. Para ello se pondera el término calculando el peso TF-IDF propuesto en el Modelo Vectorial y explicado en la sección 1.2.3 para calcular el peso de cada término.

2.3.2 Almacenamiento de documentos

Al concluir la estructuración de documentos se almacena en cada núcleo los documentos relacionados con el tipo de información que estos contienen (documentos, imágenes y videos). De ellos se guarda el vector y los metadatos que describen al documento.

2.4 Arquitectura de hardware

Se propone una arquitectura distribuida como base del flujo de procedimiento. Para la definición de la estructura y los recursos de hardware se analizaron tres variables fundamentales:

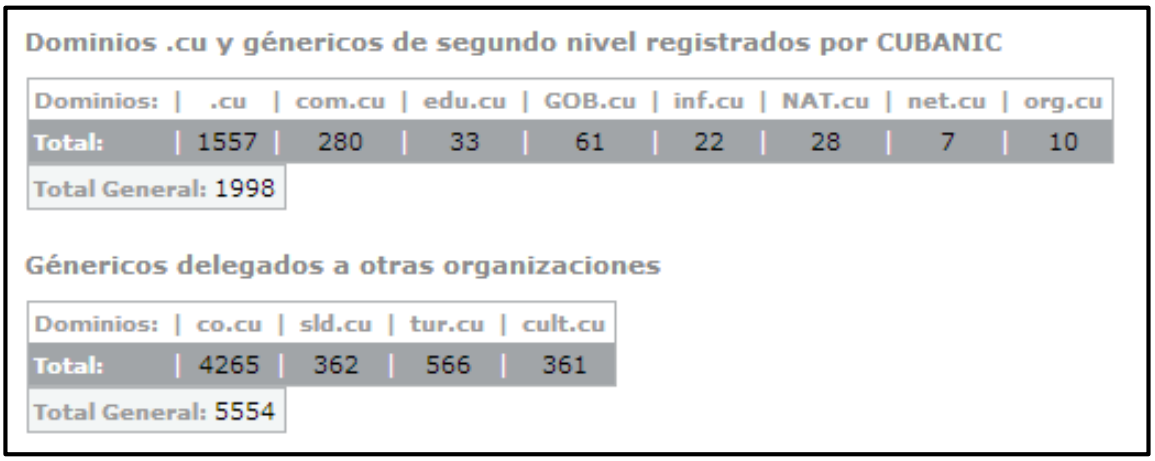
Disponibilidad: la disponibilidad es la cualidad de un sistema para mantenerse operativo principalmente ante contingencias (Antiñanco, 2014).

Balanceo de carga: clúster que permite que un conjunto de servidores compartan la carga del trabajo y del tráfico a sus clientes. Está compuesto por uno o más ordenadores llamados nodos, que actúan como front-end del clúster y se ocupan de repartir las peticiones de servicio que reciba el clúster a los otros ordenadores que forman su back-end (Sinisterra et al, 2012).

Respaldo: el respaldo de datos según Reinoso y Paulina (2007), es efectuar una copia de todos o algunos archivos que se encuentran en el medio de almacenamiento de una o varias computadoras, servidores o en otros medios diferentes, para poder recuperarlos en otro momento si se pierden o se dañan los archivos originales. Este proceso debe ser fiable, eficaz y robusto, debido a que cualquier error durante el proceso puede provocar que los datos se deterioren.

Para una correcta definición de recursos de hardware se debe hacer un estudio previo de la web para estimar su tamaño y la cantidad de usuarios potenciales que accederán al SRI. Este estudio permitirá definir variables importantes en el diseño de la arquitectura como: gigas de

almacenamiento, memoria RAM, memoria para el procesamiento de datos y organización y número de servidores para cada componente. El diseño de arquitectura distribuida que se propone fue creado tomando como objetivo la web cubana, según estadísticas extraídas del sitio CUBANIC posee un total de 6861 dominios bajo (.cu) organizados como se muestra en la figura 23.



Dominios .cu y génericos de segundo nivel registrados por CUBANIC

Dominios:	.cu	com.cu	edu.cu	GOB.cu	inf.cu	NAT.cu	net.cu	org.cu
Total:	1557	280	33	61	22	28	7	10

Total General: 1998

Génericos delegados a otras organizaciones

Dominios:	co.cu	sld.cu	tur.cu	cult.cu
Total:	4265	362	566	361

Total General: 5554

Figura 23. Estadísticas de la web cubana. Fuente: (Cubanic, 2018).

En la figura 24 se muestra la distribución de los servidores que dan soporte al procedimiento propuesto. Los servidores que ejecutan el proceso rastreo deben organizarse de forma tal que permitan el uso eficiente de los recursos de hardware con los que se cuente. El balanceo de la carga del proceso de inserción de datos en los servidores de indexación, se ejecuta a través de un balanceador de carga. Este es el encargado de comprobar el nivel de carga de los servidores de indexación maestros y decidir cuál de ellos atenderá la petición de inserción de datos. A su vez los servidores indexadores maestros se encargan de recibir el flujo de datos entrantes como resultado del rastreo y los replican en los servidores de indexación esclavos de acuerdo al tipo de información que este almacene, garantizando que en todo momento estos servidores puedan darle respuesta a las peticiones de los usuarios. Los servidores indexadores esclavos son los encargados de almacenar toda la información referente a los documentos indexados y responder a las peticiones ejecutadas a través de las interfaces web con el listado de documentos que satisfacen las consultas de los usuarios.

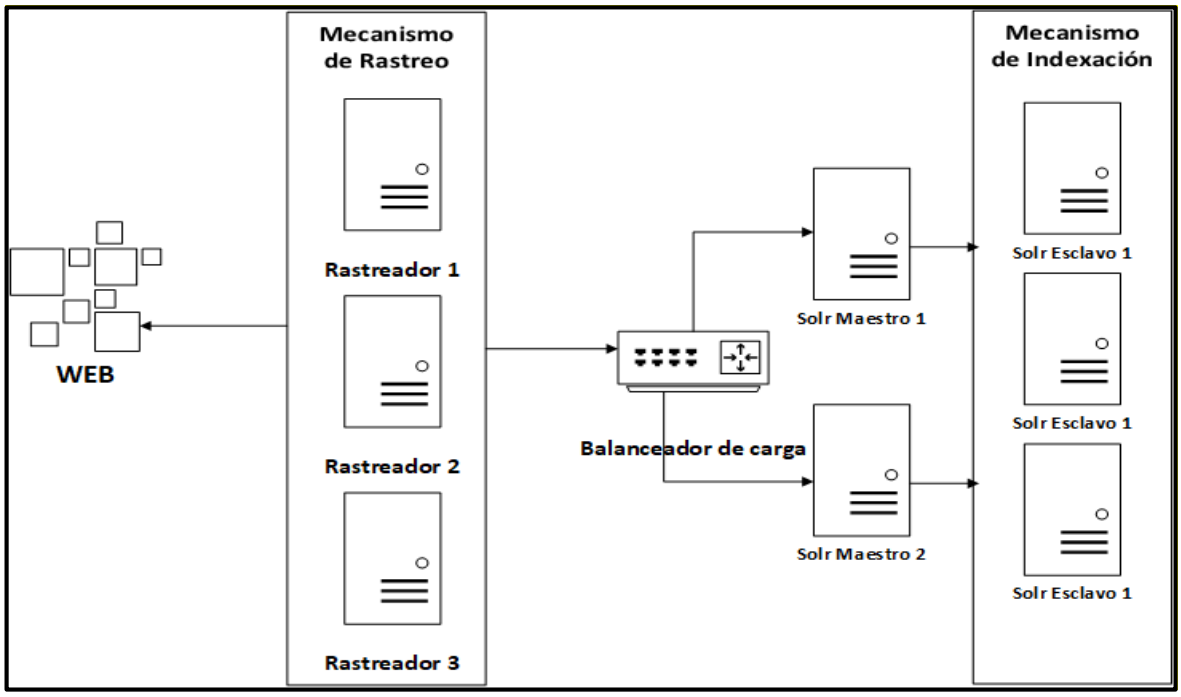


Figura 24. Distribución de servidores que brindan soporte al procedimiento.

Después de diseñada y desplegada la arquitectura, se debe realizar un monitoreo constante del estado de cada uno de los servidores. Esta actividad permitirá responder de forma rápida ante cualquier fallo. El proceso de evaluación del rendimiento permite establecer un criterio de evaluación sobre el estado de los recursos. En caso de que la evaluación sea negativa (almacenamiento insuficiente, sobrecarga en el procesamiento de los servidores, memoria RAM insuficiente, respuestas lentas a los usuarios) se debe focalizar dónde se encuentra el problema y tomar las medidas pertinentes; lo que podría traducirse en una redefinición y despliegue de recursos para mejorar el rendimiento del sistema.

2.5 Roles involucrados y responsabilidades

Especialista en rastreo:

Es el encargado de administrar todo el proceso de rastreo de la web. Además mantiene un monitoreo constante al estado de cada uno de los servidores de rastreo para evitar cualquier sobrecarga de procesamiento y solucionar los problemas en tiempo real que se presenten.

Habilidades:

- Estimación de recursos de hardware necesarios para el rastreo de la web.
- Configuración del entorno de rastreo de manera que se utilicen los recursos de hardware disponibles de la forma más eficiente.
- Evaluación del flujo de rastreo de la web.

Especialista en indexación:

Es el encargado de administrar todo el proceso de indexación velando por la correcta estructuración de los documentos y realizando verificaciones al proceso de estandarización de los mismos. Además debe mantener un monitoreo constante al crecimiento de la colección de documentos para en caso de ser necesario, aumentar el número de servidores de almacenamiento.

Habilidades:

- Estimación de recursos de hardware necesarios para la indexación de los recursos rastreados.
- Configuración del entorno de almacenamiento de documentos de manera que se utilicen los recursos de hardware disponibles de la forma más eficiente.
- Evaluación del flujo de indexación de documentos.

Conclusiones del capítulo

El procedimiento diseñado está basado en técnicas de rastreo e indexación que facilitan la correcta estructuración, almacenamiento y actualización de documentos en el SRI Orión.

La arquitectura de hardware propuesta para dar soporte al procedimiento integra conceptos de balanceo de carga y respaldo de la información con el fin de mejorar el rendimiento de los servidores.

La definición de los roles propuestos cubren las responsabilidades necesarias para ejecutar el flujo de cada una de las etapas del procedimiento.

CAPÍTULO 3: VALIDACIÓN DEL PROCEDIMIENTO PARA LA ESTRUCTURACIÓN Y ALMACENAMIENTO DE DOCUMENTOS EN EL SISTEMA DE RECUPERACIÓN DE INFORMACIÓN ORIÓN.

En este capítulo se realizan pruebas al procedimiento para la estructuración y almacenamiento de documentos en el sistema de recuperación de información Orión. Se realiza un experimento para validar la eficacia del almacenamiento y la calidad de los resultados brindados a los usuarios calculando las métricas exhaustividad, precisión y eficiencia. Se hace uso de la técnica de Iadov para medir la satisfacción de potenciales usuarios y el escalamiento de Likert para validar la completitud de los documentos teniendo en cuenta el criterio de expertos en el campo de acción. Finalmente se ejecuta una triangulación metodológica para demostrar la confiabilidad de dichos resultados.

3.1 Estrategia de validación del procedimiento propuesto

La estrategia para validar el modelo tiene como fundamento los siguientes aspectos:

- Estudios bibliográficos relacionados con el objeto de estudio, específicamente los referidos a la validación de modelos de RI (Avedaño et al., 2013; Blázquez, 2013; Roa et al., 2013; Jaramillo et al., 2014; Romá, 2014; García, 2015; Quiñones, 2015; Zhai et al., 2015; Nasution et al., 2016; Baeza-Yates y Liaghat, 2017).
- Experiencias obtenidas en la lectura de las tesis defendidas en el tribunal de automática y computación en los últimos años.
- Experiencias en temas de recuperación de información de varios equipos destacados en el desarrollo de SRI a nivel nacional.

Incluye técnicas cualitativas y cuantitativas organizadas en las siguientes tareas:

- Realización de un experimento para demostrar la capacidad del procedimiento de mejorar la calidad de los resultados brindados a los usuarios, calculando las métricas precisión y exhaustividad.
- Pruebas antes y después de aplicar el procedimiento a la arquitectura del SRI Motor de Búsqueda Orión, para evaluar la eficacia del almacenamiento y eficiencia de este buscador.
- Aplicación de la técnica Iadov para evaluar el nivel de satisfacción de usuarios potenciales que se benefician con la aplicación del procedimiento.
- Aplicación del escalamiento de Likert para validar el criterio de los expertos sobre la completitud de los datos después de ser procesados.
- Ejecución de una triangulación metodológica de las técnicas utilizadas para disminuir el sesgo que se produce al comparar resultados obtenidos en la cuantificación de variables.

3.2 Experimentación

3.2.1 Experimento para medir la eficacia del almacenamiento y la eficiencia del buscador Orión

El primer paso del experimento consiste en aplicar las pruebas a la arquitectura del Motor de Búsqueda Orión sin aplicar los cambios que propone el procedimiento para así determinar el estado de los siguientes indicadores (Farias et al., 2016; Mainegra et al., 2016):

- Eficiencia
  - cantidad de documentos rastreados en cada ronda de rastreo.
  - velocidad de respuesta a las peticiones de los usuarios.
  - respaldo de la información almacenada.
  - balanceo de carga.
- Eficacia
  - uso del espacio correcto en discos para el almacenamiento de la información.

Luego se aplican los cambios que propone el procedimiento a la distribución de los servidores y se obtienen los valores de las variables definidas para efectuar una comparación y llegar a conclusiones. Las pruebas a los servidores de interfaces web se hicieron emulando 200 usuarios, realizando peticiones con la herramienta Jmeter, que es un software especializado en pruebas de carga y estrés. Para visualizar el estado de los servidores se utilizaron las herramientas Grafana y Telegraf para la interpretación de los registros de los servidores y la visualización del consumo de los recursos.

Los resultados obtenidos en la primera etapa de la prueba se muestran en la tabla 5. La interpretación del valor de los indicadores y el estudio de la distribución de los servidores, arrojó las siguientes conclusiones:

- Los semilleros en cada servidor de rastreo eran insuficientes con respecto al volumen de enlaces totales que posee la web cubana. El promedio de enlaces era de 30; esto ocasiona que la cantidad de documentos rastreados en cada ronda fuera insuficiente.
- No existe balanceo de carga entre los servidores de rastreo y los servidores de indexación, por lo que mientras algunos servidores de indexación no respondían peticiones, otros en cambio, atendían todo el flujo proveniente de las interfaces.
- No existen mecanismos de respaldo de la información contemplados en la arquitectura.
- El promedio de tiempo de respuestas a los usuarios se considera lento, debido a que la cantidad de documentos almacenada no era muy significativa y el tiempo de respuesta promediaba los 2,5 s.

Tabla 5. Valor de indicadores medidos en la prueba.

Servidores	**RAM**	**CPU**	**ALMACENAMIENTO**
Rastreo 1	87%	83%	Aproximadamente 2500 documentos en cada ronda de rastreo.
Rastreo 2	80%	75%	
Rastreo 3	90%	58%	
Indexador maestro	90%	95%	
Indexador esclavo 1	35%	40%	
Indexador esclavo 2	40%	42%	
			Tiempo de respuesta
Servidores Web	50%	40%	2,5s de tiempo de respuesta promedio.

Luego de un estudio de la estructura del buscador Orión, se decide aplicar una serie de modificaciones en base a la distribución de hardware que propone el procedimiento para los componentes de rastreo e indexación; utilizando los recursos de RAM, CPU y almacenamiento disponibles en este SRI, tabla 6.

Tabla 6. Valores obtenidos al aplicar la prueba después de modificada la distribución de los servidores.

Servidores	**RAM**	**CPU**	**ALMACENAMIENTO**
Rastreo 1	70%	69%	Aproximadamente 7 000 documentos en cada ronda de rastreo.
Rastreo 2	80%	70%	
Rastreo 3	85%	58%	
Indexador maestro 1	70%	80%	
Indexador maestro 2	75%	83%	
Indexador esclavo 1	45%	45%	
Indexador esclavo 2	45%	42%	

Indexador esclavo 3	45%	44%	
			Tiempo de respuesta
Servidores Web 1	51%	40%	**1,5s** de tiempo de respuesta promedio

Con el análisis del nuevo diseño de arquitectura se concluye que al aplicar la distribución de servidores propuesta en el procedimiento, se solucionan los problemas detectados con anterioridad, relacionados con el balanceo de la carga, uso ineficiente del almacenamiento y detección de cuellos de botella. Esto se logra con el uso de los balanceadores de carga entre los rastreadores y los indexadores. Además creando un mecanismo de réplica de la información desde los servidores maestros de indexación hasta los indexadores esclavos. La adición de un nuevo balanceador entre los servidores de rastreo e indexación permite eliminar el cuello de botella que anteriormente existía en ese punto.

Sobre los tiempos de respuestas se concluye que después de aplicar las modificaciones propuestas en el procedimiento, a pesar de que el flujo de inserción de datos aumentó en el segundo experimento y la cantidad de documentos almacenados es mayor, el tiempo de respuestas es menor en un segundo que en el primer experimento, figura 25.

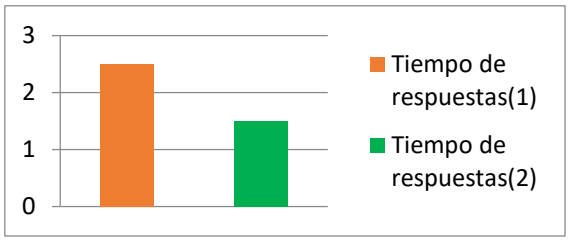


Figura 25. Comparación de tiempo de respuestas.

Sobre la cantidad de documentos indexados, se concluye que con la nueva distribución de servidores se mejora en este aspecto; ya que los valores de almacenamiento de documentos en el segundo experimento duplican a los valores obtenidos en el primero, figura 26. Los resultados obtenidos demuestran que el procedimiento mejora la eficiencia y eficacia de la arquitectura del SRI Motor de Búsqueda Orión.

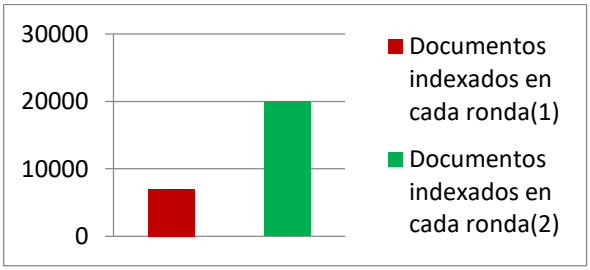


Figura 26. Comparación del promedio de documentos insertados en cada ronda.

3.2.2 Experimento para demostrar la capacidad del procedimiento de mejorar la calidad de los resultados brindados a los usuarios

Se realiza un experimento para evaluar el comportamiento de las métricas precisión y exhaustividad definidas por (Avedaño et al., 2013; Roa et al., 2013; Jaramillo et al., 2014; Romá, 2014; Quiñones, 2015; García, 2015) y compararlas antes y después de aplicar el procedimiento en el Motor de Búsqueda Orión. Para llevar a cabo el experimento, se escogieron 25 usuarios que interactúan periódicamente con SRI. Cada usuario definió su perfil de búsqueda y seleccionó una consulta de búsqueda, de 4 propuestas en una planilla (Anexo 1). Luego se creó una colección de 100 documentos ya estructurados y se seleccionaron los documentos relevantes en relación a las consultas seleccionadas por los usuarios y su perfil de búsqueda; resultando en 25 listas de documentos relevantes. El paso final del experimento consistió en que los usuarios insertaran las consultas definidas y seleccionaran los documentos que a su criterio eran relevantes. Con estos valores obtenidos se procede a calcular las métricas seleccionadas utilizando las siguientes ecuaciones (Baeza y Ribeiro, 1999; Jaramillo et al., 2014; Zhai et al., 2015; Büttcher et al., 2016; Nasution et al., 2016; Buckley, 2017; Baeza-Yates y Liaghat, 2017):

$$\text{Precisión} = \frac{\text{Documentos relevantes recuperados}}{\text{Documentos recuperados}} \quad (4)$$

$$\text{Exhaustividad} = \frac{\text{Documentos relevantes recuperados}}{\text{documentos relevantes}} \quad (5)$$

El cálculo de los valores de precisión, fórmula 4, se ejecutó sobre un total de 50 documentos de la colección definida. El cálculo de los valores de exhaustividad, fórmula 5, se ejecutó sobre la totalidad de documentos de la colección definida. Los resultados obtenidos después de promediar los valores de precisión y exhaustividad antes y después, se muestran en la tabla 7:

Tabla 7. Valores de precisión y exhaustividad obtenidos.

Precisión promedio antes de aplicar el procedimiento	**Precisión promedio después de aplicar el procedimiento**	**Exhaustividad promedio antes de aplicar el procedimiento**	**Exhaustividad promedio después de aplicar el procedimiento**
0,2964	**0,8008**	0,2468	**0,606**

Para comparar los valores de precisión y exhaustividad antes y después de aplicar el procedimiento, se utilizó el editor estadístico STATGRAPHICS. Para comprobar si lo datos se ajustan a una distribución normal se realizó la prueba de normalidad Shafiro-Wilk. El valor de $p > 0,05$ en los dos casos (valores de precisión y valores de exhaustividad), demostró que no existen problemas con la normalidad de los datos. En ambos casos los valores del sesgo estandarizado y de curtosis estandarizada, se encuentran dentro del rango esperado para datos provenientes de una

distribución normal. La aplicación de la prueba estadística T-student permitió comparar las medias obtenidas antes y después. Los valores de $p < 0,5$ obtenidos permiten rechazar la h_0 que asegura la igualdad de medias. En consecuencia queda demostrado que existe una diferencia significativa entre las medias comparadas en ambos casos (precisión antes y después y exhaustividad antes y después). De esta manera se comprueba satisfactoriamente que el procedimiento propuesto para la estructuración de los documentos mejora los resultados brindados a los usuarios en el SRI.

3.2 Valoración de los expertos sobre el procedimiento

Para obtener opiniones certeras de expertos en el tema de la RI sobre la completitud de los documentos que se almacenan en el SRI Orión después de aplicado el procedimiento, se aplica el escalamiento de Likert (Likert, 1932) a través de un cuestionario. Los aspectos a valorar son los siguientes:

1. Los documentos, imágenes y videos contienen todos los metadatos considerados imprescindibles en la propuesta.
2. Estructura general del procedimiento

3.2.1 Proceso de selección de expertos

En el proceso de selección de expertos se consultó un número de especialistas con experiencia en el campo de la recuperación de información. El primer paso de este procedimiento fue evaluar sus conocimientos sobre técnicas de RI y la estructuración y almacenamiento de documentos. El grupo de expertos seleccionados pertenecen a las siguientes instituciones:

- Departamento de Soluciones Informáticas para Internet (UCI).
- Unión de Informáticos de Cuba (UIC).

La muestra inicial fue de 30 expertos. A cada candidato se le aplicó una encuesta (Anexo 2) para determinar su coeficiente de competencia (Anexo 3), donde 19 expertos obtuvieron un coeficiente alto, 5 de ellos medio y 6 bajo, siendo seleccionados como expertos para validar el procedimiento, los 24 con coeficiente medio o alto. Los resultados de la distribución de los expertos según su coeficiente de competencia se muestran en la tabla 8.

Tabla 8. Distribución de expertos según coeficiente de competencia.

Nivel de competencia	**Cantidad**	**Por ciento**
Alto	19	63.3 %
Medio	5	16.7 %
Bajo	6	20 %
Total	**30**	**100**

3.2.2 Aplicación del escalamiento de Likert

Las preguntas del cuestionario aplicado (Anexo 4) están diseñadas para obtener valoraciones de los expertos en función de los aspectos definidos como ejes de evaluación. Cada experto expresa su valoración del procedimiento propuesto mediante la escala que se define a continuación:

5- muy de acuerdo (MA), 4- de acuerdo (DA), 3- ni de acuerdo ni en desacuerdo (Sí-No), 2- en desacuerdo (ED) y 1- completamente en desacuerdo (CD).

El procesamiento de los datos se realiza mediante la escala de Likert. Para el procesamiento de los resultados, se empleó un método que consiste en identificar la frecuencia en cada categoría de la escala de Likert definida en la encuesta realizada y se calculan los por cientos de concordancia de cada categoría de acuerdo a las características propuestas por el autor. Luego se calcula en un índice porcentual (IP), fórmula 6, que integra en un solo valor, la aceptación del grupo de evaluadores sobre las características del modelo (Orellana, 2017).

$$\mathrm{IP} = \frac{5(\%\ \text{de MA}) + 4(\%\ \text{de DA}) + 3(\%\ \text{de SI}-\text{NO}) + 2(\%\ \text{de ED}) + 1(\%\ \text{de CD})}{5} \qquad (6)$$

Los resultados obtenidos al aplicar el escalamiento de Likert (IP > 89 en todos los casos), demuestran que cada tipo de documentos que se indexaron en el SRI Orión cumple el estándar definido en el procedimiento y la estructura general del mismo tienen una alta valoración por parte de los expertos. A lo largo del proceso de evaluación se recogieron criterios positivos para el uso y aplicación del procedimiento para la estructuración y almacenamiento de documentos en el SRI Orión.

3.3 Satisfacción de potenciales usuarios con el procedimiento

Según Felipe (2017) esta técnica se basa en la aplicación de un cuestionario que tiene una estructura interna determinada, que sigue una relación entre tres preguntas cerradas y un análisis posterior de otro conjunto de preguntas abiertas. La relación entre las preguntas cerradas se establece a través del denominado Cuadro Lógico de Iadov; el cual posibilita determinar posteriormente el nivel de satisfacción del usuario y del grupo. Para aplicar el procedimiento se debe establecer una escala de satisfacción que responde a la siguiente estructura:

(1) Clara satisfacción, (2) Más satisfecho que insatisfecho, (3) No definida, (4) Más insatisfecho que satisfecho, (5) Clara insatisfacción y (6) Contradictoria.

Luego de aplicado el cuestionario y haber triangulado las preguntas cerradas, (Anexo 5), el número resultante de la interrelación de las tres preguntas cerradas indica la posición de cada cual en dicha escala de satisfacción. El resultado final de esta técnica es el índice de satisfacción grupal (ISG), que refleja el grado de satisfacción de los encuestados. Para ponderar el ISG, fórmula 7, se establece una escala numérica entre +1 y -1 (Orellana, 2017).

(+1) Máximo de satisfacción, (+0.5) Más satisfecho que insatisfecho, (0) No definido y contradictorio, (-0.5) Más insatisfecho que satisfecho y (-1) Máxima insatisfacción.

El cálculo del ISG se realiza mediante la siguiente fórmula:

$$\mathrm{ISG} = \frac{\boldsymbol{A}(+1) + \boldsymbol{B}(+0{,}5) + \boldsymbol{C}(0) + \boldsymbol{D}(-0{,}5) + \boldsymbol{E}(-1)}{\boldsymbol{N}} \qquad (7)$$

Variables:

N: cantidad de usuarios encuestados

A: cantidad de usuarios con Clara satisfacción

B: cantidad de usuarios Más satisfechos que insatisfechos

C: cantidad de usuarios No definidos

D: cantidad de usuarios Más insatisfechos que satisfechos

E: cantidad de usuarios con Clara insatisfacción

El procedimiento propuesto tiene como beneficiarios principales:

Usurarios que buscan información (UC): se seleccionaron un total de 50 usuarios que acceden frecuentemente a SRI para buscar información de diversos tipos. Las preguntas realizadas en este grupo de usuarios fueron las siguientes:

- ¿Considera usted que los resultados obtenidos satisfacen sus necesidades de búsqueda?
- ¿Recomendaría el uso de este SRI a otros usuarios?
- ¿Le satisface la forma en que son mostrados y ordenados los resultados que le brinda el SRI?

En la tabla 9 se muestran los resultados del cálculo del ISG en los dos grupos de usuarios:

Tabla 9. Valores de ISG obtenidos en la aplicación de la técnica ladov.

Grupo de usuarios	UC
ISG	0,851

Los dos valores de ISG obtenidos se encuentran en el intervalo de satisfacción, por lo que se puede concluir que la satisfacción de los usuarios que se benefician del procedimiento es alta.

3.4 Triangulación metodológica de los métodos aplicados

Según Ramírez (2016) la triangulación metodológica es una técnica para evaluar distintos puntos de referencia y definir una posición. Reduce el sesgo producido en la comparación de resultados obtenidos en la cuantificación de variables mediante un método cuantitativo, las tendencias y dimensiones que surgen de la aplicación de métodos cualitativos (Valencia, 2013). El resultado de la aplicación de esta técnica se muestra en la tabla 10.

Tabla 10. Resultados de la triangulación metodológica.

Objetivo a evaluar	Métodos cuantitativos	Métodos cualitativos	Conclusiones
Evaluar la capacidad del procedimiento desarrollado para mejorar la estructuración y almacenamiento	**Experimentos**: -pruebas de eficacia y eficiencia.	**ladov:** Alto grado de satisfacción ISG= 0.84 **Criterio de expertos:**	Los resultados arrojados por los métodos aplicados están en concordancia. Se valida de forma positiva la capacidad del procedimiento desarrollado

de los documentos en el SRI.	-pruebas aplicadas a la estructuración de los documentos. Los resultados de los dos experimentos arrojaron valores satisfactorios.	en todos los casos se obtuvo un IP>87	para mejorar la eficacia del almacenamiento y la calidad de los resultados brindados a los usuarios.

Conclusiones del capítulo

La aplicación de los métodos cualitativos y cuantitativos definidos en la estrategia de validación confirmó la capacidad del procedimiento de mejorar la eficacia del almacenamiento y la calidad de los resultados brindados a los usuarios en el SRI Orión.

La triangulación metodológica demostró la confiabilidad de los resultados que por separado se obtuvieron a través de la técnica Iadov, Likert y la experimentación.

El procedimiento desarrollado tiene impacto económico y social demostrado en:

- Su capacidad de brindar a los usuarios información que satisface sus necesidades e interrogantes.
- Brinda una guía para el diseño de arquitecturas de hardware que permiten un uso eficiente de los servidores y apoya la reducción de gastos innecesarios en recursos de hardware.
- La definición de los roles necesarios para la aplicación y mantenimiento del flujo de procesos en el procedimiento, permite una planificación adecuada del cronograma de trabajo y la asignación de las responsabilidades de cada rol.

CONCLUSIONES

- El estudio de los referentes teóricos permitió verificar que no existen procedimientos enfocados a la estructuración y almacenamiento de documentos en SRI, pero los modelos estudiados aportaron conocimientos acerca de cómo estructurar y almacenar estos documentos.
- A partir del estudio realizado de los fundamentos teóricos relacionados con la recuperación de información y procesamiento de metadatos se definieron tres grupos de metadatos asociados a los tipos de documentos almacenados por el SRI Orión (documentos, imágenes y videos), definiendo una estructura correcta para cada tipo de archivo y basando su representación en el modelo vectorial.
- El procedimiento desarrollado integra: la estructuración y almacenamiento de los documentos en el SRI Orión y las arquitecturas de hardware; lo que permite mejorar la eficacia del almacenamiento y la calidad de los resultados brindados a los usuarios en este buscador.
- La arquitectura de hardware diseñada sirve de base para los componentes rastreo e indexación del SRI Orión; es escalable y su aplicación mejora la eficiencia y eficacia de este sistema.
- Las técnicas utilizadas para la validación del procedimiento, comprobaron que los constructos del mismo están fundamentados en base a tecnologías reconocidas a nivel internacional en el campo de la RI y constataron el alto nivel de satisfacción de los usuarios con respecto al procedimiento.

RECOMENDACIONES

1. Generalizar el procedimiento propuesto en los proyectos que desarrollan SRI a nivel nacional.

2. Continuar con la investigación para definir mecanismos de definición automática de recursos de hardware, que facilite el proceso de diseño de la arquitectura de hardware necesaria para ejecutar las tareas en un proyecto enfocado a la RI.

3. Diseñar un mecanismo de retroalimentación para evaluar de forma automática la calidad del flujo de RI propuesto en el procedimiento.

REFERENCIAS BIBLIOGRÁFICAS

1. Abudaqqa, Y., & Patel, A. (2015). Distributed search engine architecture based on topic specific searches. In AIP Conference Proceedings, 1660(1), p. 090018.
2. Agre, G. H., & Mahajan, N. V. (2015). Keyword focused web crawler. In Electronics and Communication Systems (ICECS), 2015 2nd International Conference. IEEE.1089-1092. ISBN: 978-1-4799-7225-8.
3. Agredo, S. D. G., Lozada, C. A. C., & Flórez, L. C. G. (2013). Modelo de Búsqueda Web Basado en Información del Contexto del Usuario y Técnicas de Filtrado Colaborativo. Revista UIS Ingenierías, 11(1).
4. Aguilar, J., & Mosquera, D. (2015). Middleware Reflexivo para la gestión de Aprendizajes Conectivistas en Ecologías de Conocimientos (eco-conectivismo) Reflective Middleware for Managing Learning Connectivism in Knowledge Ecologies (eco-Connectivism). Latin American Journal of Computing Faculty of Systems Engineering Escuela Politécnica Nacional Quito-Ecuador, 2(2).
5. Amador, P. L. (2015). Agrupamiento de artículos científicos con formato semiestructurado basado en las referencias bibliográficas. (Disertación doctoral). Universidad Central “Marta Abreu” de Las Villas, Cuba.
6. Antiñanco, M. J. (2014). Bases de Datos NoSQL: escalabilidad y alta disponibilidad a través de patrones de diseño. (Disertación doctoral). Facultad de Informática, Universidad Nacional de La Plata. Disponible en: http://sedici.unlp.edu.ar/bitstream/handle/10915/36338/Documento_completo.pdf?sequence=5.
7. Argomedo, S., Herrera, J., Molina, K., & Relos, S. (2014). Álgebra Lineal para Ingeniería. Iniciativa Latinoamericana de Libros de Texto Abiertos (LATIn). Disponible en: http://www.proyectolatin.org/books/Algebra_Lineal_para_Ingenieria_CC_BY-SA_3.0.pdf.
8. Arora, B., & Bhardwaj, A. (2014). Analysis of Information Retrieval models. International Journal Of Engineering And Computer Science, 3(10). ISSN: 2319-7242.
9. Avedaño, N. D. O., & Ortiz, J. E. (2013). Recuperación de información utilizando ecosistemas de agentes inteligentes. Revista Vínculos, 2(1), 21-39.
10. Baena, F., Fernández, C. C., Espejo, C., & Díaz, J. (2014). Codificación y representación cartográfica de noticias. Aplicación de las humanidades digitales al estudio del periodismo de la Edad moderna. El profesional de la información, 23 (5), 519-526.

11. Baeza, R., & Ribeiro, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press. Disponible en: ftp://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap10_user_interfaces_and_visualization-modern_ir.pdf.

12. Baeza-Yates, R., & Liaghat, Z. (2017). Quality-efficiency trade-offs in machine learning for text processing. arXiv preprint arXiv:1711.02295.

13. Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: a comparison of retrieval performances. Lecture Notes on Software Engineering, 2(3). URI: http:// eprints.um.edu.my/id/eprint/13423.

14. Balamurugan, M., & Iyswarya, E. (2017). A Trend Analysis of Information Retrieval Models. International Journal of Advanced Research in Computer Science, 8(5). ISSN: 0976-5697.

15. Bandagale, P., Sawantdesai, N. R., Paradkar, R. U., & Shirodkar, P. P. (2017). Survey on Effective Web Crawling Techniques. International Journal of Computer & Mathematical Sciences, 6(10). ISSN: 2347 – 8527.

16. Benavides, D. K. (2015). Arquitectura de Google. Universidad de Costa Rica Escuela de Ciencias de la Computación e Informática M.Sc. Kryscia Daviana Ramírez Benavides. Disponible en: http://documentslide.com/documents/arquitectura-de-google-universidad-de-costa-rica-escuela-de-ciencias-de-la-computacion-e-informatica-msc-kryscia-daviana-ramirez-benavides.html.

17. Berry, M. W. (2004). Survey of Text mining: Clustering, Classification, and Retrieval. New York, USA, Springer Verlag. Disponible en: https://link.springer.com/book/10.1007/978-1-84800-046-9.

18. Blázquez, M. (2013). Técnicas avanzadas de recuperación de información: procesos, técnicas y métodos. E-Prints Complutense, Madrid. Disponible en: http://mblazquez.es/wp-content/uploads/ebook-mbo-tecnicas-avanzadas-recuperacion-informacion1.pdf.

19. Bolaños, C. S. (2015). Corpus linguistics: approaches for contemporary linguistic research. Forma y Función, 28(1), 31-54. DOI: http://dx.doi.org/10.15446/fyf.v28n1.51970.

20. Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer networks, 56(18), 3825-3833.

21. Büttcher, S., Clarke, C. L., & Cormack, G. V. (2016). Information retrieval: Implementing and evaluating search engines. Mit Press.

22. Cabeza, M. (2014). Aplicación para la recuperación de información soportada en el agrupamiento de documentos XML de artículos científicos. (Tesis Doctoral). Universidad Central “Marta Abreu” de Las Villas, Cuba.

23. Cafarella, M., & Cutting, D. (2004). Building nutch: Open source search. *Queue*, *2*(2), 54.

24. Cambazoglu, B. & Baeza, R. (2015). Scalability challenges in web search engines. Synthesis Lectures on Information Concept, Retrieval, and Services, 7(6), 1-138.

25. Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., & Quarteroni, S. (2013). The information retrieval process. In Web Information Retrieval (pp. 13-26). Springer, Berlin, Heidelberg.

26. Cubanic. (2018). Cuántos dominios hay bajo .cu. Cubanic.cu: CITMATEL. Disponible en: http://www.nic.cu/estadisticas.php.

27. Dublin Core Metadata Initiative. (2017). The DCM Initiative. U.S: ASIS&T. Disponible en: http://www.dublincore.org.

28. Durán, C; Ramírez, R. & Juganaru, M. (2014). Obtención de descripciones significativas para una memoria corporativa. Research in Computing Science, 88, 53-59.

29. Ensias, M. V. (2017). Genaum: New semantic distributed search engine. Journal of Mobile Multimedia, 12(3-4), 210-221.

30. Fagan, J. L. (2017, August). Automatic Phrase Indexing for Document Retrieval: An Examination of Syntactic and Non-Syntactic Methods. In ACM SIGIR Forum, 51(2), 51-61. DOI: 10.1145/3130348.3130355.

31. Farias, J. J. S., Frías, R. T., González, L. A. L., & Vázquez, J. I. C. (2016). Persistencia de datos con ActiveJDBC ORM. Pistas Educativas, 38(122). ISSN 1405-1249.

32. Fernández, J. M. (2013). Modelos de Recuperación de Información basados en Redes de Creencia. Repositorio Institucional Universidad de Granada. Disponible en: http://hdl.handle.net/10481/24516.

33. García, E. N. (2015). Optimización de un sistema de indexación y búsqueda de palabras clave en grandes colecciones de imágenes de texto manuscrito. (Disertación doctoral). Universidad Politécnica de Valencia, España. Disponible en: https://riunet.upv.es/bitstream/handle/10251/55625/NOYA%20-%20Optimizaci%F3n%20de%20un%20sistema%20de%20indexaci%F3n%20y%20b%FAsqueda%20de%20palabras%20clave%20en%20grandes%20coleccion....pdf?sequence=1.

34. Gómez, H. E. (2014). Diseño de un modelo para la recuperación de documentos basado en ontologías en el dominio de la ingeniería informática. (Tesis pregrado). Pontificia Universidad

Católica del Perú, Perú. Disponible en: http://tesis.pucp.edu.pe/repositorio/handle/123456789/5758.

35. Granados, R. (2013). Fusión multimedia semántica tardía aplicada a la recuperación de información multimedia. (Tesis doctoral). Universidad Nacional de Educación a Distancia (UNED), España. Disponible en: http://e-spacio.uned.es:8080/fedora/get/tesisuned:IngInf-Rgranados/Documento.pdf.

36. Gudivada, V. N., Baeza, R. A., & Raghavan, V. V. (2015). Big Data: Promises and Problems. IEEE Computer, 48(3), 20-23.

37. Gupta, K., & Goyal, A. (2015). A Comprehensive Survey on Hidden Web Crawler. International Journal Of Engineering And Computer Science, 4(06). ISSN: 2319-7242.

38. Internetlivestats.com. (2018). Total number of Websites - Internet Live Stats. Disponibl en: http://www.internetlivestats.com/total-number-of-websites/ [accedido: 8 enero 2018].

39. Jaimes, G & Vega, F. (2005) Modelos clásicos de recuperación de la información. Revista Integración, 23(1).

40. Jaramillo, S., & Londoño, J. M. (2014). Document search supported on an ontological indexing system created with mapreduce. Ciencia e Ingeniería Neogranadina, 24(2), 57-75.

41. Kausar, A; Dhaka, V. S.; Singh, S. (2013). Web crawler: a review. International Journal of Computer Applications, 63, (2).

42. Khare, R., Cutting, D., Sitaker, K., & Rifkin, A. (2004). Nutch: A flexible and scalable open-source web search engine. *Oregon State University*, *1*, 32-32.

43. Korfhage, R. (1997). Information Storage and Retrieval, New York: John Wiley. Disponible en: http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471143383.html.

44. Lafferty, J., & Zhai, C. (2017). Document language models, query models, and risk minimization for information retrieval. In ACM SIGIR Forum, 51(2), 251-259. DOI: 10.1145/3130348.3130375.

45. Lal, N., Qamar, S., & Shiwani, S. (2018). Information Retrieval and Query Ranking of Unstructured Data in Dataspace using Vector Space Model, 4(1), 17-24. ISSN: 2454-4248.

46. Leyva, P. R., Sala, H. V., & Flores, L. A. P. (2016). Componentes y funcionalidades de un sistema de recuperación de la información. Revista Cubana de Ciencias Informáticas, 10, 150-162.

47. Likert, R. (1932). A technique for the measurement of attitudes. Archives of psychology, p. 55.

48. Lizcano, L. I., & Pérez, D. A. (2016). Sistema de recuperación de información basado en el modelo vectorial. Respuestas, 6(1), 22-29.

49. López, V., Santillana, A., & Vittet, L. (2015). Plan estratégico para Google Inc. Inc. 2015-2017. Repositorio Institucional Universidad del Pacífico. Disponible en: http://hdl.handle.net/11354/1543.

50. Machado G. N. (2015). Uso de la similitud semántica para la recuperación de información geoespacial. (Tesis doctoral). Universidad de Alicante, España.

51. Mainegra, Y. B., Alonso, M. Á., Acosta, J. C. P., & Tumbarell, A. A. Í. (2016). Desarrollo de la base de datos del módulo Civil y familia del Sistema de Gestión Fiscal. Serie Científica de la Universidad de las Ciencias Informáticas, 9(1). ISSN: 2306-2495.

52. Martínez, J.A. (2006). Los modelos clásicos de recuperación de información y su vigencia. En: Tercer Seminario Hispano-Mexicano de investigación en Bibliotecología y Documentación, UNAM, Centro Universitario de Investigaciones Bibliotecológicas. pp.187-206. Disponible en: http://eprints.rclis.org/bitstream/10760/9662/1/Modelos_RI_vers_def.pdf.

53. Matías, G. A., Ledeneva, Y. N., García, R. A., & Sidorov, G. (2016). Generación automática de resúmenes independientes del lenguaje. (Tesis de maestría). Universidad Autónoma del Estado de México, México.

54. Méndez Rodríguez, E. M. (2001). Metadatos y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales.

55. Monsalve, R. (2015). Estudio de gestión de sistemas basados en grafos. Tesis de licenciatura. Biblioteca Universidad de Carlos III de Madrid. Disponible en: http://hdl.handle.net/10016/23656.

56. Mora, M. (2016). Estudio, diseño y aplicación de técnicas basadas en Soft Computing para la mejora de la búsqueda conceptual en Internet. Repositorio Universitario Institucional de Recursos Abiertos. Disponible en: https://ruidera.uclm.es/xmlui/handle/10578/8875.

57. Moreira, J. E., Michael, M. M., Da Silva, D., Shiloach, D., Dube, P., & Zhang, L. (2007, June). Scalability of the Nutch search engine. In *Proceedings of the 21st annual international conference on Supercomputing* (pp. 3-12). ACM.

58. Nasution, M. K., Noah, S. A. M., & Saad, S. (2016). Social network extraction: Superficial method and information retrieval. Proceeding of International Conference on Informatics for Development. arXiv preprint arXiv:1601.02904.

59. Nava, V; Domínguez, V & González, A. (2015). Sistema de Recomendación para la

Búsqueda Personalizada en un Repositorio de Trabajos de Titulación. Revista Latinoamericana de Ingeniería de Software, 3(6), 223-230.

60. Orión_Universidad de las Ciencias Informáticas (2017). RedCuba. Cuba: UCI. Disponible en: www.redcuba.cu.

61. Piedra, N., Chicaiza, J., Quichimbo, P., Saquicela, V., Cadme, E., López, J. & Tovar, E. (2015). Marco de trabajo para la integración de recursos digitales basado en un enfoque de web semántica. RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação, (SPE3), 55-70.

62. Pino, D. (2014). Creación de un crawler semántico y distribuible para su aplicación en un buscador web. (Tesis de maestría). Universidad Carlos III de Madrid, España. Disponible

63. Quiñones, E. (2015). Recuperación cruzada de información pública clínico-genómica a partir de consultas sobre repositorios clínicos privados. (Tesis de pregrado). Universidad Politécnica de Madrid, España. Disponible en: http://oa.upm.es/38319/7/PFG_EDUARDO_QUINONES_MATESANZ.pdf.

64. Ramírez, F. (2016). Modelo para la selección de equipos de trabajo quirúrgico en sistemas de información en salud aplicando técnicas de inteligencia organizacional. (Tesis doctoral). Universidad de las Ciencias Informáticas, La Habana, Cuba.

65. Ramírez, K. (2007). Stemming-Lematización. Obtenido de Escuela de Ciencias de la Computación e Informática; Universidad de Costa Rica: Disponible en: http://www.ecci.ucr.ac.cr/~kramirez/RI/Material/Presentaciones/Stemming.pdf.

66. Ramos, F; Velez, I. J. (2016). Integración de técnicas de procesamiento de lenguaje natural a través de servicios web. (Tesis de pregrado). Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina. Disponible en: http://www.ridaa.unicen.edu.ar/xmlui/handle/123456789/644.

67. Rautenstrauch, Ramón. (2010). Opciones avanzadas en las búsquedas de Google y Bing. *Opciones avanzadas en las búsquedas de Google y Bing.* [En línea] 28 de Octubre de 2010. [Citado el: 6 de Diciembre de 2016.] http://www.apasionadosdelmarketing.es.

68. Reinoso, Q., & Paulina, V. (2007). Definición e implementación de un modelo de respaldos de información en la compañía Transelectric SA. (Tesis de pregrado). Escuela Politécnica Nacional de Formación Tecnológica, Ecuador. Disponible en: http://bibdigital.epn.edu.ec/bitstream/15000/1475/1/CD-0990.pdf.

69. Rich, M. (2011). Report: Google Uses About 900,000 Servers. Data center Knowledge Recuperado de: https://www.datacenterknowledge.com/archives/2011/08/01/report-google-uses-about-900000-servers. Visitado el 1 de octubre del 2018.

70. Ríssola, E. A., & Tolosa, M. G. H. (2015). Gestión Eficiente del Índice Invertido para Flujos de Documentos en Tiempo Real. (Tesis de licenciatura). Universidad Nacional de Luján, Argentina. Disponible en: http://www.labredes.unlu.edu.ar/sites/www.labredes.unlu.edu.ar/files/site/data/TFL-RissolaEsteban-LG112146.pdf.

71. Roa, S. M., & Mera, M. F. (2013). Recuperación basada en contenido de imágenes microscópicas de cuello uterino infectadas con el virus del papiloma humano empleando características de textura. Revista Ingeniería Biomédica, 7(14), 69-80.

72. Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. Journal of the American Society for Information science, 27(3), 129-146.

73. Rodríguez, C. F. (2016) Sistema de búsqueda y recuperación de documentos, aplicado a los informes de tesis y proyectos de grado digitalizados de la Pontificia Universidad Católica Del Ecuador Sede Esmeraldas. (Tesis doctoral). Ecuador-PUCESE-Escuela de Sistemas y Computación, Ecuador.

74. Romá, M. T. (2014). Los sistemas de recuperación de la información (SRI) de las bases de datos documentales y la calidad de los resultados obtenidos. Documentación e Información Científica en Salud. Repositorio Institucional Unversidad de Alicante. Disponible en: http://rua.ua.es/dspace/handle/10045/42141.

75. Salton G, McGill M.J. (1983). An Introduction to Modern Information Retrieval. San Francisco: Mcgraw-Hill College. Disponible en: http://dl.acm.org/citation.cfm?id=576628.

76. Sampieri, R., Fernández, C. & Baptista, P. (2014). Metodología de la investigación. McGraw-Hill Education. México. Disponible en: https://www.esup.edu.pe/descargas/dep_investigacion/Metodologia%20de%20la%20investigaci%C3%B3n%205ta%20Edici%C3%B3n.pdf.

77. Sandoval, E. M., Montañez, C., & Bernal, L. (2015). UBOA. Una alternativa metodológica para la construcción de Objetos Virtuales de Aprendizaje. Repositorio Institucional Universidad Nacional Autónoma de México. Disponible en: http://reposital.cuaed.unam.mx:8080/jspui/handle/123456789/3936.

78. Savenkov, D., Braslavski, P., & Lebedev, M. (2011, September). Search snippet evaluation at yandex: lessons learned and future directions. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 14-25). Springer, Berlin, Heidelberg.

79. Sequera, L. J. (2010). Nueva propuesta evolutiva para el agrupamiento de documentos en sistemas de recuperación de información. (Tesis doctoral). Universidad de Alcalá. Departamento de Ciencias de la Computación, Alcalá. Disponible en: https://dialnet.unirioja.es/servlet/dctes?codigo=22028.

80. Seroubian, M. (2013). Buscadores: cómo usar las herramientas de búsqueda en Internet. Informatio. Revista del Instituto de Información de la Facultad de Información y Comunicación, 2, 43-57.

81. Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014). A latent semantic model with convolutional-pooling structure for information retrieval. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ACM. 101-110. DOI: 10.1145/2661829.2661935.

82. Singh, J., & Sharan, A. (2013, November). A comparative study between keyword and semantic based search engines. In International Conference on Cloud, Big Data and Trust (pp. 13-15).

83. Singh, V. K., & Singh, V. K. (2015). Vector space model: an information retrieval system. Int J Adv Eng Res, 141-143. ISSN: 2249–8974.

84. Sinisterra, M. M., Henao, T. M. D., & López, E. G. R. (2012). Clúster de balanceo de carga y alta disponibilidad para servicios web y mail. Informador técnico, 76, 93-102.

85. Smiley, D., Pugh, E., Parisa, K., & Mitchell, M. (2015). Apache Solr enterprise search server. Packt Publishing Ltd.

86. Sullivan, D. (2004). Major search engines and directories. Search Engine Watch (online resource) 2004.

87. Tolosa, G. H., & Feuerstein, E. (2014). Algoritmos Eficientes y Datos Masivos en Búsquedas a Gran Escala. En XVI Workshop de Investigadores en Ciencias de la Computación. Disponible en: http://hdl.handle.net/10915/42043.

88. Tolosa, H.; Bordignon, R. A. (2008). Introducción a la Recuperación de Información. E-prints in library & information sciences. Disponible en: http://eprints.rclis.org/12243/.

89. Torres, C., & Arco, L. (2016). Representación textual en espacios vectoriales semánticos. Revista Cubana de Ciencias Informáticas, 10(2), 148-180.

90. Udapure, T. V., Kale, R. D., & Dharmik, R. C. (2014). Study of web crawler and its different types. IOSR Journal of Computer Engineering, 16(1), 01-05. ISSN: 278-0661.

91. Umagandhi, R., & Kumar, A. S. (2017). Search Query Recommendations in Web Information Retrieval Using Query Logs. In Web Usage Mining Techniques and Applications Across Industries, IGI Global. 199-222. DOI: 10.4018/978-1-5225-0613-3.ch008.

92. Unocero (2017). Google usa unos 900,000 servidores. Disponible en: https://www.unocero.com/2011/08/03/google-usa-unos-900000-servidores/.

93. Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. Information Processing & Management, 50(1), 104-112. DOI: 10.1016/j.ipm.2013.08.006.

94. Valencia, M. M. A. (2013). La triangulación metodológica: sus principios, alcances y limitaciones. Investigación y educación en enfermería, 18(1).

95. Vargas, A. M. (2016). [Reseña de libro] Cibermetría. Midiendo el espacio red. Comunicación y Medios, 3, 118-119.

96. Velasco, C. (2014). Modelo basado en técnicas de procesamiento de lenguaje natural para extraer y anotar información de publicaciones científicas. (Disertación doctoral). ETSI_Informatica. Archivo Digital UPM, España. Disponible en: http://oa.upm.es/30856/.

97. Velez, O; Santos, C. (2014). Sistemas recomendadores: Un enfoque desde los algoritmos genéticos. Industrial data, 9, (1), 023-031.

98. Verma, D., & Kochar, B. (2016). Multi Agent Architecture for Search Engine. International Journal Of Advanced Computer Science And Applications, 7(3), 224-229.

99. Zhai, C., Cohen, W. W., & Lafferty, J. (2015). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In ACM SIGIR Forum, 49(1), pp. 2-9. DOI: 10.1145/2795403.2795405

100. Zhang, H., M. Yan-hong, M. Wei-jun, y B. Zhong-xian. (2013). Study of Distributed Personalized Search Engine. Advanced Materials Research. Trans Tech Publications, páginas 1035-1039.

101. Zhang, Y; Liu, S & Mathews, E. (2015). Convergence of digital humanities and digital libraries. Library management, 36, (4–5), 362–377.

ANEXOS

Anexo 1. Planilla para definir el perfil de búsqueda. Fuente: elaboración propia.

<table>
<tr><td colspan="2">Nombre y apellidos:_____
Seleccione de las categorías que a continuación se les muestran, cuál o cuáles de ellas identifica como preferidas en sus búsquedas de información. Asigne entre ellas un total de 100 puntos en dependencia del predominio de preferencia de cada una de ellas.</td></tr>
<tr><td>Categorías:
___Deporte
___Cultura
___Medio ambiente
___Economía
___Política
___Salud
___Ciencia y tecnología</td><td>Ejemplo:
_50__Deporte
_20__Cultura
_____Medio ambiente
_____Economía
_20__Política
_____Salud
_10__Ciencia y tecnología</td></tr>
<tr><td colspan="2">Seleccione una consulta de las que se muestran a continuación:
- Cuba
- Universidad de las Ciencias Informáticas
- Acontecer nacional
- Actualidad</td></tr>
</table>

Anexo 2. Encuesta para determinar el coeficiente de competencias de los expertos. Fuente: elaboración propia.

Compañero (a): _____

Usted ha sido seleccionado como posible experto para ser consultado respecto a temas relacionados con la estructuración y almacenamiento de documentos en sistemas de recuperación de información, con vista a la investigación que se está llevando a cabo. Agradecemos su valiosa cooperación.

1. Marque con una cruz (X) en la tabla siguiente el valor que se corresponde con el grado de conocimiento que usted posee sobre "el proceso de estructuración y almacenamiento de documentos". (Escala ascendente).

0	1	2	3	4	5	6	7	8	9	10

Realice una autoevaluación del grado de influencia que cada una de las fuentes que le presentamos a continuación ha tenido en su conocimiento y criterio sobre la estructuración y almacenamiento de documentos. Marque con una cruz (X) según corresponda en A (alto), M (medio) o B (bajo).

Fuente de argumento	Grado de influencia de cada una de las fuentes		
	A (alto)	M (medio)	B (bajo)
Análisis teóricos realizados			
Experiencia obtenida			
Autores nacionales			
Autores extranjeros			
Intuición			

Anexo 3. Procedimiento empleado para determinar el coeficiente de competencia de los candidatos a expertos. Fuente: elaboración propia.

Para calcular el coeficiente de competencia de los expertos se utiliza la forma propuesta por (Cañedo et al., 2006):

Kcomp = 0.5 (Kc + Ka)

Donde:

Kcomp: coeficiente de competencia.

Kc: coeficiente de conocimiento o información que tiene el experto sobre el tema de estudio. Se calcula multiplicando por 0,1 la valoración realizada por el propio experto en una escala de 0 a 10.

Ka: coeficiente de argumentación o fundamentación de los criterios del experto, obtenido como resultado de la suma de los puntos de acuerdo a la siguiente tabla patrón:

Patrón de factores para el cálculo del coeficiente de argumentación. Fuente: elaboración propia.

Nro.	Fuente de argumento	A (alto)	M (medio)	B (bajo)
1	Análisis teóricos realizados por usted	0,3	0,2	0,1
2	Experiencia profesional adquirida sobre el tema	0,5	0,4	0,2
3	Conocimiento adquirido de investigaciones y/o publicaciones nacionales e internacionales	0,05	0,04	0,03
4	Conocimiento propio sobre el estado del tema de investigación	0,05	0,04	0,03
5	Conocimientos adquiridos de cursos de superación	0,05	0,04	0,03
6	Intuición	0,05	0,04	0,03

El Competencia de competencia del experto es:

- Alto (A): Si 0,8 < Kcomp < 1,0
- Medio: Si 0,5< Kcomp < 0,8
- Bajo: si Kcomp < 0,5

RESULTADOS: Competencia de los 30 expertos que participaron en la valoración del modelo elaborado.

Expertos	Kc	1	2	3	4	5	6	Ka	Kcomp	Valor
1	0,4	0,1	0,2	0,04	0,05	0,05	0,05	0,5	0,45	Bajo
2	1,0	0,3	0,5	0,05	0,05	0,05	0,05	1,0	1,0	Alto
3	0,6	0,1	0,2	0,04	0,05	0,05	0,05	0,5	0.55	Medio
4	0,8	0,3	0,5	0,05	0,05	0,05	0,05	1,0	0,9	Alto
5	1,0	0,3	0,5	0,05	0,05	0,05	0,05	1,0	1,0	Alto
6	0,8	0,2	0,5	0,05	0,05	0,05	0,05	0,8	0,8	Medio
7	0,9	0,1	0,5	0,05	0,05	0,05	0,05	0,8	0,85	Alto
8	1,0	0,2	0,5	0,05	0,05	0,05	0,05	0,9	0,95	Alto
9	1,0	0,3	0,5	0,05	0,05	0,05	0,05	1,0	1,0	Alto
10	0,8	0,3	0,5	0,05	0,05	0,05	0,05	1,0	0,9	Alto
11	1,0	0,2	0,5	0,05	0,05	0,05	0,05	0,9	0,95	Alto
12	0,8	0,3	0,4	0,05	0,05	0,05	0,05	0,9	0,85	Alto
13	1,0	0,3	0,4	0,05	0,05	0,05	0,05	0,9	0,95	Alto
14	0,9	0,2	0,5	0,05	0,05	0,05	0,05	0,9	0,9	Alto
15	0,9	0,2	0,4	0,05	0,05	0,05	0,05	0,8	0,85	Alto
16	0,8	0,2	0,4	0,05	0,05	0,05	0,05	0,8	0,8	Medio
17	0,7	0,2	0,5	0,05	0,05	0,05	0,05	0,9	0,8	Medio
18	1,0	0,2	0,5	0,05	0,05	0,05	0,05	0,9	0,95	Alto
19	0,9	0,3	0,5	0,05	0,05	0,05	0,05	1,0	0,95	Alto
20	0,8	0,3	0,5	0,05	0,05	0,05	0,05	1,0	0,9	Alto
21	1,0	0,3	0,4	0,05	0,05	0,05	0,05	0,9	0,95	Alto
22	0,7	0,3	0,5	0,05	0,05	0,05	0,05	1,0	0,85	Alto
23	1,0	0,3	0,4	0,05	0,05	0,05	0,05	0,9	0,95	Alto
24	1,0	0,1	0,5	0,05	0,05	0,05	0,05	0.8	0,9	Alto
25	0,7	0,2	0,4	0,05	0,05	0,05	0,05	0,8	0,75	Medio
26	0,4	0,1	0,2	0,04	0,05	0,05	0,05	0,5	0,45	Bajo
27	0,5	0,1	0,2	0,03	0,03	0,03	0,03	0,4	0.45	Bajo
28	0,5	0,1	0,2	0,04	0,04	0,04	0,04	0,4	0,45	Bajo
29	0,4	0,1	0,2	0,04	0,05	0,05	0,05	0,5	0,45	Bajo
30	0,5	0,1	0,2	0,03	0,03	0,03	0,03	0,4	0.45	Bajo

Anexo 4. Cuestionario para expertos. Fuente: elaboración propia.

I- Datos generales del encuestado:

Nombre y apellidos: _____

Área donde labora: _____

Título universitario: _____

Grado científico: _____ Categoría docente: _____

Años de experiencia en el área: _____

El objetivo de la presente encuesta consiste en que usted evalúe cada uno de los indicadores que se presentan en la tabla que a continuación se le muestra, colocando el número en la casilla correspondiente y teniendo en cuenta para ello las siguientes categorías:

5- muy de acuerdo (MA)

4- de acuerdo (DA)

3- ni de acuerdo ni en desacuerdo (Sí-No)

2- en desacuerdo (ED)

1- completamente en desacuerdo (CD)

Procedimiento para la estructuración y almacenamiento de documentos en el SRI Orión

#	Afirmaciones	Respuestas
1	El procedimiento desarrollado mejora la eficacia del almacenamiento y la calidad de los resultados brindados a los usuarios en el SRI Orión.	
2	Los documentos, imágenes y videos que se almacenaron comprenden todos los metadatos definidos en la propuesta de solución.	
3	Las etapas que conforman el procedimiento y las fuentes teóricas que los sustentan, cubren aspectos relevantes en la estructuración y almacenamiento de documentos en SRI.	

Anexo 5. Cuadro Lógico de Iadov para usuarios. Fuente: elaboración propia.

	¿Considera usted que los resultados brindados satisfacen sus necesidades de búsqueda?								
	Si			No sé			No		
¿Le satisface la forma en que son mostrados los resultados que le brinda el SRI?	¿Recomendaría el uso de este SRI a otros usuarios?								
	Si	No sé	No	Si	No sé	No	Si	No sé	No
Me gusta mucho	1	2	6	2	2	6	6	6	6
Me gusta más de lo que me disgusta	2	2	3	2	3	3	6	3	6
Me da lo mismo	3	3	3	3	3	3	3	3	3
Me disgusta más de lo que me gusta	6	3	6	3	4	4	3	4	4
No me gusta nada	6	6	6	4	4	6	4	5	6
No sé qué decir	2	3	6	3	3	3	6	3	4