

Tipo de artículo: Artículo original
Temática: Técnicas de programación
Recibido: 11/12/2018 | Aceptado: 22/01/2018

Modelo computacional para el desarrollo de sistemas de recuperación de información

Computational model to develop information retrieval systems

Paúl Rodríguez Leyva^{1*}, Yennifer Delgado Mesa², Hubert Viltres Sala¹, Vivian Estrada Sentí³, Juan Pedro Febles³

¹Centro CIDI, Universidad de las Ciencias Informáticas, Carretera San Antonio de los Baños Km 2 ½, Boyeros, La Habana, Cuba. {[pleyva](mailto:pleyva@uci.cu), [hviltres](mailto:hviltres@uci.cu)}@uci.cu

² Centro CESOL, Universidad de las Ciencias Informáticas, Carretera San Antonio de los Baños Km 2 ½, Boyeros, La Habana, Cuba. ydmesa@uci.cu

³ Departamento Metodológico de Postgrado, Universidad de las Ciencias Informáticas, La Habana, Cuba. {vivian, febles}@uci.cu

*Autor para correspondencia: pleyva@uci.cu

Resumen

El campo de la recuperación de información, desde su surgimiento en el año 1950, ha aportado herramientas que permiten a los usuarios encontrar respuestas a sus necesidades e interrogantes. A pesar de confirmar en los estudios bibliográficos la existencia de varios modelos de recuperación de información, existen vacíos teóricos relacionados con aspectos como: la definición de arquitecturas escalables para el soporte de todos los componentes de hardware de un sistema de recuperación de información, fundamentación profunda de algoritmos para el cálculo de relevancia, integración de componentes de apoyo a la toma de decisiones con la estructura básica de un sistema de recuperación de información y los roles necesarios para ejecutar el desarrollo y soporte de las tareas fundamentales del proceso de recuperación de información. La presente investigación propone un modelo que integra las variables antes mencionadas y brinda mecanismos para mejorar el proceso de recuperación de información.

Palabras clave: buscadores, modelo computacional, recuperación de información, relevancia

Abstract

The field of information retrieval, since its inception in 1950, has provided tools that allow users to find answers to their needs and questions. In spite of confirming in the bibliographic studies the existence of several models of

information retrieval, there are theoretical gaps related to aspects such as: the definition of scalable architectures for the support of all the hardware components of an information retrieval system, deep foundations of algorithms for the calculation of relevance, integration of components to support decision making with the basic structure of an information retrieval system and the necessary roles to execute the development and support of the fundamental tasks of the information retrieval process. The present research proposes a model that integrates the aforementioned variables and provides mechanisms to improve the process of information retrieval.

Keywords: *computational model, information retrieval, relevance, search engine.*

Introducción

La recuperación de información (RI) se ocupa de encontrar material (normalmente documentos) de una naturaleza no estructurada (normalmente texto) que se encuentra en grandes colecciones (normalmente almacenada de forma digital) y satisface una necesidad de información (Diego, 2009; Ríssola et al., 2009). Los sistemas de recuperación de información, en lo adelante SRI, son herramientas estructuradas por componentes que emplean diferentes algoritmos y métodos para satisfacer la necesidad de información planteada por un usuario en una consulta en lenguaje natural especificada a través de un conjunto de palabras claves (James et al., 2005; Blázquez, 2013; Lafferty et al., 2017). Estos sistemas fundamentan su funcionamiento en modelos de RI destinados a establecer mecanismos para responder las necesidades de búsqueda de los usuarios (Maldonado et al., 2017). El término modelo tiene un significado bien conocido dentro del campo de la RI. Este significado no es más que una adaptación directa de su significado general (esquema teórico de un sistema o de una realidad compleja que se elabora para facilitar su comprensión y el estudio de su comportamiento) a este dominio concreto. En el campo de la RI, de forma simplificada, podemos considerar un modelo como un método para representar tanto documentos como consultas en sistemas de recuperación de la información, y comparar la similitud de esas representaciones. Para ello, los modelos de recuperación tienen que proporcionar de manera implícita o explícita una definición de relevancia. Además, pueden describir el proceso computacional, el proceso humano y las variables que intervienen en el proceso global (Diego, 2009). Empresas como Google no brindan información concreta de que modelos utilizan para desarrollar los componentes de sus buscadores (Fransson, 2010). Además, el cálculo de la relevancia de sus documentos responde a políticas comerciales y gubernamentales, por lo que resulta difícil desarrollar sistemas tan complejos como los buscadores sin poseer un modelo computacional que apoye el proceso de desarrollo de los mismos. El objetivo de esta investigación es la fundamentación de un modelo computacional para el procesamiento de documentos y apoyo a la toma de decisiones

en sistemas de recuperación de información que sirve para el diseño, desarrollo y despliegue de buscadores a nivel nacional e internacional.

Materiales y método

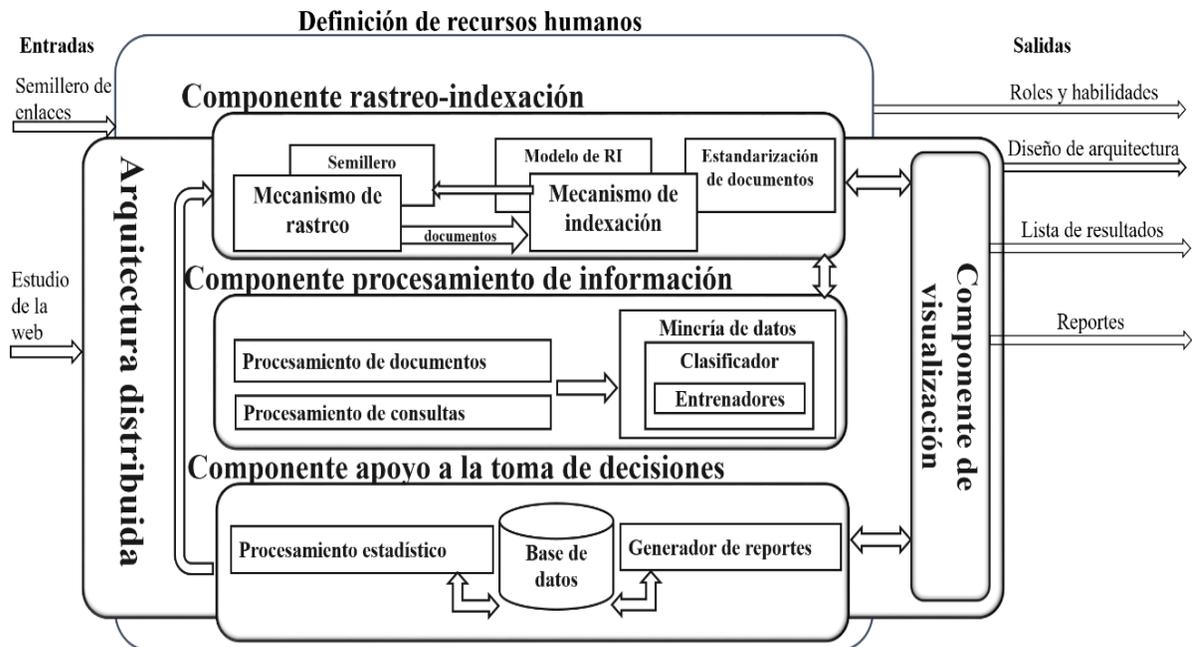


Figura 1. Modelo Computacional (Maldonado et al., 2017). (Elaboración propia)

El modelo computacional propuesto se basa principalmente en 4 principios: escalabilidad, interoperabilidad, estandarización y actualización. Su estructura está compuesta por 5 componentes (Fig. 1). El rastreo e indexación periódica de la web permite a los SRI convertirse en una fuente de constante almacenamiento de información. Los datos almacenados dependen específicamente del funcionamiento interno del buscador y de la estructura real que posee el contenido alojado en la web. Generalmente de cada documento indexado se almacenan metadatos como: url, resumen del contenido, enlaces salientes, palabras claves, el lenguaje en el que está estructurado el documento y tipo mime. El objetivo principal de este componente es rastrear la web y almacenar toda la información encontrada; para lograrlo se apoya en dos mecanismos fundamentales: mecanismo de rastreo y el mecanismo de indexación.

Mecanismo de rastreo: basa su funcionamiento en un conjunto de rastreadores web que inician un proceso de recolección, partiendo de un semillero de enlaces de sitios web que debe contener el mayor número de enlaces posible para lograr un rastreo exhaustivo. Para su correcto funcionamiento se diseña una arquitectura distribuida, (Fig. 2).

Este diseño propone organizar los servidores teniendo en cuenta las categorías de información de la web que se va a rastrear (un servidor por categoría) y su tamaño, valorando la capacidad de procesamiento y almacenamiento del hardware que se posee. Si la web es muy grande, un número pequeño de rastreadores no podría cumplir con el objetivo propuesto. La periodicidad del proceso de rastreo depende del grado de actualización de los sitios definidos en los semilleros de cada uno de los rastreadores; así los usuarios tendrían disponible la información en un período corto de tiempo después de su publicación en la web.

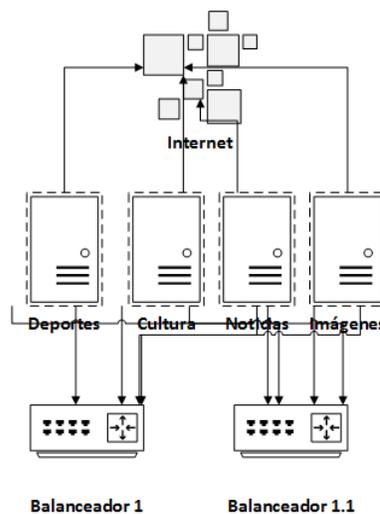


Figura 2. Arquitectura de los servidores de indexación. (Elaboración propia)

Mecanismo de indexación: es el encargado principalmente de la estructuración y almacenamiento de los documentos. Además, aplica el modelo de RI para el cálculo de la relevancia de los documentos relativa a las necesidades de búsqueda de los usuarios. Para su despliegue se propone una arquitectura distribuida, (Fig. 3).

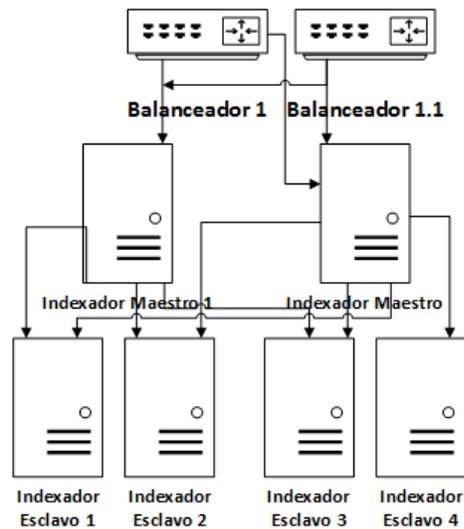


Figura 3. Arquitectura de los servidores de indexación. (Elaboración propia)

Antes de proceder al almacenamiento de los documentos rastreados se aplica un proceso de estandarización que permite definir una estructura correcta para cada uno de los documentos y evita almacenar documentos que no aporten información útil, para lo que se definen los siguientes metadatos como imprescindibles:

Tipo Mime: formato del documento (jpg, doc, html, entre otros)

Resumen: un resumen del contenido del recurso

Fecha: la fecha de elaboración del registro

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

Título: se refiere al título que lleva por nombre el documento

Sitio: el host del que procede el documento

Contenido: texto completo del documento

Enlaces salientes: enlaces a sitios externos

Lenguaje: idioma base del documento

Si los documentos que se van a indexar carecen de alguno de los metadatos definidos anteriormente no se almacenan, para evitar brindar información incompleta a los usuarios del sistema. El modelo de RI que se propone es el vectorial. Salton fue el primero en proponer los SRI basados en las estructuras de espacio vectorial a finales de los 60 dentro del

marco del proyecto SMART. Partiendo de que se pueden representar los documentos como vectores de términos, los documentos podrán situarse en un espacio vectorial de m dimensiones, con tantas dimensiones como componentes tenga el vector (Mabel, 2013). La idea fundamental en la que se basa el modelo vectorial es considerar que tanto los términos claves con respecto a un documento como las consultas, se pueden representar a través de un vector en un espacio de alta dimensionalidad. Por tanto, para evaluar la similitud entre un documento y una consulta; simplemente hay que realizar una comparación de los vectores que los representan (Sequera, 2010).

Para el cálculo de la relevancia de los documentos con respecto a las consultas se diseña un algoritmo, (Fig. 4), que permite añadir a las ecuaciones clásicas de cálculo de similitud (Coseno, Jaccard y Dice) una variable (SC) que relaciona las preferencias de búsquedas de los usuarios o perfil de búsqueda (PBU) con las categorías de los documentos almacenados (CDoc) (Baquerizo et al., 2017). Para obtener el PBU y así establecer cuál o cuáles categorías son las más buscadas por los usuarios y además obtener las categorías de los documentos almacenados CDoc, se propone la utilización de técnicas de minería web como base del funcionamiento del componente de procesamiento de información.

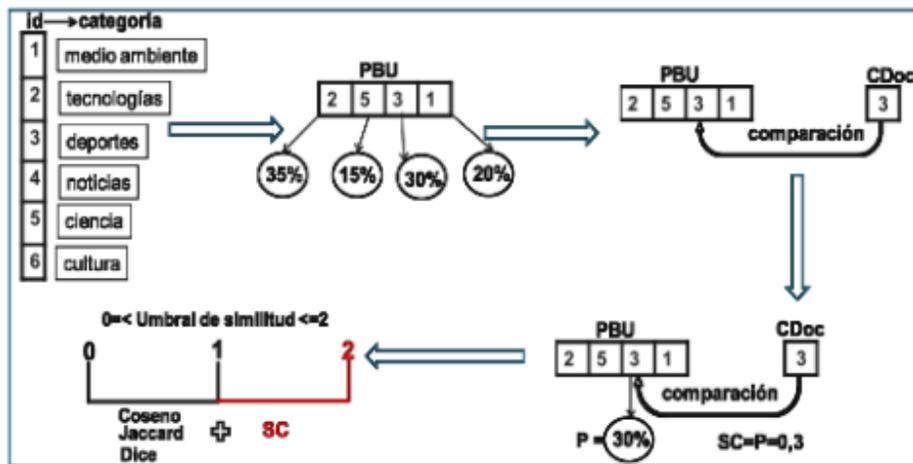


Figura 4. Algoritmo para el cálculo de la relevancia. (Elaboración propia)

La minería web (MW) se refiere esencialmente al descubrimiento y análisis de información de los usuarios en la web, con el objetivo de descubrir patrones de comportamiento (Vásquez et al., 2016). Definidas cada una de las categorías en las que se basará el proceso de categorización se debe proceder a asignar un valor numérico a cada una de ellas. Después de ser definidas las preferencias de búsqueda del usuario como resultado de la categorización de cada una de las consultas introducidas anteriormente, estas se ordenan teniendo en cuenta el porcentaje de predominio (P) de las categorías más consultadas. Si existe una sola categoría predominante, el PBU sería igual al valor numérico de esta

categoría. En caso de que más de una categoría sea predominante y posean el mismo valor de P , el PBU sería una lista con los valores de estas categorías. Con el PBU definido y los documentos categorizados, se procede a comparar los valores numéricos de la o las categorías predominantes con la categoría de cada uno de los documentos. En caso de coincidir dichos valores, la variable SC adquiere el valor de P , garantizando que el valor de SC sea mayor cuando el documento pertenezca al mismo grupo de la o las categorías definidas en su PBU; de lo contrario el valor de SC es 0 significando que la categoría del documento no es afín al PBU

Una vez calculado el valor de SC se suma dicha variable al resultado de la ecuación de similitud (Coseno, Jaccard o Dice) utilizada en el modelo vectorial aplicado, el valor resultante sería la relevancia del documento con respecto a la consulta introducida por el usuario. El umbral de similitud calculado inicialmente oscila desde 0 hasta 1, siendo los documentos más relevantes los más cercanos a 1. Al añadirse la variable SC y sumar su valor a la similitud inicial, el umbral de similitud aumenta de 0 a 2, siendo los documentos más relevantes los cercanos a 2. De esta forma se garantiza brindar a los usuarios resultados más precisos y mejor relacionados con sus preferencias de búsquedas (Baquerizo et al., 2017).

Componente de procesamiento de información

En este componente se ejecuta todo el procesamiento necesario para la categorización de cada uno de los documentos indexados y la asignación del perfil de búsqueda a cada usuario del sistema de recuperación de información. Para su despliegue se propone una arquitectura distribuida, (Fig. 5). El procesamiento en este componente se encarga de dos actividades fundamentales (Maldonado et al., 2017): Categorización de documentos: se encarga de asignar una categoría a cada uno de los documentos indexados utilizando técnicas de minería web, que se añade como un parámetro más en la estructura del documento. Definición del perfil de usuario: se encarga de clasificar las consultas insertadas por los usuarios usando técnicas de minería de texto y asignar a cada uno de ellos un parámetro que recoge todas las categorías consultadas por el usuario.

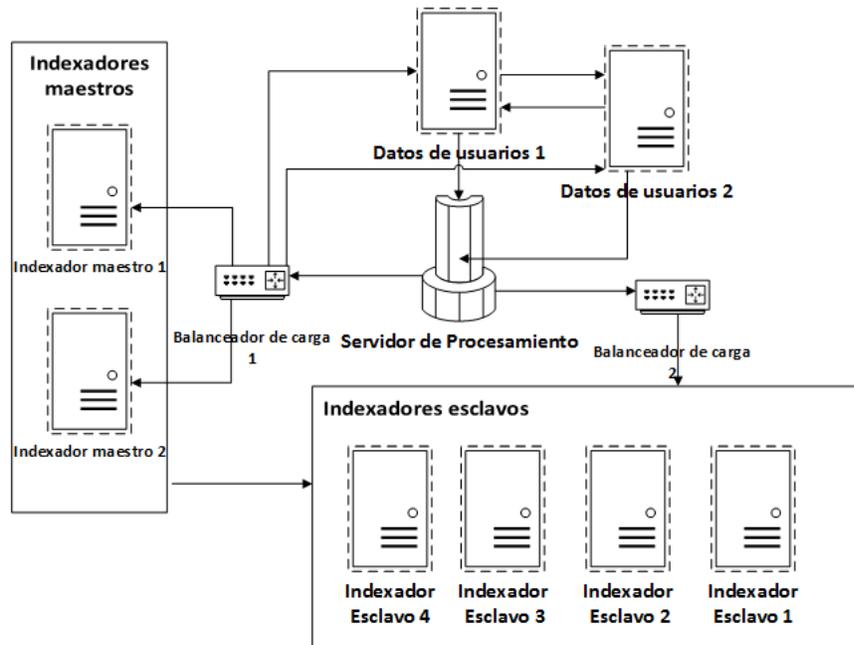


Figura 5. Arquitectura de servidores para el procesamiento de los documentos y las consultas de los usuarios. (Elaboración propia)

Componente de apoyo a la toma de decisiones

Su función principal es el procesamiento estadístico de toda la información almacenada en el sistema de recuperación de información que se debe organizar en dos tipos, los documentos almacenados y los perfiles de búsqueda de los usuarios (Maldonado et al., 2017).

La información procesada se convierte en una fuente rica de conocimiento para detectar tendencias de búsqueda de los usuarios de la red y así proporcionarles contenidos afines a sus gustos; además de brindar datos sobre los perfiles de publicación de cada uno de los sitios web. El resultado final de este componente es una serie de reportes estadísticos que manejan datos como:

- Perfil de búsqueda de los usuarios
- Perfil de publicación de los sitios
- Consultas más buscadas
- Preferencias de búsquedas por regiones geográficas o áreas institucionales
- Porcentaje de publicación de categorías de un dominio específico

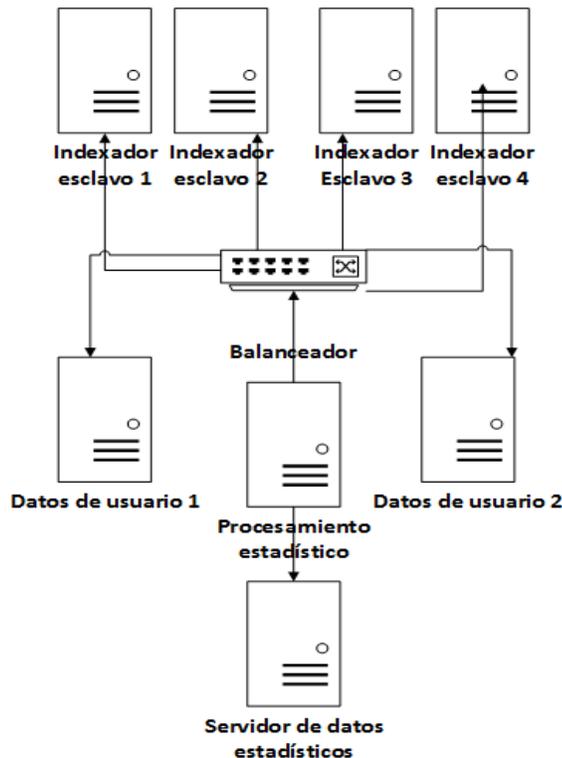


Figura 6. Arquitectura de los servidores de visualización. (Elaboración propia)

Para su despliegue se propone una arquitectura distribuida, (Fig. 6), donde se representa como el servidor de procesamiento extrae los datos relativos a los documentos de los servidores de indexación y los datos relacionados con los perfiles de usuarios de los servidores de datos de usuarios.

Componente de visualización

Es el encargado de brindar los mecanismos necesarios a los usuarios para insertar sus consultas y recibir los resultados más relevantes a través de interfaces de visualización. Estas interfaces cuentan con funcionalidades como: registro de usuarios y la búsqueda avanzada; que permite realizar una búsqueda especializada teniendo en cuenta 9 filtros principales (Leyva et al., 2016):

Con alguna de las palabras: una búsqueda que devuelve resultados que contenga una o algunas de las palabras del criterio de búsqueda.

Con todas las palabras: una búsqueda que devuelva resultados que contengan específicamente todas las palabras del criterio.

Con la frase exacta: una búsqueda que devuelva resultados que contengan específicamente la frase exacta introducida en el criterio de búsqueda.

Sin las palabras: una búsqueda que devuelva resultados que no contengan ninguna palabra introducida en el criterio de búsqueda.

Sitio: permite buscar resultados propios de sitios o dominios.

Tipo de archivo: permite obtener archivos filtrados por tipos agrupados en pdf, html, comprimidos y documentos word.

Idioma: permite obtener resultados en idioma inglés o español.

Última actualización: permite obtener resultados agrupados por intervalos de actualización tales como: en cualquier momento, últimas 24 horas, último mes, última semana y último año.

Términos que aparecen: permite obtener resultados que contengan el criterio de búsqueda en distintas aéreas de las páginas donde se encontraron: título, contenido y url de la página,

El componente de visualización se responsabiliza además de brindar distintas opciones de visualización para los reportes estadísticos brindados por el mecanismo de apoyo a la toma de decisiones. Para su despliegue se propone una arquitectura distribuida, (Fig. 7).

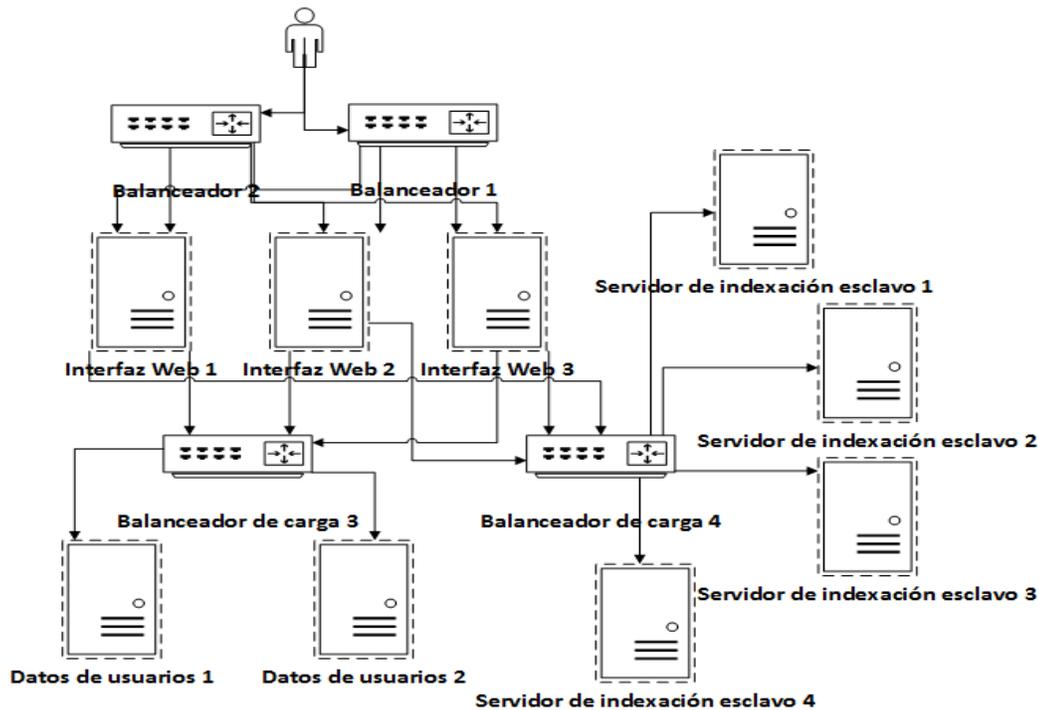


Figura 7. Arquitectura de los servidores de visualización. (Elaboración propia)

El modelo propone como roles imprescindibles para la administración de un SRI y asegurar el correcto funcionamiento de todos los componentes anteriormente abordados, las 5 especialidades que se muestran a continuación (Maldonado et al., 2017): líder de proyecto, arquitecto de software, desarrollador, especialista en rastreo, especialista en indexación, especialista en procesamiento de información, especialista en visualización de la información, especialista en procesamiento estadístico, especialista en redes y servidores y asegurador de la calidad del sistema.

Resultados y discusión

Para la validación del modelo se emplearon diferentes métodos cualitativos y cuantitativos, a continuación, se expone un resumen de los resultados obtenidos. El modelo fue aplicado en el desarrollo del Motor de Búsqueda Orión V2. Este es un SRI desarrollado en la Universidad de las Ciencias Informáticas que permite el rastreo y almacenamiento de la web cubana; para brindar resultados a los usuarios que buscan información sobre los contenidos publicados en esta web. En el momento de aplicar el modelo, el buscador Orión v2 se encontraba en la fase de pruebas de los componentes principales desarrollados para la RI (rastreo, indexación y visualización). En el momento de instanciar el modelo, el proyecto responsable de este SRI contaba con 4 roles (líder de proyecto, desarrollador, analista,

administrador de la configuración) y no existía un flujo de trabajo definido que respondiera a las necesidades reales de un proceso de RI.

Pruebas aplicadas al algoritmo para el cálculo de relevancia de documentos

Los resultados obtenidos después de promediar los valores de precisión y exhaustividad antes y después de aplicar el modelo, se muestran en la tabla 1:

Tabla 1. Valores de precisión y exhaustividad obtenidos. (Elaboración propia)

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

Para comparar los valores de precisión y exhaustividad antes y después de aplicar el algoritmo, se utilizó el editor estadístico STATGRAPHICS. Para comprobar si los 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 algoritmo propuesto para el cálculo de la relevancia mejora los resultados brindados a los usuarios en el SRI.

Pruebas aplicadas al proceso de clasificación de documentos

Debido a la importancia que posee el proceso de categorización como base del cálculo del PBU y las categorías de los documentos, ambas variables utilizadas en el algoritmo propuesto para el cálculo de la relevancia de los documentos, se decide realizar un experimento para evaluar el proceso de clasificación enfocado en las variables precisión de la categorización y la tasa de acierto, teniendo como entrada la estructura de documentos propuesta en el epígrafe: Componente de procesamiento de información. La interpretación de los resultados obtenidos, demuestra que el

categorizador utilizado tiene alto grado de precisión (0,831) y permite una tasa de % 86,564 de acierto en la clasificación de documentos. Estos datos validan de forma positiva el correcto funcionamiento del proceso de categorización propuesto.

El primer paso del experimento consiste en aplicar las pruebas de rendimiento a la arquitectura del Motor de Búsqueda Orión sin aplicar los cambios que propone el modelo para así determinar el estado de los indicadores eficiencia y eficacia. Luego se aplican los cambios que propone el modelo a la distribución de los servidores y se obtienen los valores de las variables definidas para efectuar una comparación y llegar a conclusiones.

La interpretación del valor de los indicadores y el estudio de la distribución de los servidores en el estado actual del Motor de Búsqueda Orión, permitió detectar problemas en el balanceo de la carga de los servidores, la disponibilidad de la información, el uso del almacenamiento y el tiempo de respuesta a los usuarios. Al aplicar la distribución de servidores propuesta por el modelo, se solucionan los problemas detectados con anterioridad. Esto se logra con el uso de los balanceadores de carga entre los rastreadores y los indexadores y entre las interfaces web y los servidores de indexación. 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 para responder a las peticiones de los usuarios 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 modelo, 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. 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. Los resultados obtenidos demuestran que el modelo mejora la eficiencia y eficacia de la arquitectura del SRI Motor de Búsqueda Orión.

Valoración de los expertos sobre el modelo

Los resultados obtenidos al aplicar el escalamiento de Likert ($IP > 89$ en todos los casos), figura 10, demuestran que los principios del modelo, su estructura general y detallada y su pertinencia, 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 modelo computacional para el procesamiento de documentos y apoyo a la toma de decisiones en SRI.

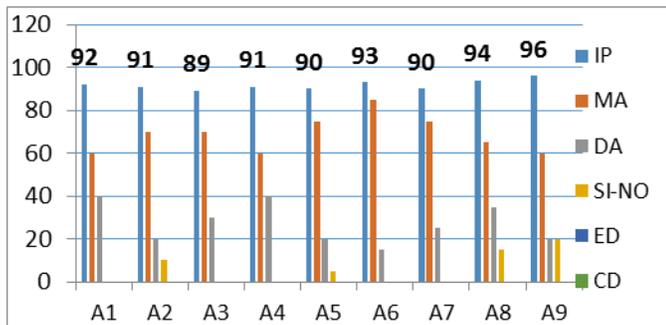


Figura 8. Resultados obtenidos en el procesamiento de las encuestas. (Elaboración propia)

Satisfacción de potenciales usuarios con el modelo

El modelo computacional propuesto tiene como beneficiarios principales, tres grupos de usuarios con características distintas, por lo que se decide aplicar la prueba en tres entornos distintos y utilizando tres cuadros lógicos de Iadov correspondientes a cada grupo. Los grupos de usuarios son los siguientes: Usuarios que desarrollan SRI (UD), Usuarios que toman decisiones (UTD), Usuarios que buscan información (UC).

En la tabla 2 se muestran los resultados del cálculo del ISG en los tres grupos de usuarios:

Tabla 2. Valores de ISG obtenidos en la aplicación de la técnica Iadov. (Elaboración propia)

Grupo de usuarios	UD	UTD	UC
ISG	0,863	0,891	0,851

Los tres 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 modelo es alta.

Conclusiones

El modelo computacional desarrollado integra las variables: arquitectura de hardware, algoritmos para el cálculo de relevancia, recursos humanos, mecanismos de apoyo a la toma de decisiones e integración de los componentes de un SRI, lo que permite mejorar la calidad de los resultados brindados a los usuarios y el apoyo a la toma de decisiones. El algoritmo diseñado y fundamentado para el cálculo de la relevancia de documentos, permite brindar resultados a los usuarios más acordes a sus necesidades de búsquedas y preferencias. La selección de técnicas de minería web permite un procesamiento de la información almacenada que sirve de base para el cálculo de la relevancia de los documentos y la extracción de conocimiento sobre la Web y su uso. La instanciación del modelo en el SRI Motor de Búsqueda Orión v2, demostró su aplicabilidad y viabilidad, además contribuye al proceso de RI en Cuba. Las técnicas utilizadas para la validación del modelo, comprobaron que los constructos del mismo están fundamentados sobre la

base de tecnologías reconocidas a nivel internacional en el campo de la RI y constataron el alto nivel de satisfacción de los usuarios con respecto al modelo.

Referencias

Baquerizo, R. P.; Leyva, P. R.; Febles, J. P.; Viltres, H.; & Estrada, V. S. Algorithm for calculating relevance of documents in information retrieval systems, 2017.

Blázquez, M. Técnicas avanzadas de recuperación de información: procesos, técnicas y métodos. E-Prints Complutense, Madrid, 2013, ISBN 978-84-695-8030-1. Disponible en: <http://mblazquez.es/wp-content/uploads/ebook-mbo-tecnicas-avanzadas-recuperacion-informacion1.pdf>.

Diego, N. Técnicas de indexación y recuperación de documentos utilizando referencias geográficas y textuales. Universidade da Coruña. Departamento de Computación, 2009, ISBN: 978-84-693-3270-2.

Disponible en:

https://books.google.com/cu/books?hl=es&lr=&id=Bd_QT3QgKLgC&oi=fnd&pg=PA5&dq=Efficient+Information+Searching+on+the+Web:+A+Handbook+in+the+Art+of+Searching+for+Information.&ots=-P3Kv0LEL6&sig=qKf7iPTHQtkIYrid5CadiZkW1g&redir_esc=y#v=onepage&q=Efficient%20Information%20Searching%20on%20the%20Web%3A%20A%20Handbook%20in%20the%20Art%20of%20Searching%20for%20Information.&f=false

Fransson, J. Efficient Information Searching on the Web: A Handbook in the Art of Searching for Information, 2010.

Jaimes, G & Vega, F. Modelos clásicos de recuperación de la información. Revista Integración, 2005, 23(1). ISSN: 2145-8472. Disponible en: <http://revistas.uis.edu.co/index.php/revistaintegracion/article/view/479>.

Lafferty, J.; & Zhai, C. Document language models, query models, and risk minimization for information retrieval. In ACM SIGIR Forum, 2017, 51(2), 251-259. DOI: 10.1145/3130348.3130375. Disponible en: <http://dl.acm.org/citation.cfm?id=3130375>.

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

Mabel, S. Buscadores: cómo usar las herramientas de búsqueda en Internet. Informatio. Revista del Instituto de Información de la Facultad de Información y Comunicación, 2013, no 2.

Maldonado, P. D. C. O.; Leyva, P. R., Febles, J. P.; Sala, H. V.; & Mesa, Y. D. Computational model for the processing of documents and support to the decision making in systems of information retrieval, 2017.

Ríssola, E. A.; & Tolosa, M. G. H. Gestión Eficiente del Índice Invertido para Flujos de Documentos en Tiempo Real. (Tesis de licenciatura). Universidad Nacional de Luján, Argentina, 2015.

Sequera, J. C. Nueva propuesta evolutiva para el agrupamiento de documentos en sistemas de recuperación de información. Tesis Doctoral. Universidad de Alcalá, 2010, página 23, 3 párrafo.

Vásquez, A. C.; Fernández, C. L. Aprendizaje de perfiles de usuario web para modelizar interfaces adaptativas. Theorema, segunda época, 2016, no 3, p. 155-164.