Revista Cubana de Ciencias Informáticas
Vol. 10 No. Especial Informática 2016, mayo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 150-162

Tipo de artículo: Artículo original
Temática: Tecnologías de la información y las telecomunicaciones
Recibido: 25/04/2016 | Aceptado: 15/05/2016

Componentes y funcionalidades de un sistema de recuperación de la información

Components and functionalities of an information retrieval system

Paúl Rodríguez Leyva[1*], Hubert Viltres Sala[1], Leiny Amel Pons Flores[1]

1 Universidad de las Ciencias Informáticas, Cuba, Carretera a San Antonio de los Baños, Km. 2 1⁄2. Torrens, municipio de La Lisa. La Habana, {pleyva, hviltres, lenny}@uci.cu

[*] Autor para correspondencia: pleyva@uci.cu

Resumen

Los sistemas de recuperación de información o buscadores, son una fuente de acceso a la información que se encuentra distribuida en el mundo de la web, así como los servicios que en esta se brindan, su importancia y necesidad en el mundo de la navegación por internet está determinada debido a que son principalmente rastreadores de información que luego es almacenada y posteriormente accesible a través de interfaces y funcionalidades de búsqueda y ordenada según algoritmos matemáticas empleados para calcular la relevancia de los resultados. Orión es un sistema de recuperación de la información diseñado e implementado en la Universidad de las Ciencias Informáticas con tecnologías libres que cumple con más de un año de utilización por la comunidad universitaria del país, contiene un gran número de funcionalidades que lo convierten en una herramienta potente para la búsqueda de información alojada en la web cubana, posee servicios desarrollados y otros en proyección de desarrollo que permitirán un avance significativo en la economía del país. Orión permite describir los componentes básicos de una herramienta de recuperación de la información, así como algunos datos de las tecnologías utilizadas para su desarrollo, además brinda un resumen de su impacto económico-social y algunos de los avales de los que ha sido meritorio este software.

Palabras Clave: indexación, rastreo, buscador, consultas, recuperación de información

Abstract

The information retrieval systems or browsers are commonly known for the Internet users, these systems are a source of access to information that is distributed in the world of the web as well as the services that are provided in this, its

importance and necessity in the world of internet surfing is determined because they are crawlers of information wich is then stored and subsequently accessible through interfaces and search functionalities and ordered according to mathematical algorithms used to calculate the relevance of the results. Orion is a system of information retrieval designed and implemented at the University of Information Science and complying with more than one year of use by the university community of the country, it contains a number of features that make it a powerful tool for finding information housed in Cuban web, has developed services and other development projection that will allow a breakthrough in the country's economy. This search engine allows us to describe the basic components of a tool for information retrieval and some data of the technologies used for its development, this article also provides a summary of its economic and social impact and some of the guarantees of this have been worthwhile software.

KeyWords: i*ndexing, crawling, searching, queries, information retrieval*

Introducción

Cuba está inmersa en un amplio proceso de informatización de la sociedad, con el objetivo de poner al alcance de todos, herramientas y vías de acceso a los servicios y tecnologías brindadas por el amplio mundo de las Tecnologías de la información y las Comunicaciones. Las herramientas de recuperación de información o buscadores como son conocidos comúnmente por los internautas son una fuente acceso a la información alojada en la web así como los servicios que en esta se brindan; su relevancia en el mundo de la navegación por internet está determinada debido a que son esencialmente rastreadores de información que luego es almacenada y posteriormente accesible a través de consultas y filtros y ordenada según algoritmos matemáticos usados para calcular la relevancia de los resultados obtenidos (Kowalski, 1998). Varios autores (Baeza et al., 1999; Cuesta, 2000; Gutiérrez, 2009; Pinto, 2011; Medina et al, 2011; Castells, 2011) plantean que los sistemas de recuperación de información o buscadores son herramientas que permiten el acceso a la información alojada en la web y tienen como principal objetivo ayudar al usuario en el proceso de recuperar información con un alto valor y calidad acorde a su necesidad. El funcionamiento de los buscadores web está asociado a los principios de la Recuperación de Información que tiene como premisas la representación, almacenamiento, organización y acceso a elementos de información (Baeza et al., 1999). Un buscador emplea 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 (Jaimes y Vega, 2005; Betancur, 2009; Blázquez, 2013). La información en la web se encuentra dispersa, para acceder a ella los buscadores emplean mecanismo de rastreos que obtienen la información y la almacenan para su posterior procesamiento. Los principales buscadores web emplean técnicas avanzadas de recuperación de información que, mediante el análisis de la necesidad

del usuario, recuperan la información disponible y selección la que consideran más relevante para los usuarios. La información que obtiene el usuario en ocasiones no es la de mayor calidad, debido principalmente a los mecanismos de ordenamientos que favorecen o penalizan en dependencia de las políticas implementadas en el buscador.

El objetivo del presente trabajo investigativo es realizar un estudio de los principales componentes de los sistemas de recuperación de información, específicamente se exponen las características del buscador Orión, desarrollado en la Universidad de las Ciencias Informáticas, haciendo uso de tecnologías libres. Este sistema tiene como meta la recuperación y visualización de la información alojada en la web cubana, actualmente se encuentra en uso por la red universitaria del país y ha sido meritorio de muchos premios y avales por la importancia de las funcionalidades que brinda para todos los usuarios.

Contenido

Los sistemas de recuperación de la información (SRI) están compuestos esencialmente por tres componentes principales como muestra la **figura 1**, la integración de estos componentes permite crear una herramienta cuyo objetivo es básicamente rastrear toda la web en busca de toda la información que se encuentra dispersa en la misma, posteriormente procede a su almacenamiento y luego brinda a través de interfaces o servicios los resultados solicitados por los usuario haciendo uso de las consultas definidas por los criterios de búsqueda insertados por los internautas.

La arquitectura del buscador Orión desarrollado en la Universidad de las Ciencias Informáticas está sustentada por los tres componentes definidos anteriormente.

El componente de recolección o *spider* web es el encargado del rastreo de la información que se encuentra dispersa por toda la red. Como *spider* se utiliza **Nutch** debido a las múltiples ventajas que ofrece a lo largo de todo el proceso de rastreo y almacenamiento de la información.

Revista Cubana de Ciencias Informáticas
Vol. 10 No. Especial Informática 2016, mayo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 150-162

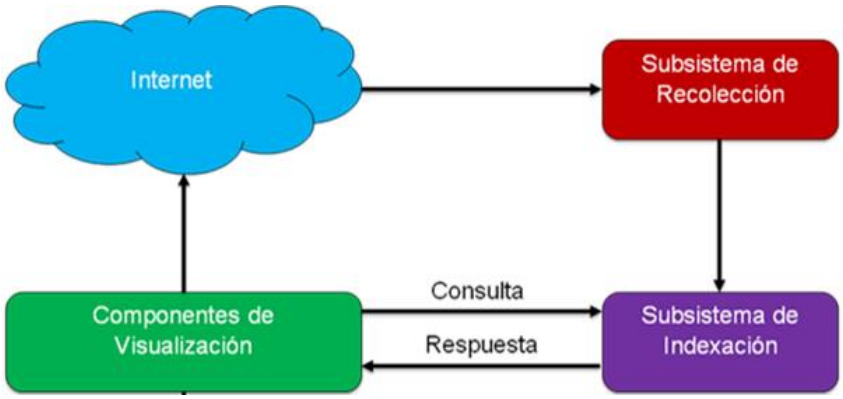

Figura 1: Componentes principales de un (SRI). Elaboración propia

Nutch: web *crawler* libre y de código abierto desarrollado en Java bajo la licencia de Apache. Este proporciona interfaces extensibles para implementaciones personalizadas. Inicialmente fue implementado sobre la base de Apache Lucene, aunque ya la versión actual es independiente de Lucene. Inicialmente fue implementado sobre la base de Apache Lucene aunque ya la versión actual es independiente de Lucene, una librería de alto rendimiento para la búsqueda basada en texto y que utiliza una modificación del algoritmo "*Vector Space Model*" (Modelo de Espacio Vectorial en español), con un enfoque booleano que restringe las estimaciones de los resultados obtenidos (Nieto, 2009). La arquitectura de Nutch es flexible permitiendo realizarle mejoras por parte de los usuarios a través de plugins. Este es independiente del servidor de indización lo que permite la integración con Solr (NutchWiki 2015). Nutch, como mecanismo de rastreo, posee ciertos componentes llamados *parsers*, los cuales se encargan de descomponer las páginas web y analizar cada uno de los recursos que la componen o tienen relación. Uno de estos *parsers* se denomina Tika, el cual puede descomponer una gran cantidad de documentos, entre ellos HTML, documentos ofimáticos, pdf y muchos más. Actualmente, aunque Tika soporta una variedad de formatos sobre una gran cantidad de tipos de documentos, no es capaz de obtener toda la información que se pudiera obtener de las imágenes que encuentra, constituyendo una de sus debilidades (The Apache Software Foundation-Tika, 2014).

Ventajas del uso de Nutch (NutchWiki 2015)**:**

- Licencia Apache Software Foundation (ASF).
- Recolección de información en formato PPT, DOC, PDF, HTML, XML, TXT, RTF, GIF, JPG, PNG, entre otros.
- Ejecución en paralelo (multihilo).
- Arquitectura distribuida.

- Extensibilidad.
- Multiplataforma.
- Variante del modelo Espacio Vectorial.
- Lenguaje Java.
- Configuración avanzada.
- Amplia y fuerte comunidad de desarrollo (Apache).
- Perspectivas de desarrollo.
- Documentación (Idioma inglés).

El componente de indexación es el encargado de recepcionar toda la información rastreada por el *spider* y luego procesarla y almacenarla (Cleverdon, 1997). En Orión se utiliza **Solr** como componente de indexación por las múltiples ventajas que ofrece este software.

Solr: es un motor de búsqueda de código abierto basado en la biblioteca Java del proyecto Lucene, con APIs en XML/HTTP y JSON, resaltado de resultados, búsqueda por facetas, caché, y una interfaz para su administración (Apache Software Fundation, 2015).

Ventajas de uso de Solr (Medina, Martínez y Delgado, 2011)**:**

- Capacidades avanzadas de búsqueda a texto completo.
- Optimizado para elevados volúmenes de tráfico.
- Interfaces abiertas basadas en estándares abiertos (XML, JSON, HTTP).
- Flexibilidad y adaptabilidad a través de extensas opciones de configuración.
- Búsquedas facetadas y filtrado.
- Análisis de texto configurable.
- Caché altamente configurable.
- Soporte para indizar varios tipos de documentos (PDF, Word, HTML, etc.)
- Extracción de metadatos.
- Soporte para varios núcleos.
- Sobresaltado de los resultados.
- “Más sobre eso” para un documento dado.
- Auto-sugerencias para completar las consultas de los usuarios (Apache Software Fundation, 2015).
- Escalabilidad – Replicación eficiente hacia otros Servidores de Búsqueda de Solr (SETA, 2010).

Revista Cubana de Ciencias Informáticas
Vol. 10 No. Especial Informática 2016, mayo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 150-162

Para el desarrollo del componente de visualización se utiliza el marco de trabajo o *framework* por su terminología en inglés, Symfony, debido a la potencia que brinda en el desarrollo web de sistemas.

Symfony: es un *framework* PHP de tipo *full-stack* construido con varios componentes independientes creados por el proyecto Symfony (GUILUZ, 2013).

Ventajas del uso de Symfony como marco de trabajo (Guiluz, 2013)**:**

- Los formularios soportan la validación automática, lo cual asegura mejor calidad de los datos en las bases de datos y una mejor experiencia para el usuario.
- El manejo de cache reduce el uso de banda ancha y la carga del servidor.
- La presentación usa *templates* y *layouts* que pueden ser construidos por diseñadores de HTML que no posean conocimientos del *framework*.
- Los *plugins* proveen un alto nivel de extensibilidad.
- Las herramientas que generan automáticamente código han sido diseñadas para hacer prototipos de aplicaciones y para crear fácilmente la parte de gestión de las aplicaciones.
- El *framework* de desarrollo de pruebas unitarias y funcionales proporciona las herramientas ideales para el desarrollo basado en pruebas.
- La barra de depuración web simplifica la depuración de las aplicaciones, ya que muestra toda la información que los programadores necesitan sobre la página en la que están trabajando.
- La interfaz de línea de comandos automatiza la instalación de las aplicaciones entre servidores.
- Es posible realizar cambios "en caliente" de la configuración (sin necesidad de reiniciar el servidor).
- El completo sistema de log permite a los administradores acceder hasta el último detalle de las actividades que realiza la aplicación.

Descripción de las interfaces y sus funcionalidades

La Interfaz principal de Orión figura 2 brinda las principales opciones de búsqueda del sistema al usuario permitiéndole acceder a sus funcionalidades más relevantes desde la primera interfaz del buscador. Esta interfaz permite insertar un criterio de búsqueda haciendo uso de un formulario con un campo de texto y brinda opciones de filtrado para la búsqueda haciendo uso del menú superior izquierdo, el filtrado permite especializar las búsquedas en tres categorías:

- Búsqueda general.
- Búsqueda de documentos.

Revista Cubana de Ciencias Informáticas
Vol. 10 No. Especial Informática 2016, mayo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 150-162

- Búsqueda de imágenes.

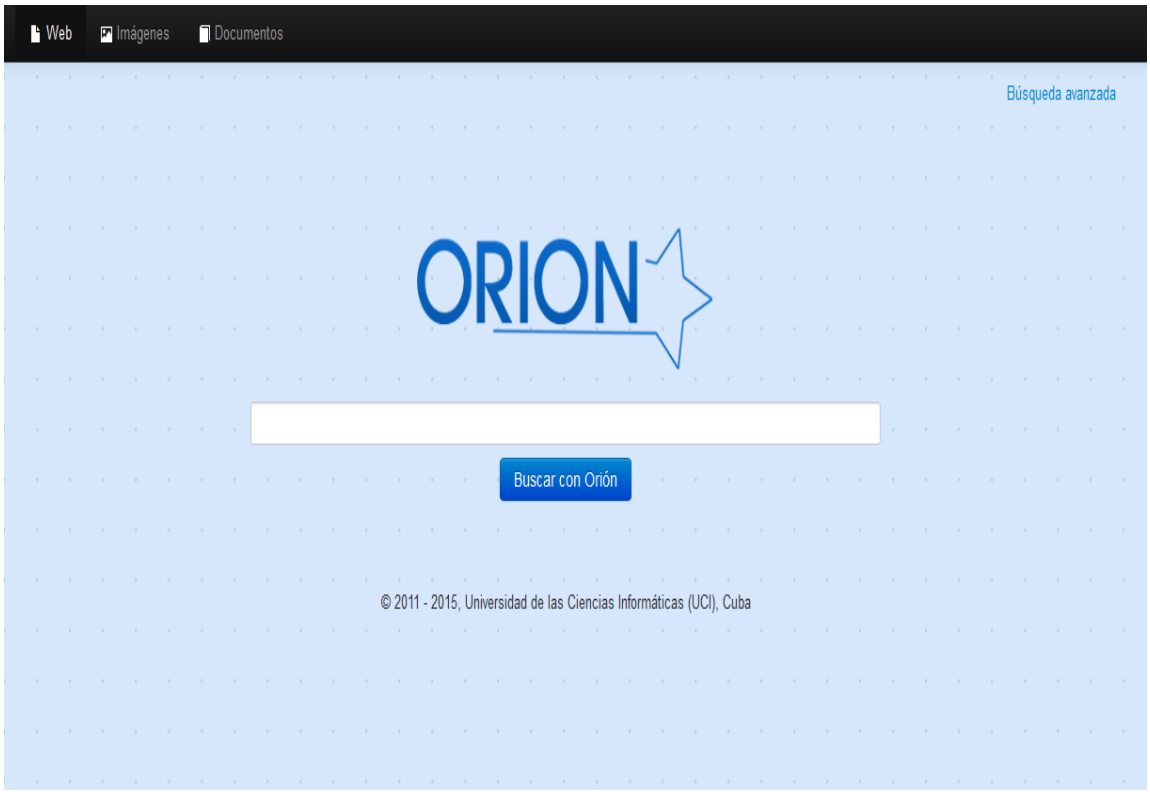

Figura 2: Interfaz principal de Orión

La funcionalidad de búsqueda avanzada figura 3 permite realizar una búsqueda especializada teniendo en cuenta 9 filtros principales:

- **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.

Revista Cubana de Ciencias Informáticas
Vol. 10 No. Especial Informática 2016, mayo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 150-162

- **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
    - Último año
- **Términos que aparecen**: permite obtener resultados que contengan el criterio de búsqueda en distintas aéreas de las páginas donde se encontraron:
    - Título de la página.
    - Contenido de la página.
    - Url de la página.

Figura 3: Interfaz de búsqueda avanzada

La interfaz de resultados **figura 4** permite al usuario obtener los resultados relativos al criterio de búsqueda insertado y ordenados mediante una relevancia calculada internamente por el sistema. El paginado ubicado en la parte inferior izquierda permite acceder a distintos intervalos de resultados, una funcionalidad muy útil teniendo en cuenta que el sistema debe contar con una usabilidad correcta.

Revista Cubana de Ciencias Informáticas
Vol. 10 No. Especial Informática 2016, mayo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 150-162

Los resultados como muestra la **figura 5** están estructurados de la siguiente manera:

- Título
- Resumen
- Url

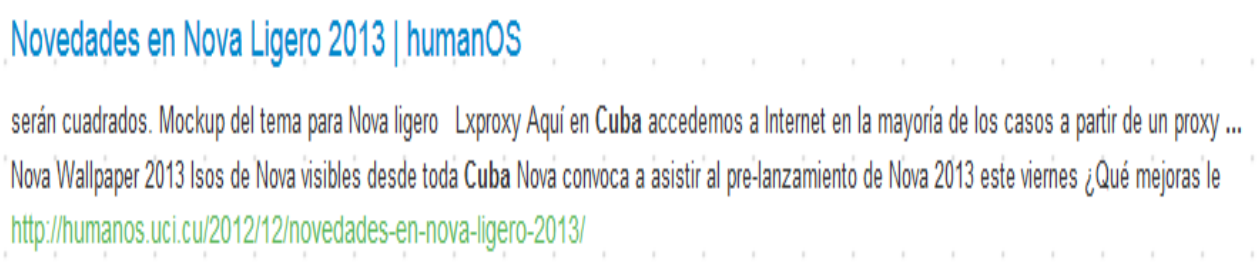

Figura 5: Estructura de resultados

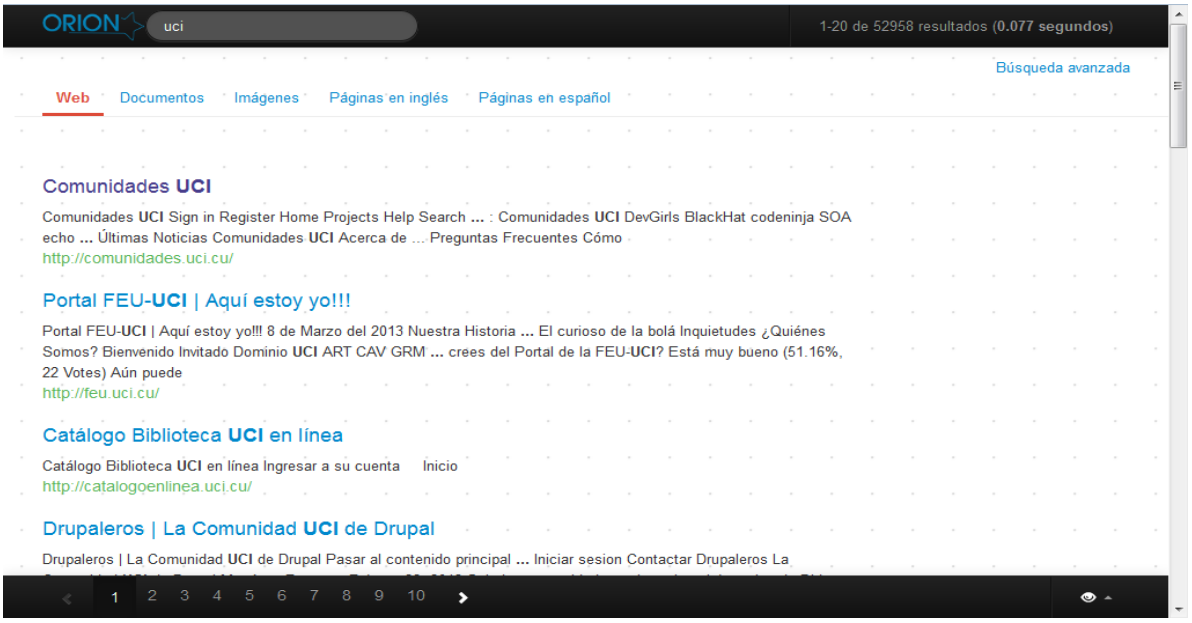

Figura 4: interfaz de resultados

En la interfaz de resultados se mantienen los filtros de búsqueda web, documentos, imágenes y se añade un filtrado para resultados en inglés o español. La **figura 6** muestra la página de resultado para búsqueda de imágenes, cada resultado muestra una miniatura de la imagen original y debajo la url de donde se obtuvo dicha imagen.

En cada una de las distintas páginas de resultados ya sea la de búsqueda general, documentos o imágenes, se mantiene una barra superior donde el usuario puede hacer uso de un formulario posicionado en la parte izquierda para insertar

Revista Cubana de Ciencias Informáticas
Vol. 10 No. Especial Informática 2016, mayo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 150-162

nuevos criterios de búsqueda y en l aparte derecha algunos datos que muestran el intervalo de resultados actual, el total de resultados para la consulta insertada y el tiempo que demoró en ejecutarse dicha consulta.

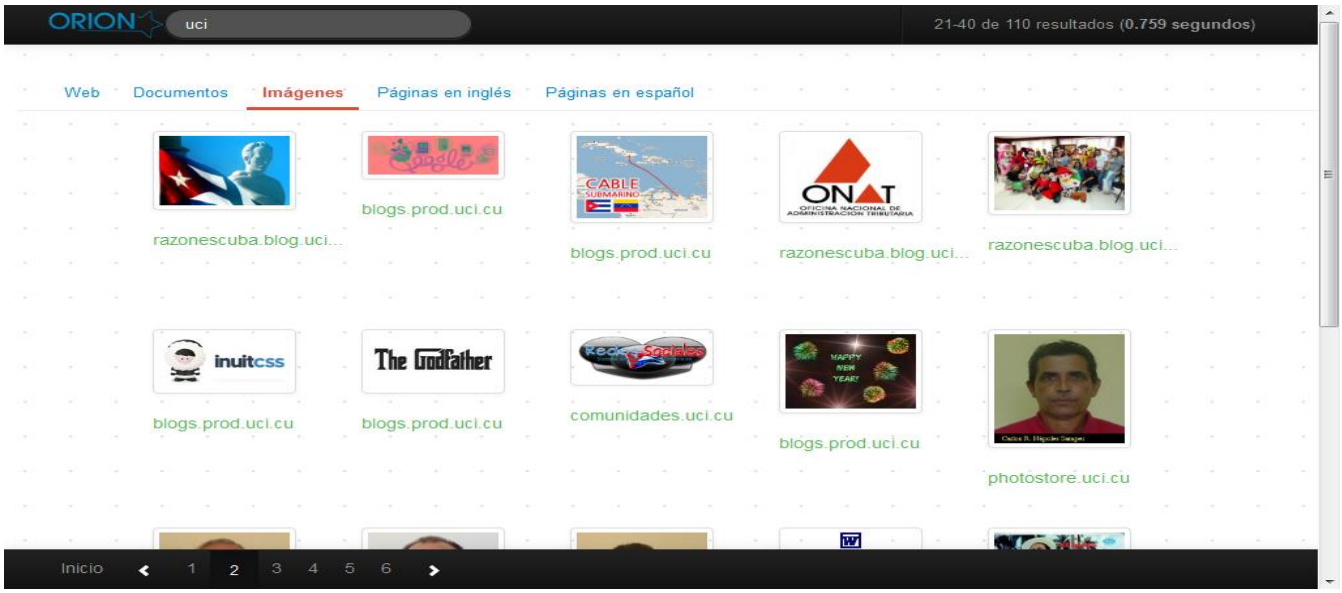

Figura 6: Resultados de imágenes

Impacto Económico y Social

El motor de búsqueda Orión pone en las manos de todos los cubanos una herramienta de búsqueda de información desarrollada bajo tecnologías del software libre que respeta las buenas prácticas de posicionamiento, garantiza resultados legítimos de nuestro país y añade una victoria más en el logro de la soberanía tecnológica de Cuba. Cumpliendo con los lineamientos 131 y 226 trazados por el Partido Comunista de Cuba enfocados en logran una informatización eficiente de la sociedad y brindar un servicio de alta calidad, los centros educacionales tendrán una herramienta más de apoyo a la formación de los estudiantes, contribuyendo al proceso de enseñanza y aprendizaje que lleva a cabo el Ministerio de Educación para cumplir con el objetivo de dotar al país de una sociedad culta e innovadora. En el ámbito de la salud podrá crearse una red de investigaciones sobre el avance científico y técnico que ha venido desarrollando Cuba a lo largo de los años de revolución, permitiendo crear un entorno colaborativo sobre investigaciones relativas a vacunas y curas de enfermedades.

En el campo investigativo permite crear un espacio centralizado que permite el acceso a publicaciones, artículos, tesis e investigaciones de corte académico que permitirá poner en manos de todos los estudiantes, profesionales e investigadores un gran cúmulo de conocimiento fruto del desarrollo investigativo alcanzado en el país.

El buscador en su esencia puede ser utilizado como medio de enseñanza en la impartición de clases y conferencias debido al rápido y confiable acceso a la información que brinda.

Actualmente se encuentran en desarrollo servicios que económicamente pueden ser una fortaleza para el país:

Módulo de publicidad y propaganda: que serviría para brindar un marco de publicidad y promoción a los servicios que se brindan en Cuba permitiéndoles ser conocidos por todos los usuarios de la red y logrando un mayor número de clientes que inviertan en sus productos.

Módulo de traducción: permite traducir contenidos y páginas web, algunos sitios sólo muestran sus servicios en idioma español, con este módulo se hacen extensivos estos servicios a los hablantes de la lengua inglesa.

Conclusiones

Se desarrolló un sistema de recuperación de la información llamado buscador Orión que cuenta con tres componentes fundamentales, rastreo, indexación y visualización. Con el despliegue del Motor de Búsqueda Orión en el país se logra poner en manos de todos los usuarios de la red de una herramienta potente de búsqueda diseñada y programada en nuestro país. La comunidad universitaria posee acceso a una forma eficiente de realizar investigaciones científicas así como estudios de maestrías y doctorados, el uso de los servicios ya programados y las nuevas proyecciones de desarrollo será un pilar fundamental en la economía del país, permitirá divulgar anuncios publicitarios que impulsarán los servicios brindados por las empresas cubana, se obtendrán datos precisos sobre el estado de la web cubana así como ideas bien definidas para la mejora de la misma, Orión es para Cuba una fortaleza más en la lucha por la soberanía tecnológica.

Referencias bibliográficas

APACHE SOFTWARE FUNDATION. Disponible en: The Apache Solr Reference Guide. [En línea]. Apache Solr – Resources, 2015. Disponible en: [http://lucene.apache.org/solr/documentation.html].

BAEZA-YATES, R. y RIBEIRO-NETO, B.: Modern InformationRetrieval. New York: ACM;Harlow, Essex: Addison-Wesley Longman, 1999

BETANCUR, D.; MORENO, J. y OVALLE, D., Modelo para la recomendación y recuperación de objetos de aprendizaje en entornos virtuales de enseñanza/aprendizaje. Revista Avances en Sistemas e Informática Vol 6 No 1. (2009)

BLÁZQUEZ OCHANDO, Manual Técnicas avanzadas de recuperación de información: procesos, técnicas y métodos. Madrid. ISBN 978-84-695-8030-1, 2013

CASTELLS, P, et al. Recuperación y almacenamiento de información en la web, Escuela Politécnica Superior Universidad Autónoma de Madrid, 2011.

CASTELLS, P., FERNANDEZ, M. y VALLET, D. (2007). AnAdaptation of the Vector-SpaceModelforOntology-BasedInformationRetrieval. IEEE TransactionsonKnowledge and Data Engineering. 19 (2). p.pp. 261-272.

CLEVERDON, C.: "The Cranfield tests on indexlanguages devices", Readings in information re-trieval. San Francisco: Morgan Kaufmann, pp. 47 -59, 1997.

CUESTA MORALES, P, et al. Aplicación de Técnicas de Recuperación de Información a un Glosario de Términos de Internet Desarrollado Utilizando Tecnología JSP, Universidad de Vigo, 2000

GUILUZ, J. Desarrollo web ágil con Symfony2, 2013. pág. 618.

GUTIÉRREZ VARGAS, M. E. et al., Los procesos de búsqueda de información, Universidad Autónoma Metropolitana-Xochimilco, 2009, pp. 9.

JAIMES, L. G y VEGA RIVEROS. F. Modelos clásicos de recuperación de la información, 2005

KOWALSKI, G.: Information retrieval systems:theory and implementation, Computers and Mathematics with Applications, vol. 5, No. 35, pp. 133,1998.

Medina García, Anay, Meylin Martínez Chong, y Yusniel Hidalgo Delgado. 2011. "Propuesta Arquitectónica de Un Sistema de Recuperación de Información Geográfica Para El Motor de Búsqueda Orión." http://repositorio_institucional.uci.cu//jspui/handle/ident/TD_04252_11.yo

NIETO, I., A., M. Universidad Nacional de Colombia. [En línea] 2009. [Citado el: 10 de octubre de 2014.]. Disponible en: [http://dis.unal.edu.co/profesores/eleon/cursos/tamd/presentaciones/nutch.pdf].

NutchWiki. 2015. "NutchTutorial - Nutch Wiki." https://wiki.apache.org/nutch/NutchTutorial.

PINTO MOLINA, MARIA. 2011. "Busqueda y Recuperación de Información." April13. http://www.mariapinto.es/e-coms/recu_infor.htm#ri2.

SETA, L., D. Apache Solr: una introducción. [En línea]. Apache Solr: una introducción - Dos Ideas. 2010. [Citado el: 3 de 10 de 2014.]. Disponible en: [http://www.dosideas.com/noticias/java/913-apache-solr-una-introduccion.html]

THE APACHE SOFTWARE FOUNDATION-TIKA. Apache Tika. [En línea]. Apache Solr, 2014. [Citado el: 13 de octubre de 2014.] Disponible en: [http://tika.apache.org/].