Análisis comparativo de herramientas de recuperación y análisis de información de acceso libre desde una concepción docente[1]

Comparative analysis of information retrieval and analysis of open access tools from an educational concept

Armando PLASENCIA-SALGUEIRO[2]
Bárbara de los Milagros BALLAGAS-FLORES[3]

Resumen

En el Instituto de Cibernética, Matemática y Física de la República de Cuba se imparte el curso "Bases de datos y biblioteca digital" dentro de la Maestría de Cibernética Aplicada. Parte esencial del curso la constituye la creación de bases de datos documentales a partir de la recuperación de información en Internet. Para poder realizar los laboratorios requeridos para un mejor aprendizaje se requiere seleccionar las herramientas de recuperación de información más idóneas, tanto desde el punto de vista docente como desde las facilidades para su adquisición. Se definieron entonces las características para evaluar esas herramientas y la metodología para realizar la selección. Como resultado, de trece herramientas de recuperación y análisis de la información de *software* libre analizadas que pudieron ser descargadas se seleccionaron ocho herramientas, *Lemur Toolkit* con *Indri*, *Sphinx*, *WebSphinx* con *Rapid Miner*, *Solr/Lucene/Hadoop/Mahout*, *Terrier*, *Dragon* lo cual permitió garantizar la calidad del curso impartido y su concatenación con otros cursos de la misma maestría.

Palabras ilave: Bases de datos. Minería de textos. Motores de búsqueda. Recuperación de información.

Abstract

In the Institute of Cybernetics, Mathematics, and Physics in the Republic of Cuba the course "Databases and digital Library" is a discipline in the Master's degree program of Applied Cybernetics. An essential part of the course is the creation of documental databases starting from information retrieval from the Internet. To equip the laboratories required for better learning, the most suitable tools for information retrieval are needed, both from an educational point of view as well as the easiness for their acquisition. Therefore, the characteristics to evaluate these tools and the methodology for selecting them were defined. As a result, of the thirteen recovery tools and data analysis from free softwares available to be downloaded, the following eight tools were selected: Lemur Toolkit with Indri, Sphinx, WebSphinx with Rapid Miner, Solr / Lucene / Hadoop / Mahout, Terrier and Dragon, which guaranteed the quality of the course and the connection with other courses in the Master's degree program.

Keywords: *Database. Text mining. Searching engines. Information retrieval.*

---

[1] Trabajo presentado en el VII Seminario Internacional sobre Estudios Cuantitativos y Cualitativos de la Ciencia y la Tecnología "Prof. Gilberto Sotolongo Aguilar" en XIII Congreso Internacional de Información - INFO' 2014. Habana, Cuba.

[2] Instituto de Cibernética, Matemática y Física, Departamento de Control Automático. Dirección Postal 10400, La Habana, Cuba. Correspondencia a nombre de/*Correspondence to*: A.P. SALGUEIRO. *E-mail*: <armando@icimaf.cu>.

[3] Empresa de Telecomunicaciones de Cuba SA., Departamento de Inteligencia Empresarial. La Habana, Cuba.

Recibido el día 10/6/2014 y aceptado para su publicación el 12/9/2014.

Introducción

La Cibernética es la ciencia sobre el control, la obtención, trasmisión y conversión de información en los sistemas cibernéticos. La principal tarea de la Cibernética es la elaboración de la estructura y los métodos de investigación, útiles para el estudio de los sistemas de control independientemente de su naturaleza (Dopico & Plasencia, 2011).

Desde el 2012 se imparte la maestría "Cibernética Aplicada" <www.icimaf.cu/maestria-ca> en el Instituto de Cibernética, Matemática y Física del Ministerio de Ciencia, Tecnología y Medio Ambiente. Esta tiene como objetivo la enseñanza y actualización de los profesionales en las técnicas de descubrimiento de conocimiento basados en datos *Knowledge Data Discovery* (KDD) y su aplicación tanto para la Inteligencia organizacional (en la Mención Minería de Datos de la Maestría), como en la modelación, la identificación, la robótica, el diagnóstico y el diseño de sistemas de control inteligente (en la Mención Control Avanzado).

La concepción del diseño curricular de la maestría se fundamentó en el enfoque sistémico, la enseñanza problémica, la interdisciplinariedad, la multidisciplinariedad, la transdisciplinariedad y la confluencia de las diferentes técnicas para el Control Avanzado y el análisis de datos (minería de datos) (Dopico & Plasencia 2011).

Para respaldar el principio de enseñanza problémica, basado en el criterio de los especialistas al abordar un problema concreto en una determinada organización, la ciencia o la tecnología, se propicia el empleo de herramientas tanto de software libre como propietarias de acceso libre para fines docentes en la realización de laboratorios y ejercicios en los cursos de la maestría con el objetivo también de afianzar los conocimientos teóricos impartidos.

En el curso "Bases de datos y bibliotecas digitales" se imparten entre otros los siguientes temas: recuperación de Información clásica (M1); lenguajes de solicitud de búsquedas (M2); indizado y búsqueda, ficheros invertidos, consultas booleanas, búsquedas secuenciales, reconocimiento de patrones, consultas estructurales, compresión (M3); recuperación de información multimedia, modelos y lenguajes. Indexado y búsqueda; buscadores y búsqueda en la *Web*.

- *Web* Semántica

- Recuperación de información en redes sociales (M4 y M5)

Se hace necesario entonces, determinar qué herramientas de software se corresponden con los laboratorios y clases prácticas que respondan a los contenidos expuestos con anterioridad.

Sin embargo, ello no es una tarea fácil. La cantidad de herramientas que existen, solo de finales de los 90 a la actualidad es considerable, sus contenidos y alcances varían prácticamente todos los años y para impartir docencia se necesita que éstas estén asequibles en internet a los estudiantes, que estén bien documentadas, que las ayudas sean comprensibles, que existan preferiblemente libros sobre las mismas, que proporcionen ejemplos y estos sean ilustrativos, comprensibles y que su nivel de complejidad sea paulatino.

Por otro lado, las herramientas existentes o son de uso comercial o están muy orientadas a la investigación lo que hace muy confusa la selección de las herramientas adecuadas.

Para darle solución a esta problemática, partimos de la premisa que es posible realizar un análisis cuantitativo y cualitativo de las características de las herramientas para seleccionar las más idóneas para el proceso docente. De ahí que los objetivos que dimanan de esta premisa es el determinar durante la preparación del curso, cuáles son las herramientas que son más adecuadas para la impartición de los laboratorios y establecer una metodología que permita monitorear este proceso de selección.

Trabajos anteriores que aborden esta problemática específicamente desde el punto de vista docente y en una escala de una cantidad suficientemente exhaustiva (más de 10) no hemos encontrado ninguno. Ya desde un punto de vista más general, se destaca el trabajo de (Middleton & Baeza-Yates, 2011) que hace una comparación de 17 motores de búsqueda de acceso libre y la colección de artículos recogidos en la recopilación de trabajos del evento "*Conference on Research and Development in Information Retrieval*" (*Association for*

 http://dx.doi.org/10.1590/0103-37862014000300008

Computing Machinery/Special Interest Group on Information Retrieval 2012) (Trotman *et al*., 2012). El inconveniente de estos trabajos es que el primero es del 2007, por lo que estas herramientas han cambiado mucho en siete años y en el segundo caso son trabajos que tratan uno, dos motores de búsqueda.

Recuperación de Información (RI) (Middleton & Baeza-Yates, 2011) es un amplio campo de estudio que abarca la representación, el almacenamiento, la organización y el acceso a los elementos de información. La RI debe de ser capaz de tratar la información de forma tal que se pueda acceder a ella de forma eficiente, enfocada a las necesidades de información del usuario. Una definición más detallada es la de que la RI es la búsqueda de material de información (generalmente documentos) de naturaleza no estructurada (generalmente textos) que satisface una necesidad de información a partir de una gran colección de datos (generalmente en los servidores de computadoras locales o en Internet).

El proceso de RI se puede desglosar en la interacción de los módulos que se dan en la Figura 1. En esta se destacan tres áreas principales, el indizado, la búsqueda y el ordenamiento:

El análisis consiste en la determinación de nuevos patrones y relaciones entre documentos, no triviales y desconocidos. La interrelación entre la RI y el análisis se puede ver en la Tabla 1.

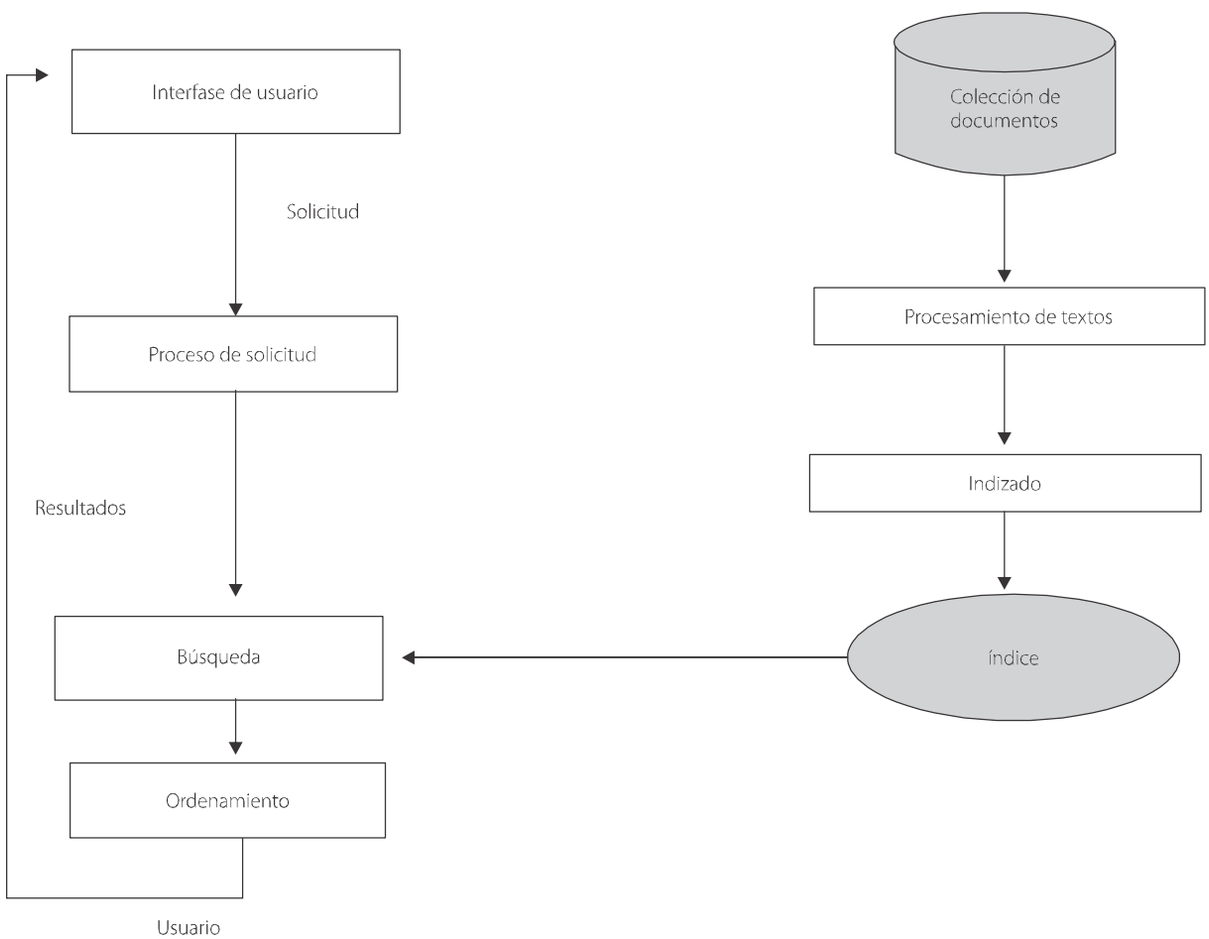

Figura 1. El proceso de recuperación de información.

Fuente: (Middleton & Baeza-Yates, 2011).

Tabla 1. Interrelación entre la recuperación de información y el análisis.

Fuentes de Datos/Información			
Propósito	Cualquier Dato	Datos textuales	Datos Relativos a la *Web*
Recuperación de datos o documentos eficiente y efectivamente	Recuperación de datos/ Bases de datos	Recuperación de Información	Recuperación de la *Web*
Hallazgo de nuevos patrones o conocimiento desconocido con anterioridad para el sistema	Minería de datos	Minería de textos	Minería *Web*

Fuente: <www.knowledgetechnologies.net/proceedings/presentations/treloar/nathantreloar.ppt>, 2008.

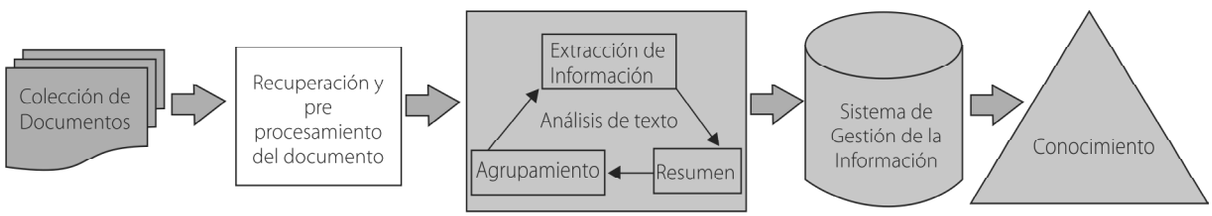

Figura 2. Minería de textos.

Fuente: (Fan *et al*., 2005).

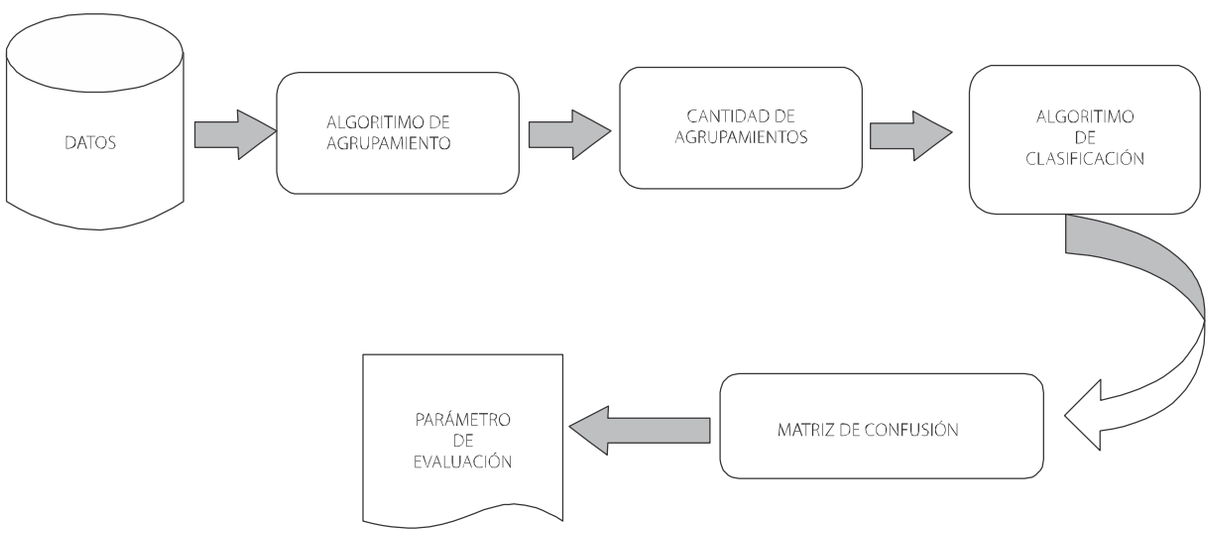

Figura 3. Técnicas de agrupamiento-clasificación.

Fuente: (Pathrey *et al*., 2013).

Se entiende por minería de textos al proceso de extracción automática de información no trivial, útil, previamente desconocida y finalmente comprensible de los depósitos de documentos textuales. La interrelación de la recuperación de información con el análisis para obtener nuevo conocimiento se da en la Figura 2.

 http://dx.doi.org/10.1590/0103-37862014000300008

Dentro de la minería de textos, los tipos de algoritmos que más se emplean son los de agrupamiento y de clasificación, en un proceso relacionado en donde el algoritmo de agrupamiento generalmente es un proceso no supervisado y la clasificación es un proceso supervisado como se muestra en la Figura 3.

Métodos

Para realizar la investigación se hizo un análisis de 13 herramientas de recuperación y análisis. Los criterios que se tomaron para la selección fueron el de que fuesen herramientas de Recuperación de información y/o análisis de información mediante algoritmos de minería de textos (aprendizaje automático) de *software* libre, con código abierto, que independientemente de ser libres no tuviesen restricciones en su descarga, lo cual se verificó mediante la descarga física de los software.

Las características de las herramientas se sometieron a un análisis cuantitativo y cualitativo. Para el análisis cuantitativo se utilizaron los parámetros de las comparaciones entre herramientas de *software* que establece el sitio *Web* (Findthebest, 2014). Estos resultados tienen la ventaja de que son contrastables por los usuarios del sitio, dinámicos y de fácil obtención. Su deficiencia radica en que no son transparentes para el usuario en relación con la procedencia inicial de los mismos como si lo pueden ser, por ejemplo, los resultados obtenidos a partir de la ejecución sobre colecciones de datos. Pero esa metodología está fuera del alcance de los recursos y de la propia investigación. Otra deficiencia es que no se incluyen todas las herramientas seleccionadas.

Entre las características que entrega el sitio tenemos: características generales de la herramienta, uso previsto, multiplataforma, multilingüe, identificación conceptual, categorización por relevancia, especificaciones de software (licencia, sistema operativo, código programado en, enlace de descarga); y se añaden por los autores los siguientes indicadores: Herramientas de Análisis, Redes Sociales, Manejo información multimedia, *Web* Semántica.

Cuyos datos se obtienen de la información técnica de los sitios de hospedaje de las herramientas.

Se incorporan también elementos cualitativos de análisis como resultado de la experiencia de los autores en la impartición del curso durante tres cesiones y la interacción con los alumnos en los laboratorios. Estos son: Especificaciones docentes: facilidad en la ejecución de los algoritmos de RI, facilidad de ejecución de los algoritmos de análisis, visualización de los resultados, comunidad de discusión, facilidad de la ayuda, existencia de ejemplos, Libro/Manual/Idioma, empleo en los cursos.

Resultados y Discusión

Relación de las herramientas de recuperación y análisis seleccionadas

Toolkit <http://www.lemurproject.org/>: Lemur es una herramienta destinada para facilitar la investigación en la modelación del lenguaje y en la recuperación de información. Para ello incluye tecnologías como recuperación ad hoc y distribuida, consultas estructuradas, RI multilingüe, indización general, filtrado y categorización. Proporciona aplicaciones de ejemplo diseñándose la herramienta para permitir la fácil programación de sus propias actualizaciones y aplicaciones.

Características (Findthebest, 2014): sofisticado lenguaje estructurado de consulta; recuperación de documentos estructurados y en XML; indizado de las páginas *Web* con capacidad de búsqueda de sitios "*out-of-the-box*"; interfases interactivas para Windows, Linux y *Web*; múltiples métodos de indizado para colecciones pequeñas, medianas y grandes (*terabytes*). Indexado incremental; realimentación relevante y seudo relevante, recuperación multilingüe. API en C++. Java y C#; recuperación de información distribuida y aplicaciones de agrupamiento de documentos.

Sphinx <http://sphinxsearch.com/>. *Sphinx* es un motor de búsqueda de acceso abierto orientado en su diseño a la indización del contenido de las bases de datos. De forma nativa incorpora como gestores de bases de datos a *MySQL* y a *PostgreSQL*. Otras fuentes de datos pueden ser indexadas mediante conexión en un formato de edición en XML.

Características (Findthebest, 2014): alta velocidad de indizado y búsqueda; incluye un ordenamiento por proximidad de frases, proporcionando una buena relevancia; también incorpora cualquier número de campos de documentos y tratamiento de las palabras vacías.

The Dragon Toolkit <http://dragon.ischool.drexel.edu>. Descripción (Zhou *et al*., 2007): es un paquete de desarrollo escrito en Java para uso académico en RI y minería de textos *Text Mining* (TM) (incluye clasificación, agrupamiento, resumen y modelación). Está destinado a los investigadores que trabajan con tareas de RI y TM a gran escala y prefieren la programación en Java. Es diferente de Lucene y de Lemur, proporcionando un soporte incorporado para RI y TM basado en semántica. No incluye RI distribuida ni es multilingüe lo cual es propio de Lucene y Lemur. Incluye la obtención de ontologías.

Otra característica importante es su escalabilidad y la inclusión de herramientas de programación en lenguaje natural *Natural Language Programing* (NLP).

Text-Garden - Text-Mining Software Tools <http://ailab.ijs.si/dunja/textgarden/>. Descripción (Grobelnik, 2007): la herramienta de TM *Text-Garden* posibilita la fácil manipulación de los documentos textuales para su análisis incluyendo la generación automática de modelos y la clasificación, agrupamiento y visualización de documentos; gestiona documentos en la *Web*, rastrea la *Web* y otras muchas funciones. Escrita en código C++ y corre en Windows.

Es un *software* multiplataforma y tiene las siguientes interfases con la misma API: C/C++, *Java*. *.NET*, *Matlab*, *Python* y R.

Apache Lucene/Solr <http://lucene.apache.org>. Descripción: Lucene es una biblioteca de RI escalable, de altas prestaciones (Hatcher *et al*., 2010). Permite adicionar capacidades de búsqueda a las aplicaciones. Implementada en Java, es un proyecto en la *Apache Software Foundation*.

Las características más importantes de Lucene: un indizado invertido para la recuperación eficiente de los documentos por los términos indizados. La misma tecnología respalda a los datos numéricos al realizarse consultas de rango. Un rico conjunto de componentes de análisis de textos encadenables, tales como señalizadores y radicalizadores de lenguajes específicos que transforman una cadena textual en una serie de términos. Una sintaxis de consulta con un analizador sintáctico y una variedad de tipos de consulta desde la consulta de un término simple hasta un exótico pareo fuzzy. Un buen algoritmo de posicionamiento basado en principios fundamentados en la Recuperación de Información (IR) para generar primero el candidato más parecido, con significados flexibles para afectar el posicionamiento. Mejoramiento de las características de búsqueda.

Solr <http://lucene.apache.org/solr/>: Solr es un servidor de búsqueda empresarial de código abierto. El servidor realiza la comunicación utilizando los estándares *HyperText Transfer Protocol* (http), *eXtensible Markup Language* (XML), y *JavaScript Object Notation* (JSON).

El motor de búsqueda subyacente de Solr es Lucene

Solr tiene características propias que van más allá del propio Lucene (Plasencia *et al*., 2012). Estas son: ficheros de configuración, en particular por su esquema de índices, los cuales definen los campos y la configuración de sus análisis de textos; un analizador sintáctico de la consulta llamado *dismax* que es más utilizable para el análisis de las consultas del usuario final que el analizador sintáctico de consultas nativo de Lucene; búsqueda geoespacial para el filtrado y el ordenamiento por la distancia; una propiedad de búsqueda distribuida y de replicación del índice para la adaptación de Solr.

Swish-e <http://swish-e.org/index.html>. Descripción: Simple *Web* Indexing System for Humans – Enhanced (Swish-e) es un sistema rápido y flexible destinado para la indización de colecciones de páginas *Web* u otros ficheros. Swish-e es ideal para trabajar con colecciones de millones de documentos o menos. Utilizando el analizador sintáctico Gnome™ libxml2 y una colección de filtros, Swish-e puede indizar textos completos, correos electrónicos, *pdf, html, Microsoft® Word/PowerPoint/Excel* y cualquier fichero que pueda convertirse a texto xml o html. Es utilizado frecuentemente para complementar bases de datos del

 http://dx.doi.org/10.1590/0103-37862014000300008

tipo MySQL® *Database Management System* (DBMS) con la búsqueda rápida a texto completo.

Características: indizado rápido de un gran número de documentos en diferentes formatos incluyendo textos, html y xml; utiliza "filtros" para indizar otros tipos de ficheros tales como pdf, gzip, o *PostScript*; incluye un rastreador *web* para el indizado de documentos remotos en http. Sigue las reglas de exclusión de los Robots (incluyendo las Meta etiquetas).

Xapian <http://xapian.org/>. Descripción: es una biblioteca de un motor de búsqueda de libre acceso, distribuida bajo licencia *General Public License* (GPL).

Esta biblioteca posibilita la creación de una herramienta muy adaptable que permite a los desarrolladores adicionar fácilmente índices avanzados y facilidades de búsqueda a sus propias aplicaciones. Incluye el modelo de RI probabilístico así como un rico conjunto de operadores de consulta booleanos.

La última versión estable es la 1.2.16, distribuida el 2013-12-4

Características: escrita en C++; enlaces que permiten el uso de Php, Perl, Python, C#, Ruby; incluye el modelo de RI probabilístico (Okapi BM25); además de la biblioteca, tiene un número de programas de ejemplo, y una aplicación desarrollada para el indizado y la búsqueda (Omega).

Web SPHINX <http://www.cs.cmu.edu/~rcm/websphinx/>. Descripción: *Website-Specific Processorsfor html INformatione Xtraction* (*Web* SPHINX) es una biblioteca de clases de Java y un entorno interactivo para los rastreadores *Web*. Un rastreador *Web* (también denominado robot o araña) es un programa que navega y procesa las páginas *Web* de forma automática (Miller & Bharat, 1998).

Web SPHINX está compuesto de dos partes el *Crawler Workbench* y la biblioteca de clases de *Web SPHINX*.

El *Crawler Workbenches* es una interfasegráfica al usuario que permite controlar y adaptar el buscador *web*, un número determinado de operaciones comunes de acceso están incorporadas al rastreador posibilitando al usuario especificar y correr rastreos simples sin necesidad de programas.

La biblioteca de clases de *Web SPHINX* proporciona los elementos para escribir rastreadores de la *Web* en Java.

RapidMiner <www.rapid-i.com>: Descripción: RapidMiner/Yale: Líder en las herramientas de minería de datos debido a la combinación de su tecnología y del umbral de funcionalidades disponibles. Cubre una amplia gama de algoritmos de Minería de Datos (MD) incluyendo los algoritmos nativos de *Waikato Environment for Knowledge Learning* (WEKA). Además de ser una herramienta flexible para aprender y explorar la minería de datos, la interfaz gráfica de usuario tiene como objetivo simplificar el uso para las tareas complejas de esta área (Cardoso *et al*., 2011).

Características de *RapidMiner*: *RapidMiner* es un sistema prototipado para el descubrimiento del conocimiento y Data Mining; usa el lenguaje de scripting xml para describir los operadores y su configuración; posee una interfaz gráfica, línea comando, y API de Java para usar *RapidMiner* desde los propios programas; una gran cantidad de extensiones (plugins); las aplicaciones incluyen: *Text Mining*, *Multimedia Mining*, entre otras.

El TextInputPlugin de Rapid Miner proporciona un rastreador *Web* simple sobre la base de *WebSPHINX Open Sourcecrawler* escrito en Java.

Indri <http://www.lemurproject.org/>. Descripción: el lenguaje de consulta Indri, basado en el lenguaje de consulta Inquery fue diseñado para ser un lenguaje robusto. Este puede trabajar tanto con una consulta simple como con una consulta muy compleja. Tal lenguaje de consulta sitúa a Indri en un lugar diferente del resto de los motores de búsqueda. Éste permite la comparación con frases complejas, sinónimos, expresiones ponderadas, filtrado booleano, campos numéricos (y fechas) y el uso extensivo de documentos estructurados (campos), entre otros.

Indri es un motor de búsqueda de textos desarrollado en la *University of Massachusetts*. Éste es parte del proyecto Lemur.

Por último, si Lucene/Solr es atractivo porque es un paquete con casi todas las características necesarias. Para realizar aplicaciones basadas en la *Web*, Indri es atractivo porque ofrece mejores resultados de búsqueda y como lenguaje de consulta altamente expresivo, lo que

permite una buena granulación del control de la búsqueda (Turtle *et al*., 2012).

TERRIER <http://terrier.org>. Descripción (Middleton & Baeza-Yates, 2011): TERabyte RetrIEveR (TERRIER) es una plataforma modular que permite un rápido desarrollo de los motores de búsqueda para la *Web*, la intranet o las PC desarrollado por la Universidad de *Glasgow* en Escocia. Éste tiene la posibilidad de indizar, consultar y evaluar las colecciones estándar de *Text REtrieval Conference* (TREC).

Características: aplicable a grandes colecciones de documentos; disponible como una aplicación de la PC, una interfase *Web Java Server Page* (JSP) y una *Application Programming Interface* (API); expansor de acrónimos; posibilita el indizado de textos, html, PDF, Microsoft Word, Excel, PowerPoint, y colecciones TREC; indizado de la información de campo (ejemplo, frecuencia de términos en el campo title); recuperación en la PC, líneas de comando y la *Web* basada en interfaces de consulta; varios modelos de ponderación de los documentos, Okapi BM25, modelación del lenguaje y tf-idf; facilidades de expansión de la consulta mediante realimentación seudo relevante; lenguaje de consulta avanzado basado en operadores booleanos, +/-, búsqueda por frases y por proximidad y búsqueda por campos.

Hadoop/Mahout/Solr. Descripción (Plasencia *et al*., 2012): *Hadoop* <http://hadoop.apache.org/>. Las bibliotecas de programas de Apache Hadoop constituyen un marco de trabajo que permite el procesamiento distribuido de grandes conjuntos de datos a través de clusters de computadoras utilizando un modelo simple de programación. Formalmente, Hadoop es un marco de trabajo de código abierto para la escritura y la corrida de aplicaciones distribuidas que procesan grandes cantidades de datos. Soporta el modelo MapReduce.

Apache Mahout <http://mahout.apache.org/>: Mahout es una biblioteca de aprendizaje automático de Apache. Los algoritmos que éste implementa caen bajo la esfera del aprendizaje automático o inteligencia colectiva. Ello significa en particular para Mahout el disponer de motores de recomendación (filtrado colaborativo), agrupamiento y clasificación. Éste es también escalable. Mahout tiene la finalidad de ser una herramienta de aprendizaje automático a seleccionar cuando la colección de datos a ser procesados es muy grande para una sola máquina.

Weka4WS/Globus Toolkit 4 (GT4) <http://grid.deis.unical.it/papers/pdf/PKDD2005.pdf>. Descripción: MapReduce no es el único método donde el aprendizaje automático distribuido tiene lugar, hay adiciones a Weka, tal como Weka4WS que están disponibles y realizan en esencia las mismas prestaciones. Weka4WS permite realizar el procesamiento distribuido de tres tareas comunes en Weka: el etiquetado, donde las etiquetas de las clases son asignadas a las instancias basadas en un modelo predictivo, prueba, donde es calculado el algoritmo de precisión de la predicción mediante el etiquetado de un conocido conjunto de instancias y la validación cruzada, un método de prueba que divide un conjunto de datos de entrada en n particiones. Estos tres métodos son ideales para una distribución simple de un conjunto de datos dividido utilizado a través de múltiples máquinas que no cambian el resultado final (Jenkin, 2009).

Weka4WS ha sido desarrollado utilizando la biblioteca de Java proporcionada por *GlobusToolkit* 4 (GT4), el cual tiene una arquitectura *Open Grid Service Architecture* (OGSA) (Bernal, 2008).

A continuación se muestran las dos tablas obtenidas de la comparación de las 13 herramientas de recuperación y análisis de información (Tabla 2).

Ahora podemos determinar cuáles son las herramientas más adecuadas para la realización de los laboratorios.

Para trabajar el Módulo 1, que incluye la introducción a las bases de datos documentales y las bibliotecas digitales, sus conceptos básicos, el proceso de recuperación de información, la modelación, la recuperación: Adhoc y el filtrado, la caracterización formal de los modelos de RI, la recuperación de información clásica, la evaluación de la recuperación y las colecciones de referencia, se seleccionó la herramienta Lemur Toolkit con Indri por la documentación en línea, la comunidad de discusión, la cantidad de ejemplos, la ayuda del programa y la utilización en otros cursos de diferentes universidades.

Para el Módulo 2, que comprende los lenguajes de solicitud de búsquedas, las solicitudes basadas en

 http://dx.doi.org/10.1590/0103-37862014000300008

Tabla 2. Comparación de los programas de recuperación de información.

1 de 1

Programas de Recuperación de Información (A)						
	Programa					
	Lemur Toolkit	Sphinx	Dragon	Text Garden	Lucene	Swish-e
Características						
Uso previsto	Motor de búsqueda	Motor de búsqueda	Biblioteca	Biblioteca	Biblioteca	Buscador
Multiplataforma	Si	Si	N/A	Si	Si	N/A
Multilingue	Si	No	No	No	N/A	N/A
Identificación conceptual	Si	N/A	N/A	N/A	N/A	N/A
Categorización por relevancia	Si	Si	Si	Si	Si	Si
Herramientas de Análisis	No	No	Si	Si	No	No
Redes Sociales	No	No	No	No	No	No
Manejo información multimedia	No	No	No	No	No	No
Web Semántica	No	No	Si	No	No	No
Especificaciones Software						
Licencia	Código libre/ Gratis	Código libre	Código libre	Código libre	Código libre	Código libre
Sistema operativo	Linux/Windows	Linux/Windows	Linux/Windows	Windows	Linux/Windows	Linux/Windows
Código programado en:	C/C++	C/C++	Java	C/C++	Java	C/C++
Enlace de descarga	<http://sourceforge.net/projects/lemur/>	<http://www.sphinxsearch.com/downloads.html>	<http://dragon.ischool.drexel.edu>	<http://ailab.ijs.si/dunja/textgarden/Txt2Cpd.zip>	<http://lucene.apache.org/>	<http://swish-e.org/download/index.html>
Versión	4.12	0.9.9	1.3	N/A	4.6	2.4.7
Especificaciones Docentes						
Facilidad en la ejecución de los algoritmos de RI	Baja	Baja	Baja	Baja	Media	Baja
Facilidad de ejecución de los algoritmos de análisis	Baja	Baja	Baja	Baja	N/a	Baja
Visualización de los resultados	No	No	No	Si	N/a	No
Comunidad de discusión	Si	N/A	N/A	N/A	Si	N/A
Facilidad de la ayuda	Media	Alta	Media	Media	Si	Media
Existencia de ejemplos	Si	Si	Si	Si	Si	Si
Libro/Manual/Idioma	No	Si (pdf)/inglés	Si (pdf)/inglés	No	Si (pdf)/inglés	No
Empleo en Cursos	Si	N/A	N/A	N/A	Si	N/A

Programas de Recuperación de Información (B)							
	Programa						
	Xapian	WebSphinx	RapidMiner	Indri	Terrier	Hadoop/Mahout	Weka4WS
Características							
Uso previsto	Biblioteca	Biblioteca	Programa	Buscador	Motor de búsqueda	Biblioteca	Programa
Multiplataforma	N/A	N/A	Si	Si	N/A	SI	Si
Multilingue	N/A	N/A	No	No	Si	N/A	N/A

Tabla 2. Comparación de los programas de recuperación de información.

Programas de recuperación de información (B)							
			Programa				
	Xapian	WebSphinx	RapidMiner	Indri	Terrier	Hadoop/Mahout	Weka4WS
Características							
Identificación conceptual	N/A	N/A	No	Si	Si	N/A	N/A
Categorización por relevancia	Si	No	No	Si	Si	N/A	N/A
Herramientas de análisis	No	No	Si	No	No	Si	Si
Redes Sociales	No	No	Si	No	Si	N/A	N/A
Manejo información multimedia	No	No	Si	No	Si (con Smart)	N/A	N/A
Web semántica	No	No	Si	No	N/A	N/A	N/A
Especificaciones software							
Licencia	Código libre	Código libre	Código libre	Código libre	Código libre	Código libre	Código libre
Sistema operativo	Windows	Windows	Linux/ Windows	Linux/ Windows	Linux/ Windows	Linux/ Windows	Linux
Código programado en:	C/C++	Java	Java	C/C++	Java	Java	Java
Enlace de descarga	<http://xapian.org>	<http://www-2.cs.cmu.edu/~rcm/websphinx/>	<http://rapidminer.com/>	<http://sourceforge.net/projects/lemur/files/latest/download?source=files>	<http://terrier.org/download/agree.shtml?terrier-3.5.tar.gz>	<https://cwiki.apache.org/confluence/display/MAHOUT/Downloads.>	<http://grid.deis.unical.it/papers/pdf/PKDD2005.pdf>
Versión	1.2.16	N/A	6.0	5.6	3.5	0.5	N/A
Especificaciones Docentes							
Facilidad en la ejecución de los algoritmos de RI	Baja	Media	Alta	Baja	Media	Media	Baja
Facilidad de ejecución de los algoritmos de análisis	Baja	N/A	Alta	Baja	N/A	Media	Media
Visualización de los resultados	No	Si	Si	No	Si	Si	Si
Comunidad de discusión	Si	No	Si	Si	Si	Si	N/A
Facilidad de la ayuda	Media	Baja	Alta	Media	Media	Alta	Baja
Existencia de ejemplos	Si	No	Si	Si	Si	Si	N/A
Libro/Manual/Idioma	No	Si	Si	No	No	Si (pdf)/inglés	N/A
Empleo en Cursos	N/A	Si	Si	Si	N/A	Si	N/A

Fuente: Elaboración propia (2014).

Nota: N/A: No disponible; RI: Recuperación de Información.

palabras claves, el reconocimiento de patrones, los protocolos de solicitud, los lenguajes textuales y propiedades, metadatos, textos, lenguajes de marcación, multimedia, operaciones con textos, preprocesamiento de documentos, agrupamiento, compresión de datos, se selecciona Sphinx por su semejanza con Lemur pero por

http://dx.doi.org/10.1590/0103-37862014000300008

incorporar además la facilidad de trabajar con PosgreSQL como gestor de bases de datos no estructuradas, lo que le da continuidad al curso de "Bases y almacenes de Datos" que se imparte con anterioridad en la misma maestría y por disponer de una buena documentación.

Para el Módulo 3, en donde se imparte el indizado y la búsqueda, los ficheros invertidos, las solicitudes booleanas, las búsquedas secuenciales, las solicitudes estructurales, la recuperación de información paralela y distribuida, las interfases al usuario y su visualización, la interacción hombre-computadora, el proceso de acceso de información, puntos de inicio, las especificaciones de las solicitudes, contexto, utilización de los juicios de relevancia y las interfases para el proceso de búsqueda, se selecciona *WebSphinx* con *Rapid Miner*, que une, en dos herramientas gráficas las utilidades de rastreo en la Web con las de análisis de los datos con minería de texto en formato XML y da continuidad al curso de "Recuperación de Información y Minería *Web*" que se imparte en la misma maestría. Para mostrar el trabajo con el procesamiento distribuido se selecciona el conjunto de herramientas "Solr/Lucene/Hadoop/Mahout" por su excelente literatura, comunidad de discusión, facilidad de descarga y el poder mostrar los ejemplos tanto en un cluster, una grid o una máquina personal.

Por último, para el Módulo 4, que incluye la recuperación de información multimedia, los buscadores y la búsqueda en la *Web*. Se propone el empleo del motor de búsqueda Terrier para el trabajo con redes sociales y datos multimedia (en conjunto con Smart) y la biblioteca Dragon para el trabajo con ontologías y la *Web* semántica.

El método expuesto de seguimiento y actualización de la Tabla 2, proporciona una metodología a seguir para el desarrollo y la evolución de las herramientas de recuperación de información, lo cual es un proceso que se debe de realizar en la preparación de cada curso si se quieren impartir contenidos actualizados a los estudiantes y elevar así la calidad de los mismos.

Conclusión

Como resultado del análisis técnico y cualitativo de acuerdo con la experiencia de los profesores del curso "Bases de datos documentales y biblioteca digital" que se imparte en la mención de Minería de datos de la maestría "Cibernética Aplicada" en el Instituto de Cibernética, Matemática y Física, en La Habana, Cuba, de trece herramientas de recuperación y análisis de la información de software libre analizadas que pudieron ser descargadas libremente de Internet se seleccionaron ocho herramientas, LemurToolkit con Indri, Sphinx, *Web*Sphinx con Rapid Miner, Solr/Lucene/Hadoop/Mahout, Terrier y Dragon lo que permite realizar un análisis comparativo entre las herramientas e incorporar la enseñanza problemática dentro de las actividades del curso para un mejor proceso de aprendizaje.

Se establece una metodología para el seguimiento y selección de nuevas herramientas ante el comienzo de un nuevo curso.

La comprensión del funcionamiento y el dominio del uso de estas herramientas es de gran ayuda a los profesionales de la información que en ocasiones realizan el trabajo de recuperación de la información de forma manual o semiautomática, o que simplemente no pueden realizar la recuperación de la información en grandes volúmenes de datos, como pueden ser los que se originan de fuentes como las redes sociales.

Los datos expuestos en la tabla son susceptibles de revisión así como las características a incorporar a la misma, para que reflejen mejor las cualidades de las herramientas mostradas. Ello deberá realizarse de forma continuada en los próximos cursos.

Referencias

Bernal, J. *Data mining and cross-validation over distributed*: *Grid enabled networks*: *Current state of the art*. Florida: Atlantic University Spring, 2008. Available from: <http://latinamericangrid.org/elgg/juan.bernal/files/2/13/Project+-+DataMining+-+CrossValidation+in+Grid+Enabled+Networks.ppt>. Cited: Dec. 20, 2013.

Cardoso, Y. *et al*. *Herramientas de minería de datos*. 2011. Disponible en: <http://www.monografias.com/trabajos92/herramientas-mineria-datos/herramientas-mineria-datos.shtml>. Acceso en: 13 enero. 2014.

Dopico, I.; Plasencia, A. Diplomado control avanzado: pertinencia y concepción curricular. In: Convención y Feria

Internacional Informática, 24., 2011. La Habana. *Resumen*… La Habana: CLAD, 2011. p.5-6.

Fan, W. *et al*. Tapping into the power of text mining. *Communications of the ACM*, v.49, n.9, p.77-82, 2005. Available from: <http://filebox.vt.edu/users/wfan/paper/text_mining_final_preprint.pdf>. Cited: Jan. 13, 2014.

FindTheBest. *Compare full text search software*. 2014. Available from: <http://full-text-search.findthebest.com/>. Cited: Jan. 13, 2014.

Grobelnik, M. *Text-Garden software suite quick overview*. Ljubljana, Slovenia: Jozef Stefan Institute. 2007. Available from: <http://www.powershow.com/view1/f5de0-ZDc1Z/TextGarden_Software_Suite_Quick_Overview_powerpoint_ppt_presentation>. Cited: Sep. 17, 2009.

Hatcher, E.; Gospodnetic, O.; McCandless, M. Lucene in action. 2nd ed. 2010. *E-book*. Stamford: Manning Publications. Available from: <dl.e-book-free.com/2013/07/lucene_in_action_2nd_edition.pdf>. Cited: Jan. 13, 2014.

Jenkin, N. *Distributed machine learning with Hadoop*. 2009. Disponible en: <http://wenku.it168.com/d_000575816.shtml> Acceso en: 13 enero 2014.

Middleton, C.; Baeza-Yates, R. *A comparison of open source search engines*. Barcelona: Universitat Pompeu Fabra, 2011. Available from: <http://wrg.upf.edu/WRG/dctos/Middleton-Baeza.pdf>. Cited: Jan. 13, 2014.

Miller, R.; Bharat, K. SPHINX: A framework for creating personal, site-specific web crawlers. In: International World Wide Web Conference, 7., 1998. Brisbane, Australia. *Proceedings*… Brisbane, Australia: Computer Network and ISDN Systems, v.30, p.119-130, 1998.

Pathrey, R. *et al*. Discovering knowledge patterns from Integration of clustering and classification techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, v.3, n.4, p.338-343, 2013.

Plasencia, A. *et al*. *Concepción de un buscador web soportado en tecnología grid e interrelación de herramientas Apache*. La Habana: Instituto de Cibernética Matemática y Física, 2012.

Trotman, A. *et al*. Towards an efficient and effective search engine. In: International ACM SIGIR Conference on Research on Development in Information Retrieval, 35., 2012, Portland. *Proceedings*… Portland: University of Otago, 2012, p.40-47.

Turtle, H; Hegde, Y.; Rowe S. Yet another comparison of Lucene and Indri performance. In: International ACM SIGIR Conference on Research on Development in Information Retrieval, 16., 2012, Portland. *Proceedings*… Portland: University of Otago, 2012, p.64-67.

Zhou, X.; Zhang, X.; Hu, X. *The dragon toolkit developer guide*. 2007. Philadelphia: Drexel University. Available from: <http://dragon.ischool.drexel.edu/tutorial.pdf>. Cited: Jan. 13, 2014.