Usos y funciones de Lexibot

Univ. Reynaldo Gonzalo Coca
Rojas Universidad Mayor de
San Andrés
Carrera de Informática
Análisis y Diseño de Sistemas de Información
r9coca6194@hotmail.com

RESUMEN

En el presente artículo se ve como ha influenciado el novedoso motor de búsqueda *Lexibot* dentro de la *Web profunda (Deep Web)*. Este motor de búsqueda se caracteriza por realizar búsquedas tanto en Web's tradicionales (o superficiales) y Web's profundas. En pocas palabras se describe su *uso, funciones, la clasificación de sitios, calidad de los resultados*.

Palabras Clave

Motor de Búsqueda, Web Profunda, Deep Web, VSM, eBIR, mEBIR, Lexibot.

1. INTRODUCCIÓN

En la actualidad con un mundo globalizado, cuando la población requiere buscar información sobre algún tema en específico, evita la lectura de libros ya que piensan que pueden encontrar su información de forma más rápida a través del Internet, el cual usa *motores de búsquedas* que permiten a los navegantes encontrar páginas con información útil.

Internet ha llegado a ser repositorio de inmensa cantidad de información, es decir la más grande del mundo y de esa forma todo el mundo aprende, investiga y a encontrar cosas de interés.

El marketing dentro de los diferentes motores de búsqueda es fácil de comprender, ya que experimentamos a diario desde la perspectiva como usuarios de la Web. En la actualidad existen tres principales motores de búsqueda disponibles en la Web (Google, Yahoo y MSN) y otros menos populares.

Pero dentro de la Web Profunda (Deep Web) aparece *Lexibot* de *BrightPlanet* insustituible para aprovechar la Web profunda y ponga sus resultados a la superficie.

La forma más sencilla de describir la *Lexibot* es un "motor de consulta dirigido." El *Lexibot* tiene otras características de gran alcance en los resultados de la calificación y clasificación, pero es esta capacidad de consultar múltiples sitios de búsqueda directa y simultánea que permite a contenido Web profunda para ser recuperada.

Por supuesto, los motores de búsqueda son a su vez las bases de datos de búsqueda. Por lo tanto, Resultados se integran fácilmente con *Lexibot* para las búsquedas dentro de la Web Profunda. Por definición, sin embargo, los motores de búsqueda son limitadas a la superficie de los documentos web que pueden ser descubiertos por el rastreo.

2. ¿QUÉ ES LEXIBOT?

Lexibot es una *herramienta de búsqueda especializado*, desarrollado por *BrightPlanet*, como una forma de buscar en la

Web profunda (la parte oculta de la Web que pueden contener 500 veces el contenido accesible a los motores de búsqueda convencionales).

3. USOS

Lexibot utiliza lo que llama *BrightPlanet* una tecnología integral e inteligente de búsqueda que permite a los usuarios realizar búsquedas utilizando texto simple, lenguaje natural, o consultas booleanas en cientos de bases de datos al mismo tiempo, filtrar y analizar los datos, y publicar los resultados como páginas Web.

Lexibot fue diseñado para realizar búsquedas complejas para identificar y recuperar contenido de todas las áreas de la Web, y para procesar la información.

Todas las recuperaciones, agregaciones y documentos caracterizaciones en este estudio utilizan la tecnología *Lexibot* de *BrightPlanet*. El *Lexibot* utiliza varios hilos para las consultas de origen simultáneas y luego documentan las descargas. En *Lexibot* todos los índices se obtienen de documentos recuperados (incluyendo el contenido de HTML).

Después de la descarga y la indexación, los documentos se califican como de relevancia utilizando cuatro diferentes *algoritmos de puntuación*, un lugar destacado de *modelización espacio vectorial (VSM)* y el estándar y recuperación de información del modelo extendido de Boole (*eBIR*).

4. FUNCIONES

El contenido de la "Web Profunda (*Deep Web*)" ha sido accesible sólo a aquellos que conocían la dirección correcta para el sitio Web.

Entonces, aunque un usuario se haya conectado a una base de datos, sus datos sólo podrían tener acceso a consultas directas individuales.

Lexibot actúa como un motor de consulta directa automatizado para hacer decenas de consultas simultáneamente a múltiples bases de datos. Las búsquedas se admiten en cerca de 600 bases de datos.

Una vez que se encuentran en los enlaces, *Lexibot* descarga los enlaces, los analiza, elimina los irrelevantes, a continuación, descarga las porciones de texto de los documentos en el disco duro del usuario.

La aplicación se puede utilizar con aplicaciones de escritorio, así como la Internet, es adaptable a las preferencias del usuario, y se puede configurar ya sea para uso simple o avanzada.

5. CLASIFICACIÓN DE SITIOS DE LA DEEP WEB

El Lexibot se utiliza para recuperar las páginas completas e índices plenamente, tanto con las fuentes iniciales únicas y las fuentes eliminadas de un enlace. Algunos 43.348 URLs resultantes se recuperan en realidad.

Aplicando un *criterio inicial de filtro* a estos sitios, permiten determinar si eran los sitios de búsqueda iniciales. Este filtro inspecciona el contenido HTML de las páginas, además de analizar el contenido de la página. Este filtro se tradujo en 17.579 URLs precalificados.

Inspeccionando el lado posterior de 700 sitios asignados al azar de esta lista identifica otros criterios de filtro. Noventa y cinco de 700, o el 13,6%, no reunía los requisitos plenamente como sitios de búsqueda. Esta corrección se ha aplicado a todo el conjunto candidato y los resultados presentados.

La prueba de los sitios calificados en una prueba automatizada en el Lexibot para sitios de búsqueda donde la clasificación es del 98% de precisión.

Además, los medios automatizados para descubrir nuevos sitios de búsqueda se han incorporado en nuestra versión base que se encuentra dentro del Lexibot.

6. ANÁLISIS DE CALIDAD DE RESULTADOS

Las comparaciones de calidad entre el contenido de la Web profunda y superficial se basaron en cinco consultas diversas, sobre temas específicos emitidos a través de Lexibot ciertos motores de búsqueda convencionales y tres sitios profundos específicos para un tema. Las cinco áreas temáticas fueron la agricultura, la medicina, las finanzas / negocios, la ciencia y el derecho.

Las consultas fueron diseñadas específicamente para limitar los resultados devueltos por el total de cualquiera de las seis fuentes para un máximo de 200 para asegurar la recuperación completa de cada source.

La determinación de "calidad" se basa en un promedio de VSM del Lexibot y métodos de calificación lingüística computacional mEBIR. El medidor de "calidad" se fijó en la puntuación 82 para Lexibot, empíricamente determinado como más o menos precisión de millones de puntuaciones Lexibot anteriores de documentos subidos en Web tradicional.

La Web profunda vs las Web's superficiales se obtuvieron mediante el uso de la selección de la opción de Lexibot fuente y luego contar los documentos y los documentos que superen el medidor de puntuación de la calidad total.

7. CONCLUSIONES

Se da a conocer el motor de búsqueda Lexibot dentro de la Deep Web, ya que es novedosa y efectiva al momento de recuperar los registros de datos en una página Web.

Su algoritmo se basa en dos observaciones importantes sobre los registros de datos en la web y un algoritmo de coincidencia de cadenas.

Es automático y por lo tanto no requiere ningún esfuerzo manual. Además, es capaz de descubrir registros de datos no contiguos, que no pueden ser manejados por las técnicas existentes, ni siquiera por los motores de búsqueda tradicionales que se conocen actualmente.

8. BIBLIOGRAFÍA

- [1] Rouse, M. *What is LexiBot?*, (Sep. 2005), <http://whatis.techtarget.com/definition/LexiBot>
- [2] BrightPlanet. *The Deep Web: Surfacing Hidden Value*, White Paper, (Jul. 2000), <http://brightplanet.com/wp-content/uploads/2012/03/12550176481-deepwebwhitepaper1.pdf>
- [3] Grasso, A. *La 'deep web', el lado oscuro de la internet que no conocemos*, (Jul. 2013), http://www.eset-la.com/pdf/prensa/concurso/677_AgustinaGrasso_DeepWeb_Perfil-AR.pdf