ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

“Observatorio de Internet para la ESPOL”

TESINA DE SEMINARIO

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:
INGENIERO EN CIENCIAS COMPUTACIONALES ESPECIALIZACIÓN SISTEMAS DE INFORMACIÓN

Presentado por:

Adrián Fabricio Ponguillo Intriago
Eddy Eduardo Ponguillo Intriago

Guayaquil – Ecuador
2010

AGRADECIMIENTO

Agradecemos a Dios por todas las bendiciones recibidas y el permitirnos realizar este trabajo. A nuestra madre, nuestro padre, familiares y amigos que han caminado junto a nosotros en los momentos difíciles. A la vida.

DEDICATORIA

A Dios, a nuestros padres, hermanos, familiares y amigos por el respaldo permanente que nos brindaron.

TRIBUNAL DE SUSTENTACIÓN

_____

Ing. Fabricio Echeverría
PROFESOR DEL SEMINARIO DE GRADUACIÓN

_____

Ing. Juan Moreno Velasco
PROFESOR DELEGADO DEL DECANO

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de este Trabajo de Graduación, nos corresponde exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL".

(Reglamento de Graduación de la ESPOL).

Adrián Fabricio Ponguillo Intriago

Eddy Eduardo Ponguillo Intriago

RESUMEN

El presente trabajo es una herramienta WEB para la ESPOL que permite tener información de varios parámetros SEO (Search Engine Optimization) de los sitios web de las distintas universidades del país.

Entre estos parámetros tenemos el tráfico WEB, Pagerank, Alexa Ranking, Google Indexed, Google Backlink, Yahoo! Indexed, Yahoo! Backlink, contador de Spam, entre otros.

La aplicación permite registrar la URL del sitio WEB de la universidad y, mediante la misma, se conecta a los servidores de Alexa y Yahoo! mediante sus diferentes APIs, obtenemos la información requerida, la cual es almacenada en la base de datos para efectos de consultas a la misma.

En la URL **http://bilalbhatti.com/checksite.php?extra=yes&url=**, mediante consulta en el parámetro url, extraemos el contenido de la página devuelta.

Dentro de la implementación del sistema se emplea un Demonio que cada 24 horas lanza procesos para actualizar las diferentes métricas.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

INTRODUCCIÓN

El desarrollo de un sitio WEB no termina con la puesta en producción del mismo, sino que debe someterse a un constante monitoreo que determine, en un momento dado, el grado de interés que produce. El uso de estándares WEB no basta para tener una posición consolidada en internet. Hay que considerar para el efecto varios indicadores, que llevados adecuadamente, lograrán que los motores de búsquedas nos coloquen en las mejores posiciones.

Una de las grandes dificultades para lograrlo es el desconocimiento de los diferentes parámetros que ayudan a un sitio WEB a posicionarse.

Este trabajo pretende ser una herramienta que monitoree el posicionamiento de la ESPOL dentro de las universidades del Ecuador.

Así mismo, persigue desarrollar una herramienta que permita conocer el estado de la ESPOL dentro del WEB en el contexto de las universidades ecuatorianas, obteniendo información de calificación SEO de los principales buscadores WEB que existen.

Esto sugiere explicar las barreras de posicionamiento que los desarrolladores WEB deben tener en cuenta, para darle más visibilidad a su sitio.

El análisis SEO que se realiza a un sitio comprende una serie de parámetros que son importantes conocer para alguien que trabaja como Webmaster, sea en una PYME o una corporación, o incluso una institución académica. El presente proyecto está pensado para convertirse en una herramienta que mida cómo nos ven en el Internet, que se ha convertido en una métrica para evaluar cómo estamos y/o cuál es el interés que generamos como institución, individualmente y dentro del dominio de las universidades del país.

Capítulo 1

1. ANÁLISIS DE LOS PARÁMETROS SEO

1.1. Introducción

SEO, Search Engine Optimization (optimización para motores de búsqueda), es una tarea que implica ajustar la información referente a un sitio para ubicarla en las primeras posiciones en los resultados de la búsqueda. Esta tarea consiste en aplicar varias técnicas para obtener una posición alta (primeros lugares) en los resultados para determinados términos o palabras claves de la búsqueda. A través de estos procesos de optimización se logra mejorar la visibilidad de un sitio para los buscadores, quitando las barreras de rastreo para los contenidos. Estas prácticas SEO son continuas, lo que trae como consecuencia un incremento en el tráfico de nuestro sitio.

A la hora de construir un sitio, no sólo debemos tener en cuenta que el mismo sea visible y usable para el usuario, sino también accesible para los buscadores, para que las páginas WEB formen parte de los índices de los buscadores.

1.2. ¿Qué es un buscador?

"Un motor de búsqueda es un sistema informático que indexa archivos almacenados en servidores WEB gracias a su «spider» (o WEB crawler)."[1]

Existen diferentes tipos de buscadores, pero los más usados son: Google, Yahoo!, Live, Ask y AOL.

Los buscadores realizan dos procesos básicamente: búsqueda e indexación (construcción/actualización del índice). Para la construcción de los índices los buscadores deben rastrear la WEB en busca de páginas. Para lograr este cometido, los buscadores utilizan un sistema de robots, arañas o rastreadores. Los robots avanzan enlace a enlace por todo el contenido de la red. Este rastreo no es lineal sino que se efectúa sobre las páginas que más cambian o las más relevantes para hacerlo lo más óptimo posible. En este proceso se analiza no sólo las palabras claves o la temática sino varios parámetros que miden la calidad e importancia (criterios de relevancia).

[1] WIKIPEDIA. 2009. Motor de Búsqueda. (Disponible en: http://es.wikipedia.org/wiki/Motor_de_busqueda. Consultado el: 12 de octubre de 2009).

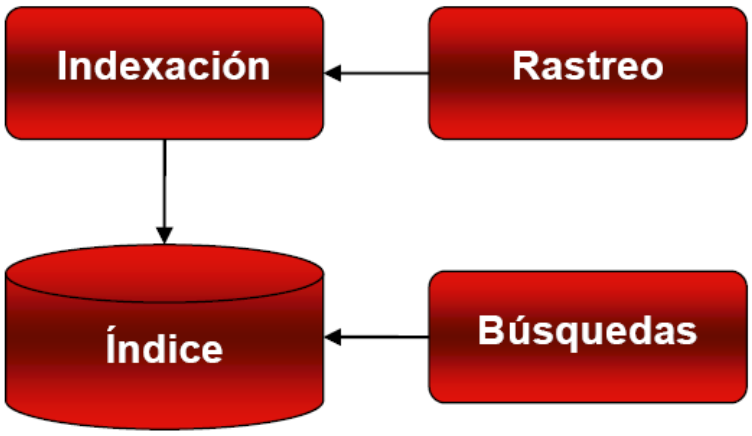

Figura 1: Proceso de búsquedas

Por lo general, los buscadores muestran la información por algún criterio de relevancia, estimando la probabilidad de que una página determinada pueda satisfacer la necesidad del usuario. A modo de resumen, el buscador recibe las palabras a buscar desde la caja de texto del mismo, realiza las consultas en el índice identificando las páginas en donde están presentes dichas palabras. Al mismo tiempo, lee los criterios de relevancia que se crearon en la fase de indexación y las ubica de acuerdo a dichos parámetros de importancia, y los presenta al usuario.

Estos criterios de relevancia no son revelados, pero de la experiencia de los usuarios se han detectado dos grandes grupos de incidencia:

- Los intra-site, que se refieren a la información que se puede obtener desde el propio sitio.

- Los extra-site, que son los enlaces externos que apuntan hacia nuestro sitio.

Factores intra-site

- Ubicación de la búsqueda. No se encuentra lo mismo en google.com.ec que en google.com porque la primera es una búsqueda localizada y tienen más relevancia las páginas localizadas en Ecuador.

- Edad del sitio. Es una métrica reciente y es utilizada para luchar contra el spam. Resulta una barrera para los sitios con menor tiempo de creación.

- Palabras claves en el dominio. Se utiliza para posicionar el sitio, como www.monografias.com, que busca competir con los sitios que ofrecen trabajos monográficos o de investigación.

- Palabras claves en URLs. Estudios aseguran que cuando la palabra clave se encuentra en la URL, el usuario tiende a dar click sobre el enlace.

- Palabras en el TITLE. Por lo general, los buscadores asignan un peso significativo al título de la página para efecto de otorgar la carga descriptiva de un contenido. Por otro lado es el campo utilizado por los buscadores para ubicarlo como título de la

respuesta en las páginas de resultados, lo que agregado a estar ubicado en la URL, motiva al click.

- Uso de secciones h1, h2,... El contenido ubicado entre estas etiquetas, es considerado como texto relevante y descriptivo. Pero también el uso excesivo puede conllevar a la penalización por parte del buscador.

- Frecuencia de aparición de las palabras. Este criterio también debe cuidarse, ya que es motivo de penalización por el uso indebido o excesivo. Este parámetro se fundamenta en el hecho de que, por ejemplo, un sitio de venta de computadoras, tendrá frecuentemente la palabra computadora en el contenido.

- Anatomía del sitio. Se refiere a la profundidad de las páginas en el sitio. A mayor profundidad, menor relevancia y viceversa. Podemos decir de forma general que la profundidad y la relevancia NO SON directamente proporcionales. Conviene tener en cuenta que los buscadores prefieren estructuras simples, sin muchos niveles de profundidad y que algunos expertos consideran que una página no puede tener más de cien (100) enlaces[2].

[2] Instituto Nacional de Tecnologías de la Comunicación - INTECO. 2009. Guía de recomendaciones "SEO" de posicionamiento en Internet. pág 15.

Factores extra-site

- Número de enlaces externos. Aquí debemos considerar también el hecho de que no todos los enlaces externos tienen el mismo peso. Difiere el peso de un sitio muy relevante de uno completamente anónimo.

- Relevancia de las páginas. Esto sugiere que la relevancia de los enlaces externos hacia un sitio de una temática similar es muy diferente que los del mismo sitio hacia una WEB con temática completamente diferente.

- Palabras utilizadas en los enlaces. Nuestra página puede ser mostrada por el buscador por medio de otros resultados que contengan la palabra buscada, y que hagan referencia a nuestro sitio por medio del enlace. Ejemplo:

    - sitio a. no contiene la palabra "venta de computadoras", pero es un website de compra y venta de partes y piezas de computadoras.

    - sitio b. contiene la palabra "venta de computadoras" y hace referencia al sitio a.

- Diversidad y dispersión de las palabras. Los seres humanos elaboran textos anclas ligeramente diferentes a los creados

automáticamente por mecanismos de engaño a buscadores que suelen ser similares. Estos últimos tienen formas para distinguir entre estas formas de generar los textos anclas mediante el análisis de la dispersión de estos textos anclas.

¿Qué son los Meta-Tags?

Respecto a los "meta-tags", la wikipedia indica:

> Las **metatags** son etiquetas HTML que se incorporan en el encabezado de una página WEB y que resultan invisibles para un visitante normal, pero de gran utilidad para navegadores u otros programas que puedan valerse de esta información.
>
> Su propósito es el de incluir información (metadatos) de referencia sobre la página: autor, título, fecha, palabras clave, descripción, etc.
>
> Esta información podría ser utilizada por los robots de búsqueda para incluirla en las bases de datos de sus buscadores y mostrarla en el resumen de búsquedas o tenerla en cuenta durante las mismas y será invisible para un visitante normal.
>
> Estas etiquetas también se usan para especificar cierta información técnica de la cual pueden valerse los navegadores para mostrar la página, como el grupo de caracteres usado, tiempo de expiración del contenido, posibilidad de dejar la página en cache o calificar el contenido del sitio ("para adultos", "violento"...)[3]

Estos tags son:

- Keywords

[3] WIKIPEDIA. 2009. Metatag. (Disponible en: http://es.wikipedia.org/wiki/Metatag. Consultado el: 30 de octubre de 2009).

- Description
- Robots

Por motivo de los spam, en la actualidad estos meta-tags son obviados, utilizándose únicamente el "description" que se emplea como pequeño texto de descripción del resultado de la búsqueda, y el "robots" que sirven de guía para el rastreo de las páginas[4].

Como ejemplos podemos citar los siguientes:

<meta name="keywords" content="palabra1, palabra2, palabra3" />

<meta name="description" content="Aqui describo mi sitio" />

<meta name="robots" content="index,follow" />

1.3. Barreras de rastreo

La barreras de rastreo son las dificultades técnicas que tienen los robots de los buscadores para posibilitar el rastreo de un sitio WEB, es decir nos imposibilitan estar visibles para los buscadores. De manera general podemos decir que las barreras que los robots de un determinado buscador son:

[4] Instituto Nacional de Tecnologías de la Comunicación - INTECO. 2009. Guía de recomendaciones "SEO" de posicionamiento en Internet. pág 15

- Código javascript: Poner cuidado a los enlaces creados por medio de este código, ya que los robots no pueden determinar las rutas formadas de esta manera.

- Flash: No pueden indexar el contenido embebido en estos recursos multimedia.

- Applets de java: Los buscadores no pueden interpretar el código creado en estos componentes java.

- Contenido embebido en imágenes: Al igual que las animaciones en flash y en los applets de java, el contenido visualizado como imagen no puede ser accedido por los robots de los buscadores.

- Formularios: No se puede colocar información en los campos, para obtener nuevas páginas para indexarlas por medio de los robots.

1.4. ¿Qué es el spam en buscadores (spamdexing) ?

Debido al factor económico asociado a la posición de un sitio WEB, sectores de la red concentran su esfuerzo en idear la forma de engañar a los buscadores para hacer creer que una página tiene más relevancia de la que realmente la tiene.

La Wikipedia nos habla del spam de la siguiente manera:

Hay dos tipos de **spam**: la publicidad encubierta como artículos y los enlaces externos masivos. Los artículos considerados publicidad incluyen a aquellos que son solicitudes para un negocio, producto o servicio, o son textos de relaciones públicas diseñados para promocionar una empresa o individuo. Los artículos de spam generalmente utilizan un lenguaje comercial y tiene enlaces a una WEB comercial. Sin embargo, se debe diferenciar entre los artículos de spam y los artículos legítimos sobre entidades comerciales[5].

1.5. Prácticas SEO

Para efectos de ubicar o posicionar un sitio WEB, se debe realizar algunas tareas como las siguientes:

- Lograr que otros sitios con temática similar, enlacen con nuestro sitio

- Darse de alta en directorios como Dmoz[6] (http://www.dmoz.org/), y Yahoo! (http://dir.yahoo.com/).

[5] WIKIPEDIA. Wikipedia:Spam. 2009. (Disponible en: http://es.wikipedia.org/wiki/Wikipedia:Spam. Consultado el: 1 de noviembre de 2009)

[6] MÁSTER ONLINE EN BUSCADORES. Selección de unidades didácticas 2008. (Disponible en http://www.masterenbuscadores.com/mbus_2007_2008.pdf. Consultado el: 1 de noviembre de 2009)

- Construir una amplia base de datos de keywords y frases claves con relación al negocio. "Google Trends"[7] nos puede ayudar mucho al realizar el análisis sobre la lista de palabras claves sobre las que deseamos realizar SEO, con el objeto de determinar cómo las personas buscan dentro de nuestro modelo de negocio. Una buena alternativa al momento de determinar las palabras claves que irán en el sitio es utilizar la herramienta para palabras clave de google "Keyword Tool External" visitando la dirección https://adwords.google.com.ec/select/KeywordToolExternal.

- Limitar el uso de flash, javascript, frames en la página, ya que esto limita la búsqueda o rastreo del robot en la página al ser visto como espacio plano por el cual no pueden navegar.

- Crear títulos y descripciones pertinentes del contenido de cada página. Estos son como la tarjeta de presentación de la página frente a los buscadores, donde los title y description son puntos iniciales para identificar los términos relevantes.

- Generar un "sitemap" que permita al buscador desplazarse por el sitio de manera ordenada y clara, mejorando su visibilidad.

- Actualizar la página con contenido original de calidad.

[7] http://google.com/trends

- Generar un archivo "robots" de forma correcta. Este archivo indica qué páginas deben indexarse por parte del motor de búsqueda y cuáles bloquearse.

1.6. Alexa

Alexa es propiedad de Amazon y fue fundada en 1996 por Brewster Kahle y Bruce Gilliat, su nombre, un acrónimo de Address Lookup EXperts Authority. Provee estadística del número de visitas de un sitio, y proporciona una gráfica del tráfico WEB que muestra el incremento/decremento del mismo con datos de los últimos tres meses dentro del top 100.000 que maneja[8].

Alexa recolecta las estadísticas de tráfico de los usuarios a través de la barra de Alexa (como las barras de Google o Yahoo!) que estos tienen instalada. Si un usuario no tiene la barra de Alexa instalada, los sitios visitados por la misma son detectados por Alexa y por lo tanto no se contabiliza la visita del sitio.

Para elaborar la clasificación, alexa.com, recoge dos tipos de datos:

1. el número de usuarios únicos que acceden a una página

[8] WIKIPEDIA. 2009. Alexa Internet. (Disponible en: http://es.wikipedia.org/wiki/Alexa_Internet. Consultado el: 3 de noviembre de 2009)

2. el número de veces que un usuario solicita una página determinada.

Para evitar la generación de número de visitas por los motores automáticos, alexa considera varias peticiones por un mismo usuario y un mismo día como una visita.

La compra de alexa por parte de Amazon en 1999[9] cambió el modelo de negocio del mismo.

Entre los servicios que ofrece Alexa tenemos[10]:

- Sitios relacionados con el sitio que se está visitando
- Información del sitio WEB: tráfico, ranking, tendencias.
- Top sites mundial, por país y por categoría.

1.7. Yahoo!

La Wikipedia se refiere a Yahoo! de la siguiente manera:

> **Yahoo! Inc.** es una empresa global de medios con sede en Estados Unidos, cuya misión es "ser el servicio global de

[9] WIKIPEDIA. 2009. Alexa Internet. (Disponible en: http://en.wikipedia.org/wiki/Alexa_Internet. Consultado el: 3 de noviembre de 2009)

[10] TUFUNCION. 2009. 10 cosas que deberías saber de Alexa. (Disponible en: http://www.tufuncion.com/10_trucos_alexa. Consultado el: 3 de noviembre de 2009)

> Internet más esencial para consumidores y negocios". Posee un portal de Internet, un directorio WEB y una serie de servicios, incluido el popular correo electrónico Yahoo!. Fue fundada en enero de 1994 por dos estudiantes de postgrado de la Universidad de Stanford, Jerry Yang y David Filo. Yahoo! se constituyó como empresa el 2 de marzo de 1995 y comenzó a cotizar en bolsa el 12 de abril de 1996. La empresa tiene su sede corporativa en Sunnyvale, California, Estados Unidos[11].

En su estrategia de abrir su motor de búsqueda, lanzó Yahoo! Boss (Acrónimo para Build Your Own Search Service), el que define un servicio WEB abierto de búsqueda, tiene como objetivo fomentar la innovación en la industria de búsqueda y convertirla en un servicio WEB.

El API de Yahoo! BOSS invita a los desarrolladores a construir sus propias aplicaciones de búsqueda WEB. Esta API permite a los desarrolladores manejar las preguntas, imágenes, noticias, resultados con XML o en formato JSON. Permite aprovechar sus algoritmos e infraestructura.

Entre las cosas que pueden hacerse por medio del API, está la manipulación de los rankings que a diferencia de Google que vienen ordenados por el PageRank. Uno puede crear su propio algoritmo de

[11] WIKIPEDIA. 2009. Yahoo! (Disponible en: http://es.wikipedia.org/wiki/Yahoo! Consultado el: 3 de noviembre de 2009)

ordenación, la apariencia de las páginas de búsqueda, crear mashups de los datos con otras fuentes de datos, entre otras.

Esta API recibe una URL así:

```
http://boss.yahooapis.com/ysearch/spelling/v1/{query}?appid=xyz&format=xml
```

Esta información es procesada y los resultados son devueltos en formato XML.

El API es utilizado por[12]:

- Hakia, buscador semántico
- Me.dium Search, buscador social
- Daylife, plataforma de publicación de contenidos
- Cluuz, clusteriza y ordena de forma gráfica los resultados

1.8. Google

Google Inc. es la empresa propietaria de la marca Google, cuyo principal producto es el motor de búsqueda del mismo nombre. Fue fundada el 4 de septiembre de 1998 por Larry Page y Sergey Brin (dos estudiantes de

[12] CORTIZO PÉREZ, José. En la práctica: Lucene y Yahoo! BOSS (Disponible en: http://www.slideshare.net/jccortizo/sinai-ejemplos-prcticos-con-lucene-y-yahoo-boss-presentation. Consultado el: 3 de noviembre de 2009)

doctorado en Ciencias de la Computación de la Universidad de Stanford)[13].

PageRank

Sergey Brin y Larry Page publicaron en 1998 un algoritmo que permite aproximar este cálculo. Este sistema de ranking tiene un valor entre 0 y 10 y la Wikipedia enuncia: “PageRank confía en la naturaleza democrática de la WEB utilizando su vasta estructura de enlaces como un indicador del valor de una página en concreto. Google interpreta un enlace de una página **A** a una página **B** como un voto, de la página A, para la página B. Pero Google mira más allá del volumen de votos, o enlaces que una página recibe; también analiza la página que emite el voto. Los votos emitidos por las páginas consideradas "importantes", es decir con un PageRank elevado, valen más, y ayudan a hacer a otras páginas "importantes". Por lo tanto, el PageRank de una página refleja la importancia de la misma en Internet”[14].

1.9. Ask

Conocido también como Ask Jeeves y fundada en 1996[15], tenía como idea original permitir que los usuarios obtengan las respuestas a las

[13]WIKIPEDIA. 2009. Google. (Disponible en: http://es.wikipedia.org/wiki/Google. Consultado el: 3 de noviembre de 2009)
[14]WIKIPEDIA. 2009. PageRank. (Disponible en: http://es.wikipedia.org/wiki/PageRank. Consultado el: 3 de noviembre de 2009)
[15]WIKIPEDIA. 2009. Ask.com. (Disponible en: http://es.wikipedia.org/wiki/Ask.com. Consultado el: 3 de noviembre de 2009)

preguntas que se formulan a diario en un lenguaje natural. Jeeves es una creación ficticia de un personaje (mayordono) que intenta resolver cualquier pregunta en el lenguaje natural, personaje que es eliminado en el 2006 de ask.com.

Con el paso del tiempo y la eficiencia de otros motores de búsqueda como Google, fue perdiendo adeptos. Ask es lento para indexar las páginas, hecho que permite que no sufra de spam[16].

> "Ask.com también posee la tecnología de búsqueda basado en temas de popularidad para calcular el grado de autoría en un resultado. La tecnología fue nombrada como Teoma, el cual posee también una licencia para calcular la popularidad de cada click, que proviene directamente desde el buscador DirectHit, el cual fue comprado por Ask Jeeves en enero de 2000. El 26 de febrero de 2006, Teoma fue renombrado y redirigido directamente a ask.com"[17].

1.10. Bing

Es el nombre del buscador de Microsoft, anteriormente Live Search, Windows Live Search y MSN Search. Fue presentado el 28 de mayo del

[16]WIKIPEDIA. 2009. Ask.com. (Disponible en: http://es.wikipedia.org/wiki/Ask.com. Consultado el: 3 de noviembre de 2009)
[17]Id.

2009 por Steve Ballmer[18], CEO de Microsoft y puesto en línea el 3 de junio del 2009. Se basa en tres puntos[19]:

1. Acceso al conocimiento, dejando de ser un simple buscador y ofrecer respuestas, eliminando los “enlaces malos” que generan pérdida de tiempo al usuario.

2. No limita los resultados a enlaces, sino que da imágenes, videos, sitios, perfiles, etc.

3. Que los resultados se conviertan realmente en aportes para convertirse en puntos de decisión.

El 29 de julio del 2009, Microsoft y Yahoo! anunciaron el acuerdo al que llegaron en el que Bing reemplazaría a Yahoo! Search[20].

El nombre "Bing" es una palabra que imita el sonido que representa y de fácil memorización. Bing censura los resultados de búsqueda con términos como "sexo" para algunas de las regiones incluyendo India, República Popular de China, Alemania y arábiga. Esta censura se realiza basada en la legislación local de esos países. Sin embargo, Bing permite

[18]WIKIPEDIA. 2009. Bing (motor de búsqueda). (Disponible en: http://es.wikipedia.org/wiki/Bing_%28motor_de_b%C3%BAsqueda%29. Consultado el: 3 de noviembre de 2009)
[19]OJOBUSCADOR. 2009. Bing se reinventa. (Disponible en: http://www.ojobuscador.com/noticias/bing-se-reinventa/. Consultado el: 3 de noviembre de 2009)
[20]WIUKIPEDIA- 2009. Bing (motor de búsqueda). (Disponible en: http://es.wikipedia.org/wiki/Bing_%28motor_de_b%C3%BAsqueda%29. Consultado el: 3 de noviembre de 2009)

a los usuarios simplemente cambiar su preferencia de país o región a otros países donde no existe restricciones de las leyes locales tales como Estados Unidos o Australia a renunciar a esta censura[21].

1.11. Robots.txt

Conocido también como “crawler” o “spider” [22], es un archivo de texto plano que debe ser colocado en el directorio raíz como sigue:

- http://dominio/robots.txt

En este archivo se incluye instrucciones para el crawler – programa que inspecciona el WEB – para indicar qué páginas se deben seguir e indexar y a cuáles se deben bloquear el acceso.

Un tema a considerar en la optimización de páginas WEB para los buscadores es el hecho de que por lo general difieren en algo respecto a los criterios para posicionarse. Por esta razón, lo que se acostumbra hacer es optimizar ciertas páginas para ciertos buscadores. Esto conlleva a un problema. Las páginas serán diferentes ligeramente, lo que podría provocar que cuando los spiders lleguen al sitio a través de las páginas optimizadas para dicho buscador, al notar la similitud con las otras páginas "piensen" que son spam, lo que generaría la eliminación de la

[21] WIKIPEDIA. Bing (motor de búsqueda). (Disponible en: http://es.wikipedia.org/wiki/Bing_%28motor_de_b%C3%BAsqueda%29. Consultado el: 30 de octubre de 2009).
[22] ROBOTSTXT. (Disponible en: http://www.robotstxt.org/. Consultado el: 30 de octubre de 2009).

WEB en su buscador o la penalización haciéndola descender en su posición.

El archivo robots.txt tiene la siguiente forma[23]:

- User-Agent: *
- Disallow: /privatefolder/
- Disallow: /privatefile.html

El User-Agent es el nombre del spider del buscador y Disallow es el nombre del archivo que no deseamos que el spider indexe. Ejemplo:

- User-Agent: Googlebot (El spider de Google)
- Disallow: directorio/pagina-00.html
- Disallow: directorio/pagina-01.html
- Disallow: directorio/pagina-02.html
- Disallow: pagina-03.html

Este código prohíbe el acceso al spider de Google a dos páginas optimizadas para Yahoo! y dos páginas optimizadas para Altavista. De

[23] SEOMOZ. 2009. The Web Developer's SEO Cheat Sheet. (http://www.seomoz.org/user_files/SEO_Web_Developer_Cheat_Sheet.pdf) p2

tener permiso Google para acceder a dichas páginas se corre el riesgo de ser penalizados o eliminados de sus búsquedas.

Por ejemplo, el robots.txt de la ESPOL (http://www.espol.edu.ec/robots.txt) es el siguiente:

User-agent: *

Disallow:

Unos cuantos buscadores a continuación:

- Excite - ArchitextSpider
- Altavista - Scooter
- Lycos - Lycos_Spider_(T-Rex)
- Google – Googlebot y Freshbot[24] que, para páginas como las de noticias, que se actualizan más frecuentemente, son indexadas de una manera especial mediante este último.
- Alltheweb - FAST-WebCrawler/
- Ask - AskJeeves
- Bing - MSNbot

[24] CONOCIMIENTOSWEB. Descubriendo a Google. (Disponible en: http://www.conocimientosweb.net/portal/html.php?file=cursos/Google/g4.htm. Consultado el: 30 de octubre de 2009)

- Yahoo! - Slurp

Se puede encontrar una lista más detallada de los diferentes robots, visitando la dirección “http://www.robotstxt.org/db.html”.

Si insertamos información en los siguientes elementos, ésta no será reconocida y por ende no aparecerá en las búsquedas:

- JavaScript
- DHTML
- Flash
- Frames
- Session IDs
- Applets de Java
- Imágenes: no insertes textos dentro de ellas.

1.12. Sitemap.xml

El sitemap (mapa de sitio WEB, mapa de sitio o mapa WEB) es un archivo que contiene una lista de las páginas del sitio y que se encuentra accesible para los spiders y los usuarios, organizado de forma jerárquica.

Esto ayuda a los visitantes y a los robots de los motores de búsqueda a encontrar las páginas del sitio.

Este archivo ayuda a mejorar el posicionamiento en los buscadores, garantizándonos que las páginas puedan ser encontradas. Es importante sobre todo si se utiliza menú en Flash o Javascript que contienen los enlaces HTML.

Otorgan también una ayuda a la navegación al mostrar una vista general del contenido del sitio WEB.

Los sitemaps suelen usar XML, aunque admiten también fuentes WEB RSS y archivos de texto como formato[25].

Se recomienda poner el archivo sitemap en la raíz[26] como sigue:

http://dominio/sitemap.xml

[25] WIKIPEDIA. Mapa de sitio web. (Disponible en: http://es.wikipedia.org/wiki/Mapa_de_sitio_web. Consultado el: 3 de noviembre de 2009)
[26] SITEMAPS. Preguntas frecuentes. ¿Dónde puedo colocar mi Sitemap? (Disponible en: http://www.sitemaps.org/es/faq.php#faq_sitemap_location. Consultado el: 3 de noviembre de 2009)

Capítulo 2

2. DISEÑO E IMPLEMENTACIÓN

2.1. Introducción

La adquisición de métricas para evaluar la calidad o cantidad de contenido WEB de las universidades ecuatorianas es el principal problema que presenta nuestra solución. Las métricas a considerar son tomadas de diversos servidores, algunos accedidos mediante APIS provistos por ellos, otras capturando los valores mostrados como contenidos de algún sitio WEB que las provee.

El trabajo explicado más adelante, luego de un análisis donde por importancia se seleccionó determinados indicadores, es llevado a la plataforma WEB por la particularidad de esta de poder ser accedida con recursos mínimos como una conexión a la red y la presencia de un browser en el lado del cliente.

Nuestra solución permite que cada usuario registrado pueda darle seguimiento a su o sus sitios de interés obteniendo información actualizada, en su mayor parte, con retardo de un día; en el peor de los casos, muestra información atrasada con siete días.

La mejor manera de explicar el significado de un indicador es mediante gráficos; por ello nuestra aplicación provee al usuario la facilidad de interpretación con estas herramientas. Estos al ser muy heterogéneos hacen complejo el entendimiento al usuario. Por esto diseñamos un dashboard o consola de parámetros donde se explican con semáforos y flechas el estado de cada uno.

2.2. Esquema general de la aplicación

El programa SEO ESPOL es una aplicación WEB, construida para poder ser accedido por usuarios registrados comunes y un usuario administrador. Además se pueden consultar sin necesidad de ser un usuario registrado el estado de spam de cualquier sitio WEB.

Es una aplicación multicapa, desarrollada de esta forma para poder dar mejor mantenimiento y escalabilidad a la misma. Siguiendo el patrón MVC y cumpliendo con estándares WEB para mejor visualización en los diferentes navegadores, intenta ser una herramienta útil de manera que haga fácil el entender la situación de los diferentes sitios WEB registrados.

Revisemos el esquema general de la aplicación:

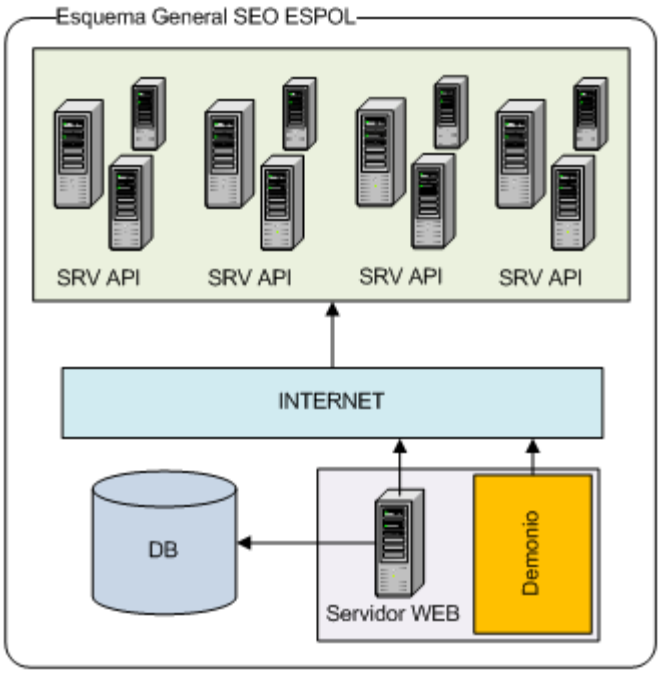

Figura 2: Arquitectura de la aplicación

2.3. Implementación de la aplicación WEB

Se decidió como plataforma de nuestra aplicación la WEB, gracias a que basta con un browser en el lado del cliente para poder hacer uso de la misma.

SEO ESPOL consta de diferentes módulos que describimos a continuación:

Módulo de Ingreso al sistema

Permite acceder a las funcionalidades principales del sistema de acuerdo al rol al que pertenezca el usuario. Existen dos roles:

1. Administrador, que habilita las funciones de control del sistema, permitiendo, dar de alta o baja a usuarios, establecer límites objetivos en las métricas, controlar los valores de búsqueda spam.

2. Usuario, es el módulo general que habilita el seguimiento de los sitios WEB registrados y observados mediante el dashboard.

Para acceder al sistema se debe ingresar el usuario y password registrados y validados por el administrador. La figura 3 muestra la pantalla de ingreso al sistema.

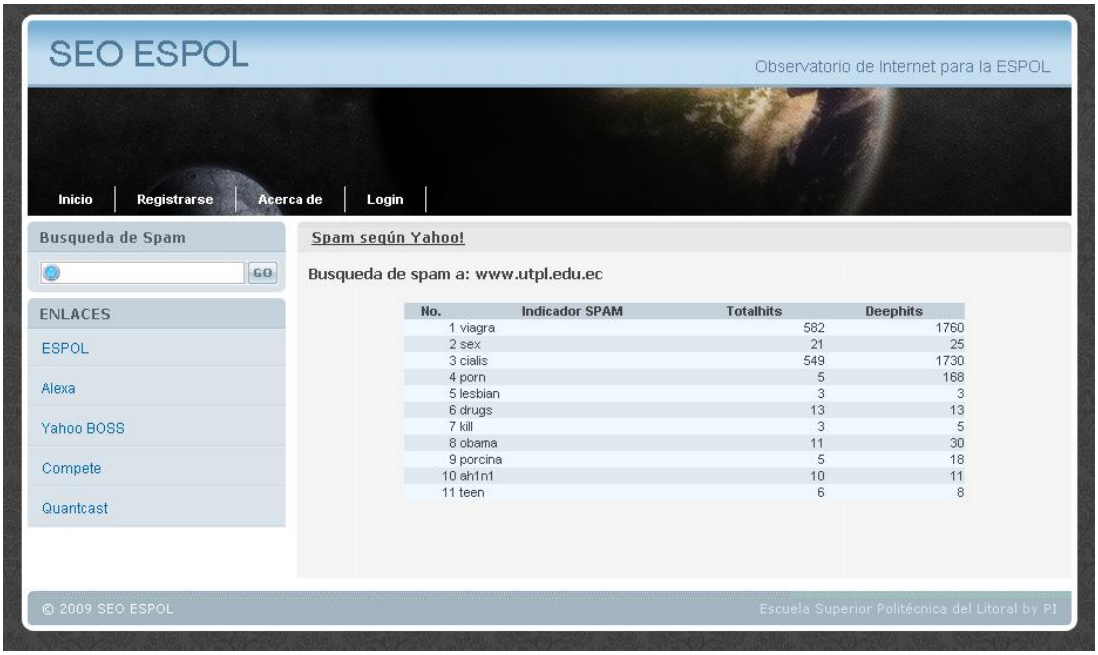

Figura 3: Búsqueda de Spam

Módulo de Presentación de datos

Considerado el corazón de la aplicación, debido a la importancia del mismo de acuerdo a los objetivos, se encarga de mostrar al usuario la lista de sitios registrados por él, para poder examinar el comportamiento actual del sitio seleccionado de manera fácil y precisa.

Aquí podremos apreciar las diferentes métricas seleccionadas con los valores adquiridos por el módulo de adquisición de datos.

Podemos apreciar en la figura 4 la pantalla de presentación, conocida como dashboard.

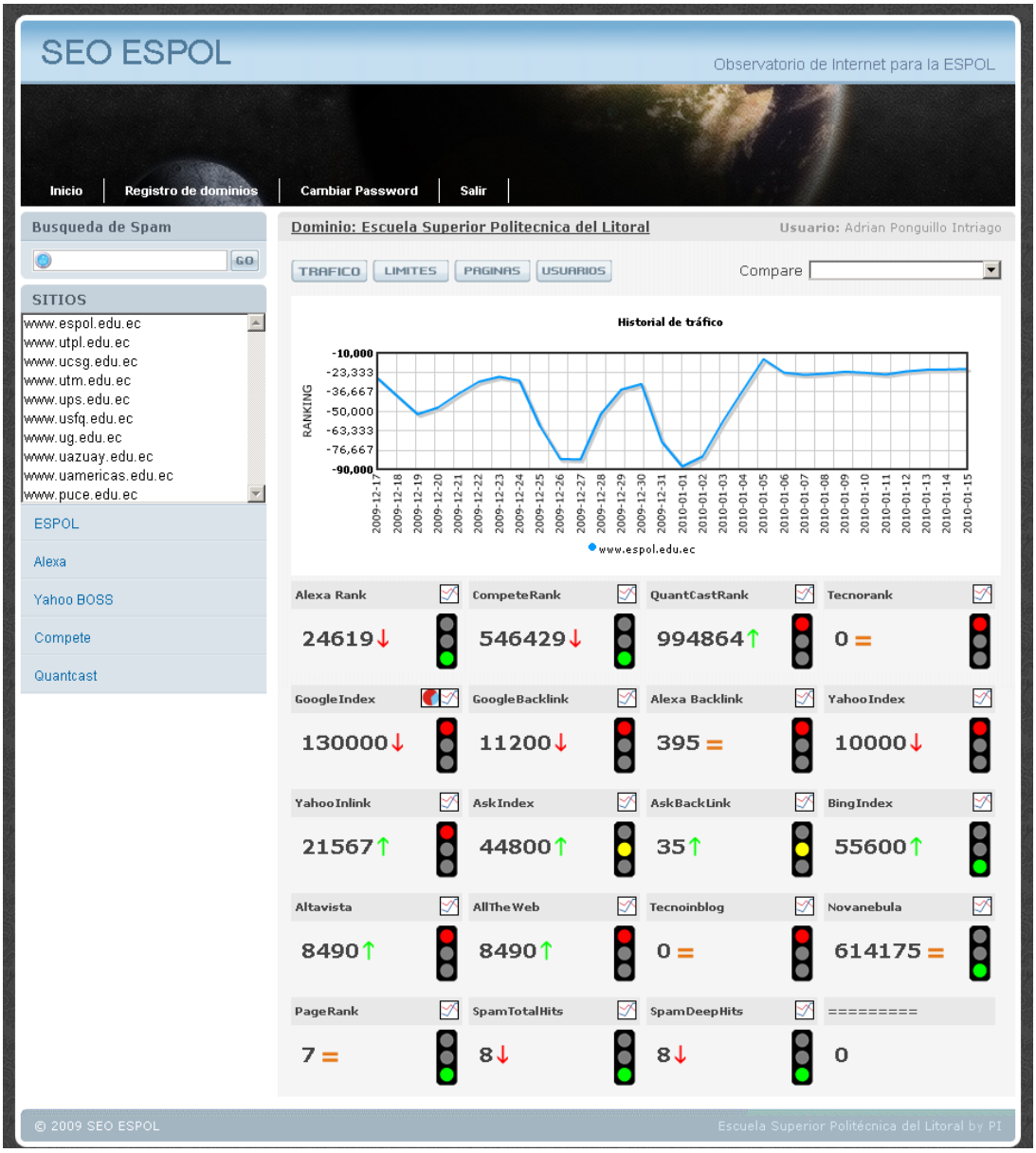

Figura 4: Dashboard

Módulo de Registro de Sitios

El módulo que permite registrar sitios a usuarios previamente registrados y aceptados por el administrador.

Podemos observar la figura 5:

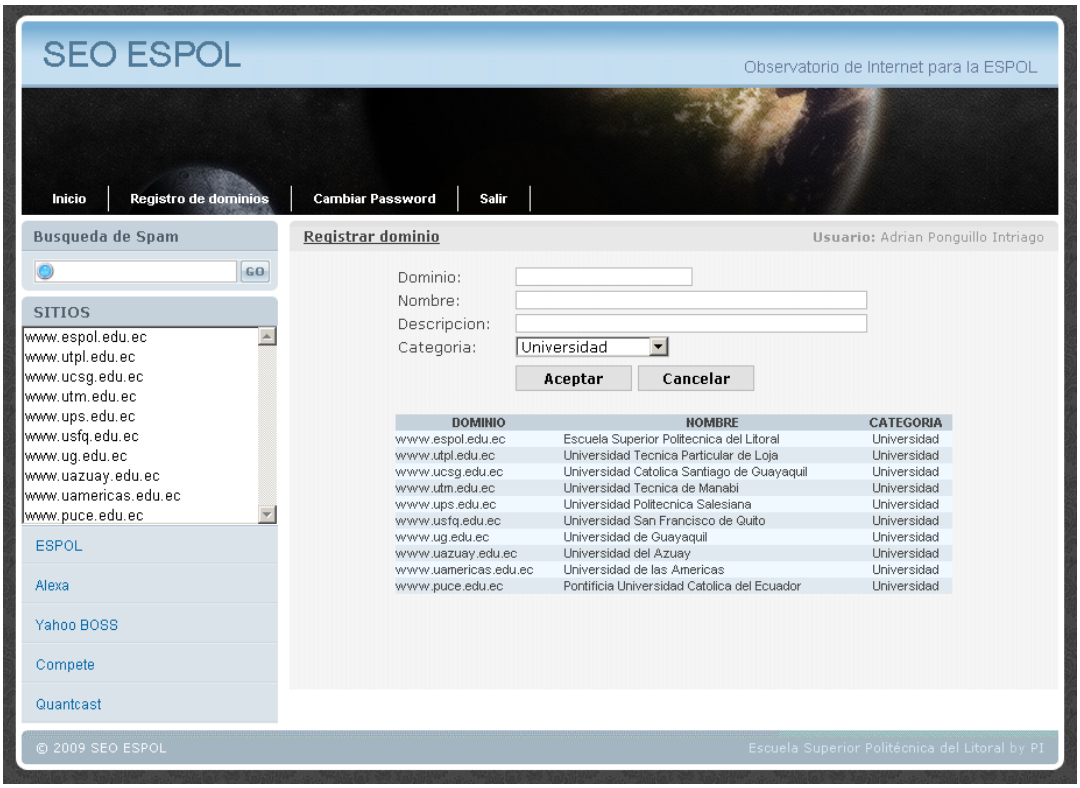

Figura 5: Registro de dominios

Módulo de Filtro Spam

Permite al administrador del sistema ingresar nuevos términos para búsqueda de spam mediante el servicio y criterios de Yahoo!, como también deshabilitar términos que pasaron de moda y por ende no son atractivos para el spam.

Podemos ver la pantalla en la figura 6:

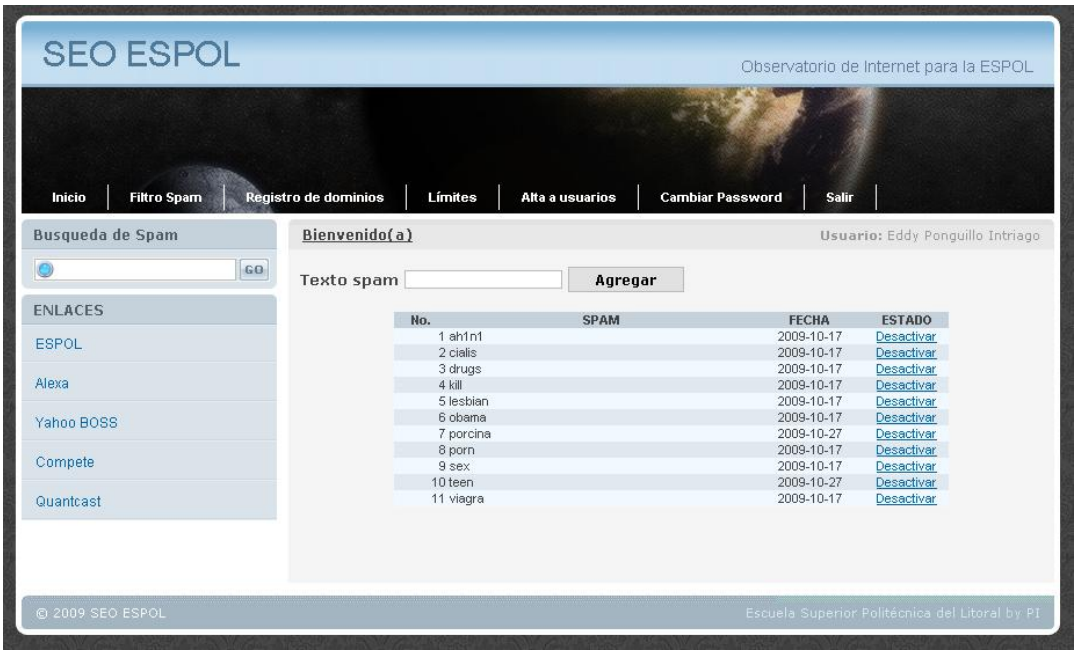

Figura 6: Filtro de Spam - Administración

Módulo de Configuración de Límites para métricas

Permite al administrador establecer los límites mínimos y máximos de las métricas registradas. De esta manera podemos mantener un control real de los objetivos deseados en el análisis de nuestros sitios.

Podemos ver la pantalla en la figura 7:

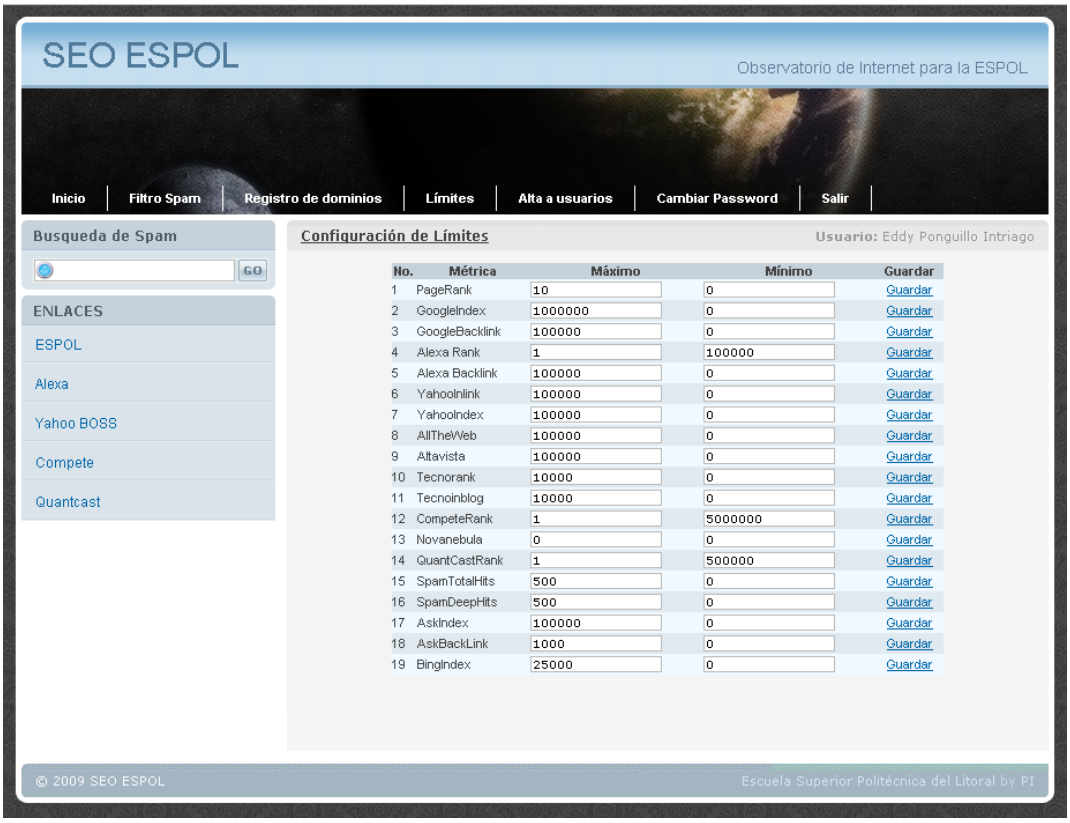

Figura 7: Administración de métricas

Módulo Adquisición de datos

Nuestro sistema es en sí una base de datos del comportamiento histórico en el WEB de los diferentes dominios registrados. Ordenados por métricas según grupos de interés, periódicamente se necesita de una actualización de estas métricas para ver y decidir qué camino seguir de manera que pueda optimizarse el sitio WEB.

Para esto se decidió implementar en resumen:

1. Un demonio lanzador de procesos de actualización de métricas,

2. Procesos encargados de consultar las métricas establecidas y guardarlas en la base de datos.

3. Base de datos de métricas para el dashboard.

4. El dashboard o panel que muestra las métricas actualizadas.

Demonio

Un demonio es un programa que se ejecuta en segundo plano sin necesidad de interacción o acción de algún usuario. Es un proceso especial que actúa indefinidamente y esperando la o las condiciones necesarias para su actuar.

Ejemplos comunes de demonios en sistemas Linux son: httpd, sendmail, ftpd, etc.

Nuestro sistema se enfrenta a la necesidad de obtener información de diferentes servicios para los diferentes sitios registrados en la aplicación.

En nuestras primeras pruebas nos dimos cuenta de que la actualización de métricas por sitio WEB se tomaba entre 45 y 300 segundos aproximadamente. Esto depende de la demora en la respuesta del servicio consultado, así como en la complejidad de la respuesta al ser procesada.

Es por esto que se decidió implementar como solución un procesamiento por hilos. En los cuales a sabiendas que cada petición se procesa independientemente y en paralelo se coloco las llamadas a los API en hilos que son llamados cada cierto tiempo.

Para controlar la llamada a los APIs de manera que no se requiera del comportamiento humano para su efectividad, se decidió establecer un HILO del tipo TIMER. Este hilo tiene como particularidad que se auto ejecuta cada determinado tiempo.

Por las diferencias entre las métricas decidimos separarlas dentro del Timer en dos grupos. Las que se ejecutan a diario y la que se ejecutan cada 7 días. El criterio para establecer esta diferenciación se basa en el costo. El costo de las peticiones tienen dos dimensiones, la dimensión precio/consulta y la dimensión bytes/consulta.

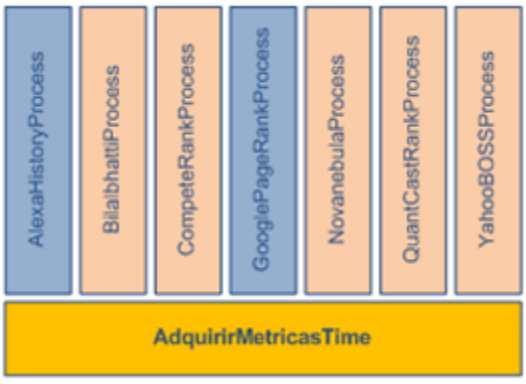

Figura 8: Procesos de adquisición

En el gráfico podemos apreciar al demonio AdquirirMetricasTime que es el encargado de lanzar los diferentes procesos que se encargan por cada sitio registrado en la base de datos, en consultar a los servicios respectivos. Cada proceso a su vez actualiza las métricas establecidas en la base de datos.

AlexaHistoryProcess

Este proceso se ejecuta cada siete días y obtiene la información histórica de tráfico WEB de los diferentes sitios registrados en la base de datos. Hace uso del API de Alexa el cual detallo a continuación.

Alexa Web Information Services.-

Alexa Web Information Services es una plataforma basada en servicios WEB que pretende brindar a los desarrolladores mediante su API información amplia acerca de sus sitios WEB.

El acceso a esta información se realiza mediante peticiones de cualquiera de estos dos tipos:

- Query request.- Son peticiones HTTP GET con respuestas XML
- SOAP request.- Son peticiones cuyo paso de parámetros son encapsulados en formato SOAP

La arquitectura de nuestro sistema maneja únicamente Query request.

Ejemplo:

http://awis.amazonaws.com/?AWSAccessKeyId=9876543212345123&Timestamp=2007-01-26T01%3A15%3A38.000Z&Signature=oQkiPZUtQ9PlTl2l4OTRA8fjYsM%3D&Version=2005-07-11 &Action=UrlInfo &ResponseGroup=Rank&Url=espol.edu.ec

En nuestro ejemplo podemos observar diferentes parámetros formando la url, que es pasada al servlet para ser procesada.

Estos parámetros son:

- **AWSAccessKeyId**.- Representa el AlexaID de un usuario registrado en AWIS. En dicho registro debe el usuario definir un método de pago. Esta definición se realiza agregando a su cuenta la información de una o más tarjetas de crédito, de donde mensualmente se debitará los valores correspondientes a las consultas realizadas.

  El paquete de 1000 consultas tiene un costo de 0.15USD, sin embargo el usuario del servicio solo pagará la parte proporcional a su consumo.

  Este AlexaID por los motivos expuestos debe manejarse con cuidado.
- **Timestamp**.- Se refiere al tiempo actual en formato UTC. Es tomada en consideración por AWIS para responder o no a la petición, ya que

si difiere de 15 minutos en relación a los servidores de Amazon la respuesta se denegará. Este parámetro pasa codificado.

- **Action**.- Indica lo que se va a pedir en la URL. En general define lo que se puede pedir. Los valores que puede tomar son:

- **CategoryBrowse**: Devuelve la lista de sitios ordenados por categorías por traffic rank.

- **CategoryListing**: Parecido a CategoryBrowse pero basado en contenido.

- **SitesLinkingIn**: Devuelve una lista de los sitios WEB que enlazan a el sitio WEB observado. Este dato es actualizado cada dos meses.

- **UrlInfo**: Obtiene información acerca del sitio WEB analizado tal como es: que tan popular es, sitios relacionados, información de contacto.

- **TrafficHistory**: Esta acción retorna el ranking diario de Alexa, límites por millones de usuarios y páginas vistas. Los datos que almacena AWIS pueden ser consultados siempre que comiencen no antes de Junio del 2007.

- **Signature**.- Es la firma de la URL. Es calculada con la concatenación del parámetro Action con el Timestamp en conformidad con el

estándar RFC-2104 HMAC-SHA1, para encriptar esta cadena, usando como clave el Secret Access Key provisto por Amazon.

- **Versión**.- Permitirá según Amazon al API seguir consultando incluso cuando la versión del mismo cambie. Actualmente es la 2005-07-11 y también es un parámetro opcional.
- **ResponseGroup**.- Varía de acuerdo al Action y se puede requerir varios separados con coma. Permite filtrar los datos del Action y es obligatorio de acuerdo al Action.
- **URL**.- Se refiere al sitio a consultar.

Una vez ejecutada la petición, AWIS devuelve un XML con la respuesta. El cual se lo parsea para extraer los datos que interesan y almacenarlo.

Nuestro proceso que es lanzado por el demonio cada siete días, se encarga de consultar el ranking histórico de los últimos días faltantes disponibles. Es decir consulta para cada sitio los datos desde el último día guardado más uno hasta la última fecha que provea el servicio. Generalmente debe ser 7 días.

BilalbhattiProcess

Este proceso es lanzado cada 24 horas por el demonio para consultar en el sitio http://bilalbhatti.com/ las siguientes métricas para cada sitio registrado en la base de datos:

- Google Pagerank
- Google backlink
- Google indexed
- Alexa popularity o rank
- Alexa backlink
- AllTheWeb result
- Altavista result
- Technorati rank
- Technorati inblogs

Se estableció este sitio como base de esta información por las pruebas que se realizaron que demostraron la fidelidad de los datos. Además de la necesidad imperiosa de minimizar el tiempo de adquisición de métricas por sitio.

Para procesar este requerimiento debemos enviar una solicitud HTTP GET mediante la siguiente URL base:

http://bilalbhatti.com/checksite.php?extra=yes&url=[sitio]

Donde [sitio] es la dirección de la que se requiere estas métricas.

La respuesta del sitio es en formato HTML, y ha sido minimizada al código estrictamente necesario para mostrar los valores de las métricas.

CompeteRankProcess

Compete es un sitio que ofrece un API para la adquisición libre de su ranking que es basado en el parámetro estadístico de mediciones web llamado Unique Visitors o Reach, que considera solamente una vez al usuario por determinado tiempo, incluso si este ha ingresado varias veces al sitio web, generalmente es un mes.

Para consultar a este API se requirió de un registro en el sitio, donde se nos otorgo acceso mediante una clave dada y que se requiere suministrarla cada vez que el llamado al API es realizado.

Este llamado es mediante una solicitud HTTP GET y requiere de una URL de la siguiente forma:

http://api.compete.com/fast-cgi/MI?d=[sitio]&ver=3&apikey=vdny39yptsbukrcm3wcfsjhy&size=large

Donde [sitio], el dominio a ser observado en este servicio.

La respuesta del servicio es mediante XML.

GooglePageRankProcess

Este proceso lanzado por el demonio cada siete días. El pagerank como ya hemos mencionado no se actualiza a diario sino entre dos y tres meses, por lo cual no necesitamos lanzar este proceso diariamente, sin embargo por la no difusión de la última vez que se actualizo este parámetro por parte de Google, hemos decidido que se haga en periodos de siete días.

Este servicio lo consumimos del sitio http://icons.geek-tools.org/, mediante una petición HTTP GET y con respuesta de texto plano.

La url de llamado al API es:

http://icons.geek-tools.org/api/show_pr.php?url=[sitio]

NovanebulaProcess

Este proceso es lanzado cada día para consultar en Novanebula que es un sitio que basado en diferentes parámetros SEO calcula un precio para el sitio web referido. Para acceder a este API debemos hacer una petición HTTP GET mediante la siguiente url:

http://www.novanebula.net/money/?site=[sitio]

Donde sitio es el dominio al que queremos estimar su valor.

La respuesta del API es dada en formato XML.

QuantCastRankProcess

Quantcast es un sitio que se dedica a realizar análisis web. Para adquirir el valor del ranking según quantcast.com el demonio lanza este proceso cada día mediante una petición HTTP GET cuya url es:

http://www.quantcast.com/[sitio]

Donde sitio es el dominio que analizaremos mediante esta herramienta. El valor de respuesta es HTML, el cual es procesado mediante este proceso.

YahooBOSSProcess

Este proceso lanzado cada 24 horas, hace uso del API de Yahoo! llamado BOSS (Build your own search services). Este proyecto es una iniciativa de Yahoo! para que terceros puedan hacer uso de su infraestructura y logren crear productos de búsqueda haciendo uso de sus datos.

Hacemos uso de este API para obtener:

- Cantidad de spam aproximado por sitio web.
- Páginas indexadas en Yahoo!.
- Páginas que enlazan a nuestro sitio.

El llamado a este servicio implica enviar una solicitud HTTP GET mediante la siguiente plantilla de url:

http://boss.yahooapis.com/ysearch/se_inlink/v1/[sitio]?appid={yourBOSSappid}&format=xml&count=1

Donde [sitio] es la dirección que vamos a revisar, appid es el ID provisto por el servicio una vez que hemos procedido debidamente con el registro, **format** indica el formato que debe usar el servidor para devolver la respuesta y **count** indica el índice del primer dato devuelto.

Nosotros hemos decidido que la respuesta sea recibida en XML, para mantener la consistencia con los demás servicios.

Una vez ejecutados los procesos, estos se encargan de almacenar en la base de datos por sitio, métrica y fecha el valor obtenido en cada caso. Cabe indicar que esta toma de datos al hacerse en segundo plano no interfiere en ningún momento con el manejo del sistema, pero si podría no mostrar los datos actualizados hasta que los hilos terminen su ejecución.

AskProcess

Mediante este hilo podemos traer a nuestra base de datos valores desde los servidores de ASK, sitio muy reconocido por contener buenos índices de contenido del web.

Este API no brinda a la aplicación dos importantes métricas como son:

- Ask index: Que se refiere a las páginas que están indexadas en el servidor de ASK.
- Ask backlink: Que se refiere a las páginas de terceros que hacen referencia al sitio consultado por nosotros.

Para poder acceder a este servicio vamos a realizar un Query Request de tipo HTTP GET, donde para cada una de las métricas arriba señaladas tenemos un formato de URL a indicar:

http://es.ask.com/web?q=site:[url1]

Donde [url1] es la dirección del sitio que estamos monitoreando la cantidad de contenido indexado.

http://es.ask.com/web?q=inlink:[url2]

Donde [url2] es la dirección del sitio que estamos monitoreando la cantidad de sitios que lo refieren.

BingProcess

El API del buscador Bing, requiere de un registro para la obtención de un API ID que se enviará cada vez que se requiera hacer una petición de información.

Por ser un buscador joven no presenta las mismas ventajas ni madurez de los API de Yahoo!. Sin embargo logramos obtener de él, el número de páginas indexadas por su robot.

Es por esto que para ver como se encuentra indexado el contenido publicado por las universidades del país, hacemos uso de este API también que tiene la siguiente URL para realizar el QUERY REQUEST de tipo HTTP GET.

http://api.search.live.net/xml.aspx?Appid=4A8D5344BDC9A7A86819CE8046FA6827E3D43D7A&sources=web+image&query=[url3]

Donde Appid es el identificador provisto por live.net para realizar las consultas sobre los datos de los datacenters de Bing.

También [url3] define la dirección que estamos necesitando obtener la cantidad de contenido publicado y accesible en internet y que ha sido indexado por Bing.

2.4. Diseño de la base de datos

Hemos diseñado la base de datos del sistema que permite almacenar:

- Los usuarios del sistema
- Las métricas por sitio

- El ranking de Alexa

- El control de spam

Podemos revisar su estructura en la figura 9:

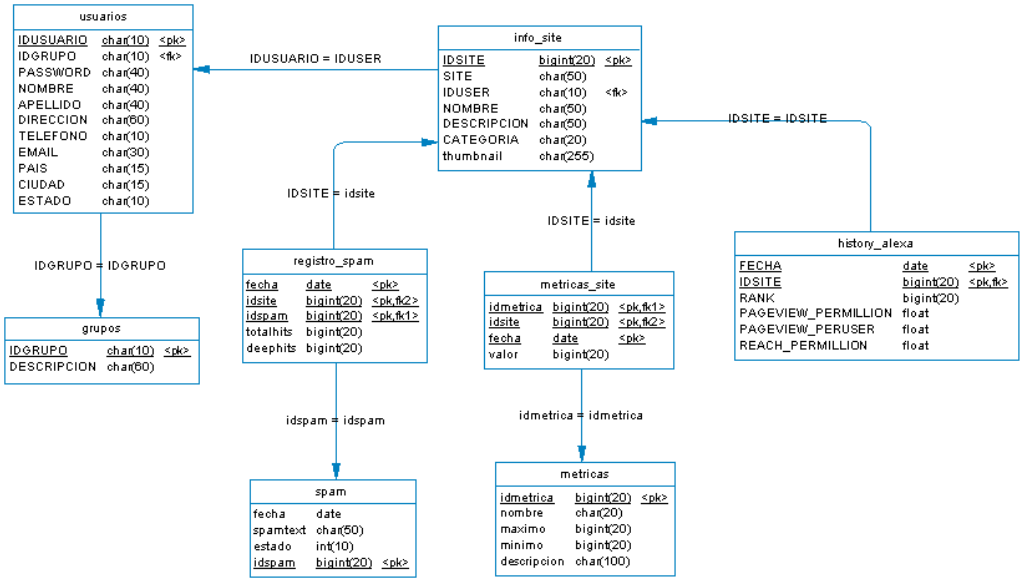

Figura 9: Diseño físico de la base de datos

Capítulo 3

3. RESULTADOS

3.1. Introducción

Durante la fase de implementación y pruebas, se presentaron algunos detalles de factibilidad y desempeño que detallamos en el presente capítulo. Se menciona la aplicación de hilos para la extracción de los datos y la carga de los mismos al dashboard. De la misma forma, la utilización de los API's para el análisis de tráfico WEB como los inconvenientes suscitados. En las pruebas se incluyó a universidades como:

- Escuela Superior Politécnica del Litoral
- Universidad Técnica Particular de Loja

- Universidad Politécnica Salesiana
- Universidad Católica Santiago de Guayaquil
- Universidad de Guayaquil
- Universidad Técnica de Manabí

3.2. Resultados

Google Indexed

Hemos obtenidos resultados de las páginas indizadas en los datacenter de Google mediante el API, con diferencias de hasta un 2,56% relativo a las búsquedas desde el browser.

Google Inlink

De acuerdo a las observaciones realizadas el error relativo que se encuentra al buscar sitios que enlacen a las diferentes universidades del País es de hasta el 7,89% con respecto a la búsqueda desde el browser.

Yahoo! BOSS

La no actualización de los datos por la API de Yahoo!. Al revisar los datos proporcionados desde el sitio WEB de Yahoo! y al realizar el estudio comparativo con los datos proporcionados desde el API de Yahoo! Boss, se presentaron inconsistencias no significativas (el margen de diferencia

es pequeño). Esto podría darse por el sitio de origen de la información, ya que estas grandes empresas de motores de búsquedas no manejan un sólo servidor sino una granja de servidores, lo que en un instante específico pudieran no estar sincronizados con los datos que proporcionan.

Compete API inestable

El API de compete es inestable, es decir puede devolver datos pero no de todas las url ingresadas, por lo que hasta el momento no se ha podido ver grandes variaciones en el ranking provisto por este.

Demora en la carga del dashboard

Durante el proceso de pruebas, se observó retardo en la presentación de la información dentro del dashboard, ya que cada elemento del mismo era alimentado por el mismo proceso. Para ello se tuvo que implementar un hilo para cada proceso, de tal manera que sea presentado de forma independiente y eliminar así el cuello de botella que generan algunos elementos dentro de la interfaz del dashboard del usuario.

CONCLUSIONES Y RECOMENDACIONES

Toda información que esté en el web, debe aspirar a ser accesible y visible a los usuarios, principalmente porque estamos en un mundo tan competitivo. Los buscadores son los que canalizan el tráfico que se genera en la red y que día a día crece en volumen. De allí la importancia de aparecer en los primeros lugares dentro de las búsquedas de los internautas.

1. Se han desarrollado un conjunto de herramientas SEO (Search Engine Optimizer) que servirán de apoyo al web máster de la ESPOL en el proceso de optimización, del website de la universidad, en los motores de búsqueda. De esta manera, se implementa una solución que representa una ventaja competitiva frente a sitios similares, al poder evaluar cómo estamos en popularidad frente a sitios similares que representan nuestra competencia para, a partir de este conocimiento, incorporar las

acciones necesarias que nos permitan seguir a la vanguardia de las universidades del país.

2. El hecho de tener un observatorio implica poder mirar más allá, y de esta observación obtener una ventaja en el entorno. Pretendemos brindar a las universidades las métricas suficientes para que vean como pueden crecer y hacer que la competencia por el bien de la educación nacional siempre esté presente.

3. Pero la competencia debe ser leal siempre. Existen lastimosamente prácticas malsanas que provocan que existan valores medidos con incorrecta información, pues de una u otra manera se ha permitido que el spam WEB afecte a estos parámetros, cosa que hemos detectado y que intentamos con nuestro programa ayudar a medir.

Las recomendaciones son:

Adicionalmente al empleo de herramientas que muestren cómo estamos en el internet, debemos ser fieles a las buenas prácticas web para mejorar nuestra posición dentro del mismo. Estas prácticas incluyen el uso correcto de los títulos en las páginas, utilizar texto en atributos "alt" para contenido multimedia como flash para ayudar al robot a que nos indexen.

También invitarnos a los responsables de los dominios de las universidades observadas que presentan índices altos de spam WEB, revisen muy coherentemente la seguridad de sus servidores. Este spam produce que los buscadores serios como Google eviten la indización de contenido e incluso ser motivo de bloqueos por los buscadores.

REFERENCIAS BIBLIOGRÁFICAS

QUINTAS, Jorge. 2008. Fundamentos SEO: Conceptos, Factores de posicionamiento y Herramientas SEO. (Disponible en: http://www.slideshare.net/jorge.quintas/fundamentos-seo-subflash-2008-tanta-comunicacion-presentation. Consultado el: 1 de noviembre de 2009)

INSTITUTO NACIONAL DE TECNOLOGÍAS DE LA COMUNICACIÓN - INTECO. 2009. Guía de recomendaciones "SEO" de posicionamiento en internet. (Disponible en: http://www.inteco.es/Accesibilidad/Formacion_6/Manuales_y_Guias//guia_seo. Consultado el. 1 de noviembre de 2009)

BING WEBMASTER CENTER TEAM, Microsoft®. 2009. Bing: New Features Relevant to Webmasters. (Disponible en http://www.bing.com/community/blogs/webmaster/archive/2009/06/

11/bing-white-paper-for-webmasters-amp-publishers-released.aspx. Consultado el: 10 de noviembre de 2009)

GOOGLE. Google's Search Engine Optimization Starter Guide. 2008. (Disponible en: http://www.google.com/webmasters/docs/search-engine-optimization-starter-guide.pdf. Consultado el: 1 de noviembre de 2009)

CORTIZO PÉREZ, José. En la práctica: Lucene y Yahoo! BOSS. (Disponible en: http://www.slideshare.net/jccortizo/sinai-ejemplos-prcticos-con-lucene-y-yahoo-boss-presentation. Consultado el: 10 de noviembre de 2009)

VERA CARTES, Luis. Chile. Normas para la redacción de referencias bibliográficas UACh. (Disponible en: http://www.biblioteca.uach.cl/pdf/referencias_lvera.pdf. Consultado el: 10 de noviembre de 2009)