Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

DISEÑO DE INVESTIGACIÓN EN IMPLEMENTACIÓN DE UN SITIO WEB QUE FACILITE INFORMACIÓN PERSONALIZADA SOBRE BECAS DE POSGRADO PARA PROFESIONALES GUATEMALTECOS

Mónica Ivett Picén Castañeda
Asesorado por el MBA. Ing. Leonardo Antonio Telón García

Guatemala, marzo de 2014

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

DISEÑO DE INVESTIGACIÓN EN IMPLEMENTACIÓN DE UN SITIO WEB QUE FACILITE INFORMACIÓN PERSONALIZADA SOBRE BECAS DE POSGRADO PARA PROFESIONALES GUATEMALTECOS

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

MÓNICA IVETT PICÉN CASTAÑEDA
ASESORADO POR EL MBA. ING. LEONARDO ANTONIO TELÓN GARCÍA

AL CONFERÍRSELE EL TÍTULO DE

INGENIERA EN CIENCIAS Y SISTEMAS

GUATEMALA, MARZO DE 2014

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO	Ing. Murphy Olympo Paiz Recinos
VOCAL I	Ing. Alfredo Enrique Beber Aceituno
VOCAL II	Ing. Pedro Antonio Aguilar Polanco
VOCAL III	Inga. Elvia Miriam Ruballos Samayoa
VOCAL IV	Br. Walter Rafael Véliz Muñoz
VOCAL V	Br. Sergio Alejandro Donis Soto
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Murphy Olympo Paiz Recinos
EXAMINADOR	Ing. Marlon Antonio Pérez Türk
EXAMINADOR	Ing. Víctor Hugo de León Barrios
EXAMINADORA	Inga. Virginia Victoria Tala Ayerdi
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

DISEÑO DE INVESTIGACIÓN EN IMPLEMENTACIÓN DE UN SITIO WEB QUE FACILITE INFORMACIÓN PERSONALIZADA SOBRE BECAS DE POSGRADO PARA PROFESIONALES GUATEMALTECOS

Tema que me fuera asignado por la Dirección de la Escuela de Estudios de Posgrado, en septiembre de 2013.

Mónica Ivett Picén Castañeda

USAC
TRICENTENARIA
Universidad de San Carlos de Guatemala

ESCUELA DE ESTUDIOS DE POSTGRADO
USAC

Escuela de Estudios de Postgrado
Facultad de Ingeniería
Teléfono 2418-9142 / Ext. 86226

AATT-MTIPP-0003-2014

Guatemala, 23 de enero de 2014.

Director:
Marlon Antonio Pérez Turk
Escuela de Ingeniería en Ciencias y Sistemas
Presente.

Estimado Director:

Reciba un atento y cordial saludo de la Escuela de Estudios de Postgrado. El propósito de la presente es para informarle que se ha revisado los cursos aprobados del primer año y el Diseño de Investigación del estudiante **Mónica Ivett Picén Castañeda** con carné número **2001-13117**, quien opto la modalidad del **"PROCESO DE GRADUACIÓN DE LOS ESTUDIANTES DE LA FACULTAD DE INGENIERÍA OPCIÓN ESTUDIOS DE POSTGRADO".** Previo a culminar sus estudios en la **Maestría de Tecnologías de la información y la Comunicación.**

Y si habiendo cumplido y aprobado con los requisitos establecidos en el normativo de este Proceso de Graduación en el Punto 6.2, aprobado por la Junta Directiva de la Facultad de Ingeniería en el Punto Decimo, Inciso 10.2, del Acta 28-2011 de fecha 19 de septiembre de 2011, firmo y sello la presente para el trámite correspondiente de graduación de Pregrado.

Sin otro particular, atentamente,

"Id y enseñad a todos"

Marlon Antonio Pérez Turk
Ingeniero En Ciencias y Sistemas
Colegiado 4402

MBA. Ing. Leonardo Antonio Telón G.
Asesor (a)

Leonardo Antonio Telón García
Ingeniero en Ciencias y Sistemas
Colegiado No. 9926

MSc. Ing. Marlon Antonio Pérez Turk
Coordinador de Área
Aplicación y transferencia tecnológica

Dra. Mayra Virginia Castillo Montes
Directora
Escuela de Estudios de Postgrado

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
ESCUELA DE POST-GRADO
FACULTAD DE INGENIERIA

Cc: archivo
/la

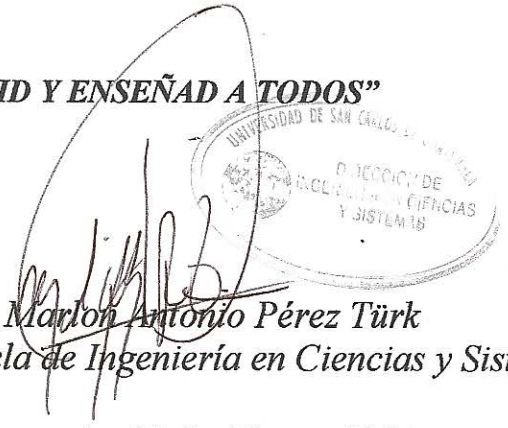
ESCUELA DE CIENCIAS Y SISTEMAS

UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÌA
ESCUELA DE CIENCIAS Y SISTEMAS
TEL: 24767644

El Director de la Escuela de Ingeniería en Ciencias y Sistemas de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor con el visto bueno del revisor y del Licenciado en Letras, del trabajo de graduación ***"DISEÑO DE INVESTIGACIÓN EN IMPLEMENTACIÓN DE UN SITIO WEB QUE FACILITE INFORMACIÓN PERSONALIZADA SOBRE BECAS DE POSGRADO PARA PROFESIONALES GUATEMALTECOS",*** *realizado por la estudiante MÓNICA IVETT PICÉN CASTAÑEDA, aprueba el presente trabajo y solicita la autorización del mismo.*

"ID Y ENSEÑAD A TODOS"



Ing. Marlon Antonio Pérez Türk
Director, Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 27 de febrero 2014

Universidad de San Carlos
de Guatemala

Facultad de Ingeniería
Decanato

DTG. 099.2014

El Decano de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería en Ciencias y Sistemas, al Trabajo de Graduación titulado: **DISEÑO DE INVESTIGACIÓN EN IMPLEMENTACIÓN DE UN SITIO WEB QUE FACILITE INFORMACIÓN PERSONALIZADA SOBRE BECAS DE POSGRADO PARA PROFESIONALES GUATEMALTECOS,** presentado por la estudiante universitaria: **Mónica Ivett Picén Castañeda,** autoriza la impresión del mismo.

IMPRÍMASE:

Ing. Murphy Olympo Paiz Recinos
Decano

Guatemala, 4 de marzo de 2014

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANO
FACULTAD DE INGENIERIA

/gdech

ACTO QUE DEDICO A:

Dios	Por permitirme llegar hasta este momento, acompañándome siempre y enseñándome el camino correcto a seguir. Este logro es para Él.
Mis padres	Por su apoyo constante para poder alcanzar mis metas; por el cariño con que me enseñaron a trabajar para ser mejor cada día.

AGRADECIMIENTOS A:

Dios y la Virgen María	Por ser mi guía y apoyo cuando más los necesito. Gracias por las constantes bendiciones para mí y mi familia.
Universidad de San Carlos de Guatemala y mis catedráticos	Por formarme como profesional.
Mis padres	Por sus sacrificios constantes; este logro es suyo, a ustedes les debo todo lo que soy.
Mis hermanos	Por su apoyo incondicional; porque juntos hemos enorgullecido a nuestros padres.
Mi abuela, Cleotilde Castañeda	Por haber cuidado de mí y mis hermanos y ser apoyo de mis padres cuando ellos más lo necesitaron.
Mi tía, Glenda Rivera, primos y sobrinos	Juntos formamos una gran familia, tanto para los momentos buenos como para los malos. Gracias por compartir con nosotros.

Mis amigos

A los compañeros de estudio y hermanos de desvelos y mis compañeros de trabajo quienes me brindaron una nueva visión de la vida. Gracias por su amistad.

Escuela de formación "Taller de Nazaret"

En especial a su director, maestros, servidores y alumnos; ha sido una gran experiencia ser parte de ustedes.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

GLOSARIO

Crawler	Un programa que sistemáticamente busca en la Word Wide Web (WWW) para crear un índice de datos.
CSS	*Cascade Style Sheet.* Conjunto de instrucciones HTML que definen la apariencia de uno o más elementos de un conjunto de páginas web con el objetivo de uniformizar su diseño.
Dominio Web	Sistema de denominación de hosts en internet el cual está formado por un conjunto de caracteres el cual identifica un sitio de la red accesible por un usuario.
HTML	Siglas del inglés *hypertext markup language* (lenguaje de marcado hipertexto). Es un lenguaje para crear documentos de hipertexto para uso en la Word Wide Web (WWW).
HTML5	Es el nombre que se usa para referirse a la quinta revisión del lenguaje HTML. Es el resultado de agrupar las especificaciones relacionadas con el desarrollo de páginas web.

HTTP	Protocolo de transferencia de hipertexto (HTTP). Tiene la ligereza y velocidad necesaria para distribuir y manejar sistemas de información hipermedia. Ha sido usado por los servidores World Wide Web desde su inicio en 1993.
Open source	Código fuente abierto o software libre, se refiere a un programa cuyo código fuente está disponible al público general para usar y/o modificar.
Software	Se refiere a programas en general, aplicaciones, juegos, sistemas operativos, utilitarios, antivirus, etc. Lo que se pueda ejecutar en la computadora.
URL	*Uniform Resource Locator.* Es el sistema de direcciones en internet. El modo estándar de escribir la dirección de un sitio específico o parte de una información en el sitio web.
World Wide Web	Conocido como WWW. Es el sistema de información basado en hipertexto, cuya función es almacenar y acceder a documentos a través de la red, de forma que un usuario pueda verlos usando un navegador web. Creada a principios de los años 90 por Tim Berners-Lee, investigador en el CERN, Suiza. La información transmitida por el WWW puede ser de cualquier formato: texto, gráfico, audio y vídeo.

RESUMEN

El presente estudio es el diseño para la creación e implementación de un sitio web que facilite a los profesionales guatemaltecos encontrar becas de su interés.

Se identificó la necesidad de una herramienta de este tipo, porque no existe un sitio web que centralice todas las fuentes de becas de posgrado y es complicado encontrar información necesaria para aplicar a una beca en el tiempo adecuado.

Para esto, se utilizará un buscador especializado, el cual se configurará para que busque datos especificados por cada persona. Se definirán los parámetros que un profesional deberá completar como parte de su perfil y el perfil de las becas en las que está interesado, la herramienta realizará la búsqueda en las fuentes de información previamente seleccionadas, y, al terminar la búsqueda, presentará un listado de todas las becas que mejor se ajusten al perfil ingresado.

Se espera que con la implementación de un proceso automático y periódico, se obtenga información reciente, completa y oportuna, para el mejor aprovechamiento de las oportunidades de becas que están disponibles para Guatemala.

OBJETIVOS

General

Implementar un sitio web amigable que permita a las personas registrar su perfil e intereses sobre becas de posgrado para que encuentren la beca que mejor se adapte a sus intereses.

Específicos

1. Definir los parámetros que una persona deberá completar como parte de su perfil profesional e implementar la herramienta, que permitirá almacenar y actualizar dicho perfil para la identificación de becas que sean de su interés.

2. Diseñar e implementar un proceso automático y periódico de búsqueda de información de becas de posgrado, con la finalidad de obtener información reciente, completa y de forma oportuna, para que el estudiante pueda aprovecharlas.

3. Centralizar en un solo lugar la información de becas, identificando y manteniendo un registro de los sitios que ofrecen el mismo tipo de información en Guatemala, así como los elementos de información que se obtendrá de cada uno de ellos.

INTRODUCCIÓN

Gracias a internet se tiene al alcance gran cantidad de información sobre cualquier tema, ya que muchas personas, oficial o extraoficialmente, publican páginas, sitios y blogs, exponiendo sus opiniones o estudios que cubren cualquier contenido.

Como consecuencia de esto, la investigación sobre algún tema específico requiere de una gran inversión de tiempo para filtrar la información presentada e identificar la que realmente se necesita. Por ejemplo, si se desea encontrar información sobre becas para estudiantes de posgrado, se realiza una consulta en un motor de búsqueda sobre el tema, y se presentan miles de *links* diferentes sobre becas.

Para ilustrar lo anterior, se utiliza el motor de búsqueda Google y se ejecuta una búsqueda con las palabras “becas de posgrado en Guatemala"; los resultados muestran 568,000 *links* que coinciden con esas palabras, pero no toda la información que se presenta da detalles de becas o incluye información que no es la que las personas necesitan específicamente.

Es por esto que se desea proporcionar una herramienta tecnológica que se dedique a realizar la búsqueda de información sobre becas de posgrado; para que sea esta la que realice el esfuerzo e invierta tiempo en encontrar las becas que cumplan con los criterios que el usuario especifique, en el menor tiempo posible.

El presente documento, que constituye el protocolo de trabajo de graduación, propone crear esta herramienta y cumplir con los requerimientos específicos de estudiantes de posgrado que desean encontrar los detalles necesarios para aplicar a una beca.

En el informe final se presentará lo siguiente: el primer capítulo incluye el marco teórico, en el cual se dará una explicación técnica del buscador automatizado del sitio web, se presentará la definición de qué es, cuáles son sus características, funcionalidades y los tipos que existen. Se explicará también qué es un sitio web y las características con las que debe cumplir.

En el capítulo 2 se presentará el análisis necesario y el diseño de la aplicación, previo a su creación; se describirán los componentes necesarios para la creación y funcionamiento del sitio web, también las pantallas que se crearán para que el usuario pueda interactuar amigablemente con el sitio.

En el capítulo 3 se explicará la forma en que se configuró el buscador que la aplicación utilizará, qué reglas se utilizaron y si fue necesario realizar modificaciones según las pruebas realizadas; por otro lado, se describirá cómo se eligieron los sitios web sobre becas que se utilizarán para realizar las búsquedas de información y los parámetros a los cuales deben ajustarse los sitios para ser incluidos como fuentes de búsqueda.

En el capítulo 4 se dan a conocer las opciones principales del sitio web ya terminado; se presentarán gráficamente las secciones del sitio y el formato que tendrá el mensaje de respuesta con los resultados de las búsquedas de cada usuario.

1. ANTECEDENTES

Cuando una persona está interesada en solicitar una beca, si no opta por presentarse físicamente a los lugares destinados para tal información, utiliza los motores de búsqueda de internet para recopilar información y selecciona los sitios en los cuales se especifican mejor los datos que necesita para realizar su solicitud. Existen sitios web que se encargan de recopilar información y presentan datos de becas como sus requisitos y ofrecimientos, fechas para aplicar, entre otros; algunos de ellos permiten realizar búsquedas por medio de ciertos criterios dentro del catálogo que manejan. A continuación, se detallará algunos de los sitios web dedicados a presentar información de becas:

- Secretaría de Planificación y Programación de la Presidencia (SEGEPLAN): es “una institución de apoyo a las atribuciones de la Presidencia de la República.” (SEGEPLAN, Secretaría de Planificación y Programación de la Presidencia, 2013). El sitio web de esta institución es bastante amplio y dentro de las opciones que presenta, se encuentra la opción “Sistema de Becas”. Las opciones de su menú principal se refieren a becas disponibles, oferentes de becas y la de registro a sus boletines. Se presentan listados con los detalles de becas, que están agrupados en distintas categorías, ya que existen becas reembolsables y crédito educativo, carreras de pregrado y posgrado, maestrías, doctorados y diplomados. Este es el sitio que presenta la mayoría de becas disponibles para Guatemala.

- Instituto para el Desarrollo de la Educación Superior en Guatemala (INDESGUA): es una asociación civil, formada por un grupo de profesionales en el año 2007, que ayudan a asesorar a personas en busca de educación superior y técnica. También ayudan a “gestionar y/o otorgar becas y crédito educativo a estudiantes y profesionales” (INDESGUA, Instituto para el Desarrollo de la Educación Superior en Guatemala, 2013). En su sitio, las personas pueden encontrar información sobre becas listadas por país, también provee una guía para escoger mejor el tipo de posgrado para cursar.

- Fundación Carolina: inició en el año 2000, con el propósito de promover relaciones culturales y de cooperación, educativa y científicamente, entre España y los países de la Comunidad Iberoamericana de Naciones. Con su “Programa de Formación”, ofrece becas a graduados universitarios, procedentes de los países de América Latina, miembros de la Comunidad Iberoamericana de Naciones y Portugal, para especialización y actualización de conocimientos por medio de posgrados y doctorados en España (Fundación Carolina, 2013).

Otros sitios se dedican específicamente a financiar a las personas becadas, como La Fundación Guatefuturo (Guatefuturo, 2013) que se esfuerza por financiar a guatemaltecos que ya han sido aceptados por universidades extranjeras, es una colaboración de los sectores públicos y privados. Los becados deben presentar su solicitud y si cumplen con los requisitos que cada financiamiento necesita, se aprueban. El sitio ofrece opciones de asesorías y consejos para los interesados en cómo aplicar a una beca y lo que deben realizar una vez obtenida la misma, para continuar con el proceso.

Los sitios web de las universidades nacionales tienen una sección dedicada a ofrecer información sobre las becas que cada universidad patrocina o con las que está relacionada. Ellos indican los requisitos para solicitarlos, detalles de las mismas y fechas para presentar solicitudes.

También hay testimonios de personas que han sido beneficiadas. Algunos ejemplos son: Universidad del Valle de Guatemala (Universidad del Valle de Guatemala, 2013), Universidad San Carlos de Guatemala (Universidad de San Carlos de Guatemala, 2013), Universidad Rafael Landívar (Universidad Rafael Landívar, 2013), Universidad Francisco Marroquín (Universidad Francisco Marroquín, 2013).

Todos estos sitios hacen referencia a SEGEPLAN ya que es el vocero principal de las becas nacionales. Cabe mencionar que un objetivo común de los sitios mencionados anteriormente, es ayudar a la superación de Guatemala fomentando la educación de profesionales.

2. PLANTEAMIENTO DEL PROBLEMA

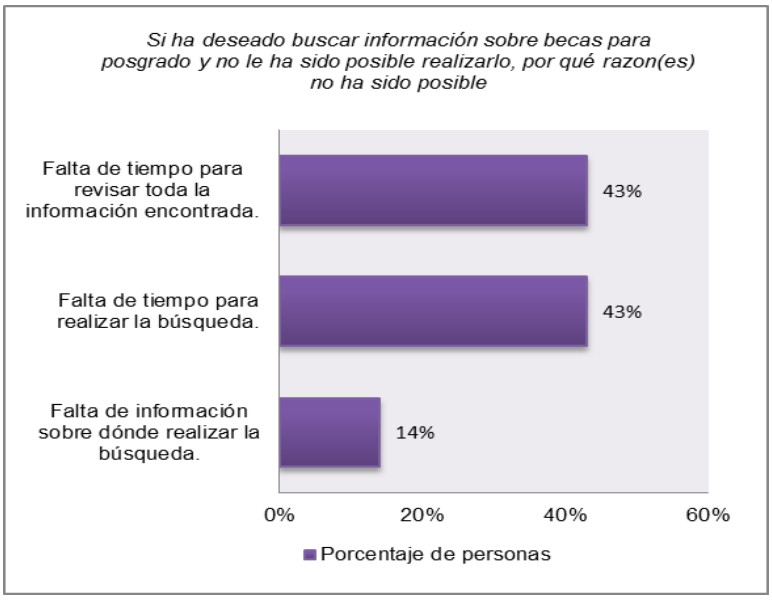
En Guatemala, existen sitios web que ofrecen becas de posgrado para profesionales que están interesados en aplicar a ellas. En una encuesta realizada a 28 profesionales universitarios (Picén, 2013) que han utilizado alguno de estos sitios web, se pueden desglosar los siguientes resultados:

- Al preguntar si han deseado realizar búsquedas de información sobre becas para posgrado y no le ha sido posible realizarlo, en las respuestas indicaron que: no tienen tiempo para realizar una búsqueda de becas correctamente (43 %), falta de tiempo para revisar toda la información que encuentra sobre el tema (43 %) y el resto de encuestados indicaron desconocer dónde realizar la búsqueda (ver figura 1).

- Al preguntarles respecto del mayor problema que tienen para encontrar información sobre becas, los encuestados indicaron que no encuentran algo de su interés (64 %) y el resto indica que la información no está completa o no está actualizada (ver figura 2).

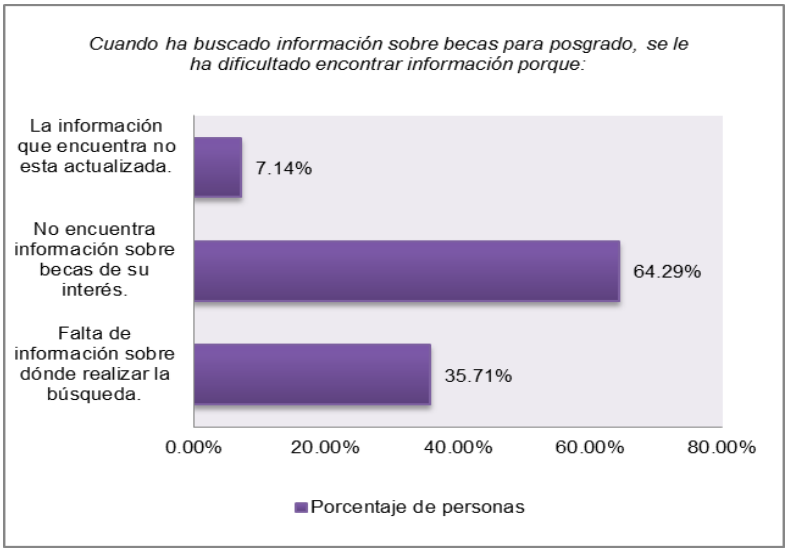
- Al consultarles si han utilizado los buscadores de los sitios web sobre becas, el 86 % indicó que los resultados no son precisos pero sí muestran cierta información; el 7% indica que la información es precisa y el 7% restante, señala que la búsqueda no cumplió los parámetros (ver figura 3).

Figura 1. **Resultados de la pregunta No. 1**



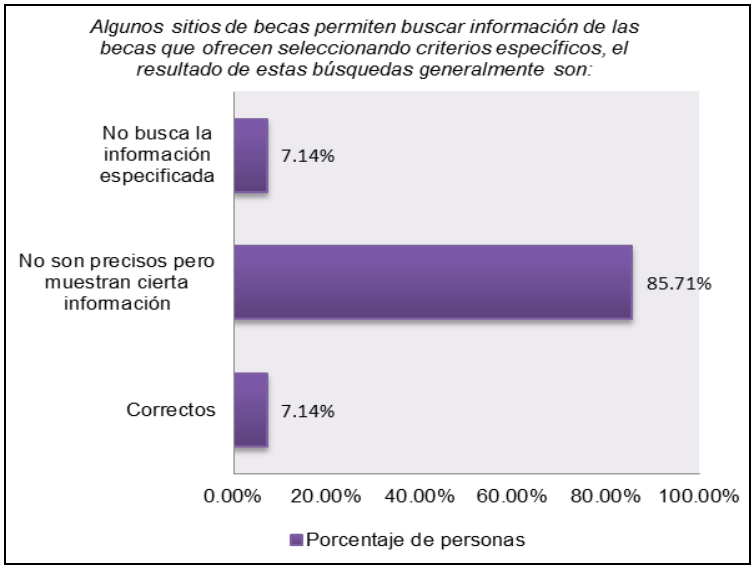
Fuente: elaboración propia. Encuesta realizada en mayo de 2013.

Figura 2. **Resultados de la pregunta No. 2**



Fuente: elaboración propia. Encuesta realizada en mayo de 2013.

Figura 3. **Resultados de la pregunta No. 3**



Fuente: elaboración propia. Encuesta realizada en mayo de 2013.

Con estos datos, se identifica que encontrar información en relación con las becas es complicado, ya que no existe un lugar centralizado que provea dicha información.

Al plantearse la pregunta ¿Existe una herramienta informática que encuentre información de becas específicas al perfil de las personas interesadas?, se identifica que al realizar la búsqueda de información, en la mayoría de los casos, no se toman en cuenta los parámetros personales ni académicos ingresados por los profesionales, lo que evita que se presente la información específica para cada persona.

Esto provoca que no sean utilizadas las oportunidades de estudio para posgrado, específicamente, becas que serían de interés a personas que ya

tienen un título universitario y desean alguna especialización o ampliar sus conocimientos.

El proceso de investigación se basa en las siguientes preguntas auxiliares:

- ¿Cómo se pueden optimizar los procesos de búsqueda de información de becas para lograr resultados que sean de interés y se apeguen al perfil profesional de los estudiantes?

- ¿Cómo se puede lograr la automatización de las búsquedas de información de becas de posgrado, que permita a un estudiante obtener información actualizada y en el tiempo adecuado para poder aplicar a ellas?

- ¿Cómo es posible contar con información completa y centralizada de becas para los estudiantes de posgrado de Guatemala para su mejor aprovechamiento?

3. JUSTIFICACIÓN

Cuando una persona desea obtener información de becas, debe navegar por muchas páginas de internet, ya que no hay un sitio que permita obtener la información que se ajuste a los criterios específicos que cada persona necesita. Luego, debe filtrar estos datos para encontrar las becas que cumplen con sus criterios personales, eliminando cualquier información que no les sea útil. También debe invertir tiempo corroborando la información recolectada, porque algunos sitios no la actualizan o no es real.

Según el periódico *Plaza Pública* (Goyzueta, 2013), hasta en marzo del 2013 el 81 % de las becas no han sido aprovechadas por distintas razones, tales como la falta de recursos o por compromisos laborales que no permiten ausentarse para estudiar, y también expone que por falta de información no se aprovechan estas oportunidades de estudio.

Una herramienta que realice estas búsquedas y que haga llegar los resultados, utilizando un medio fácil y rápido, a las personas interesadas, ayudaría a aprovechar mejor las oportunidades de estudio que ofrecen a Guatemala los distintos oferentes de becas y los profesionales graduados no tendrían que interrumpir sus actividades para obtener información de posibles becas que sean de su interés.

El siguiente trabajo se presenta como una investigación en la línea de tecnologías de la información y la comunicación, que permita innovar la industria de buscadores para cumplir las funciones previamente descritas.

El resultado esperado es el desarrollo de una herramienta que utilice las tecnologías existentes en las áreas de información, comercio y comunicaciones, diseñadas bajo una arquitectura de sistemas que permita integrarlas y enfocarlas en el cumplimiento de los objetivos propuestos. El componente de software principal, el buscador tipo *crawler,* que con base en el perfil y los intereses de becas de posgrado de los usuarios, filtrará la información de las fuentes especificadas. Actualmente no existen *crawlers* que realicen esta tarea.

La información que se obtenga del nuevo uso de la tecnología de los *crawlers* se podrá utilizar en otros temas relacionados, como búsquedas de información de colegios o tutores que cumplan con ciertos criterios; también podría utilizarse para búsqueda de cursos específicos, lo que ayudará a los padres de familia guatemaltecos a buscar la forma de educación que mejor se adapte a sus horarios.

4. NECESIDADES A CUBRIR Y ESQUEMA DE SOLUCIÓN

La aplicación que se desea implementar realizará la tarea de búsqueda de información de becas de posgrado por el usuario; es decir, la aplicación hará el trabajo por él. Por esto la aplicación deberá utilizar las herramientas y recursos necesarios para poder cumplir correctamente con los requerimientos del usuario.

Para crear la herramienta será necesario:

- Seleccionar un buscador *crawler* de tipo *open source,* el cual se configurará para que realice búsquedas y encuentre la información que cumpla con los parámetros específicos que se le indicarán; así, cada búsqueda será personalizada para cada usuario de la aplicación. Estas búsquedas se configurarán para que se realicen periódicamente, así la información que se presente al usuario será lo más actualizada posible.

- Se escogerán las fuentes de búsqueda de información de becas de posgrado para Guatemala; en estas fuentes el *crawler* comparará la información que ingrese el usuario y las becas disponibles que cumplan con dicha información.

- Se obtendrá un dominio web con el nombre EncuentraTuBeca.com.

- Se creará un sitio web utilizando HTML5, para poder ser utilizado en cualquier dispositivo, y herramientas tipo *open source*, para no incurrir en gastos de licenciamiento.

- El sitio web tendrá cuatro secciones:

    - La primera sección permitirá al usuario suscribirse al sitio para poder recibir el servicio de búsquedas automatizadas de becas. Para esto, tendrá que pagar un monto especificado.

    - La segunda sección será para obtener la información del usuario, específicamente su información académica y correo electrónico, donde podrá ser contactado.

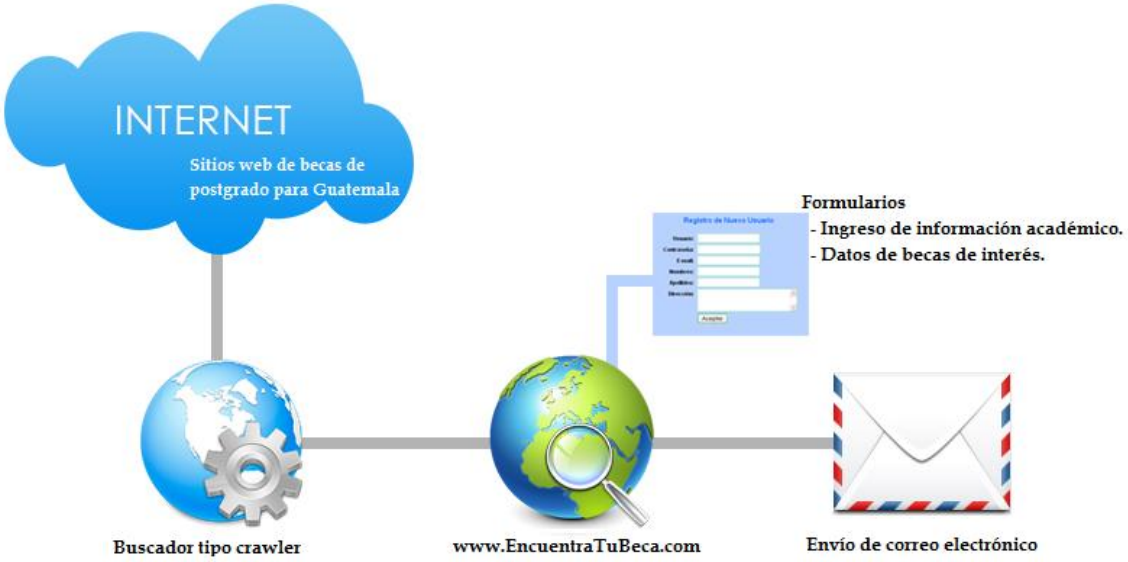
    - En la tercera sección se solicitará la información de las becas, de las cuales el usuario está interesado, para poder delimitar la búsqueda a este perfil.

    - Debe existir una sección para solicitar el pago del servicio al usuario, utilizando tarjeta de crédito o pagos en línea.

- Se enviará la información de becas encontradas al correo electrónico del usuario. Para esto será necesario crear un formato de la estructura del mensaje para presentar la información de forma completa, que contenga los detalles más importantes de la beca que se adecúa a su perfil y que sea entendible para el lector.

A continuación se presenta en forma gráfica cómo se relacionarán los componentes descritos previamente.

Figura 4. **Componentes del sitio web www.EncuentraTuBeca.com**



Fuente: elaboración propia, con programa Photoshop.

5. ALCANCES

El estudio especial de graduación en su carácter investigativo es descriptivo, y se enfocará en la creación de una herramienta web que provea el servicio de búsqueda de información de becas de posgrado, para profesionales graduados guatemaltecos.

El aspecto técnico de este estudio se enfoca en crear e implementar un sitio web que realizará la búsqueda de información de becas que cumpla con los criterios ingresados por los usuarios, criterios que se toman de su información personal y académica, y los detalles de becas que le interesaría conocer. Los beneficiarios de esta herramienta serán los profesionales graduados guatemaltecos que desean actualizarse o especializar sus estudios, ya sea en Guatemala o en el extranjero.

Los resultados esperados del estudio especial de graduación corresponderán a la implementación del sitio web que contendrá el buscador configurado, para enfocarse solo en fuentes que contengan información de becas de posgrado para guatemaltecos y que busque periódicamente dicha información, tomando como parámetros los perfiles de cada usuario. Se realizarán las pruebas necesarias para garantizar el funcionamiento correcto del sitio web, que presente la información requerida en el formato correcto.

Esta herramienta ayudará a que se divulgue la información de becas a una gran cantidad de personas, para que sean aprovechadas en su mayoría, en el momento oportuno.

6. MARCO TEÓRICO

6.1. Definiciones generales

Para la comprensión de todos los aspectos tecnológicos relacionados con la búsqueda de becas, a continuación se incluyen algunas definiciones.

6.1.1. Agentes inteligentes

Un agente inteligente es “una entidad software que, basándose en su propio conocimiento, realiza un conjunto de operaciones destinadas a satisfacer las necesidades de un usuario o de otro programa, bien por iniciativa propia o porque alguno de estos se lo requiere” (Hípola y Vargas-Quesada, 1999).

En el documento *Agentes de información*, (Julián, et al, 1999), “un agente viene definido por su flexibilidad, entendiendo por flexible que un agente sea:

- Reactivo: responda al entorno en que se encuentra;
- Proactivo: que sea capaz de intentar cumplir sus propios objetivos;
- Social: sea capaz de comunicarse con otros agentes mediante algún tipo de lenguaje”.

Es necesario especificar que un agente inteligente es un programa, pero no todos los programas que ejecutan búsquedas son agentes inteligentes.

Pueden ser denominados entidades individuales, ya que tienen control sobre sí mismos o partes de sí mismos, mediante procesos que les indican qué

hacer y cómo hacerlo. Para resolver un problema adecuadamente, se comunican con otros agentes.

Se tiende a etiquetar a un agente en función del papel que desempeña.

6.1.1.1. Agentes de información

Según Julián, Rebollo y Carrascosa, en su documento *Agentes de información* de 1999, los agentes inteligentes de información o agentes de internet pueden definirse como “aquellos sistemas software de computación que tienen acceso a múltiples y heterogéneas fuentes de información que están distribuidas geográficamente”, por esto, son útiles para búsquedas de información relevante.

Los agentes de información son una solución para el problema de encontrar la información deseada en el menor tiempo posible. Por esto los agentes deben proveer acceso a las fuentes de información en internet, recuperar, analizar, manipular e integrar la información solicitada.

Para cumplir con su función, los agentes deben de almacenar, aprender y manipular las preferencias y gustos de los usuarios; también ser flexibles para manejar los cambios de su entorno. Esto quiere decir que puede aprender y comunicarse con otros agentes con características similares; por esto, es necesario desarrollar agentes como entidades que constituyen un sistema; a este sistema se le llama multiagente.

Las interacciones entre agentes, como informar o consultar entre ellos, permiten la conversación entre ellos y razonar acerca del papel que ejecutan otros agentes.

Otros términos para referirse a los agentes de información son: *interface agents, system agents, advisory agents, filtering agents, retrieval agents, navigation agents, monitoring agents, recommender agents, profiling agents,* entre otros.

En el documento citado anteriormente, se proponen cuatro clases de agentes de información, tomando en cuenta las características que poseen como agentes, cada clase no es exclusiva; así que un agente puede pertenecer a varias de ellas al mismo tiempo:

- Agentes de información cooperativos o no cooperativos: si los agentes cooperan entre ellos para ejecutar tareas.

- Agentes de información adaptativos: se adaptan a los cambios en las redes y/o en la información.

- Agentes de información racionales: actúan por sí mismos o colaboran para aumentar su propio beneficio. Estos agentes se utilizan principalmente en el comercio electrónico, como los sistemas para subastas mediadas por agentes en la web.

- Agentes de información móviles: “son capaces de viajar de forma autónoma a través de internet y permiten balancear la carga en redes de gran escala o reducir la transferencia de datos entre servidores de información.”

Otra clasificación que propone el documento *Agentes de información* es la clasificación de los agentes de información según su función, y estos grupos, al

igual que en la clasificación anterior, no son excluyentes, y un mismo agente puede pertenecer a varios de ellos:

- Agentes de búsqueda (*retrieval agents*): son agentes que buscan, recuperan y proporcionan la información como si fueran auténticos gestores de información y documentación (*information brokers*).
  Para Hípola y Vargas-Quesada en su obra *Agentes Inteligentes: definición y tipología. Los agentes de información* (1999) explican que los sistemas expertos fueron diseñados para ejecutar consultas en una sola base de datos; luego, con la aparición de internet surgieron miles de bases de datos almacenadas en diferentes direcciones. Por esto surge un sistema descentralizado de recuperación de información basado en agentes inteligentes para localizar, recuperar y almacenar un resultado para un usuario en concreto. Independientemente del tipo de información que se desea encontrar, los agentes de búsqueda pueden diferenciarse por la entidad o persona para la que trabajan: usuarios y/o consultas y/o bases de datos. También se pueden distinguir por su forma de interactuar, si se relacionan libremente todos los agentes para resolver las consultas o solo con unos pocos agentes.

  - Los agentes de búsqueda inteligente para la web tienen la capacidad de hacer transparente la complejidad de la información almacenada en internet, filtrando lo más relevante sobre el tema solicitado. Las consultas pueden ser textuales o por las distintas partes en que la WWW se representa (Hípola y Vargas-Quesada, 1999).

- Agentes de filtrado (*filtering agents*): se utilizan para reducir el exceso de información, eliminando los datos que no son deseados, por ejemplo, los datos que no satisfacen completamente el perfil de usuario.

- Agentes de monitorización (*monitoring agents*): “proporcionan al usuario la información cuando sucede un determinado acontecimiento, por ejemplo, cuando los datos han sido actualizados, trasladados de lugar o borrados”.

6.1.2. *Internet bot*

Un *bot* de internet se refiere a aplicaciones de software que ejecutan tareas automatizadas. Los *bots* de internet realizan tareas que son simples y estructuralmente repetitivas, a un mayor rango del que sería humanamente posible. También son llamados *robots web, robots www* o simplemente *bots* (Peña, 2005).

6.1.3. Indexación automática

Es utilizar un programa de software o algoritmo para recorrer archivos, documentos y sitios web en busca de palabas claves o *keywords*. Los documentos en internet usualmente tienen un tema específico y claves recurrentes que presentarán este tema. Un programa de indexación automática revisará el documento y lo categorizará, basado en estas palabras.

Donde más se utiliza este tipo de indexación es en los motores de búsqueda, porque estos deben emparejar las palabras claves que el usuario ingresa con todos los sitios web en existencia. Sin esta característica sería muy difícil que las personas pudieran encontrar los sitios relevantes.

Las ventajas de la indexación automática es que la computadora puede indexar y buscar un documento mucho más rápido que una persona, buscar palabras en muchos sitios web y categorizar documentos después de encontrarlos.

La desventaja de estos programas o algoritmos es que un ser humano debe crearlo e indicar la forma en que se emparejan las palabras, lo que puede llevar a errores por mala programación, como omitir puntos clave. Por otro lado, el indexador no puede distinguir algunas palabras o puede tener muchas imprecisiones por palabras (*wisegeek: what is automatic indexing*?).

6.2. Información en internet

La *World Wide Web* (WWW) se puede considerar como una biblioteca, ya que se tiene mucha información de una gran variedad de temas: está disponible desde cualquier parte del mundo y a cualquier hora. Otra característica importante es su accesibilidad, ya que varias personas pueden consultar el mismo archivo al mismo tiempo, y a su vez, pueden agregar y modificar los documentos que forman parte de esta “biblioteca”.

Esto enriquece la web pero también puede causar que no todo lo que forme parte de ella sea de calidad relevante o se adapte a las necesidades de los tema buscados, por lo que, generalmente, provoca que las personas deban navegar bastante tiempo para poder discriminar qué contenido es útil y cuál no (Alvarez, et al, 2012).

Es importante mencionar que no solo documentos forman parte de la web, existen muchos formatos, como texto, imágenes, audio y vídeo (Álvarez Díaz, 2007).

La información contenida en la WWW es tan grande que se estima que un motor de búsqueda solo tiene acceso al 15 % o 16 % de toda la web. Tanta información en un solo lugar hace muy difícil la búsqueda de un tema específico, y tan importante es que internet contenga información como encontrar dicha información. Es por esto que los navegadores deben priorizar el tipo de información que les interesa para que se enfoquen solo en temas específicos (Peña, 2005) y (Montero, 2009).

Otro factor que se debe tomar en cuenta es que la información en internet cambia constantemente y aunque los buscadores encuentren la información un día, al siguiente es posible que ya no la puedan identificar, ya que pudo haber sufrido cambios que hacen que se clasifique la información de una forma diferente (Peña, 2005).

6.2.1. La web oculta

Una gran parte de la información de la WWW no es alcanzable por la mayor parte de los motores de búsqueda, debido a las tecnologías que utilizan, como las navegaciones a través de menús emergentes, diferentes capas de datos que se ocultan o hacen visibles dependiendo de las acciones del usuario, sistemas de redirecciones o mecanismos de mantenimiento de sesión (Álvarez, et al, 2011). Esta parte de la web se conoce como web oculta, *Hidden web*, que forma parte de la información que no puede ser accesible al utilizar un motor de búsqueda.

Las características del contenido de la web oculta son: cobertura importante y muy amplia, alta calidad y el contenido de este es mayor a todos los contenidos impresos (Madaan, et al, 2010).

6.3. Motores de búsqueda

Para el desarrollo del presente proyecto se construirá un sitio web para el registro del perfil de los usuarios que deseen utilizarlo, y un *crawler* para los procesos de búsqueda de la información que se desea. En este apartado se explicará con mayor detalle en qué consiste un sitio web y el *crawler*.

6.3.1. Sitios web

"Es un conjunto de páginas que se encuentran alojadas en un servidor web. Generalmente estas páginas estarán escritas en el lenguaje HTML, las cuales son accedidas a través del protocolo HTTP, que permite transmitir los documentos del servidor web hacia el navegador del cliente." (Arrecis y Telón, 2009). Existen diferentes clasificaciones del sitio web, dependiendo del contenido que se desea presentar a los usuarios, como los sitios de noticias, educativos, entretenimiento, entre otros.

6.3.1.1. Diseño de un sitio web

El diseño de un sitio web es la fase más importante de su desarrollo ya que en este momento se identifican los requisitos del sitio, qué es lo que se desea presentar en él, tanto para el desarrollador como para los clientes (Mcheley, 2008). Hay cinco preguntas claves que se deben realizar en esta fase:

- ¿Cuál es el propósito y la meta del sitio web? Definir el propósito del sitio ayudará a enfocar los recursos para obtenerlo.

- ¿Quiénes serán la audiencia? Los usuarios que utilizarán el sitio web ¿serán usuarios expertos o inexpertos? Esto permite crear el sitio adecuado, atractivo, para los usuarios que lo utilizarán.

- ¿Cuáles serán los indicadores de éxito? Se debe medir el nivel de éxito o fracaso, así se puede monitorear cómo el proyecto va progresando luego de su implementación. Dependiendo del tema del sitio, estos indicadores pueden ser el número de ventas, cuántos usuarios los visitan por hora, cuántos referencian al sitio, entre otros.

- ¿Cómo se comercializará el sitio? Es necesario promocionar el sitio para que sea conocido por el mercado objetivo; se puede utilizar el posicionamiento orgánico, utilizando las redes sociales o creando campañas de mercadeo, dependiendo de cuáles sean las necesidades del sitio web, lo importante es que el sitio sea promocionado.

- ¿Cuáles son los requerimientos de diseño? Se debe definir si utilizarán elementos visuales para identificarse, como logos o una paleta de colores. Este diseño debe estar en todo el sitio para poder establecer la “marca” del sitio y así los clientes puedan distinguirlo fácilmente.

6.3.1.2. Estilos de sitios web

- Estáticos: un sitio web estático presenta el mismo contenido siempre, independientemente del usuario o cualquier otra característica. Las páginas contenidas en estos sitios se encuentran almacenadas de la misma forma que serán presentadas al usuario. El proceso de actualización del contenido del sitio debe hacerse manualmente, a través de editores o cualquier otro software, teniendo en cuenta que la persona

que lo realice debe tener conocimientos sobre el lenguaje de programación (Arrecis & Telón, 2009).

- Dinámicos: se les llama así cuando un sitio web presenta información dependiendo del perfil del usuario accediendo a él, también conocido como “aplicación web”. Este no tiene páginas almacenadas en la forma en la que el usuario las verá, sino que se generan automáticamente, dependiendo de diferentes parámetros; además, bajo los mismos parámetros entre diferentes solicitudes por el mismo usuario, el sistema puede retornar diferente información por actualizaciones que puedan ocurrir dentro del sitio.

  Una aplicación web igualmente realiza procesos que alteran el contenido del sitio web como ingresar nueva información, imágenes, documentos, entre otros; esto a través de otras herramientas dentro del propio sitio. La aplicación permitirá al usuario generar información dinámicamente y ejecutar tareas como búsquedas de contenidos, actualizaciones e ingreso de nueva información (Arrecis & Telón, 2009).

6.3.2. *Crawlers*

Un *crawler*, también conocidos como *Web Crawler, Web Spider o Web Walker*, en español Araña Web, es un programa de software que recorre páginas de la web en forma automática y sistemática.

Estos son un tipo especializados de internet *bot*, con la tarea específica de recorrer páginas web, descargarlas y procesarlas (Montero, 2009).

Es un programa automatizado relativamente simple o un script que metódicamente escanea o “se arrastra” a través de páginas de internet para crear un índice de los datos que se están buscando; estos programas usualmente son creados para hacer uso de ellos una sola vez, pero pueden ser programados para su uso a largo plazo.

Se utilizan para los motores de búsqueda, con los cuales se aseguran que sus bases de datos están actualizadas, pero también tienen otros usuarios como lingüistas que investigan cuál es la palabra más utilizada actualmente o investigadores de mercado para determinar y evaluar tendencias en un mercado determinado y cualquiera que necesite buscar información de internet de una forma organizada (*Wisegeek: what is a web crawler?).*

“Se utilizan también para tareas de mantenimiento automatizado de un sitio web, tales como comprobar enlaces o validar código HTML. Incluso pueden emplearse para reunir tipos específicos de información procedente de páginas web, como es el caso de recolectar direcciones de correo electrónico para enviar correos basura” (Dixxinet: Crawler, Rastreador, 2012).

Es un método importante para recolectar datos en internet, que se expande tan rápidamente. Un vasto número de páginas de internet son agregadas continuamente y la información cambia constantemente.

Es importante mencionar que los *crawlers* convencionales solo pueden acceder a las páginas que son estáticas y públicas en la web, que son una porción de toda la información que está en internet. La información generada dinámicamente por un servidor basado en las respuestas de un usuario no puede ser accedida por los *crawlers.* También las páginas web utilizan lenguajes *script* que son ejecutados del lado del cliente web.

Por otro lado, los *crawlers* se puede diferenciar por el ámbito en que funcionan: globales, que son los que recuperan toda o gran parte de la información de la web, y dirigidos o focalizados, que están orientados a recuperar una parte concreta o más reducida de la web (Álvarez, et al, 2011).

6.3.2.1. Funcionamiento

Un *crawler* dispone de "un conjunto inicial de URLs, conocidas como semillas, va descargando las páginas web asociadas a las semillas y buscando dentro de estas otras URLs. Cada nueva URL encontrada se añade a la lista de URLs que el *crawler* debe visitar. A este proceso se le denomina *recolección de URLs*. Existen distintas políticas para escoger la siguiente URL que el *crawler* visitará. En general, estas políticas se basan en las respuestas a preguntas tipo:

- ¿Es importante la página en la que estoy?
- ¿Es importante el sitio en el que se encuentra la página web actual?
- ¿He visitado ya alguna página web del dominio de la página a la que tengo intención de dirigirme?" (Montero, 2009)

Cuando el *crawler* accede a una nueva URL, la página web asociada es descargada al servidor donde se ejecuta la búsqueda; luego, estas son parseadas y procesadas; esto quiere decir que en este momento se decide si el texto o sus enlaces o solo sus imágenes o tal vez todo, es útil para la búsqueda.

Es importante mencionar que ningún *crawler* puede acceder a todas las URLs que hay en internet, pues el número de páginas existentes es demasiado grande, por esto es importante definir las políticas para escoger las URLs correctas.

Tras el parseo, el *crawler* aplica algún tipo de algoritmo para obtener la información deseada. Por ejemplo, comprobar la disponibilidad de un enlace o verificar el tamaño de una imagen.

Algunas de las dificultades a las que los *crawlers* se deben enfrentar, son las mencionadas en la sección información en internet: “enormes cantidades de páginas que recorrer, elevado número de actualizaciones de páginas existentes, páginas que crean su contenido de forma dinámica, redireccionamientos, entre otros” (Montero, 2009).

Es por esto que existen *crawlers* para la web oculta del lado del servidor y para la del lado del cliente. La web oculta del lado servidor se ocupa de la amplia cantidad de sitios web en los cuales se accede al contenido mediante formularios; este tipo de contenido es de gran cantidad y calidad.

Respecto de la del lado cliente, básicamente se resumen en las siguientes dos aproximaciones: acceder al contenido y enlaces mediante intérpretes que permitan ejecutar los *scripts* o utilizar mininavegadores (Álvarez, et al, 2011).

6.3.2.2. Protocolo de exclusión de robots

Los sitios web tienen el control de qué información es pública en internet. Cada sitio web puede indicar qué partes de ellos son privadas y cuáles son públicas, para ser presentadas en las búsquedas de los navegadores de internet. La forma en que los sitios web especifican que algunas partes son privadas y no pueden ser accedidas por navegadores es por medio del Protocolo de Exclusión de Robots (*The Robots Exclusion Protocol*, REP), que es un mecanismo sencillo pero muy poderoso.

El archivo robots.txt define el protocolo para un sitio web, las políticas de privacidad se colocan en este archivo. En él se controla cómo los motores de búsqueda acceden a los sitios web. Con este archivo se pueden controlar el acceso a muchos niveles (todo el sitio web), directorios individuales, tipos específicos de páginas, páginas individuales, entre otros. También define las directivas que excluyen a los robots web de los directorios o archivos de un sitio web, define directivas de *crawling*, no directivas de indexación (Crow, 2007).

Los buenos robots web se adhieren a las directivas del archivo robots.txt, los que no son buenos, *spambots*, no lo hacen. Este archivo es de acceso público, no se recomienda incluir archivos o carpetas con información crítica de la empresa, tampoco se debe confiar en él para proteger información privada de los navegadores (Clay, 2010).

Requerimientos clave: el archivo debe estar escrito en letras minúsculas, y ser accesible al público; el tipo debe estar en un formato estándar de archivo (como ASCII o UTF-8), estar localizado en la carpeta raíz del *host* del sitio web y el archivo debe ser válido para las versiones seguras del dominio (https). Algunos motores de búsqueda pueden tener limitaciones de tamaño para el archivo robots.txt (Clay, 2010).

6.3.2.3. Características

De los documentos *Introduction to Information Retrieval, Chapter 20: Web crawling and indexes* (Manning, et al, 2008) y de *WebCrawlers* (Peña, 2005) se dan a conocer las siguientes características:

- Robusto: el contenido de la web crea “trampas” que, ya sea intencionalmente o no, provocan que los *crawlers* no puedan seguir su

búsqueda de información, perdiéndolos en una cantidad infinita de páginas en algún dominio. Por esto, es necesario que el diseño de los *crawlers* permita evadirlas o recuperarse, una vez hayan caído en esas trampas.

- Escalable: su arquitectura debe permitir que crezcan los niveles de *crawleo,* agregando máquinas extras y ancho de banda.

- Bajo costo, alto rendimiento: el *crawler* debe hacer uso eficiente de varios recursos a la vez, como procesamiento, almacenamiento y ancho de banda de la red.

- Cortesía: debe tener políticas implícitas y explícitas para regular el rango en el que el *crawler* puede buscar información en un servidor web en específico; estas políticas siempre deben ser respetadas.

- Distribuido: debe tener la habilidad de ejecutarse en un ambiente distribuido, utilizando varias máquinas.

- Calidad: dado que gran cantidad de las páginas web no son de utilidad para una búsqueda específica, el *crawler* debe orientarse a encontrar primero las páginas con información útil.

- Actualizado: en muchas aplicaciones, el *crawler* debe ejecutarse continuamente para obtener información actualizada. Para esto, debería visitar la página en una frecuencia que se aproxime a la frecuencia que la página se refresca en el servidor.

- Extensible: los *crawlers* deberían de diseñarse para ser extensibles, acoplarse a nuevos formatos, nuevos protocolos, entre otros. La arquitectura del *crawler* debe ser modular.

6.3.2.4. Políticas de comportamiento

El comportamiento de un *crawler* es el resultado de una combinación de políticas:

- Política de selección: especifica qué páginas descargar
- Política de cortesía: especifica cómo evitar sobrecarga a los sitios web
- Política de revisita: indica cuándo revisar las páginas por cambios
- Política de paralelización: señala cómo coordinar *crawlers* distribuidos

6.3.2.4.1. Políticas de selección

También llamadas estrategias de ordenamiento de páginas.

"Se requiere de una métrica de importancia para dar prioridad a las páginas webs. La importancia de una página web está en función de su calidad y del número de visitas que tiene, entre otros factores. El diseño de la política de selección tiene un inconveniente que es el hecho de que el sistema completo de páginas web a visitar no se conoce durante la ejecución del web *crawler,* por lo que se debe trabajar con información parcial" (Del Coso Santos, 2009).

Según Baeza-Yates, Castillo, Marín, & Rodriguez, 2005, en su documento *Crawling a Country: Better strategies than breadthfirst for web page ordering*" si se está revisando una porción de la web, que ya fue analizada previamente, unas semanas o meses antes, se tiene disponible información histórica.

Ese es el caso típico de los motores de búsqueda y la información histórica, puede ser usada para dirigir el *crawleo* hacia páginas que tenían un alto *pagerank* (es la cantidad de veces que una página web es referenciada por otras páginas) en ese último *crawleo*.

También se toma en cuenta la estimación de calidad de página. Dicha estimación se puede combinar con los cambios de información de la página, porque en algunos casos un *crawler* puede preferir un programa de descargas, priorizando páginas de alta calidad que sufren cambios constantes.

Baeza-Yates, et al (2005), indican que existen tres tipos de estrategias, tomando en cuenta la cantidad de información que pueden utilizar:

- Estrategias sin información extra:

    - *Breadth-first*: el *crawler* empieza visitando todas las páginas de inicio de todos los sitios web semilla y las páginas web se acumulan y guardan de tal forma, que las nuevas páginas se agregan al final. Hay estudios que confirman que esta estrategia captura páginas iniciales de alta calidad.

    - *Backlink-count*: visita primero las páginas con la mayor cantidad de links apuntando hacia ella; así la nueva página para ser visitada es la más referenciada por las páginas ya descargadas.

    - *Batch-pagerank*: calcula una estimación de *pagerank,* usando las páginas revisadas hasta ese momento (cada K páginas descargadas); las próximas K páginas a descarga son las que

poseen el mayor *pagerank* estimado. Esta estrategia es mejor que *backlink-count*, pero puede ser inexacta.

- *Partial-pagerank*: es como *batch-pagerank*, pero entre los recálculos de *pagerank*, un *pagerank* temporal es asignado a las nuevas páginas, usando la suma del *pagerank* de las páginas, apuntando hacia ella y dividiéndola por el número de links hacia otras páginas, que las contienen.

- *OPIC*: esta estrategia esta basa en el algoritmo OPIC - *On line page importance computation*; puede verse como una estrategia *backlink-count* ponderado. Todas las páginas empiezan con la misma cantidad de "efectivo". Cada vez que una página es visitada, su "efectivo" se divide entre las páginas que enlaza. La prioridad de una página no visitada es la suma del "efectivo" que recibió de las páginas que apuntan a ella. Esta estrategia es similar a *pagerank,* pero no tiene *links* al azar y el cálculo no es iterativo, así que es mucho más rápido.

- *Lager-sites-first*: la meta de esta estrategia es evitar tener muchas páginas pendientes en cualquier sitio web y tener solamente un pequeño número de grandes sitios web que pueden causar pérdida de tiempo por las políticas de cortesía. El *crawler* utiliza el número de páginas no visitadas encontradas hasta el momento, como prioridad para escoger un sitio web, y empieza con los sitios con el mayor número de páginas pendientes.

- Estrategias con información histórica: estas utilizan el *pagerank* de un *crawleo* anterior, como una estimación del *pagerank* para el nuevo *crawleo* y empieza con las páginas con el mayor *pagerank*. Se aplican estrategias que tratan las páginas encontradas en el *crawleo* actual, que no fueron encontradas en el *crawleo* anterior. Entre estas están:

    - *Historical-pagerank-omniscient*: a nuevas páginas se les asigna un *pagerank* tomado de un oráculo que conoce todo el grafo.

    - *Historical-pagerank-random*: a nuevas páginas se les asigna un valor de *pagerank* seleccionado uniformemente al azar entre los valores obtenidos de *crawleos* anteriores.

    - *Historial-pagerank-zero*: a nuevas páginas se les asigna un *pagerank* de cero, páginas viejas son visitadas primero y luego las nuevas.

    - *Historical-pagerank-parent*: a nuevas páginas se les asigna el *pagerank* de una página padre (la página en la que el *link* se encontró) dividido por el número de *links* hacia otras páginas de la página padre.

- Estrategias con toda la información

    - *Ominiscient*: esta estrategia puede buscar un oráculo que conoce completamente el grafo de la web y tiene calculado el *pagerank* actual de cada página. Cada vez que necesita priorizar una descarga pregunta, para luego descargar la página con el mayor *pagerank*. Esta estrategia está unida a las mismas restricciones

como las otras y solo puede descargar páginas si previamente había descargado la página que apunta a ella.

6.3.2.4.2. Políticas de cortesía

Para optimizar recursos de red y evitar la saturación de servidores, el *crawler* debe espaciar las solicitudes a un sitio web. Estas políticas son muy importantes para evitar obstruir las tareas del sitio y así encontrar la información deseada (Del Coso Santos, 2009). Existen distintas políticas para configurar un *crawler* (Peña, 2005):

- Intervalo de 10 segundos, propuesto por Cho y García-Molina
- Intervalo de 15 segundos, propuesto por WIRE *Crawler*
- Mercator propone que si tardó t segundos en bajar un documento, entonces espera 10*t segundos antes del siguiente

6.3.2.4.3. Políticas de revisita

Desde el punto de vista de un motor de búsqueda, hay un costo asociado a no detectar un evento y por ello, tener una copia no actualizada de un recurso. Las funciones de costo más usadas son:

- Frescura: es una medida que indica si la copia local está actualizada o no.
- Edad: esta es una medida que indica qué tan desactualizada es la copia local.

El objetivo del *crawle*r es mantener el promedio de frescura de las páginas en su colección, lo más alto posible, o mantener el promedio de edad de páginas tan bajo como sea posible. Estos objetivos no son equivalentes; en el

primer caso, el *crawler* solo se preocupa de cuantas páginas desactualizadas existen, mientras que en el segundo caso, el *crawler* se preocupa de qué tan antiguas son las copias locales. Se explicarán dos políticas de revisita (Peña, 2005):

- Uniforme: revisita todas las páginas con la misma frecuencia, independientemente de su tasa de cambio.

- Proporcional**:** revisita más las páginas que cambian más frecuentemente. Es decir, la frecuencia de visita es proporcional al estimado de actualizaciones de la página.

6.3.2.4.4. Políticas de paralelización

Un *crawler* paralelo corre múltiples procesos, valga la redundancia, en paralelo. El objetivo es maximizar la velocidad de descarga mientras se minimiza la sobrecarga de paralelización y se evita descargas de la misma página. Para eso, el *crawler* necesita una política para asignar las nuevas URLs encontradas durante el *crawling*, ya que la misma URL puede ser encontrada por dos procesos *crawling* diferentes al mismo tiempo. Existen dos tipos de políticas (Del Coso Santos, 2009):

- Asignación dinámica: en la que existe un servidor central que se encarga de asignar las URLs a visitar a los distintos *crawlers*, lo que permite balancear la carga que tienen los robots.

- Asignación estática: en la que hay una regla definida desde el comienzo de la ejecución que define cómo asignar las nuevas URLs.

6.3.2.5. *Crawler* focalizado

También llamado *crawle*r dirigido o *focused crawler.* Este tipo de *crawler* localiza, obtiene, indexa y mantiene páginas relacionadas con un conjunto de temáticas determinadas, que representan segmentos relativamente limitados de la web. Comienza con un pequeño conjunto de páginas relacionadas con un tema; un "clasificador" determina qué enlaces seguir para cada página obtenida en función de su relevancia potencial para dicho tema.

El “clasificador” es el encargado de decidir la dirección de exploración del *crawler.*

Existen tres aproximaciones para la construcción de *crawler* focalizado, en función de la información utilizada para el “clasificador”, para determinar si las páginas se ajustan al tema específico de una tarea de *crawling* concreta: los que hacen solo uso del contenido de las páginas obtenidas; los que utilizan la estructura de enlaces existentes entre los documentos web para obtener la relevancia de las páginas o el nivel de confianza en una ruta de exploración determinada.

“Las diferentes aproximaciones no son exclusivas, sino que pueden complementarse para mejorar la efectividad del sistema final” (Álvarez Díaz, 2007).

Un *crawler* focalizado busca de forma selectiva las páginas que son relevantes para un conjunto predefinido de temas; esto supone un ahorro significativo en recursos de *hardware*, mejorando además la relevancia de los documentos indexados.

Otras ventajas que provee este tipo de *crawler* frente a un *crawler* estándar son (Alvarez, et al, 2012):

- Disponer de una evaluación temprana de recursos.
- Disponer de versiones segmentadas de la web, ya que se dispone de segmentos de la web que favorecen su posterior análisis.
- Mejorar la precisión de las recomendaciones realizadas por el *crawler.*

6.3.2.6. *Crawler* ilegal

Es importante mencionar que “existen *crawlers* con fines ilegales, se les denomina *spambots* y tienen un propósito malicioso”. “También tratan de *hackear* los sitios web para obtener información que no es gratuita o pública”. Para lograr esto utilizan técnicas como la “falsificación de identidad” (Montero, 2009).

Existen análisis para detectar el *web spam,* ya que se han utilizado las tecnologías del lado de cliente para engañar a los *crawlers.* Las técnicas comunes para *web spam* son *cloaking* y *redirection spam*. *Cloaking* tiene la función de detectar cuándo es un usuario normal o un *crawler,* el que realiza la petición de la página; si el que realiza la petición es un *crawler*, el sitio web mostrará un contenido diferente al que mostraría a un navegador de un usuario.

Las técnicas de *redirection spam* ocultan las redirecciones para ser ejecutadas únicamente en un navegador. En ambos casos se consigue “mentir” al buscador de forma que indexe unos contenidos diferentes a los que realmente son (Álvarez, et al, 2012).

7. PROPUESTA DE ÍNDICE DE CONTENIDOS

8. METODOLOGÍA

8.1. Matriz de operacionalización de variables

En la tabla que se presenta a continuación, se definen las variables e indicadores del estudio especial de graduación.

Tabla I. **Matriz de operacionalización de variables**

Variables	Definición conceptual	Subvariables	Indicadores	Dimensiones
Información de becas para posgrado.	La dificultad para encontrar información de becas para posgrado.	Becas no utilizadas	Cantidad de becas perdidas por mes y año.	Becas pérdidas.
		Links con información útil	Cantidad de links útiles al realizar la búsqueda.	Links con información útil.
Buscador parametrizable	Los buscadores tipo *crawler* pueden realizar búsquedas de muchas páginas en segundos, por lo que puede consultar muchas fuentes distintas de información en poco tiempo.	Buscadores automatizados	Número de tipos de *crawlers*.	• Crawlers pagados. • Crawlers open source. • Otros.
		Fuentes de información	Fuentes distintas de información	• SEGEPLAN • Sitios Web Embajadas • Sitios Web de Universidades • Otros
		Búsquedas periódicas	Período en que el buscador repetirá la búsqueda	• Segundos • Minutos • Horas

Fuente: elaboración propia.

8.2. Diseño, tipo y alcances del estudio

Al diseñar el tipo de investigación se determinó que se hará una investigación cualitativa; se seleccionará una herramienta web que permita facilitar el proceso para obtener información de una beca, con opciones preestablecidas.

Tabla II. **Diseño, tipo y alcances del informe final de graduación**

Tipo	Alcances
Investigación cualitativa	Descriptivos
Se implementará una herramienta web que ayude a las personas interesadas en becas de posgrado a obtener información de forma más fácil y rápida de lo que existe actualmente.	Los usuarios obtendrán resultados que se adhieren a su perfil por lo que obtendrán información personalizada y actualizada.

Fuente: elaboración propia.

8.3. Fases del estudio

- Observación: se identificó que los sitios web existentes proveen información de becas que no está centralizada y, en general, estos sitios no tienen buscadores de información que ingresen criterios específicos. Así se identificó la necesidad de una plataforma web que centralice la información de dichos sitios y que provea una forma fácil de encontrar la información que se necesita.

- Análisis del entorno: se estudiarán los sitios ya existentes sobre becas, para identificar los pros y los contras de cada uno. Estas serán las principales fuentes de información para la aplicación, creándose un registro de los que proveen la información más completa y actualizada.

- Técnicas para recolección de investigación: para la investigación se utilizarán fuentes secundarias de información. Se realizará una investigación bibliográfica de las herramientas que se utilizarán para crear la aplicación.

- Revisión documental: se realizó una búsqueda en internet para encontrar información sobre los temas relacionados con este estudio, se leyeron superficialmente y se escogieron los documentos con fecha después del 2007 y que su información se complementara una con otra. Luego se leyeron las fuentes escogidas, detenidamente, y se seleccionaron las que proveyeran la información más completa. En el caso de algunos temas, fue necesario buscarlos específicamente; estos no tuvieron la restricción de tener fecha reciente.

- Trabajo de campo: se publicará una versión de prueba del sitio web con funcionalidad básica, así se podrá estudiar si dicho sitio provee la funcionalidad necesaria y se podrán corregir o agregar nuevas opciones en la versión final. Se realizará una observación directa de los sitios web de becas que existen actualmente, de las secciones de becas de los distintos sitios web de las universidades y de las embajadas de algunos países, para analizar la información que pueden proveer y utilizarlos como fuentes para los buscadores.

- Población y muestra: el producto de este estudio será para uso de profesionales graduados guatemaltecos. Se tomará como muestra de profesionales graduados a los estudiantes de la Maestría de Tecnologías de la Información y Comunicación de la Universidad de San Carlos de Guatemala.

9. TÉCNICAS DE ANÁLISIS DE INFORMACIÓN

9.1. Estadística descriptiva e inferencial

Se utilizarán estas herramientas para procesar las respuestas de las encuestas realizadas a los usuarios del sistema y evaluar la opinión que tengan de la herramienta.

9.1.1. Gráficas

Se elaborarán gráficas para visualizar la información obtenida de las estadísticas. Se presentarán gráficas del diseño de las distintas secciones del sitio web, para que así sea posible visualizar mejor cómo estará creado y cómo funcionará al estar implementado.

9.1.2. Interpretación del contenido investigado

Se presentará la información obtenida luego de revisar las fuentes bibliográficas escogidas para sustentar esta investigación y los sitios web que proveen información de becas en general. Los sitios web de becas que se utilizarán como fuentes de las búsquedas automáticas, se escogerán tomando en cuenta estas características:

- Los oferentes de las becas que presentan deben ser confiables.
- Los estudiantes beneficiados de las becas deben ser profesionales graduados guatemaltecos.
- Deben ofrecer becas para posgrado.

- Debe ser un sitio web actualizado al presente año y el periodo para aplicar a las becas ofrecidas debe estar activo.

9.1.3. Tablas comparativas

Se realizará un cuadro comparativo con la información obtenida en el punto anterior, para comparar las distintas características de los sitios web y así visualizar mejor qué fuentes serán las más adecuadas para realizar las búsquedas de datos de becas de posgrado.

10. CRONOGRAMA

A continuación se presentan las actividades que abarcan las actividades del estudio especial de graduación, la creación de la aplicación hasta la creación y entrega del informe final.

Tabla III. **Tareas programadas**

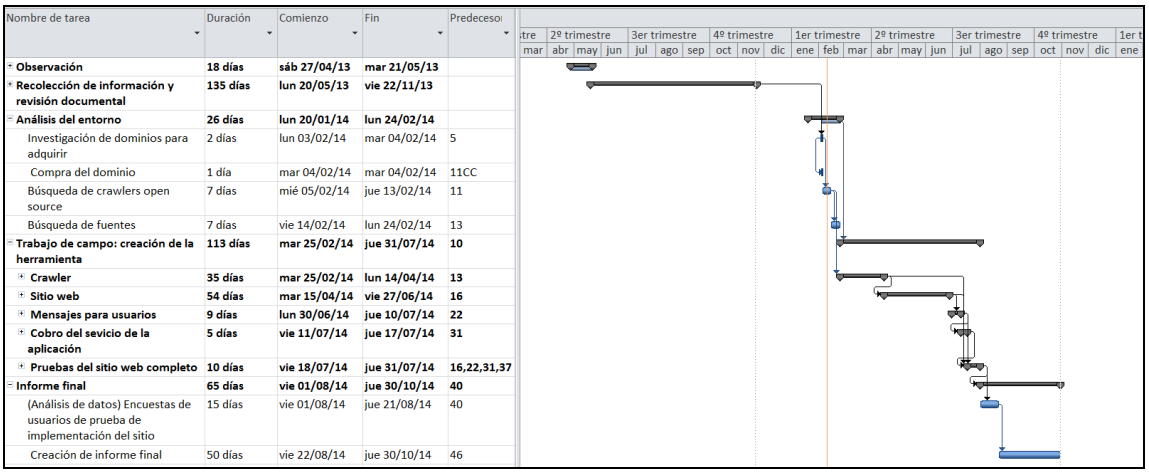
Nombre de tarea	Duración	Comienzo	Fin
Observación	18 días	sáb 27/04/13	mar 21/05/13
Recolección de información y revisión documental	135 días	lun 20/05/13	vie 22/11/13
Análisis del entorno	26 días	lun 20/01/14	lun 24/02/14
Investigación de dominios para adquirir	2 días	lun 03/02/14	mar 04/02/14
Compra del dominio	1 día	mar 04/02/14	mar 04/02/14
Búsqueda de *crawlers open source*	7 días	mié 05/02/14	jue 13/02/14
Búsqueda de fuentes	7 días	vie 14/02/14	lun 24/02/14
Trabajo de campo: creación de la herramienta	113 días	mar 25/02/14	jue 31/07/14
Crawler	35 días	mar 25/02/14	lun 14/04/14
Sitio web	54 días	mar 15/04/14	vie 27/06/14
Mensajes para usuarios	9 días	lun 30/06/14	jue 10/07/14
Cobro del servicio de la aplicación	5 días	vie 11/07/14	jue 17/07/14
Pruebas del sitio web completo	10 días	vie 18/07/14	jue 31/07/14
Informe final	65 días	vie 01/08/14	jue 30/10/14

Continuación de la tabla III.

(Análisis de datos) encuestas de usuarios de prueba de implementación del sitio	15 días	vie 01/08/14	jue 21/08/14
Creación de informe final	50 días	vie 22/08/14	jue 30/10/14

Fuente: elaboración propia.

Figura 5. **Diagrama de Gantt de tareas programadas**



Fuente: elaboración propia.

11. FACTIBILIDAD DEL ESTUDIO

11.1. Factibilidad operativa

- Recursos humanos
  - Asesor del trabajo de investigación
  - 2 programadores *senior*

- Acceso a la información: la información que se utilizará es información pública. Se tomará de los distintos sitios web y se hará la respectiva referencia a cada uno de ellos.

- Equipo e infraestructura
  - Dos computadoras con los componentes necesarios para crear la aplicación
  - Acceso a internet
  - Se utilizará un servidor en la nube para crear ambientes de desarrollo, pruebas y producción
  - Se adquirirá un dominio para publicar la aplicación

Analizando estos puntos, es factible realizar la investigación desde el punto de vista operativo.

11.2. Factibilidad técnica

- Se utilizarán herramientas *open source*, las cuales proveen mayor flexibilidad para crear aplicaciones.

- Las herramientas que se utilizarán son conocidas ya por los programadores, también podrán invertir tiempo para la investigación, en caso sea necesario.

- Se tiene planeado que el tiempo para el desarrollo de la aplicación será de 4 meses, invirtiendo 15 horas a la semana.

Técnicamente, es factible realizar la investigación porque se cuenta con las herramientas y tiempo necesarios.

11.3. Factibilidad económica

- El costo por hora trabajada de los programadores *senior* será de $10.00.

- El costo por hora de la consulta del asesor es de $20.00.

- Se adquirirá un dominio, el cual tendrá un costo de entre $10.00 y $11.00. Esta será la inversión inicial del proyecto.

- Las herramientas que se utilizarán son *open source*, por lo que no se pagará licenciamiento.

- El servidor que se utilizará será uno en la nube; este provee sus servicios gratis, por lo que tampoco habrá inversión económica.

Tabla IV. **Costos del proyecto**

Gastos	Costos subtotales		Costos totales	
Equipo informático	$2,500.00	Único	$2,500,00	Único
Dominio	$11.00	Anual	$22,00	2 años
Luz	$100.00	Mensual	$1 300,00	13 meses
Internet	$50.00	Mensual	$650,00	13 meses
Desarrolladores	$10.00	Hora	$2 500,00	250 horas
Asesor	$20.00	Hora	$1 600,00	80 horas
Otros gastos	$200.00	Anual	$400,00	2 años
Total de gastos			$8 972,00	

Fuente: elaboración propia.

Para el proyecto, no cobrarán los programadores, ya que el trabajo que realizarán será una inversión para obtener las ganancias de la aplicación cuando ya esté implementada. El equipo que se va a utilizar, como computadoras, electricidad e internet, son costos que los programadores absorberán como parte de esta inversión. El asesor también donará su tiempo.

El costo inicial será de $11,00; que es el pago por adquirir el dominio.

Se cuenta con la factibilidad económica, ya que se dispone del equipo para trabajar; se utilizarán componentes gratuitos y la inversión inicial que se necesita ya está contemplada para realizar la investigación.

12. BIBLIOGRAFÍA

1. ÁLVAREZ DÍAZ, Manuel. 2007. Arquitectura para crawling dirigido de información contenida en la web oculta. [en línea]. <http://ruc.udc.es/dspace/bitstream/2183/1000/1/AlvarezDiaz_Manuel_td_2007.pdf> [Consulta: 2 de agosto de 2013].

2. ÁLVAREZ, Manuel; et al. 2011. Web oculta del lado cliente: escala de Crawling, Departamento de Tecnologias de la Información y las Comunicaciones. Universidade da Coruña. [en línea] <http://www.dl.kuis.kyotou.ac.jp/~rafael.lopez/publications/JITEL_2011_2_CR.pdf> [Consulta: 3 de agosto de 2013].

3. ALVAREZ, Mauricio; et al. 2012. Crawler focalizado dinámico, fondef ContentCompass: identificación semántica y composición automática de material didáctico para dominios especializados. [en línea]. <http://lahuen.dcc.uchile.cl/~orojas/fondef_cc/Informe_ContentCompass_2012_Hito4.pdf> [Consulta: 04 de agosto de 2013].

4. ARRECIS, Leonel; TELÓN, Leonardo. 2009. Sistema para la administración de documentos de tesis. Universidad de San Carlos de Guatemala, Ingeniería en Ciencias y Sistemas, Guatemala: USAC, 2009. 107 p.

5. BAEZA-YATES, Ricardo; et al. 2005. Crawling a Country: Better Strategies than BreadthFirst for Web Page Ordering. Universidad de Chile. [en línea].<https://www.google.com.gt/url?sa=t&rct=j&q=&esrc=s&source=web&cd=5&cad=rja&ved=0CFUQFjAE&url=https://www.google.com.gt/url?sa=t&r

ct=j&q=&esrc=s&source=web&cd=5&cad=rja&ved=0CFUQFjAE&url=http%3A%2F%2Farachnode.net%2Fcfsfilesystemfile.ashx%2F__key%2FCommun > [Consulta: 20 de julio de 2013].

6. CLAY, Bruce. 2010. Robots exclusion protocol guide. [en línea]. <http://www.bruceclay.com/seo/robots-exclusion-guide.pdf> [Consulta: 10 de agosto de 2013].

7. CROW, Dan. 2007. The Robots Exclusion Protocol. Google Official Blog. [en línea] <http://googleblog.blogspot.com/2007/01/controlling-how-search-engines-access.html> [Consulta: 20 de julio de 2013].

8. DEL COSO SANTOS, Ana. 2009. Desarrollo de infraestructuras para el modelado de usuarios. Departamento de Informática, Escuela Politécnica Superior, Universidad Carlos III de Madrid. [en línea]. <http://earchivo.uc3m.es/bitstream/10016/8544/1/PFC_Ana_Coso_Santos.pdf> [Consulta: 20 de julio de 2013].

9. Dixxinet. Crawler, rastreador. [en línea] <http://www.dixxinet.com/crawler-rastreador> [Consulta: 10 de agosto de 2013].

10. Fundación Carolina. 2013. [en línea]. http://www.fundacioncarolina.es/es-ES/becas/posgrado/solicitarbeca/Paginas/solicitarbeca.aspx [Consulta: 20 de julio de 2013].

11. GOYZUETA, Juan Miguel. Becas en Guatemala. PlazaPublica. [en línea]. <http://www.plazapublica.com.gt/content/becas-en-guatemala. [Consulta: julio de 2013].

12. Guatefuturo. 2013. [en línea] <http://www.guatefuturo.org> [Consulta: 20 de julio de 2013].

13. HÍPOLA, Pedro y VARGAS-QUESADA, Benjamín. Agentes inteligentes: definición y tipología. Los agentes de información. El profesional de la información. [en línea] <http://www.elprofesionaldelainformacion.com/contenidos/1999/abril/agentes_inteligentes_definicion_y_tipologia_los_agentes_de_informacion.html> [Consulta: 10 de agosto de 2013].

14. Instituto para el Desarrollo de la Educación Superior en Guatemala, INDESGUA. 2013. [en línea] <http://www.indesgua.org/> [Consulta: 20 de julio de 2013].

15. JULIÁN, Vicente; REBOLLO, Miguel; CARRASCOSA, Carlos. 1999. Agentes de información. Profesores de la Facultad de Informática. Universidad Politécnica de Valencia. [en línea]. <http://www.upv.es/sma/teoria/aplicaciones/Aginformacion.pdf> [Consulta: 10 de agosto de 2013].

16. MADAAN, Rosy, et al. 2010. A framework for incremental hidden web crawler. Engg Journals Publications, 2010, International Journal on Computer Science and Engineering, Vol. 2, págs. 753-758. [en línea] <http://web.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=13b3a5f7-722-493a-b73c-fc553a0c37a8%40sessionmgr4&vid=1&hid=9> [Consulta: 4 de agosto de 2013].

17. MANNING, Christopher; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. Introduction to Information Retrieval, Chapter 20: Web crawling and

indexes. [en línea]. <http://nlp.stanford.edu/IR-book/pdf/20crawl.pdf> [Consulta: 20 de julio de 2013].

18. MCHELEY. What is a Web Site – Part I (Discovery). Graphtek Interactive. [en línea]. <http://blog.graphtek.com/2008/07/11/what-is-a-web-site-parti/#more-3> [Consulta: 31 de agosto de 2013].

19. MONTERO, José. 2009. Recuperación y Organización de la Información, Arañas Web (Crawlers). Universidad Carlos III de Madrid. [en línea]. <http://recuperacionorganizacioninformacionacces.net78.net/aranas_web_(crawlers)/aranas_web_(crawlers)_introduccion.html> [Consulta: 10 de agosto de 2013].

20. PEÑA, Rafael. 2005. WebCrawlers. Instituto de Matematicas y Facultad de Ciencias, UNAM. [en línea]. <http://www.matem.unam.mx/~rajsbaum/cursos/web/crawlers.pdf> [Consulta: 20 de julio de 2013].

21. PICÉN, Mónica. Mayo de 2013. Encuestas sobre Sitios Web de Becas. Guatemala, Guatemala. [en línea]. <http://www.surveymonkey.com/s/PZ2J6JK> [Consulta: mayo 2013].

22. Secretaría de Planificación y Programación de la Presidencia, SEGEPLAN. 2013. Sistema de Becas, pluralidad, transparencia y solidaridad. [en línea]. <http://becas.segeplan.gob.gt/becas/index.php> [Consulta: 20 de julio de 2013].

23. Universidad de San Carlos de Guatemala. [en línea]. <http://www.usac.edu.gt/secundario_dua.php?c=4957&f=coope> [Consulta: 06 de julio de 2013].

24. Universidad del Valle de Guatemala. 2013. Becas. [en línea]. <http://www.uvg.edu.gt/ai/becas.html> [Consulta: 06 de julio de 2013].

25. Universidad Francisco Marroquín. 2013. Becas y oportunidades. [en línea]. <http://epri.ufm.edu/becas-y-oportunidades/> [Consulta: 06 de julio de 2013].

26. Universidad Rafael Landívar. Becas 2013. [en línea]. <http://www.url.edu.gt/PortalURL/Contenido.aspx?o=4603&s=101&sm=c22> [Consulta: 06 de julio de 2013].

27. Wikipedia: Internet Bot. (S. f.). [en línea]. <http://en.wikipedia.org/wiki/Internet_bot> [Consulta: 10 de agosto de 2013].

28. Wikipedia: Web Crawler. (S. f.). [en línea]. <http://en.wikipedia.org/wiki/Web_crawler> [Consulta: 10 de agosto de 2013].

29. Wisegeek: What is a Web Crawler? (S. f.). [en línea]. <http://www.wisegeek.org/what-is-a-web-crawler.htm> [Consulta: 10 de agosto de 2013].

30. Wisegeek: What is Automatic Indexing? (S. f.). [en línea]. <http://www.wisegeek.com/what-is-automatic-indexing.htm> [Consulta: 10 de agosto de 2013].