Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría de Tecnologías de la Información y Comunicación

PROTOTIPO DE UN SISTEMA EXPERTO DE ORIENTACIÓN UNIVERSITARIA UTILIZANDO EL CONCEPTO DE RECUPERACIÓN DE INFORMACIÓN DE LA INTELIGENCIA ARTIFICIAL

Ing. William Samuel Guevara Orellana
Asesorado por la Inga. MSC. María Elizabeth Aldana Díaz

Guatemala, junio de 2020

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

PROTOTIPO DE UN SISTEMA EXPERTO DE ORIENTACIÓN UNIVERSITARIA UTILIZANDO EL CONCEPTO DE RECUPERACIÓN DE INFORMACIÓN DE LA INTELIGENCIA ARTIFICIAL

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

ING. WILLIAM SAMUEL GUEVARA ORELLANA
ASESORADO POR EL INGA. MSC. MARÍA ELIZABETH ALDANA DÍAZ

AL CONFERÍRSELE EL TÍTULO DE

MAESTRO EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

GUATEMALA, JUNIO DE 2020

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Christian Moisés de la Cruz Leal
VOCAL V	Br. Kevin Armando Cruz Lorente
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANA	Inga. Aurelia Anabela Cordova Estrada
DIRECTOR	Ing. Edgar Darío Álvarez Cotí
EXAMINADOR	Ing. Marlon Antonio Pérez Türk
EXAMINADOR	Ing. Edwin Estuardo Zapeta Gómez
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:

PROTOTIPO DE UN SISTEMA EXPERTO DE ORIENTACIÓN UNIVERSITARIA UTILIZANDO EL CONCEPTO DE RECUPERACIÓN DE INFORMACIÓN DE LA INTELIGENCIA ARTIFICIAL

Tema que me fuera asignado por la Dirección de la Escuela de Estudios de Postgrado, con fecha julio de 2011.

Ing. William Samuel Guevara Orellana

USAC
TRICENTENARIA

Decanato
Facultad de Ingeniería
24189101 - 24189102

DTG. 146.2020.

La Decana 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 Estudios de Postgrado, al Trabajo de Graduación titulado: **PROTOTIPO DE UN SISTEMA EXPERTO DE ORIENTACIÓN UNIVERSITARIA UTILIZANDO EL CONCEPTO DE RECUPERACIÓN DE INFORMACIÓN DE LA INTELIGENCIA ARTIFICIAL,** presentado por el Ingeniero **William Samuel Guevara Orellana,** estudiante de la **Maestría en Tecnologías de la Información y Comunicaciones** y después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

Inga. Anabela Cordova Estrada
Decana

Guatemala, julio de 2020.

AACE/asga

Escuela de Estudios de Postgrado, Edificio S-1
Teléfono: 2418-9142 / 24188000 ext. 1382
WhatsApp: 5746-9323
Email: informacion_eep@ing.usac.edu.gt

Guatemala, Junio de 2020

EEPFI-612-2020

En mi calidad de Director de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen y verificar la aprobación del Revisor y la aprobación del Área de Lingüística al Trabajo de Graduación titulado: **"PROTOTIPO DE UN SISTEMA EXPERTO DE ORIENTACIÓN UNIVERSITARIA UTILIZANDO EL CONCEPTO DE RECUPERACIÓN DE INFORMACIÓN DE LA INTELIGENCIA ARTIFICIAL"** presentado por el Ingeniero **William Samuel Guevara Orellana** quien se identifica con Carné **100020945**, correspondiente al programa de **Maestría en Artes en Tecnología de la Información y la Comunicación** ; apruebo y autorizo el mismo.

Atentamente,

"Id y Enseñad a Todos"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA · FACULTAD DE INGENIERÍA · DIRECTOR · ESCUELA DE POSTGRADO ·

Mtro. Ing. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería
Universidad de San Carlos de Guatemala

ESCUELA DE ESTUDIOS DE
POSTGRADO
FACULTAD DE INGENIERÍA

Escuela de Estudios de Postgrado, Edificio S-1
Teléfono: 2418-9142 / 24188000 ext. 1382
WhatsApp: 5746-9323
Email: informacion_eep@ing.usac.edu.gt

https://postgrado.ingenieria.usac.edu.gt

Guatemala, junio de 2020

EEPFI-611-2020

Como Coordinador de la Maestría en Artes en Tecnología de la Información y la Comunicación doy el aval correspondiente para la aprobación del Trabajo de Graduación titulado: **"PROTOTIPO DE UN SISTEMA EXPERTO DE ORIENTACIÓN UNIVERSITARIA UTILIZANDO EL CONCEPTO DE RECUPERACIÓN DE INFORMACIÓN DE LA INTELIGENCIA ARTIFICIAL"** presentado por el Ingeniero **William Samuel Guevara Orellana** quien se identifica con Carné **100020945.**

Atentamente,

"Id y Enseñad a Todos"

MARLON ANTONIO PEREZ TURK
INGENIERO EN CIENCIAS Y SISTEMAS
COLEGIADO No. 4492

Mtro. Ing. Marlon Antonio Pérez Turk
Coordinador de Maestría
Escuela de Estudios de Postgrado
Facultad de Ingeniería
Universidad de San Carlos de Guatemala

Guatemala, Junio de 2020

En mi calidad como Asesor del Ingeniero en ciencias y sistemas **William Samuel Guevara Orellana** quien se identifica con Carné **100020945** procedo a dar el aval correspondiente para la aprobación del Trabajo de Graduación titulado: **"PROTOTIPO DE UN SISTEMA EXPERTO DE ORIENTACIÓN UNIVERSITARIA UTILIZANDO EL CONCEPTO DE RECUPERACIÓN DE INFORMACIÓN DE LA INTELIGENCIA ARTIFICIAL"** quien se encuentra en el programa de **Maestría en Artes en Tecnologías de la Información y la Comunicación** en la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala.

Atentamente,

"Id y Enseñad a Todos"

María Elizabeth Aldana Díaz
Ingeniera en Ciencias y Sistemas
No. de Colegiado 9,188

Mtra. Inga. María Elizabeth Aldana Díaz
Asesora

ACTO QUE DEDICO A:

Dios	Porque todo lo que tengo y soy es por la misericordia de Él y todo se lo debo a Él, porque es merecedor de toda la honra y gloria, por los siglos de los siglos.
Mi esposa	Porque es un apoyo incondicional, mi inspiración; es y seguirá siendo el amor de mi vida, hasta que la muerte nos separe.
Mi hijo	Por ser la fuente de inspiración y el motivo por el cual todo mi esfuerzo se encaminó.
Mi mamá	Por su preocupación por mí, por siempre brindarme lo mejor y darme todo su esfuerzo y enseñanza.
Mi papá	Por ser mi ejemplo a seguir como hombre, como esposo y como hijo de Dios; por sus consejos y por su preocupación para que yo salga adelante.
Mis hermanas	Por siempre apoyarme en todas las decisiones y ser siempre una parte de mi vida.
Mis sobrinos	Para que este trabajo de graduación, le sea de motivación para alcanzar muchos éxitos en sus vidas, y honren así a sus padres. Esto mismo es, para futuros sobrinos o sobrinas.

Mis abuelos	Por siempre estar pendientes de mí y de mis triunfos.
Mi familia	Porque me instan a seguir adelante; por su aprecio hacia mí; por estimarme; porque los aprecio.
Mis amistades	Porque les aprecio, y para que este trabajo de graduación, les motive a seguir superándose en la vida.
Mis colegas	Para que puedan usar este material como referencia en sus ámbitos académicos; para recibir su retroalimentación acerca del sistema de información INFUNISA INVENTORY, que es objeto de este trabajo de graduación.

AGRADECIMIENTOS A:

Dios	Por brindarme de su sabiduría e inteligencia; por ayudarme a culminar otro éxito en mi vida; por su fidelidad y misericordia hacia mí; por resolver mis problemas y aflicciones.
Mi esposa	Priscila María López García de Guevara por su apoyo incondicional y su manera de amarme en cualquier situación.
Mi hijo	Samuel Isaac Guevara López, porque su sonrisa y su dulzura fue lo que me dio una motivación y las fuerzas necesarias para este éxito.
Mis padres	William Guevara y María Orellana, porque sus consejos y sus correcciones me llevaron a ser el hombre que soy.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES ……..……..……..…V
LISTA DE SÍMBOLOS…...……..VII
GLOSARIO……...……..IX
RESUMEN ……...……..……..……....XI
PLANTEAMIENTO DEL PROBLEMA Y FORMULACIÓN DE
PREGUNTAS ORIENTADORAS ……..……..XIII
OBJETIVOS …...……..……..……XV
RESUMEN DE MARCO METODOLÓGICO …...……..……..…XVII
INTRODUCCIÓN ……...……..……...…XXIII

1. ANTECEDENTES ……..……1

2. JUSTIFICACIÓN ...……...….....5

3. ALCANCES …...……..……..…….7
   3.1 Perspectiva investigativa ……...……..…..7
   3.2 Perspectiva técnica ……..……...……...7
   3.3 Resultados esperados ……...……..…..8

4. MARCO TEÓRICO ……..…...……9
   4.1 Sistema de búsqueda de respuesta …...……..…..……..9
      4.1.1 Características importantes en un sistema de
      búsqueda de respuesta ……..9
   4.2 Sistemas de recuperación de información ……..……..11
   4.3 Arquitectura comúnmente utilizada en sistemas QA ……...15

4.3.1 Análisis de la pregunta ……..……..15
4.3.2 Lingüística de la pregunta ……..……17
4.3.3 Proceso de búsqueda ……..…..……....18
4.3.4 Recuperación y selección de la información ……..……..20
4.3.5 Extracción y generación de respuestas ……..……..22
4.3.6 Técnicas de procesamiento del lenguaje natural ……..…22

5. PRESENTACIÓN DE RESULTADOS ……...25
5.1 Construcción del prototipo……...…25
5.2 Procedimiento para actualizar fuente de información de un sistema experto ……..……....27
5.3 Arquitectura de un sistema experto ……..….30
5.3.1 Análisis de la pregunta ……..……..33
5.3.2 Recuperación de la información ……..36
5.3.2.1 Indexación de la colección …….……..36
5.3.2.2 Recuperación de documentos ……..……..38
5.3.2.3 Selección de pasajes relevantes ……..……...39
5.3.3 Extracción de la respuesta ……..……39

6. ANÁLISIS Y DISCUSIÓN DE RESULTADOS ……..……..41
6.1 Robustez ……..……41
6.1.1 Actualización de la fuente de información …….……44
6.2 Fiabilidad ……..……45
6.3 Rapidez ……..……..…47
6.4 Impactos económicos, tecnológicos y sociales ……..……..…48
6.4.1 Impacto económico ……..……48
6.4.2 Impacto tecnológico ……..……..…49
6.4.3 Impacto social …….……..……49

CONCLUSIONES ……51
RECOMENDACIONES ……,,,……53
REFERENCIAS BIBLIOGRÁFICAS ……..…….....55

ÍNDICE DE ILUSTRACIONES

FIGURAS

1. Proceso de recuperación de información……..…….....12
2. Operaciones para la recuperación de documentos……..……...12
3. Sistema de búsqueda genérico……..…….....15
4. Taxonomía de preguntas…….....16
5. Sistema de búsqueda de respuestas genérico……..……..19
6. El proceso documental……....……..……..….21
7. Prototipo del agente virtual...…….…..27
8. Proceso de actualización de la fuente de información……..…..……28
9. Componentes de la arquitectura propuesta……...……...31
10. Arquitectura general del sistema QA……....…...…….……..…31
11. Arquitectura de almacenamiento para flexibilidad de aplicaciones……..33
12. Indexación de archivos……..……..…….……..….37
13. Indexación por swish-e……..……...…...……..... 38

TABLAS

I. Definición de variables, subvariables e indicadores……..…..…XIX
II. Pasajes descritos en los documentos……...……..……40
III. Criterios de evaluación……..…..……..42
IV. Análisis de riesgos de prototipo……..…...…43
V. Análisis de tiempos y tratamientos de solución de riesgos……..…44

LISTA DE SÍMBOLOS

Símbolo	Significado
%	Porcentajes
Gb.	Gigabytes

GLOSARIO

Arquitectura de software	Organización de un sistema y cómo debe de diseñarse en forma global, es la primera etapa en el proceso de diseño de software y adicionalmente describe la forma en que se organiza el sistema como un conjunto de componentes en comunicación.
BR	Búsqueda de respuestas.
FAQ	*Frequently asked questions.*
ICMI	International Customer Management Institute.
Inteligencia artificial	Es una inteligencia diseñada para simular el razonamiento humano en un ámbito computacional o lenguaje de máquinas, de modo que estas puedan aprender, tomar decisiones y, finalmente, resolver un problema.
LSA	Análisis semántico de texto.
Prototipo	Ejemplar que se diseña de un software y que sirve de modelo para la construcción, pruebas y validación de requerimientos funcionales y no funcionales.
QA	*Question answering.*

QALL-ME	*Question answering learning technologies in a multilingual and multimodal environment.*
P-R	Pregunta-respuesta.
PLN	Procesamiento del lenguaje natural con TTS, *text to speech.*
RI	Recuperación de información.
RSV	*Retrieval status value.*
Sistema experto	Sistema basado en el concepto de inteligencia artificial diseñado para resolver problemas o tomar decisiones en un dominio en particular de un modo similar al del razonamiento humano.
SRI	Sistema de recuperación de información

RESUMEN

Hoy, las organizaciones e instituciones públicas requieren sistemas inteligentes que ayuden a cumplir los objetivos estratégicos o reducir la complejidad de los procesos de negocio. Dejan esa responsabilidad al proceso automatizado, que puede incluir un sistema experto.

Los sistemas expertos y basados en la inteligencia artificial no son más que procesos automatizados que cobran vida de forma inteligente de acuerdo a una entrada en el mismo. De acuerdo a esa entrada, estos generan información al usuario de forma lógica, la cual servirá para incluirla en otro proceso o bien al finalizar el mismo. Adicionalmente, los sistemas expertos proporcionan un mecanismo mucho más eficaz y eficiente de hacer las cosas, ya que pueden funcionar las 24 horas del día sin descanso alguno y sin necesidad de la intervención humana. Solamente se suspende el servicio si hay algún mantenimiento, actualización, cambio, corrección, etc. del sistema.

El hecho de que un sistema experto no necesite intervención humana hace pensar que existe un reto aún más grande, que es modelar la mente humana en dichos sistemas como entes pensantes y analíticos. Esto conlleva crear un sistema experto que resolverá un prototipo de sistema inteligente, capaz de ser eficaz, eficiente y confiable al usuario.

La confiabilidad de los sistemas expertos es cada vez más aceptada dado los algoritmos que se implementan en dichos sistemas. Cada vez más existen sistemas expertos que analizan y sirven la información de manera acertada a los usuarios.

En el trabajo de graduación se analizó y diseñó un prototipo que pueda generar respuestas acertadas y coherentes a preguntas de usuarios. Este diseño contempla un mecanismo de entender la pregunta, analizarla, buscar las posibles respuestas y resolverlas para luego dar una información confiable al usuario.

El prototipo del sistema experto se diseñó bajo una arquitectura basada en la inteligencia artificial, utilizando el concepto de recuperación de la información. Los procesos de dicha arquitectura nos ofrecen un sistema final de preguntas y respuestas que podrá ser utilizado sobre cualquier plataforma. Esto significa que la arquitectura o el diseño del prototipo contempla el conjunto de actividades y diseño de procesos para implementarlo en cualquier sistema de preguntas y respuestas; sin embargo, este trabajo se hará bajo la premisa de utilizarlo en el proceso de orientación estudiantil de cualquier universidad.

PLANTEAMIENTO DEL PROBLEMA Y FORMULACIÓN DE PREGUNTAS ORIENTADORAS

En el ámbito de desarrollo de sistemas de software se ha notado un crecimiento en el desarrollo de sistemas expertos o inteligentes, dado que cada vez más estos permiten realizar operaciones que comúnmente realizaría un humano, solamente que los sistemas expertos son “incansables” en tiempo, esfuerzo y dinero. Esto se logra entender al observar el trabajo común de una persona, ya que un humano tiene límites de atención, tiempo y esfuerzo en los alcances definidos en una empresa.

Según los experimentos realizados con sistemas expertos de preguntas y respuestas, se puede observar la incapacidad de dichos sistemas en resolver y responder preguntas sencillas en un tiempo considerable de un tema en específico y responder preguntas complejas dando las respuestas exactas. En muchas ocasiones existen problemas de redacción dando lugar a que las preguntas no sean entendidas por el sistema experto. La mala redacción influye mucho en la búsqueda de las respuestas. La falta de información o información desactualizada en dicho sistema experto ocasiona muchos desaciertos y todo esto ocasiona que los sistemas mencionados sean inexactos y tardados en responder preguntas específicas.

Por tanto, es necesario determinar y centrarnos en una pregunta principal: ¿es posible desarrollar un prototipo de un sistema experto basado en una arquitectura centrada en la recuperación de información, utilizando la inteligencia artificial con el fin que pueda responder adecuadamente a preguntas frecuentes?

Algunas otras preguntas auxiliares involucradas para realizar este tipo de sistemas son:

- ¿Cuál sería el procedimiento adecuado para actualizar la fuente de información de un sistema experto?

- ¿Cuál es el diseño y arquitectura de búsqueda de respuestas en un sistema experto?

- ¿Cuál sería el proceso interno óptimo, eficaz y eficiente que conlleva a un sistema experto buscar una respuesta correcta?

- ¿Es posible diseñar una arquitectura centrada en la recuperación de información para resolver preguntas simples y complejas?

OBJETIVOS

Objetivo general

Desarrollar un prototipo de un sistema experto basado en una arquitectura centrada en la recuperación de información utilizando la inteligencia artificial.

Objetivos específicos

- Determinar el procedimiento adecuado para actualizar la fuente de información de un sistema experto.
- Analizar la arquitectura que permita la búsqueda eficiente de respuestas en un sistema experto.
- Determinar el proceso interno óptimo, eficaz y eficiente que conlleve a un sistema experto en buscar una respuesta correcta.
- Diseñar una arquitectura centrada en la recuperación de información para resolver preguntas simples y complejas.

RESUMEN DE MARCO METODOLÓGICO

A continuación se presenta el marco metodológico:

- Tipo de estudio

  El tipo de estudio de la investigación es cualitativo y cuantitativo.

  - Desde el punto de vista cuantitativo, la investigación se enfocó en la rapidez de la búsqueda de resultados, tomando como parámetro de medición el tiempo de respuesta por evento y transacciones por segundo procesadas.

  - Desde el punto de vista cualitativo, la investigación mostrará los resultados obtenidos en una búsqueda con relación a la arquitectura propuesta y su prototipo asociado.

- Diseño

El trabajo se considera experimental, debido a que el prototipo, basado en una arquitectura de recuperación de información, será puesto a prueba con un conjunto de datos de entrada o búsqueda simples y complejas. De acuerdo a esto se analizará si la salida o respuesta tiene cierto grado de exactitud. Adicionalmente, este conjunto de entradas o búsquedas pondrá a prueba la eficiencia del prototipo en cuanto a su tiempo de respuesta.

- Alcances

El estudio tiene un alcance descriptivo, debido a que no se descubre la causa que origina la inconsistencia de respuestas en base a preguntas realizadas en un sistema experto. También se toma en cuenta que pueden existir variables que afecten las respuestas del sistema experto como el lenguaje, la forma de redacción, la ortografía, semántica, etc. Sin embargo, el estudio se enfoca en describir la forma en que un sistema experto puede recuperar información de acuerdo a una arquitectura y prototipo propuesto, tomando en cuenta las variables que afectan anteriormente mencionadas.

El prototipo se realizó basado en el concepto de recuperación de la información, con reglas definidas que fueron utilizadas para encontrar la información correcta; sin embargo, las reglas pueden ser evaluadas, analizadas, diseñadas y optimizadas para crear un conjunto de posibles respuestas alternativas.

- Definición de variables, subvariables e indicadores

Tabla I. **Definición de variables, subvariables e indicadores**

Variable	Definición	Indicadores
Robustez	Se refiere a la forma de recuperación del prototipo de fallas en el sistema.	Tiempo de reinicio después de falla. Porcentaje de eventos que causan falla.
Fiabilidad	Representa la salida correcta de acuerdo a la entrada o pregunta.	Tasa de ocurrencia de falla
Rapidez	Se refiere al tiempo en que el prototipo tarda en responder a una serie de entradas o transacciones.	Tiempo de respuesta usuario/evento Transacciones/segundo procesadas

Fuente: elaboración propia

- Técnicas de recolección de información

Se recolectará la información por medio de la observación y consulta de datos de fuentes de investigación que puedan servir para la solución de la arquitectura y prototipo.

- Fases del estudio

    - Fase de revisión bibliográfica. Recolectar y analizar la información de diversas fuentes como tesis de maestría, libros o publicaciones científicas para determinar el procedimiento adecuado para

actualizar la fuente de información de un sistema experto. Estas fuentes fueron analizadas para definir la arquitectura del sistema de información. La información que se buscó estaba asociada a la arquitectura QALL-ME, sistemas de búsqueda y sistemas de recuperación de información.

- Fase de análisis. En esta fase se analizó la información relevante recolectada en la primera fase, con el fin de encontrar una solución de diseño de una arquitectura. Se realizó la búsqueda para determinar la arquitectura adecuada en función de los componentes que posee el sistema. Se realizó un proceso de clasificación de algoritmos, tomando en cuenta que estos debieran analizar un planteamiento expuesto.

  Se analizaron los procedimientos existentes para la recuperación de la información de un sistema experto. Se analizaron los algoritmos y procedimientos que permitieran la extracción de la información basados en un diseño QALL-ME. Se llevó a cabo un proceso de análisis de cómo los sistemas expertos se basan en el contexto de la pregunta para realizar una búsqueda de información.

- Fase de diseño de la arquitectura. Se diseñó un modelo de representación de la pregunta, tomando en cuenta la información relativa de la pregunta, respuesta y contexto. Se diseñó un modelo que soporte un modo de pregunta-respuesta complejo y dialogado. Finalmente se tomó la taxonomía de preguntas para el diseño de la solución.

- Fase de diseño del prototipo. Se analizó el prototipo diseñado, sus ventajas y desventajas.

INTRODUCCIÓN

El trabajo de graduación está enfocado en el prototipo de un sistema experto de preguntas y respuestas de orientación universitaria, por medio de una arquitectura de software basada en la recuperación de la información de la inteligencia artificial.

La investigación se centra en utilizar un diseño de arquitectura de pregunta y respuesta, llevando a cabo un análisis de la pregunta. Posteriormente se recupera la información relevante para finalmente extraer la respuesta correcta y presentarla al usuario. Todo este proceso será dirigido por el contexto en cual se hace la pregunta.

En el capítulo 1 se analizan y describen los antecedentes y se presentan los datos de investigaciones anteriores que han sido utilizadas para desarrollo de sistemas basadas en la inteligencia artificial.

En el capítulo 2 se aborda la justificación del proyecto, que se desarrolla con base en la pregunta principal formulada: ¿Es posible desarrollar un prototipo de un sistema experto basado en una arquitectura centrada en la recuperación de información utilizando la inteligencia artificial con el fin que pueda responder adecuadamente a preguntas frecuentes?. Describe la importancia del proyecto sobre la eficacia y eficiencia de un sistema de preguntas y respuestas basado en la recuperación de información de la inteligencia artificial.

El capítulo 3 se refiere al alcance del proyecto, donde se clarifican los requisitos del sistema, tanto funcionales como no funcionales. También se

aclaran las restricciones del proyecto que sean relevantes para usuarios y clientes del sistema.

El capítulo 4 presenta toda la información acerca de sistemas expertos basados en la recuperación de información. Se define qué es un sistema experto en términos de sus componentes, así como un sistema experto basado en la recuperación de la información.

En el capítulo 5 se desarrolla y muestra el resultado del diseño de la arquitectura de un sistema de pregunta-respuesta desde una arquitectura básica; el proceso de análisis de la pregunta dirigido por el contexto en la cual se realiza, para luego enfocarse en los procesos de recuperación de información. Concluye con el proceso de extracción de la respuesta relevante al usuario. También se aborda la arquitectura QALL-ME basada en servicios web como un modelo para desarrollar nuevas aproximaciones para incorporar mayor usabilidad a los sistemas de pregunta-respuesta. También ayuda a definir una infraestructura compartida para las búsquedas de respuestas (QA) en dominio abierto multilingüe y multimodal, apoyándose en tecnologías y algoritmos de aprendizaje automático. Como un punto final, se propone el diseño del prototipo de solución basado en las investigaciones y documentos previos a este capítulo, con la presentación de un análisis de ventajas y desventajas del mismo.

El capítulo 6 muestra los puntos importantes del desarrollo del proyecto. Se expone el análisis de los resultados que ayudan a entender el conocimiento obtenido del proyecto.

1. ANTECEDENTES

Actualmente existen varias compañías que utilizan sistemas expertos de preguntas frecuentes para dar asesoría y ayudar a los usuarios que ingresan a sus sitios. Algunas de ellas se mencionan a continuación:

La Universidad de Granada utiliza un agente virtual llamado Elvira que da la asesoría y responde preguntas frecuentes acerca de la Universidad. Una de sus principales atribuciones y funciones es dar de la mejor forma posible la información necesaria a las personas que ingresan al sitio y tienen dudas acerca de cualquier tema de la Universidad de Granada. Una de sus fortalezas es que responde a la mayor parte de preguntas de forma muy atinada y sobria; sin embargo, como cualquier sistema experto posee debilidades y una de ellas es que el agente virtual no está 100 % actualizado con la información del periodo actual de la Universidad. Por tal motivo la información solicitada no coincide con algunas áreas de información de la universidad actualmente (Universidad Granada, 2011).

Natalia es un agente virtual que trabaja como Asesora en Capacitación en Proaxion; consultora líder en Contact Centers y Global Partner de ICMI, International Customer Management Institute, en Latinoamérica. Ella puede ayudar a conocer en detalle los cursos que la compañía ofrece, su misión y más. Además este agente virtual tiene estudios en Administración de Empresas y asiste de manera amena y cordial a los clientes con sus necesidades de entrenamiento. Puede además conversar sobre temas de interés general. Natalia utiliza el "cerebro artificial" más avanzado desarrollado por BotGenes con más de 400.000 reglas de decisión, lo cual le permite manejar miles de regionalismos,

errores de escritura, ortográficos y gramaticales, y contextualizar respuestas. Ella integra tecnologías de PLN, procesamiento del lenguaje natural con TTS, text to speech, texto a voz; ofrece además de chat, audio streaming y un avatar animado desde los servidores de BotGenes. Una de las debilidades de este agente virtual es que no está actualizado con la información reciente de los cursos que ofrecen y de los precios. Esto se puede observar al comparar la información enviada por medio del personal administrativo y la que proporciona el agente (Proaxion, 2012).

Micro Lending Argentina (MILA) es un agente virtual con información del mercado de créditos y seguros. Sofía es un agente virtual de MILA que brinda atención al cliente sin intervención humana a través de un chat integrado. Provee la posibilidad de extender el horario de atención al cliente las 24 horas del día durante todo el año, y liberar recursos para tareas de mayor valor agregado. Lo interesante de este agente virtual es que integra un sistema de control de calidad donde realiza una encuesta de su servicio y pregunta cuál hubiera sido la respuesta correcta, con el fin de aprender. Posee la debilidad de no tener la información actualizada para informarle al usuario acerca de un tema y sus respuestas no son precisas (Micro Lending Argentina, 2014).

Distinta, que es el centro de Esperanza, un agente virtual desarrollado por BotGenes, trabaja como asistente virtual en Una Mirada Distinta; Centro de la Municipalidad de San Isidro para la inclusión de personas con discapacidad (Municipalidad de San Isidro, 2013).

Ella tiene estudios en comunicación social y da información sobre el centro a los interesados. Utiliza el "cerebro artificial" en idioma español más avanzado desarrollado, con más de 400.000 reglas de decisión, lo cual le permite manejar

miles de regionalismos, errores de tipeo, ortográficos y gramaticales, y contextualizar respuestas.

Ella integra tecnologías de PLN, procesamiento del lenguaje natural, con TTS, texto a voz. Ofrece, además de chat, audio streaming y un avatar. Este agente virtual posee la debilidad de manejar tiempos de respuestas muy largas al momento de ingresar la pregunta. Las respuestas no son precisas y correctas; confunde sinónimos de palabras y en muchas ocasiones no resuelve la pregunta de acuerdo al negocio.

Ezequiel es el agente virtual del Municipio de Tigre, Buenos Aires, Argentina. Basado en inteligencia artificial, Ezequiel posibilita la atención al ciudadano sin intervención humana las 24 horas durante todos los días del año. Provee información y atiende las cuestiones de los vecinos acerca de trámites, reclamos, actividades, teléfonos útiles, cursos y talleres dentro del municipio. Es uno de los más activos de Buenos Aires. Utiliza el cerebro artificial desarrollado por BotGenes que cuenta con más de 500.000 reglas de decisión, incluyendo tecnologías de procesamiento del lenguaje natural (PLN), con TTS (texto a voz) y un avatar 3D animado (Municipio de Tigre, 2014).

Una de las debilidades de este agente virtual es su velocidad de respuesta, ya que para preguntas sencillas su respuesta es lenta y mucho más para preguntas complejas. Tampoco posee información puntual acerca de información específica de usuarios, esto se debe a que el sistema no está actualizado con la información del usuario.

En los últimos años se ha producido un notable crecimiento en el área de la inteligencia artificial, a tal punto que muchas compañías están dejando de utilizar su recurso humano y lo sustituyen por máquinas expertas o sistemas

expertos de la información. Acerca de sistemas expertos de preguntas y respuestas, pueden tener cierto grado de desconfianza ya que, comparado con la mente humana, tienen más probabilidades de no coincidir en la respuesta que el usuario requiera. Esto se debe al diseño de estos sistemas, ya que se basan en ciertas reglas para contestar acertadamente y la utilización del lenguaje de cada país. El análisis semántico de texto LSA (Knowledge Analysis Technologies, 1998, IEA), el procesamiento natural del lenguaje NPL (Educational Testing Service, 1999, E-rater) para buscar algún tipo de respuesta, ya que si se escribe una pregunta donde se utilice la palabra “día”, el sistema inteligente probablemente asociará dicha palabra a una pregunta que indique el día donde ocurrió u ocurrirá cierto evento, y así sucesivamente. El sistema inteligente analiza cada palabra para asociarla y brindar la respuesta que más se acerque a la pregunta realizada por el usuario.

2. JUSTIFICACIÓN

La línea de investigación se centraliza en sistemas para aumentar la experiencia del usuario a través de la tecnología. Utiliza el concepto de sistemas de preguntas y respuestas basado en recuperación de información de la inteligencia artificial, con el fin de crear un sistema experto especializado para el uso en orientación universitaria. El sistema experto tiene como objetivo obtener la respuesta de acuerdo con un algoritmo de recuperación de información.

Es importante tomar en cuenta que se debe procesar los elementos clave que permiten hacer la búsqueda, como los índices, palabras clave y los fenómenos que se pueden dar en el proceso, que causen ruido.

Para recuperar la información se hará uso de la técnica de lógica difusa, la cual permite hacer búsquedas tomando en cuenta frases normales. La máquina, al realizar la búsqueda, elimina signos de puntuación, artículos, conjunciones, plurales, tiempos verbales, palabras comunes (que suelen aparecer en todos los documentos), y deja aquellas palabras que el sistema considera relevantes.

La fase de diseño incluye una solución tecnológicamente viable para desarrollar el sistema. El diseño de la arquitectura puede dar como resultado un sistema confiable y eficiente en el manejo de preguntas basadas en técnicas de la recuperación de información

Las ventajas de utilizar sistemas expertos son:

- Disponibilidad alta de información.
- Agilizar y optimizar los procesos de negocio en el área.
- Generar rentabilidad al reducir costos administrativos.

3. ALCANCES

3.1 Perspectiva investigativa

Este trabajo define una investigación para seleccionar la arquitectura que permita la búsqueda eficiente de respuesta en un sistema experto. Se definieron varios diseños, entre los que podemos mencionar la arquitectura basada en conocimiento y motores de inferencia, basados en reglas, en metaconocimiento o meta-reglas y en la recuperación de la información por medio de documentos. Esta última se definió para ser la solución propuesta para cubrir la necesidad de una búsqueda eficiente.

Los componentes definidos en la investigación para diseñar una arquitectura enfocada a la recuperación de la información para resolver preguntas simples y complejas son el componente de análisis de la pregunta y extracción de la respuesta en lenguaje natural; el componente para reconocer la búsqueda y la forma de encontrarla en los documentos y el componente RI de los documentos relevantes que recupera la información en base a la respuesta que desea el usuario.

3.2 Perspectiva técnica

Este trabajo define el diseño de un prototipo de un sistema experto basado en la recuperación de la información tomando en cuenta las siguientes características:

- Procedimiento para actualizar la fuente de información de un sistema experto.

- Diseño de la arquitectura basada en servicios web para extracción de la información por medio de un sistema de búsqueda eficiente de respuesta en la base del sistema experto.

- Planificador central óptimo, eficaz y eficiente que busca la respuesta correcta.

- Diseño de un sistema de recuperación de información que resuelve e interpreta preguntas simples y complejas.

3.3 Resultados esperados

En esta investigación se define un procedimiento adecuado para actualizar la fuente de información de un sistema experto.

- Diseño de la arquitectura que permite la búsqueda eficiente de respuestas en un sistema experto.

- Proceso interno óptimo, eficaz y eficiente que conlleve a un sistema experto en buscar una respuesta correcta.

- Diseño de la arquitectura centrada en la recuperación de información para resolver preguntas simples y complejas

4. MARCO TEÓRICO

4.1 Sistema de búsqueda de respuesta

Los sistemas pregunta-respuesta (sistemas P-R) son también denominados sistemas de búsqueda de respuestas (sistemas BR) y son conocidos también por su término en inglés como Question-Answering Systems (QA systems), (Martínez, P., 2007).

Un sistema de búsqueda de respuesta tiene como objetivo devolver una respuesta a una pregunta o consulta planteada por un usuario en un lenguaje natural. De esta forma, los sistemas modernos de QA, debido a la complejidad de esta tarea, utilizan técnicas de procesamiento del lenguaje natural para resolver y brindar al usuario una respuesta adecuada (Vicente-Díez, M. T., 2009).

4.1.1 Características importantes en un sistema de búsqueda de respuesta

Un grupo de 19 investigadores logró un aporte importante en el contexto de sistemas QA con un programa que dota de características importantes y útiles para el usuario, aumenta las expectativas en función de la respuestas que se obtengan y los factores que giran en torno al proceso interno (Martínez, P., 2007), tales como:

- Respuesta en tiempo razonable: el resultado a una pregunta debe de darse en un tiempo real, a pesar de que el sistema tenga varias peticiones de preguntas y varios usuarios accedan simultáneamente.

- Precisión: esta característica es muy importante, ya que se considera que una respuesta errónea es peor que no responder. Los sistemas deben enfocarse en evaluar la corrección de las respuestas proporcionadas, e incluir métodos para determinar que la respuesta no está disponible. Para ser más preciso, un sistema QA debería de incorporar métodos que imiten la inferencia del sentido común.

- Usabilidad: el conocimiento de un sistema QA debe ser estructurado a la medida para que satisfaga las necesidades específicas de los usuarios. Los conocimientos específicos de un dominio deben ser incorporadas en un sistema QA, y ser capaz de permitir al usuario describir el contexto de la pregunta.

- Completitud: se refiere a que la respuesta ante una pregunta de un usuario debe ser completa. Generar una respuesta completa depende de muchos factores, ya que las personas pueden expresarse de diferente forma o por la dispersión de la información. Además, se debe relacionar el conocimiento del mundo con el conocimiento específico del dominio, y encontrar la forma de razonar con ambos. Esto quiere decir que un sistema QA debe tener la capacidad de razonamiento y usar bases de conocimiento de alto rendimiento. En muchas ocasiones se tendrá que encontrar la relación con otras preguntas, y su razonamiento debe tomar en cuenta el contexto definido por el usuario y el contexto del perfil del usuario. Esto permitirá obtener una retroalimentación desde el usuario, muy útil en el proceso de la búsqueda de la respuesta.

- Relevancia: la respuesta debe ser relevante en un contexto específico; esto quiere decir que en muchas ocasiones será necesario realizar una

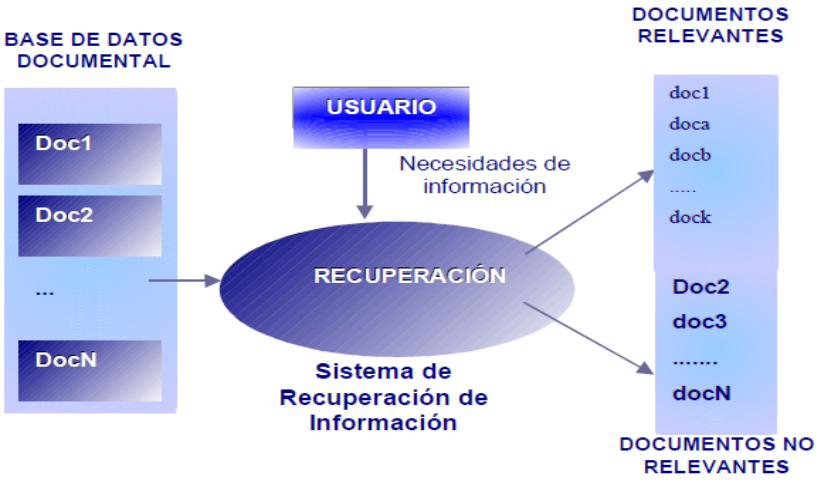
búsqueda interactiva, en la que una serie de preguntas ayudan a definir y clarificar la información necesaria. El crecimiento de un sistema de QA debe estar centrado en el usuario. Los humanos son los jueces de la utilidad y relevancia del sistema QA y de la facilidad de uso.

4.2 Sistemas de recuperación de información

La recuperación de información (RI) se puede concretar como el problema de la elección de información desde un dispositivo de almacenamiento en respuesta a consultas realizadas por el usuario.

Los sistemas de recuperación de información (SRI) tienen una base de datos compuesta por documentos y procesan las consultas de los usuarios, a quienes permite obtener la información relevante en un tiempo apropiado. Estos fueron inicialmente desarrollados a partir de los años 40. Un SRI puede recuperar información, anteriormente almacenada, por medio de un conjunto de consultas a la base de datos. Estas consultas son oraciones formales de expresión de necesidades de información y regularmente pueden ser expresadas en un lenguaje de consulta o natural (Rodríguez M., 2005). El proceso de recuperación de información puede verse en la figura 1.

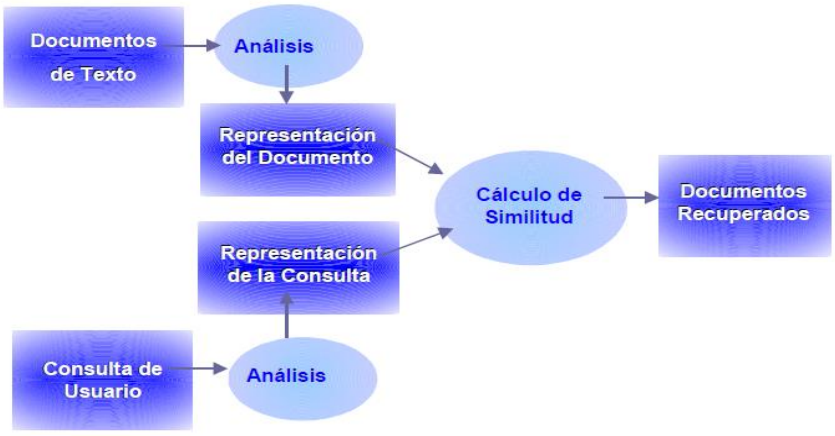
Figura 1. **Proceso de recuperación de información**



Fuente: Rodríguez (2005). *Modelos de recuperación de información.*

Un SRI puede procesar un conjunto de operaciones sobre los documentos almacenados, ya sea para insertar nuevos documentos, modificarlos o eliminarlos como se muestra en la figura 2. Los SRI cuentan con métodos para localizar uno o varios documentos para mostrarle al usuario.

Figura 2. **Operaciones para la recuperación de documentos**



Fuente: Rodríguez (2005). *Modelos de recuperación de información.*

- Extracción de información: la forma de extracción de la respuesta conlleva la localización de la respuesta en las partes relevantes. El problema y la complejidad de esta actividad de extracción depende el tipo de respuesta que el usuario espera. Por ejemplo, la respuesta a una pregunta ¿Quién…? puede ser tan simple como encontrar una entidad de tipo persona en un fragmento relevante. El punto importante en un SRI es encontrar el mecanismo en que a partir de la pregunta pueda inferirse información sobre la respuesta. Esta inferencia pregunta-respuesta puede realizarse de las siguientes maneras:

    - Por medio de modelos de traducción automática del lenguaje de las preguntas al lenguaje de las respuestas.

    - Por medio de técnicas de aprendizaje automático a partir de pares pregunta/respuesta, obtenido desde ficheros de FAQ donde aprende los términos, frases y patrones que se espera aparezcan en los documentos a recuperar.

    - Mediante patrones léxico-semánticos, que utilizan 361 patrones léxico-semánticos para localizar la respuesta.

- Búsqueda de respuesta: los sistemas de búsqueda de respuestas comúnmente combinan métodos basados en la recuperación de la información y extracción de la información. Aprovechan la rapidez de algoritmos resultantes y la redundancia de la información, y algunos utilizan métodos basados en información lingüística.

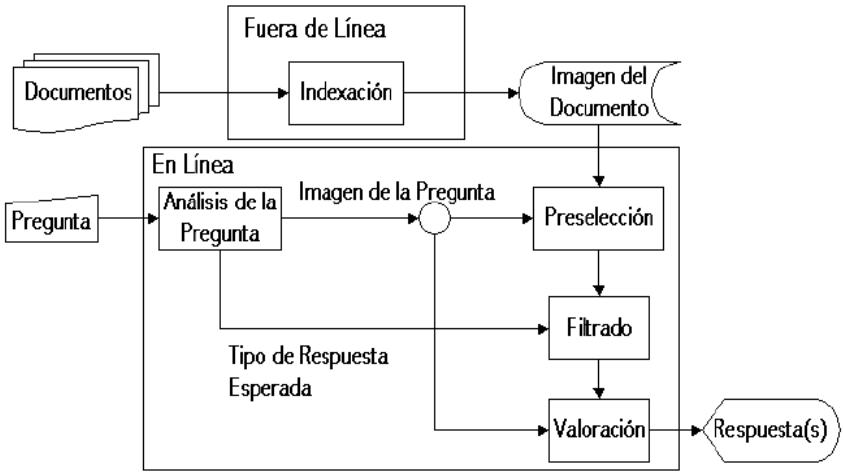
Los sistemas de búsqueda de información son capaces de procesar documentos de texto automáticamente y, de esa forma, encontrar la respuesta a

preguntas en tiempo real. Este tipo de sistemas es cada vez más utilizado y su desarrollo es motivado por la disponibilidad de la información, crecimiento en la potencia de los ordenadores y la disponibilidad de recursos y herramientas para desarrollarlos. La cantidad de textos disponibles ha alcanzado enormes proporciones y sigue creciendo años tras año. Un ejemplo es la cantidad de documentos indexados por Google. Las empresas hoy poseen documentos almacenados digitalmente más que los almacenados físicamente. Los sistemas actuales de RI son utilizados para encontrar la respuesta exacta en vez de documentos relevantes, en especial cuando el usuario desea encontrarlo en un corto tiempo y en una posible gran lista de documentos (Molla, D. 2003).

Los sistemas de búsqueda de información tienen que ser capaces de encontrar la respuesta del usuario en una cantidad de texto relativamente grande. Para tal efecto, estos sistemas aprovechan la redundancia de información que se encuentra en dicho texto. De ese modo no intentan examinar detalladamente el texto sino que se conforman con buscar patrones o contar la frecuencia de ciertas palabras (Molla, D. 2003).

En una arquitectura común, los sistemas reducen la cantidad de información a procesar y en cada fase utilizan métodos más intensivos de análisis como se muestra en la figura 3, en donde puede observarse el sistema de búsqueda genérico.

Figura 3. **Sistema de búsqueda genérico**



Fuente: Mollá (2003). *Hacia el uso de la información sintáctica y semántica en los sistemas de búsqueda de respuestas.*

4.3 Arquitectura comúnmente utilizada en sistemas QA

Los componentes de la arquitectura son:

4.3.1 Análisis de la pregunta

La clasificación de la pregunta consiste en Identificar, etiquetar y categorizar la misma. Se refiere a una serie limitada de clasificaciones. La forma más básica diferencia una de la otra por medio de la partícula interrogativa. Clasificaciones más complejas vinculan el tipo de pregunta con el tipo de respuesta.

La clasificación de la pregunta contiene 3 niveles: el primer nivel define el tipo básico de la pregunta. El segundo nivel contiene el tipo de pregunta, adiciona información sobre el contexto y define el tipo de respuesta esperado. Al obtener

el segundo nivel ya es posible vincular el tipo de respuesta esperada con el tipo de objeto que se busca.

Por ejemplo, como se muestra en la figura 4, si obtenemos el primer nivel por la partícula *WHICH*, el segundo nivel analizará el contexto de la pregunta. Se distingue entre *WHICH-WHEN* que espera un tipo de respuesta temporal de *WHICH-WHERE* que espera un tipo de respuesta espacial o *WHICH-WHO* que espera una respuesta de tipo persona. En el tercer nivel para WHICH-WHEN se espera un objeto tipo *DATE* o en *WHICH-WHERE* se espera un objeto tipo lugar.

Figura 4. **Taxonomía de preguntas**

Q-class	Q-subclass	A-type
WHAT	basic what what-who what-when what-where	money | number | definition | title | nnp | undefined person | organization date location
WHO		person | organization
HOW	basic how how-many how-much how-far how-tall how-rich how-large	Maner number money | price distance number undefined number
WHERE		Location
WHEN		Date
WHICH	which-who which-where which-when which-what	Person location date nnp | organization
NAME	name-who name-where name-what	person | organization location title | nnp
WHY		Reason
WHOM		person | organization

Fuente: Barco (2007). *Sistema pregunta-respuesta.*

Se clasifica bien una pregunta como "*Which city has the oldest relationship as sister-city with Los Angeles?*". El primer nivel se puede realizar mediante concordancia de patrones o técnicas estadísticas. Para el segundo nivel es

necesario realizar un análisis semántico y determinar que "*city*" implica localización y, por tanto, se trata del subtipo *WHICH-WHERE*. En el tercer nivel se sabe que se debe buscar un objeto tipo lugar. La clasificación puede necesitar cierto nivel de análisis semántico de las palabras.

4.3.2 Lingüística de la pregunta

Para conseguir un análisis lingüístico correcto se emplea diversidad de técnicas. Algunas de ellas son las siguientes:

- *Bag of words*: considera la pregunta como una lista de palabras sueltas que se introducen en el recuperador de información sin contemplar ningún tipo de extensión ni orden de importancia entre ellas.

- *Stemming*: esta técnica consiste en aminorar una palabra a su forma raíz; es decir, la parte de la palabra que es invariable a todas sus formas flexionadas eliminando los sufijos. El *stemming* en el proceso de análisis de la pregunta se aplica a los diferentes términos de la entrada y lo reduce a la búsqueda de la forma básica de la palabra en lugar de su forma derivada. Uno de los algoritmos más utilizados en *Stemming* es el algoritmo de Porter (1980), originalmente diseñado para el inglés.

- Lematización: técnica computacional para determinar el lema de una palabra. Este proceso ya implica determinar la categoría gramatical de la palabra, por lo que se requiere de una gramática de la lengua y de un diccionario.

4.3.3 Proceso de búsqueda

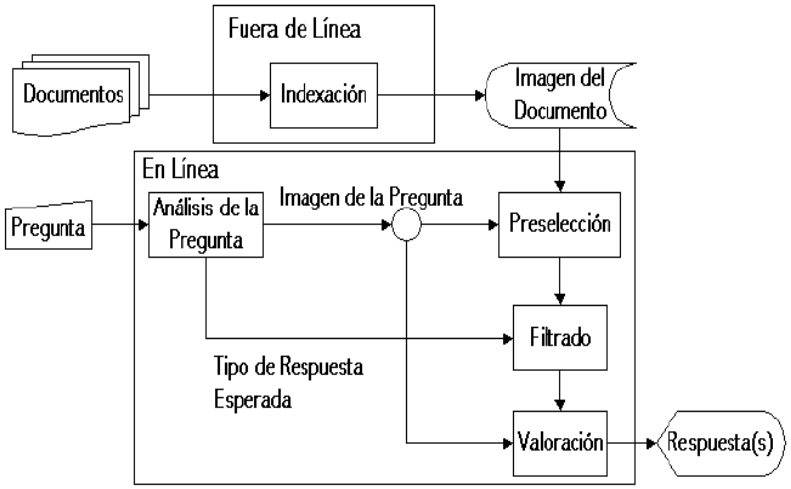
Cualquier organización dispone de documentos guardados digitalmente, de esta manera almacena mucho más documentos digitales que físicos. Esta información almacenada digitalmente se vuelve complicada de buscar y encontrar el dato exacto. Los sistemas actualmente desarrollados que recuperan la información son utilizados para tal efecto de búsqueda, y de esa manera, encontrar respuestas exactas en lugar de documentos relevantes, dado que el usuario dispone de muy poco tiempo para encontrar los datos deseados en una gran lista de documentos.

Existen herramientas de búsqueda que facilitan la misma. Se pueden utilizar analizadores sintácticos con gramáticas de gran cobertura, como Link Grammar (Sleator, D., 1993) y Conexor FDG (Jarvinen, 1997). También existen paquetes estadísticos y herramientas para administrar colecciones de textos, diccionarios electrónicos y otras herramientas léxicas como WordNet (Fellbaum, C., 1998) módulos de tratamientos de textos como separadores de palabras, etiquetadores de entidades y herramientas para comprobación de argumentos lógicos como OTTER (Kalman, A., 2001). Estas herramientas colaboran en la construcción de sistemas de búsquedas de respuestas y apoyan en la interpretación de lenguaje natural en un tiempo corto y de forma asequible.

En una arquitectura normal los sistemas reducen la cantidad de información a procesar y en cada etapa utilizan métodos más intensivos de análisis. De esa forma, un indexador analiza los documentos y saca toda la información, como las palabras claves y sus posiciones relativas. Durante la sesión de preguntas y respuestas, un analizador clasifica la pregunta y determina el tipo de respuesta. Este analizador también devuelve un análisis de la pregunta más detallado que la información que el indexador había producido de los

documentos. La información puede ser una simple lista de palabras o una forma lógica de cierta complejidad. El siguiente módulo es el pre seleccionador, el cual determina qué documentos pueden contener la respuesta. Por lo general usa técnicas de recuperación de la información como el modelo de espacios vectoriales o modelos booleanos. Estos modelos consideran los documentos como un set de palabras sin estructura ni dependencias (Voorhees, M., 2001). Los documentos que se devuelven pasan por un filtro que toman las oraciones o pasajes que tengan más probabilidad de tener la respuesta. Finalmente, un valorizador analiza los fragmentos y les concede un valor relacionado con la probabilidad de que la respuesta se encuentre en dicho texto. El valorizador utiliza toda la información que dispone. Otros sistemas de búsquedas utilizan mecanismos para evaluar la calidad de la salida de cada una de las etapas. Otros sistemas, como el que se muestra en la figura 5, utilizan métodos de patrones de texto que pueden tener la respuesta para determinar la respuesta final. Estos patrones pueden ser complejos y estar organizados jerárquicamente.

Figura 5. **Sistema de búsqueda de respuestas genérico**



Fuente: Mollá (2003). *Hacia el uso de la información sintáctica y semántica en los sistemas de búsqueda de respuestas.*

4.3.4 Recuperación y selección de la información

Un SRI está constituido por tres componentes básicos: 1. La base de datos documental, 2. El subsistema de consultas y 3. El mecanismo de emparejamiento o evaluación.

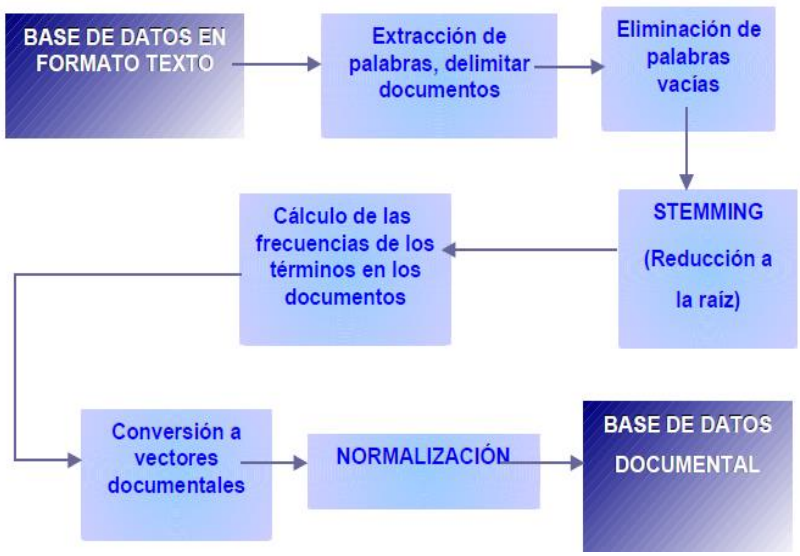
- La base de datos documental: un documento es un objeto de datos de naturaleza textual. Estos objetos no se insertan en el SRI, sino que estarán representados por unos elementos llamados descriptores. Esto proporciona una mayor eficiencia a la base de datos y hace que el tiempo de búsqueda en ella sea mucho menor. El documento se compondrá de un conjunto de descriptores. Desde la perspectiva matemática, la base es una matriz en la que cada columna indica las asignaciones de un descriptor y cada fila es un documento. En cada fila aparecen “unos” en las columnas relativas a los descriptores asignados al documento y “ceros” en las restantes; de esa manera, cada documento estará representado por un vector de unos y ceros.

Los documentos de tipo textual se pueden representar por una componente estructurada en campos (título, autor, resumen, palabras clave...), por una componente no estructurada o el texto literal. La representación textual de cada documento se basará en los términos de indización (que pueden ser tanto palabras como frases), los cuales son identificadores de los propios documentos.

El primer paso para la construcción de la base consiste en extraer los términos del texto del documento, como se muestra en la figura 6. Cada una de las palabras se comparará con una lista de palabras vacías, que eliminará las que no tienen interés o carecen de significado propio. Después, las palabras

podrán sufrir un proceso de recorte de sus raíces (por ejemplo, palabras como informática, información... pueden reducirse a la raíz “infor”). Luego, se aplica una función de ponderación para obtener los pesos asociados a cada término en los vectores de documentos y se introducen estos en la base de datos documental.

Figura 6. **El proceso documental**



Fuente: Rodríguez (2005). *Modelos de recuperación de información.*

- El subsistema de consultas está integrado por la interfaz que permite al usuario formular consultas al SRI y por un analizador sintáctico que toma la consulta escrita por el usuario y la divide en partes integrantes. Para realizar esta tarea, el componente incluye un lenguaje de consulta que analiza las reglas y genera consultas apropiadas y la forma para seleccionar los documentos relevantes. La interfaz ofrecerá una mejor experiencia al usuario en el momento de formular la pregunta. También mostrará al usuario el resultado de la búsqueda al ser procesada la consulta. En muchas ocasiones, los usuarios de SRI realizan sus peticiones basándose en la estructura de consultas booleanas (con operadores booleanos; es decir, y, o, no). La consulta del usuario no puede

procesarse directamente en su forma original. Se debe desglosar la consulta en sus componentes básicos y comprobar si el formato es correcto; es decir, si su composición cuadra con las reglas del lenguaje de consulta. Esta comprobación se lleva a cabo tanto a priori como a posteriori. Finalmente, la consulta se indizará o vectorizará y será enviada al mecanismo de evaluación para estudiar qué documentos se consideran relevantes para las necesidades de información que representa.

4.3.5 Extracción y generación de respuestas

Acerca del mecanismo de evaluación, en este punto ya se tiene una representación del contenido de los documentos en la base documental y también de las consultas que se desean realizar proveniente del subsistema de consulta. Queda por resolver la selección de los documentos que se considera relevantes, contenidos en la base documental, de acuerdo con los criterios de la consulta. De esto se encarga el mecanismo de evaluación: evalúa el grado en el que las representaciones de los documentos satisfacen los requisitos definidos en la consulta y recupera aquellos documentos que son relevantes a la misma. Este grado es lo que se denomina RSV (Retrieval Status Value). Principalmente, existen dos modalidades de evaluación: sistemas que emparejan los documentos individualmente con la consulta, uno por uno y otros que los emparejan en su conjunto. Se dedicará la sección siguiente a analizar los modelos de RI más conocidos.

4.3.6 Técnicas de procesamiento del lenguaje natural

La técnica de desambiguación del sentido de las palabras proporciona el sentido exacto del término contra una ontología previamente establecida. Los sistemas de QA de dominio restringido pueden definir su propia ontología

semántica para términos relacionados con el dominio, de esta forma se puede reducir enormemente las posibilidades de elección para el desambiguador y, con ello, sus errores; también se puede descartar aquellos términos que no son relevantes para el dominio. Además, una vez determinado el sentido de la palabra es posible añadirle otras propiedades semánticas derivadas de sus relaciones con otras palabras.

El análisis sintáctico superficial proporciona información importante al análisis de la pregunta. Mediante este análisis se pueden detectar datos básicos necesarios para la búsqueda de información como los grupos nominales y verbales. La salida que utiliza este tipo de herramientas es una lista plana de constituyentes básicos.

5. PRESENTACIÓN DE RESULTADOS

5.1 Construcción del prototipo

El prototipo fue diseñado y construido en una computadora local con las siguientes características:

- Sistema Operativo Windows 10
- .Net Framework 3.5
- Disco Duro de 228 GB de estado sólido
- Memora de 4 GB
- Procesador Core i7 vPro 7th Gen

El prototipo recibe como parámetros de entradas las preguntas realizadas por los usuarios de acuerdo al dominio; en este caso, preguntas relacionadas a orientar al estudiante de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala con las actividades que desee saber de los procesos de la carrera de Ingeniería en Ciencias y Sistemas.

De acuerdo a los parámetros o preguntas ingresadas en el sistema, por medio de una caja de texto o un chat, el prototipo consulta a la base de datos de conocimiento para luego, si fuera necesario y si existiera documento asociado a la pregunta, consultar a la base de datos documental y devolver el documento asociado.

La base de datos contiene toda la información previamente cargada de los procesos y actividades de cada facultad de ingeniería que es asociada con la

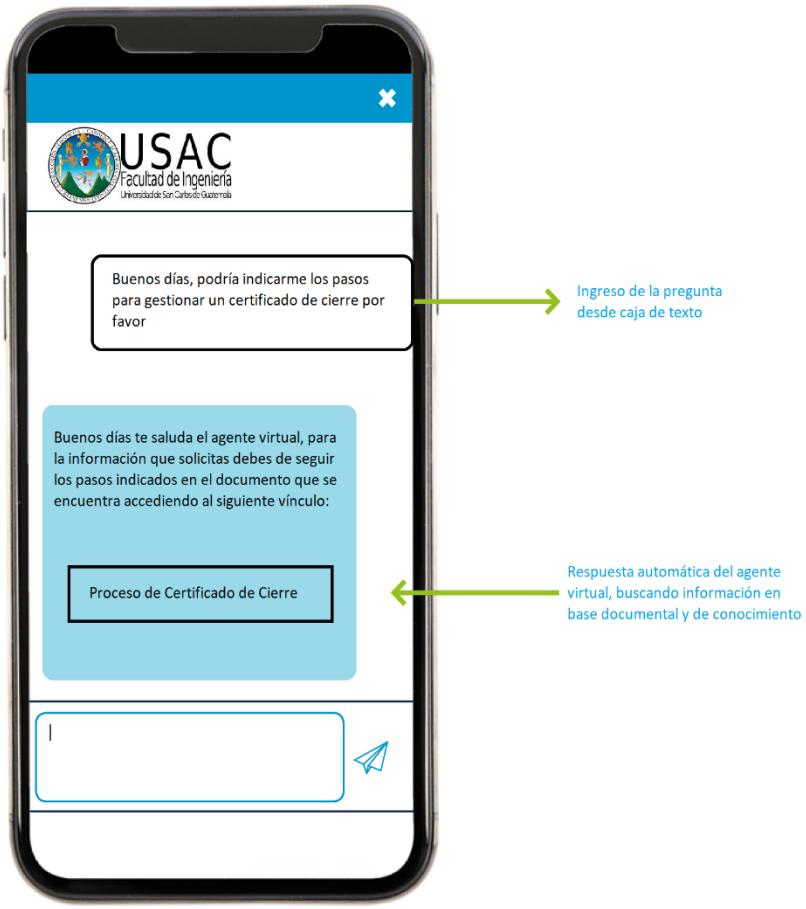
información de los documentos relacionados con las mismas. La información en la base de conocimiento es la siguiente:

- Información de procesos de ingreso a la Facultad.
- Información de procesos generación de certificado de cierre de pensum
- Información de procesos de cada escuela de la Facultad.
- Información de procesos de trámite de título.

La información previamente cargada en el prototipo y el proceso llevado a cabo para la búsqueda de la información correcta, de forma eficiente y eficaz, es el que lleva a los resultados presentados y analizados en esta sección.

El prototipo de esta aplicación se muestra en la figura 7.

Figura 7. **Prototipo del agente virtual**

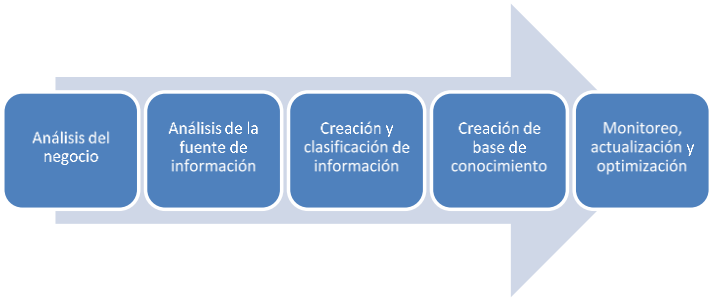


Fuente: elaboración propia

5.2 Procedimiento para actualizar fuente de información de un sistema experto

El procedimiento adecuado utilizado para actualizar la fuente de información de un sistema experto se muestra en la Figura 8.

Figura 8. **Proceso de actualización de la fuente de información**



Fuente: elaboración propia

- Análisis del negocio: el primer paso para una actualización de fuentes correcta es familiarizarse con el negocio y analizar el dominio que pertenece. Para ello se determinó la unidad académica a analizar para el prototipo. En esta investigación se tomó como unidad académica la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala. Este proceso se realiza de forma manual.

- Análisis de la fuente de información: para esta etapa se consideraron varias fuentes principales y base de datos de negocio que contienen información y documentos relevantes al usuario; por ejemplo, base de datos de estudiantes, fuentes de información de la unidades académicas que contienen fechas, descripciones de horarios, procesos, procedimientos, documentos informativos, documentos explicativos, etc.. Este proceso se realiza de forma manual.

- Creación y clasificación de la información: en esta etapa se clasificó de forma ordenada los documentos e información recopilada en la etapa uno. Esta etapa forma parte de un proceso de internalización en una base de datos no estructurada utilizando MongoDB como la base orientada a documentos. De esa manera se puede consultar o preguntar acerca de algún tema en particular. La clasificación ayudará a buscar la información

en la clasificación correcta, de modo que el usuario pueda extraer la información de forma óptima. Este proceso se realiza sin utilizar ninguna herramienta automatizada; sin embargo la información se obtuvo con el motor de búsqueda de la base de datos no estructurada.

- Creación de base de conocimiento: después de clasificar y organizar la información y documentos se procedió a ingresar dicha información a una base de datos de conocimiento, la cual contendrá la información de las fuentes de forma ordenada y clasificada. Con esto podemos garantizar la extracción por medio de algoritmos y reglas de inferencia que se utilizarán para buscar una respuesta correcta. Para realizar este proceso se utilizan varias herramientas como la base de datos documental; en esta, un documento es un objeto de datos que contienen elementos que se denominan descriptores. Estos le dan mayor eficiencia a la base de datos y están representados por una matriz en donde las columnas son los descriptores y las filas, los documentos.

- Monitoreo, actualización y optimización: la base de conocimiento es monitoreada, actualizada y optimizada, dado que el negocio es cambiante. De no tener un proceso correcto de actualización y optimización, el sistema no tendrá una fuente actualizada y por tal motivo el sistema no será fiable. Un proceso correcto de actualización de la información debe incluir las siguientes actividades realizadas de forma manual:

    - Mantenimientos de información diarios, cada hora si fuera necesario.

    - Verificación de memoria del servidor de base de datos del sistema.

- Verificación de utilización de CPU del servidor de base de datos del sistema.

- Verificación de entradas y salidas a disco del servidor de base de datos.

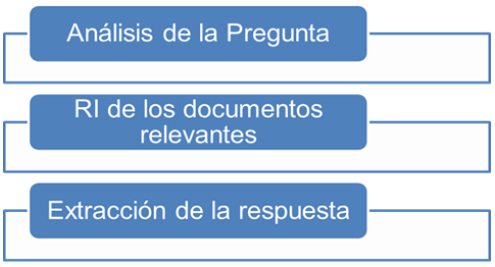
- Verificación y optimización de tablas y consultas que utilicen mucha memoria.

5.3 Arquitectura de un sistema experto

La arquitectura tiene como fundamentos los conceptos de la recuperación de la información, aunque el fin de un sistema Question-Answer (QA) es diferente a un sistema de recuperación de información (RI). Un sistema QA está orientado por respuesta y no por la pregunta, como es el caso de los sistemas RI. Por ejemplo, si se hace una consulta como "¿Cuándo es la próxima fecha de evaluación específica en la facultad de ingeniería?", un sistema RI buscaría documentos relacionados a la pregunta aunque la respuesta estuviera o no en dicho documento, mientras que en un sistema QA busca el documento y específicamente se debe encontrar en estos documentos la fecha específica a la que hace referencia la pregunta.

La arquitectura QA utilizada en el prototipo está basada en 3 componentes básicos que pueden verse en la figura 9. Los componentes de análisis de la pregunta y extracción de la respuesta se utilizan para reconocer la búsqueda y la forma de encontrarla en la base documental, mientras que el componente RI de los documentos relevantes se utiliza para recuperar la información en base a la respuesta que desea el usuario, tal como se mencionó.

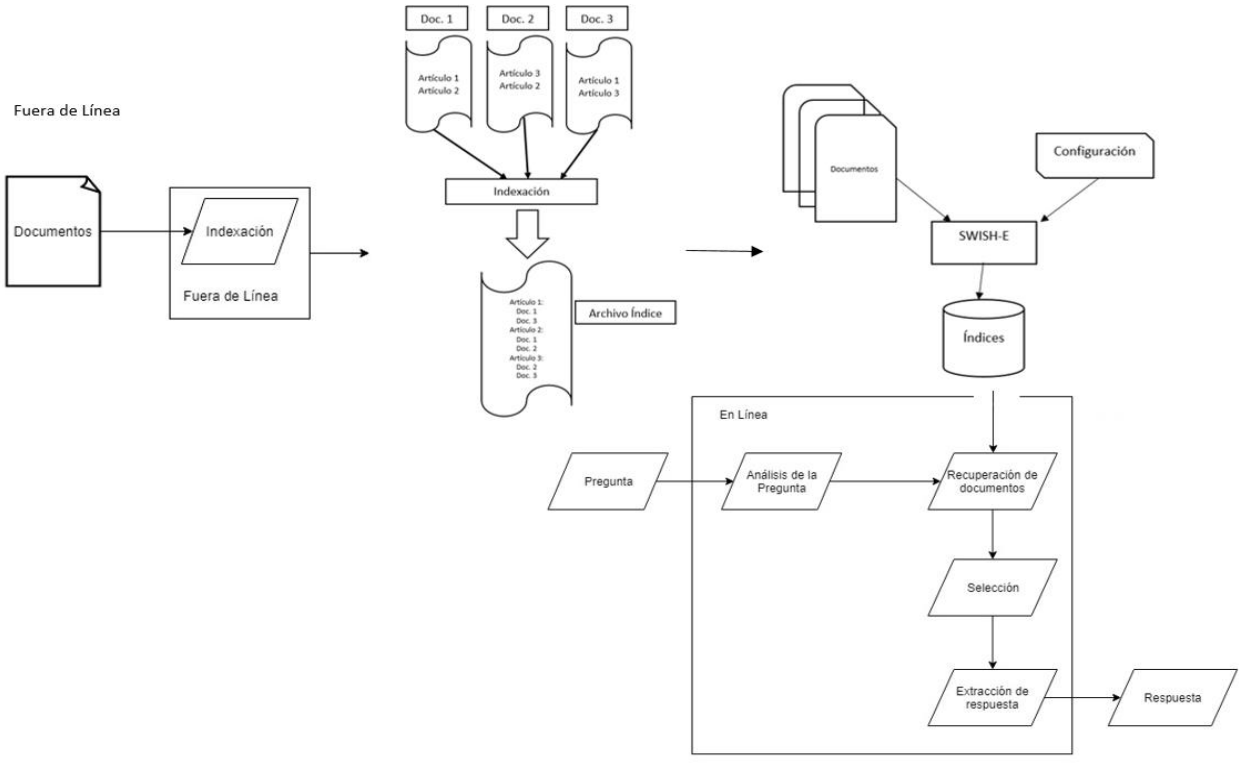
Figura 9. **Componentes de la arquitectura propuesta**



Fuente: elaboración propia

La arquitectura general del sistema propuesto se muestra en la figura 10:

Figura 10. **Arquitectura general del sistema QA**



Fuente: elaboración propia

La tecnología utilizada para que funcione correctamente la arquitectura es MongoDB. Esta herramienta se utiliza para guardar la base documental y para la implementación de índices, recuperación de documentos y extracción de respuestas. La figura 11 muestra la arquitectura y los componentes de la herramienta.

El diseño de MongoDB combina las capacidades de las bases de datos no estructuradas denominadas NoSQL y las bases de datos estructuradas o relacionales. Algunas características que se ofrecen son:

- Proporciona un lenguaje de consulta para que los usuarios puedan acceder y manipular los datos de una manera sofisticada. Adicionalmente, para acceder a la información proporciona índices que mejoran el desempeño en las búsquedas y extracción de la información, nativos de la herramienta.

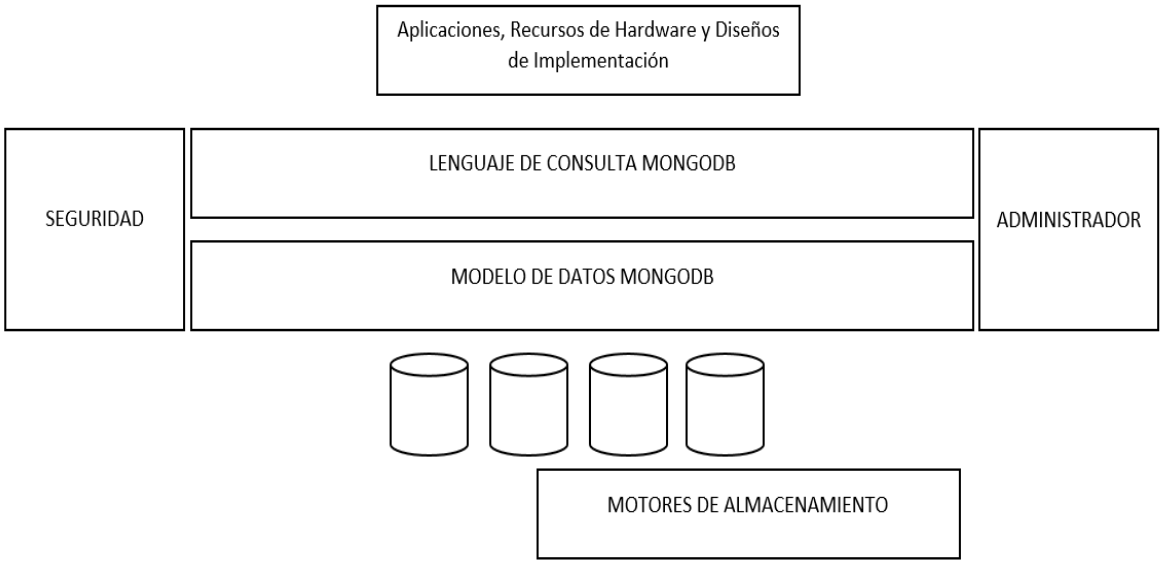
- Maneja una fuerte consistencia de datos. Esto hace que las aplicaciones puedan acceder a la información de manera inmediata.

- La herramienta permite una integración empresarial, de modo que la base de datos pueda monitorearse, automatizarse e integrarse a la infraestructura tecnológica existente, procesos y personal. Incluye analistas de datos, administradores de base de datos y equipo de operaciones.

- Almacenamiento de documentos en forma binaria o Binary Jason (BSON). Estos documentos tienen todos los datos para un registro dado.

- La herramienta permite la conectividad de varias aplicaciones, recursos de hardware y múltiples diseños de implementación de software con una sola base de datos, con el mismo lenguaje de consulta, seguridad, modelo de base de datos y administración, alimentada por varios motores de almacenamiento, como se muestra en la figura 10.

Figura 11. **Arquitectura de almacenamiento para flexibilidad de aplicaciones**



Fuente: elaboración propia

5.3.1 Análisis de la pregunta

Este componente fue utilizado para establecer el tipo de pregunta, determinar la respuesta, establecer el foco de la pregunta, determinar los términos para efectuar la consulta al sistema RI y el contexto semántico de la respuesta.

El objetivo de este componente se cumplió al seguir los siguientes pasos:

- Listar las palabras del sistema en la base de datos documental por medio de procedimientos almacenados y lenguaje de programación Visual Studio .Net C#.

- Como segundo paso se utilizó la técnica de *stemming* (es un método para reducir una palabra a su raíz (en inglés, a un *stem*) y se armonizó una palabra a su forma raíz; es decir, la parte de la palabra que es invariante a todas sus formas flexionadas, eliminando los sufijos. Se aplica a los diferentes términos de la entrada y lo reduce a la búsqueda de la forma básica de la palabra en lugar de su forma derivada. Para esta tarea se utilizó el algoritmo de Porter y un lenguaje de programación Visual Studio .Net C#.

    - Por ejemplo, el algoritmo colaboró en la identificación de la frecuencia de una palabra y asegurar que la forma de la misma no infiera en la frecuencia de la misma en el texto. Por ejemplo, en el texto “Aquel es un caballo de la caballería militar, los otros caballos no.”, la frecuencia del término “caball” es 3; con esto se remueven los sufijos comunes morfológicos de las palabras diferentes pero con un stem común.

- Como tercer paso se eliminaron las palabras de paso por medio de un análisis cuidadoso del texto. Se eliminaron artículos y algunas preposiciones; sin embargo, se tuvo el cuidado de no eliminar “palabras de paso” que ayudarían a determinar la respuesta correcta dentro del sistema QA. Para un sistema RI este paso se realiza sin ningún tipo de restricciones; sin embargo, para un sistema QA, algunas preposiciones, partículas interrogativas o pronombres pueden tener injerencia en el momento de interpretar la pregunta y extracción de la respuesta. Este paso

fue realizado por medio de la base de datos documental y la utilización de procedimientos almacenados.

- Como cuarto paso se realizó un análisis sintáctico de dependencias y sus relaciones, y quedó un árbol de análisis de dependencias.

- Como quinto paso se utilizó la técnica de desambiguación del sentido de la palabra. Esto se realizó con una ontología semántica para términos relacionados con el dominio, que ayudó a reducir la elección y, por consiguiente, los errores.

- El sexto paso consistió en clasificar la pregunta con la taxonomía de Moldovan; esto ayudó a relacionar el tipo de pregunta con el tipo de respuesta. Al encontrar el tipo de pregunta se pudo encontrar el tipo de objeto por buscar. Por ejemplo, en la pregunta “¿Dónde puedo encontrar el pénsum de ingeniería en sistemas?”, en el primer nivel se identificó el tipo básico de la pregunta que, en este caso, se basa en la partícula interrogativa. En el segundo nivel se añadió el contexto para determinar el tipo de respuesta esperado. Con la palabra “Dónde” se espera un tipo de respuesta espacial, el contenido puede encontrarse en un lugar físico o virtual. En el tercer nivel ya se puede relacionar el tipo de respuesta esperada con el tipo de objeto. En este ejemplo, el objeto es tipo lugar, con una dirección física o virtual.

- Por último, se estableció el foco de la pregunta y se determinaron lo términos claves. Esto se realizó por medio de un análisis programado de la pregunta con base en los pasos anteriores. En ese sentido, se determinó posteriormente la lista de términos a recuperar. Por ejemplo, en la pregunta “En el año 2019, ¿en qué día de la semana caerá la inscripción

de ingeniería en sistemas?”, el foco es “día de la semana”. Es el término focal o más importante de la pregunta; sin embargo, se debe tomar en cuenta que es muy poco probable que se recupere la respuesta correcta. Por lo tanto, un foco que sí puede ayudar sería el término “día” o “fecha”; esto puede ayudar a recuperar la información de forma correcta.

Con estos análisis, pasos y herramientas utilizadas, el prototipo es creado para un nivel de eficiencia y fiabilidad necesario para ser utilizado por los usuarios.

5.3.2 Recuperación de la información

Con este componente se encuentran los documentos relevantes a la pregunta utilizando una base de datos no estructurada. Los pasos que se siguieron fueron los siguientes:

5.3.2.1 Indexación de la colección

Para realizar esta tarea se realizó un proceso fuera de línea del conjunto de documentos relevantes, con el objetivo de determinar un índice y reconocer los documentos importantes o relevantes sin tener la necesidad de realizarlo en tiempo real.

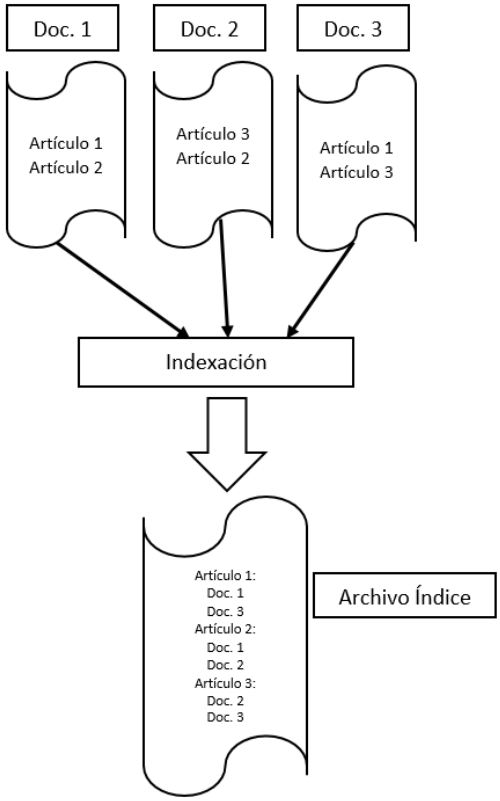
Por lo tanto, se creó un archivo que contiene todas las palabras de los documentos. En cada una de ellas se creará un vínculo a los documentos en los que la contiene. El indexar es un proceso que conlleva los siguientes pasos:

- Análisis de documentos.
- Extracción de palabras o grupos de palabras existentes.

- Ponderación de documentos por importancia.
- Extracción de las palabras de un archivo índice.

La figura 12 muestra cómo se forma un archivo índice tomando 3 archivos (doc. 1, doc. 2 y doc. 3) que contienen algunas palabras sueltas. Generalmente hay una lista de palabras que no se toman en cuenta por la carencia de significado (*Stop-Words*).

Figura 12. **Indexación de archivos**



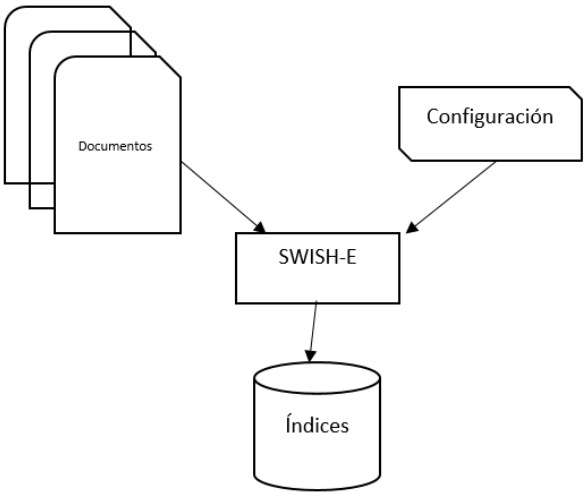
Fuente: elaboración propia.

Para indexar se utilizó la herramienta Swish-e con la capacidad de indexar textos planos, formatos como PDF o PostScrit. Esta herramienta es rápida, tanto

para indexar como para la búsqueda. Utiliza un nivel bajo de consumo de memoria; es muy fácil la configuración y se integra con otras herramientas.

Para la construcción del índice se utilizó un archivo de configuración y los tipos de archivos que se indexaron como se muestra en la figura 13.

Figura 13. **Indexación por swish-e**



Fuente: elaboración propia.

5.3.2.2 Recuperación de documentos

Con los términos claves de la pregunta se buscó la palabra en el archivo índice y se seleccionan los documentos más relevantes. La búsqueda se realiza sobre palabras, frases o palabras truncadas. Los documentos son devueltos de acuerdo a una puntuación.

5.3.2.3 Selección de pasajes relevantes

La presentación de un resultado en un sistema RI puede hacerse en forma de documento; sin embargo, para un sistema QA el resultado es una respuesta concreta a la pregunta realizada. Por lo tanto, se definió un proceso de refinamiento en donde se buscan los pasajes relevantes a la pregunta dentro del documento. Luego se utilizó la recuperación de la información para obtener los segmentos más relevantes, considerando el porcentaje de palabras clave que aparecen en el documento y teniendo en cuenta la obligatoriedad de algunas palabras como, por ejemplo, el foco de la pregunta.

Los pasajes obtenidos son enviados al subproceso de extracción de la respuesta en modo de texto para la comparación del pasaje con la pregunta.

5.3.3 Extracción de la respuesta

Se realizó al localizar la respuesta en los fragmentos relevantes, de la siguiente manera:

- Con técnicas de análisis superficial.

- Al validar la respuesta de acuerdo a la precisión, eficiencia y simplicidad de la pregunta, basadas en la proximidad y patrones dentro de la base documental.

Si se considera la siguiente pregunta:

¿Quién firma los certificados de cursos aprobados?

Respuesta candidata: Secretario Académico

En una validación por proximidad se toman los términos relevantes de la pregunta y se busca en los documentos para obtener los pasajes descritos en la tabla 2. Cada uno de ellos tiene una proximidad entre los términos: secretario académico, firma y certificado de cursos aprobados. Con esto determinamos que es la respuesta correcta.

Tabla II. **Pasajes descritos en los documentos**

El **secretario académico** es el responsable de **firmar** finalmente el **certificado de cursos aprobados**.
Una de las funciones del **secretario académico** es **firmar** los **certificados de cursos aprobados.**
Para **firma** de **certificados de curso aprobados** el responsable será el **secretario académico.**
…

Fuente: elaboración propia.

También se valida por patrones donde la pregunta se podría reformular como la siguiente afirmación: “El secretario académico firma los certificados de cursos aprobados”. Si se busca exactamente esta afirmación en la base documental se encuentra la existencia varias veces, con esto se puede dar como respuesta correcta.

6. ANÁLISIS Y DISCUSIÓN DE RESULTADOS

Las variables identificadas en la metodología, que son robustez, fiabilidad y rapidez, ayudaron a determinar el prototipo adecuado para el proyecto. Esto quiere decir que el prototipo cumple con los indicadores de dichas variables; se validaron los tiempos de reinicio de falla, tasa de ocurrencia de la falla y tiempos de respuesta por evento y transacción.

6.1 Robustez

Para obtener el tiempo de recuperación de fallas y evaluar los diferentes eventos que causan la falla del prototipo se realizó un cuadro de análisis de riesgos (ver tabla 3), en donde se clasificaron los tipos de riesgo del equipo y del sistema. Estos ayudaron a determinar el tiempo de respuesta ante cualquier riesgo de falla.

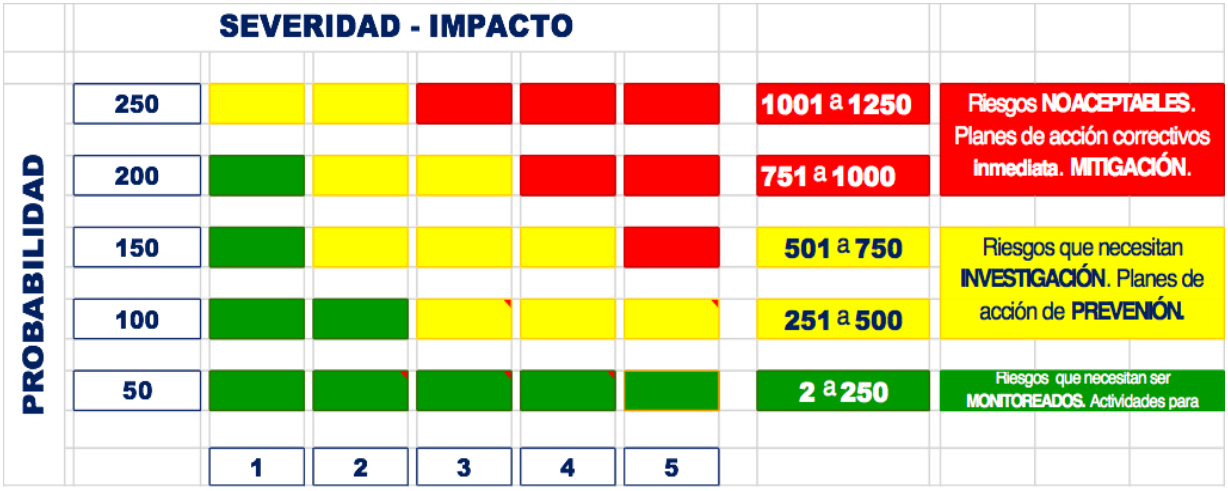
La tabla muestra dos tipos de riesgos, los de equipo y de sistema, de los cuales se derivan los siguientes riesgos:

- Problemas en la conectividad.
- Problemas que pueden ocasionar quedar fuera de línea.
- Daños en la fuente, batería y discos.
- Conato de incendio.
- Pérdida de datos.
- Distribución de información relevante y confidencia.
- Desastres naturales.
- Problemas con aíre acondicionado.

- Sistema operativo sin soporte.
- Riesgo de ataque cibernético.
- Pérdida de información en la base de datos.
- Inconsistencia de información.

El criterio de evaluación se basa en una matriz de calor donde se clasifican los distintos riesgos en base a probabilidad y severidad o impacto:

Tabla III. **Criterios de evaluación**



Fuente: elaboración propia.

Los criterios utilizados para la severidad o impacto se ven reflejados con las siguientes letras en la tabla X: F, S, P, E, A y V.

De esta forma cada riesgo es evaluado y, al final, el resultado se basa en la siguiente fórmula: probabilidad X severidad. El resultado ubica en qué rango de criterio se encuentra y cómo se priorizan los riesgos, todo en base a la tabla 3 de criterios de evaluación.

Tabla IV. **Análisis de riesgos de prototipo**

	F (CRITERIO DE FUNCIÓN)	S (CRITERIO DE SUSTITUCIÓN)	P (CRITERIO DE PROFUNDIDAD)	E (CRITERIO DE EXTENSIÓN)	A (CRITERIO DE AGRESIÓN)	V (CRITERIO DE VULNERABILIDAD)	Puntuación sobre 50	Puntuación sobre 25
RIESGO	¿Cómo afecta a nuestra actividad la materialización del Riesgo?	¿En qué grado puede sustituirse el bien afectado?	¿Cómo afecta a la imagen de la empresa, tiene efectos psicológicos dentro de ella?	¿Hasta dónde llegan las repercusiones de la materialización del Riesgo?	¿Qué probabilidad real existe de que el Riesgo se manifieste?	¿Si el Riesgo se manifiesta cuál es la probabilidad de que se produzca un daño?	Severidad - Impacto	Probabilidad
NO Conectividad con el Enlace redundante de Datos	Gravemente	Difícilmente	Perturbaciones Grandes	De carácter Nacional	Alta	Alta	32	16
Probabilidad de quedar fuera de línea	Gravemente	Difícilmente	Perturbaciones Limitadas	De carácter Nacional	Alta	Alta	28	16
Daño físico de componentes del equipo (fuente, batería, discos)	Gravemente	Sin Muchas Dificultades	Perturbaciones Muy Leves	De carácter Local	Alta	Muy Alta	14	20
Incendio en el Data Center	Muy Gravemente	Muy Difícilmente	Perturbaciones Grandes	De carácter Nacional	Baja	Muy Alta	41	10
Perdida o daño de información	Muy Gravemente	Difícilmente	Perturbaciones Grandes	De carácter Internacional	Muy Alta	Muy Alta	40	25
Fuga de Información	Muy Gravemente	Difícilmente	Perturbaciones Grandes	De carácter Nacional	Alta	Muy Alta	36	20
Colisión Aérea contra Instalaciones	Muy Gravemente	Muy Difícilmente	Graves Perturvaciones	De carácter Nacional	Alta	Muy Alta	45	20
Terremoto	Muy Gravemente	Muy Difícilmente	Perturbaciones Grandes	De carácter Nacional	Media	Muy Alta	41	15
Inundación por condenzadora del aire acondicionado	Muy Gravemente	Difícilmente	Perturbaciones Grandes	De carácter Nacional	Media	Muy Alta	36	15
Obsolescencia de Sistema Operativo por avances tecnológicos	Gravemente	Difícilmente	Perturbaciones Grandes	De carácter Nacional	Media	Alta	32	12
Robo	Gravemente	Difícilmente	Perturbaciones Grandes	De carácter Nacional	Media	Alta	32	12
Ataque Cibernetico	Muy Gravemente	Difícilmente	Perturbaciones Grandes	De carácter Nacional	Media	Alta	36	12
Saturación de Recursos (Procesador, memoria, disco)	Gravemente	Difícilmente	Perturbaciones Grandes	De carácter Nacional	Alta	Alta	32	16
Falla en la búsqueda de respuestas en Base de Datos	Muy Gravemente	Muy Difícilmente	Graves Perturvaciones	De carácter Nacional	Alta	Muy Alta	45	20
Bloqueo de tablas	Medianamente	Difícilmente	Perturbaciones Leves	De carácter Nacional	Alta	Alta	20	16
Pérdida de información en la base de datos	Medianamente	Difícilmente	Perturbaciones Leves	De carácter Nacional	Alta	Alta	20	16
Data inconsistente	Gravemente	Difícilmente	Perturbaciones Limitadas	De carácter Nacional	Alta	Alta	28	16
Mala organización de la fuente de información en la base de datos	Gravemente	Sin Muchas Dificultades	Perturbaciones Muy Leves	De carácter Local	Alta	Muy Alta	14	20
Eliminar, reemplazar, cambiar tablas de la Biblioteca	Gravemente	Sin Muchas Dificultades	Perturbaciones Muy Leves	De carácter Local	Alta	Muy Alta	14	20
Modificación de tablas	Gravemente	Sin Muchas Dificultades	Perturbaciones Muy Leves	De carácter Local	Alta	Muy Alta	14	20
Perdida de información por replicación en línea	Muy Gravemente	Difícilmente	Perturbaciones Grandes	De carácter Nacional	Alta	Muy Alta	36	20
No replicación de información	Muy Gravemente	Difícilmente	Perturbaciones Grandes	De carácter Nacional	Alta	Muy Alta	36	20

Fuente: elaboración propia.

De acuerdo a la tabla 4 que contiene los riesgos identificados, se hicieron pruebas y análisis de tratamientos de los riesgos para validar los tiempos de respuesta ante las fallas identificadas. Se propusieron las alternativas de solución a dichos fallos. En la tabla 5 se puede observar los tiempos de recuperación y las soluciones a los fallos identificados.

Los tiempos de recuperación fueron medidos con simulaciones y fuentes de información histórica de eventos pasados o similares. El tratamiento de los riesgos fue también investigado de acuerdo con una fuente de información histórica y experiencias pasadas en donde el riesgo se materializó en un problema.

Tabla V. **Análisis de tiempos y tratamientos de solución de riesgos**

No.	Tipo	RIESGO	Tiempo de Recuperación	Tratamiento del Riesgo
1	Equipo	NO Conectividad con el Enlace redundante de Datos	4 hrs.	Contar con enlace reduntante de datos, através de un proveedor diferente
2	Equipo	Probabilidad de quedar fuera de línea	4 hrs.	Planta eléctrica y UPS´s, inversores
3	Equipo	Daño físico de componentes del equipo (fuente, bateria, discos)	8 hrs.	Contar con contrato activo de soporte de Hardware y stock local de hardware
4	Equipo	Incendio en el Data Center	24 hrs.	Se cuenta con sistema supresor de incendios, al cual se le da mantenimeinto y pruebas de funcionamiento 3 veces al año (ver control)
5	Equipo	Perdida o daño de Información	8 hrs.	Replicación de información a Sitio Alterno, respaldos diarios de información, Auditorias , Contrato de responsabilidad y confidencialidad
6	Equipo	Fuga de Información	8 hrs.	Implementación de políticas de DLP (Data Los Prevention) y análisis de Logs
7	Equipo	Colisión Aérea contra Instalaciones	16 hrs.	Replicación en el Sitio Alterno del 100% de la información
8	Equipo	Terremoto	48 hrs.	Trasladar el Data Center a otra instalación más segura
9	Equipo	Inundación por condenzadora del aíre acondicionado	8 hrs.	Replicación de información a Sitio Alterno, respaldos diarios de información, BCP, Traslado del Data Center a otras Instalaciones
10	Equipo	Obsolescencia de Sistema Operativo por avances tecnológicos	8 hrs.	Mantener en el presupuesto el rubro de extensión de Garantía de los Equipos, así como los planes de mantenimiennto de hardware y software
11	Equipo	Robo	8 hrs.	Controles de acceso, monitoreo a través de cámaras de seguridad, Seguro contra Robo
12	Equipo	Ataque Cibernetico	4 hrs.	Seguridad perimetral, gestión de alta y baja de usuarios, administración de accesos.
13	Equipo	Saturación de Recursos (Procesador, memoria, disco)	2 hrs.	Bitácoara de Cierre diario, depuración de logs y programas innecesarios
14	Sistemas	Falla en la búsqueda de respuestas en Base de Datos	2 hrs.	Análisis de errores en bitacora, revisión de pregunta alterna
15	Sistemas	Bloqueo de tablas	2 hrs.	Revisión periodica de base de datos y bloqueos, por medio de alertas. Revisión de índices de base de datos y manejo de buenas prácticas.
16	Sistemas	Pérdida de información en la base de datos	3 hrs.	Replicación en tiempo real
17	Sistemas	Data inconsistente	3 hrs.	Evaluación de datos y análisis de datos desde la fuente de información.
18	Sistemas	Mala organización de la fuente de información en la base de datos	5 hrs.	Revisión periodica de la fuente de información, actualización y optimización.
20	Sistemas	Eliminar, reemplazar, cambiar tablas de la Biblioteca	8 hrs.	Revisión periodica de base de datos y bloqueos, por medio de alertas. Revisión de índices de base de datos y manejo de buenas prácticas.
21	Sistemas	Modificación de tablas	8 hrs.	Alertas y bitacoras de tablas, adicionalmente replicaión a sitio alterno
22	Sistemas	Perdida de información por replicación en línea	8 hrs.	Revisión de datos replicados diariamente.
23	Sistemas	No replicación de información	8 hrs.	Revisión de datos replicados diariamente.

Fuente: elaboración propia.

6.1.1 Actualización de la fuente de información

El proceso de actualización de la fuente de información es fundamental para el prototipo. Por lo tanto, se realizó un análisis para seleccionar la fuente

correcta, ya que se usará para extraer la información de la pregunta; en dado caso esta fuente sea incorrecta, la información extraída no le será útil al usuario. El proceso de selección y actualización de la fuente de información debe tener un tiempo prudente y adecuado, para definir un conjunto de listas de fuentes a utilizarse.

Si la información requerida por el usuario no existe en la fuente de información, esta no será presentada; por lo tanto, deberá de ser creada. Se utilizará y seleccionará otra fuente de información.

6.2 Fiabilidad

La fiabilidad del prototipo es una parte importante; por lo tanto, se necesitaron mecanismos de control que ubiquen la respuesta correcta dentro de la fuente de información, así como el tiempo de respuesta del prototipo. Estos mecanismos garantizan que todas las escrituras en un solo documento se lleven a cabo en su totalidad y que los clientes nunca reciban una vista incoherente de los datos.

El prototipo usa bloqueo de granularidad múltiple que permite que las operaciones se bloqueen a nivel global, de base de datos o de colección.

Un punto importante en el prototipo es que se definieron acciones para crear, actualizar, eliminar y consultar los documentos en la base de datos del prototipo. Esto ayudará a que el mecanismo de acceso a la información sea coherente y los datos sean encontrados de una manera eficaz y eficiente, con la menor cantidad de recursos y desplegando los datos correctos a los usuarios.

Para agregar nuevos documentos a una colección, se debe verificar si la colección no existe actualmente; entonces, las operaciones de inserción crearán la colección.

El prototipo proporciona los siguientes métodos para insertar documentos en una colección:

- db.collection.insertOne()
- db.collection.insertMany()

Las operaciones de inserción se dirigen a una sola colección. Todas las operaciones de escritura son atómicas a nivel de un solo documento.

Las operaciones de lectura recuperan documentos de una colección; es decir, realizan consultas a una colección de documentos. El prototipo proporciona los siguientes métodos para leer documentos de una colección:

- db.collection.find()

Se puede especificar filtros de consulta o criterios que identifiquen los documentos que serán devueltos.

Las operaciones de actualización modifican los documentos existentes en una colección. El prototipo proporciona los siguientes métodos para actualizar documentos de una colección:

- db.collection.updateOne()
- db.collection.updateMany()
- db.collection.replaceOne()

Las operaciones de actualización se dirigen a una sola colección. Todas las operaciones de escritura son atómicas a nivel de un solo documento.

También se puede especificar criterios o filtros que identifiquen los documentos para actualizar. Estos filtros usan la misma sintaxis que las operaciones de lectura.

Para eliminar documentos de una colección, existen los siguientes métodos:

- db.collection.deleteOne()
- db.collection.deleteMany()

Las operaciones de eliminación se dirigen a una sola colección . Todas las operaciones de escritura son atómicas a nivel de un solo documento.

También se puede especificar criterios o filtros que identifiquen los documentos para eliminar. Estos filtros usan la misma sintaxis que las operaciones de lectura.

6.3 Rapidez

Las pruebas de rendimiento se basan en la forma en que se utilizan los índices en la aplicación a nivel de base de datos.

Los índices admiten la ejecución eficiente de consultas. Sin índices, se debería realizar un escaneo de colecciones; es decir, escanear cada documento en una colección, para seleccionar los que coincidan con la declaración de consulta. Si existe un índice apropiado para una consulta, este puede usarse para

limitar la cantidad de documentos que debe inspeccionar. Este es el verdadero valor que da el mejor rendimiento al prototipo.

Los índices devuelven estructuras de datos especiales que almacenan una pequeña porción en una forma fácil de recorrer. El índice almacena el valor de un campo específico o conjunto de campos, ordenados por el valor del mismo. La base de datos puede devolver resultados ordenados mediante el uso del orden en el índice.

Para que el prototipo funcione de forma eficiente y el rendimiento sea óptimo se definen índices a nivel de colección y se admiten índices en cualquier campo o subcampo de los documentos en una colección.

6.4 Impactos económicos, tecnológicos y sociales

Los impactos son los siguientes:

6.4.1 Impacto económico

El impacto económico que atrae a la sociedad e impacta en la misma en el desarrollo de un proyecto de inteligencia artificial es fuerte y atractivo, dado que la mayor parte de industrias, organizaciones e instituciones lucrativas y no lucrativas tratan de evitar gastos y costos innecesarios en todo sentido. Este proyecto impactaría en la reducción de gastos salariales, pasivo laboral, bono 14 y aguinaldo, ya que el sistema sustituiría el recurso humano.

6.4.2 Impacto tecnológico

El campo tecnológico de la inteligencia artificial y del prototipo propuesto es amplio, ya que es utilizable en cualquier tipo de institución, organización o industria que desea tecnificar su servicio al cliente o su mesa de ayuda. El campo se abre a múltiples plataformas y tipos de servicio, ya que la herramienta propuesta es configurable para cualquier área de negocio que deseen. Solo se debe ingresar la información correcta de manera que pueda ser localizable por el usuario final.

6.4.3 Impacto social

El impacto social de diseñar, modelar e implementar un sistema basado en una arquitectura de preguntas y respuestas centrado en el concepto de la recuperación de la información es importante ya que en la actualidad muchos sistemas inteligentes son utilizados y proveen información inmediata en cualquier momento y lugar. Esto crea un mayor impacto social en los usuarios, ya que apoya el concepto de autoservicio en donde los usuarios son capaces de consultar la información para satisfacer su necesidad de información, sin acudir a un lugar físico en donde deban esperar para informarse y puedan hacerlo desde cualquier lugar y en cualquier momento.

CONCLUSIONES

1. Se desarrolló el prototipo de un sistema experto más eficiente basado en una arquitectura centrada en la recuperación de información con la inteligencia artificial.

2. Los procesos de análisis y clasificación de la información utilizados en esta investigación logran un procedimiento adecuado para la actualización de la fuente de información de un sistema experto.

3. Se analizaron los componentes arquitectónicos para resolver preguntas complejas en un sistema QA dirigido por la respuesta y no por la pregunta. La arquitectura basada en estos componentes se analizó en base a preguntas complejas y la búsqueda eficiente de documentos que lleva a la respuesta con mayor precisión.

4. El procedimiento eficaz para la búsqueda de información está definido por la indexación, recuperación de documentos, selección de pasajes relevantes y la extracción de la respuesta.

5. Los componentes de análisis de la pregunta, RI de los documentos relevantes y la extracción de la respuesta son parte de la arquitectura centrada en la recuperación de la información para resolver preguntas simples y complejas.

RECOMENDACIONES

1. Determinar la arquitectura de despliegue de servidores y dimensionar los servidores que alojen la arquitectura propuesta. Este dimensionamiento se deberá analizar en base a las necesidades del negocio y a la cantidad de usuarios que podrán utilizar dicha arquitectura.

2. Implementar un prototipo de la arquitectura que permita realizar las pruebas del agente virtual en un ambiente controlado de pruebas.

3. Continuar descubriendo variables que permitan establecer parámetros efectividad, actualización y recuperación de la información en un sistema experto basado en una arquitectura centrada en la recuperación de la información e inteligencia artificial.

4. Analizar los métodos y procedimientos adecuados en la rama de la inteligencia artificial para enriquecer y mejorar el prototipo y, de esa manera, implementar y ampliar el presente trabajo para futuras investigaciones.

5. Optimizar constantemente los procesos de actualización de la información, especialmente del proceso de monitoreo, actualización y optimización de la fuente, ya que constantemente existen nuevos procesos en las áreas de negocio que se necesitan actualizar, por ejemplo, nuevos cursos, nuevas y horarios, entre otros.

6. Implementación de un componente *speech to text* que pueda resolver preguntas por medio del traslado de voz a texto.

REFERENCIAS BIBLIOGRÁFICAS

1. Bäck, T. (1996). *Evolutionary Algorithms in Theory and Practice*. Estados Unidos: Oxford University Press.

2. Bordogna, G. P. (2001). *An Ordinal Information Retrieval Model.* Estados Unidos: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems.

3. Boughanem, M. C. (2002). *On Using Genetic Algorithms for Multimodal Relevance Optimization in Information Retrieval.* Estados Unidos: Journal of the American Society for Information Science and Technology.

4. Fellbaum, C. (1998). *Word- Net: an electronic lexical database. Language, Speech, and Communication.* Estados Unidos: MIT Press, Cambrige, MA.

5. Jarvinen, T. y. (1997). *A dependency parser for english. Informe Técnico TR-1, Department of Linguistics. Informe Técnico TR-1*. Estados Unidos: Department of Linguistics, University of Helsinki, Helsinki.

6. Kalman, J. A. (2001). *Automated Reasoning with OTTER.* Estados Unidos: Rinton Press, Paramus, NJ.

7. Martínez-Barco, P., Vicedo, J. L., Saquete, E., & Tomás, D. (2007). *Sistemas de Pregunta-Respuesta. Grupo de Procesamiento del*

Lenguaje y Sistemas de Información. España: Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante. Recuperado de https://rua.ua.es.

8. Moldovan, D. H. (2000). *The Structure and Performance of an Open-Domain Question Answering System. In Proceedings of the Conference of the Association for Computational Linguistics.* Estados Unidos: En Voorhees y Harman.

9. Molla, D. (2003). *Hacia el Uso de la Información Sintáctica y Semántica en los Sistemas de Búsqueda de Respuestas.* Australia: Centre for Language Technology Division of Information and Communication.

10. Morales González, V. (2012). *APRENDIZAJE EN MAQUINAS CON INTELIGENCIA ARTIFICIAL*. (Tesis de licenciatura). Universidad de Buenaventura CALI, España.

11. Rodríguez, M. L. (2005). *Modelos de Recuperación de la Información basados en Información Lingüística Difusa y Algoritmos Evolutivos. Mejorando la Representación de las Necesidades de Información.* (Tesis doctoral) Universidad de Granada, España.

12. Sleator, D. D. (1993). *Parsing English with a link grammar. En Proc. Third International Workshop on Parsing Technologies.* Estados Unidos: En Proc. Third International Workshop on Parsing Technologies.

13. Tomás, D. (2010). *Sistemas de Clasificación de Preguntas Basados en Corpus para la Búsqueda de Respuestas.* España: Depto. de Lenguajes y Sistemas Informáticos - Universidad de Alicante.

14. Velasco, G. (2014). *Modelo basado en ténicas de procesamiento de lenguaje natural para extraer y anotar información de publicaciones científicas.* (Tesis doctoral). Universidad Politécnica de Madrid, España.

15. Vicente-Díez, M. T., Paloma, M., González, Á., & Fernández, J. L. (2009). *Application of temporal information extraction techniques to question.* España: Universidad Carlos III de Madrid.

16. Voorhees, E. M. (2001). *The Tenth Text REtrieval Conference.* Estados Unidos: National Institute of Standars and Technology Gaithersburg, MD 20899.