Universidad Nacional del Altiplano

FACULTAD DE INGENIERÍA MECÁNICA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

"RECUPERACIÓN DE LA INFORMACIÓN EMPLEANDO EL MODELO DE ESPACIO VECTORIAL EN LA GESTIÓN DOCUMENTARIA PARA LA UNIDAD DE RESOLUCIONES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO - PUNO"

PRESENTADO POR:

OSCAR RUBÉN HUACANI MAMANI

PARA OPTAR EL TÍTULO PROFESIONAL DE:
INGENIERO DE SISTEMAS

PUNO – PERU

2017

UNIVERSIDAD NACIONAL DEL ALTIPLANO

FACULTAD DE INGENIERÍA MECÁNICA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

"RECUPERACIÓN DE LA INFORMACIÓN EMPLEANDO EL MODELO DE ESPACIO VECTORIAL EN LA GESTIÓN DOCUMENTARIA PARA LA UNIDAD DE RESOLUCIONES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO - PUNO"

TESIS

PRESENTADO POR:

OSCAR RUBÉN HUACANI MAMANI

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

PUNO - PERÚ

2017

Universidad Nacional del Altiplano

FACULTAD DE INGENIERÍA MECÁNICA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

"RECUPERACION DE LA INFORMACION EMPLEANDO EL MODELO DE ESPACIO VECTORIAL EN LA GESTIÓN DOCUMENTARIA PARA LA UNIDAD DE RESOLUCIONES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO - PUNO"

TESIS PRESENTADA POR:
OSCAR RUBÉN HUACANI MAMANI

PARA OPTAR EL TÍTULO PROFESIONAL DE: INGENIERO DE SISTEMAS

APROBADA POR EL JURADO REVISOR CONFORMADO POR:

PRESIDENTE :

M.Sc. EDELERE FLORES VELÁSQUEZ

PRIMER MIEMBRO :

M.Sc. Ing. WILLIAM EUSEBIO ARCAYA COAQUIRA

SEGUNDO MIEMBRO :

M.Sc. Ing. PABLO CESAR TAPIA CATACORA

DIRECTOR :

M.Sc. Ing. ELMER COYLA IDME

Área: Sistemas de información
Tema: Sistema Administrativo y de gestión

PUNO – PERÚ

2017

AGRADECIMIENTO

Agradecer a Dios en primer lugar por la vida concedida, un agradecimiento a las autoridades universitarias que conllevaron y conllevan a la institución como pionera y líder en el ámbito de la región del sur peruano; a los compañeros de estudios por haber aprendido a convivir en los claustros universitarios y asimilar de ellos esas agradables experiencias como estudiante; a todos los docentes de la Escuela Profesional de Ingeniería de Sistemas, por la sapiencia que inculcan a las generaciones de estudiantes, del cual fuimos parte, que con un anhelo concretizan sus sueños en realidad de ser profesionales. A los compañeros de la Oficina de Tecnología Informática y Telecomunicaciones de la UNA por haberme colaborado en ese pequeño grano de arena.

Un eterno agradecimiento a mis padres, por haberme brindado todo el apoyo incondicional y haberme sabido entender en los momentos malos y haberme dado la complacencia en los mejores momentos.

Agradecer a todo el Personal Administrativo de la Unidad de Resoluciones de la Universidad Nacional del Altiplano, por haberme facilitado la información objetiva concerniente a todo el proceso de elaboración de resoluciones.

DEDICATORIA

A mis padres Juan y Justina por haberme apoyado en el transcurso de mis estudios porque fueron el soporte, y fueron los que encaminaron y enrumbaron mi vida para ahora verme convertido en una persona de bien.

A mis hermanos Elizabeth y Hugo, por haberme entendido y haberme dado su tiempo y haber dado a conocer que si se logran los objetivos es con perseverancia y voluntad.

A mi esposa Luzma y mi hija Angie por darme el apoyo, los consejos sobre todo por hacer que persevere en el logro de mis objetivos.

CONTENIDO

ÍNDICE DE CUADROS

ÍNDICE DE GRÁFICOS

ÍNDICE DE ECUACIONES

RESUMEN

Con el presente trabajo lo que pretendemos es diseñar un primer modelo de un sistema que pueda mejorar el proceso de la generación de resoluciones en la Oficina de Resoluciones. La arquitectura soportada es la de cliente – servidor, esta tecnología permitirá que los interesados de la documentación puedan acceder de manera oportuna vía Web a la obtención de las resoluciones, desde el lugar donde se encuentren, desechándose el riesgo, el costo o la pérdida de tiempo. Para la elaboración del modelo de sistema, se ha realizado el análisis respectivo de la situación real y de los problemas que se presentan en la Unidad de Resoluciones durante el proceso de la elaboración de resoluciones; luego se procedió con el diseño conceptual y posterior diseño de lógico de la base de datos, al mismo tiempo se utilizó el lenguaje UML (Lenguaje Unificado de Modelado) y los Casos de Uso para entender de manera lógica la situación del problema. Por otro lado se hizo uso de un paquete de software libre el mismo que contiene el sistema de gestión de base de datos MySQL, el servidor Web Apache y los interpretes para lenguaje de script, herramienta exclusiva para el desarrollo del software y para la búsqueda de información se utilizó el Zend Framework el mismo que soporta la interfaz de programación de aplicaciones *Lucene*. Este trabajo sea el reflejo para que la entidad pueda integrar la información especialmente administrativa en un sistema documental, donde puedan acceder los interesados, y que sirva como instrumento para la mejor toma de decisiones y la mejora de la gestión administrativa.

Palabra Clave: Gestión Documentaria, Búsqueda, Modelo Espacio Vectorial

ABSTRACT

With the present work we intend to design a first model of a system that can improve the process of generating resolutions in the Bureau of Resolutions. The supported architecture is client - server, this technology will allow documentation stakeholders to access in a timely manner via the Web to obtain resolutions, from wherever they are, discarding the risk, cost or loss of weather. For the elaboration of the system model, the respective analysis of the real situation and of the problems presented in the Resolutions Unit during the process of the elaboration of resolutions has been carried out; Then proceeded with the conceptual design and subsequent design of the logical database, at the same time used the UML (Unified Modeling Language) and Use Cases to logically understand the problem situation. On the other hand was made use of a package of free software the same that contains the database management system MySQL, the Apache Web server and the interpreters for scripting language, exclusive tool for the development of the software and for the search of Information was used by the Zend Framework itself which supports the Lucene application programming interface. This work is the reflection so that the entity can integrate the especially administrative information in a documentary system, where they can access the interested parties, and that serves as instrument for the better decision making and the improvement of the administrative management.

Keyword: Document Management, Search, Model Vector Space

INTRODUCCIÓN

El presente trabajo de investigación se enfoca en un ambiente teórico –práctico, donde se emplea y aplica tecnologías Web el mismo que sirve para mejorar los procesos del sistema real denominado Unidad de Resoluciones de la Universidad Nacional del Altiplano.

En el contexto de la actual trabajo, indicamos que la problemática es el manejo documentario y la forma de mejorar los procesos mediante la automatización, el que actualmente trae consigo desventajas y contratiempos en la elaboración, búsqueda y distribución de la documentación (resoluciones rectorales), al mismo tiempo que existe inseguridad en el resguardo de la documentación propenso a daños, pérdidas o hurtos o simplemente susceptibles a modificaciones ilegales por terceros.

En el Capítulo I, se presenta el Problema de Investigación, en este describimos el por qué y la razón del problema como la justificación; basándonos en los objetivos que se pretende alcanzar al final de la investigación.

En el Capítulo II, damos a conocer el soporte teórico y conceptual del trabajo en los que se menciona el Marco Teórico destacándose conceptos relevantes al entorno de la investigación y a los objetivos establecidos. Por Otro lado se ha considerado el glosario de términos los mismos que respaldan al marco teórico; también se ha considerado dentro del capítulo la hipótesis de investigación como la operacionalización de variables.

En el Capítulo III, se tiene el Diseño Metodológico de la investigación, donde se considera el tipo de investigación, el diseño del problema de investigación, la población y muestra, dándose la cantidad de documentos digitalizados como

población y la cantidad de documentos elaborados dentro de un año como la muestra; también está la recolección de información, uso de técnicas e instrumentos como la entrevista, encuestas y otros. Para el desarrollo del sistema se tomó en cuenta la metodología UML, diagramas de caso de uso, modelado de datos, implementación SQL, diagramas de secuencias.

En el Capítulo IV, tenemos el Análisis e Interpretación de los Resultados, para ello nos valemos del marco teórico, del diseño metodológico y se tiene el contraste entre la investigación realizada en campo con la aplicación teórica en función a la hipótesis previamente formulada.

Finalmente se ha considerado las conclusiones y sugerencias a las que se ha arribado en el trabajo de investigación, concluyéndose que la implementación del Sistema de Gestión Documental, considerándose el proceso de elaboración, almacenamiento, búsqueda y distribución de la información será más eficiente y oportuna empleándose el Modelo de Espacio Vectorial. La sugerencia que se hace, es que no basta solo con dársele un equipo informático al personal que hace la labor administrativa, si no también formarlo con conocimiento y manejo de las TIC´s y que puedan desempeñar una labor racionada y no simplemente que unos pocos administrados sean los privilegiados en el manejo de los sistemas, esto por un lado y por el otro lado a los usuarios externos facilitarles el acceso a la información de acuerdo a la Ley Nº 27806 Ley de Transparencia

CAPITULO I

PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN

1.1 DESCRIPCIÓN DEL PROBLEMA DE INVESTIGACIÓN

En la actualidad, todavía no se ha acostumbrado a trabajar sin papeles, por tanto, la gestión de los documentos en papel no debe olvidarse cuando se propone un sistema de Gestión Electrónico, aunque no suene excesivamente "moderno". En la mayoría de las instituciones públicas, la gestión de los documentos en papel es un problema, no sólo por el espacio que ocupan, sino por la facilidad con que se producen las copias de los mismos y porque, de alguna manera, se han perdido los principios básicos de archivo. Además, la convivencia con los documentos electrónicos, lejos de disminuir su número o su importancia, la ha aumentado. La mayor parte de las personas que utilizan las aplicaciones o herramientas ofimáticas en el entorno actual; consideran que para guardar o archivar un documento electrónico hay que imprimirlo, lo que va, unido a las facilidades de impresión, usualmente causa un gran número de documentos en papel.

Aunque la validez legal de los documentos electrónicos es una realidad, todavía nos queda mucho tiempo en que determinados documentos, especialmente dentro de la administración pública, tendrán que conservarse en papel.

La Unidad de Resoluciones de la Universidad Nacional del Altiplano dependiente de la Oficina de Secretaria General, cuya función principal es la de elaborar Resoluciones de índole Rectoral, provenientes de Facultades, Unidades Administrativas, Consejo Universitario, Asamblea Universitaria, de personas

naturales y jurídicas, se ve en la necesidad de mejorar la elaboración, resguardo distribución y de búsqueda de manera digital, y que en lo posterior pueda facilitar realizar consultas mucho más rápidas de lo que se hace actualmente, al mismo tiempo para salvaguardar la gran cantidad de información. Esta documentación de resoluciones se vienen acumulando desde hace más de cuatro décadas, estos documentos están en formato papel, clasificados en archiveros los mismos que son susceptibles de deterioro, hurto o a cualquier otro tipo de accidente o incidente que puede ocurrir, incluso pudiéndose poner intereses personales sobre los intereses institucionales.

Las Facultades, las Oficinas (entes internos), personas naturales o jurídicas (entes externos) solicitan la emisión de resolución, estos presentan a la Unidad de Tramite Documentario o directamente a Rectorado con la documentación y sustento respectivo, es en esta instancia donde se verifica el sustento, si este es argumentado y aceptado debidamente pasa a la Oficina de Secretaria General donde es registrado en un libro de registro de ingreso de documentos.

Estando el documento sustentado en la Unidad de Resoluciones, este es verificado por el Jefe de la Unidad y distribuida de manera verbal al personal de acuerdo al tipo de resolución para su elaboración, al mismo tiempo el jefe es quien también elabora las resoluciones más importantes; una vez elaborado, este se registra en el Libro de Registros de Resoluciones de manera correlativa y secuencialmente, es el jefe de la Unidad quien asigna el número a la Resolución, los datos que se ingresan en el libro son: número de resolución, fecha de ingreso a la Unidad, documentos precedentes a la resolución, carácter de la resolución; posteriormente este pasa a la Oficina de Secretaria General

para la firma respectiva por el Jefe de la Oficina y luego es derivada a Rectorado para que se le dé, el visto bueno y la firma respectiva por parte de la autoridad.

En el transcurso del proceso de elaboración del documento, este se guarda en el disco duro del equipo el mismo que está propenso a la pérdida del documento por un ataque malicioso o de terceros, puesto que no se realizan la copias de respaldo o backups de los archivos; ni mucho menos escaneos de las resoluciones ya elaboradas, luego los documentos son impresos en papel y estos son archivados en archiveros que están clasificados por años, el almacenamiento del documento no es el más seguro, pudiéndose en algunos casos deteriorar y/o extraviar, al mismo tiempo como estos son únicos, no existe copia de respaldo.

Luego de que el documento este oleado y sacramentado, parte de este, fundamentalmente la parte resolutiva y los nombres implicados; son considerados e ingresados a una base de datos. El problema trasciende cuando los documentos no siempre llegan a tiempo o de inmediato a las manos de los interesados trayendo consigo perjuicios en algunos casos en la toma de decisiones o molestias a quienes va dirigido los documentos, más aún si los interesados viven fuera de la ciudad de Puno.

El problema en la Unidad de Resoluciones se manifiesta desde el ingreso de datos, su procesamiento o elaboración, las salidas plasmadas en “Resoluciones y Certificaciones” fundamentalmente la difusión y el medio adecuado de propagar a todos los entes involucrados o implicados con cada uno estos documentos.

¿Cómo incide la recuperación de la información empleando el modelo de espacio Vectorial en la gestión documentaria para la Unidad de Resoluciones de la UNA Puno?

1.2 JUSTIFICACIÓN DEL PROBLEMA

En la actualidad, todavía no se han acostumbrado a trabajar sin papeles, por tanto, la gestión de los documentos en papel no debe olvidarse cuando se propone un sistema de gestión electrónico, aunque no suene excesivamente "moderno". En la mayoría de las instituciones públicas, la gestión de los documentos en papeles es un problema, no sólo por el espacio que ocupan, sino por la facilidad con que se producen las copias de los mismos. La mayor parte de las personas que utilizan las aplicaciones o herramientas ofimáticas en el entorno actual; consideran que para guardar o archivar un documento electrónico hay que imprimirlo, lo que va, unido a las facilidades de impresión, usualmente causa un gran número de documentos en papel.

Por otro lado, se desconoce el ciclo de vida que tiene un documento, a veces estos que ya no tienen validez los mismos que ocupan cantidades de espacio.

A partir de cierto volumen de texto escrito se hace imprescindible un sistema organizativo que posibilite la localización de la información que se precise en cualquier momento. Esta necesidad ha estado cubierta por técnicas que no han variado en 200 años, básicamente hasta la disponibilidad de ordenadores cada vez más potentes, dispositivos de almacenamiento más rápidos y de mayor capacidad y las redes de ancho de banda han producido una explosión de la información que no pueden ser afrontada sin un amplio conjunto de nuevas

técnicas de almacenamiento, acceso, interrogación y manipulación de esa información (Lizcano Luis Ignacio, 2001).

El desarrollo de los sistemas automatizados de recuperación de información se inició con el objetivo de facilitar el manejo de la enorme cantidad de literatura científica surgida de los años 40 del siglo pasado (Mañas 1994). No ha quedado restringida a este campo sino que se ha extendido a otras áreas: cualquier disciplina que base su trabajo en la utilización de documentos puede beneficiarse de las técnicas de recuperación de información textual. En los últimos 30 años se han desarrollado estructuras de datos eficientes para almacenamiento de índices, sofisticados algoritmos de interrogación, métodos de compresión; según Frakes y Beaza-Yates (1992) e incluso hardware específico, más recientemente se han aplicado técnicas de procesamiento de lenguaje natural en aspectos tales como la extracción de información, formulación de interrogaciones amigables y la generación de respuestas (Rijsbergen, 1979). La búsqueda de cadenas tanto exacta como aproximada, los métodos de construcción y manipulación de diccionarios (Baeza-Yates y Ribeiro, 1999).

El problema actual que se afronta en la Unidad de Resoluciones es que los documentos elaborados no llegan a tiempo al usuario o a veces nunca les llega, este tiene que apersonarse a la unidad e indagar por el documento, trayendo como consecuencia pérdida de tiempo y costos innecesarios para los usuarios
Al realizar el proyecto se dinamizará el proceso de elaboración de resoluciones desde que ingresa el documento petitorio, hasta que este llegue al usuario para el cual se resuelve la resolución en un plazo determinado menor al actual.

Lo que se pretende es que los usuarios accedan a los documentos (resoluciones) desde el lugar donde estén, para ello se elaborará un sistema de gestión y administración documentaria, bajo plataforma cliente servidor, y tenga un sistema de búsqueda que permita dinamizar la búsqueda del documento por cualquier referencia.

Con el proyecto de investigación lo que se pretende es conocer el conjunto de normas y técnicas para administrar el flujo de documentos en la unidad de resoluciones y su trascendencia en la entidad, al mismo tiempo conocer cómo se debe representar un documento para su recuperación utilizando técnicas como filtrado, recuperación, indexado, representación de la información en vectores y cálculo de relevancia de información; asegurar la conservación de los documentos. Estos conocimientos permitirán que se desarrolle un prototipo de sistematización de los documentos y la aplicación de la recuperación de la información en el sistema.

Con el desarrollo del proyecto de investigación se pretende beneficiar al personal que labora en la unidad de resoluciones fundamentalmente a todos los usuarios que hacen uso de esta unidad para conocer en tiempo real la situación de su documento y su implicancia.

1.3 OBJETIVOS DE INVESTIGACIÓN

1.3.1 OBJETIVO GENERAL

Determinar la incidencia de la recuperación de la información empleando el Modelo de Espacio Vectorial en la gestión documentaria, para la Unidad de Resoluciones de la UNA - PUNO.

1.3.2 OBJETIVOS ESPECÍFICOS

- Desarrollar el prototipo de Gestión documentaria en la administración del flujo de documentos para la Unidad de Resoluciones.
- Analizar y desarrollar el sistema de Recuperación de información empleando el Modelo de Espacio vectorial para la búsqueda de información.

CAPITULO II

MARCO TEORICO

2.1 ANTECEDENTES DE INVESTIGACIÓN

La recuperación de la información es un área de investigación que se viene estudiando aproximadamente desde 1945, con la propuesta de Vanne var Bush accesibilidad a la información, usa tecnología para organizar y recuperar los libros, las publicaciones y las notas. Bush desarrollo su sistema denominado MEMEX (**Mem**ory - ind**ex**), dispositivo para almacenamiento y búsqueda de información en fichas, donde presento una interfaz sencilla, basada enteramente en cuadros de páginas etiquetadas con códigos.

En 1983 Salton G., propuso el concepto de la recuperación de la información moderna, es entendida como el área de conocimiento que concierne a la representación, el almacenamiento, el tratamiento y el acceso automatizados o a sus sustitutos, por otro lado, también se le ha añadido el proceso de tratamiento (pensando principalmente en la indización automática) y el carácter automático en todos ellos.

En 1996 Ellis D., realiza un enfoque del progreso y los problemas en la recuperación de la información, esto es la recuperación de todos los documentos relevantes y al mismo tiempo rechazar todos los documentos irrelevantes ante la formulación de una consulta por parte del usuario; desde el modelo booleano (por tratarse de una de las vías más simples desde un punto de vista teórico), hasta la aplicación de técnicas de inteligencia artificial.

En 1997 Tramillas, ha generado un estudio de nuevas disciplinas como son las Ciencias de la Información y la Documentación y sus aplicaciones particulares como la Biblioteconomía, Documentación y la más reciente la documentación automática (Documática) donde el problema central para la Recuperación de Información es la representación de los documentos, determinándolos relevantes de acuerdo a la necesidad de información planteada.

En el 2001, Zazo Ángel, Figueroa Carlos G., Berrocal José Luis; del Departamento de Informática y Automática de la Universidad de Salamanca Participan en el Taller CLEF-2001 con la investigación, Recuperación de Información utilizando el Modelo Vectorial. En este documento se describe el proceso seguido para la construcción del sistema de recuperación de información. El sistema utiliza el conocido modelo vectorial. Se analiza el problema de recuperación de información, el modelo vectorial y la evaluación sobre la colección de pruebas. Seguidamente se describe el procesado léxico realizado sobre el contenido de documentos y consultas. Se finaliza con los resultados obtenidos en los experimentos y conclusiones.

En el 2002 Martínez Méndez Francisco Javier, desarrolla de un modelo para la evaluación de la recuperación de información en internet, primeramente identificando las variables, posteriormente llega aquellos Sistemas de Recuperación de Información se sintetizan en cuatro grupos: (1) las capacidades de búsqueda y la interface del sistema, (2) la precisión y la exhaustividad (calculadas normalmente en términos de promedios), (3) el tamaño del índice del

motor de búsqueda y (4) el grado de solapamiento y el número de error es detectados en los motores de búsqueda.

En el 2004, Docentes de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos, realizarón un estudio y evaluación de los sistemas de recuperación de información. El trabajo se desarrolla en el marco del proyecto de investigación “Sistema de Recuperación de Información”, cuyo objetivo es diseñar un SRI para la biblioteca de la Facultad de Ingeniería de Sistemas e Informática, y posteriormente para la Biblioteca Central de la universidad. En el trabajo se han desarrollado básicamente; a) un estudio detallado de las principales técnicas, modelos y arquitecturas, así como de los criterios de la evaluación de estos sistemas, b) análisis de las técnicas de indexación, necesarios para el almacenamiento de los documentos; c) el trabajo ha permitido la selección de cuatro aplicaciones de SRI para su análisis y evaluación: Karpanta, SISA, Dialog, SMART.

En el 2006 López Herrera Antonio, desarrolla la mejora de los Sistemas de Recuperación de Información, usando técnicas de modelado lingüístico difuso como proceso de evaluación de consultas que realizan dichos sistemas, donde aplica la representación de información difusa, el modelo lingüístico difuso, al mismo tiempo propone nuevos mecanismos de evaluación de consultas.

En el 2009, Mendoza Mendoza Roberto C, implanta una solución a partir de un modelo de agentes móviles que realiza una recuperación semántica de información, basado en meta datos Dublín Core con descripciones de un dominio

específico, de acuerdo a la consulta del usuario en un sistema de biblioteca digital distribuido.

En 1997 Romero Flores Robert A., inicia con las primeras investigaciones por automatizar los sistemas en la UNA – PUNO, que hasta entonces eran manuales y mecánicos., (egresado de la CCPP de Ingeniería de Sistemas, UNA-PUNO) sustenta la tesis "Automatización del Sistema de Administración Académica de la UNA - PUNO". Este ilustra la importancia que tiene la administración de la información académica de la UNA-PUNO, así como el diagnóstico situacional, en base al desempeño de los sistemas existentes en ese entonces, identificando situaciones no óptimas en el tratamiento de la información, como alternativa se propone un sistema automatizado. Como aporte en la investigación pretende contribuir a mejorar el conocimiento del Sistema de Administración Académica y la problemática que implica su administración, evaluando el impacto el impacto que tenga sobre los usuarios. Así también determinar cuán importante es el apoyo que brinda la tecnología como solución a situaciones "mecánicos - repetitivos" que muchas veces tiene que efectuar el personal administrativo, así también la precisión en el desarrollo de procedimientos.

En el 2014 Ramos Pacara Henry S., realiza la tesis "Prototipo del sistema de información para optimizar los procesos de manejo de información en la Unidad de Pensiones y Liquidaciones de la UNA-PUNO 2012", siendo el objetivo, aplicar la tecnología computacional e informática, para que los trabajadores de la universidad puedan obtener los documentos solicitados, como es la información precisa de su situación laboral y económica, otro aspecto importante que

considera aparte de lo mencionado es que con esta investigación puede darse el inicio de un manejo tecnológico más eficiente de las actividades administrativas en la Universidad.

En el 2014 Vilca Quisocala Jonatan V. y Alferez Vilca Romel A.; sustentan la tesis “Aplicación Web de Trámite Documentario para la mejora y agilización de Trámite en el Edificio Administrativo de la Universidad Nacional del Altiplano”, donde el objetivo principal de la investigación es determinar, la mejora y la accesibilidad a la información de los trámites documentarios, como también el seguimiento a los documentos, concluyendo que la aplicación Web de Trámite documentario si mejora y agiliza la accesibilidad a la información.

2.2 SUSTENTO TEÓRICO

2.2.1 GESTIÓN DOCUMENTAL

Conjunto de actividades administrativas y técnicas tendientes a la planificación, manejo y organización de la documentación producida y recibida por las entidades, desde su origen hasta su destino final, con el objeto de facilitar su utilización y conservación. Para agilizar el proceso de la gestión documental dentro de una organización nos valemos de un sistema de gestión documental SGD, la misma que debe asegurar la asignación de recursos económicos, humanos y físicos que permitan el desarrollo armónico de las distintas fases archivísticas bajo criterios operativos.

Con un Sistema de Gestión Documental – SGD, según Rhoadas James B. (1989) obtiene:

- ✓ Resaltar la importancia del papel de los documentos y archivos, como lenguaje natural de la administración, para el funcionamiento de la misma, elementos de apoyo decisivos para la transparencia y el control de la gestión y garantía de los derechos individuales y colectivos.
- ✓ La racionalización y la normalización de la documentación desde se producción hasta su destino final.
- ✓ Normalizar la utilización de materiales, soportes y equipos de calidad y que a la vez preserven el cuidado del medio ambiente.
- ✓ Lograr una acertada normalización en los procedimientos de la documentación, mediante la utilización de sistemas eficientes.
- ✓ Facilitar la recuperación de la información en forma rápida y oportuna.
- ✓ Encaminar los archivos para que sean verdaderos centros de información, útiles para la administración e importancia para la cultura.
- ✓ El manejo integral de los documentos y de la información como base para la toma de decisiones y la preservación de la memoria institucional.
- ✓ La evaluación y valoración de la documentación para evitar acumulación innecesaria de información y reducir costos en la producción y conservación de acervo documental.
- ✓ La simplificación de trámites en los procesos administrativos con miras al flujo normal y eficaz de la información.

Dentro del proceso de la gestión documental, las entidades, han centrado su atención y preocupación en aquella documentación que se encuentra almacenada en los llamados Archivos Centrales, puesto que existe desconocimiento y poca importancia sobre el documento y su valor de información que pueda tener para la entidad.

Un SGD permite tener una visión exacta y completa de las políticas, funciones, programas y servicios de una entidad, lo cual se ve reflejado en un sistema institucional de archivos plenamente organizado y definido que garantiza el flujo y disposición de la información en forma ágil y oportuna.

2.2.2 ALTERNATIVAS ESTRATÉGICAS DE UN SISTEMA DE GESTIÓN DE DOCUMENTOS

Según las investigaciones que hace Alberch Fugueras Ramón (2016), propone alternativas de un sistema de gestión de documentos:

- El establecimiento, implementación, mantenimiento y mejora de un Sistema de Gestión de Documentos es un reto de carácter transversal e interdisciplinar, de manera que precisa del trabajo cooperativo de profesionales que son expertos en los temas de su competencia pero que, usualmente, no tienen el hábito de trabajar conjuntamente y de manera concertada.
- Las consideraciones derivadas de las normas ISO de gestión documental es imprescindible la concurrencia de:
  - La alta dirección que debe impulsar el SGD y garantizar su sostenibilidad mediante la asignación de los recursos humanos y económicos adecuados.
  - El servicio de tecnologías en tanto que responsable del modelo tecnológico de la organización.
  - El servicio de Gestión de Documentos y Archivo en tanto que órgano responsable de las políticas encaminadas a planificar, implementar y administrar un SGD.

- Los servicios jurídicos que deben garantizar la formalidad legal de los procesos y procedimientos.
- Los responsables de organización deben modelar los circuitos administrativos e impulsar la reingeniería de procesos
- Los responsables de formación que deben garantizar la difusión de la información necesaria para entender y aplicar las directrices del SGD y el conocimiento preciso de los instrumentos que lo desarrollan.
- El personal, de cualquier nivel, implicado en las tareas de gestión de documentos y que deberá aplicar el modelo a implantar en las acciones y trámites cotidianos y, en especial, mantener documentos precisos y completos de sus actividades, con la obligación de informar sobre su labor.

2.2.3 NORMATIVIDAD PARA LA GESTIÓN DE DOCUMENTOS

En el 2003 Alberch, toma en consideración; que actualmente se cuenta con un entramado normativo que consolida y afirma las acciones de gestión de documentos electrónicos. En concreto, y a manera de síntesis, hace mención de tres bloques normativos esenciales y de obligado cumplimiento.

a) La aparición de normas de gestión de documentos fomentadas desde instancias internacionales, muy especialmente las normas ISO y las directrices emanadas de gobiernos nacionales o federales y el Consejo Internacional de Archivos, como las Moreq2010 en caso de la Unión Europea, las ISO 15489 1 y 2 de gestión de documentos, las normas ISO 30300, 30301 y 30302 de sistemas de gestión para documentos y, finalmente, de las normas 16175 1, 2 y 3 de gestión de documentos en entornos de oficinas.

b) La existencia de otras normas ISO que complementan adecuadamente las citadas directivas de gestión de documentos. Actualmente existen con normas referidas a metadatos, procesos, digitalización (migración y conversión de documentos), preservación digital (Open Archival Information Systems - OAIS), seguridad y gestión de riesgos, evidencias electrónicas y calidad.

c) La generación de una legislación específica para el desarrollo del gobierno electrónico, esencialmente de normativa sobre firma electrónica, libertad de información y transparencia, protección de datos, interoperabilidad y seguridad.

2.2.4 IMPORTANCIA DE LOS DOCUMENTOS

Sin documentos cualquier administración organizada dejaría rápidamente de funcionar. Los documentos y específicamente la información que contienen, son uno de los recursos fundamentales que las instituciones necesitan para poder realizar sus operaciones eficazmente.

Así como una organización podría seguir funcionando con escasos recursos humanos, económicos o materiales, no podría funcionar si no mantuviera sus documentos y fuera posible el acceso a los mismos. En el contexto de las actividades administrativas de las organizaciones los documentos ayudan a:

- Proporcionar una memoria corporativa
- Formular políticas
- Tomar decisiones apropiadas

- Alcanzar más eficiencia, productividad y coherencia
- Cumplir con los requisitos legales y las regulaciones vigentes
- Proteger los intereses de la organización y aquellos de su personal y sus clientes
- Reducir los riesgos relacionados con la falta de pruebas de decisiones y acciones
- Documentar actividades y logros

Los documentos se realizan en una variedad de formas físicas, como unidades documentales simples en papel, expedientes, legajos, mapas, fotografías, microformas y datos en forma electrónica. De hecho, los documentos producidos o mantenidos en un ordenador no tienen forma física y existen solamente en tanto conjuntos lógicos de datos electrónicos, sin embargo, son documentos.

Los documentos son más que los datos que contienen. Los documentos ofrecen datos e informaciones, significado y contexto al estar intrínsecamente ligados a la actividad que documentan y de la que dan testimonio. Sólo los documentos sirven como prueba de la realización de trámites.

2.2.4.1 CARACTERÍSTICAS ESENCIALES DE LOS DOCUMENTOS

Los documentos reflejan las actividades de las que son producto. Debido a que derivan de sucesos reales, representan una imagen *congelada* la que fija una acción en su contexto particular de: función, autoridad, lugar y momento. Es posible establecer ciertas características esenciales.

Los documentos:

- Son estáticos en su forma
- Tienen autoridad
- Son singulares cuando están en su contexto
- Son auténticos

2.2.4.2 CALIDAD DE LOS DOCUMENTOS

Así como es necesario crear y conservar los documentos para proporcionar una prueba documental, esos documentos tienen que tener calidad e integridad suficientes. Para que las instituciones estén en condiciones de realizar sus funciones administrativas eficientemente y puedan ser responsables, tienen que mantener documentos íntegros y fieles. Sin ellos los funcionarios no podrían realizar sus cometidos adecuadamente. La puesta en marcha de ese cometido no podría ser convenida y los derechos financieros, legales y de otra índole de la organización, sus clientes y otras personas afectadas por sus acciones y sus decisiones no podrían ser protegidos.

Para llevar a cabo su propósito de proporcionar una prueba confiable, los documentos deben ser fieles, completos e integrales.

Los documentos, ya sean en papel o en forma electrónica deben ser:

- **Integrales:** deberá producirse un documento para cada transacción administrativa, de la que dará testimonio.
- **Fidelidad:** un documento debe registrar con precisión la transacción para la cual fue creado

- **Adecuado:** un documento debe ser adecuado para los propósitos para los cuales se conserva (es decir, el documento debe contener la información necesaria para dar prueba de la transacción que documenta)
- **Completo y significante:** así como debe contener información suficiente para documentar una transacción, un documento debe incluir información suficiente sobre el contexto en el cual fue creado y usado, sobre su estructura o su forma física y sobre sus vínculos con otros documentos, para que sea posible comprender su contenido
- **Comprensible y utilizable:** debe ser posible extraer del documento la información que contiene y que pretende comunicar; y ser posible usarlo sin que se pierda información
- **Auténtico:** debe ser posible comprobar que el documento es lo que dice ser.
- **Inalterado:** ninguna información del documento habrá sido borrada, alterada o perdida, bien sea intencionalmente o por accidente, una vez que la transacción que dio origen al mismo se ha finalizado (en otras palabras, los documentos se mantendrán seguros y se evitará el acceso a ellos o a su uso, no autorizado).

De esto se desprende que los documentos deben ser integrales y completos, confiables y auténticos, deber ser administrados con sistemas de gestión que los controlan durante su ciclo de vida, desde su creación hasta su disposición final y su tratamiento en los archivos históricos.

2.2.5 LA DOCUMENTACIÓN EN LA UNIVERSIDAD NACIONAL DEL ALTIPLANO

Los archivos están formados por un conjunto de documentos en formato impreso, digital y de diferente fecha de creación producidos por las unidades académicas, unidades administrativas y los órganos de rectorado y vicerrectorado, así también como personas involucradas con el que hacer universitario como las de actividad administrativa, académica y de investigación.

Para efectos de conceptualización de un Sistema de Gestión Documental, se tiene en cuenta diversos procesos que mantienen un orden secuencial como son:

- La producción documental, lo cual comprende los aspectos del origen, creación, diseño de formatos y documentos, conforme al desarrollo de las funciones en la Universidad Nacional del Altiplano.
- Recepción de documentos (solo impresos) se realiza manualmente registrándose estos en el denominado “Libro de Registros”, pasa al ente superior para la decisión que este tiene sobre el asunto del documento.
- El cumplimiento de los documentos, está dada por la lógica del documento es decir que se siempre se tiene que dar de esa manera, también se da por el argumento o sustento que tiene el documento.
- La resolución y distribución del documento, la resolución está en función al argumento establecido previamente y la distribución es la consecuencia del anterior proceso, este último se hace de persona a persona, es decir quien entrega y quien recibe.

2.2.6 CLASIFICACIÓN DE LA DOCUMENTACIÓN EN LA UNA-PUNO

- Documentos de los órganos de gobierno de la universidad
- Documentos relativos al desarrollo normativo, jurídico y legal de la universidad (resoluciones rectorales, resoluciones de asamblea).
- Expedientes académicos de los estudiantes de todas las escuelas profesionales como los de Post Grado.
- Documentos del presupuesto, contabilidad y logística.
- Expediente del personal (docente y administrativo)
- Plan Operativo Anual (POA).
- Plan Estratégico Institucional (PEI).
- Memorias anuales, publicaciones institucionales
- Expedientes de investigación
- Tesis
- Y todos aquellos documentos y expedientes que reflejen las actividades de la universidad y sus miembros.

2.2.7 PROPÓSITO DE LA GESTIÓN DE DOCUMENTOS

En líneas generales, puede decirse que, la Gestión de Documentos se ocupa de todos los procesos por los cuales la información registrada ayuda a las instituciones a satisfacer sus necesidades operativas y administrativas.

La Gestión de Documentos no tiene un fin en sí misma: es un componente decisivo de la actividad administrativa y, en un sentido más amplio, un pilar para el funcionamiento eficiente de las organizaciones, y con el tiempo, de la sociedad en general.

La eficiencia de las instituciones se basa en el acceso oportuno a la información cuando esta es requerida. La Gestión de Documentos es fundamental para la formulación de políticas, la toma de decisiones, las operaciones administrativas y la rendición responsable de la organización. El proceso de la Gestión de Documentos capta pruebas de las transacciones de una organización, documenta sus actividades y sus decisiones, y proporciona un acceso fácil a esas pruebas.

La Gestión de Documentos promueve la agrupación y la distribución de la información, y contribuye al buen uso de los antecedentes y de la experiencia de la organización. La Gestión de Documentos también hace posible que la organización esté en condiciones de controlar el volumen de los documentos que produce, recibe y almacena. Esto no sólo es importante por razones de economía y eficacia ya que el mantenimiento de documentos es costoso; también promueve la eficiencia operativa al mejorar el acceso a la información mediante la baja de los documentos que ya no se necesitan en las transacciones corrientes. Por último, la Gestión de Documentos con programas de evaluación controla el retiro y la disposición de los documentos, una vez que su valor para los propósitos administrativos se ha extinguido.

Tomando en cuenta los conceptos de la administración de documentos mencionados anteriormente, podemos resumir los propósitos de la Gestión de Documentos de la manera siguiente:

- Gestionar los Documentos durante todo su ciclo de vida, comenzando por el diseño de un programa de atención y conservación de documentos

hasta la baja de los documentos o su transferencia y atención en los archivos históricos.

- Proporcionar servicios para satisfacer las necesidades y proteger los intereses de la organización, su personal y sus clientes o usuarios.
- Lograr documentación completa, precisa, confiable y utilizable para satisfacer sus necesidades legales, de regulación, probatorias y de rendición de cuentas.
- Gestionar los documentos como recursos documentales.
- Fomentar la eficiencia y la eficacia mediante prácticas de mantenimiento de documentos bien concebidos.

2.2.7.1 AGRUPACIÓN DE LOS DOCUMENTOS

Para organizar un archivo de oficina es necesario distinguir claramente los distintos grupos de documentos de archivo que a continuación se detallan:

a) Correspondencia

Características:

- Es un tipo de documentación que reciben todas las unidades administrativas de la universidad
- Pueden ser originales o copias
- No está vinculada a ningún procedimiento administrativo
- Aporta información de muy distinto tipo

Organización

- La correspondencia no se debe clasificar y ordenar por entradas y salidas, sino que cada una debe ir unida a su respuesta.

- La correspondencia, tanto emitida como recibida que forma parte de un expediente, se ordena con dicho expediente
- La correspondencia que acompaña a una información importante, facturas, informes u otros documentos, se archiva en función de esta información adjunta.
- La correspondencia que se organiza como tal es la que no forma parte de un expediente ni acompaña a una información importante, es decir que se limita a una información más genérica

b) Libros de registro

El registro es un instrumento jurídico, cuya finalidad es conseguir un sistema de control y de garantía interna y externa de los documentos que se presentan en la administración y de los documentos oficiales que se envían a otros órganos o a particulares. El registro permite certificar la existencia de un documento, aunque este no se haya conservado.

En la Universidad Nacional del Altiplano, existe una Unidad de Archivo Central, y una Oficina de Archivo y Registro Académico, a los que recurrirán las unidades académicas, administrativas, órganos de gobierno, cuando requieran registrar documentos oficiales o con “valor probatorio”.

c) Expedientes

Constituyen la unidad documental básica en todo tipo de archivos.

Se entiende por expediente administrativo una unidad documental formada por un conjunto de documentos generado, orgánica y

funcionalmente, por un ente productor en la resolución de un mismo asunto.

La estructura del expediente debe ser, lógica, coherente y cronológica, ordenándose los documentos de acuerdo con el procedimiento o trámite seguido. Los criterios de formación del expediente administrativo deberán ser uniformes y conocidos por todo el personal involucrado en dicha tarea.

Los tipos de documentos que forma parte de un expediente administrativo podrían ser:

Documentos de Decisión (contiene una declaración de voluntad de un órgano administrativo sobre materias de su competencia	RESOLUCIONES ACUERDOS
Documentos de Transmisión (comunican la existencia de hechos o actos a otras personas, órganos o entidades)	COMUNICACIONES NOTIFICACIONES PUBLICACIONES
Documentos de Constancia (Contienen una declaración de conocimiento de un órgano administrativo, cuya finalidad es la acreditación de actos, hechos o efectos)	ACTAS CERTIFICADOS CERTIFICACIONES ACTO PRESUNTO
Documentos de Juicio (contienen una declaración de juicio de un órgano administrativo persona o entidad pública o privada sobre las cuestiones de hecho o de derecho que sean objeto de procedimiento administrativo)	INFORMES

Tabla Nº 1: Tipos de Documentos
Fuente: Elaboración propia.

d) Series Documentales

Constituyen el segundo nivel de agrupación de documentos en los archivos de gestión y se forman a partir del conjunto ordenado de los expedientes o unidades documentales que se producen de

manera continuada como resultado de una misma actividad o función.

2.2.7.2 BENEFICIOS DE LA GESTIÓN DE DOCUMENTOS DE ARCHIVO

La gestión de documentos de archivo regula las prácticas efectuadas tanto por los responsables de su gestión como por cualquier otra persona que cree o use documentos en el ejercicio de sus actividades. La gestión de documentos de archivo en una organización incluye: (Norma ISO/IEC 15489, para la Gestión de Documental en las Organizaciones).

a) El establecimiento de políticas y normas
b) La asignación de responsabilidades y competencias.
c) La fijación y promulgación de procedimientos y directrices
d) La presentación de una serie de servicios relacionados con su gestión y uso
e) El diseño, la implementación y la administración de sistemas especializados
f) La integración de la gestión de documentos de archivo en los sistemas y procesos de la organización.

Un sistema de gestión de documentos de archivo, se convierte en una fuente de información sobre las actividades de la organización que puede servir de apoyo a posteriores actividades y toma de decisiones, al tiempo que garantiza la asunción de responsabilidades frente a las partes interesadas presentes y futuras. Los documentos de archivo permiten a las organizaciones:

- Realizar sus actividades de una manera ordenada, eficaz y responsable.
- Prestar servicios de un modo coherente y equitativo.

- Respaldar y documentar la creación de políticas y la toma de decisiones a un nivel directivo.
- Proporcionar coherencia, continuidad y productividad a la gestión y a la administración
- Facilitar la ejecución eficaz de las actividades en el seno de la organización.
- Garantizar la continuidad en caso de desastre.
- Cumplir con los requisitos legislativos y normativos, incluidas las actividades archivísticas, de auditoria y las de supervisión.
- Mantener una memoria corporativa personal o colectiva.

2.2.8 RECUPERACIÓN DE INFORMACIÓN RI

La RI es la disciplina encargada de la representación, almacenamiento y la organización de la información y su posterior acceso y recuperación de información que corresponda a las necesidades de un usuario.

La RI tiene sus orígenes en las bibliotecas y centros de documentación en los que se requerían búsquedas bibliográficas de libros y artículos de revista. El objetivo principal de cualquier centro de documentación es satisfacer las necesidades reales y potenciales de información de todos los usuarios, proporcionándoles la información veraz, pertinente, justo a tiempo y al menor costo.

Debido a motivos históricos, los documentos en esos centros se representan utilizando un conjunto de términos índice o palabras clave. Existen fichas (manuales o electrónicas) en las que se rellenan los campos apropiados

con esa información. En esos campos se incluyen datos con el título, autor fecha de publicación, etc., del documento en cuestión. Pero también se incluyen otros términos que dan una indicación de su contenido y que normalmente quedan reflejados en el campo *materia*. Uno o varios especialistas asignan la materia de acuerdo a criterios más o menos subjetivos.

El objetivo de la recuperación de información es, dada una necesidad de información y un conjunto de documentos, ordenar los documentos demás a menos relevantes para esa necesidad y presentarlos al usuario. Para ilustrar el problema nos valemos de la siguiente figura (Rijsbergen, 1979).

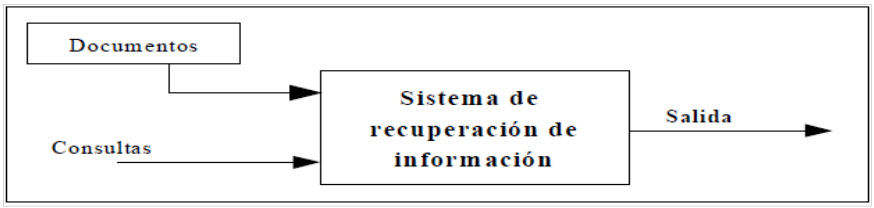


Figura Nº 1: Modelo General de un Sistema de Recuperación de Información.
Fuente: Rijsbergen, 1979, Recuperación de Información

En esta figura se ha supuesto que el sistema de recuperación de información SRI es una caja negra que acepta documentos y consultas, y que obtiene una salida, que es el conjunto de documentos que satisfacen la consulta. El problema principal en este sistema es obtener representaciones homogéneas de documentos y consultas y procesar convenientemente esas representaciones para obtener la salida. De otro lado también es muy importante la evaluación de la salida, para determinar si esta coincide con las necesidades informativas del usuario.

En el proceso de recuperación de información se suelen distinguir las siguientes etapas:

a) Obtener representación de documentos. Generalmente los documentos se representan utilizando un conjunto más o menos grande de términos índice. La elección de dichos términos es el proceso más complicado.
b) Identificar la necesidad informativa del usuario. Se trata de obtener la representación de esa necesidad y plasmarla formalmente en una consulta acorde con el sistema de recuperación.
c) Búsqueda de documentos que satisfagan la consulta. Consiste en comparar las representaciones de documentos y la representación de la necesidad informativa para seleccionar los documentos pertinentes.
d) Obtención de resultados y presentación al usuario.
e) Evaluación de los resultados por parte del usuario.

Un sistema de recuperación de información SRI, permite la recuperación de información, previamente almacenada, por medio de consultas a los documentos contenidos en la base de datos. Esta serie de preguntas se conceptúan como sentencias formales de expresiones de necesidades de información y suelen venir expresadas por medio de un lenguaje de interrogación. Un documento es un objeto de datos, de naturaleza textual generalmente, aunque la evolución tecnológica ha propiciado la profusión de documentos multimedia, incorporándose al texto fotografías, ilustraciones gráficas, video, audio, etc.

Un SRI debe soportar una serie de operaciones básicas sobre documentos almacenados en el mismo, como son: introducción de nuevos documentos, modificación de los documentos almacenados y eliminación de los mismos. Los SRI implementan estas operaciones en formatos muy diversos lo que provoca

una diversidad en lo relacionado con la naturaleza de los mismos. (Lizcano Luis Ignacio, 2001).

Con el incremento del número de documentos en formato electrónico, se hace necesario contar con herramientas informáticas adecuadas para la recuperación de documentos. Las técnicas manuales han demostrado ser ineficaces, básicamente consisten en la elaboración manual de una descripción del contenido temático de cada uno de los documentos, siendo este un trabajo costoso en tiempo, que además tiene abundante inconsistencia (Hooper, 1965).

En estos sistemas cada documento se puede representar por todas sus palabras, tanto nombres, como verbos, adjetivos, adverbios, etc. A pesar de ello, no todos los términos poseen la misma utilidad para describir el contenido de un documento. De hecho, hay términos más importantes que otros, pero no es tarea fácil decidir la importancia de cada término. Desde el punto de vista de la recuperación de la información existen palabras casi vacías de contenido semántico, como los artículos, preposiciones o conjunciones, que parecen poco útiles en el proceso. Sin embargo, algunos estudios actuales en el campo de la lingüística indican que el estudio de los artículos incorpora un nivel semántico adicional al nombre que acompañan y que por tanto deben tenerse en cuenta.

Independientemente de ello, tampoco es útil para las tareas de recuperación de información aquellos términos que se repiten con mucha frecuencia en toda la colección de documentos, pues son términos poco discriminatorios en relación con una consulta dada. En conclusión, en estos sistemas no solo se persigue

encontrar aquellos términos que mejor representen a los documentos, sino además aquellos que permitan diferenciar unos respecto de otros.

2.2.9 VISTA FUNCIONAL DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN

En 1983 Salton, se entiende que la recuperación de información mejora cuando la información procesada son documentos, con el fin de diferenciar a los sistemas encargados de su gestión de otro tipo de sistemas, como los gestores de bases de datos relacionales; piensa que cualquier Sistema de Recuperación de Información puede ser descrito como un conjunto de ítems de información (Docs), un conjunto de peticiones (Reqs) y algún mecanismo (Similar) que determine que item satisfacen las necesidades de información expresadas por el usuario en la petición.

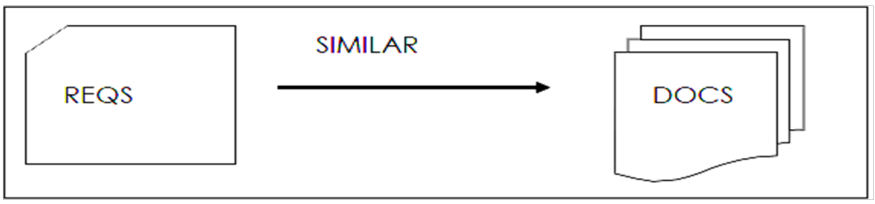


Figura Nº 2: Esquema Simple de un SRI.
Fuente: Salton 1983, Introducción a la Recuperación de Información Moderna.

Según el autor, reconoce que en la práctica este esquema resulta muy simple y precisa ampliación, porque los documentos suelen convertirse inicialmente a un formato especial, por medio del uso de una clasificación o de un sistema de indización que denomina *Lang.*

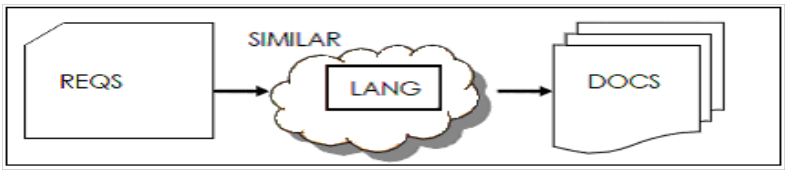


Figura Nº 3: Esquema avanzado de un SRI
Fuente: Salton 1983, Introducción a la Recuperación de Información Moderna.

En el gráfico se observa que el proceso establecido entre la entrada Reqs y Similar es el proceso de formulación de la búsqueda y el establecido entre Similar y el conjunto de documentos Docs es el proceso de recuperación. Similar es el proceso de determinación de la similitud existente entre la representación de la pregunta y la representación de los ítems de información.

En 1999 Chowdhury, identifica el siguiente conjunto de funciones principales en un SRI:

1. Identificar las fuentes de información relevantes a las áreas de interés de las solicitudes de los usuarios.
2. Analizar los contenidos de los documentos
3. Representar los contenidos de las fuentes analizadas de una manera que sea adecuada para compararlas con las preguntas de los usuarios.
4. Analizar las preguntas de los usuarios y representar las de una forma que sea adecuada para comparar con las representaciones de los documentos de la base de datos.
5. Realizar la correspondencia entre la representación de la búsqueda y los documentos almacenados en la base de datos.
6. Recuperar la información relevante.
7. Realizar los ajustes necesarios en el sistema basados en la retroalimentación con los usuarios.

2.2.9.1 COMPONENTES DE LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

a) La Base de Datos Documental

Desde un punto de vista matemático, la base de datos es una tabla o matriz en la que cada fila representa un documento y cada columna indica la presencia o no, de un determinado descriptor en el documento correspondiente. En principio, en cada fila aparecen "unos" en las columnas relativas a los descriptores asignados al documento y "ceros" en las restantes. Así cada documento estará representado por un vector de ceros y unos (Van Rijsbergen C. J., 1979).

Esta representación se podría mejorar introduciendo información numérica sobre la asignación de un descriptor al documento en lugar de simplemente 0 y 1. Se puede considerar una base documental *D*, compuesta por documentos di, indizada por un conjunto de términos, *T*, formado por *n* términos t_j, en la que cada documento d_i contiene un número no especificado de términos de indización t_j. De esta forma, sería posible representar cada documento como un vector perteneciente a un espacio n-dimensional, siendo *n* el número de términos de indización que forman el conjunto *T*:

$$di=(t_{i1},t_{i2},t_{i3},\ldots,t_{in}) \qquad (1)$$

Donde cada uno de los elementos t_{ij} de este vector puede representar la presencia o ausencia del termino t_j en el documento di en la indización binaria, la relevancia del termino t_j en el documento d_i en el modelo de Espacio Vectorial o

la probabilidad de que el documento d_i sea relevante al termino t_j en el modelo probabilístico.

La indización (proceso de construcción de los vectores documentales) puede realizarse de forma manual o automática. En este último caso, la base de datos documental comprende un módulo llamado indicador que se encarga de generar automáticamente la representación de los documentos extrayendo los contenidos de información de los mismos. La labor del módulo indicador consistirá en asociar automáticamente una representación a cada documento en función de los contenidos de información de este, es decir, determinar los pesos de cada término en el vector documental su función de indización o ponderación será:

$$F: DxT \rightarrow [0,1] \qquad (2)$$

La representación de cada vector tendrán componentes, de los cuales los que estén referenciados en el documento tendrán un valor diferente de 0, mientras que los que no estén referenciados tendrán el valor nulo o 0. Es importante señalar que la indización juega un papel fundamental en la calidad de recuperación, siendo crucial la elección apropiada del método de indización.

Se debe aclarar que un Sistema de Gestión de Base de Datos Documentales – SGBDD, se ocupa de la gestión de documentos optimizando el almacenaje y felicitando su recuperación. A diferencia de cualquier otro Sistema de Gestión de Base de Datos - SGBD, un SGBDD no realiza ningún tratamiento sobre la información. Simplemente la almacena y posibilita su recuperación.

b) Análisis Documental

Un SGBDD ofrece herramientas para realizar dos funciones básicas:

- Almacenar la Información documental y
- Facilitar su recuperación.

Para tal fin se emplea técnicas de análisis documental, siendo sus conceptos básicos:

- La indexación
- El lenguaje documental

El análisis documental se realiza a través de tres niveles de detalle:

- Asiento: se determinan los identificadores de la información (título, autor, …), para el caso de recuperación en la Unidad de Resoluciones será el número de resolución, tipo de resolución, fecha de resolución
- Descriptores: se extraen las palabras clave más representativas (indexación) para nuestro caso de estudio se considerará la parte resolutiva de la resolución.
- Resumen: resumen analítico del texto íntegro para hacer más fácil la consulta, para nuestro caso se empleará el Modelo Vectorial.

c) Lenguajes Documentales

Es un lenguaje artificial que permite realizar una representación formal de documentos considerando para ello la manera en que se realizan las demandas o consultas por parte de los usuarios.

En la práctica, un lenguaje documental viene hacer un conjunto estructurado de términos mediante los cuales se indexan los documentos.

Tesauros: es un lenguaje documental que incluye relaciones semánticas de tres tipos:

- Relaciones de equivalencia (sinónimos)
- Relaciones Jerárquicas (término general y término específico)
- Relaciones Asociativas (términos relacionados)

d) Indexación

Operación orientada a poner de manifiesto los temas del documento entresacando los elementos que lo representan para su posterior localización. Considera dos fases:

1. Extracción de descriptores, dispone de tres métodos
   - Estadístico (análisis de frecuencias)
   - Por asignación (listas normalizadas)
   - Sintáctico (análisis morfológico y semántico)
2. Traducción al lenguaje documental (Tesauro).

d.1) Métodos de Indexación:

Persiguen dos objetivos

- Mínimo espacio de almacenamiento y
- Máxima rapidez en indexación y respuesta.

d.2) Tipos de indexación

- Por las palabras: se indexa cada palabra del documento (excepto las incluidas en el diccionario de palabras vacías).

- De “string” se pueden indexar palabras o frases completas. Este método da lugar a índices menos extensos.

e) Diccionarios de palabras vacías

Conjunto de palabras que no deseamos que formen parte de los índices como pueden ser los artículos, las preposiciones y todas aquellas palabras que no son relevantes en la recuperación de la información y que solo contribuyen a ocupar espacio y disminuir la velocidad de acceso a la información.

f) Filtros

Herramientas que permiten igualar caracteres para que el sistema de búsqueda los considere iguales.

- Mayúsculas y minúsculas (A = a, …)
- Vocales acentuadas y sin acentuar (á = a, …).

2.2.10 MODELO CONCEPTUAL DE LA RECUPERACIÓN DE INFORMACIÓN RI

En principio, la RI, engloba las acciones encaminadas a identificar, seleccionar y acceder a los recursos de información útiles al usuario, el objeto documental se ha organizado y presentado, utilizando una serie de normas y convenciones, en un soporte informático, mediante el diseño, creación y mantenimiento de bases de datos.

Como los SRI, implementan una gama diversa de estructuras de datos, algoritmos y técnicas de recuperación de información, se precisa de un modelo conceptual donde se determinan el tipo de ficheros, operaciones sobre los

términos, modelos de búsqueda con base patrones exactos o los modelos inexactos los cuales contendrán las técnicas probabilísticas, los espacios vectoriales (Fernández 1998, Baeza-Yates y Ribeiro 1999).

Se ha tratado de mejorar el rendimiento de los SRI por medio de la distribución estadística de los términos, en tanto que la frecuencia de aparición de un término en un documento o conjunto de documentos podría considerarse relevante a la hora de establecer la similitud entre la consulta y el dato que identifica el documento. La distribución de frecuencia de un término se implementa dentro de algunos modelos estadísticos como es el caso del modelo espacio vectorial, o el modelo probabilístico (Mañas 1994, Fernández 1998).

Una decisión fundamental a tomar en el diseño de un SRI es que tipo de estructura de fichero se va a usar para la base de datos subyacente, podemos enunciar: ficheros de patrones de bits, árboles PAT y grafos (Frakes y Baeza-Yates 1992, Fernández 1998, Baeza-Yates y Ribeiro 1999).

2.2.11 ARQUITECTURA DEL SISTEMA DE RECUPERACIÓN DE INFORMACIÓN

La recuperación de información RI puede definirse como la representación, almacenamiento, organización y el acceso a elementos de información (Fernández 1998).

El campo de RI envuelve un conjunto bastante grande de conceptos, estructuras y métodos. Para seguir un orden lógico se verán las fases en las que se ve involucrado el tratamiento de la información que se puede resumir en: *el modelo conceptual, la indexación, la transformación de consultas, las*

operaciones sobre los términos y la gestión de documentos (Frakes y Baeza-Yates 1992, Fernández 1998, Baeza-Yates y Ribeiro 1999).

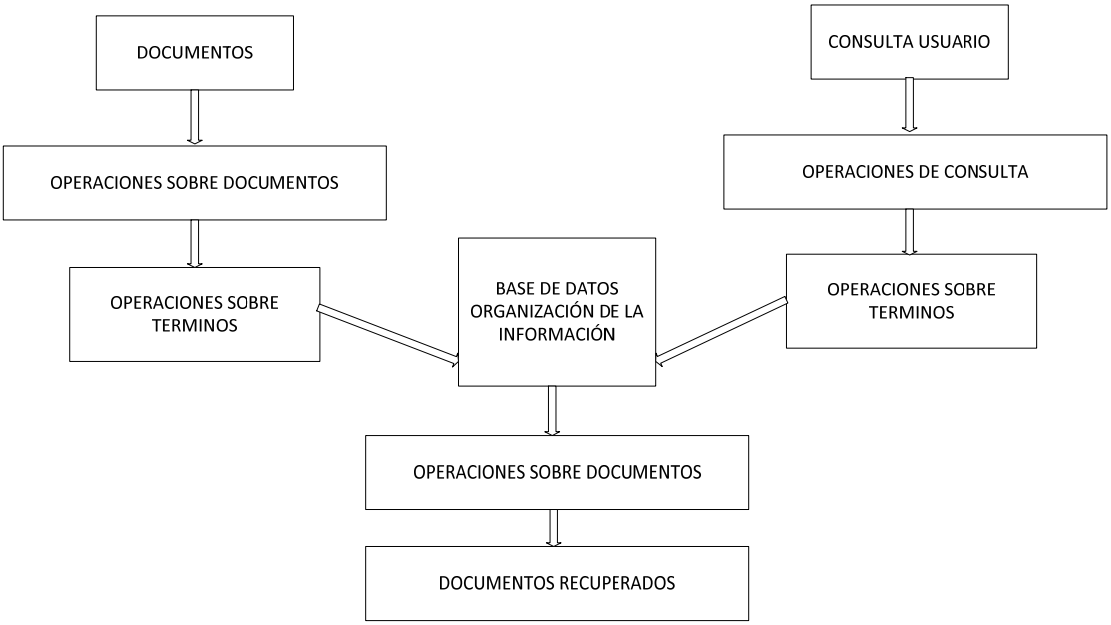


Figura Nº 4: Arquitectura de un sistema de recuperación
Fuente: Frakes y Baeza-Yates 1992, Information Retrieval

Los SRI se apoyan en dos módulos: uno de indexación, que construye los vectores de los documentos, y otro de consulta, que calcula la similaridad con una consulta dada (Figueroa, Berrocal y Zazo 2000).

Tanto los documentos como los vectores resultantes, así como productos intermedios y auxiliares se almacenan en una base de datos relacional; en la actualidad se ha superado algunos de los inconvenientes en las bases de datos como el manejo de los campos de tamaño variable (los conocidos como los campos memo), posibilidad de almacenar datos binarios (imágenes, sonido, referencia a objetos externos).

Básicamente el proceso de RI se divide en dos etapas: la indexación y búsqueda.

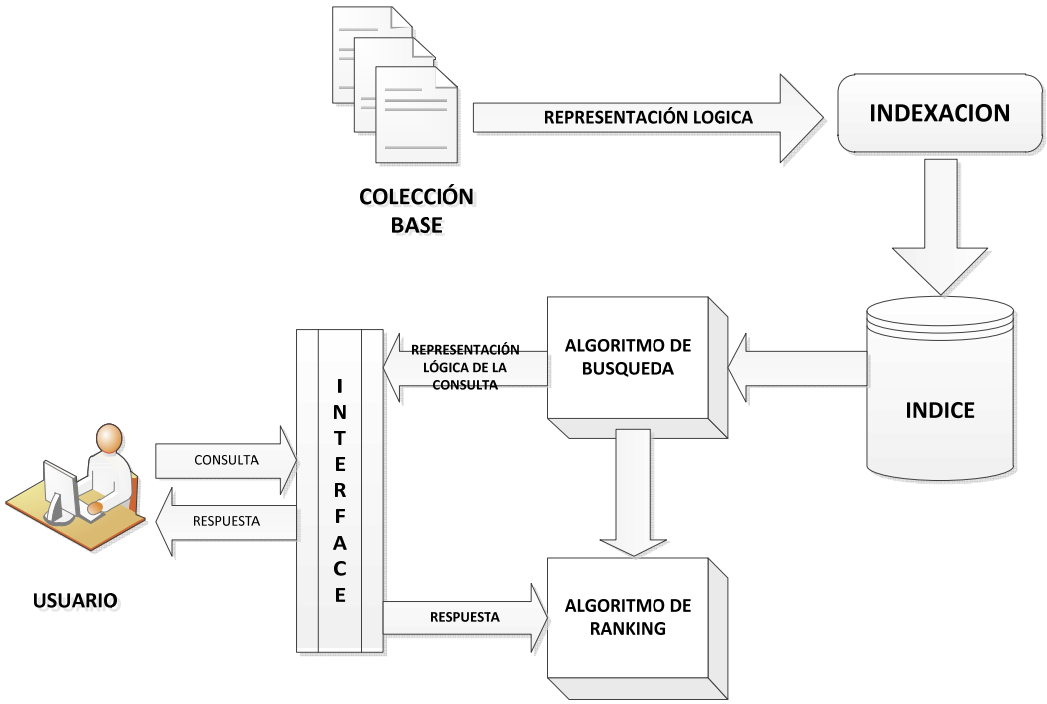


Figura Nº 5: Modelo simplificado de un Sistema de Recuperación de Información.
Fuente: Elaboración propia

2.2.12 MODELOS PARA LA RECUPERACIÓN DE INFORMACIÓN RI

El diseño de un SRI se realiza bajo un modelo, donde queda definido: como se obtienen las representaciones de los documentos y de la consulta, la estrategia para evaluar la relevancia de un documento respecto a una consulta y los métodos para establecer la importancia (orden) de los documentos de salida. (Villena Román, 1997).

En esta sección se analizarán varios modelos de RI. Formalmente se puede definir un modelo de RI de la siguiente manera: (Baeza-Yates, Ribeiro-Neto, 1999).

Definición 1: *Un modelo de recuperación de información es una cuaterna <D,Q,F,sim> en la que:*

- *D es un conjunto de vistas lógicas de documentos.*
- *Q es un conjunto de vistas lógicas de las necesidades de información de los usuarios. Los elementos de Q se denominan "preguntas" o "consultas".*
- *F es un marco que permite modelar los documentos, las preguntas y las relaciones entre ellos.*
- *Sim: D X Q* $\rightarrow R$, *siendo* R *el conjunto de los elementos reales, es una función de ordenación que asocia un número real a una pregunta* $q_j \in$ Q *y un documento* $d_i \in D$. *Esta función define un orden sobre los documentos respecto a su relevancia (o similitud) a la pregunta* q_j.

En consecuencia, tal como indican Baeza-Yates y Ribeiro-Neto (1999), un modelo RI consiste en: i) unas especificaciones para la información que el sistema tiene de los documentos y preguntas, ii) un marco que determina como se representan estos objetos en el modelo y que debe facilitar la construcción de una función de ordenación, y iii) una función de ordenación que permite la estimación de la relevancia de los documentos para las preguntas.

Definición 2: Sea $D = \{d_1,\ldots,d_m\}$ un conjunto de vistas lógicas de documentos y Q un conjunto de vistas lógicas de preguntas. Además, sea $T = \{t_1,\ldots,t_n\}$ el conjunto de términos de indexación para el conjunto D y R el conjunto de número reales.

El conjunto *D* se denomina colección de documentos, *m* es el número de documentos en la colección y *n* es el número de términos de indexación utilizados en las vistas lógicas de *D.*

Por razones de formalización, se representa cada documento $d_i \in D$ y cada pregunta $q \in Q$ mediante un vector sobre el conjunto de los términos de indexación, $\overrightarrow{dtf_i} = (tf_{i1}, \dots, tf_{in})$ y $\overrightarrow{qtf} = (tf_1, \dots, tf_n)$, respectivamente, siendo $tf_{ij}, tf_i \in R$ las frecuencias de aparición del término t_j en el documento d_i y en la pregunta *q*. Estos vectores se denominan vectores de frecuencias de términos o simplemente vectores de frecuencias.

Existen varias propuestas de clasificación de modelos, según Dominich (2000), clasifica en cinco grupos:

Modelo	**Descripción**
Modelos clásicos	Incluye los tres más común mente citados: booleano, espacio vectorial, y probabilística.
Modelos alternativos	Están basados en la lógica Fuzzy
Modelos lógicos	Basados en la lógica formal. La RI es un proceso inferencial.
Modelos basados en la interactividad	Incluyen posibilidades de expansión del alcance de la búsqueda y hacen uso de retroalimentación por la relevancia de los documentos recuperados.
Modelos basados en la inteligencia Artificial	Bases del conocimiento, redes neuronales, algoritmos genéticos y procesamiento del lenguaje natural.

Tabla Nº 2: Clasificación de los Modelos de Recuperación de Información.
Fuente: Dominich 2000, Métodos Matemáticos

A continuación se presenta la descripción de las características de los distintos modelos.

a) ***Modelo Booleano*** Es un modelo de recuperación simple, basado en la teoría de conjuntos y el álgebra booleana. Dadas su inherente simplicidad y su pulcro formalismo ha recibido gran atención y ha sido adoptado por muchos de los primeros sistemas bibliográficos comerciales. Su estrategia de recuperación está basada en un criterio de decisión binario (pertinente o no pertinente) sin ninguna noción de escala de medida, sin noción de un emparejamiento parcial de las condiciones de la pregunta.

El modelo Booleano, es uno de los modelos más simples e intuitivos. Por ello, ha sido adoptado en la mayoría de los primeros sistemas de recuperación de información y, todavía hoy en día, muchos sistemas lo utilizan (Salton 1983, Baeza-Yates, Ribeiro-Neto 1999).

El Modelo Booleano está basado en la Teoría de Conjuntos. Se considera que los documentos están representados mediante conjuntos de palabras claves que son subconjuntos del conjunto de términos de indexación. Haciendo uso de la definición 2, estos conjuntos se pueden representar mediante vectores binarios n-dimensionales. Cada documento $d_i \in D$ se representa con un vector $\overrightarrow{d_i}=(w_{i1}, \dots, w_{in})$ sobre el conjunto de términos de indexación *T*, siendo

$$w_{ij} = \begin{cases} 1 & si\ g_j(\overrightarrow{dtf_i}) > 0 \\ 0 & En\ caso\ contrario \end{cases} \qquad (3)$$

Las preguntas están especificadas por expresiones Booleanas. Una pregunta *q* consiste en un conjunto de palabras claves que están

conectadas mediante los operadores lógicos “AND”, “OR” y “NOT” (representados por los signos $\wedge, \vee, \neg$). En el proceso de recuperación, la función de ordenación, sim, devolverá un valor de 1 para todos los documentos, que cumplan la expresión formulada en la pregunta, y un valor de 0 para todos los demás. De esta forma si para un documento d_i y una pregunta q el valor de $sim(d_i,q)$ es 1, entonces se estima que d_i es relevante para q y se añade d_i al conjunto de los resultados.

El modelo Booleano es el modelo de RI más similar a la recuperación de datos. La especificación de las preguntas es exacta y refleja criterios de búsqueda precisos en lugar de una descripción ambigua de la información buscada. La comparación de la pregunta con los documentos se realiza de forma exacta y no a nivel conceptual o semántico. Eso también implica que los resultados son más susceptibles a posibles errores en la formulación de la pregunta. El resultado es un conjunto de aquellos documentos que cumplen la expresión lógica especificada, ya que se considera que estos son los documentos relevantes. Por tanto, no se determinan distintos niveles de relevancia y no es posible establecer una ordenación entre los documentos recuperados.

Las principales desventajas de este modelo son dos. En *Primer lugar*, se trata la relevancia como un criterio binario. Eso, por un lado, impide la posibilidad de facilitar a los usuarios algún criterio adicional acerca de la utilidad de los distintos documentos recuperados, y, por otro lado, no tienen en cuenta aquellos documentos que cumplen los criterios de la

pregunta de forma parcial. Estos últimos se consideran simplemente irrelevantes. Este comportamiento, en parte, también es debido al uso de pesos binarios para los términos en las representaciones de los documentos. Una única aparición de cierto término en un documento puede ser accidental, pero si el mismo término aparece varias veces es más probable que refleje el contenido del documento. En segundo lugar, la traducción de una necesidad de información en una expresión Booleana no siempre es fácil, en gran medida por que las expresiones son exactas mientras las necesidades de información muchas veces no lo son.

b) ***Modelo Probabilístico***, Es uno de los modelos considerado como clásico. Como indica; Sparck Jones y Willett (1997), "*La función principal de un sistema de RI es clasificar los documentos en una colección en orden de probabilidad decreciente de relevancia a la necesidad de información de un usuario*". El modelo de recuperación probabilístico se basa en la equiparación probabilística, dados un documento y una pregunta, es posible calcular la probabilidad de que ese documento sea relevante para esa pregunta.

El modelo Probabilístico enfoca, el problema de la recuperación de información mediante la teoría de probabilidad. Los primeros trabajos en este sentido datan del principio de los años 60 y han sido: Maron y Kuhns (1960). En su artículo, los autores buscan soluciones a las limitaciones del modelo Booleano, primero un número de relevancia a los documentos que indican la probabilidad de relevancia de los documentos y permite la ordenación; segundo, permitiendo la especificación de pesos para los

términos de la pregunta, y, tercero, utilizando una expansión del conjunto de documentos recuperados mediante i) la recuperación de documentos "parecidos" y ii) la expansión de la pregunta.

El modelo probabilístico clásico fue desarrollado años más tarde, por Robertson y Sparck Jones (1976). Este modelo, que se conoce también como *modelo de recuperación de independencia binaria* ("binary Independence retrieval model"- modelo BIR).

En el modelo BIR la suposición básica es que los términos de indexación están distribuidos de forma desigual entre los documentos relevantes e irrelevantes. Se considera que tanto los documentos como las preguntas están representados mediante conjuntos de términos de indexación; es decir, tanto un documento d_i como una pregunta q están representados mediante vectores binarios sobre el conjunto de términos de indexación.

$\overrightarrow{d_i}=(w_{i1},\ldots,w_{in})$ y $\vec{q}=(w_1,\ldots,w_n)$, siendo w_{ij}, $w_j \in \{0;1\} \forall j \in [1,n]$. (4)

Dada una pregunta q, un documento d_i debe ser recuperado por su probabilidad de pertenecer al conjunto de documentos relevantes para q es mayor que la probabilidad de pertenecer al conjunto de documentos irrelevantes. Con el objetivo de ordenar los documentos de una colección por su probabilidad de ser relevantes.

Las limitaciones básicas del modelo probabilístico clásico, pueden resumirse en tres puntos:

- El modelo usa una recuperación binaria de los documentos y preguntas. No se tiene en cuenta la frecuencia de aparición de los términos en los documentos. Los primeros intentos de incluir la frecuencia de aparición sólo han tenido éxito limitado.
- Aunque el modelo tiene una base fundada para estimar los parámetros se requiere información sobre los conjuntos de documentos relevantes e irrelevantes para una pregunta. Esta se puede obtener, sólo de forma parcial, de la interacción con los usuarios en la recuperación con realimentación sobre relevancia. No obstante, en muchos escenarios esta interacción no es posible o no es deseable.
- El tercer punto es la suposición de la independencia condicional de la aparición de los términos en los documentos. En el sentido semántico, esta suposición equivale a considerar que no existen relaciones entre términos, o, en otras palabras, la existencia de ciertos términos en un documento no indica la existencia de otros términos semánticamente relacionados.

2.2.13 CÁLCULO DE PESO DE UN TÉRMINO

Existen varios esquemas propuestos para propiciar la jerarquización de documentos como es el uso de pesos de cada término, es decir la frecuencia de los términos en los documentos, incluso otros proporcionan una normalización; incluido en las búsquedas.

Para el cálculo de pesos del término se hará uso del TF-IDF (Term Frequency – Inverse Document Frequency), frecuencia de término – frecuencia inversa de documento (frecuencia de ocurrencia del término en la colección de

documentos), es una medida numérica que expresa cuan relevante es una palabra para un documento en una colección. Esta medida se utiliza a menudo como factor de ponderación en la Recuperación de la Información. El valor tf-idf aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero es compensada por la frecuencia de la palabra en la colección de documentos, lo que permite manejar el hecho de que algunas palabras son generalmente más comunes que otras (Vuotto A., Bogetti C. 2015).

Variaciones del esquema de peso tf-idf son empleadas frecuentemente por los motores de búsqueda como herramienta fundamental para medir la relevancia de un documento dada una consulta del usuario, estableciendo así una ordenación o ranking de los mismos. Para obtener el peso de los términos de un documento es preciso utilizar la fórmula:

$$w_{i,j} = f_{i,j} \times idf_i \tag{5}$$

En donde $w_{i,j}$ representa el peso del término *i* en el documento *j*, y así $f_{i,j}$ es la frecuencia normalizada del término *i* en el documento *j*, por último idf_i representa la frecuencia invertida del documento *i*. Dichas medidas pueden ser obtenidas utilizando las fórmulas:

$$f_{i,j} = \frac{frec_{i,j}}{Max\ frec_j} \tag{6}$$

$$idf_i = log \frac{N}{n_i} \tag{7}$$

Donde $frec_{i,j}$ es la frecuencia del término *i* en el documento *j* (número de veces que aparece el término en un documento) y *max* $frec_j$ representa la

máxima frecuencia sobre todo los términos del documento *j*. *N* es el número de documentos que contiene la colección y n_i es el número de documentos que contiene el término *i*.

2.2.14 MODELO VECTORIAL

En 1960 Gerald Salton, fue primero en proponer los SRI basados en Espacio Vectorial SRI-EV, dentro del marco de proyecto SMART, en la Cornell University, partiendo de que se puede representar los documentos como vectores de términos, los documentos podrán situarse en un espacio vectorial de *n* dimensiones, es decir, con tantas dimensiones como elementos tenga el vector. Situado en ese espacio vectorial, cada documento cae entonces en un lugar determinado por sus coordenadas, al igual que en un espacio de tres dimensiones cada objeto queda bien ubicado si se especifican sus tres coordenadas espaciales. Se crean así grupos de documentos que quedan próximos entre sí a causa de las características de sus vectores

En 1989 Salton, menciona que el modelo vectorial MV, es uno de los modelos más populares y que ha encontrado más atención en el área de recuperación de información. Este modelo es aplicable, tanto cuando la indexación de los documentos es manual, mediante la asignación de palabras claves, como cuando es automática por texto completo. Comparado con el modelo Booleano, el MV permite la definición y el uso de pesos para los términos de indexación e implementa la técnica de ordenación por relevancia. Con ello, el resultado de un proceso de recuperación es una lista de documentos ordenados de mayor a menor relevancia respecto a la pregunta.

Según este modelo, cada documento es representado mediante un vector de n elementos, siendo n igual al número de términos indizables que existen en una colección documental, existe un vector para cada documento, y en cada vector, un elemento para cada término o palabra susceptible de aparecer en el documento. Cada uno de esos elementos es cubierto u ocupado con un valor numérico. Si la palabra no está presente en el documento, ese valor es igual a 0. En caso contrario, ese valor es calculado teniendo en cuenta diversos factores, dado que una palabra puede ser más o menos significativa, este valor se conoce con el nombre de peso del término en el documento.

En el MV, tanto los documentos como las preguntas de los usuarios se representan mediante vectores en un espacio cuyo conjunto de vectores de base corresponde al conjunto de términos de indexación. Cada documento d_i de una colección de D está representado mediante un vector $\overrightarrow{d_i} = (w_{i1}, \dots, w_{in})$ y, de igual forma, cada pregunta q está representada por un vector $\overrightarrow{q_i} = (w_1, \dots, w_n)$. Cada elemento $w_{ij}(w_j)$ refleja su importancia del término t_j en la descripción del documento d_i (la pregunta q). Si $w_{ij} = 0$, el término t_j no ha sido seleccionado para representar el contenido, mientas que si $w_{ij}>0$, el w_{ij} es el peso de t_j en la descripción del documento (pesos negativos no se utilizan).

La idea del modelo vectorial es que la relevancia de un documento a una pregunta puede estimarse a través de la correlación o de la similitud de sus vectores. Por eso, las funciones que cuantifican estas correlaciones – en realidad las funciones de ordenación de la definición 1 – se suele llamar “funciones o medidas de similitud”.

Según el Modelo de Espacio Vectorial, el proceso de recuperación para una pregunta consiste en calcular las similitudes entre cada uno de los documentos de la colección (sus vectores) y el vector de la pregunta. Las consultas son representadas también mediante un vector de las mismas características que la de los documentos (variando los valores numéricos de cada elemento en función de las palabras que forman parte de la consulta). Esto permite calcular fácilmente una función de similaridad dada entre el vector de una consulta y los de cada uno de los documentos. El resultado de dicho cálculo mide la semejanza entre la consulta y cada uno de los documentos, de manera que aquellos que, en teoría, se ajustan más a la consulta formulada, producen un índice más alto de similaridad. Naturalmente se asume que la consulta se formula en lenguaje natural, se deduce que el resultado de la consulta consiste en una lista de documentos ordenada en orden decreciente en función de su similaridad con la consulta.

2.2.14.1 CARACTERIZACIÓN DEL MODELO VECTORIAL

Ordenando los documentos recuperados en orden decreciente a este grado de similitud, el modelo de recuperación vectorial toma en consideración documentos que solo se emparejan parcialmente con la pregunta, así el conjunto de la respuesta con los documentos alineados es mucho más preciso (en el sentido que empareja mejor la necesidad de información del usuario) que el conjunto recuperado por otro modelo. Los rendimientos de alineación del conjunto de la respuesta son difíciles de mejorar.

La mayoría de los motores de búsqueda lo implementan como estructura de datos y que el alineamiento suele realizarse en función del parecido (o similitud) de la pregunta con los documentos almacenados.

2.2.14.2 FUNCIONAMIENTO DEL MODELO VECTORIAL

La idea básica de este modelo de recuperación vectorial, reside en la construcción de una matriz (podría llamarse tabla) de términos y documentos, donde las filas fueron estos últimos y las columnas corresponderían a los términos incluidos en ellos. Así las filas de esta matriz (que en términos algebraicos se denominan vectores) serían equivalentes a los documentos que se expresarían en función a las apariciones (frecuencia) de cada término. De esta manera un documento podría expresarse de la manera:

$$d1 = (1,2,0,0,\ldots, \ldots, \ldots,1,3) \qquad (8)$$

Siendo cada uno de estos valores el número de veces que aparece cada término en el documento.

Un documento d_j también, se modeliza como un vector $d_j = (w_{1,i}, \ldots w_{t,i})$, donde $w_{i,j}$ es el peso del termino t_j en el documento d_j.

La longitud del vector de documentos sería igual al total de términos de la matriz (el número de columnas).

Según el Modelo de Espacio Vectorial, las consultas son representadas también mediante un vector de las mismas características que las de los documentos (variando los valores numéricos de cada elemento en función de las palabras que forman parte de la consulta). Esto permite calcular fácilmente una

función de similaridad dada entre el vector de una consulta y los de cada uno de los documentos. El resultado de dicho cálculo mide la semejanza entre la consulta y cada uno de los documentos, de manera que aquellos, en teoría, se ajustan más a la consulta formulada producen un índice más alto de similaridad. Naturalmente se asume que la consulta se formula en lenguaje natural (podría ser, incluso un documento de muestra, para recuperar los que fuesen parecidos a él). Se deduce que el resultado de la consulta consiste en una lista de documentos ordenada en forma decreciente en función de su similaridad con la consulta.

Simplificando la función *sim*$_{angulo}$, obteniendo el *coeficiente de coseno.*

$$sim_{cos}(d_i, q) = \frac{\vec{d_i} * \vec{q}}{|\vec{d_i}| \cdot |\vec{q}|} = \frac{\sum_{j=1}^{n} g_j(\vec{d_i}).g_j(\vec{q})}{\sqrt{\sum_{j=1}^{n} g_j(\vec{d_i})^2}\sqrt{\sum_{j=1}^{n} g_i(\vec{q})^2}} \quad (9)$$

"*" denota el producto escalar entre dos vectores, y $|\vec{d_i}|$ y $|\vec{q}|$ son las longitudes euclídeas o normas de los vectores $\vec{d_i}$ y $\vec{q}$. Esta función calcula el coseno del ángulo entre los dos vectores y tiene resultados en el intervalo [0;1] ($sim_{cos}(d_i, q) = 0$ si los vectores son ortogonales y el ángulo es máximo, y $sim_{cos}(d_i, q) = 1$ si el ángulo es 0 y los vectores tienen la misma dirección en el espacio).

2.2.14.3 CÁLCULO DE LA SIMILITUD

Se dispone de varias fórmulas que nos permiten realizar este cálculo, la más conocida es la Función del Coseno, que equivale a calcular el producto escalar de dos vectores de documentos (A y B) y dividirlo por la raíz cuadrada

del sumatorio de los componentes del vector A multiplicada por la raíz cuadrada del sumatorio de los componentes del vector B.

De esta manera se calcula este valor de similitud. Como es obvio si no hay coincidencia alguna entre estos componentes, la similitud de los vectores será cero ya que el producto escalar será cero (circunstancia muy frecuente en la realidad ya que los vectores llegan a tener miles de componentes y se da el caso de la no coincidencia con mayor frecuencia de lo que cabría pensar).

La función del Coseno no es la única función de similitud. Existen otras (el producto escalar, el coeficiente de Dice, el coeficiente de Jaccard, el coeficiente de coincidencia), las cuales no son difíciles de calcular, sino más bien de interpretar y que por tanto son menos aplicadas en Recuperación de Información.

El modelo vectorial sólo modela los procesos de indexación y recuperación sobre la base de una representación vectorial de los documentos y de las preguntas, y una variante concreta de este modelo se puede definir de la siguiente manera:

Definición 3 Una variante del modelo vectorial es una quinterna $<pl_D,pl_Q,pg_D,pg_Q,sim>$, en la que:

- pl_D es una función que determina los pesos locales de los términos en los documentos.
- pl_Q es una función que determina los pesos locales de los términos en las preguntas.
- pg_D es un esquema de pesos globales que se aplica a los documentos.

- pg_Q es un esquema de pesos globales que se aplica a las preguntas.
- *sim* es una función de ordenación que estima la relevancia de un documento a una pregunta.

En muchos casos, se usa la misma función para obtener los pesos locales y el mismo esquema de pesos globales para los documentos y las preguntas, es decir, $pl_D = pl_Q$ y $pg_D = pg_Q$. No obstante, en el caso general estos parámetros pueden ser distintos.

Las ventajas del modelo vectorial respecto al modelo Booleano son obvias: el uso de la relevancia como un criterio de valores continuos, que admite que documentos sean “parcialmente relevantes” y, por tanto, permite el uso de la técnica de ordenación. Otra ventaja del modelo es su gran flexibilidad que consiste en la posibilidad de utilizar distintos esquemas de pesos y distintas funciones de similitud.

Como principal inconveniente del MV clásico hay que destacar que la estimación de la relevancia se basa en su esencia, en una comparación léxica entre preguntas y documento. No se realiza una comparación semántica o conceptual y un documento es considerado más relevante si un mayor número de sus palabras coinciden con la pregunta. No obstante, este problema no es en realidad del modelo en sí, sino más bien de las variantes tradicionalmente utilizadas.

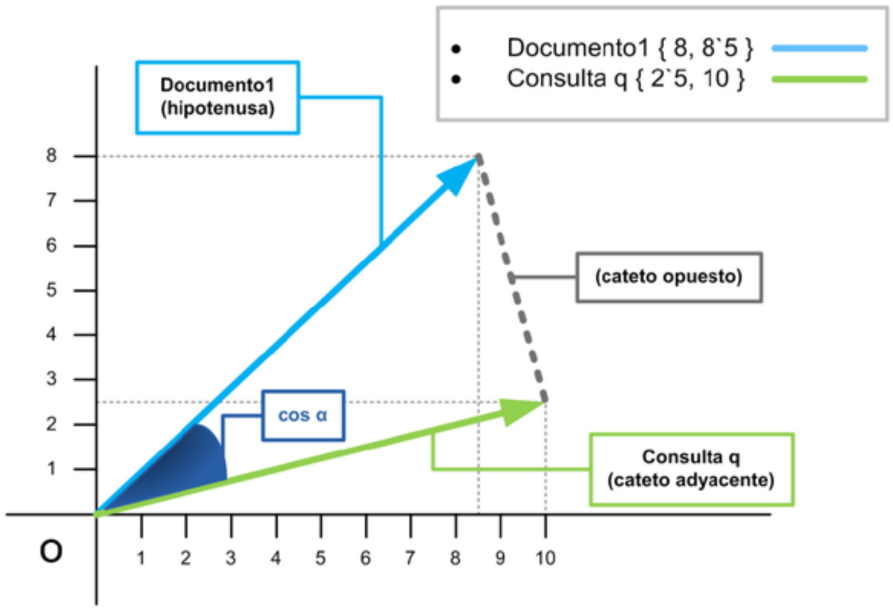


Figura Nº 6: Representación gráfica de la similaridad de dos documentos. Fuente: Blázquez M. 2013, Técnicas Avanzadas de Recuperación de Información.

2.2.14.4 MÓDULO DE INDEXACIÓN

La vista lógica de un documento o una pregunta consiste en una bolsa de términos de indexación. Un problema que debe resolver cualquier Sistema de Recuperación de Información SRI por texto completo es la selección de los términos de indexación que representan las vistas lógicas. El espectro de posibilidades para esta selección es amplio: desde la utilización de todas las palabras existentes en los documentos hasta el uso de un conjunto muy reducido de los lexemas de las palabras más significativas.

La mayoría de los sistemas de indexación por texto completo usan un enfoque intermedio, en el que se obtienen los términos de indexación mediante un proceso que filtra las palabras más importantes y convierte las palabras a sus formas más elementales, las raíces o lexemas. De esta forma, aparte de eliminar

palabras cuya selección como términos de indexación puede empeorar la calidad de recuperación, se reduce el vocabulario que el sistema tiene que mantener, lo que lleva a menores requerimientos de memoria y mejores tiempos de cálculo

En la indexación de los documentos, no todas las palabras o términos que componen se incluyen en los índices. A los términos que se incluyen en el índice se les llama elementos de indexación. Además, hay que considerar que dichos elementos pueden sufrir una serie de transformaciones antes de acabar el índice.

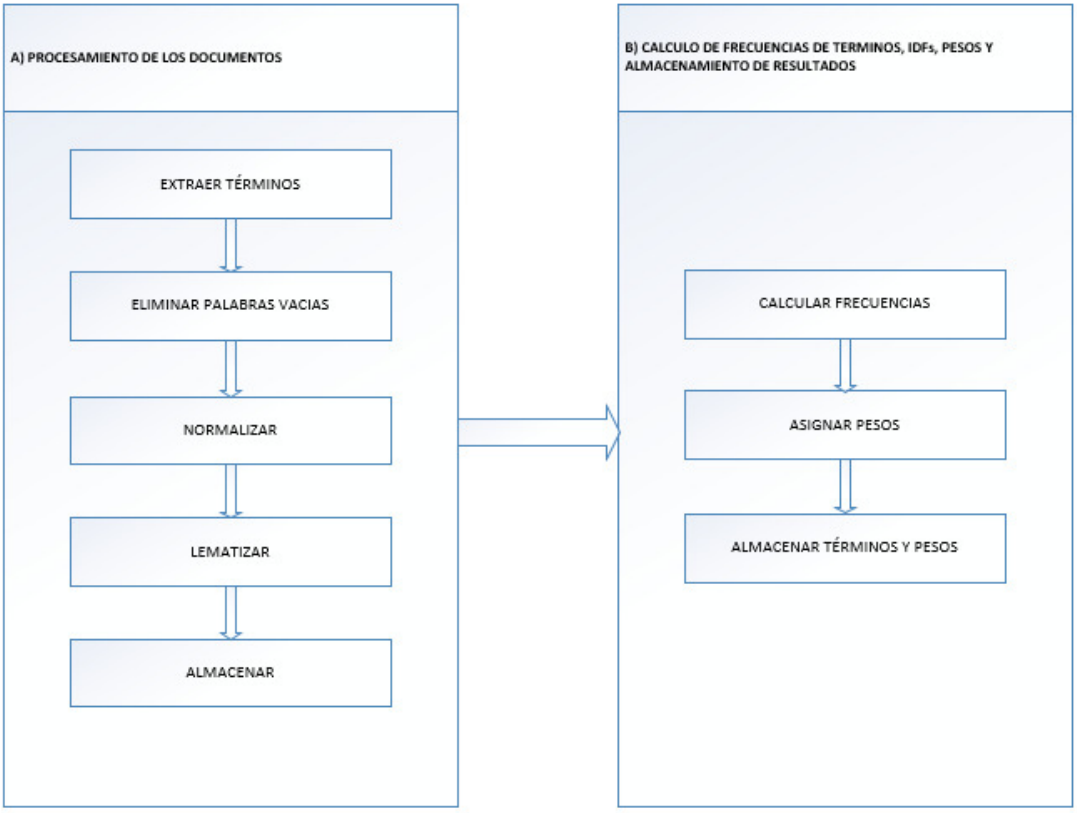


Figura Nº 7: Módulo de indexación
Fuente: Elaboración propia

2.2.14.5 ANÁLISIS LÉXICO

El análisis léxico es la obtención de las palabras propiamente dichas a partir de un texto. En este proceso se deben de identificar las palabras de un texto teniendo en cuenta todos los separadores posibles, como caracteres en blanco o caracteres especiales o de puntuación. Sin embargo, como indica Fox (1992), hay algunos aspectos que deben analizarse con cuidado: i) números, ii) caracteres de puntuación, iii) guiones y iv) letras en mayúscula y minúscula.

Normalmente, los números son eliminados porque no son muy útiles para reflejar los contenidos de los documentos. Eso, sin embargo, implica que, si un número es parte de la pregunta, éste es ignorado. Por ejemplo, las búsquedas con preguntas que incluyen fechas no devolverán los resultados deseados. Por tanto, la utilidad de los números depende en gran medida del contexto y en algunas tareas de RI como la contestación a preguntas, o cuando se integra la recuperación de datos con la recuperación de información el uso de números como términos de indexación puede ser beneficioso.

Aparte de los números propiamente dichos hay que definir el tratamiento de cadenas que incluyen dígitos, como, por ejemplo, “Windows7”, “resolución rectoral 2016”, “jdk1.3”. Este tipo de palabras se refiere muchas veces a entidades muy concretas o a descriptores técnicos (por ejemplo, nombres de productos o nombres de modelos). También en la decisión de incluir o no este tipo de términos debe considerarse el contexto en el cual se usa el sistema.

El problema de los caracteres de puntuación ocurre cuando estos no representan estructuras gramaticales en el texto, sino que son parte de palabras

como “jdk1.3”, “fichero.dat” o casos similares. Normalmente, los signos de puntuación son usados como delimitadores de palabras y se eliminan del texto. Esta eliminación, sin embargo, no debe llevarse a cabo en los casos mencionados sino se quiere perder el significado original de estos términos.

El caso de los guiones es algo parecido. Normalmente son usados para separar palabras al final de una línea. Por tanto, el tratamiento obvio es la eliminación del guion y el agrupamiento de las palabras adyacentes. No obstante, los guiones también pueden ser parte de palabras, algunas palabras pueden escribirse con o sin un guion intermedio, como ocurre en el inglés.

Respecto a la diferenciación de caracteres en mayúsculas y minúsculas generalmente se opta por una de las dos variantes. Eso, implica que, por ejemplo, las palabras “Administración” y “administración” son tratadas como iguales. Sin embargo, también existen casos en los que la misma palabra se refiere a cosas distintas dependiendo del uso de mayúsculas y minúsculas. Eso, por ejemplo, ocurre en algunos lenguajes de programación y, por tanto, en documentos que traten de estos lenguajes.

2.2.14.6 FILTRADO Y ELIMINACIÓN DE PALABRAS VACÍAS

Desde el punto de vista de los documentos, las palabras vacías son muy frecuentes y, por tanto, no son buenos indicadores para separar el conjunto de documentos relevantes del conjunto de documentos irrelevantes. En 1975, Salton, Wong y Yang; indican que en el Modelo Vectorial, la efectividad de la recuperación esta inversamente relacionada con la densidad del espacio de documentos de una colección. Es decir, se obtienen mejores resultados cuando

los documentos son menos similares entre sí. Basándose en esta hipótesis, Salton, Wong y Yang, analizan el efecto que produce la inclusión o no inclusión de un término en el conjunto de términos de indexación en la densidad del espacio de los documentos. El resultado de este análisis es que tanto las palabras que aparecen en muchos documentos como las que aparecen en muy pocos son los peores discriminadores. La inclusión de estos términos lleva a un espacio de documentos más denso y reduce la efectividad de la recuperación. Basándose en estas conclusiones, se puede considerar palabras vacías todas aquellas que aparecen con mucha frecuencia en un cuerpo suficientemente grande y, de hecho, en la práctica, muchas listas de palabras vacías incluyen simplemente las palabras más frecuentes.

2.2.14.7 REDUCCIÓN DE LAS PALABRAS A SUS LEXEMAS ("STEMMING")

La reducción de las palabras a sus lexemas se refiere al proceso en el que se agrupan las diferentes palabras con la misma raíz bajo un único representante, su lexema ("stem"). La idea subyacente es que es la raíz la que contiene la información sobre los conceptos asociados a las palabras. Las diferentes palabras construidas sobre el mismo lexema sólo introducen pequeños matices respecto a los conceptos o, muchas veces, representan estructuras gramaticales o puramente sintácticas.

La idea de reducir las palabras a sus partes significativas ha sido aplicada en los SRI desde la automatización del proceso de indexación. El objetivo inicial ha sido la mejora de la efectividad de los sistemas al considerar palabras de una misma base como iguales, pero el fin también ha sido la reducción de los recursos de los sistemas en cuanto a memoria y tiempo de cálculo.

2.2.14.8 PESOS DE TÉRMINOS

La representación de los documentos y de las preguntas se basa, en un gran número de modelos, en conjuntos de términos de indexación. Es decir, internamente, cada documento esta descrito por un conjunto de términos, cada una de ellos corresponde a una palabra, y la semántica subyacente de estas palabras refleja en cierta medida el contenido del documento.

En la indexación automática, el conjunto de términos asignado a un documento corresponde directamente a las palabras que aparecen en el texto o se deriva de éstas. Sin embargo, no todas las palabras de un documento son igualmente útiles para reflejar su contenido. Algunas son menos específicas y tienen menos carga semántica que otras y, por tanto, deberían tener menos peso en la representación, por ejemplo, las palabras vacías, tienen esta característica. Estas palabras normalmente no se usan como términos de indexación, suelen haber otras, con características similares, que no se eliminan previamente. De manera similar, palabras con una frecuencia de aparición media normalmente son más útiles a la hora de diferencias entre documentos relevantes e irrelevantes a una pregunta.

Por otro lado, el trabajo descrito por Salton, Wong y Yang (1975), obtiene resultados similares, pero considera la distribución de las palabras sobre los documentos de una colección. También desde este punto de vista, las palabras que aparecen en muy pocos o en demasiados documentos son los peores términos para la indexación.

2.2.14.9 MÓDULO DE CONSULTA

El módulo de consulta en lenguaje natural ha de ser tratada como un documento cualquiera, requiere las mismas operaciones: obtención de palabras, eliminación de palabras vacías, normalización de caracteres, lematización, cálculo de pesos de los términos de la consulta, utilizando los datos de IDF almacenados en una tabla en la operación de indexado, cálculo de similaridad entre la consulta y cada uno de los documentos, mediante una simple sentencia SQL. Para realizar el cálculo de similaridad entre dos vectores existen diversas funciones, siendo la más conocida la del producto escalar de dos vectores y los coeficientes del coseno.

Al hacer el cálculo del coeficiente de similaridad de los documentos y del vector de búsqueda y someterlos a una comparación sistemática, se está en condiciones de establecer un orden descendente, colocando en primer término el documento cuyo valor es más cercano al del vector de búsqueda y así hasta concluir con todos los registros resultantes.

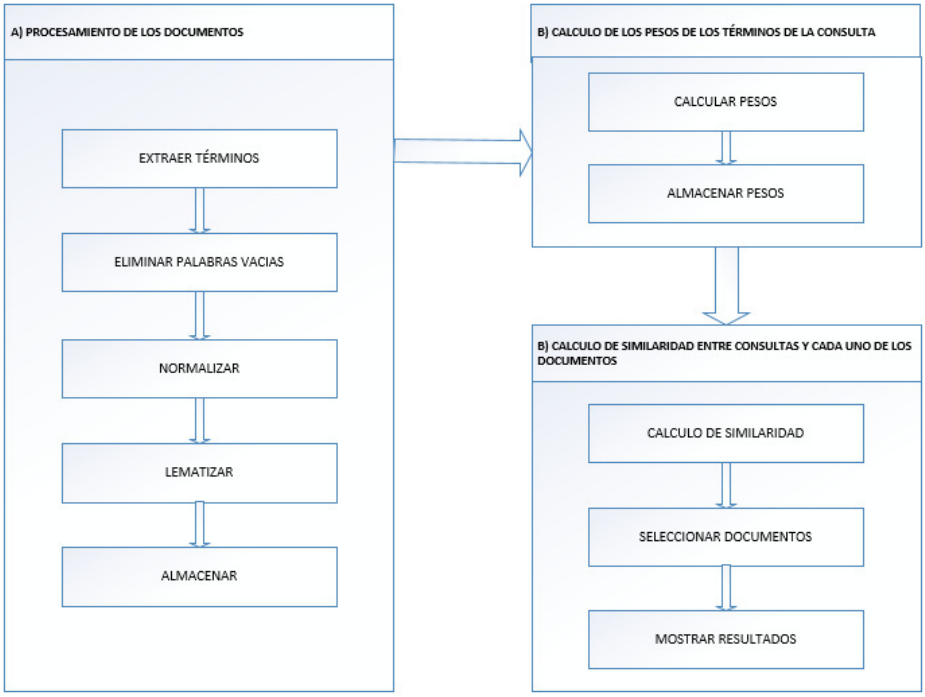


Figura Nº 8: Módulo de búsqueda
Fuente: Elaboración propia

2.2.14.10 ARQUITECTURA DEL MODELO VECTORIAL

En la Figura 11, se presenta la gráfica de la vista funcional del modelo, en donde se realizan las siguientes tareas:

a) Se analizan los documentos y se transforman a una representación interna de cada uno.

b) Se analiza la consulta y se transforma a una representación interna.

c) A partir de las representaciones obtenidas en los pasos anteriores se calcula el grado de similitud entre cada documento y la consulta.

d) Se recuperan los documentos que guardan mayor similitud con la consulta del usuario.

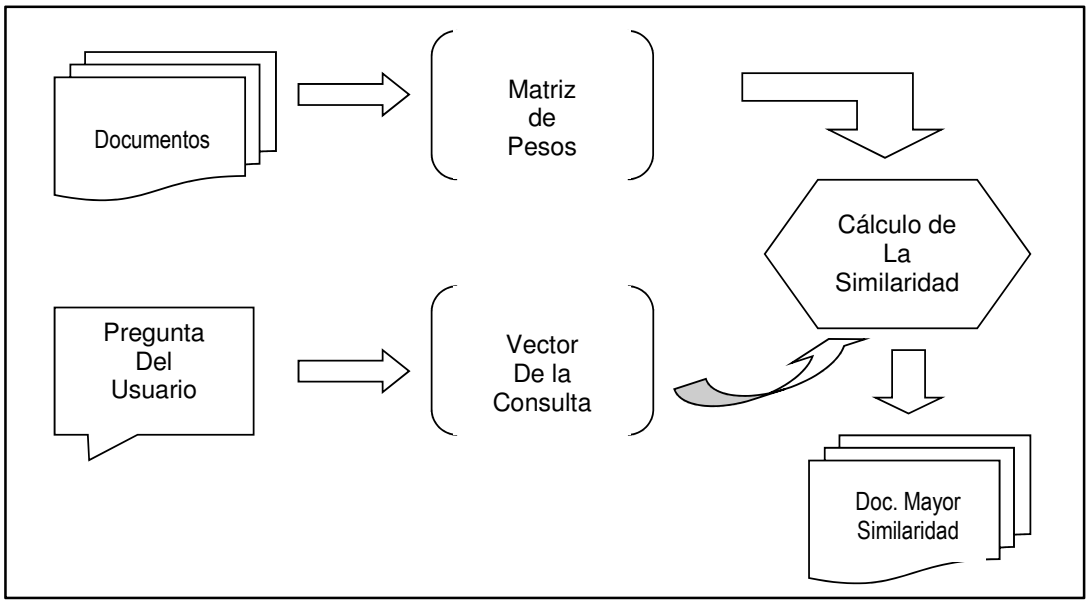


Figura Nº 9: Representación Gráfica de la Vista Funcional del modelo
Fuente: Elaboración propia

Los documentos están formados por "pesos de términos". El primer paso es escoger los términos. Por ejemplo, seleccionamos como términos cada una de las palabras en los siguientes documentos:

Doc1 = "Sistemas de Recuperación de la Información".

Doc2 = "Clasificación de los SRI".

Doc3 = "Motores de búsqueda".

2.2.15 EVALUACIÓN DE SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

Los SRI, como cualquier otro sistema, son susceptibles de ser sometidos a evaluación con el fin de que los usuarios se encuentren en condiciones de valorar su efectividad y de este modo adquieran confianza en los mismos.

Las dos medidas más utilizadas acostumbran a ser el índice de exhaustividad (*recall*) y el índice de precisión (*precisión*).

Las fórmulas para estos dos índices son las siguientes:

$$\varepsilon = \frac{n_{drr}}{n_{tdrp}} \times 100 \qquad (10)$$

Donde:

ε :	Exhaustividad
n_{drr} :	Número de documentos relevantes recuperados
n_{tdrp} :	Número total de documentos relevantes presentes

$$\rho = \frac{n_{drr}}{n_{tdr}} \times 100 \qquad (11)$$

Donde:

P :	Precisión
n_{drr} :	Número de documentos relevantes recuperados
n_{tdr} :	Número total de documentos recuperados

Mientras el índice de exhaustividad proporciona una medida de la habilidad del sistema para recuperar documentos relevantes, el índice de precisión proporciona una medida de la habilidad del sistema para evitar el ruido.

Naturalmente, el objetivo consiste en diseñar sistemas que proporcionen al mismo tiempo un 100% de exhaustividad y un 100% de precisión, es decir, sistemas que recuperen todos los documentos relevantes y tan sólo los documentos relevantes, pero, en la práctica, estos dos indicadores se comportan de manera antagónica, ya que las medidas para incrementar la exhaustividad tienden a disminuir la precisión y al revés.

La razón es la siguiente, si queremos asegurar la precisión del sistema adoptaremos medidas tendentes a aumentar la especificidad de la indización.

2.3 GLOSARIO DE TÉRMINOS BÁSICOS

Administración

Proceso que consiste en las actividades de planeación, organización, dirección y control para alcanzar los objetivos establecidos utilizando para ello recursos económicos, humanos materiales y técnicos a través de herramientas y técnicas sistematizadas.

Administración de documentos

Se refiere al adecuado manejo de la información producida y recibida por la administración pública o privada en el ejercicio de sus funciones, bajo criterios y técnicas archivísticas y se basa en el concepto de ciclo vital de la documentación y el principio de conservación selectiva de información. Es la planificación, control, dirección, organización, capacitación, promoción y otras actividades gerenciales relacionadas con la creación, uso, conservación y disposición de documentos.

Gestión

Hace a la referencia a la acción y al efecto de gestionar o administrar, el mismo que también implica un conjunto de trámites que se llevan a cabo para resolver un asunto o concretar un proyecto. La gestión es también la dirección o administración de una empresa o de un negocio.

Gestión documental

Conjunto de normas técnicas y prácticas usadas para administrar el flujo de documentos de todo tipo en una organización, permitir la recuperación de información desde ellos, determinar el tiempo que los documentos

deben guardarse, eliminar los que ya no sirven y asegurar la conservación indefinida de los documentos más valiosos, aplicando principios de racionalización y economía.

Área de gestión responsable de un control eficaz y sistemático, de la creación, la recepción, el mantenimiento, el uso y la disposición de documentos de archivos incluidos los procesos para incorporar y mantener en forma de documentos la información y prueba de las actividades y operaciones de la organización (Norma ISO/IEC 15489, para la Gestión de Documentos).

Modelo de Recuperación de información RI

Un modelo de RI es la especificación sobre como representar documentos y consultas y cómo comparar unos y otros. El objetivo de todo modelo es obtener un orden (ranking) de los documentos recuperados que refleje la relevancia de estos con la consulta del usuario.

Motor de búsqueda

Herramienta que basa su funcionamiento en palabras clave que tienen el objetivo de realizar búsquedas dentro de una base de datos.

Recuperación de información

La recuperación de la información es la disciplina encargada de la representación, almacenamiento y la organización de la información, y su posterior acceso y recuperación de información que responda a las necesidades de un usuario.

Sistema de Apoyo a Decisiones (SAD) basados en Web

Se están desarrollando SAD basados en Web, para apoyar la toma de decisiones, al proporcionar acceso en línea a diversas bases de datos y almacenamientos de información, junto con software para análisis de datos. Algunos de estos SAD van dirigidos a la administración, pero otros se han creado para atraer a clientes, ya que proporcionan información y herramientas que los ayudan a elegir productos y servicios.

Sistema de Gestión Documental

Sistema de información que incorpora, gestiona y facilita el acceso a los documentos de archivo a lo largo del tiempo. (Norma ISO/IEC 15489, 2000).

2.4 HIPÓTESIS DE LA INVESTIGACIÓN

2.4.1 HIPÓTESIS GENERAL

La Recuperación de la información empleando el Modelo de Espacio Vectorial incide positivamente en la Gestión Documentaria para la Unidad de Resoluciones de la UNA PUNO.

2.4.2 HIPÓTESIS ESPECÍFICOS

- El prototipo de gestión documentaria permite administrar positivamente el flujo de documentos para la Unidad de Resoluciones.

- Con la implementación de la recuperación de la información empleando el modelo de espacio vectorial permitirá mayor pertinencia y precisión en la búsqueda de la información.

2.5 OPERACIONALIZACIÓN DE VARIABLES

	DIMENSIONES	INDICADORES
Variable Independiente		
Recuperación de la información utilizando el Modelo de Espacio Vectorial	Representación de la información	* Representación vectorial del documento
	Representación de la consulta	* Representación vectorial de la consulta
	Búsqueda de documentos	* Normalización de vectores * Calcular el peso de cada elemento del vector * Calcular el grado de similitud entre los vectores
	Obtención de resultados	* Cálculo del coseno del ángulo de ambos vectores
	Evaluación de los resultados	* Colección de pruebas, precisión - exhaustividad
Variable Dependiente		
Gestión Documentaria para la Unidad de Resoluciones de la UNAP	Representación de los documentos	* Base de datos documental * Cantidad de documentos * Indización y clasificación
	Calidad de los documentos	* autenticidad * fiabilidad * integridad * accesibilidad * disponibilidad
	Clasificación de documentos	* Registro de documentos * Codificación de los documentos * Varios tipos de formato * Documentos confidenciales
	Almacenamiento de documentos	* Volumen y tasa de crecimiento * Métodos de protección y copia * Costos relativos de almacenamiento * Accesibilidad
	Seguridad en la documentación	* Patrones de seguridad y protección * Plan de recuperación de documentos * Categorías de acceso * Restricciones

Tabla Nº 3: Operacionalización de variables

Fuente: Elaboración Propia

CAPITULO III

DISEÑO METODOLÓGICO DE LA INVESTIGACIÓN

3.1 TIPO Y DISEÑO DE INVESTIGACIÓN

3.1.1 TIPO DE INVESTIGACIÓN

El tipo de problema de investigación es; experimental categoría cuasiexperimental**,** puesto que dentro de esta categoría se manipula la variable independiente, además el grupo de investigación se formó antes del experimento.

3.1.2 DISEÑO DEL PROBLEMA DE INVESTIGACIÓN

El diseño del problema se ha enfocado desde el modelo propuesto, realizándose el:

- Análisis de la situación actual en que se encuentra el sistema.
- Verificación de los procesos de tránsito de documentos.
- La elaboración de los documentos (resoluciones).
- La distribución de las resoluciones.

Respecto a la *variable Independiente*: Recuperación de la Información se ha enfocado de la siguiente manera:

- La implementación del Zend Framework 2.5, como plataforma de soporte para la Recuperación de la Información.
- Adecuación de la libreria Lucene en el Zend Framework, como aplicación para la búsqueda de información, el mismo que hace uso del Modelo de Espacio Vectorial.

- Asignación de campos de la Base de Datos *Sisresol* a Lucene, para la búsqueda de información.

Respecto a la *variable Dependiente*: Gestión Documentaria para la Unidad de Resoluciones, se ha enfocado:

- El diseño de la base de datos, Modelo Entidad - Relación.
- Diseño de los procesos y procedimientos utilizando metodologías y técnicas como son los casos de uso, diagrama de secuencias.
- Diseño de la interface del sistema.

Modelo del problema de investigación

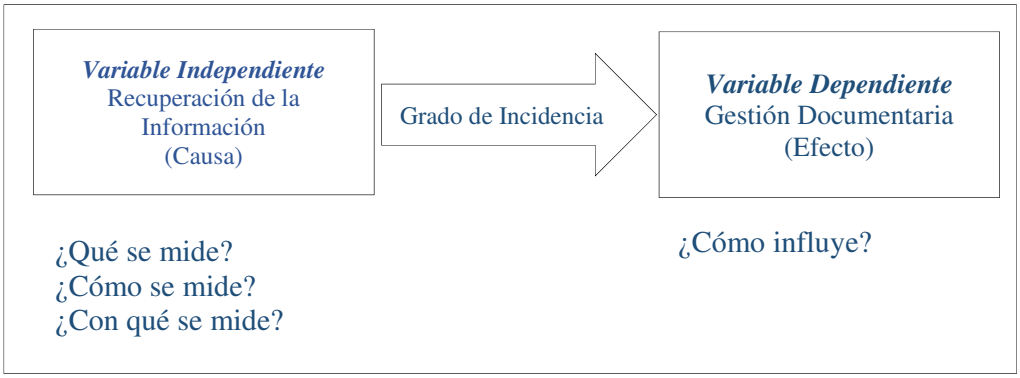


Figura Nº 10: Modelo que relaciona las variables a investigarse
Fuente: Elaboración propia.

3.2 POBLACIÓN Y MUESTRA DE INVESTIGACIÓN

Cantidad total de documentos electrónicos (resoluciones) por año que se elabora es más de cuatro mil, es decir cuatro mil usuarios sean estos de internos (Unidades académicas, administrativas), y externos (personas jurídicas, naturales, entidades, organizaciones privadas, etc.). Por lo que la población considerada en la investigación es 4000 mil documentos que se emite en el año.

3.3 MUESTRA DE LA INVESTIGACIÓN

La técnica de muestreo utilizada es el muestreo probabilístico en la modalidad de Muestreo Aleatorio Simple.

El nivel de confianza está considerado entre +2 y -2 sigmas de la Curva de Distribución Normal de Gauss, a partir de la media, está incluido el 95% de la población. El Error de estimación, se considera el máximo error tolerable, que es de 5%.

Para la determinación del tamaño de la muestra, se emplea la siguiente fórmula:

$$n = \frac{Z^2 * p * q * N}{e^2(N-1) + Z^2 * p * q} \qquad (12)$$

Donde:

n: Número de elementos que debe poseer la muestra.
Z^2: Nivel de confianza.
p: Probabilidad que ocurra un evento.
q: Probabilidad que no se realice el evento.
e: Error permitido.

Calculando:

$$n = \frac{1.92^2 * 0.50 * 0.50 * 4000}{0.05^2(4000 - 1) + 1.92^2 * 0.50 * 0.50}$$

n = 337

La muestra considerada será de 337 resoluciones.

3.4 UBICACIÓN Y DESCRIPCIÓN DE LA POBLACIÓN

Toda la información (cantidad de documentos) analógica y digital es facilitada y proporcionada por la Unidad de Resoluciones de la Oficina de Secretaría General de la UNA, información obtenida *in situ*.

3.5 TÉCNICAS E INSTRUMENTOS PARA RECOLECTAR INFORMACIÓN

Respecto al uso de técnicas de recolección de información en el proceso de elaboración de la resolución, se realizará la entrevista a cada uno del personal que labora en dicha unidad.

Para la investigación se hará uso de los siguientes instrumentos presénciales

- Entrevistas al personal que labora en la unidad, a través de cuestionarios
- Seguimiento del proceso de elaboración de la resolución.

3.6 TÉCNICAS PARA EL PROCESAMIENTO Y ANÁLISIS DE DATOS

Tratamiento de los datos

- Preparar la información para facilitar su posterior análisis
- Fases: Codificación de la información
- Almacenamiento de datos

Estrategias de análisis

- Metodología UML
- Ingeniería de requisitos
- Ingeniería de software

3.7 PLAN DE TRATAMIENTO DE DATOS

Para determinar el comportamiento y el tratamiento de los datos nos valdremos de la Estadística Inferencial, puesto que buscamos obtener información sobre una determinada población basándonos en el estudio de datos de una muestra tomada a partir de ella.

El Sistema de Gestión Documentaria, determinará la efectividad del proceso en la elaboración de resoluciones, es decir el manejo de esta variable, con alguna distribución de probabilidad, es decir medirá la rapidez en la elaboración de resoluciones promedio (que es un parámetro (μ) de dicha distribución). De manera específica, interesa mencionar si esta rapidez promedio en la elaboración de resoluciones, es de 15 resoluciones por día (con el proceso actual con que se cuenta).

3.8 DISEÑO ESTADÍSTICO PARA LA PRUEBA DE HIPÓTESIS

El sistema de Gestión Documentaria influye en la búsqueda de información.

El planteamiento formal de la situación se realiza en términos de:

- **Formulación de la Hipótesis**
  - Hipótesis nula H_0 (Proposición que se quiere poner a prueba)
    - Hipótesis a probar
  - Hipótesis alternativa H_1
    - Alternativa se aceptará si se rechaza la hipótesis nula.

Hipótesis Nula:	H_0:	$\mu = 15$resol/día
Hipótesis Alterna:	H_1:	$\mu > 15$resol/día

- **Identificación del Estadístico de prueba y su distribución**

  Estadístico de Prueba:

$$z = \frac{\bar{x}-\mu_0}{\sigma}\sqrt{n} \quad (13)$$

  Donde:

  z: Es la distancia desde la media en relación con la desviación estándar de la media.

  $\bar{x}$: Media de la muestra

  μ_0: Media de la población hipotética

  σ : Desviación de la población

  n : Tamaño de la muestra

  Distribución normal estándar

- **Nivel de significancia**

  Nivel de confianza 95%

  α = 0.05

  σ = 1 (no existe dispersión entre los datos)

  n = 337

  μ_0 = 15

  $\bar{x}$ = 16

- **Formulación de la regla de decisión**

  La aceptación de H_0, si un valor de la media muestral $\bar{X}$, es igual o menor a 15 resol/día, es una evidencia que se apoya en la hipótesis nula. De lo contrario, se acepta H_1.

- **Decisión estadística**

  Evaluar el estadístico de prueba

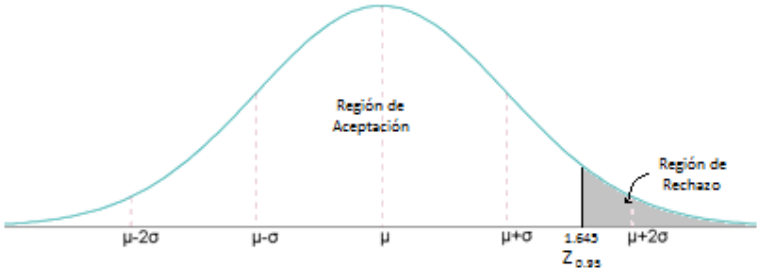


Figura Nº 11: La decisión estadística.
Fuente: Elaboración propia.

Calculando:

$$Z_{1-\alpha} \quad \frac{\bar{X}-\mu_0}{\sigma} \sqrt{n}$$

$$\frac{16-15}{1}\sqrt{337} = 18.35$$

Reemplazando valores obtenemos: 18.35, este valor pertenece a la región de rechazo, por lo que aceptamos la Hipótesis Alterna.

De donde concluimos que con la implantación de un Sistema de Gestión Documentaria el proceso en la elaboración y distribución de las resoluciones serán más eficientes y por tanto las búsquedas se harán de manera oportuna, empleando el modelo de espacio vectorial.

CAPITULO IV

ANALISIS E INTERPRETACIÓN DE RESULTADOS DE LA INVESTIGACIÓN

4.1 MODELO INCREMENTAL

Para el desarrollo del Software, se ha hecho uso del Modelo Incremental, porque combina elementos del Modelo Lineal Secuencial con la Filosofía Interactiva de construcción de prototipos, este modelo aplica secuencias lineales de forma escalonada mientras progresa el tiempo calendario, Cada secuencia lineal produce un incremento de software. El primer incremento generalmente es un producto esencial denominado núcleo.

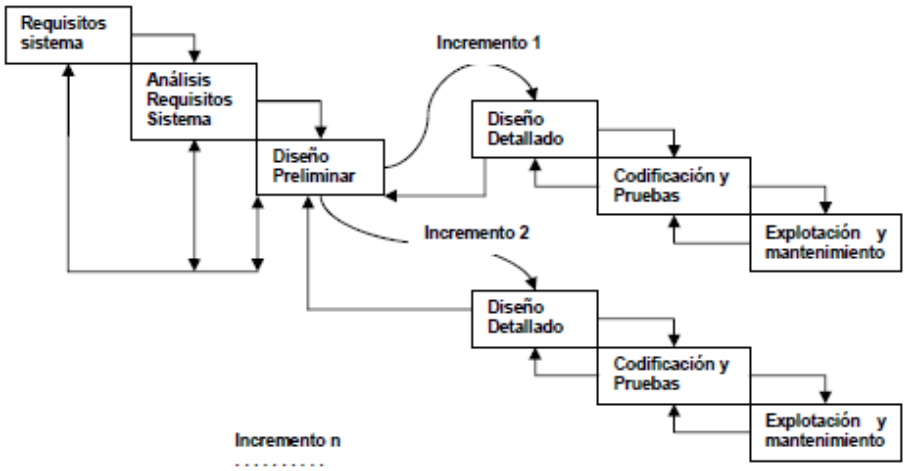


Figura Nº 12: Modelo incremental
Fuente: Universidad la Salle, 2015, Ingeniería de Software.

4.2 LENGUAJE DE MODELO UNIFICADO - UML

El Lenguaje de Modelo Unificado, es la sucesión estándar de una serie de métodos de análisis y diseños orientados a objetos. UML es llamado un lenguaje de modelado. El Lenguaje de modelado es la notación (principalmente gráfica) que usan los métodos para expresar un diseño. El proceso indica los pasos que se deben seguir para llegar a un diseño.

a) Requerimientos del Sistema

La fase de requerimientos es la primera en desarrollarse, el objetivo es delimitar el sistema y capturar los requisitos y características principales con que contaría la herramienta software. Para la formulación de los requisitos, ha sido necesario la entrevista con todos los miembros de la Unidad de Resoluciones, con el fin de revisar, ver las necesidades de estos, permitiendo así recoger los requerimientos preliminares y básicos para el desarrollo del software.

Durante esta fase se conceptualizaron las características principales del sistema, a través de los cuales se identificaron los diferentes actores y casos de uso de este, permitiendo de esta manera a través de los actores representar los distintos roles que los usuarios realizan en el sistema.

Con la especificación de cada uno de los casos de uso, se obtuvo un mayor detalle de los requerimientos y la funcionalidad del sistema, lo que permitió crear las primeras interfaces, al ejecutar cada uno de los casos de uso, permitiendo aclarar los requerimientos y evitando errores en las posteriores sucesiones de la implementación.

b) Requisitos candidatos

Determinado los requisitos funcionales del sistema se espera definir el comportamiento entradas, procesos, salidas, cálculos, detalles técnicos, manipulación de datos y características del sistema, los requisitos no funcionales especifican las propiedades emergentes del sistema, la

respuesta en el tiempo, capacidad de almacenamiento, software, hardware necesarios para la implementación.

c) Requerimientos funcionales

A continuación, se presenta una tabla con el listado de los requerimientos funcionales, candidatos para la comprensión del software. La tabla estará compuesta por los siguientes campos:

- Id del requerimiento.
- Nombre del Requerimiento.
- Descripción.

Id	Nombre del Requerimiento	Descripción del Requerimiento
RF1	Registro de los usuarios del sistema	Permitir que los miembros/usuarios de la Unidad de Resoluciones y Certificaciones URC, puedan registrarse como usuarios del sistema
RF2	Autenticación para el ingreso al sistema.	Identificarse mediante un login (usuario y contraseña) a los miembros de la URC, para que puedan acceder al sistema y hacer uso de este como usuario.
RF3	Cambiar Contraseñas	Permitir a los usuarios del sistema, poder cambiar su contraseña a fin de resguardar el acceso.
RF4	Ingreso de documentos origen a la base de datos	Permitir que el documento fuente se ingrese a la base de datos.
RF5	Disponibilidad y clasificación de datos nativos	Hace que del documento fuente se disponga para su distribución, clasificación y su posterior elaboración.
RF6	Procesamiento de la documentación de acuerdo a la disponibilidad y clasificación de datos nativos	Elaborar la resolución rectoral o la certificación en función al documento respectivo de acuerdo a la clasificación y la responsabilidad en la elaboración.
RF7	Mostrar la producción de los miembros del grupo.	Determina el reporte de la cantidad de documento elaborado por cada uno de los integrantes del grupo.
RF8	Manejo de la producción	Una vez elaborado la documentación –resolución

		rectoral- se determina el almacenamiento en la base de datos y el propósito que cumplirá como documento.
RF9	Publicación de la Documentación.	Desde que se almacena en la base de datos los documentos, antes de la publicación estos son confrontados y validados posteriormente publicados en la web.
RF10	Administración de la publicación de la documentación	Durante la publicación el administrador puede subir, eliminar, la información documentaria, bajo un criterio determinado.
RF11	Búsqueda y descarga de la información haciendo uso del Web.	El usuario externo o el interesado por la resolución rectoral puede hacer la búsqueda en la web, posteriormente si desea descargar esta información para sus intereses.
RF12	Manejo de los archivos de información	Estos archivos de información contendrán la documentación elaborada, para mantener integra estos, se deberá hacer el backup respectivo cada cierto período de tiempo.

Tabla Nº 4: Requerimientos Funcionales.
Fuente: Elaboración propia.

d) Requerimientos no Funcionales

Mediante este se pretende definir las propiedades y restricciones que tendrá el sistema, como funcionalidad, usabilidad, confiabilidad, compatibilidad con hardware, software y especificaciones del sistema.

Tipo de requisito	Descripción
Interfaz del Usuario Interno	• Desarrollo de una interface de usuario interno orientado a la Web. • Interfaz dinámica acorde a las funciones que cumplen cada uno de los miembros de la Unidad de Resoluciones • Interacción eficiente entre la información que ingresa y sale de la

	base de datos con el proceso que realiza el usuario. • Cumplir con patrones como Modelo, Vista, Controlador y acceso a diferentes navegadores.
Interfaz del Usuario Externo	• Desarrollo de una interface de usuario externo orientado a la Web. • Interfaz dinámica que permita interactuar el usuario al realizar las búsquedas con la base de datos empleando el modelo vectorial. • Cumplir con los estándares de desarrollo de las interfaces Web y permitan accesibilidad.
Requerimientos de Hardware	• Necesidad de contar con un servidor en la Unidad de Resoluciones y no se tenga inconvenientes en el funcionamiento del sistema.
Requerimientos de Software	• El servidor tenga instalado las aplicaciones como PHP, Gestor y Administrador de Base de Datos.
Restricciones de implementación	• Se hizo uso del lenguaje de programación PHP, orientado a objetos.

Tabla Nº 5: Requisitos No Funcionales.
Fuente: Elaboración propia

e) Riesgos Críticos

A continuación, se presentan los riesgos críticos que se han identificado, que mediante en una tabla se da la descripción del riesgo, la probabilidad de ocurrencia, el impacto en el desarrollo del proyecto y el plan de contingencia.

Los campos que se consideran son:

ID Riesgo: ID identificación del Riesgo

Probabilidad de Ocurrencia: Alta/Media/Baja

Impacto de la ocurrencia: Crítico/Significativo/Desconsiderado

Descripción: Explicación del riesgo

Presentación del riesgo: Como se presenta el riesgo

Monitoreo: Encargado de realizar el seguimiento del riesgo

Contingencia: Función para mitigar el riesgo

Riesgo ID 01	Probabilidad: Media	Impacto: Significativo
Nombre: Falta de Conocimiento en las herramientas de desarrollo		
Descripción: Problemas con el manejo de las herramientas de programación utilizadas para el desarrollo del software, caso como PHP Orientado a Objetos, HTML, CSS, Javascript		
Presentación del Riesgo: • Problemas en la instalación de servidor Apache, PHP, SGBD • Limitaciones en el diseño • Limitaciones en la implementación • Complejidad en el desarrollo del sistema.		
Monitoreo: Evaluar los conocimientos de las herramientas en la elaboración del proyecto.		
Contingencia: • Estudio de los lenguajes de programación, gestor y administrador de base de datos, herramientas de diseño • Realizar la correcta documentación del proyecto • Consultar la documentación disponible de las herramientas.		

Tabla Nº 6: Riesgo ID 01: Falta de Conocimiento en las herramientas de desarrollo
Fuente: Elaboración propia.

Riesgo ID 02	Probabilidad: Media	Impacto: Critico
Nombre: El servidor no cuente con las aplicaciones y software debidos.		
Descripción: El servidor no cuente con el software y/o las últimas versiones requerido para la implementación del proyecto.		
Presentación del Riesgo: • Errores en la instalación de la plataforma del sistema operativo. • Errores en la ejecución del software de aplicación. • Inaccesibilidad al gestor y administrador de base de datos.		
Monitoreo: Gestión por parte de los administradores de servidores y elaborador del proyecto		
Contingencia: • Estar debidamente capacitado en la instalación de servidores. • Gestionar la instalación de las aplicaciones como PHP, Mysql, mediante el administrador de servidores. • Gestionar la actualización permanente de las herramientas software		

Tabla Nº 7: Riesgo ID 02: Riesgo de que el servidor no cuente con el software requerido.
Fuente: Elaboración propia.

Riesgo ID 03	Probabilidad: Alta	Impacto: Critico
Nombre: Modelamiento inadecuado para la implementación del sistema		
Descripción: El modelamiento inadecuado de los requerimientos y necesidades por parte de los usuarios, que no reflejan las necesidades requeridas en la implementación del sistema.		
Presentación del Riesgo: • Carencia en la concertación de necesidades del grupo durante el proceso de elaboración del sistema. • Los avances presentados no coinciden con las necesidades planteadas. • Modificaciones en el avance del software en etapas avanzadas en la elaboración del sistema lo que trae consigo demora y molestias.		
Monitoreo: Está a cargo el ejecutor del software, administrador o jefe de la Unidad de Resoluciones, miembros.		
Contingencia: • Establecer inicialmente los requerimientos consolidados de información, los mismos que se consoliden al final de la ejecución del proyecto. • Utilizar herramientas adecuadas como el UML, los mismos que permitan visualizar, especificar, construir y documentar un sistema robusto.		

Tabla Nº 8: Riesgo ID 03: Modelamiento inadecuado para la implementación del sistema.
Fuente: Elaboración propia.

Riesgo ID 04	Probabilidad: Alta	Impacto: Significativo
Nombre: Renuencia de los usuarios a utilizar el sistema		
Descripción: Renuencia, insatisfacción o indiferencia por la inutilidad que puede generar el sistema una vez concluido.		
Presentación del Riesgo: • Inadecuada estructuración en la formulación del modelo del sistema. • Inadecuada construcción y desarrollo del software. • Carencia de coordinación entre los usuarios y administrador de desarrollo de software. • Información no consolidada de ciertas acciones inherentes al sistema.		
Monitoreo: Está a cargo del analista de sistemas, ejecutor de software conjuntamente con los miembros de la Unidad de Resoluciones		
Contingencia: • Antes de la implementación, se debe realizar un análisis de la situación real de las necesidades que tiene el sistema. • Los procesos de uso del software deben ser lo más entendible y asequible al usuario, presentándose una interfaz atractiva. • Influir en los usuarios en el uso del software para el cual está diseñado y el propósito que este tiene en el cumplimiento de fines y objetivos de la Unidad de Resoluciones.		

Tabla Nº 9: Riesgo ID 04: Riesgo a la renuencia de los usuarios a utilizar el sistema.
Fuente propia.

4.3 MODELOS DE CASOS DE USO

Mediante los Casos de Uso podemos modelar los requerimientos funcionales del sistema, además dirigir el proceso durante todos los flujos de trabajo de las distintas fases de desarrollo.

a) Actores

Cada actor modelará e interactuará con el sistema que necesita intercambiar información. Los usuarios que interactúan con el sistema sean representados mediante el actor. Los actores no están restringidos a ser personas físicas, o representar sistemas externos que ejecutan tareas en el sistema.

En la siguiente figura se muestra la representación del sistema y los actores, representando al sistema como una caja negra y los actores como entes externos.

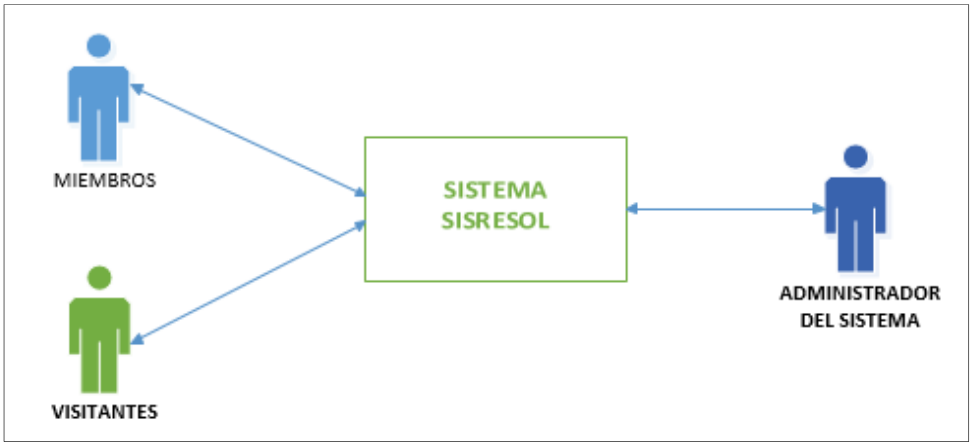


Figura Nº 13: Delimitación del sistema y posición de los actores

Fuente: Elaboración propia.

En las tablas siguientes se muestran la descripción de cada actor, determinándose el rol o función que cumple cada uno de ellos en el entorno del sistema y los Casos de Uso con los que el actor interactúa.

La composición de la tabla está dada por los siguientes campos:

- Id Actor : ID Identidad del actor.
- Actor : Nombre del actor.
- Descripción : Describe al actor en el entorno.
- Casos de Uso asociados: Determina la interrelación del actor mediante casos de uso.

ACT ID 01	Actor Visitante
Descripción:	Este hace referencia a todas las personas que acceden a la Web para consultar o realizar búsquedas de la documentación desde la base de datos, es aquí donde se emplea el Modelo de Espacio Vectorial.
Caso de uso asociado:	Mostrar la documentación que requiere el usuario, como las resoluciones rectorales, clasificados por nombre, fecha, número, parte resolutiva o interesado.

Tabla Nº 10: ACT ID 01 Especificación del actor visitante.
Fuente: Elaboración propia

ACT ID 02	Actor Miembro
Descripción:	Este hace referencia a los miembros integrantes de la Unidad de Resoluciones, registrados cada uno en sistema, los mismos que para acceder previamente se registran con un usuario y contraseña.
Caso de uso asociado:	Están asociados con el ingreso de información básico o fuente a la base de datos, distribución de la documentación a los miembros del grupo, elaboración del resolución, envían la información clasificada a la Web.

Tabla Nº 11: ACT ID 02 Especificación del actor miembro.
Fuente: Elaboración propia.

ACT ID 03	Actor Administrador
Descripción:	Esta encargado de realizar la dirección, control, administración y monitoreo permanente del sistema.
Caso de uso asociado:	Administrar la base de datos de la documentación fuente, controlar la distribución de la resolución y que estos se consoliden en la base de datos. Determinar que la elaboración de la documentación no presente errores durante este proceso, monitorear que la información web sea la más oportuna.

Tabla Nº 12: ACT ID 03 Especificación del actor administrador.
Fuente: Elaboración propia.

A continuación, se presentan los módulos elementales organizados luego en función a los módulos se organizan los casos de uso obtenidos a través de los requerimientos del sistema.

Figura Nº 14: Organización de los Módulos según Casos de Uso.
Fuente: Elaboración propia.

Módulo: Usuarios del Sistema

Contiene los casos de uso que se relacionan con los miembros y usuarios del sistema, en la figura se muestran los casos de uso principales.

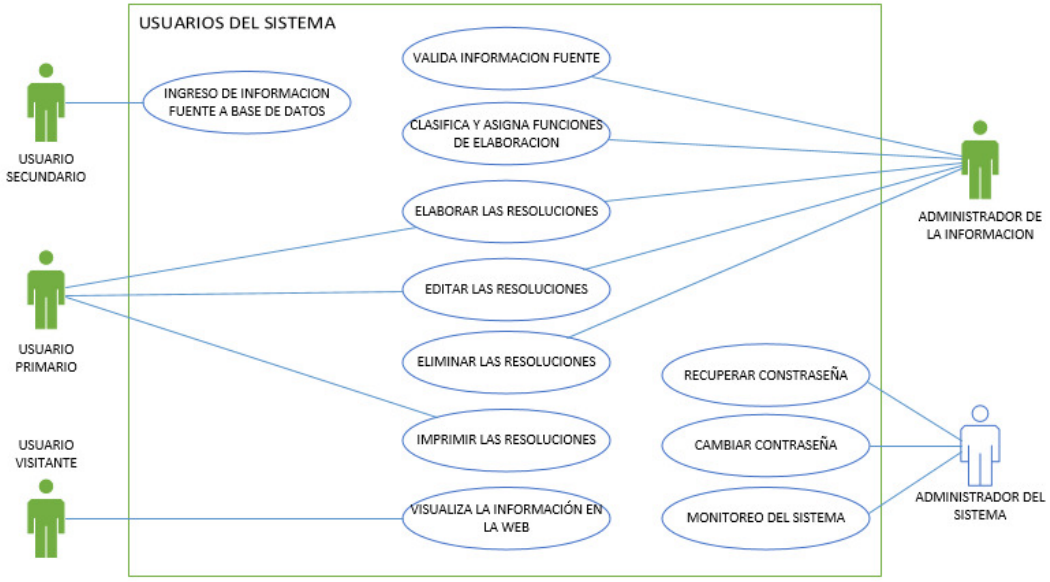


Figura Nº 15: Caso de Uso, Usuarios del Sistema.
Fuente: Elaboración propia.

Módulo: Ingreso de Información Básica y Referencial a la Base de Datos

Este módulo permite ver casos de uso cuando, el usuario secundario ingresa la información básica, elemental a la Base de Datos.

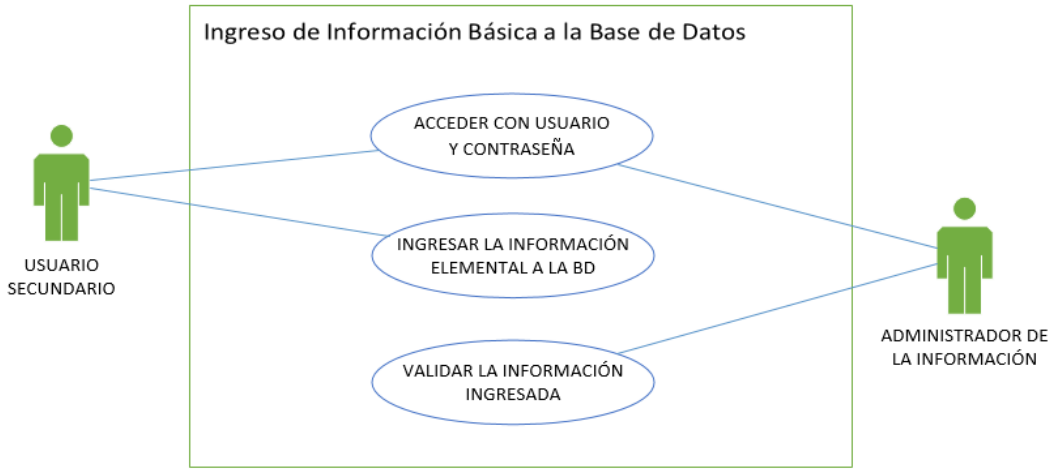


Figura Nº 16: Caso de Uso, Ingreso de Información Básica a la Base de Datos.
Fuente: Elaboración propia.

Módulo: Distribución y Elaboración de la Resolución Rectoral

Este módulo permite ver Caso de Uso, para la Distribución y elaboración de la Resolución Rectoral, responsabilidad del usuario primario.

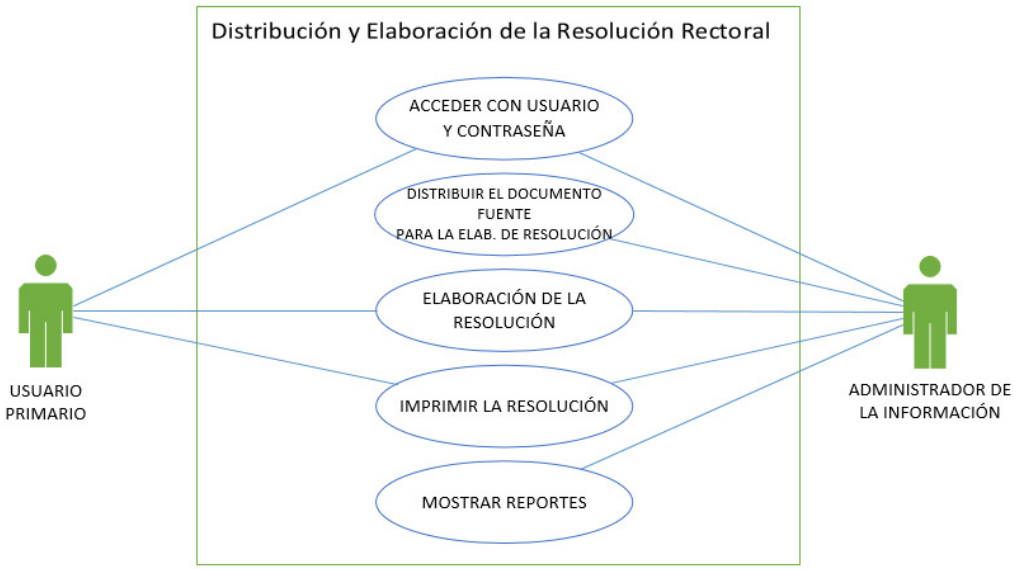


Figura Nº 17: Caso de Uso, módulo Distribución y Elaboración de la Resolución.
Fuente: Elaboración propia

Módulo: Búsqueda de la Información en la Web

En este caso son los usuarios visitantes los que realizan la búsqueda de la documentación de acuerdo a razones básicas.

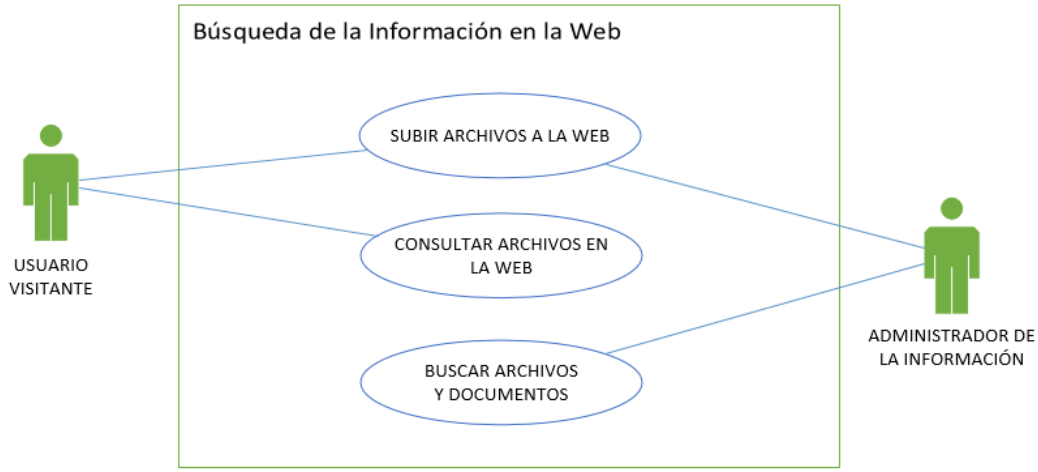


Figura Nº 18: Caso de Uso, módulo Búsqueda de la Información en la Web.
Fuente: Elaboración propia

4.4 MODELO DE DOMINIO DEL PROBLEMA

Las clases de dominio permiten crear un modelo de clases global, a través del cual se puede obtener una representación visual de los conceptos u objetos del mundo real. Se identificaron conceptos característicos del problema que permiten la identificación de los objetos del negocio, conceptos que tuvieran propiedades o atributos y relaciones con otros objetos del dominio.

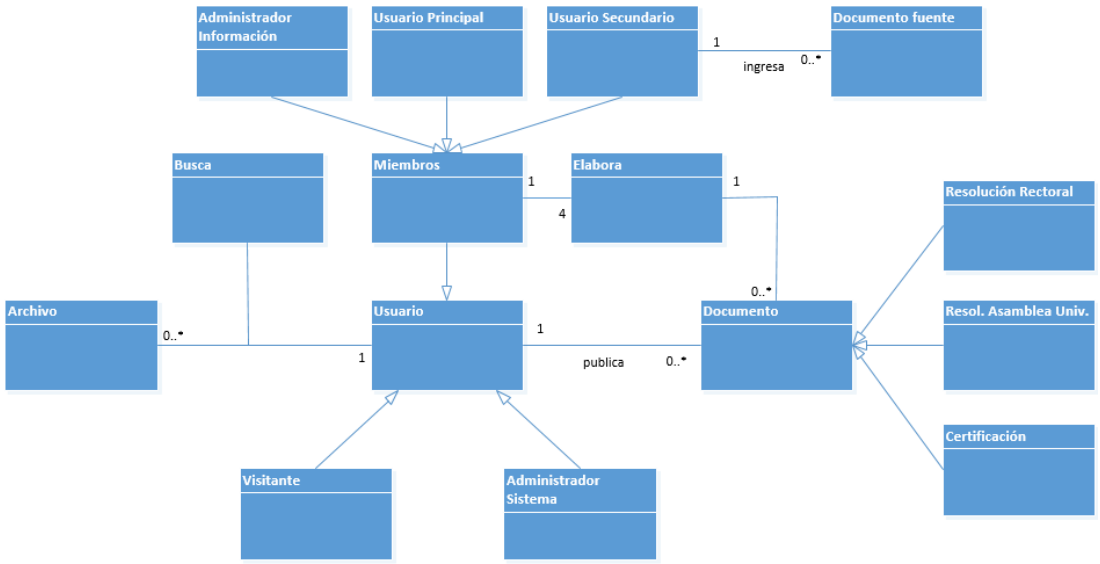


Figura Nº 19: Diagrama de clases de dominio.

Fuente: Elaboración propia.

Atributos

Se realizó la identificación de los atributos, el que permite entender mejor el dominio y la base para el desarrollo del proyecto. En la tabla que se presenta a continuación se enumeran los atributos para cada una de las clases del dominio.

Clases	Atributos
Usuario	Id, Usuario, contraseña, estado
Miembros	Id, Nombres, apellidos, email, tiempo de servicio
Administrador de la información	Clasifica_doc, clasifica_miembros
Usuario principal	Id, documento_asignado, fecha_asignacion
Usuario secundario	Id
Administrador del sistema	Id, nombres apellidos
Visitante	Id, nombres apellidos, email
Documento_fuente	Id, nro_doc, procedencia, fecha_recepcion, tipo_documento
Documento	Id, tipo_documento, nro_doc_fuente, nro_documento, fecha_documento, contenido,fecha_asignacion,fecha_emision,id_miembros,
Resolución Rectoral	Estado (elaborado, no elaborado)
Resolución Asamblea Universitaria	Estado (elaborado, no elaborado)
Certificaciones	Estado (elaborado, no elaborado)
Elabora	Id, id_adm_info, id_usuario_principal, id_documento
Archivo	Id, nombre, titulo, fecha, autor, texto
Busca	Tipo_documento,

Tabla Nº 13: Atributos que se identifican de acuerdo a sus clases.
Fuente: Elaboración propia.

4.5 DICCIONARIO DE CLASES

Permite hacer la descripción textual de las clases que componen el dominio del problema, manteniendo el lenguaje común entre las personas involucradas.

A continuación, se presenta una descripción de las clases que se identifican en el dominio del proyecto.

- Usuario: es aquel miembro del grupo que se encuentra registrado en el sistema, para el acceso al sistema, cuenta con un login (usuario y contraseña).
- Miembros: constituido por los usuarios que forman parte del grupo, cumplen varias labores específicas dentro del sistema.
- Administrador de la información: Usuario encargado de validar, distribuir, clasificar y designar la tarea para elaboración del documento.
- Usuario Principal: miembro encargado de elaborar la documentación de acuerdo a la designación.
- Usuario Secundario: miembro encargado de ingresar la información de procedencia de las unidades administrativas, académicas u otros a la base de datos.
- Administrador del Sistema: miembro encargado de monitorear el normal funcionamiento de todo el sistema.
- Visitante: es un usuario del sistema que se interesa por encontrar la información requerida.
- Documento Fuente: Es el documento de procedencia de una unidad académica, unidad administrativa, Consejo Universitario, Asamblea Universitaria u otros, base para la elaboración de la resolución.
- Documento: constituido por los tipos de resolución y certificación ha elaborarse.
- Resolución Rectoral: documento principal que constantemente es elaborado, constituido por el número de resolución, fecha, el contenido, las firmas y su posterior publicación.

- Resolución de Asamblea Universitaria: documento principal que periódicamente se elabora, frente a un cambio estructural, organizacional que tenga que ver con los intereses de la institución, posee el número de resolución, fecha, el contenido y las firmas respectivas.
- Certificaciones: documento relacionado con requerimientos de las unidades académicas. Constituido por fecha, nombre del interesado, contenido, firmas.
- Elabora: una vez clasificada y distribuida la información, comienza el proceso de la elaboración del documento, para ello se hace invoca a plantillas prediseñadas de Ms. Word, donde se considera el número de resolución, la fecha, el contenido.
- Archivo: es el documento elaborado (resolución o certificación), culminado refrendado con las firmas delas autoridades de la institución, el mismo que se muestra en la Web, y puede ser descargado para fines e intereses de los usuarios, antes de la descarga el usuario debe ingresar sus datos personales con los siguientes campos; nombres y apellidos, correo electrónico y DNI.
- Busca: clase que se aplica a la búsqueda de los archivos, es en esta, donde se emplea el modelo de espacio vectorial exclusivo para realizar búsquedas por cualquier tipo de campo. Para la búsqueda de consultas por parte del usuario externo se ha hecho uso del Zend Framework 2.5, con el API Lucene.

4.6 CLASES DEL DOMINIO SEGÚN MÓDULOS

- **Usuarios del sistema.**

  En esta figura se presenta las clases y atributos pertenecientes al módulo miembros del sistema.

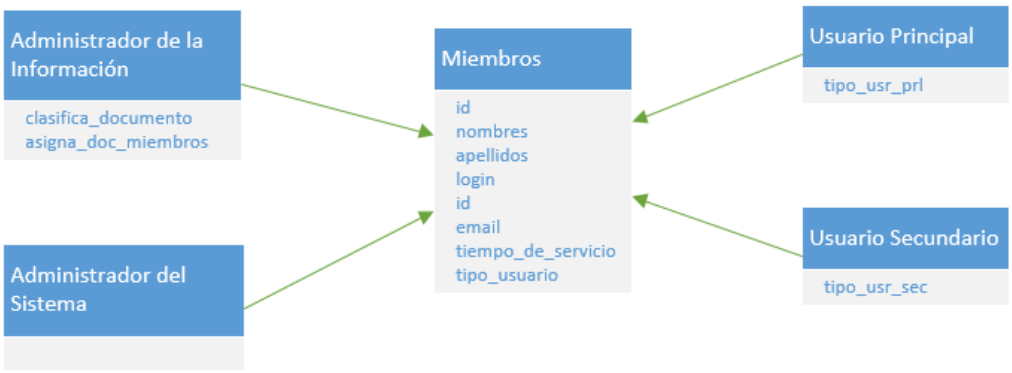


Figura Nº 20: Clases de dominio, módulo usuario del sistema.
Fuente: Elaboración propia.

- **Ingreso de la información básica y referencial a la base de datos.**

  Se presenta las clases y atributos del ingreso de información de procedencia a la base de datos por el usuario secundario.

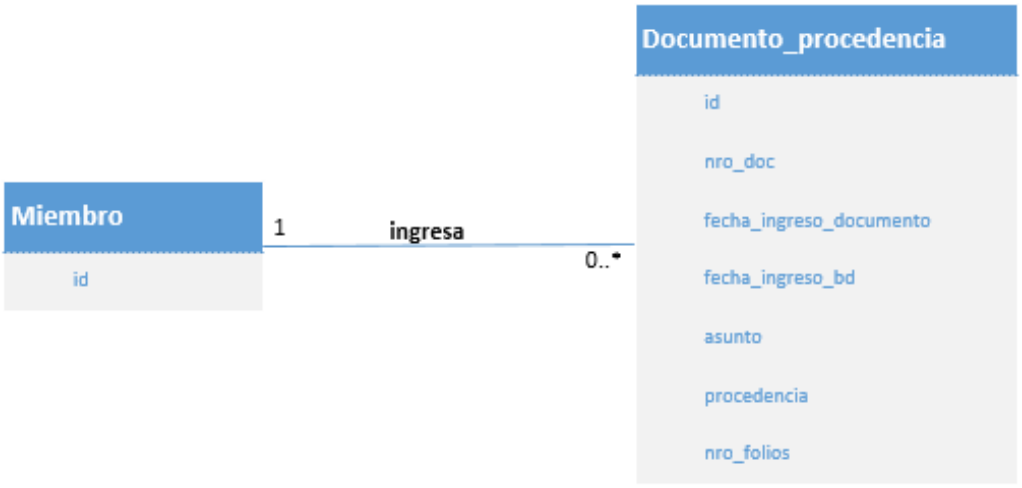


Figura Nº 21: Clases de dominio, módulo de ingreso de información básica.
Fuente: Elaboración propia.

- **Distribución de la elaboración de la documentación**.

  Se presenta las clases y atributos de la distribución para la elaboración de la documentación, cuya distribución es efectuada por el administrador de la información y la elaboración está a cargo de los usuarios principales.

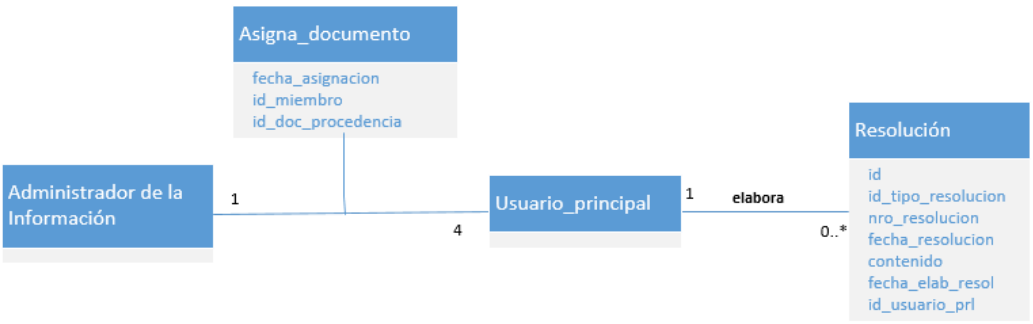


Figura Nº 22: Clases de dominio, Distribución y elaboración del documento.
Fuente: Elaboración propia.

- **Búsqueda de la información en la Web**.

  Se presentan las clases y atributos de las clases búsqueda y archivo, cuyo proceso de búsqueda es efectuado por el usuario visitante, aquí es donde se hace uso del modelo de espacio vectorial, la misma que es incluida en el API Lucene.

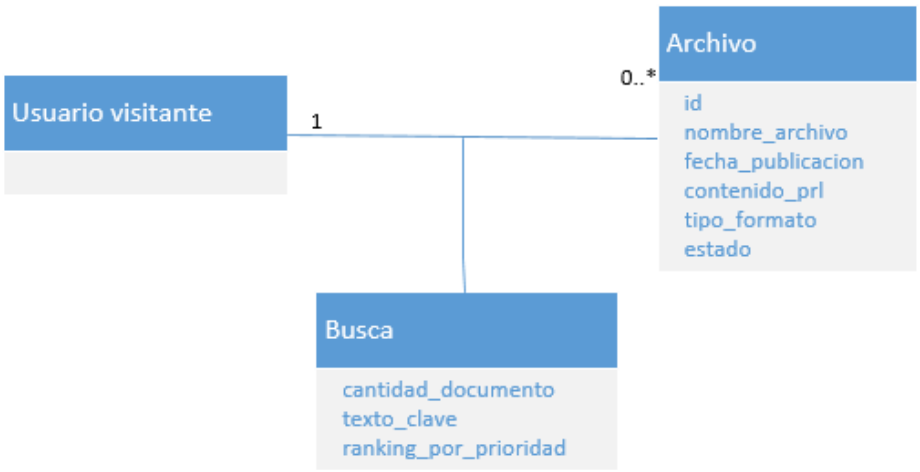


Figura Nº 23: Clases de dominio, módulo de búsqueda de información.
Fuente: Elaboración propia.

4.7 ANÁLISIS DEL SISTEMA

El objetivo de esta fase es comprender y generar una arquitectura de objetos para el sistema, basado en la fase de requerimientos. En esta fase se realizó la identificación de los diferentes objetos necesarios para la implementación de los diagramas de secuencia, la identificación de los objetos y la definición de su interacción permitió lograr la comprensión más precisa de los requisitos, lo que hace que esta fase sea de gran importancia ya que es la base para definir la arquitectura del sistema tanto estructural como funcional.

En esta fase de realizó los Diagramas de Secuencias, al mismo que se identificaron los módulos del sistema.

4.8 DIAGRAMAS DE SECUENCIAS

A continuación, se muestra el diagrama de secuencias para el caso de uso.

- **Ingreso de información básica de procedencia a la base de datos**.

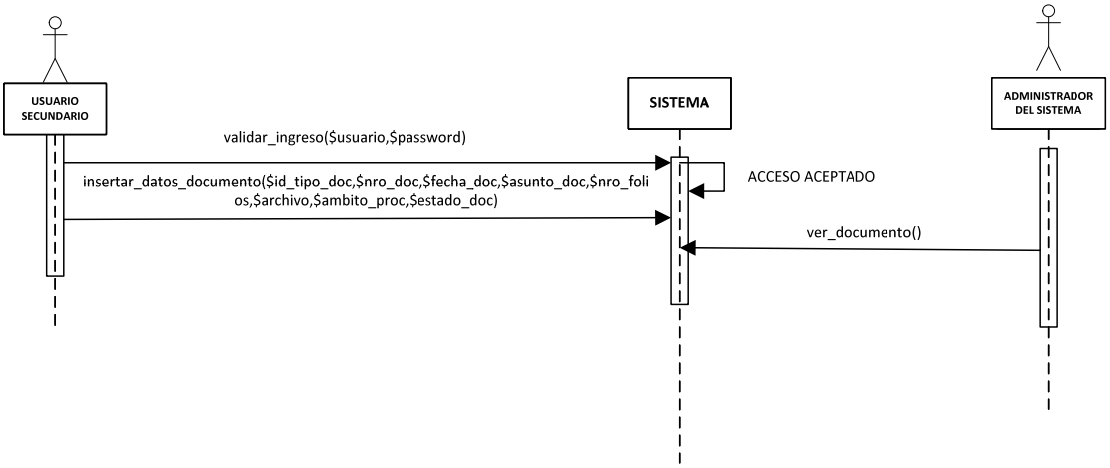


Figura Nº 24: Diagrama Secuencia – Ingreso de Información.
Fuente: Elaboración propia.

- **Distribución del documento.**

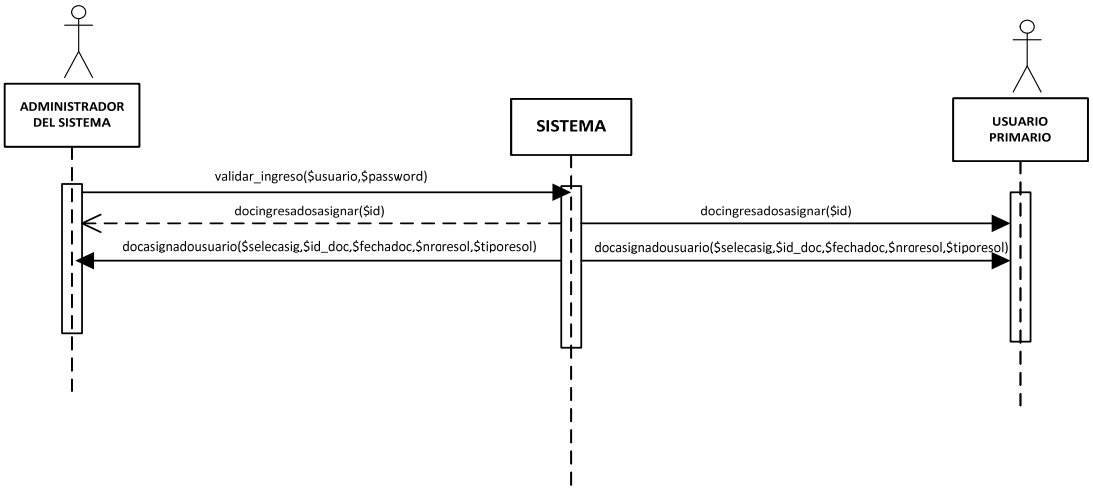


Figura Nº 25: Diagrama Secuencia – Distribución de documento.
Fuente: Elaboración propia.

- **Elaboración del documento.**

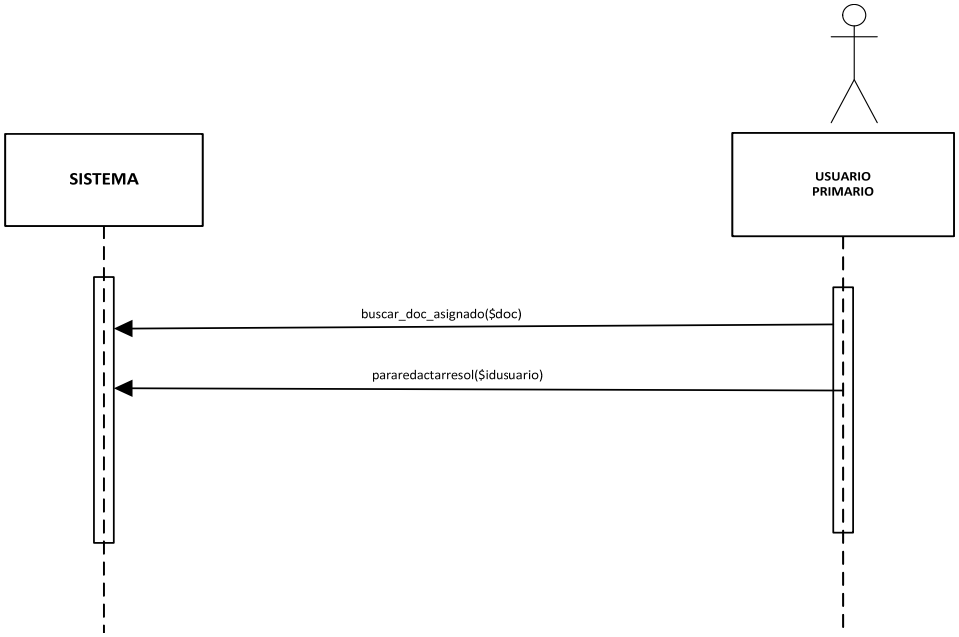


Fig. Nº 26: Diagrama Secuencia – Elaboración de Resolución.
Fuente: Elaboración propia

4.9 DISEÑO DEL SISTEMA – MODELO DE DATOS

En el diseño del sistema se considera al diseño de la base de datos como soporte a la estructura del flujo, manejo y gestión de la información.

- Diseño de la Base de Datos: Modelo Conceptual

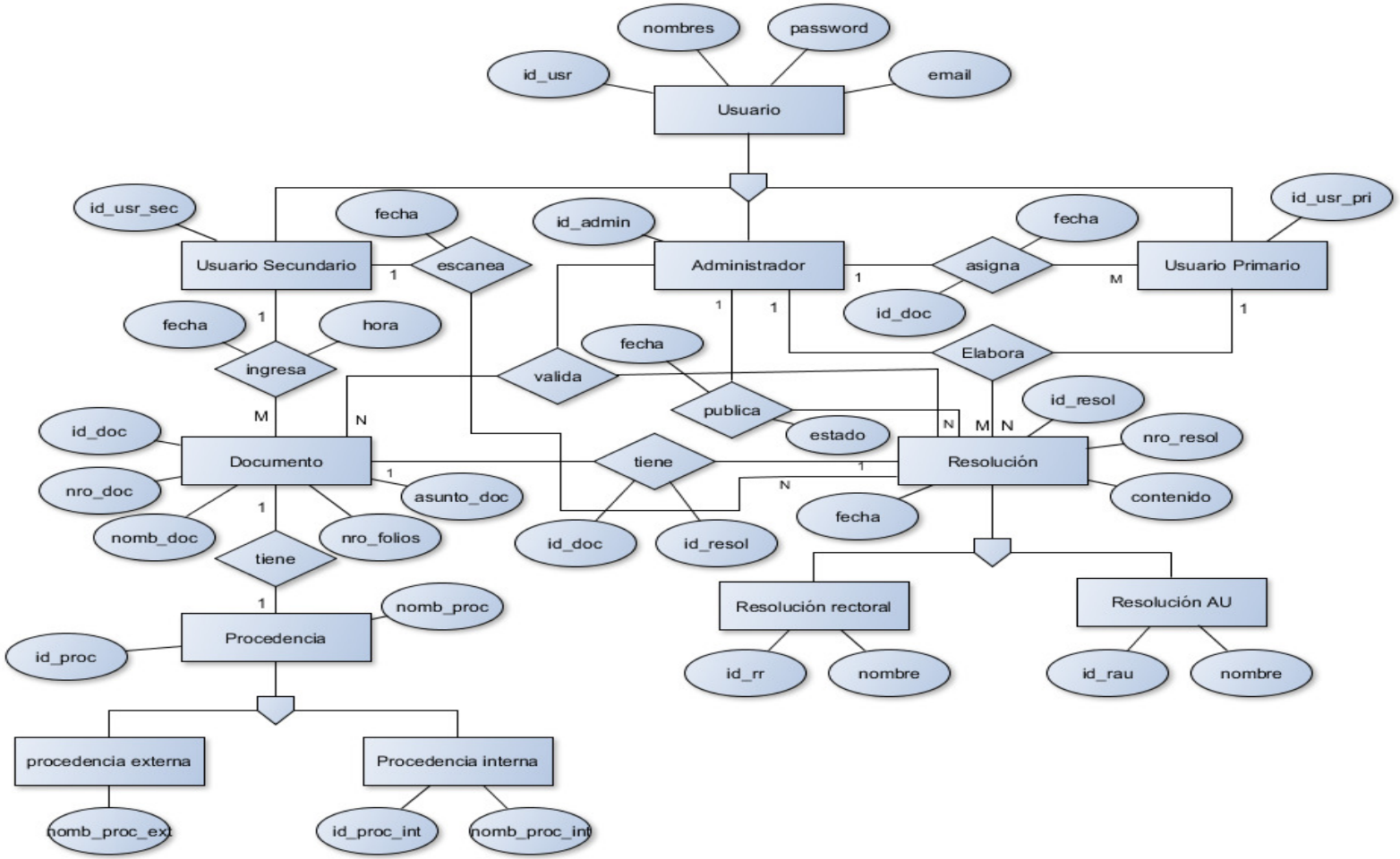


Figura Nº 27: Modelo Conceptual de la Base de Datos.
Fuente: Elaboración propia.

- Diseño de la Base de Datos de Modelo Lógico

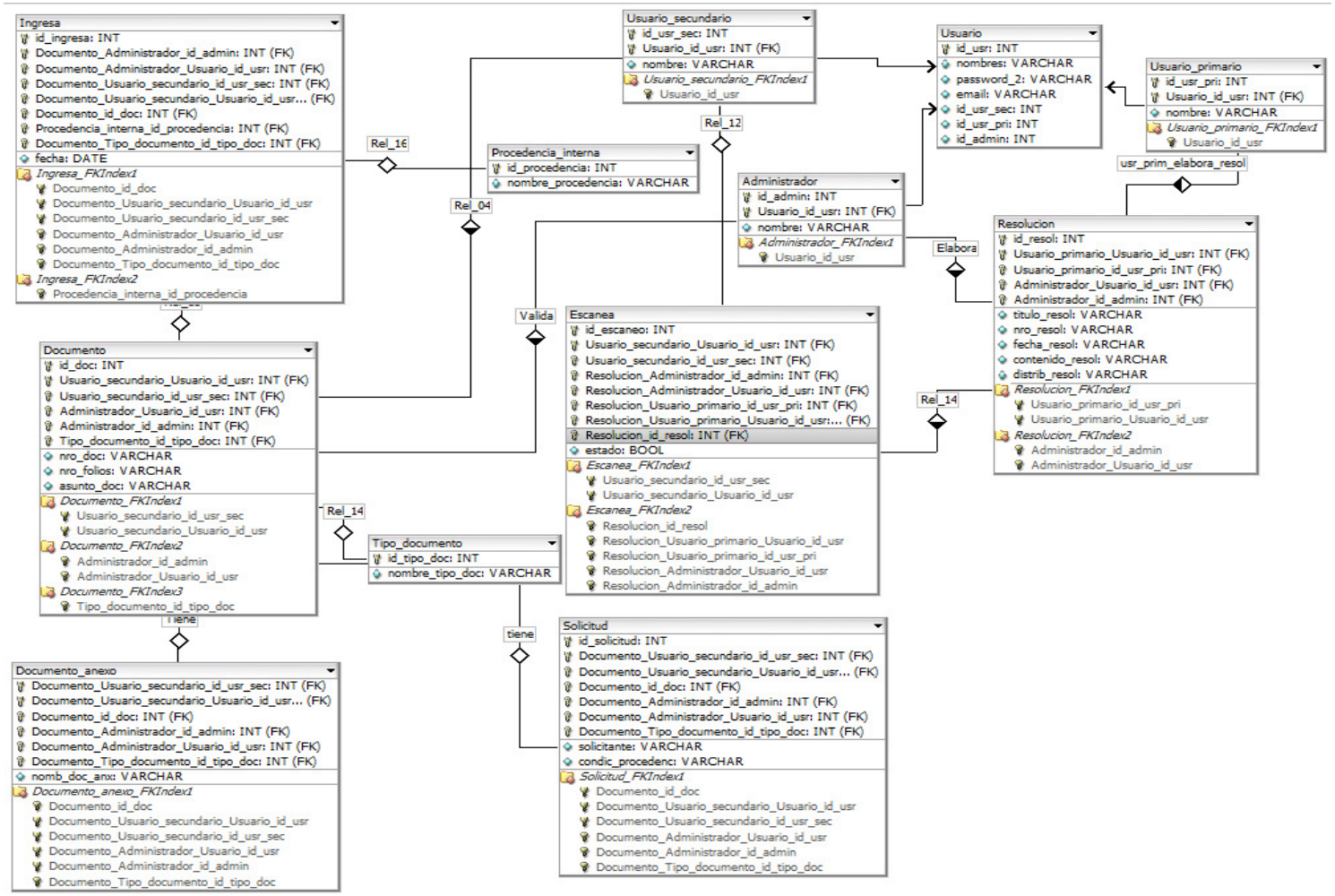


Figura Nº 28: Modelo Lógico de Datos
Fuente: Elaboración propia.

4.10 IMPLEMENTACIÓN DEL SISTEMA

Para la implementación del sistema se ha realizado:

- Análisis del sistema, el mismo que consiste en determinar todos los procesos de la elaboración de una resolución, con los responsables de la Unidad de Resoluciones.
- Diseño, está en función al proceso anterior a la realidad en que se encuentra actualmente, llegándose a diseñar el Diseño de la Base de Datos – Modelo Conceptual y Modelo Lógico.
- Codificación, Para la codificación del sistema se utilizó PHP con el Gestor de Base de Datos, contenidos en el compilado XAMPP, al mismo tiempo para el diseño de la interface se utilizó sentencias CSS, también se hizo uso de AJAX, las librerías de phpword, para generar archivos “resoluciones” Word, para el desarrollo y la implementación de la búsqueda de la información se hizo uso del Zend Framework 2.5 y al mismo tiempo se hizo la implementación de la librería API Lucene.

Pruebas del Sistema

En el desarrollo del prototipo del sistema permanentemente se ha ido realizando pruebas, dándose consistencia a este y el mismo que determina la robustez de la información.

Interfaces del sistema

Diseño de las interfaces del sistema

Para el diseño de las interfaces del sistema, se utilizó los estilos en cascada (CSS), puesto que HTML queda relegado en algunos aspectos de diseño, lo que permitió separar el diseño, del código HTML de la página. Utilizando hojas de estilo, se permitió definir colores, fuentes, espacios, para la mejor presentación de las interfaces y estandarizándose y uniformizándose estas.

Para el proyecto se diseñaron cuatro tipos de plantillas. A continuación, se presenta.

- **Interfaz de Acceso al Sistema**.

  Todos los usuarios al sistema acceden mediante un logueo y todos lo realizan por esta misma ventana.

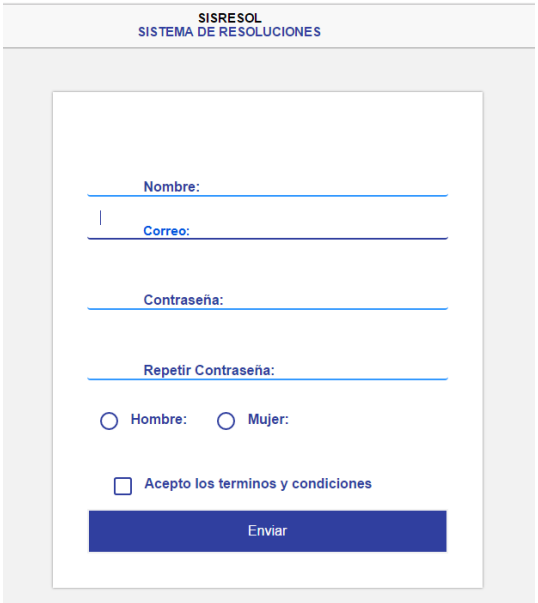


Figura Nº 29: Login de usuario.
Fuente: Elaboración propia.

- **Interfaz del ingreso de información.**

Información la que proviene de la Oficina de Secretaria General, es decir la petición para la generación de Resolución.

Figura Nº 30: Presentación de la interface de ingreso a la información.
Fuente: Elaboración propia.

Menú: Documento

Muestra toda la documentación que se ingresa al sistema, sea interna o externa en el lado derecho se observa los campos “Editar” y “Ver Documento”.

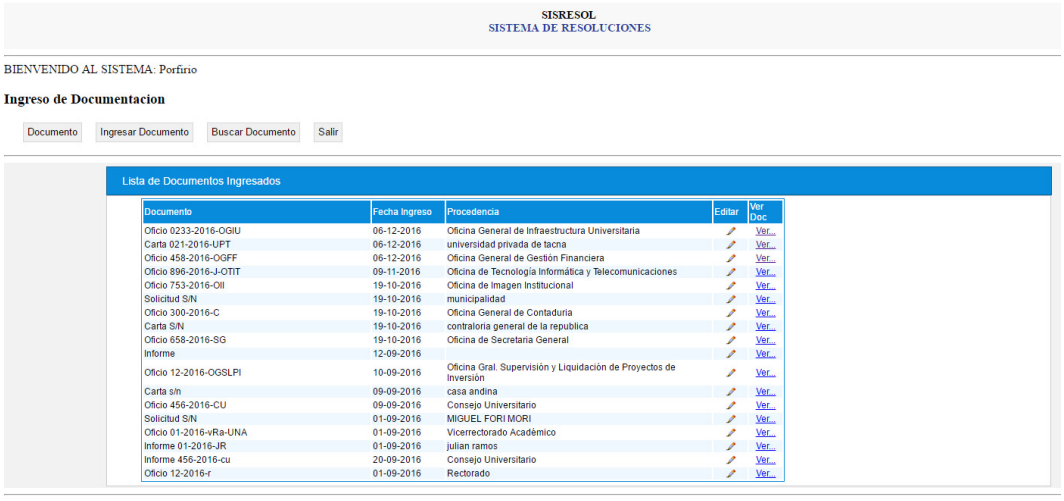


Figura Nº 31: Relación de la documentación ingresado al sistema.
Fuente: Elaboración propia.

Menú: Ingresar Documento

Aquí se ingresan las solicitudes mediante diferentes tipos de documentos.

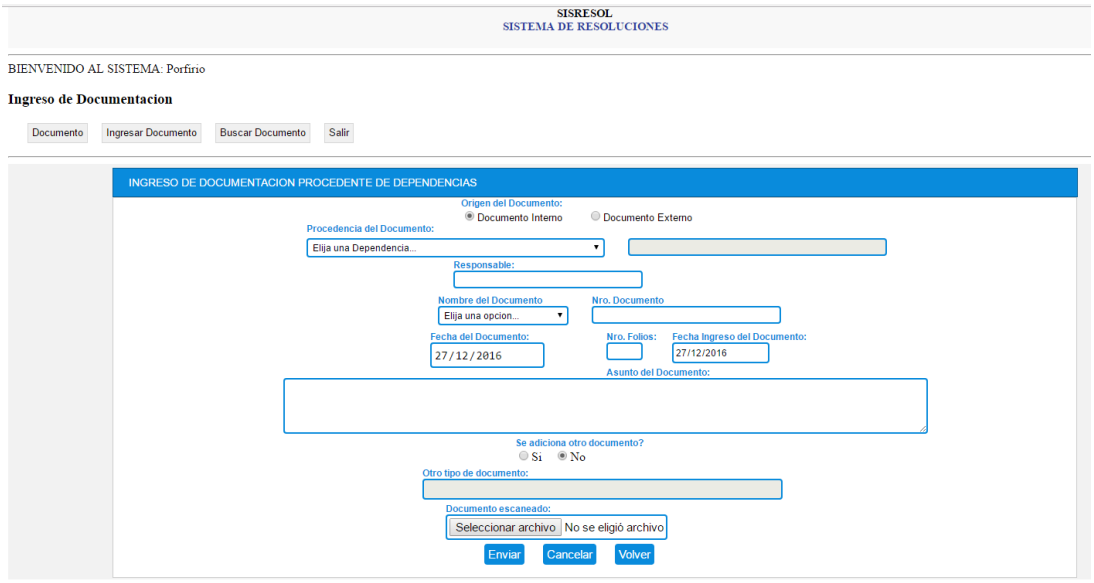


Figura Nº 32: Interface para el ingreso de la documentación al sistema.
Fuente: Elaboración propia.

Menú: Buscar Documento

Opción que permite buscar documento para su edición

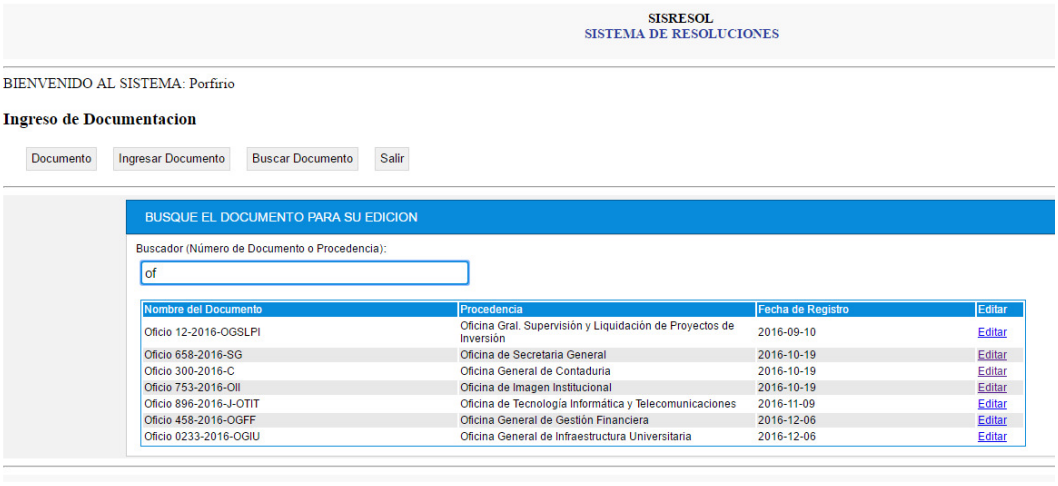


Figura Nº 33: Interface para la búsqueda del documento.
Fuente: Elaboración propia.

- **Interface de elaboración de la documentación**.

Figura Nº 34: Interface de la administración del documento.
Fuente: Elaboración propia.

Menú: Documentos Ingresados

Muestra los documentos ingresados en la fase anterior, en los dos últimos campos se encuentra “asignar” y “mostrar”

Figura Nº 35: Documentos ingresados.
Elaboración propia.

Enlace/campo: Asignar

Muestra el documento que se ingresa y el usuario al cual se le va asignar el documento para la elaboración de la resolución.

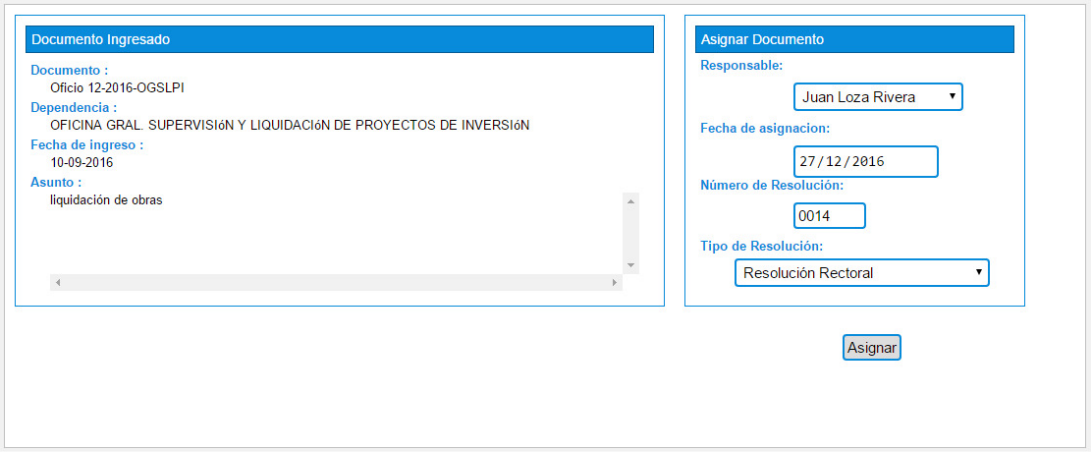


Figura Nº 36: Interface de documentos que se asignan al responsable.
Fuente: Elaboración propia.

Enlace: Mostrar

Muestra el documento ingresado, y el documento escaneado

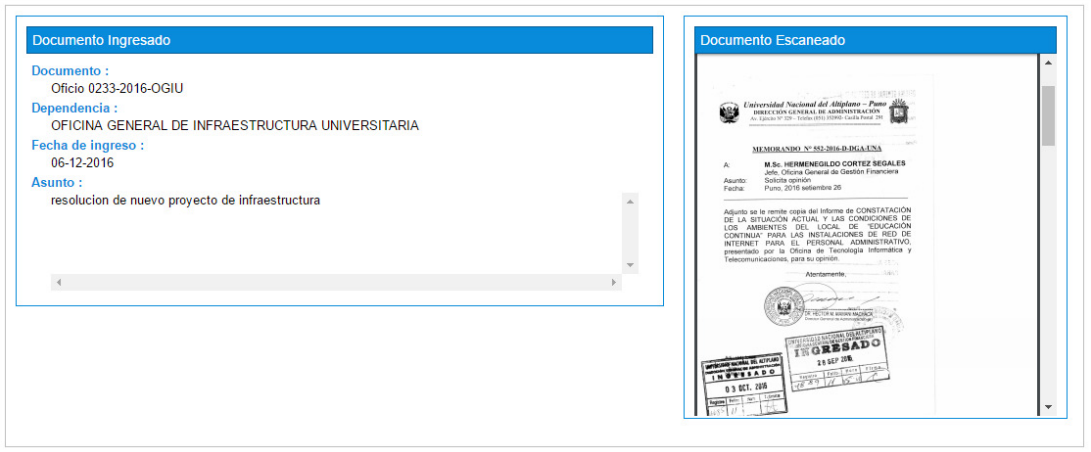


Figura Nº 37: Interface del documento ingresado.
Fuente: Elaboración propia.

Menú: Búsqueda de Documentos Asignados

Realiza la búsqueda para poder editar la asignación, sea por el número del documento o responsable.

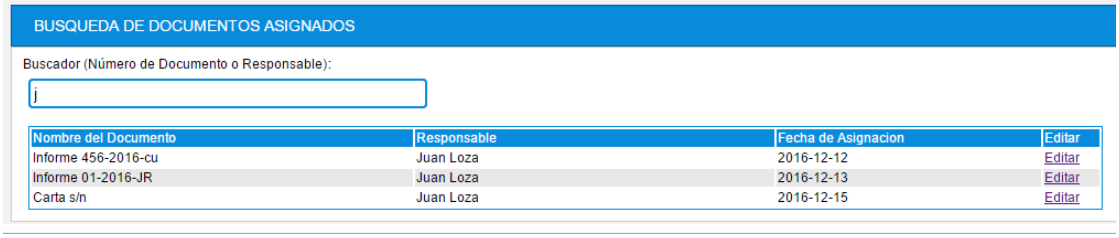


Figura Nº 38: Interface de búsqueda de documentos asignados.
Fuente: Elaboración propia.

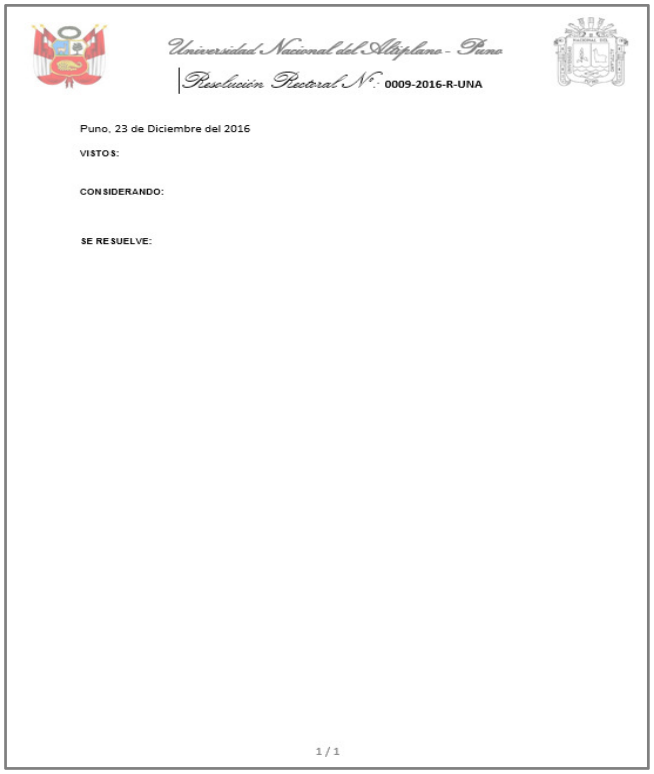
Menu: Generar Resolución

Permite generar la resolución en formato Word para su edición.

Figura Nº 39: Interface que permite generar Resolución.
Fuente: Elaboración propia.

Enlace: Generar Resolución

Genera en documento Word la Resolución listo para su edición



Universidad Nacional del Altiplano - Puno

Resolución Rectoral N°: 0009-2016-R-UNA

Puno, 23 de Diciembre del 2016

VISTOS:

CONSIDERANDO:

SE RESUELVE:

1 / 1

Figura Nº 40: Documento nuevo generado.
Fuente: Elaboración propia.

Búsqueda de Resoluciones, aplicando Lucene

Para la Búsqueda de Documentos se hizo uso del Zend Framework 2.5, porque este en su librería posee una Interfaz de Programación de Aplicaciones API Lucene, de código abierto para la Recuperación de la Información. Esta API es útil para cualquier aplicación que requiera indexado y búsqueda a texto completo y es ampliamente usado por su utilidad en la implementación de motores de búsqueda (Fundación de Software Apache).

Este motor de búsqueda utiliza el modelo de Espacio Vectorial.

Se ha utilizado la base de datos *Sisresol*, *Tabla Resoluciones* del Proyecto, para la búsqueda de documentos.

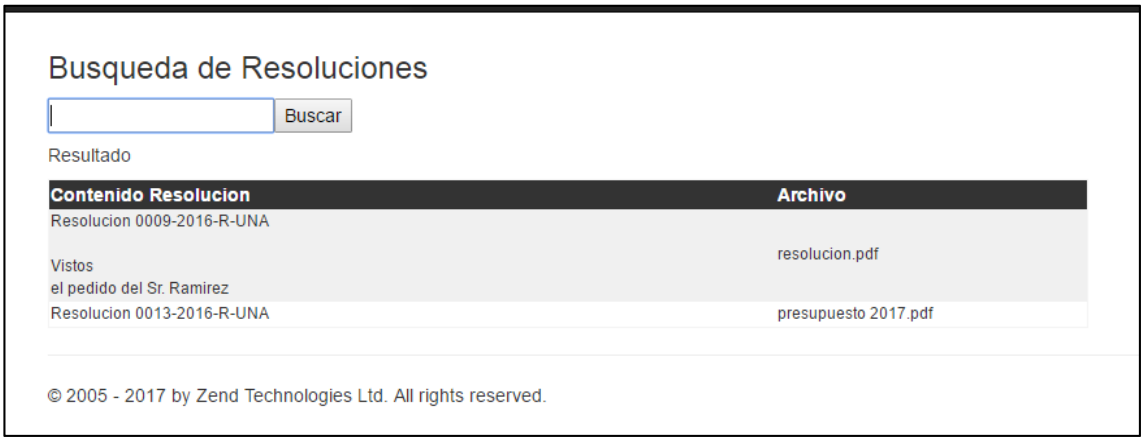


Figura Nº 41: Búsqueda de Resoluciones con Lucene.
Fuente: Elaboración propia.

Generación de índices con Lucene, durante el proceso de búsqueda

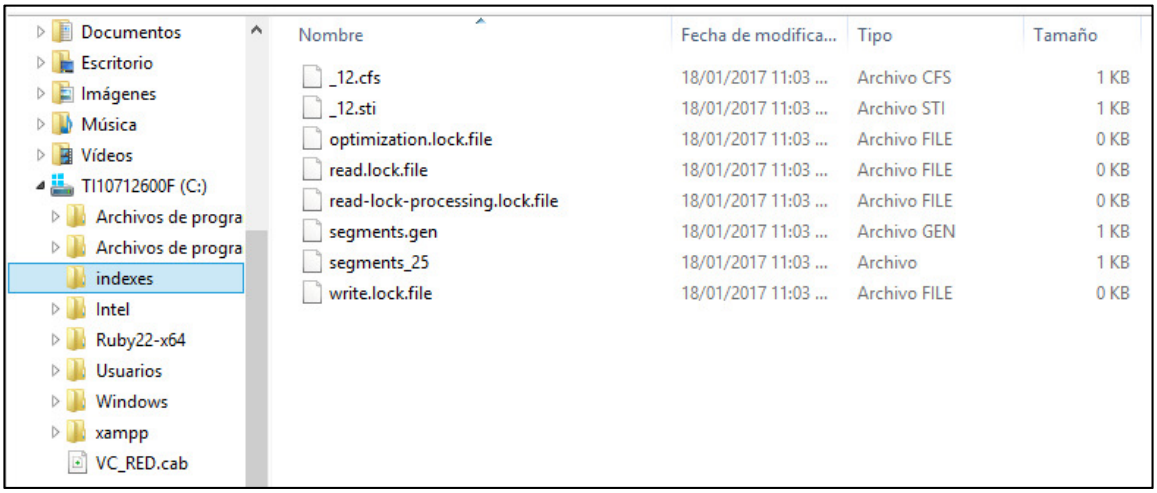


Figura Nº 42: Generación de Índices con Lucene.
Fuente: Elaboración propia.

CONCLUSIONES

Primera: Con el Prototipo del sistema desarrollado se puede determinar que toda la documentación ingresada y generada en la Unidad de Resoluciones puede ser digitalizada, determinándose así una mejor gestión en el control de la información y la documentación, al mismo resguardándose la información en una base de datos, ajena al deterioro, pérdida, falsificación u otro tipo de inconveniente que pueda surgir, y ser una herramienta que se utilice la mejor toma de decisiones por la rapidez con que se puede generar una resolución.

Segunda: El prototipo de sistema sirva como un efecto multiplicador en todas las dependencias de la universidad, con el fin de salvaguardar todo el legajo de documentación impreso y en digital que tiene la entidad, desde su reapertura y en un futuro cercano llegar a consolidar toda la información en uno solo, para una mejor administración y toma de decisiones oportuna.

Tercera: El Modelo de Espacio Vectorial en la búsqueda de información es adecuada y oportuna y versátil cuando es para búsquedas con términos y consultas especializadas como el caso de las resoluciones, generando rankings de precisión fundamentalmente cuando se tiene grandes cantidades de registros en la base de datos.

Cuarta: El implementar la búsqueda de información, nos encamina a continuar haciendo aplicaciones de búsqueda más eficientes que permitan facilitar información a los usuarios de manera oportuna y precisa.

SUGERENCIA

Primera: Se sugiere que la información que se ingresa debe ser la más real, objetiva es decir que la información impresa en papel debe ser el mismo que la información digital y/o puesto que de no ser así acarrearía cruces de información incoherentes y generaría problemas a los usuarios del sistema como a los interesados.

Segunda: Para que una resolución sea mostrada en la web como documento legal, objetivo y coherente, este debe estar refrendado por los respectivos sellos y firmas correspondientes a las autoridades universitarias, y en lo posterior desarrollar sistemas de documentales con firmas digitales. Si bien es cierto existen documentos denominados confidenciales, no todos se mostrarán en la Web, sino que estará supeditado en función al grado o nivel que tenga la resolución, manejándose con criterio objetivo la publicación de las resoluciones.

Tercera: Promover la cultura digital y hacer que el personal administrativo esté involucrado con las nuevas tecnologías, para ello se debe realizar la capacitación en Tecnologías de Información y Comunicaciones, porque no basta proporcionársele equipos de última generación si no se sabe hacer uso de los sistemas de información, o que simplemente que uno cuantos sean los privilegiados en el manejo de las TIC´s y se continúe generando la brecha digital entre los administrados de la entidad.

BIBLIOGRAFIA

Billhardt Holger, Fusión de Modelos Vectoriales y Contextuales para la Recuperación de Información, Departamento de Inteligencia Artificial, Facultad de Informática.2003

Blázquez Ochando Manuel, Técnicas avanzadas de recuperación de información: procesos, técnicas y métodos Madrid, 2013.

Castro Rueda Oscar J.. Tesis: Diseño e Implementación del Prototipo de un Sistema de Información Basado en Web para el Grupo de Investigación Ciencias de Materiales Biológicos y Semiconductores (CIMBIOS), Bucaramanga, 2008.

Chowdhury G. Introducción a la Moderna Recuperación de Información, Londres, Editorial: Association Publishing, 1999.

Dominich S. Documentos, sobre métodos matemáticos / formales de Recuperación de Información, Buckinhamshire Chilterns University College, Faculty of Technology, 2000.

Epifanio Tula Luis Gerónimo, Medeo Matias Daniel, sistema de Recuperación de Información, Motor de búsqueda Innuendo, Universidad Tecnológica Nacional, Facultad Regional Córdova.
http://www.jidis.frc.utn.edu.ar/papers/0b1f0cf7432b488ea2273481bfbf.pdf

Fernández Iparraguirre Jaddy, Segundo Seminario internacional sobre Gestión de Información y Transparencia, Conferencia, Video Importancia de la Gestión Documental en el Desarrollo de Proyectos para la administración Pública, México. 2015 https://www.youtube.com/watch?v=pcUtuycrtl4

Fox C, Lexical analysis and stoplists, W. Frakes y R. Baeza-Yates, editors, Information Retrieval: Data Structures & Algorithms, paginas 102-130 Prentice Hall, Englewood Cliffs, NJ, 1992.

Frakes y Baeza-Yates, Information Retrieval: Information Retrieval:Data Structures and Algorithms, editorial PRENTICE HALL, 1992.

Fugueras Ramón Alberch, Sistema de Gestión Documental, Transparencia y Preservación Digital, Video de Charla Dictada en la Biblioteca del Congreso Nacional de Chile. 2016.

Honrado A., R. Leon, O'Donnel R. y D. Sinclair. A Word stemming algorithm for the spanish language. Páginas 139-145 A Coruña, España, setiembre 2000, IEEE CS Press.

Lizcano B. Luis I., Sistema de Recuperación de Información Basado en el Modelo Vectorial, Dpto. de Sistemas Universidad Francisco de Paula Santander - Colombia, 2001

Luhn H. P., The automatic creation of literature abstracts. IBM Journal of Research and Development, 159-165, 1958.

Maron M. E. y Kuhns J. L., On relevance, probabilistic indexing an information retrieval. Journal of the ACMA, 1960.

Porter M. F., An algorithm for suffix stripping. Program, 130-137, 1980.

Quintero Belkys P., Fuentes Paola A., Diseño de un Sistema de Gestión Documental para el Programa de Ingeniería de Sistemas de la Universidad de San Buenaventura – Sede Bogota. – Tesis, 2007.

Rhoadas James B. La función de la gestión de documentos y archivos en los sistemas nacionales de información, estudio de RAMP, París UNESCO, 1989

Robertson S. E. y Sparck Jones K.. Relevance weighting of search terms. Journal of the American SocietyforInformationSciencia, 1976.

Sparck Jones K. y Willett P., editores. Readings in Information Retrieval. Morgan Kaufmann Publishers Inc., 1997.

Salton G., The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

Salton G., Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, Reading, 1989.

Salton G. y Buckley C., Term-weighting approaches in automatic text retrieval. Information Processing &Managenent, 1988.

Salton G. y M. J. McGill. Introducción a la Recuperación de Información Moderna. McGraw-Hill, New York, 1983

Salton G., A. Wong y C. S. Yang A vector space model for automatic indexing. Communications of the ACM 613-620, noviembre 1975.

Shasankar V. Krishna. Zend Framework 2.0 by Example, Beginner's Guide, Packt Publishing, Birmingham, julio 2013.

Sparck Jones K., A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 1972.

Van Rijsbergen C. J., Recuperación de Información, Butterworths, Londres Segunda Edición 1979.

Villena Roman J., Sistemas de Recuperación de Información, Valladolid, Departamento de Ingeniería de Sistemas Telemáticos, Universidad Politécnica de Madrid, 1997.

Vuotto Andres, Bogetti Celeste, Aplicación del factor TF-IDF en el análisis semántico de una colección documental, Universidad Nacional del Mar de Plata – Argentina, 2015, http://biblios.pitt.edu/ojs/index.php/biblios/article/viewFile/227/230.

Zipf H. P., Human Behavior and the Principle of Least Effort. Addison – Wesley, Cambrige, MA, 1949.