UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE CIENCIAS DE LA INGENIERÍA

CARRERA INGENIERÍA EN SISTEMAS

Proyecto de Investigación previo a
la obtención del título de Ingeniero
en Sistemas.

Título del Proyecto de Investigación:

"SISTEMA DE RECOMENDACIÓN DE MATERIAL BIBLIOGRÁFICO DIRIGIDO A LAS CARRERAS QUE OFERTA LA UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO."

Autor:
Juseh Roger Alcívar Cansiong

Director de Proyecto de Investigación:
Ing. Yeikier Mendez Socorro, Msc.

Quevedo – Los Ríos - Ecuador.

2015

DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS

Yo, **Juseh Roger Alcivar Cansiong,** declaro que la investigación aquí descrita es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Universidad Técnica Estatal de Quevedo, puede hacer uso de los derechos correspondientes a este documento, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

f. _____

Juseh Roger Alcivar Cansiong

C.C. 1206106427

CERTIFICACIÓN DE CULMINACIÓN DEL PROYECTO DE INVESTIGACIÓN

El suscrito, **Ing. Yeikier Mendez Socorro, Msc**, Docente de la Universidad Técnica Estatal de Quevedo, certifica que el estudiante **Juseh Roger Alcivar Cansiong**, realizó el Proyecto de Investigación de grado titulado **"Sistema de recomendación de material bibliográfico dirigido a las carreras que oferta la Universidad Técnica Estatal de Quevedo."**, previo a la obtención del título de Ingeniero en Sistemas, bajo mi dirección, habiendo cumplido con las disposiciones reglamentarias establecidas para el efecto.

……..

Ing. Yeikier Mendez Socorro, Msc

DIRECTOR DEL PROYECTO DE INVESTIGACIÓN

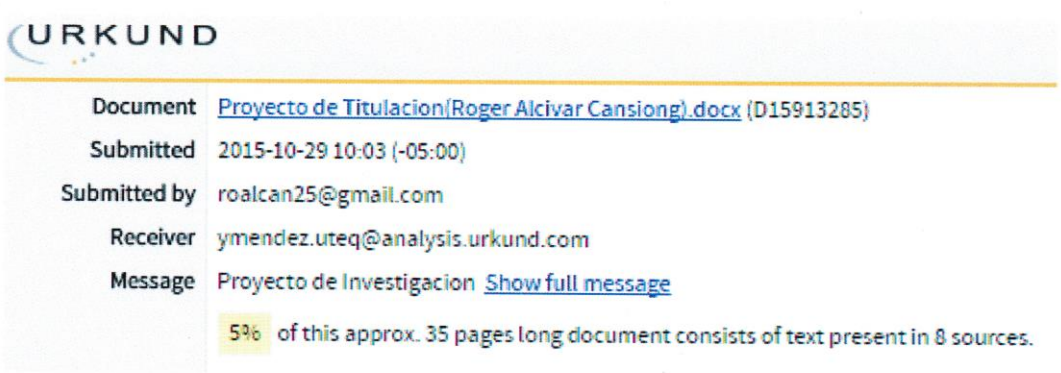
CERTIFICADO DEL REPORTE DE LA HERRAMIENTA DE PREVENCIÓN DE COINCIDENCIA Y/O PLAGIO ACADÉMICO

ING. JORGE MURILLO
DECANO DE LA FACULTAD DE CIENCIAS DE LA INGENIERIA.
Presente.-

De mis consideraciones:

ING. YEIKIER MÉNDEZ SOCORRO, en calidad de Director de Trabajo de Titulación cuyo tema es: **"SISTEMA DE RECOMENDACIÓN DE MATERIAL BIBLIOGRÁFICO DIRIGIDO A LAS CARRERAS QUE OFERTA LA UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO.",** me permito manifestar a usted y por intermedio del Consejo Directivo lo siguiente:

Que, el señor **Juseh Roger Alcívar Cansiong**, egresado de la Facultad de Ciencias de la Ingeniería, en la **Carrera de Ingeniería en Sistemas**, ha cumplido con las correcciones pertinentes, de acuerdo al reglamento de Graduación de Pregrado de la UTEQ, e ingresado su Trabajo de Titulación al sistema **URKUND**, tengo bien certificar la siguiente información sobre el informe del sistema reflejando un porcentaje del 5%


URKUND

Document	Proyecto de Titulacion(Roger Alcivar Cansiong).docx (D15913285)
Submitted	2015-10-29 10:03 (-05:00)
Submitted by	roalcan25@gmail.com
Receiver	ymendez.uteq@analysis.urkund.com
Message	Proyecto de Investigacion Show full message

5% of this approx. 35 pages long document consists of text present in 8 sources.

ING. YEIKIER MÉNDEZ SOCORRO

Director de Trabajo de Titulación

UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE CIENCIAS DE LA INGENIERÍA

CARRERA DE INGENIERIA EN SISTEMAS

PROYECTO DE INVESTIGACION

Título:

"Sistema de recomendación de material bibliográfico dirigido a las carreras que oferta la Universidad Técnica Estatal de Quevedo."

Presentado a la Comisión Académica como requisito previo a la obtención del título de Ingeniero en Sistemas.

Aprobado por:

_____

Phd. Amilkar Puris Cáceres
PRESIDENTE DEL TRIBUNAL

_____ _____

Ing. Iván Jaramillo Chuqui
MIEMBRO DEL TRIBUNAL

Ing. Kenya Guerrero
MIEMBRO DEL TRIBUNAL

QUEVEDO – LOS RIOS – ECUADOR

2015

AGRADECIMIENTO

Mis agradecimientos a todas las personas que estuvieron involucradas directa o indirectamente, en el camino de este proceso académico, siendo más específico.

A cada docente que brindaron sus conocimientos, sus historias y experiencias que aportaron lo suficiente en mi vida profesional y personal, aquellos que llevaron su forma de enseñanza fácil, complicada o abrumadora, que no se conformaban con nada, los que se conformaban con todo y a los que no tenían ni idea que querían, pero todos ellos me enseñaron de alguna manera esforzarme cada día, dar lo mejor de mí y no rendirme ante cualquier tropiezo o error y poder volverlo a intentar, a todos ellos gracias.

A mis compañeros al principio, que luego se volvieron parte de mi vida siendo mis amigos, con aquellos que se trabajaba sin importar la hora o el día, que en el rumbo de mi vida algunos se fueron alejando, aquellos que permanecieron en la lucha casi imposible "como así lo decíamos", aquellos que esa lucha les jugó una dura pasada, a todos ellos gracias.

A mis familiares, que me brindaron su apoyo, su paciencia, que pensaban que lo único que hacía era arreglar computadoras e impresoras, que no entendían lo que yo hacía, que se preocuparon por mis exámenes, que no paraban de recordarme si ya me había graduado, aquellos que estuvieron pendientes de mí, a todos ellos gracias.

Al Msc. Yeikier Mendez y al Ing. Jorge Guanin, por su contribución brindada en este proyecto de fin de carrera, a ellos gracias.

DEDICATORIA

Dedico logro obtenido, a mis padres que estuvieron día a día junto a mí, que me apoyaron en esos momentos de lucha y compromiso, en esos momentos de errores cometidos pero que ellos estuvieron apoyando, a mi madre Pilar Cansiong por su paciencia, por esas madrugadas que se levantaba a ver si aún estaba estudiando, por el carriño incondicional, a mi Padre Juseh Alcivar por sus consejos, por su paciencia y preocupación, por el apoyo y esfuerzo brindado hacia mí, por todo ello y por el hecho de haberme dado la vida, esto es por Ustedes, muchas gracias.

También va dedicado a mi hermano Bryan Alcivar por eso días en los que comprendía que dedicarle tiempo a ayudarlo me era difícil, a mis grandes amigos John Cruz y Lorna Jácome, ellos quienes juntos luchamos día tras día por superar los problemas e inconvenientes, con quienes se pasaron días de penas, melancolía, tristezas, y risas, esas risas sarcásticas y esas risas de felicidad, ellos quienes a pesar de los rumbos diferentes que se tomaron en una época de estudios volvimos a unirnos hasta el final de nuestro objetivo profesional, a ellos también va dedicado.

RESUMEN Y PALABRAS CLAVES

Los sistemas de búsqueda de información son capaces de presentar recomendaciones de contenido relevante, estos procesos comúnmente requieren acciones explicitas de parte de los usuarios, como valoraciones, perfiles de usuario, etc. Un sistema de recomendación basado en la obtención de información del usuario de forma implícita, permite crear procesos de manera automática capaces de recomendar contenidos a una colección de usuarios, basados en perfiles asociados a sus estudios académicos, carreras y materias que le pertenecen. Se presenta a continuación un estudio de investigación que describe los procesos de búsqueda, utilizando información de estudiantes en la institución realizando procesos mediante técnicas de búsqueda de información siguiendo el modelo de espacios vectoriales de los sistemas de recomendación basada en contenido. Se realizó pruebas de diferentes consultas dando como resultado una serie de comparaciones entre los modelos de similitud empleados, demostrando así, resultados relevantes y de utilidad en los niveles de estudio para los estudiantes de la Universidad Técnica Estatal de Quevedo. De esta manera los resultados arrojados ofrecen una buena alternativa a información apta para ser recomendada a estudiantes siendo este un prototipo de un sistema capaz de ser manipulado y fusionado en un gran sistema automatizado.

Sistema, recomendaciones, bibliografía, estudiantes, materias.

ABSTRACT AND KEYWORDS

The information search systems are able to submit recommendations of relevant content, these processes commonly require explicit actions on the part of users, as valuations, user profiles, etc. A recommendation system based on the obtaining of information of the user implicitly, it allows to create processes automatically able to recommend content to a collection of users, based on profiles associated with their academic studies, careers and subjects that belong to them. Presented below is a research study that describes the search process, using information from students in the institution making processes through information search techniques following the model of vector spaces of the recommendation systems based on content. It conducted tests for various consultations resulting in a series of comparisons among the models of similarity used, thus demonstrating, the relevant and useful results in the levels of study for students of the Universidad Técnica Estatal de Quevedo. In this way the results offer a good alternative to information suitable to be recommended to students being this a prototype of a system capable of handling and merged into a large automated system.

System, recommendations, bibliography, students, subjects.

TABLA DE CONTENIDO

INDICE DE FIGURAS

ÍNDICE DE TABLAS

CODIGO DUBLIN

Título:	Sistema de recomendación de material bibliográfico dirigido a las carreras que oferta la Universidad Técnica Estatal de Quevedo.			
Autor	Alcivar Cansiong, Juseh Roger			
Palabra clave:	Sistema	re-comendador	bibliografía	Material
Fecha de Publicación :	25-mar-15			
Editorial:	Quevedo: UTEQ, 2015.			
Resumen:	Los sistemas de búsqueda de información son capaces de presentar recomendaciones de contenido relevante, estos procesos comúnmente requieren acciones explicitas de parte de los usuarios, como valoraciones, perfiles de usuario, etc. Un sistema de recomendación basado en la obtención de información del usuario de forma implícita, permite crear procesos de manera automática capaces de recomendar contenidos a una colección de usuarios, basados en perfiles asociados a sus estudios académicos, carreras y materias que le pertenecen. Se presenta a continuación un estudio de investigación que describe los procesos de búsqueda, utilizando información de estudiantes en la institución realizando procesos mediante técnicas de búsqueda de información siguiendo el modelo de espacios vectoriales de los sistemas de recomendación basada en contenido. Se realizó pruebas de diferentes consultas dando como resultado una serie de comparaciones entre los modelos de similitud empleados, demostrando así, resultados relevantes y de utilidad en los niveles de estudio para los estudiantes de la Universidad Técnica Estatal de Quevedo. De esta manera los resultados arrojados ofrecen una buena alternativa a información apta para ser recomendada a estudiantes siendo este un prototipo de un sistema capaz de ser manipulado y fusionado en un gran sistema automatizado.			
Descripción:				
URI:				

INTRODUCCIÓN

En el dominio de la educación existe gran cantidad y diversidad de material que puede contribuir al proceso enseñanza-aprendizaje. Un objeto de aprendizaje (OA) es “cualquier recurso digital que puede ser utilizado repetidamente para facilitar el aprendizaje” [1]. A menudo a los estudiantes les resulta complejo buscar materiales bibliográficos en su institución que aporten en su proceso de aprendizaje.

Los sistemas de recomendación surgen para proporcionar sugerencias personalizadas para ayudar al usuario a escoger o seleccionar elementos entre distintas opciones. En concreto, este trabajo se centra en proporcionar recomendaciones de materiales bibliográficos como libros, siendo documentos susceptibles a ser sugeridos (recomendados).

Existen muchos tipos de sistemas de recomendación, entre los que se destacan los Sistemas de recomendación basado en contenido, colaborativos, basado en conocimiento e híbridos [2]. Cada uno de ellos tienen sus inconvenientes por lo que se necesitará encontrar el que mejor (o su combinación) se ajuste para obtener los resultados más relevantes. La finalidad es obtener información de aquellos materiales bibliográficos (contenidos en una base de datos) que se asemejen a un perfil académico obtenidos mediante realimentación implícita (analizando los registro de unidades de aprendizaje de la facultad) de esta manera sugerir (recomendar) los mejores resultados.

“La misión de la Biblioteca está comprometida con el desarrollo intelectual y el apoyo a la investigación científica, tecnológica de la comunidad universitaria a través de la gestión y administración de la información y el conocimiento tanto en soporte impreso como digital para alcanzar la excelencia académica y en el avance de las innovaciones tecnológicas.” [3] La Universidad Técnica Estatal de Quevedo (UTEQ) busca el desarrollo intelectual en sus estudiantes ofreciendo apoyo continuo y disponible, impulsando el uso de las Tecnologías de la Información y Comunicación (TICs), de esta manera pretende aumentar la curva de aprendizaje y el interés de los estudiantes de cada facultad.

Por tal motivo, es necesario el desarrollo de sistemas de información para la UTEQ que faciliten la búsqueda y recuperación de contenido bibliográfico que aporten al conocimiento

y avance académico de los estudiantes, aplicando técnicas de filtrado de información, para lograr aciertos útiles y confiables para usuarios interesados en aumentar su desarrollo intelectual e investigativo.

CAPITULO I

CONTEXTUALIZACIÓN DE LA INVESTIGACIÓN

1.1. Problema de la investigación

1.1.1. Planteamiento del problema.

Diagnostico

En la UTEQ se presenta algunos problemas relacionados con la falta de motivación e interés de parte del estudiante hacia los materiales bibliográficos existentes en la Institución, dando como resultado que el estudiante desconozca sobre el contenido de su unidad de aprendizaje, y así se haga poco o ningún uso de las instalaciones y portales web disponibles para aumentar la participación a la investigación y uso de los recursos propios de la Institución.

Debido a la gran cantidad de material bibliográfico en la institución a veces resulta difícil para los administradores encontrar los libros que los estudiantes necesitan debido a los contenidos amplios que pueden contener cada texto guía, dando como resultado que los estudiantes pierdan el interés y busquen otros medios sin aprovechar los existentes en la biblioteca.

Frente a ello, el estudiante al ignorar los distintos materiales de apoyo con los que cuenta la biblioteca institucional provoca una baja actividad en el uso de recursos propios de la institución y su afán de mejorar las capacidades investigativas hacia sus textos por parte de la Universidad.

En este contexto, se requiere fomentar el uso de materiales de apoyo y el uso de textos guías que el estudiante necesita. La UTEQ no cuenta con un sistema capaz de recomendar libros de su propio repositorio de textos, es por eso que se necesita de un sistema capaz de brindar a los estudiantes recomendaciones de textos guías o libros que faciliten la búsqueda de información en sus tareas diarias.

Pronóstico.

La institución podría verse afectada de manera parcial, debido a que los estudiantes carecerán de información sobre los materiales necesarios para sus estudios, la falta de motivación he incentivos provoca el desinterés hacia la lectura de libros físicos y utilización de los propios recursos de la institución y debido a la falta de un sistema capaz de recomendar a los estudiantes materiales bibliográficos que aporten el conocimiento e interés a utilizar recursos propios de la institución.

1.1.2. Formulación del problema

¿Cómo contribuir al mejoramiento en la selección de materiales bibliográficos en las carreras que oferta la Universidad Técnica Estatal de Quevedo?

1.1.3. Sistematización del problema

- ¿Qué información se necesitará para el análisis y utilización de las entradas del Sistema de recomendaciones de materiales bibliográficos?
- ¿Cuál modelo implementar capaz de obtener resultados relevantes para los estudiantes de la UTEQ?
- ¿De qué forma determinar las mejores respuestas de los algoritmos de recomendación y sus parámetros?
- ¿De qué manera se puede comparar los resultados de los algoritmos de los sistemas de recomendación basado en contenido mediante realimentación implícita?

1.2. Objetivos

1.2.1. Objetivo General

Desarrollar un sistema de recomendación basado en técnicas de filtrado de información para obtener una selección de materiales bibliográficos acorde a las carreras que oferta la Universidad Técnica Estatal de Quevedo.

1.2.2. Objetivos Específicos

- Analizar y utilizar la información académica de los estudiantes, unidades de aprendizaje y materiales bibliográficos como entradas que alimenten al Sistema de recomendaciones de materiales bibliográficos (SRMB) a desarrollar para así brindar resultados esperados.
- Implementar modelos de sistemas de recomendación capaces de obtener resultados relevantes para los estudiantes en las unidades de aprendizaje.
- Estudiar las respuestas de los algoritmos de recomendación y sus parámetros para determinar cuál o cuáles muestran los resultados más adecuados.
- Comparar los resultados obtenidos mediante realimentación implícita apoyados en el algoritmo de sistema de recomendación basado en contenido.

1.3. Justificación

Una definición formal de sistema de recomendación es la siguiente: se trata de aquel sistema que tiene como principal tarea seleccionar ciertos objetos de acuerdo a los requerimientos del usuario. [4]

El uso de sistemas de recomendación como técnica de filtrado de información, pretende solucionar el problema de la gran cantidad de textos guías, estos nos ayudan a filtrar contenidos obteniendo características de cada unidad de aprendizaje, además de la base de información de los materiales bibliográficos se podrá descubrir contenidos relevantes y eficientes para los estudiantes.

De esta manera ayudará a filtrar la información disponible sobre los estudiantes para obtener información de mayor interés y valiosa, permitiendo descubrir contenidos relevantes que servirán para obtener resultados esperados que sirvan para el trabajo diario de los estudiantes.

Uno de los beneficios o ventajas que brinda el uso de un sistema de recomendaciones de materiales bibliográficos, es el proceso automático de búsqueda de información que se basa en los perfiles de estudiantes y en el contenido digital de los textos guías para así obtener los mejores resultados para cada materia de las carreras de la institución.

Es necesario contar con un sistema que recomiende a los estudiantes materiales bibliográfico como libros o textos en general, de tal manera que el estudiante tenga a disposición contenido de vital interés en el desarrollo de su carrera profesional, fomentando la participación e interés por la lectura y aprendizaje, además de incrementar el uso de los recursos con los que cuenta la Universidad Técnica Estatal de Quevedo.

CAPÍTULO II

FUNDAMENTACIÓN TEÓRICA DE LA INVESTIGACIÓN

2.1. Marco conceptual.

2.1.1. Sistema de recomendación.

En este capítulo se pretende repasar los conceptos de los sistemas de recomendación y sus distintas técnicas para su implementación, empezando por la introducción seguido de una breve historia de lo que ha significado los sistemas de recomendación a lo largo de la informática, sus distintas técnicas y áreas correspondientes donde se encontrará formas y maneras diferentes de obtener resultados que satisfagan las necesidades de los involucrados, revisará las ventajas y desventajas de los diferentes temas tratados en este apartado además se mostraran ejemplos de sistemas en el mundo real.

2.1.2. Introducción.

Los sistemas de recomendación recopilan información de los usuarios para generar recomendaciones de ítems que pueden ser interés para cada usuario de forma personalizada. Este proceso lo realiza obteniendo información de los usuarios siendo la clave principal, ya que sin esa información sería imposible realizar las predicciones o recomendaciones de interés.

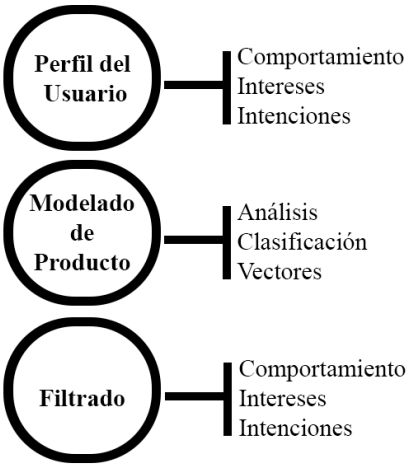
De una manera más formal, los sistemas de recomendación forman parte de un tipo de específico de técnicas de filtro de información, los cuales presentan distintos tipos de temas o ítems de información (películas, música, libros, noticias, imágenes, páginas web, etc.) que son del interés de un usuario en particular. [2]

Para sugerir nuevos ítems a los usuarios es necesario recolectar información de ellos, para esto existen dos formas de retroalimentación para hacerlo posible. Una es la forma explícita, es decir el usuario expresa de forma voluntaria y directa cuales contenidos les gustan de esta manera asignándose una puntuación a cada ítem o elemento. La segunda forma es recoger la información de manera implícita, en este caso el usuario desconoce que son evaluados, sin su intervención directa se obtiene la información relevante para el sistema.
Las tareas básicas que usan en los sistemas de recomendación son: la obtención del perfil de

usuario, el modelado de los productos y el filtrado. [5]

- Perfil de usuario: El sistema de recomendación necesita obtener información del usuario, el mismo que será usado para encontrar y recomendar los productos más relacionados a su perfil. Existen enfoques dentro de la tarea del perfil de usuario modelado de comportamiento, modelado de intereses o modelado de intenciones. De acuerdo a esta información obtenida de los modelados se podrá analizar y clasificar los usuarios, agruparlos, realizar predicciones, comportamientos y muchos más. [5]
- Modelado de productos: Para necesitar conocer la información de los productos del conjunto de datos obtenidos a partir del perfil del usuario, identificar el conjunto de información que sea necesaria para el modelamiento de los productos estos pueden ser palabras claves, etiquetas o cualquier otro campo que sea de interés. De esta forma realizar análisis, clasificación o modelamiento matemático como vectores.
- Filtrado: Una vez conocidos los perfiles de usuarios y de los productos se pretende realizar un filtrado basándose en como de adecuado es un producto a un usuario concreto (utilidad) o predecir la valoración que un usuario daría a un producto. [5] En el que se utiliza distintas técnicas de filtrado como, colaborativo, contenido, conocimiento, etc.

Figura.- 1 Tareas básicas de los sistemas de recomendación



FUENTE: UNIVERSIDAD DE JAÉN

ELABORADO: JORGE CASTRO GALLARDO

2.1.3. Historia.

En un comienzo los sistemas de recomendación datan de principios de los años 90. El termino fue acuñado en 1992 para un sistema de filtrado de correo electrónico no automatizado. En 1994 se desarrolló el primer "workshop" en Berkeley donde se vio la utilidad en diversas áreas de los primeros algoritmos simples de este tipo. También se identificaron algunas cuestiones importantes para el desarrollo de estos algoritmos: Escalabilidad, Viabilidad económica, puntuaciones implícitas y explicitas. Uno de los grupos de investigación pioneros en el desarrollo del filtrado colaborativo fue el proyecto GroupLens de la universidad de Minnesota que a´un permanece muy activo y que ha proporcionado una gran parte de la base algorítmica de muchos sistemas de recomendación. [6]
Para aquella época se realizaban recomendaciones basadas en recuperación de información, tales como:

- Filtrado basado en características: Se basa en la idea de que es posible capturar que características que le gustan (o no) a un usuario con respecto a cierto ítem y entonces es posible proceder a realimentar al usuario con información de diversos ítems. [7]
- Filtrado de productos sin personalización: Con esta técnica se calcula una lista de productos que se corresponde con los productos mejor valora- dos (mayor valoración media) o con los productos más populares del sistema (el más comprado). [5]
- Filtrado con datos generales de los consumidores: Utilizan datos generales que el sistema conoce, sin realizar ninguna operación con los datos del usuario activo1. Se han utilizado técnicas como la recomendación a partir otros productos. [5] Un claro ejemplo es la tienda de comercio electrónico Amazon donde sus productos son recomendados en base la interacción que tiene el usuario.

2.1.4. Técnica de retroalimentación de información.

Un sistema de recomendación no debe ser una entidad estática, sino evolucionar en el tiempo en cuanto a la cantidad de sus recomendaciones y pronósticos en base a la experiencia y nueva información adquiridas. Para conseguir este objetivo se utilizan mecanismos de

realimentación entre el sistema y los gustos de los usuarios. Existen dos tipos de mecanismos de realimentación. [8]

2.1.4.1. Realimentación implícita.

Un mecanismo de realimentación implícita es aquel que proporciona información al sistema de recomendación acerca de los gustos de los usuarios sin que éstos sean conscientes de esta situación. Por lo tanto este tipo de realimentación no es directa sino que se realiza mediante diversos tipos de medidas como pueden ser el tiempo de visualización del objeto, el número de veces que el objeto es solicitado, etc. [8]

Esta realimentación implícita tiene el problema de depender en demasía del contexto y de ser excesivamente hipotética (podemos suponer que solicitar la visualización de un objeto muchas veces indica un especial interés por parte del usuario pero no tiene por qué ser de esa manera) por lo que no resulta ser la más apropiada para todas las situaciones de recomendación. [8]
Ejemplos de recolección de datos de forma implícita.

- Guardar un registro de los temas que el usuario ha visto en una tienda online Analizar el número de visitas que recibe un artículo
- Guardar un registro de los artículos que el usuario ha seleccionado
- Obtener un registro de las asignaturas de un usuario para la obtención de referencias a libros digitales o físicos.
- Obtener un listado de los artículos que el usuario ha seleccionado o visto en su computadora.
- Analizar las redes sociales de las que el usuario forma parte y de esta manera conocer sus gustos y preferencias. [2]

2.1.4.2. Realimentación explícita.

Un mecanismo de realimentación explícita es aquel basado en la acción directa por parte del usuario para indicar que objetos determinados del sistema son de su interés. Esta interacción directa se puede realizar mediante votaciones numéricas o más sencillas aún, que el usuario diga si el objeto es o no de su agrado. [8]

Este mecanismo tampoco se encuentra exento de problemas como pueden ser la voluntariedad del cliente o el tiempo consumido. [8]

Algunos ejemplos de recolección de datos de forma explícita:

- Solicitar al usuario que pondere en base a una escala proporcionada, algún tema en particular. [2]
- Solicitar al usuario que pondere un conjunto de temas de una lista de favoritos. [2]
- Presentar al usuario dos temas, y solicitarle que seleccione uno de ellos Solicitar al usuario que cree una lista de temas de su preferencia. [2]

2.1.5. Clasificación de los Sistemas de Recomendación.

Los sistemas de recomendación pueden ser clasificados en diferentes tipos, de acuerdo al tipo de información que se utiliza.

2.1.5.1. Sistema de Recomendación basado en filtrado colaborativo.

En el enfoque más reciente, el filtrado colaborativo es un método para hacer predicciones automáticas (filtrado) sobre los intereses de un usuario mediante la recopilación de las preferencias o gustos de información de muchos usuarios (colaborador). [7]

Dentro del enfoque de otro autor los sistemas basados en filtrado colaborativo en el caso más simple, estos sistemas predicen las preferencias de un usuario como una suma ponderada de

las preferencias de otros usuarios, en los cuales los pesos son proporcionales a las correlaciones sobre el conjunto de objetos comunes evaluados por dos personas. [9]

Nótese que estas predicciones son específicas para el usuario, pero utilizan la información obtenida de muchos usuarios. Esto difiere del enfoque más simple de otorgarle una puntuación promedio (poco específico) para cada elemento de interés, por ejemplo sobre la base de su número de votos. [10]

El filtrado colaborativo utilizó inicialmente el concepto de "vecinos más cercanos", donde un usuario que accede al sistema es confrontado con una base de datos de valoraciones para descubrir otros usuarios que poseen un historial similar al de él. Así, los ítems recomendados van a ser aquellos que han gustado a los usuarios que poseen intereses similares al "usuario que recibe la recomendación", al cual en la literatura se suele referir por Usuario Activo. En el caso de que el usuario no tenga historial de navegación (usuario nuevo), se suele simular, utilizándose diferentes tipos de técnicas, un historial de navegación, donde se asume que el usuario ha visitado determinados ítems, según [11].

2.1.5.2. Sistema de Recomendación basado en contenido.

En los sistemas de recomendación basado en contenido se tiene información sobre las características de cada producto y se intentan extraer relaciones entre estas y la valoraciones de preferencia de un usuario. Para ello, se construye un perfil de usuario a partir de los productos que ha valorado y se utiliza para evaluar los productos no experimentados. Este tipo de SR no trata de predecir la valoración que un usuario daría, sino que asigna una puntuación a los productos, que indica la idoneidad del producto para el usuario. [5]

$$Utilidad_{(u,p)} = Puntuacion_{(Perfil_{(u)},Contenido_{(p)})}$$

La información que estos sistemas utilizan se describe en la Tabla.- 1: [5]

- $a_{x_{(u,1)}}, \ldots . a_{x_{(u,nu)}}$ son los productos que el usuario u ha valorado. $x_{(u,1)}$ denota el primer producto que el usuario (u) valoró.
- $r_1^u, \ldots r_{nu}^u$ son los respectivos valores de preferencia que el usuario ha dado.

- $c_1, \ldots\ c_m$ son las caracter´ısticas que tienen dichos productos.
- $v_1^{x_{ul}}, \ldots\ v_m^{x_{ul}}$ son los valores de las características para el producto x.

Tabla.- 1 Datos de un Sistema de Recomendación Basado en Contenido

	$\mathbf{c_1 \ldots c_j \ldots c_m}$
$\mathbf{a_1}$	$v_1 \ldots v_j^1 \ldots v_m^1$
...	... ...
$\mathbf{a_n}$	$v_1^n \ldots v_j^n \ldots v_m^n$

FUENTE: UNIVERSIDAD DE JAÉN
ELABORADO: JORGE CASTRO GALLARDO

Los sistemas de recomendación basados en contenido presentan distintas técnicas, las cuales se presentan a continuación:

- Modelado en Espacios Vectoriales. Esta técnica procede del área de la recuperación de información. Para calcular las recomendaciones, el sistema calcula un vector asociado al usuario y obtiene la similitud (coseno, correlación) con los vectores asocia- dos a los productos que el usuario no ha valorado. Finalmente se recomiendan los N productos cuya similitud con el perfil del usuario sea mayor. [5]

En sección 2.1.8 se explica con más detalles el área de recuperación de información, ya que fue inspiración para este proyecto.

- Redes Bayesianas. También conocidas como redes causales probabilísticas, sistemas expertos bayesianos, sistemas expertos probabilísticos, redes causales, redes de creencia o diagramas de influencia son herramientas estadísticas orientadas al modelado gráfico que pertenecen a la familia de los Sistemas Estocásticos Altamente Estructurados. [12]

La red bayesiana contiene nodos para cada valor de cada característica y para cada producto. Utilizando las valoraciones que otros usuarios dan a productos similares, se calcula una puntuación para el producto. Se recomiendan los k productos con mayor puntuación. [5]

2.1.5.3. Sistema de Recomendación basado en conocimiento.

Los sistemas basados en contenido buscan utilizar la información disponible tanto del ítem como del usuario para calcular la recomendación más adecuada.

Los Sistemas de Recomendación Basados en Conocimiento son sistemas que utilizan una base de conocimiento que describe como los distintos productos satisfacen las necesidades de un usuario y en qué medida (conocimiento funcional del entorno). De esta manera, el sistema encuentra el producto (o productos) que se ajusta a las necesidades que el usuario ha especificado. Esta búsqueda de productos a partir de las necesidades se realiza mediante un proceso de inferencia de algún tipo. Esta arquitectura permite a estos sistemas obtener y almacenar la información de distintas formas. [5]

Para realizar un proceso de recomendación correcto con pocas cantidades de información sobre los usuarios, estos sistemas necesitan tener un conocimiento adicional sobre el entorno en el que se utilizan. Además de tener información sobre las características de los productos, necesitan información sobre las relaciones entre los mismos, su capacidad para satisfacer las necesidades del usuario, etc. Esto puede resultar una tarea costosa, ya que se necesita realizar un proceso de Ingeniería del Conocimiento para obtenerla. [5]
Por ello, la aplicación de sistema basado en conocimiento en distintos dominios (recomendación de películas, coches, libros, música, noticias) no resulta trivial y requiere un gran esfuerzo para trasladarlo de un dominio a otro. [5]

2.1.5.4. Sistema de Recomendación demográficos.

Clasifican a los usuarios en grupos demográficos basándose en atributos personales y les proporcionan recomendaciones potencialmente interesantes para cualquier persona perteneciente a dicho grupo demográfico. [9]

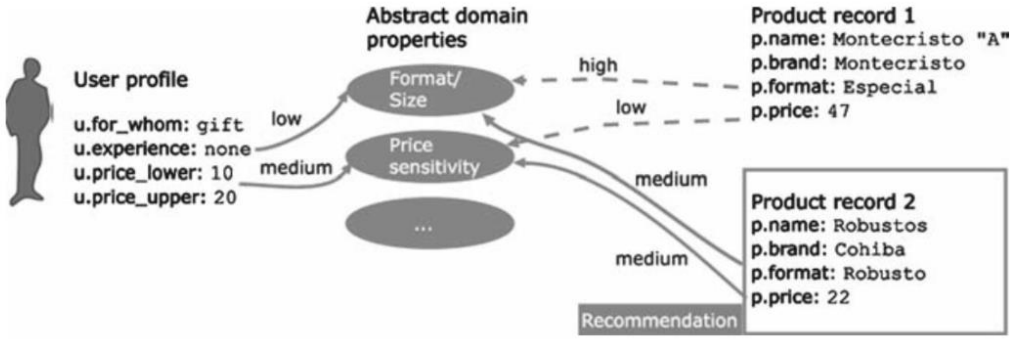
Visto desde otro punto de vista de otro autor Los Sistemas de Recomendación Demográficos (SRD) se basan en la idea de que personas con unas características demográficas dadas (edad, sexo, nivel de educación, domicilio) tengan gustos similares a otras personas con características demográficas similares. A continuación se exponen algunos trabajos que

proponen distintos modelos para aprovechar la información demográfica de los usuarios para mejorar el proceso de recomendación. [5]

2.1.5.5. Sistema de Recomendación basado en utilidad.

En los Sistemas de Recomendación Basados en Utilidad la información que se tiene sobre el usuario es una función que el usuario define para otorgar un valor de utilidad a los productos del catálogo. Esta función de utilidad se define mediante la combinación de los valores de los atributos de los productos. El SRBU evalúa todos los productos del catálogo y recomienda los N productos con mayor valor de utilidad al usuario [5].

Figura.- 2 Recomendación de productos basada en utilidad



FUENTE: UNIVERSIDAD DE JAÉN
ELABORADO: JORGE CASTRO GALLARDO

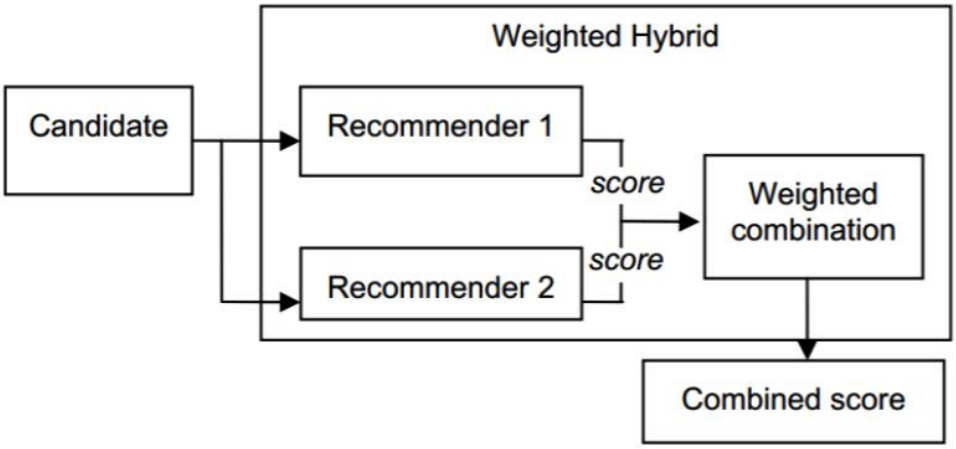
Una característica común de estos SR es que tienen un alto grado de interactividad con el usuario, realizando un dialogo con el usuario a través de formulario. De esta manera, el usuario puede especificar las características relevantes de los productos, así como los valores deseados de las mismas. [5]

2.1.5.6. Sistema de Recomendación híbridos.

Los sistemas híbridos combinan las técnicas de filtrado colaborativo y basado en contenido con el fin de mejorar las recomendaciones resultantes o superar problemas como el arranque en frío que presentan algunos de los métodos anteriores al ser utilizados por separado. [13]

Existen distintos métodos que describen la combinación del filtrado basado en contenido y el filtrado colaborativo. Los métodos parten de la obtención de la información para la creación de los perfiles de usuario. El perfil del usuario indica la información necesaria de las preferencias relacionadas a los ítems que son de interés del usuario. La creación del perfil puede ser realizada por un método de ponderación manual en el que cada usuario da un peso a los atributos de los ítems de acuerdo a sus gustos o puede ser automática donde la ponderación se calcula usando características de los atributos e información previa del usuario [13].

Figura.- 3 Esquema general de hibridación por ponderación



FUENTE: UNIVERSIDAD DE JAÉN

ELABORADO: JORGE CASTRO GALLARDO

2.1.6. Utilidad de los Sistemas de Recomendación.

Los sistemas de recomendación resultan muy importantes ya que reducen el tiempo de búsqueda de ítems, consiguen una mayor efectividad en las búsquedas y, por lo tanto, una mayor satisfacción de los usuarios.
Los sistemas de recomendación efectúan dos fases las cuales son:

- Predecir: En esta fase los sistemas de recomendación hacen referencia a estimar que valoración daría el usuario a cada ítem.
- Recomendación: En esta fase los sistemas de recomendación se refieren a extraer los N ítems más relevantes o recomendables.

2.1.7. Ejemplos y casos reales de estudios de sistemas de recomendación.

Actualmente existen una gran cantidad de sistemas de recomendación que son utilizados en diferentes áreas, ya sea comercial, científico o experimental. A continuación se resumirán algunos de los sistemas más importantes. [14]

2.1.7.1. Sistema de recomendación de Amazon.

Es el sistema de recomendación utilizado por la tienda virtual de Amazon.com y sus variantes, es un poderoso sistema que utiliza una mezcla de algoritmos, los cuales son basado en contenido y basado en filtro colaborativo. Este sistema se basa en la búsqueda de artículos similares o de relación y no de usuarios similares.

Figura.- 4 Sistema de recomendación de la web de Amazon

FUENTE: AMAZON

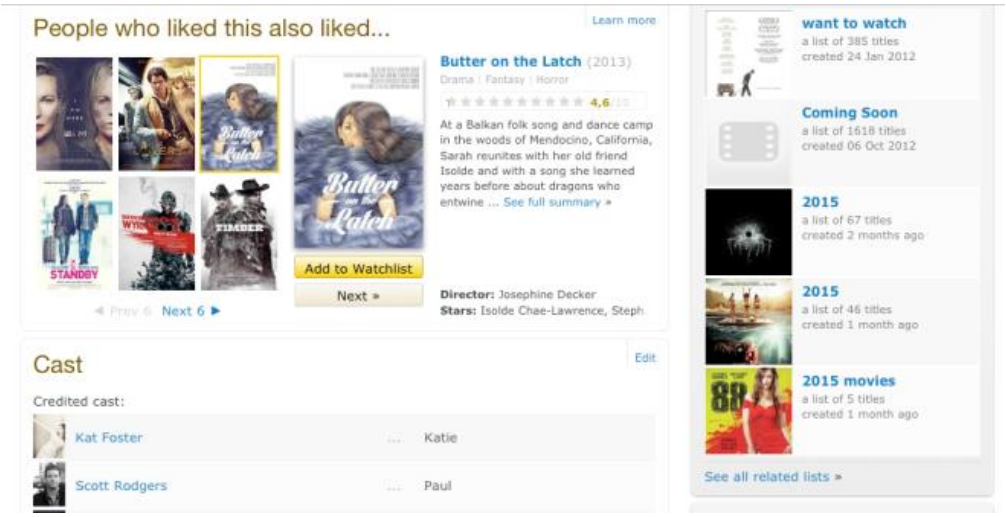
ELABORADO: AMAZON

2.1.7.2. Sistema de recomendación IMDB Recommendation Center.

El sitio web imdb.com es uno de los más populares del mundo teniendo así también la mayor base de datos de cine y televisión., ofreciendo a sus usuarios Recomemendation Center un sistema de recomendación basado en filtrado por contenido.

El usuario introduce la película o show televisivo que más le guste y el sistema le ofrece una lista con diez recomendaciones. Como método de realimentación o feedback con el sistema, el usuario puede señalar las recomendaciones con las que no esté de acuerdo y proponer recomendaciones nuevas con lo que el algoritmo se va depurando con la interacción del usuario.

Figura.- 5 Interfaz del Centro de Recomendación de IMDB



FUENTE: IMDB

ELABORADO: IMDB

2.1.7.3. Sistema de Recomendación Tapestry.

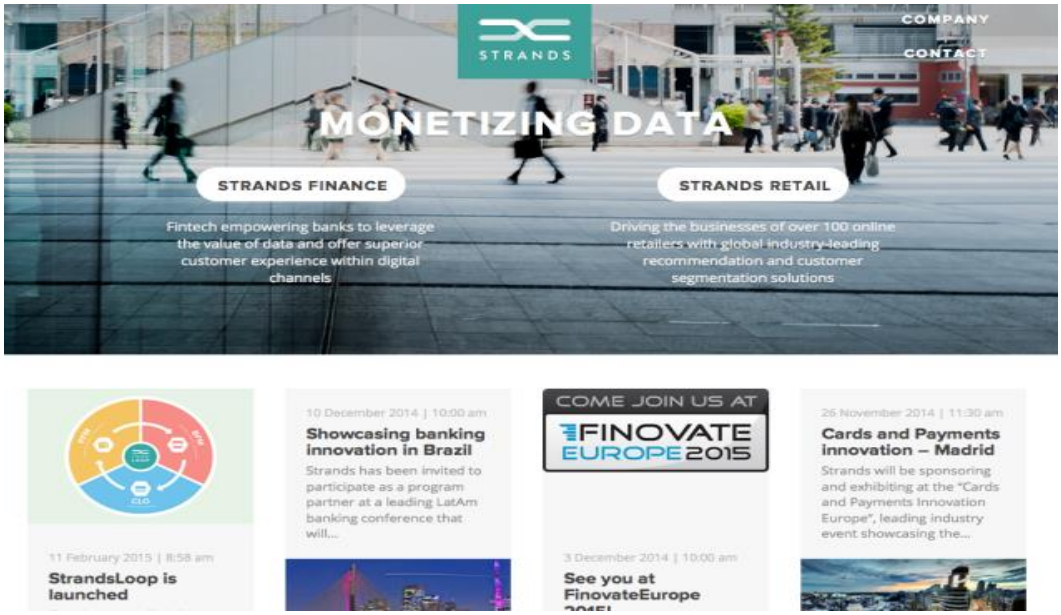
Es un sistema que permite almacenar el feedback de los usuarios sobre los artículos o noticias que éstos han leído y posteriormente ser utilizado por otros usuarios que aún no han leído el artículo o noticia, para establecer si la información del documento es relevante o no.

En un principio este tipo de sistemas fue adoptado con el nombre de filtro colaborativo (collaborative filter) dado que permite que los usuarios creen filtros a través de sus ítems de interés (en el caso de Tapestry, artículos o noticias), y colaborativo pues los usuarios añaden las anotaciones con las opiniones sobre los documentos. [15]

2.1.7.4. Sistema de recomendación Strands.

Como se muestra en la Figura 6 Strand es un sistema de recomendación que permite personalizar la experiencia en línea de los usuarios y generar recomendaciones de productos de acuerdo a su perfil. También es utilizado en una es una red social dedicada a deportistas que permite que éstos compartan sus programas de entrenamiento y recomienda programas de otros usuarios. [14]

Figura.- 6 Sistema de recomendación del sitio web Strands



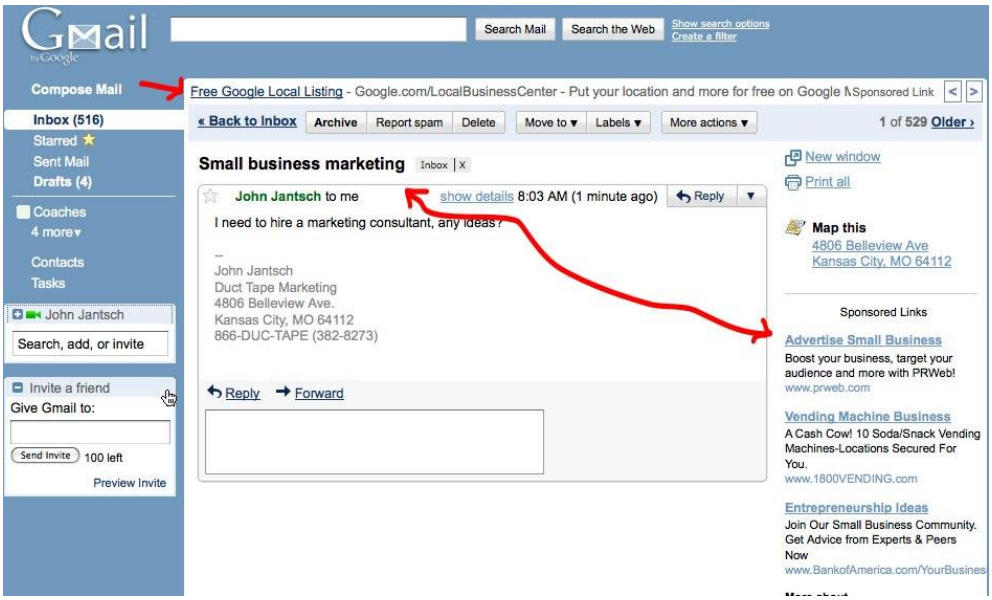
FUENTE: WEB STRANDS

ELABORADO: WEB STRANDS

2.1.7.5. Google Adsense.

Esta herramienta de Google permite insertar en las páginas web de las diferentes aplicaciones de Google (gmail, docs, etc.) publicidad acorde con los gustos y preferencias de cada usuario particular. La información para conocer a los usuarios se obtiene, por ejemplo, analizando los correos electrónicos que escriben. Esta información se procesa de manera automática, manteniendo el anonimato del usuario, y se utiliza para mostrarle al usuario anuncios sobre productos que puedan serle de interés. [14]

Figura.- 7 Sistema de recomendación de Google Adsence



FUENTE: GOOGLE ADSENCE

ELABORADO: GOOGLE ADSENCE

2.1.8. Recuperación de información

La investigación en Recuperación de Información pretende entender el complejo proceso de la búsqueda de información con la finalidad de diseñar, construir y testar sistemas cada vez más eficientes. [16]

Una definición del autor el artículo, dice, la Recuperación de Información (RI, a partir de ahora) es la disciplina que estudia la representación, la organización y el acceso eficiente a la información que se encuentra registrada en documentos. [17]

De las operaciones propias de la RI, sin duda la más característica consiste en la selección de documentos, bien a partir de las características de su contenido, (los temas tratados), bien a partir de características de su contexto (p.e. la fecha de publicación,) bien a partir de alguna combinación de ambas cosas (p.e: "documentos sobre desarrollo humano publicados por UNESCO entre 2003 y 2005"). [17]

Ahora bien, para que la RI tenga sentido se presupone un entorno en el cual no es trivial, precisamente, el hecho de acceder a los documentos por su contenido. Este contexto lo genera, típicamente, cualquier fondo documental a partir del momento que contenga unos centenares o unos miles de documentos. Empresas pequeñas, medianas o grandes, con ejecutivos, abogados, químicos o ingenieros que necesitan encontrar una información en fondos internos o externos es un ejemplo. Universitarios e investigadores que necesitan consultar bases de datos bibliográficas para asegurarse de que no reinventan la rueda es otro. Finalmente, la Web, que en realidad es un enorme sistema de información documental con varios miles de millones de documentos es el ejemplo extremo de contexto característico de RI. [17]

2.1.8.1. Normalización e indexación.

Se pretende formatear los términos y documentos a un estado aceptable para los modelos de recuperación de información. Un documento indexado es una representación del documento original. En la práctica, consiste en una lista de términos o conceptos normalizados, de alto valor semántico, con información adicional asociada (por ejemplo, su frecuencia de aparición o posición en el texto). Los términos pertenecientes al índice pueden estar en su forma original o lematizados y pueden ser palabras simples, multipalabras, siglas o nombres propios. [18]

En general, la indexación de base no lingüística se fundamenta en el análisis de la frecuencia de los términos y su distribución dentro de los documentos. Este análisis tiene como objeto establecer criterios que permitan determinar si una palabra es un término de indexación válido, fundamentalmente porque permite discriminar el contenido de los documentos y –de alguna manera – aporta información. [18]

2.1.8.2. Eliminación de palabras vacías o “stopwords”.

Los términos que ocurren en casi todos los documentos de una colección no son buenos para la recuperación de información por su nulo poder para discriminar documentos. Las palabras que aparecen en más del 80% de documentos no deberían seleccionarse como términos de

indexación. Este conjunto de términos se lo conoce como palabras vacías o *stopwords*. Esta categoría de palabras está formada – generalmente – por artículos, preposiciones, conjunciones y forman lo que se conoce como diccionario negativo o anti- diccionario. La lista de palabras vacías depende de cada lenguaje. [18]

En la Tabla.- 2 véase un ejemplo del artículo [19] muestra las palabras vacías que podrían usar para la eliminación.

Tabla.- 2 Muestra de palabras vacías del español

el, la, los, les, las, de, del, a, ante, con, en, para, por, y, o, u, tu, te, ti, le, que, al, ha, un, han, lo, su, una, estas, esto, este, es, tras, suya, a, acá, ahí, ajena, ajenas, ajeno, ajenos, al, algo, algún, alguna, algunas, alguno, algunos, allá, alli, allí, ambos, empleamos, ante, antes, aquel, aquella, aquellas, aquello, aquellos, aqui, aquí, arriba, asi, atras, aun, aunque, bajo, bastante, bien, cabe, cada, casi, cierta, ciertas, cierto, ciertos, como, cómo, con, conmigo, conseguimos, conseguir, consigo, consigue, consiguen, consigues, contigo, contra, cual, cuales, cualquier, cualquiera, cualquieras, cuan, cuán, cuando, cuanta, cuánta, cuantas, cuántas, cuanto, cuánto, cuantos, cuántos, de, dejar, del, demás, demás, demasiada, demasiadas, demasiado, demasiados, dentro, desde, donde, dos, el, él, ella, ellas, ello, ellos, empleáis, emplean, emplear, empleas, empleo, en, encima, entonces, entre, era, éramos...

FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID
ELABORADO: MANUEL BLÁZQUEZ OCHANDO

2.1.8.3. Conversión de minúsculas a mayúsculas y eliminación de acentos.

El pre-procesado léxico inicial ha consistido en convertir la cadena de entrada, esto es, el texto de cada documento, en un conjunto de palabras, que puedan servir luego como términos índice. [20]

El tratamiento de vocales acentuadas es también importante. En general, las vocales acentuadas incluyen una carga semántica importante a la palabra, pero en recuperación de

información no suelen considerarse; el motivo no es otro que el alto grado de errores ortográficos que se cometen con los acentos, por lo que, generalmente, en el pre-proceso léxico se convierten a vocales no acentuadas aquellas que lo estén en el texto. [20]

Tabla.- 3 Conversión de vocales con acentos a su representación a usar

[óíáéúüÓÍÁÉÚÜàèìòùÀÈÌÒÙâêîôûÂÊÎÔÛäëïöÄËÏÖåÅãÃõÕ]
[oiaeuuOIAEUUaeiouAEIOUaeiouAEIOUaeioAEIOaAaAoO]

FUENTE: UNIVERSIDAD DE SALAMANCA
ELABORADO: ZARO, FIGUEROLA, ALONSO, GÓMEZ

2.1.9. Las técnicas automáticas de Recuperación de información.

Con la frase técnicas automáticas se hace referencia al conjunto de procedimientos y recursos que se aplican para que el sistema explote capacidades que el usuario no posee, lo alivie de las tareas rutinarias y trabajosas, o complemente y amplíe sus capacidades. Aplicando este criterio amplio, muchas de las cosas que se tratan en la bibliografía sobre RI son técnicas. Por ejemplo, la base de datos es un recurso de almacenamiento y recuperación de información que supera ampliamente a la memoria humana. Una interfaz de búsqueda gráfica es un procedimiento que permite el acceso temático a una colección de miles de documentos en solo 2 o 3 pantallas. Todas estas técnicas son empleadas para lograr una interacción exitosa entre un usuario, que puede ser humano o máquina, con cierta necesidad informativa; y una masa de información variada, registrada en documentos digitales o no, susceptible de satisfacer dicha necesidad y que puede o no haber sido sometida a algún proceso de descripción previo. [16]

2.2. Sistema de Recomendación basado en contenido mediante el método de espacios vectoriales.

En el presenta trabajo de investigación se presenta como propuesta el uso de métodos de espacios vectoriales, estos sistemas trabajan especialmente con contenido lingüísticos, que

pueden tomar mediantes valores dentro de un dominio, quienes están definidos en un conjunto de características. [5]

Dentro de esta área aborda varios métodos como son los modelos booleanos, modelos probabilísticos y modelo vectorial, en el cual este último se aborda en este proyecto de investigación.

2.2.1. Modelo Booleano.

La idea principal de este modelo es que los documentos y las interrogaciones de los usuarios podrán ser representados por uno o más términos. Cada término debe pensarse como un conjunto cuyos elementos son los documentos que lo tienen asignado como término de indización. Se recordará que en el capítulo anterior se expuso que los sistemas
17clásicos utilizan un índice de términos con la referencia a los documentos que lo poseen. Cuando se expresa el término kn en la interrogación, para el sistema se está expresando el conjunto de los documentos que lo contienen. Si se pueden pensar los términos como conjuntos con elementos, se puede entonces establecer operaciones entre ellos. Los términos de la interrogación son susceptibles de ser enlazados por los operadores pertenecientes al álgebra de Boole: [16]

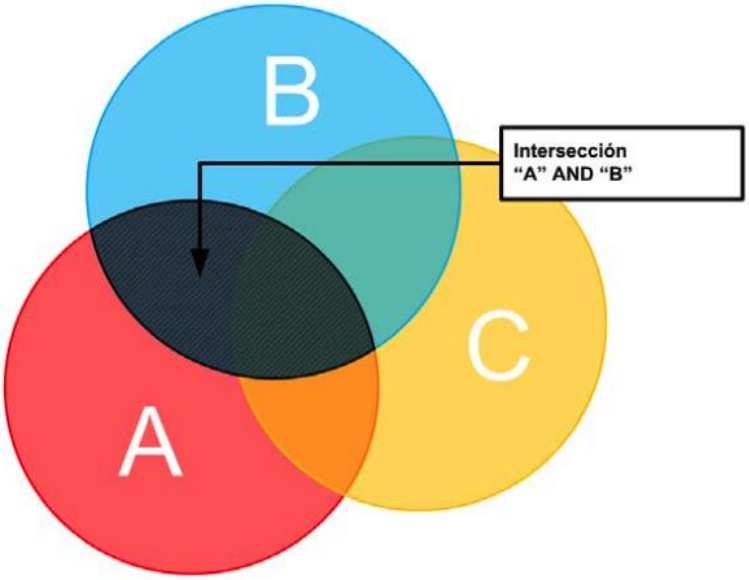
• Y (AND) - Intersección
• (OR) - Suma o unión
• NO (NOT) – Resta o negación

Explicado de otra manera en el artículo [19], el álgebra de Boole aplicada a la recuperación de información consta de una serie de casos básicos a los que pueden añadirse múltiples cadenas de resolución booleana. Dicho de otra forma, pueden llevarse a cabo operaciones verdaderamente complejas dependiendo de la cantidad de conjuntos a dirimir.

Operador AND. Q = "A" AND "B"
El operador AND es el encargado de intersecar o especificar que dos condiciones, premisas o términos tienen que cumplirse obligatoriamente, simultáneamente o a la vez. Esto significa

que si no se produce de esta forma, el sistema de recuperación no devolverá resultado alguno. Según lo que se muestra en la Figura.- 8, sólo los documentos que posean el término A y B (zona sombreada) se recuperarán, desechando por lo tanto aquellos términos que o bien sólo contengan A o bien sólo contengan B. [19]

Figura.- 8 Intercepción de documentos A y B



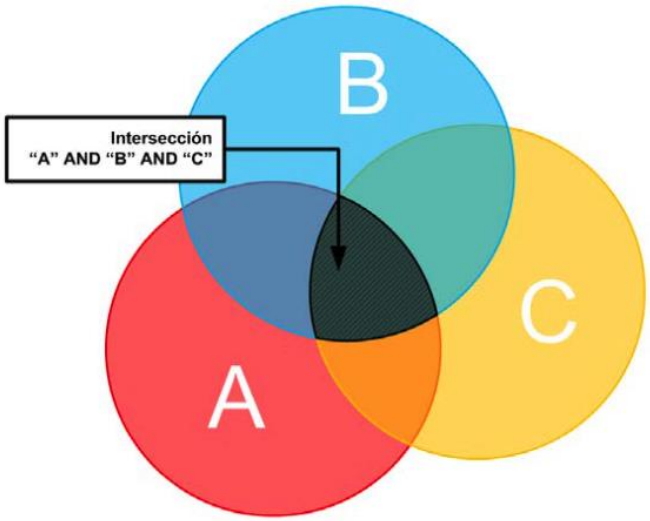
FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

Operador AND. Q = "A" AND "B" AND "C"

En este caso, la consulta propuesta implica la intersección de 3 términos. Por lo tanto, la recuperación sólo se efectuará para aquellos documentos que tengan presentes los términos A, B y C. Si faltase uno de estos términos el documento no se recuperaría, véase Figura.- 9. [19]

Figura.- 9 Intercepción de documentos A, B y C

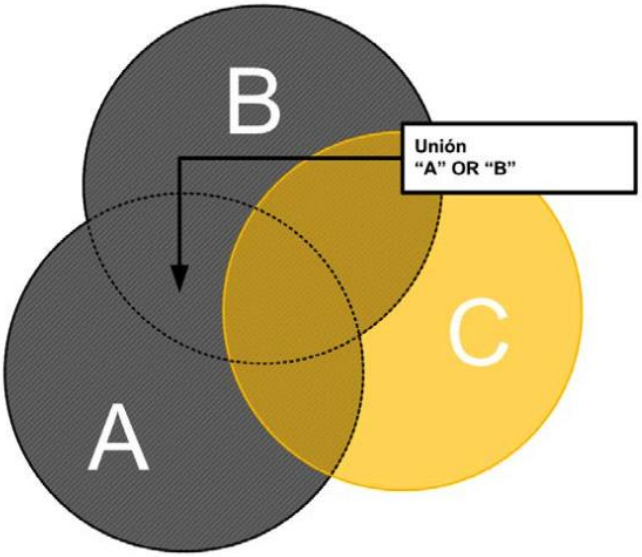


FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID
ELABORADO: MANUEL BLÁZQUEZ OCHANDO

Operador OR. Q = "A" OR "B"

El operador OR implica unión, alternativa y adición. Esto significa que dos conjuntos conectados por el operador OR se sumarán o unirán y si constan de elementos comunes, éstos también se recogerán. En recuperación de información significa que para una consulta de términos A OR B, se recuperarán aquellos documentos que tengan presencia de A, presencia de B y presencia de A y B a la vez. Por ello la consulta AND es más específica y restrictiva que OR, mucho más amplia Figura.- 10. [19]

Figura.- 10 Unión de documentos A, B y C



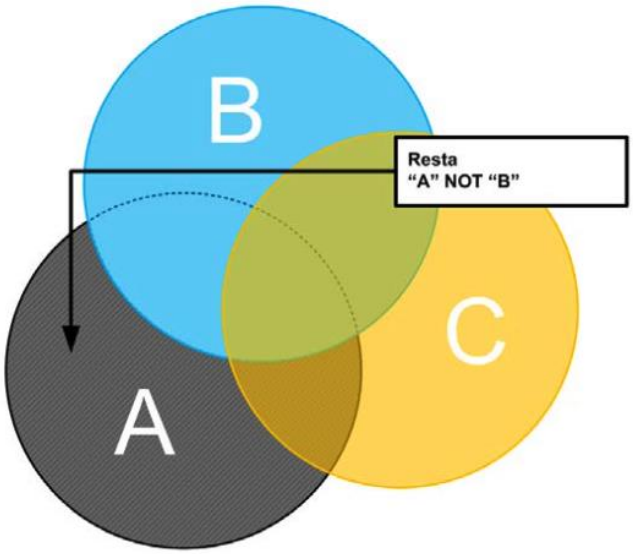
FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

Operador NOT :: Q = "A" NOT "B"

El operador NOT también conocido como de negación, implica resta, diferencia, reducción o sustracción. Esto es restar a un conjunto de documentos aquellos que contenga el término B. Obsérvese la Figura.- 11, en la que sólo se recuperan aquellos documentos que contengan los términos A pero no los términos B. [19]

Figura.- 11 Documentos que contienen el término A pero no el B



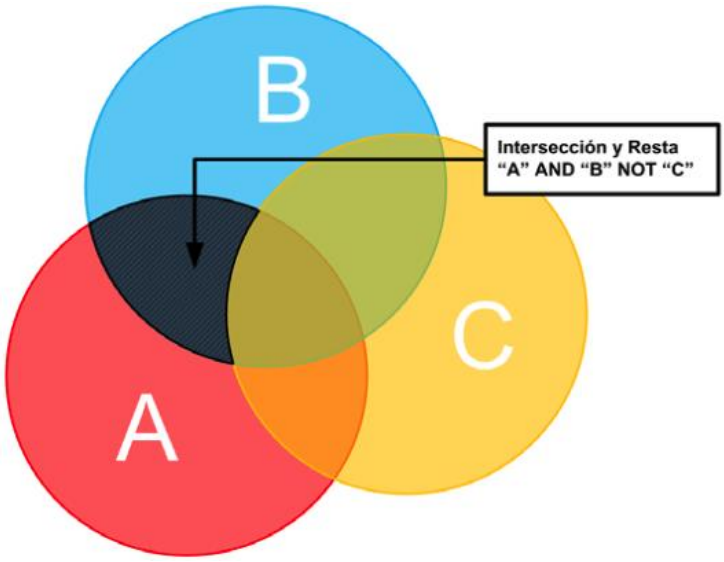
FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

Operador AND y NOT. Q = "A" AND "B" NOT "C"

La flexibilidad del lenguaje booleano permite combinar distintos operadores para obtener resultados más restringidos. Según se muestra en este caso, véase Figura.- 12, el operador AND y NOT pueden precisar la distinción de términos basándose en la negación de un tercero. De esta forma la consulta Q recuperará aquellos documentos en los que esté presente el término A y B pero no C. [19]

Figura.- 12 Documentos que contienen el término A y B pero no el C

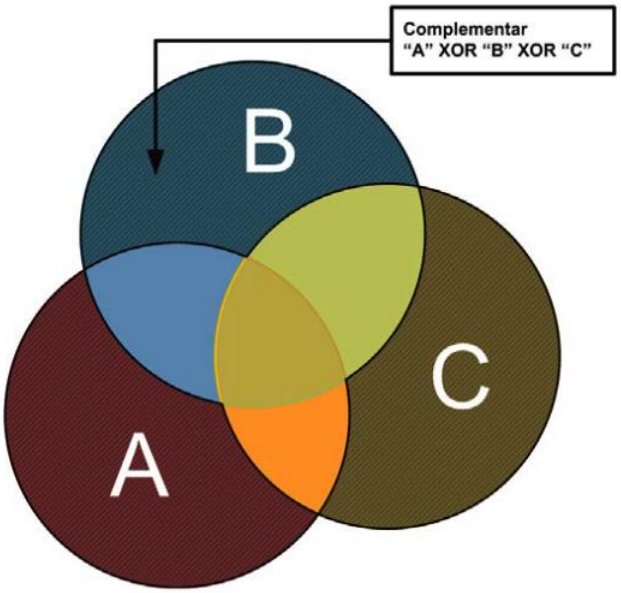


FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID
ELABORADO: MANUEL BLÁZQUEZ OCHANDO

Operador XOR :: Q = "A" XOR "B" XOR "C"

El operador XOR se utiliza para seleccionar todos los elementos complementarios de los conjuntos. Dicho de otra forma, evita las intersecciones. En la Figura.- 13, se observa que la zona de documentos que serán recuperados es aquella en la que no se combinan los términos A, B y C. Así por ejemplo, de esta forma la expresión A XOR B es equivalente a (A AND (NOT B)) OR ((NOT A) and B). [19]

Figura.- 13 Documentos cuyos términos complementarios sean A, B y C



FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

2.2.1.1. Ventajas del Modelo Booleano.

- El modelo booleano permite procesar colecciones muy grandes rápidamente. Resulta sistemático y ello supone una gran velocidad de recuperación. [19]
- Entraña ventajas para efectuar una recuperación de información igualada, en el sentido de que el sistema de información presente la mejor respuesta a una necesidad de información expresada por ciertas palabras clave. [19]

2.2.1.2. Desventajas del Modelo Booleano.

- En muchos casos, las necesidades de información son complejas y ello entraña cierta dificultad a la hora de expresar las consultas mediante fórmulas lógicas que pueden incluso llegar a concatenarse. [19]
- El volumen de resultados no se puede controlar, ya que la consulta plantea una resolución absoluta para toda la colección en la que se aplica. Esto significa que el resultado puede ser excesivamente grande o pequeño. [19]

- El carácter binario (consideración exclusivamente de la presencia/ausencia de los términos en los documentos) es el principal responsable de la equiparación exacta, siendo considerado la principal desventaja del modelo. [21]
- Los resultados obtenidos pueden ser perfectamente relevante o absolutamente irrelevante, no hay gradación o término medio, ya que el funcionamiento del modelo booleano se basa en equiparación exacta. Es decir que no ordena los documentos por orden de relevancia, tal como se llevaría a cabo en un modelo basado en pesos o ponderación de los términos. [19]

2.2.2. Modelo Probabilístico.

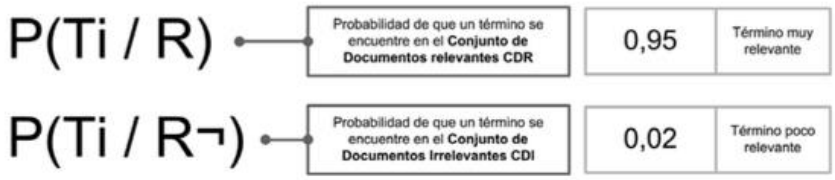
Desarrollado por Robertson y Sparck Jones, fue introducido entre 1977 y 1979 y es conocido como modelo probabilístico o de independencia binaria (BIR). [21] Se fundamenta en la representación binaria de los documentos, al igual que en el modelo de recuperación booleano, indicando presencia o ausencia de términos mediante 0 y 1. Su diferencia radica en el método estadístico y en las premisas bajo las que se constituye su funcionamiento estableciendo las siguientes aseveraciones: [19]

- Según la consulta planteada por el usuario, los documentos de la colección se clasifican en dos grupos; 1) Conjunto de Documentos Relevantes y 2) Conjunto de Documentos Irrelevantes. [19]
- Si el usuario supiese los términos de indización que permiten caracterizar tal subconjunto de documentos relevantes (porque aparecen en ellos y no aparecen en el resto de los documentos de la colección), tendríamos el problema resuelto. Como vemos, el modelo probabilístico parte exclusivamente de la presencia o ausencia de los términos en los documentos de la colección. Se trata, pues, también de un modelo binario, como el modelo booleano. [21]
- Aunque a priori se desconoce cuál es la Consulta Ideal (el usuario no tiene porqué conocerla), sí se sabe que es una combinación de 0 y 1 por ser un modelo binario de recuperación. Se desconocen por tanto los términos que se deberían introducir para obtener el Conjunto de Respuesta Ideal. [19]

2.2.2.1. Ponderación.

El objetivo del modelo probabilístico es tomar la consulta del usuario para ser refinada sucesivamente hasta obtener el conjunto de respuesta ideal, mediante la reformulación sucesiva de los términos de su consulta, empleando para ello la ponderación de los términos. Esto significa que se modifican los valores 1 (presencia) por un número (peso) que permita acercar la consulta imperfecta a una consulta ideal. El proceso de ponderación de los términos de la consulta es el cálculo de probabilidad de que exista dicho término en el conjunto de los documentos relevantes y la probabilidad de que se encuentre presente en el conjunto de los documentos irrelevantes. [19]

Figura.- 14 El cálculo de probabilidades como base para la ponderación de los términos



FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID
ELABORADO: MANUEL BLÁZQUEZ OCHANDO

El proceso de ponderación de los términos de la consulta es el cálculo de probabilidad de que dicho término en el conjunto de los documentos relevantes y la probabilidad de que se encuentre presente en el conjunto de los documentos irrelevantes. [19] Véase la Figura.- 14.

Figura.- 15 Método estándar para el cálculo de pesos de los términos de la consulta en el modelo probabilístico de independencia binaria

$$W_{(Ti)} = \log_{10} \frac{P(Ti / R)}{1 - P(Ti / R)} + \log_{10} \frac{1 - P(Ti / R\neg)}{P(Ti / R\neg)}$$

FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID
ELABORADO: MANUEL BLÁZQUEZ OCHANDO

El cálculo del peso para el término de la consulta "Ti" de Figura.- 15, incluye la suma de logaritmos de las probabilidades de presencia y ausencia en los conjuntos de documentos relevantes CDR (primera parte de la ecuación) y las probabilidades de presencia y ausencia en los conjuntos de documentos irrelevantes CDI (segunda parte de la ecuación). Aplicando los valores de aproximación inicial propuestos anteriormente, [19] su formulación se asemejaría a la que se muestra en la Figura.- 16.

Figura.- 16 Método estándar para el cálculo de pesos de los términos de la consulta en el modelo probabilístico de independencia binaria

$$W_{(Ti)} = \log_{10} \frac{0,5}{1 - 0,5} + \log_{10} \frac{1 - (ni / N)}{(ni / N)}$$

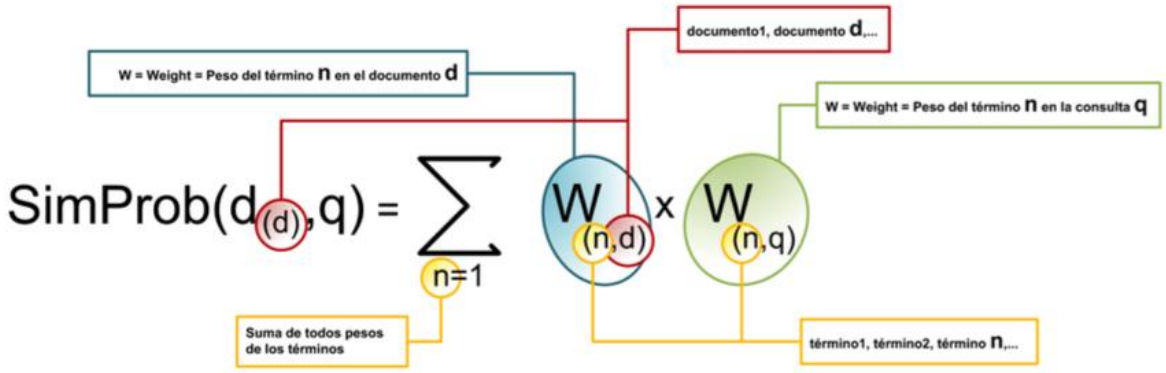
FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

2.2.2.2. El cálculo de la similitud.

Para cuantificar la similitud de los documentos de la colección con la consulta expresada por el usuario se emplea la siguiente formulación, véase Figura.- 17, que pone en relación el peso de los términos de la consulta del usuario con los del documento. Se trata de una variante del cálculo de similitud mediante el producto escalar, en la que el único elemento variable es el peso de la consulta.

Figura.- 17 Cálculo de similitud del modelo probabilístico de independencia binaria



FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

2.2.2.3. Ventajas del modelo probabilístico.

- De esta manera el modelo probabilístico supera el gran inconveniente puesto de manifiesto en el modelo booleano, a saber, la equiparación exacta. [21]
- En efecto, el modelo probabilístico, aun siendo un modelo binario, efectúa equiparación parcial, lo que permite ordenar los documentos de la respuesta conforme a su probabilidad de relevancia. [21]

2.2.2.4. Desventajas del modelo probabilístico.

- Mantiene el modelo binario de recuperación de información, no teniendo en cuenta todos los términos del documento como ocurriría en el modelo vectorial.
- Asigna pesos a los términos, permitiendo recuperar los documentos que probablemente sean irrelevantes.
- Requiere alta capacidad de computación, resultando complejo de implementar.
- Necesita efectuar una hipótesis inicial que no siempre resulta acertada.
- No tiene en cuenta la frecuencia de aparición de cada término en el documento, tal como lo haría un modelo vectorial. [19]

2.2.3. Modelo Vectorial.

El modelo de espacio vectorial se basa en el grado de similitud de una consulta dada por el usuario con respecto a los documentos de la colección cuyos términos fueron ponderados mediante TF-IDF. [19] El modelo vectorial no solo se identifica por representar los términos de la consulta por un conjunto ordenados de números, es decir la diferencia a modelos anteriores es que ya no se emplea un solo valor de 1 para identificar la presencia de un término excepto el valor de 0 que en este caso representaría la ausencia de un término.

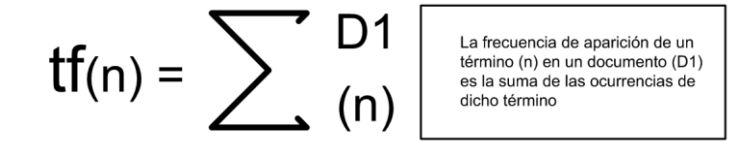
2.2.3.1. Proceso de ponderación TF-IDF.

La función de peso TF-IDF tiene la capacidad de asignar valores numéricos a los documentos basados en muy pocos factores. Estos incluyen la frecuencia de términos, el número de términos total en ese documento, el número de documentos en los que una palabra concreta aparece en la colección, y el número total de documentos. [22]

Frecuencia de términos.

Se denomina a las veces que aparece el término en un documento, véase Figura.- 18.

Figura.- 18 Frecuencia de término

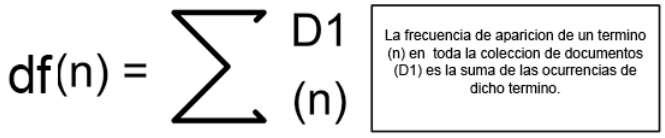


FUENTE: UNIVERSIDAD DE SRI LANKA
ELABORADO: CHAAMINDA MANJULA WIJEWICKREMA

Frecuencia de documento.

Corresponde al número de veces que aparece un término a lo largo de toda la colección de documentos, véase Figura.- 19.

Figura.- 19 Frecuencia de documento



FUENTE: UNIVERSIDAD DE SRI LANKA
ELABORADO: CHAAMINDA MANJULA WIJEWICKREMA

Frecuencia inversa del documento para un término.

Podría decirse que la capacidad de recuperación de un término es inversamente proporcional a su frecuencia en la colección de documentos. [20] Véase Figura.- 20.

Figura.- 20 Frecuencia inversa del documento para un término

$$IDF_{(n)} = \log \frac{N}{DF} + 1$$

FUENTE: UNIVERSIDAD DE SRI LANKA
ELABORADO: CHAAMINDA MANJULA WIJEWICKREMA

El factor (idf) es único para cada colección, su resultado será el logaritmo de 10 de la división de N el número de documentos de la colección para el (idf) y todo esto sumado a 1, debido a un factor de corrector de valores muy bajos.

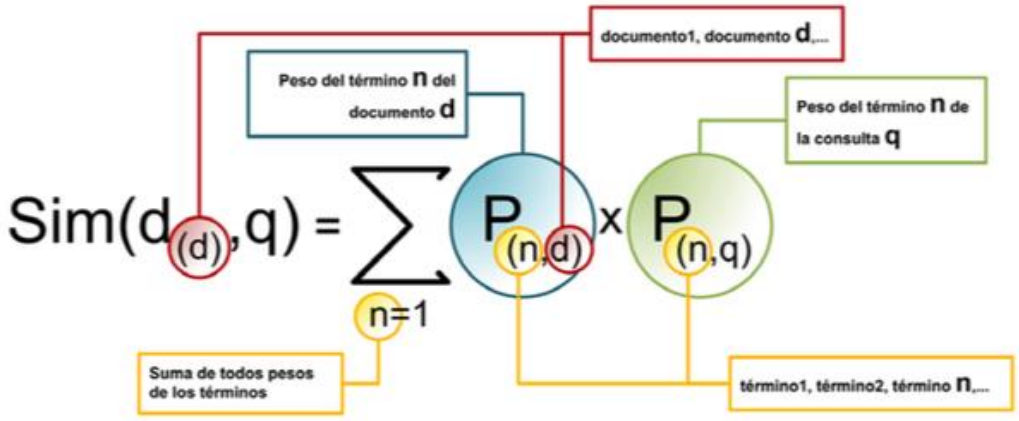
Finalmente se obtiene el peso de un término dado el productos de su TF x IDF como se muestra.

$$TF - IDF = TF\ x\ IDF$$

2.2.3.2. Equiparación mediante producto escalar.

Los procesos de equiparación de los documentos de la colección con respecto a la consulta del usuario, en el modelo booleano, se efectúan mediante cálculos de similitud. Existen muchas modalidades de comparación o equiparación mediante similitud, en este caso se presenta una de las más sencillas por su simplicidad y sistematización inmediata. [19]

Figura.- 21 Similitud de un documento "d" y la consulta "q" mediante producto escalar



FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

La similitud de un producto se obtiene de la sumatoria del producto de sus pesos, dado que $P_{(n,d)}$ es el peso del término relevante en el documento, $P_{(n,q)}$ pertenece al peso del término relevante en la consulta del usuario, este método es aplicable tanto en la modalidad de pesos binarios como en los pesos TF-IDF. [19]

2.2.3.3. Equiparación mediante la fórmula de coseno.

Es posible medir cuál es la desviación de un documento con respecto a una consulta, por el número de grados del ángulo que forman. Esto es posible porque crean una estructura triangular a la que se aplica el cálculo del ángulo que forma la hipotenusa (en este caso el vector del documento1) y el adyacente (el vector q de la consulta dada por el usuario) que resulta ser el coseno del triángulo. Cuando ambos vectores se muestran tan próximos como para superponerse, implicará que el ángulo que forman será menor y que su nivel de coincidencia será superior. De hecho, un coseno de 0o implicaría una similitud máxima. [19]

Figura.- 22 Fórmula para el cálculo de la similitud del coseno

$$SimCos(d_{(d)},q) = \frac{\sum_{n=1} \left(P_{(n,d)} \times P_{(n,q)} \right)}{\sqrt{\sum_{n=1} \left(P_{(n,d)} \right)^2 \times \sum_{n=1} \left(P_{(n,q)} \right)^2}}$$

FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID
ELABORADO: MANUEL BLÁZQUEZ OCHANDO

Por lo tanto, la fórmula aplicada para calcular el coeficiente de similitud del coseno entre un documento y una consulta es aquella que permite poner en relación los vectores de la consulta y del documento. De hecho el coseno de alfa de un triángulo cualquiera siempre es igual al cateto adyacente entre la hipotenusa. Tomando como clave esa idea, Figura.- 22 muestra la misma relación pero esta vez con los pesos que forman los vectores del documento y la consulta. De hecho el numerador no deja de ser un producto escalar entre los pesos del documento y la consulta; y el denominador la raíz cuadrada del producto del sumatorio de los pesos del documento y la consulta al cuadrado. La formulación del denominador con raíz cuadrada y cálculo de cuadrados, se diseñó para conseguir un resultado final de la división, inferior a 1, de tal manera que el coeficiente fuera de fácil manejo y lectura. [19]

2.2.3.4. Equiparación mediante Coeficiente de Dice.

El cálculo del coeficiente de similitud según Lee Raymond Dice es una adaptación del cálculo del coeficiente del coseno, véase la Figura.- 23.

Figura.- 23 Fórmula para el cálculo del coeficiente de similitud de Dice

$$SimDice(d_{(d)},q) = \frac{2 \times \sum_{n=1} \left(P_{(n,d)} \times P_{(n,q)} \right)}{\sqrt{\sum_{n=1} \left(P_{(n,d)} \right)^2 + \sum_{n=1} \left(P_{(n,q)} \right)^2}}$$

FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID
ELABORADO: MANUEL BLÁZQUEZ OCHANDO

2.2.3.5. Equiparación mediante el coeficiente de Jaccard

El cálculo del coeficiente de similitud de Jaccard al igual que el de Dice, resultan deudores del coeficiente de similitud del coseno. Su aplicación, centrada en usos estadísticos, también se aplica a recuperación de información y mide la similitud entre conjuntos. [19]

Figura.- 24 Fórmula para el cálculo del coeficiente de similitud de Jaccard

$$SimJacc(d_{(d)},q) = \frac{\sum_{n=1} \left(P_{(n,d)} \times P_{(n,q)} \right)}{\sum_{n=1} \left(P_{(n,d)} \right)^2 + \sum_{n=1} \left(P_{(n,q)} \right)^2 - \sum_{n=1} \left(P_{(n,d)} \times P_{(n,q)} \right)}$$

FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID
ELABORADO: MANUEL BLÁZQUEZ OCHANDO

2.2.4. Ventajas del modelo vectorial.

- El modelo vectorial es muy versátil y eficiente a la hora de generar rankings de precisión en colecciones de gran tamaño, lo que le hace idóneo para determinar la

equiparación parcial de los documentos. [19]

- Tiene en cuenta los pesos TF-IDF para determinar la representatividad de los documentos de la colección. [19]

2.2.5. Desventajas del modelo vectorial.

- El modelo vectorial por producto escalar tiene la desventaja de que sólo tiene en cuenta la intersección de los términos del documento con respecto a la consulta, por lo que la gradación de los resultados no es tan precisa como en el caso del cálculo del coseno. [19]
- Necesita de la intersección de los términos de la consulta con los documentos, en caso contrario no se produce la recuperación de información. [19]
- Al ser un modelo estadístico-matemático, no tiene en cuenta la estructura sintáctico-semántica del lenguaje natural. [19]

2.3. Marco referencial.

En el siguiente trabajo de investigación busca obtener información de materiales bibliográficos mediante técnicas avanzadas de recuperación de información para lo cual se ha revisado las siguientes publicaciones con el fin de dar un mayor aporte al actual trabajo.

Cada uno de los trabajos citados contribuye a la elaboración de cada etapa en la elaboración de este trabajo de investigación brindando información de fórmulas, procesos y técnicas que ayuden con la continuidad de dicho proyecto.

2.3.1. Aplicación de modelos de recuperación de información.

A la hora de comenzar con los procesos de recuperación de información para las recomendaciones de materiales bibliográficos, se debe tomar en cuenta el conjunto de datos que se utilizará, ya que la representación adecuada de un documento es clave para los procesos siguientes. Por razones históricas, los documentos han sido generalmente representados como conjuntos de términos. Dichos términos, denominados términos índice, palabras clave o etiquetas, son generados manualmente por especialistas por ejemplo, el caso de las fichas de una biblioteca, o bien automáticamente a partir del contenido del documento, como lo explica en [23], y además se expone con más detalle en la sección 2.1.9.

2.3.2. Proceso de depuración e indexación de términos de la consulta y documentos.

Antes de la manipulación de las consultas, términos, etiquetas y documentos se debe limpiar dicho contenido de palabras vacías o stopwords, ejemplo una consulta dada “Fundamentos de programación y técnicas de aprendizaje” tendría que ser depurada a “Fundamentos programación técnicas aprendizaje”. Si tales etiquetados no son eliminados, no se puede iniciar el procesamiento de la información y su correspondiente tratamiento. Por ello, se demuestra la importancia de aplicar mecanismos de depuración del código fuente, que facilite la extracción limpia de los textos, que serán la materia prima con la que se componen las colecciones sobre las que se recupera la información. [19]

En este sentido se trata de reducir al máximo el tamaño de las consultas y valores lingüísticos de los documentos para conseguir la mejor relación entre las consultas y el tiempo en que tarda en responder, como lo describe el autor de este apartado, a esta misión se la denomina “compresión de la indexación” y en ella se circunscriben los procesos de depuración que se han mostrado en el apartado anterior, tales como la supresión de palabras vacías[1] o stopwords. [19]

[1] Se eliminan del texto palabras que no son significativas en el proceso de selección de términos, denominadas *palabras vacías.* [29]

Para llevar a cabo la depuración e indexación, se requiere de unas prestaciones de hardware y capacidad de computación altas, lo suficiente como para soportar el procesamiento de este tipo de información debido a la gran cantidad filtros que se utiliza, esto aumenta en el caso de usar también para indexar en otros idiomas.

2.1.3 Sistema basado en contenido con modelado en espacios vectoriales.

En este artículo se describe el modelo usado como referencia para la realización del presente proyecto donde se explica la técnica utilizada en la propuesta del trabajo.

Se considera al modelo en espacios vectoriales o modelo vectorial dentro del área de recuperación de información perteneciente a los sistemas basados en contenido. Las recomendaciones son calculadas mediante el vector asociado al usuario obteniendo una similitud[2] (Producto Escalar, Coseno, Dice, Jaccard) con los vectores asociados a los ítems o documentos, finalmente se recomienda los N ítems cuyo perfil de usuario tenga una mayor similitud. [5]

2.1.4 Ponderación y obtención de valores TF-IDF de los términos de la consulta y consulta del usuario.

Para la aplicación de los procesos siguientes a este enunciado se debe ponderar los términos de la consulta y la consulta del usuario el resultado a estos conocimientos se lo detalla en el apartado Resultados y discusión. Los valores TF-IDF permiten calcular el grado de relevancia que tendrá cada término con respecto a cada documento [24], como se ha podido demostrar su significado y definición en el marco teórico de este proyecto de investigación.

[2] El término “similaridad” suele usarse para describir su significado en la jerga profesional española, pero debido a que es un término no admitido en la Real Academia Española se recomienda usar el término “similitud”. [28]

2.1.5 Modelo vectorial como técnica utilizada.

Como se ha podido mencionar las distintas técnicas de recuperación de información en la sección 2.1.9, se considera que tanto el modelo booleano y el modelo probabilístico efectúan equiparación exacta y parcial respectivamente, dando lugar a características negativas, en ambos modelos representan los términos por modelos binarios mientras el modelo vectorial tienen en cuenta la frecuencia de aparición de los términos en el conjunto de documentos o ítems obtenidos, como se menciona en el artículo [21].

En el modelo de espacio vectorial se emplea el peso para cada término del documento representado en la colección, en la Figura.- 25 se muestra los pesos binarios pertenecientes a cada término de la consulta del usuario en referencia a un documento, permitiendo su representación en el espacio vectorial permitiendo su tratamiento matemático. [19]

Figura.- 25 Representación del vector de documento

Id	Término	Documento 1	
		Peso binario	Peso TF-IDF
T1	Clima	1	1,452
T2	Biblioteca	0	0
T3	Universidad	1	2,122
T4	Alcalá	1	3,564
T5	España	1	4,123
T6	Libros	0	0
T7	Geografía	0	0
T8	Población	1	2,342
T9	Electricidad	0	0
T10	Ciencia	0	0
T11	Social	0	0
T12	Luz	1	1,975
T13	Unamuno	1	4,543
T14	Física	0	0
T15	Fluidos	1	6,134
T16	Literatura	1	2,234
Vector del documento1			
Documento1 { $Clima_{(1,452)}$, $Biblioteca_{(0)}$, $Universidad_{(2,122)}$, $Alcalá_{(3,564)}$, $España_{(4,123)}$, $Libros_{(0)}$, $Geografía_{(0)}$, $Población_{(2,342)}$, $Electricidad_{(0)}$, $Ciencia_{(0)}$, $Social_{(0)}$, $Luz_{(1,975)}$, $Unamuno_{(4,543)}$, $Física_{(0)}$, $Fluidos_{(6,134)}$, $Literatura_{(2,234)}$ }			

FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

Todos los documentos necesitan ser representados mediantes pesos TF-IDF, la consulta del usuario también requiere de dicho tratamiento. Ello significa que se tiene que ponderar la importancia de los términos de la consulta para poder generar el Vector de la consulta del usuario. Este paso es imprescindible para poder efectuar el Proceso de Equiparación de la consulta con los documentos de la colección y determinar cuáles de ellos son más relevantes [19], obsérvese la Figura.- 26.

Figura.- 26 Documento 1 y consulta del usuario con sus pesos

Cadena de consulta original del usuario				
Los libros y la literatura de Unamuno en la biblioteca de la Universidad de Alcalá				
Depuración de la consulta del usuario				
Libros Literatura Unamuno Biblioteca Universidad Alcalá				
Fichero diccionario		**Documento1**		**q = pesos de la consulta del usuario**
Id	**Término**	**Peso binario**	**Peso TF-IDF**	
T1	Clima	1	1,452	0
T2	Biblioteca	0	0	1,345
T3	Universidad	1	2,122	1,453
T4	Alcalá	1	3,564	1,987
T5	España	1	4,123	0
T6	Libros	0	0	2,133
T7	Geografía	0	0	0
T8	Población	1	2,342	0
T9	Electricidad	0	0	0
T10	Ciencia	0	0	0
T11	Social	0	0	0
T12	Luz	1	1,975	0
T13	Unamuno	1	4,543	3,452
T14	Física	0	0	0
T15	Fluidos	1	6,134	0
T16	Literatura	1	2,234	4,234

FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

2.1.5.1 Producto escalar como proceso en la técnica de modelo vectorial.

El producto escalar siendo un modelo de equiparación de los documentos con respecto a los términos de la consulta, el cual mediante el cálculo de similitud se podrá obtener resultados de forma más sencilla, en la sección 2.2.3.2 se puede observar la ecuación que forma parte de este modelo.

2.1.5.1.1 Modalidad de pesos binarios.

En la modalidad de pesos binarios, cada termino se identifica con valores booleanos es decir 1 y 0. Es decir para la presencia de un término dentro del documento o de la consulta del usuario se determina con el 1 y la ausencia del mismo con el valor de 0, esto da lugar a no realizar ningún cálculo respecto a estos valores ya que no inciden en el mismo. Siendo así que la sumatoria del producto de sus valores binarios obteniendo el valor de similitud. [19]

Figura.- 27 Pesos binarios del producto escalar

Cadena de consulta original del usuario			
Los libros y la literatura de Unamuno en la biblioteca de la Universidad de Alcalá			
Depuración de la consulta del usuario			
Libros Literatura Unamuno Biblioteca Universidad Alcalá			
Fichero diccionario		**Documento1**	**q = pesos binarios de la consulta del usuario**
Id	**Término**	**Peso binario**	
T1	Clima	1	0
T2	Biblioteca	0	1
T3	Universidad	1	1
T4	Alcalá	1	1
T5	España	1	0
T6	Libros	0	1
T7	Geografía	0	0
T8	Población	1	0
T9	Electricidad	0	0
T10	Ciencia	0	0
T11	Social	0	0
T12	Luz	1	0
T13	Unamuno	1	1
T14	Física	0	0
T15	Fluidos	1	0
T16	Literatura	1	1
Proceso de equiparación mediante el producto escalar de pesos binarios			
$\mathbf{Sim(doc1,q)} = {}^{Clima}(1*0) + {}^{Biblioteca}(0*1) + {}^{Universidad}(1*1) + {}^{Alcalá}(1*1) + {}^{España}(1*0) + {}^{Libros}(0*1) + {}^{Geografía}(0*0) + {}^{Población}(1*0) + {}^{Electricidad}(0*0) + {}^{Ciencia}(0*0) + {}^{Social}(0*0) + {}^{Luz}(1*0) + {}^{Unamuno}(1*1) + {}^{Física}(0*0) + {}^{Fluidos}(1*0) + {}^{Literatura}(1*1) = \mathbf{4}$			

FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

Como se puede analizar en la Figura.- 27 los términos Universidad, Alcalá, Unamuno, y Literatura, son los términos más relevantes tanto en la consulta del usuario como en el documento. Por lo tanto, en una escala de 6 debido a ser el número de términos de la consulta y que su producto es 1, el documento1 tiene un alto grado de similitud permitiendo ser un ítem susceptible a ser recomendado.

Como lo menciona un autor en su artículo, La asignación de pesos binarios 1 y 0, no es lo más apropiado para estos modelos. De hecho, las investigaciones que lo generaron siempre dedicaron especial esfuerzo al desarrollo de funciones de ponderación que son las que realmente le brindan mejor "performance" en la recuperación. [16]

2.1.5.1.2 Modalidad de pesos TF-IDF.

El método de modalidad de pesos TF-IDF, presenta resultados muchos más precisos en relación a la modalidad de pesos binarios dada por su simple función, obsérvese en la Figura.- 28 la representación del método empleado en varios escenarios. [19]

Figura.- 28 Representación de pesos TF-IDF en producto escalar

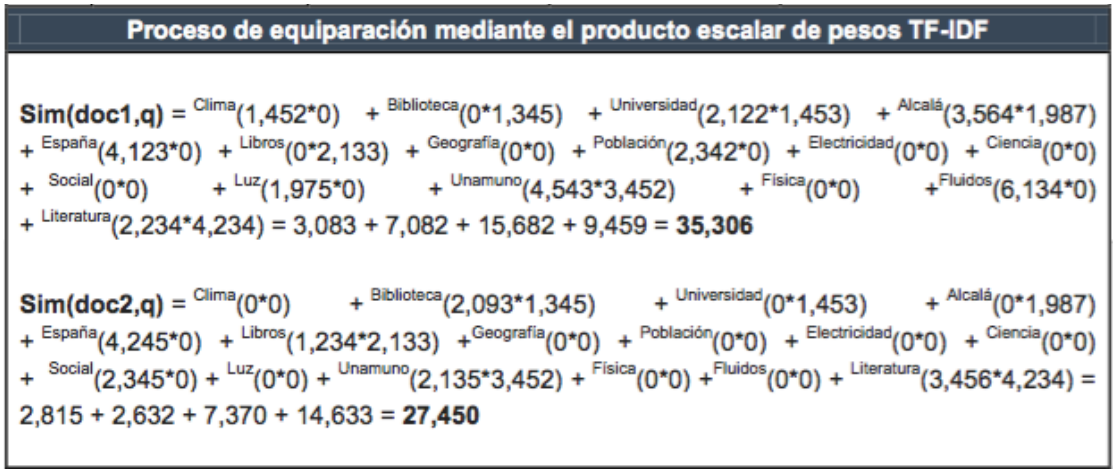
Cadena de consulta original del usuario				
Los libros y la literatura de Unamuno en la biblioteca de la Universidad de Alcalá				
Depuración de la consulta del usuario				
Libros Literatura Unamuno Biblioteca Universidad Alcalá				
Fichero diccionario		**Documento1**	**Documento2**	**q = pesos de la consulta del usuario**
Id	**Término**	**Peso TF-IDF**	**Peso TF-IDF**	
T1	Clima	1,452	0	0
T2	Biblioteca	0	2,093	1,345
T3	Universidad	2,122	0	1,453
T4	Alcalá	3,564	0	1,987
T5	España	4,123	4,245	0
T6	Libros	0	1,234	2,133
T7	Geografía	0	0	0
T8	Población	2,342	0	0
T9	Electricidad	0	0	0
T10	Ciencia	0	0	0
T11	Social	0	2,345	0
T12	Luz	1,975	0	0
T13	Unamuno	4,543	2,135	3,452
T14	Física	0	0	0
T15	Fluidos	6,134	0	0
T16	Literatura	2,234	3,456	4,234

FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

Mediante el proceso anteriormente mencionado se puede observar en la Figura.- 28 la representatividad de cada documento en relación con los términos de la consulta.

Figura.- 29 Muestra el cálculo de similitud de cada documento



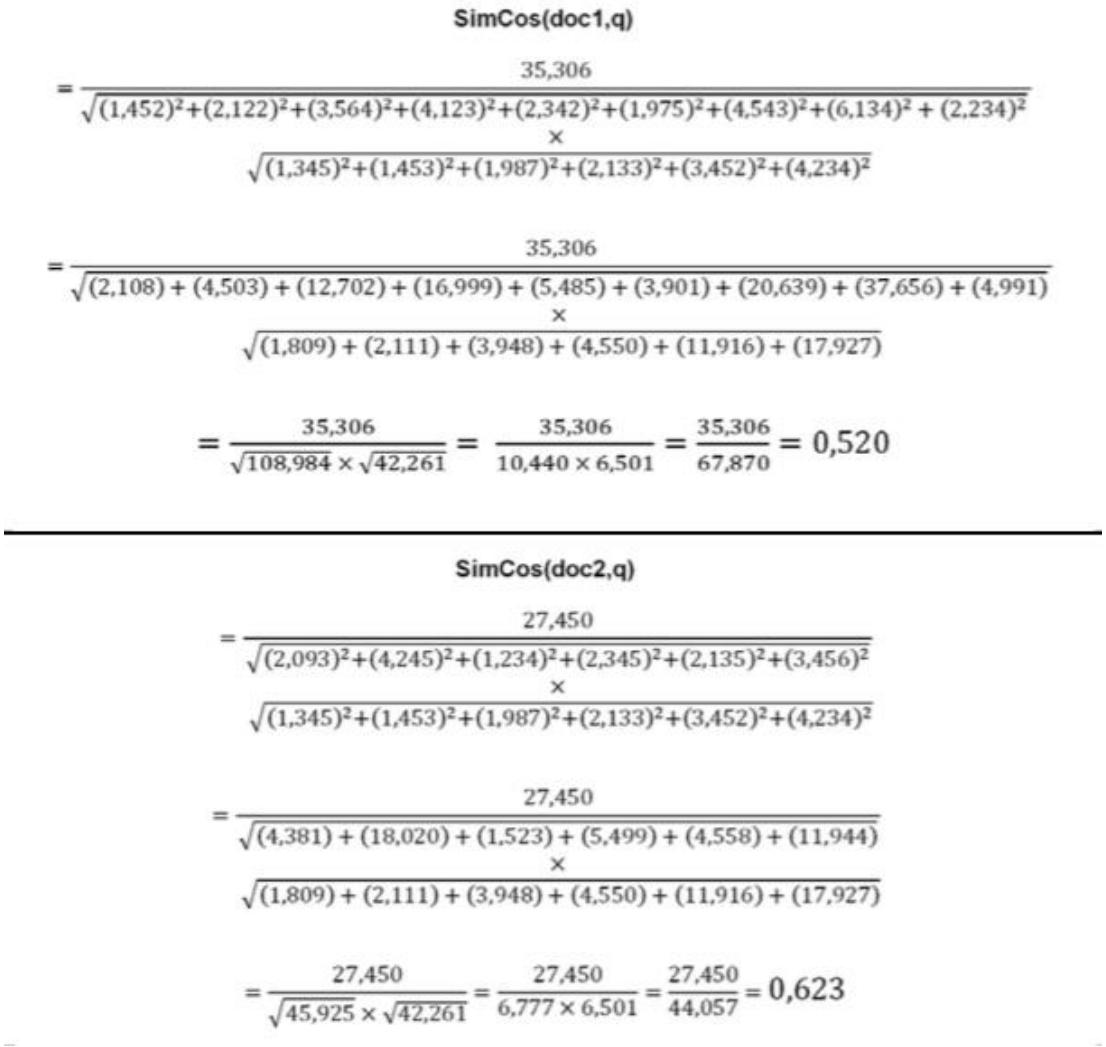
FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID
ELABORADO: MANUEL BLÁZQUEZ OCHANDO

Realizando un análisis de los resultados de Figura.- 29, para el docimento1 la similitud de la consulta del usuario será diferente con el documento2. De la misma manera que ocurría con los pesos binarios, tienen, mayor incidencia los términos de la consulta hacia los pesos de los documentos, debido a su múltiplo. La similitud del documento1 es superior a la del documento2 siendo este más susceptible a ser empleado en las recomendaciones, así su uso es más preciso que usar solo valores binarios. [19]

2.1.5.2 Formula del coseno como proceso en la técnica de modelo vectorial.

En el estudio [19] explica la función del ángulo de coseno este permite poner en relación el resultado de los vectores de los términos de la consulta y la consulta del usuario, del cual nos ubica cada vector en relación a un ángulo. Cada vector de documento mientras más cercano este del vector de la consulta indicando mayor coincidencia. En artículo [25] explica como el resultado que se obtiene debe variar entre 0 y 1, mientras el vector del documento este más cercano a 1, tiene más posibilidades de coincidencia con la consulta del usuario. Se puede observar la formula en la sección 2.2.3.3.

Figura.- 30 Muestra los resultados luego del proceso matemático de la fórmula del coseno



FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

En la figura podemos observar un ejemplo que el autor del artículo [19] explica, en este caso utilizando la fórmula del coseno, como la Figura.- 30 del producto escalar, el documento2 queda más representativo a razón del documento 1.

2.1.5.3 Coeficiente de Dice como proceso en la técnica de modelo vectorial.

El coeficiente de Dice es una adaptación del coeficiente del coseno, como se explica en este artículo [19], en la sección 2.2.3.4 se muestra la manipulación hecha por medio de la formula

dada, esta adaptación incluye en su enumerador un múltiplo de 2 con relación a los formula anteriormente expuesta en la fórmula del coseno. En Figura.- 31 se puede observar el ejemplo el autor del artículo ha realizado [19].

Figura.- 31 Muestra los resultados luego del proceso matemático de coeficiente de Dice

Proceso de equiparación mediante el coeficiente de Dice

SimDice(doc1,q)

$$= \frac{2 \times 35{,}306}{\sqrt{(1{,}452)^2+(2{,}122)^2+(3{,}564)^2+(4{,}123)^2+(2{,}342)^2+(1{,}975)^2+(4{,}543)^2+(6{,}134)^2+(2{,}234)^2} + \sqrt{(1{,}345)^2+(1{,}453)^2+(1{,}987)^2+(2{,}133)^2+(3{,}452)^2+(4{,}234)^2}}$$

$$= \frac{70{,}612}{\sqrt{(2{,}108)+(4{,}503)+(12{,}702)+(16{,}999)+(5{,}485)+(3{,}901)+(20{,}639)+(37{,}656)+(4{,}991)} + \sqrt{(1{,}809)+(2{,}111)+(3{,}948)+(4{,}550)+(11{,}916)+(17{,}927)}}$$

$$= \frac{70{,}612}{\sqrt{108{,}984} + \sqrt{42{,}261}} = \frac{70{,}612}{10{,}440 + 6{,}501} = \frac{70{,}612}{16{,}941} = 4{,}168$$

SimDice(doc2,q)

$$= \frac{2 \times 27{,}450}{\sqrt{(2{,}093)^2+(4{,}245)^2+(1{,}234)^2+(2{,}345)^2+(2{,}135)^2+(3{,}456)^2} + \sqrt{(1{,}345)^2+(1{,}453)^2+(1{,}987)^2+(2{,}133)^2+(3{,}452)^2+(4{,}234)^2}}$$

$$= \frac{54{,}900}{\sqrt{(4{,}381)+(18{,}020)+(1{,}523)+(5{,}499)+(4{,}558)+(11{,}944)} + \sqrt{(1{,}809)+(2{,}111)+(3{,}948)+(4{,}550)+(11{,}916)+(17{,}927)}}$$

$$= \frac{54{,}900}{\sqrt{45{,}925} + \sqrt{42{,}261}} = \frac{54{,}900}{6{,}777 + 6{,}501} = \frac{54{,}900}{13{,}278} = 4{,}135$$

FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

2.1.5.4 Coeficiente de Jaccard como proceso en la técnica de modelo vectorial.

El coeficiente de Jaccard al igual que el coeficiente de Dice utilizan la misma estructura del coeficiente de la fórmula del coseno. Pero se identifica algunos cambios, como en su

denominador se excluye la raíz cuadrada agregando a la suma de sus pesos de los términos de la consulta y la consulta del usuario el resto de la sumatoria de sus pesos. Los cambios realizados se pueden estudiar en la sección 2.2.3.5.

Figura.- 32 Muestra los resultados luego del proceso matemático de coeficiente de Jaccard

Proceso de equiparación mediante el coeficiente de Jaccard (Tanimoto)

SimJaccard(doc1,q)

$$= \frac{35{,}306}{(1{,}452)^2+(2{,}122)^2+(3{,}564)^2+(4{,}123)^2+(2{,}342)^2+(1{,}975)^2+(4{,}543)^2+(6{,}134)^2+(2{,}234)^2 + (1{,}345)^2+(1{,}453)^2+(1{,}987)^2+(2{,}133)^2+(3{,}452)^2+(4{,}234)^2 - 35{,}306}$$

$$= \frac{35{,}306}{(2{,}108)+(4{,}503)+(12{,}702)+(16{,}999)+(5{,}485)+(3{,}901)+(20{,}639)+(37{,}656)+(4{,}991) + (1{,}809)+(2{,}111)+(3{,}948)+(4{,}550)+(11{,}916)+(17{,}927) - 35{,}306}$$

$$= \frac{35{,}306}{108{,}984 + 42{,}261 - 35{,}306} = \frac{35{,}306}{115{,}939} = 0{,}305$$

SimJaccard(doc2,q)

$$= \frac{27{,}450}{(2{,}093)^2+(4{,}245)^2+(1{,}234)^2+(2{,}345)^2+(2{,}135)^2+(3{,}456)^2 + (1{,}345)^2+(1{,}453)^2+(1{,}987)^2+(2{,}133)^2+(3{,}452)^2+(4{,}234)^2 - 27{,}450}$$

$$= \frac{27{,}450}{(4{,}381)+(18{,}020)+(1{,}523)+(5{,}499)+(4{,}558)+(11{,}944) + (1{,}809)+(2{,}111)+(3{,}948)+(4{,}550)+(11{,}916)+(17{,}927) - 27{,}450}$$

$$= \frac{27{,}450}{45{,}925 + 42{,}261 - 27{,}450} = \frac{27{,}450}{60{,}736} = 0{,}452$$

FUENTE: UNIVERSIDAD COMPLUTENSE DE MADRID

ELABORADO: MANUEL BLÁZQUEZ OCHANDO

CAPÍTULO III
METODOLOGÍA DE LA INVESTIGACIÓN.

3.1. Localización.

El siguiente proyecto ha sido desarrollado en la ciudad de Quevedo provincia Los Ríos, en el cual forma parte de la Universidad Técnica Estatal de Quevedo donde se ha obtenido información de la institución como son: bases de datos e información necesaria para llevar a cabo este proyecto investigativo.

3.2. Tipo de investigación.

El presente trabajo corresponde al tipo de investigación diagnóstica debido a que se recopilará información de los distintos departamentos de la Universidad Técnica Estatal de Quevedo que aporten a la realización del sistema de recomendación de material bibliográfico tales como Unidad de Planeamiento Académico, Decanato y Coordinación de Carrera de la Facultad Ciencias de la Ingeniería, además de criterios recogidos al coordinador de Biblioteca de la institución educativa.

3.3. Métodos de investigación.

3.3.1. Método inductivo.

Mediante el análisis de los resultados obtenidos del sistema de recomendación de material bibliográfico y los procesos utilizados actualmente en la biblioteca de la Universidad Técnica Estatal de Quevedo se podrá determinar la eficiencia del sistema para realizar una búsqueda tratando así de incentivar el uso de la biblioteca a los estudiantes.

3.3.2. Método deductivo.

Por medio de la aplicación del presente método se podrá deducir las características que rigen al problema planteado por la Coordinación de Biblioteca de la UTEQ así como también los tiempos de espera que tarda un estudiante para encontrar los materiales bibliográficos que

estén acorde a las unidades de aprendizaje de su carrera contribuyendo así al desarrollo de la investigación.

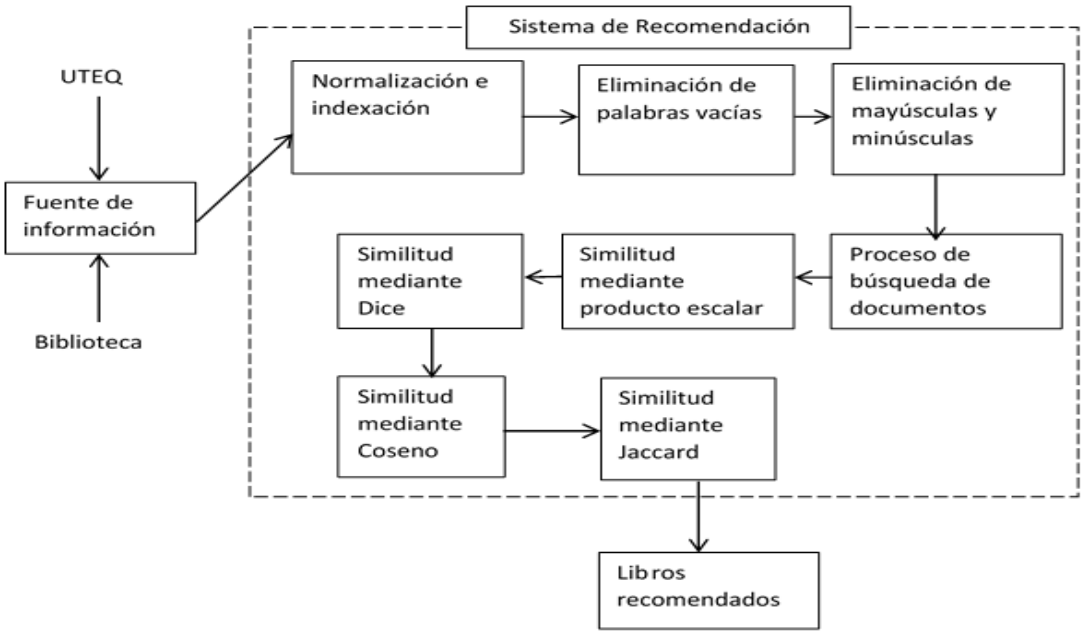
3.3.3. Método analítico.

Con la aplicación del método analítico se podrá realizar un análisis del proceso actual que se realiza dentro de la biblioteca para conseguir un determinado libro además de las variables que aportan al desarrollo del sistema de recomendación de material bibliográfico de la Universidad Técnica Estatal de Quevedo.

3.3.4. Modelo conceptual.

Por medio de un modelo se podrá mostrar la interacción de los procesos más importantes en la realización de un sistema capaz de mostrar el flujo de los procesos internos del sistema de recomendación de materiales bibliográficos, a continuación se muestra como modelo el diagrama de bloques del sistema:

Figura.- 33 Diagrama de bloques del sistema de recomendación de material bibliografico



FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

3.4. Fuentes de recopilación de información.

Las características de la presente investigación fueron obtenidas por medio de una entrevista realizada al Coordinador de biblioteca de la UTEQ en el año 2015 Ing. Jorge Guanín convirtiéndose así en la fuente primaria del trabajo investigativo planteado.

En el desarrollo del sistema de recomendación de material bibliográfico se utilizarán datos proporcionados por la Biblioteca en la cual se puede observar los libros con lo que cuenta la misma, así mismo la Unidad de Planeamiento Académico otorgando datos sobre las carreras de la distintas facultades, del Decanato y Coordinación de Carrera se obtuvo los datos correspondientes a los estudiantes que se encuentran en las distintas carreras que ofertan así como también de algunas mallas de estudio.

La elaboración del documento del proyecto de investigación se basa tomando en cuenta materiales bibliográficos de libros, revistas científicas, páginas web, bibliotecas virtuales y otros repositorios de información que aportan a la realización teórica del trabajo de tesis.

3.5. Diseño de investigación.

La presente investigación corresponde a la metodología cuasi experimental, ya que se realizarán diferentes grupos de pruebas de los resultados obtenidos de la búsqueda de materiales bibliográficos acorde a su carrera por parte de los estudiantes.

Con la utilización de varios grupos a ser observados se podrá inferir si existe mejoramiento en las respuestas del proceso aplicado actualmente frente al uso del sistema de recomendación de material bibliográfico de la UTEQ acorde a la malla de estudio de las carreras que oferta la misma.

3.6. Instrumento de investigación.

En la presente investigación que tiene como fin desarrollar un sistema de recomendación de material bibliográfico de la Universidad Técnica Estatal de Quevedo se utilizan los siguientes instrumentos de investigación.

3.6.1. Entrevista.

La entrevista fue aplicada al Coordinador de Biblioteca el mismo que manifiesta el proceso para elegir los diferentes materiales bibliográficos de la Universidad Técnica Estatal de Quevedo que reposan en la biblioteca de la institución educativa.

Mediante el uso de este instrumento de investigación se le permite al autor del trabajo de proyecto de investigación determinar las características que pueden aportar a reducir los tiempos de respuesta para consultar los materiales bibliográficos tratando de mejorar los procesos aplicados actualmente.

3.6.2. Observación.

Mediante la observación se puede analizar los tiempos de respuestas del proceso de búsqueda de material bibliográfico de la biblioteca de la Universidad Técnica Estatal de Quevedo frente a los tiempos de respuesta que se obtienen con el uso del sistema re-comendador.

Con el análisis de las observaciones se podrá deducir si existe mejora en el proceso de consultar materiales biográficos por parte de los estudiantes de las distintas carreras que oferta la UTEQ.

3.7. Tratamiento de los datos.

Los datos trabajados en este proyecto de investigación provienen de los repositorios de datos con los que cuenta la Universidad Técnica Estatal de Quevedo de manera específica de las áreas tales como Biblioteca, Unidad de Planeamiento Académico, Decanato y Coordinación de Carrera de la Facultad Ciencias de la Ingeniería (FCI).

Mediante los datos proporcionados por la Coordinación de la Biblioteca de la UTEQ se da por inicio al proceso de minería de datos, dentro de este conjunto de datos se obtendrá la información referente a los libros con los que cuenta el departamento de la institución educativa antes mencionada, la Unidad de planeamiento académico concedió información como lo es las diferentes facultades, carreras y mallas de estudio, mientras que la FCI proporciono información de los estudiantes que se encuentran matriculados en el periodo 2015 de las diferentes carreras.

EL conjunto de datos con los que se cuenta se le aplica minería de datos para realizar análisis y limpieza de los mismos permitiendo así obtener un almacén de datos que integre toda la información recolectada de los distintos departamentos los cuales se les aplicará un algoritmo de búsqueda.

3.8. Recursos humanos y materiales.

En el desarrollo del sistema re-comendador de material bibliográfico de las carreras que oferta la UTEQ se utilizan los siguientes recursos.

3.8.1. Recursos Humanos.

Tabla.- 4 Recursos Humanos

PERSONAL	DESCRIPCIÓN
AUTOR	Roger Alcívar
DIRECTOR DE TESIS	Ing. Yeikier Mendoza
COORDINADOR DE BIBLIOTECA AÑO 2014	Ing. Jorge Guanín
DECANO FCI	Ing. Jorge Murillo
COORDINADOR DE CARRERA INGENIERÍA EN SISTEMAS	Ing. Carlos Márquez de la Plata

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

3.8.2. Recursos de Software.

Tabla.- 5 Recursos de software

CANTIDAD	DESCRIPCIÓN	VALOR
1	Visual Studio C#	$ 0.00
1	SQL Server 2012	$ 0.00
1	Libre Office	$ 0.00

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

3.8.3. Recursos de hardware.

Tabla.- 6 Recursos de software

CANTIDAD	NOMBRE	DESCRIPCIÓN
1	Computador	Hp Compaq Intel Core i7 8 Gb.RAM 1000 Gb Disco Duro
1	Impresora	Canon Pixma MG3500
1	Celular	Samsung Core Prime

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

3.8.4. Diferentes recursos.

Tabla.- 7 Diferentes Recursos

CANTIDAD	MATERIAL	VALOR
	Libros	$ 00.00
	Útiles de oficina: Hojas, Carpetas, Copias, tinta	$ 40.00
3 meses	Servicio de internet	$ 80.00
	Movilización	$ 100.00
300 min	Llamada	$ 20.00
# unid.	Anillados	$ 64.00
# unid	Empastados	$ 45.00
TOTAL		$ 349.00

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

4.1. Resultados.

En este capítulo se realiza una exhaustiva explicación de los métodos y técnicas empleados en la búsqueda de información de la cual se implementa algoritmos que puedan resolver y mostrar los resultados esperados, como materiales de estudios se utiliza documentos de la Biblioteca que puedan enriquecer los conocimientos de los estudiantes de la Universidad Técnica Estatal de Quevedo.

Se detalla su funcionamiento, las entradas como parámetros para ejecutar los procesos y todas las operaciones que demuestran el logro de los objetivos planteados, las salidas que exponen las recomendaciones para los estudiantes de determinadas carreras y su correspondiente año de estudio. Adicionalmente, se presenta las herramientas, modelos y técnicas empleadas en el desarrollo integral del sistema del cual se describe el conjunto de datos usados como entradas que alimentan el sistema de recomendación.

4.1.1. Conjunto de datos.

El conjunto de datos fue obtenido de varias bases de datos proporcionadas por la Universidad Técnica Estatal de Quevedo. La fuente información fue obtenida en formato xls(Excel) los documentos, malla curricular en formato PDF, de esta manera se hacen modificaciones necesarias para el flujo del sistema de la cual se explican con más detalle en la sección 4.1.1.1, a continuación se detallan características y se añaden nuevas:

- Documentos:

Los documentos son las entidades principales de utilización en las recomendaciones de la cual contiene los principales atributos, número de documento con identificador individual, etiquetas o palabras claves añadidas.

- Usuarios y entorno principal.

La palabra estudiantes y/o usuarios se empleará para los términos a quienes se recomendará los documentos, en el caso de las materias será el grupo directo y entorno directo a la obtención de información hacia los usuarios principales, las unidades de estudios o materias tienen las siguientes características:

- Carreras:

Es la entidad de la cual le corresponde a cada estudiante conteniendo los siguientes atributos, número de carrera, nombre y relación con la facultad.

- Facultad:

La facultad corresponde a la especialidad o carrera que corresponde un estudiante de esta manera los atributos correspondientes son el número de facultad, nombre.

- Malla:

Corresponde a las materias o unidades de aprendizajes impartidas durante un periodo sus atributos son número de materias con sus respectivos periodos y carrera perteneciente.

4.1.1.1. Modificaciones realizadas.

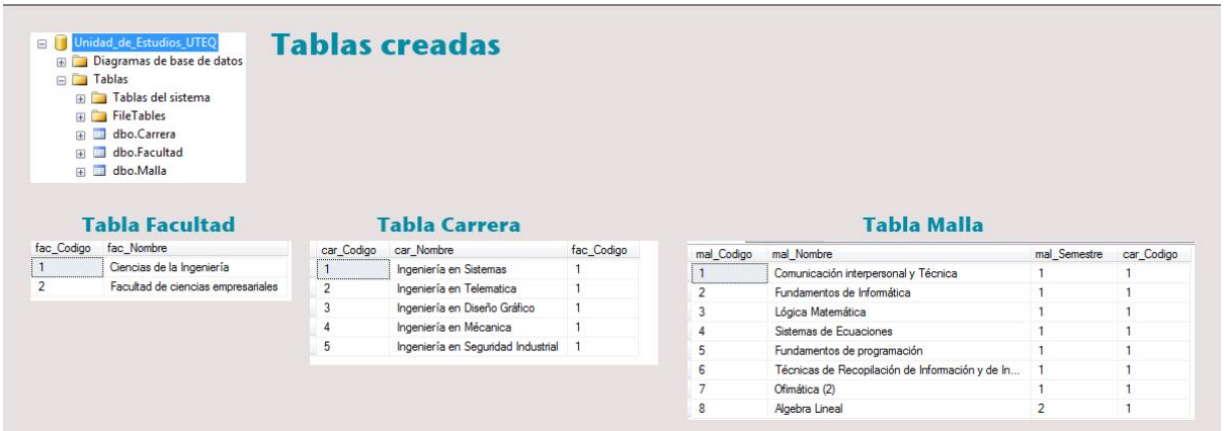
En vista que las fuentes de información obtenida contiene poco o nada de parámetros suficientes para realizar un correcto filtrado en las técnicas utilizadas, en este proyecto se ha procedido a la modificación y/o añadidos a la estructura para un mejor modelamiento de este tipo de información.

Se realiza una centralización de información en el cual se obtuvo la información de diferentes orígenes siendo estos migrados a Sql Server 2012 para una mejor ejecución de sus procesos, cabe mencionar que esta estructura no es igual a la original de la institución debido a que se necesitó modelar a la manera más oportuna para la adaptación de las técnicas empleadas en este proyecto de investigación, las modificaciones realizadas se adapta a cualquier parte del sistema mientras los parámetros que lo requiere se le asigne oportunamente, en los apartados siguientes se detallan los cambios realizados.

4.1.1.2. Cambio en la base de datos.

Tras un análisis exhaustivo, se determinó que la base de datos inicial, era muy básica, limitada y no cumplía con los requerimientos necesarios para el tratamiento de los datos, de esta manera se da a conocer la estructura más adaptable para el sistema. En la Figura.- 34 se aprecia la estructura definida de las facultades, carreras y malla curricular.

Figura.- 34 Muestra los cambios realizados en la estructura de las tablas



FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

De esta manera se puede obtener mejores resultados en la búsqueda información, se unió en una sola base de datos las tablas con sus correspondientes entidades facultad, carrera y malla curricular logrando así separar tanto las unidades de aprendizaje y los documentos dentro de la base de datos de la biblioteca.

La estructura principal de la base de datos de la biblioteca se ha hecho pocos cambios como se muestra en Figura.- 35 de los cuales se incluyó el campo “etiqueta” de esta manera podremos obtener mejor precisión en la búsqueda de información relevante para los estudiantes.

Figura.- 35 Muestra los cambios realizados en la estructura de las tablas en modo diseño

Nombre de columna	Tipo de datos
N	float
Titulo_del_Libro	nvarchar(255)
Volumen	float
Tomo	float
N_de_Ejemplar	float
Autor_es	nvarchar(255)
Editorial	nvarchar(255)
Ciudad	nvarchar(255)
Anio	float
Area_de_Conocimiento	nvarchar(255)
Carrera	nvarchar(255)
Precio	nvarchar(255)
Estado	nvarchar(255)
Etiquetas	nvarchar(1000)

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

El campo añadido brinda un mejor filtro para la búsqueda y comparación de los términos de la consulta del usuario, este campo “etiquetas” contiene las palabras claves o keywords que nos ayudaran a encontrar resultados más precisos, así como su utilidad en páginas web que brinda la posibilidad de encontrar de mejor manera lo que buscamos por ejemplo, podremos hacer una búsqueda en el buscador Google varias palabras como “Vacaciones en Brasil”, los resultados arrojados serán muchos, pero todo referente a la búsqueda realizada.

Figura.- 36 Muestra los campos más importantes en la tabla de Documentos

N	Titulo_del_Libro	Area_de_Conocimiento	Etiquetas
4	ACCOUNTING INFORMATION SYSTEMS	OBRAS GENERALES.	reportes, entidades, dispositivos, contable, public...
3	ACCESO A LA WAN	Informática.	redes, clasificacion, topologias, cables, intercone...
180	FUNDAMENTOS DESARROLLO WEB CON PHP, APACHE Y MYSQL	Informática.	fundamentos, guia, basica, introduccion, progra...
166	FUNDAMENTOS DE BASES DE DATOS	PROGRAMACION DE COMPUTADORAS, RPOGRAMAS, DATOS.	fundamentos, guia, basica, introduccion, base, d...
167	FUNDAMENTOS DE BASES DE DATOS	BASE DE DATOS.	fundamentos, guia, basica, introduccion, base, d...
168	FUNDAMENTOS DE BASES DE DATOS	PROGRAMACION DE COMPUTADORAS, RPOGRAMAS, DATOS.	fundamentos, guia, basica, introduccion, base, d...
174	FUNDAMENTOS DE PROGRAMACION ALGORITMOS, ESTRUCTU...	Ingeniería de Sistemas y Automática.	fundamentos, guia, basica, introduccion, algoritm...
176	FUNDAMENTOS DE PROGRAMACION, Algoritmos, estructura de dato...	Informática.	fundamentos, guia, basica, introduccion, algoritm...
219	INTRODUCCION A LA PROGRAMACION CON C++	Programas básicos.	fundamentos, codigo, guia, basica, programas, in...
197	GUIA DE PROGRAMACION 80386	Informática.	fundamentos, codigo, guia, basica, programas, in...
175	FUNDAMENTOS DE PROGRAMACION	Programas básicos.	fundamentos, codigo, guia, basica, programas, in...
165	FUNDAMENOS DE PROGRAMACION	Informática.	fundamentos, codigo, guia, basica, programas, in...
60	C/C++ COMO PROGRAMAR	Informática.	fundamentos, codigo, guia, basica, programas, in...
256	MANUAL DE VISUAL BASIC 5	Ingeniería de Sistemas y Automática.	fundamentos, codigo, guia, basica, programas, in...
35	APRENDA PRACTICANDO VISUAL BASIC 2005 USANDO VISUAL S...	PROGRAMACION DE COMPUTADORAS, PROGRAMAS, DATOS.	fundamentos, codigo, guia, basica, programas, in...

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Esta forma puede ser adaptada en su propia interfaz y así el administrador o administradores podrán incluirlas manualmente de acuerdo a cada documento definido en la base de datos. En la Figura.- 36 podemos observar que también en la tabla de las materias o malla curricular se agregó un campo "mal_Clave" que corresponde de igual forma a las keywords, cada palabra clave contiene parte del contenido de dicha materia importante para los modelos planteados en este proyecto de investigación.

Figura.- 37 Muestra los campos más importantes en la tabla de Malla

mal_Codigo	mal_Nombre	mal_Clave
1	Comunicación interpersonal y Técnica	relaciones, educación, concepto, manual, habilida...
2	Fundamentos de Informática	conceptos, algoritmos, programas, hardware, soft...
3	Lógica Matemática	lenguaje, recursion, formulas, analisis, funciones, ...
4	Sistemas de Ecuaciones	notacion, lineales, matriz, gauss
5	Fundamentos de programación	codigo, guia, basica, algoritmos, programas, introd...
6	Técnicas de Recopilación de Información y de In...	fases, proyecto, investigacion, metodos, experime...
7	Ofimática (2)	word, excel, hojas, access, administraccion, indic...
8	Algebra Lineal	vectores, matrices, polinomios, formulas, espacios,...
9	Grafos y Arboles	binario, estrella, subgrafos, recorridos, isomorfismo
10	Circuitos eléctricos	carga, electrica, corriente, magnitudes, basicas, le...
11	Programación Orientada a Objetos para aplicacio...	puntero, this, sobrecarga, operadores, basica, gui...
12	Contabilidad Comercial	contabilidad, comercio, numeros, estados, contabl...
13	Estructura de Datos	internas, estaticas, dinamicas, arreglos, matrices, li...
14	Ingles Elemental (2)	Hablar, conversar, leer, escribir, medio, basico
15	Base de Datos (2)	diseño, columnas, indices, consultas, anidadas, tri...

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

4.1.2. Descripción del escenario.

Por medio de la experimentación se busca estudiar el comportamiento de los algoritmos empleados en este proyecto de investigación según los parámetros de entradas o ajustes realizados. Cada resultado obtenido se evaluará y analizará las veces que sea necesaria para finalmente obtener conclusiones de cada experimento realizado y así poder determinar el algoritmo que más se adapte a las necesidades y objetivos de este proyecto. Cabe recordar que el sistema de recomendación realizado en este proyecto de investigación fue realizado como un prototipo, es decir no como un sistema final para la utilización de usuarios si no como una parte de un sistema que puede ser usado para un sistema completo.

4.1.3. Normalización e indexación de términos de consulta.

El proceso de generación de términos de la consulta lleva a cabo una serie de procedimientos que dan lugar a la limpieza y transformación de los textos de entrada, esta serie de modificaciones se enfoca en la reducción de textos innecesarios y así facilitar la búsqueda de información realizadas de consultas específicas.

4.1.3.1. Eliminación de StopWords.

En este proceso se elimina esas palabras, espacios y acentos que albergan dentro de las consultas obtenidas, que por su excesiva frecuencia y escasa utilidad no forman parte del contenido útil para el filtrado de información.

Tabla.- 8 Muestra la consulta original sin eliminar stopwords

Consulta del usuario Original
Ingeniería en Sistemas Fundamentos de programación código guía básica algoritmos programas introducción objetos clases

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO
ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Tabla.- 9 Muestra la consulta depurada eliminado los stopwords

Consulta del usuario Depurada
Ingeniería Sistemas Fundamentos programación código guía básica algoritmos programas introducción objetos clases

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO
ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

4.1.3.2. Eliminación de minúsculas y signos de puntuación.

En este proceso se transforma la información de tal manera que se cambie de minúsculas a mayúsculas, ya que así se evita la falla o inconsistencias en las comparaciones que se realizan en las técnicas de búsqueda de información. De esta forma se realiza la eliminación de acentos que puedan interrumpir el acceso a información congruente que facilite la comparación y esquematización de información relevante.

Tabla.- 10 Muestra la consulta depurada final

Consulta del usuario depurada
INGENIERIA SISTEMAS FUNDAMENTOS PROGRAMACION CODIGO GUIA BASICA ALGORITMOS PROGRAMAS INTRODUCCION OBJETOS CLASES

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO
ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

4.1.4. Preparación de los algoritmos.

En primera instancia para la realización y obtención de resultados se emplea el siguiente sistema de recomendación con sus respectivas técnicas:

Sistema de recomendación basado en contenido, este sistema fue escogido y analizado por el autor debido a que cumple con las condiciones necesarias y específicas que determinan el propósito y función del sistema.

Los sistemas basados en contenido buscan utilizar información disponible tanto del usuario como de documentos de esta manera realizar las recomendaciones más adecuadas, para este trabajo de investigación se usó técnicas avanzadas de recuperación de información en la cual forman parte del concepto y estructura que se lleva a cabo en los sistemas de recomendación basado en contenido, este campo brinda las posibilidades de manejar la información de manera más eficiente y aplicando distintas técnicas podemos determinar y evaluar los diferentes resultados o documentos obtenidos de un conjunto de datos, así ordenar la información obtenida en complemento con lo requerido y brindar las mejores respuestas.

Los sistemas de recomendación colaborativos y otros utilizan valores cuantitativos para

poder realizar el proceso de cálculos, por ejemplo, en una estructura de trabajo web donde los usuarios pueden valorar películas, estos valores se almacenan en forma numérica donde una película podría tener más peso dependiendo del número de valoraciones recibidas o el número de “me gusta” que recibió.

En muchos casos la información inicial que proporcionan los usuarios es inconsistente e irregular, de tal manera se determinó la obtención de información de forma implícita. Creando así consultas donde no hay interacción del usuario directamente, por este motivo se obtiene información de un grupo de información(Materias) y carreras creando un perfil que en conjunto forma una consulta general que servirá para todos los estudiantes que cursen una carrera en especial y por consiguientes sus respectivas materias.

De la misma forma se obtiene información de los ítems o documentos por el perfil de dicho usuario o en este caso de la materia, se trata sobre la identificación de la información de ítems o documentos y las preferencias o perfil del usuario, de esta manera se debe definir los principales esquemas de funcionamiento.

4.1.5. Modelamiento de preferencias.

4.1.5.1. Consulta del usuario.

En esta etapa obtenemos las preferencias del usuario o grupo de información (Materias), como se ha mencionó en el capítulo 4.1.1.1, la información o preferencias es obtenidas directamente de la materia perteneciente a una carrera en especial, estas preferencias forman parte del conjunto de información que servirá para el procesamiento y modelamiento de información necesaria para llevar a cabo la ejecución de los algoritmos de recomendación.

Las preferencias lingüísticas obtenidas brindan una mejor forma de procesar la información ya que es obtenida de forma implícita y no de forma directa de parte del usuario dando valores numéricos a la información, de esta manera podemos modelar la información teniendo variables y valores lingüísticos, en la siguiente Tabla.- 11 se muestra un ejemplo de la consulta del usuario o en este caso la información obtenida de la materia y carrera.

Tabla.- 11 Muestra el resultado de la obtención de la consulta del usuario

Carrera	Materia	Etiquetas
Ingeniería en Sistemas	Fundamentos de programación	código guía básica algoritmos programas introducción objetos clases
Consulta del usuario		
Ingeniería en Sistemas Fundamentos de programación código guía básica algoritmos programas introducción objetos clases		

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO
ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

4.1.5.2. Términos de la consulta.

Los términos de la consulta corresponden a la acción de crear un fichero diccionario de forma automática o manual de un conjunto de datos. Este proceso es necesario para obtener puntos de referencias con la consulta del usuario, esto significa que a cada palabra de la consulta del usuario se le asigna pesos, estos determinan si dicha palabra aparece en el documento.

Para este proyecto de investigación se utiliza todos los términos que aparecen en el conjunto de datos de la colección perteneciente a la consulta realizada, de esta manera se podrá ponderar los términos en los documentos [19] en la Tabla.- 12 se puede observar un ejemplo de los términos de la consulta, se usa de ejemplo como resultado del experimento la consulta del usuario en el apartado anterior.

Tabla.- 12 Muestra términos de la consulta del usuario sin palabras claves

Fichero Diccionario
Términos
COMUNICACION
INTERPERSONAL

TECNICA
FUNDAMENTOS
INFORMATICA
LOGICA
MATEMATICA
SISTEMAS
ECUACIONES
PROGRAMACION
RECOPILACION
INFORMACION
INVESTIGACION
OFIMATICA

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO
ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

En este ejemplo se muestra el fichero de términos obtenido de una consulta relacional de la facultad “Ciencias de la Ingeniería” seguido de la carrera “Ingeniería en sistemas” y del primer semestre. En la Tabla.- 13 se muestra los términos de la consulta completa, es decir junto a las palabras claves o etiquetas correspondientes a la materia seleccionada “Fundamentos de programación” como se muestra en Figura.- 37.

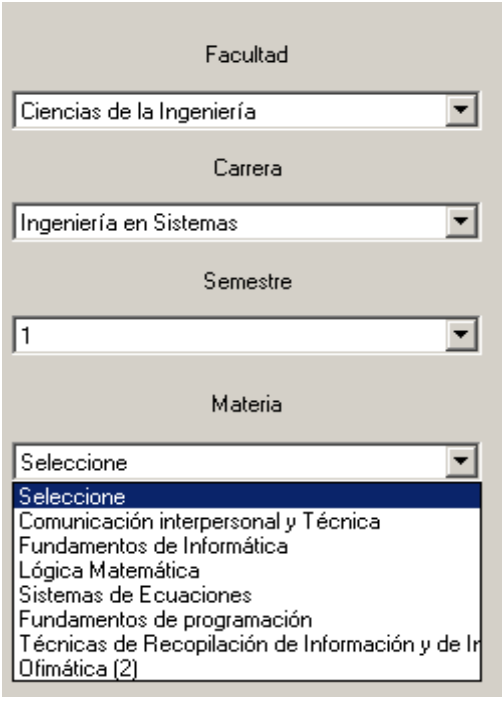
Tabla.- 13 Muestra términos de la consulta del usuario con palabras claves

Fichero Diccionario
Termino
INGENIERIA
SISTEMAS
COMUNICACION
INTERPERSONAL
TECNICA
FUNDAMENTOS
INFORMATICA
LOGICA
MATEMATICA
ECUACIONES
PROGRAMACION
TECNICAS
RECOPILACION
INFORMACION
INVESTIGACION
OFIMATICA
CODIGO
GUIA
BASICA
ALGORITMOS
PROGRAMAS
INTRODUCCION
OBJETOS
CLASES

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Figura.- 38 Muestra el origen de los términos



FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Los términos de la consulta forma parte en este caso de todas las palabras de la consulta de todo el periodo académico además se añade las etiquetas de la materia seleccionada y debidamente depurada.

4.1.5.3. Modelamiento de la información de los ítems.

Los elementos que se pueden recomendar se encuentran en una base de datos, en la Figura.- 39 se muestra unos simples registros de documentos que describen las propiedades más importantes para la manipulación y extracción de contenido.

Los nombres de columnas “Titulo”, “Area_de_conocimiento” y “Etiquetas” son los criterios que se utilizan para el filtro de información.

Figura.- 39 Muestra los campos a utilizar

N	Titulo_del_Libro	Area_de_Conocimiento	Etiquetas
4	ACCOUNTING INFORMATION SYSTEMS	OBRAS GENERALES.	reportes, entidades, dispositivos, contable, public...
3	ACCESO A LA WAN	Informática.	redes, clasificacion, topologias, cables, intercone...
180	FUNDAMENTOS DESARROLLO WEB CON PHP, APACHE Y MYSQL	Informática.	fundamentos, guia, basica, introduccion, progra...
166	FUNDAMENTOS DE BASES DE DATOS	PROGRAMACION DE COMPUTADORAS, RPOGRAMAS, DATOS.	fundamentos, guia, basica, introduccion, base, d...
167	FUNDAMENTOS DE BASES DE DATOS	BASE DE DATOS.	fundamentos, guia, basica, introduccion, base, d...
168	FUNDAMENTOS DE BASES DE DATOS	PROGRAMACION DE COMPUTADORAS, RPOGRAMAS, DATOS.	fundamentos, guia, basica, introduccion, base, d...
174	FUNDAMENTOS DE PROGRAMACION ALGORITMOS, ESTRUCTU...	Ingeniería de Sistemas y Automática.	fundamentos, guia, basica, introduccion, algoritm...
176	FUNDAMENTOS DE PROGRAMACION, Algoritmos, estructura de dato...	Informática.	fundamentos, guia, basica, introduccion, algoritm...
219	INTRODUCCION A LA PROGRAMACION CON C++	Programas básicos.	fundamentos, codigo, guia, basica, programas, in...
197	GUIA DE PROGRAMACION 80386	Informática.	fundamentos, codigo, guia, basica, programas, in...
175	FUNDAMENTOS DE PROGRAMACION	Programas básicos.	fundamentos, codigo, guia, basica, programas, in...
165	FUNDAMENOS DE PROGRAMACION	Informática.	fundamentos, codigo, guia, basica, programas, in...
60	C/C++ COMO PROGRAMAR	Informática.	fundamentos, codigo, guia, basica, programas, in...
256	MANUAL DE VISUAL BASIC 5	Ingeniería de Sistemas y Automática.	fundamentos, codigo, guia, basica, programas, in...
35	APRENDA PRACTICANDO VISUAL BASIC 2005 USANDO VISUAL S...	PROGRAMACION DE COMPUTADORAS, PROGRAMAS, DATOS.	fundamentos, codigo, guia, basica, programas, in...

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Cada registro contiene un valores lingüísticos para cada criterio, este valor junto con las preferencias del usuario forman un valor único que determinan la búsqueda y filtración de información para la obtención de los resultados.

En este apartado se mostrará una visión general de las técnicas de recomendación más utilizadas, como se ha mencionado anteriormente en este proyecto de investigación se utilizó el sistema de recomendación basado en contenido con más detalles se explica en el la sección 4.1.4.

En los sistemas basados en conocimiento se ha aplicado distintas técnicas, muchas de ellas son totalmente aplicables al esquema planteado, a continuación se presenta las técnicas más utilizadas.

Dentro del área de recuperación de información se utiliza el Modelo de espacios vectoriales debido a las ventajas que ofrece, siendo muy versátil y eficiente al utilizar colecciones de información de gran tamaño generando ranking de precisión, lo que es idóneo en la equiparación de documentos con la consulta del usuario, se detalla más en la sección 2.2.4.

En el Modelo de espacios vectoriales se hará la ponderación del diccionario de términos el cual se utiliza para los procesos matemáticos que determinarán el peso de cada término en relación a la colección de documentos. Para los primeros cálculos se necesita obtener el (DF) es decir la frecuencia de documento, donde tendremos las veces que se repite el término a lo largo de toda la colección de documento, en los anexos se detalla más el proceso en el motor del programa y script de generación.

Tabla.- 14 Muestra el valor frecuencia de documento (DF) de cada término

Diccionario de Términos	DF
INGENIERIA	33
SISTEMAS	48
COMUNICACION	3
INTERPERSONAL	0
TECNICA	2
FUNDAMENTOS	32
INFORMATICA	30
LOGICA	1
MATEMATICA	4
ECUACIONES	0
PROGRAMACION	38
TECNICAS	2
RECOPILACION	0
INFORMACION	9
INVESTIGACION	3
OFIMATICA	0
CODIGO	19
GUIA	26
BASICA	21
ALGORITMOS	12
PROGRAMAS	23
INTRODUCCION	29
OBJETOS	22
CLASES	11

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

El valor (DF) se utilizará a lo largo de todos los procesos matemáticos de aquí en adelante, ya que su utilización es vista en la obtención de los valores TF-IDF de cada término.

Tabla.- 15 Resultados de la ponderación inicial TF, IDF

				Documento 1	
N	**Diccionario de Términos**	**DF**	**Nº Documentos de la colección**	**TF**	**IDF**
1	INGENIERIA	33	136	0	1,615
2	SISTEMAS	48		0	1,452
3	COMUNICACION	3		0	2,656
4	INTERPERSONAL	0		0	0,000
5	TECNICA	2		0	2,833
6	FUNDAMENTOS	32		2	1,628
7	INFORMATICA	30		1	1,656
8	LOGICA	1		0	3,134
9	MATEMATICA	4		0	2,531
10	ECUACIONES	0		0	0,000
11	PROGRAMACION	38		1	1,554
12	TECNICAS	2		0	2,833
13	RECOPILACION	0		0	0,000
14	INFORMACION	9		0	2,179
15	INVESTIGACION	3		0	2,656
16	OFIMATICA	0		0	0,000
17	CODIGO	19		1	1,855
18	GUIA	26		2	1,719
19	BASICA	21		1	1,811
20	ALGORITMOS	12		1	2,054
21	PROGRAMAS	23		1	1,772
22	INTRODUCCION	29		1	1,671
23	OBJETOS	22		1	1,791
24	CLASES	11		1	2,092

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Para el ejemplo en este apartado se usa un documento obtenido de la colección de los mismos, en la Tabla.- 15 podemos observar el número de documentos obtenido de la colección que previamente filtrada por los valores de la consulta del usuario. Luego tenemos

la frecuencia de termino (TF) obtenido del conteo de repetición del termino dentro de los campos de del documento seguido de la frecuencia inversa de documento (IDF) que mediante el cálculo de los valores anteriores dan como resultado se puede ver en la Tabla.- 16 como se calculó dicho valor.

Tabla.- 16 Ejemplo de resultado frecuencia de documento inversa

Fundamentos	$\boldsymbol{IDF = Log_{10} \frac{N}{DF} + 1}$
	$IDF = Log_{10} \frac{136}{32} + 1$
	IDF = 1,628

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO
ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

El resultado de este ejemplo nos muestra el valor calculado de IDF como se comprueba en la Tabla.- 16, luego se procede con el cálculo del TF – IDF como muestra en la Tabla.- 17.

Tabla.- 17 Ejemplo de resultado ponderación TF – IDF al documento 1

		Documento 1
N	**Diccionario de Términos**	**TF-IDF**
1	INGENIERIA	0,000
2	SISTEMAS	0,000
3	COMUNICACION	0,000
4	INTERPERSONAL	0,000
5	TECNICA	0,000
6	FUNDAMENTOS	3,257
7	INFORMATICA	1,656
8	LOGICA	0,000
9	MATEMATICA	0,000
10	ECUACIONES	0,000
11	PROGRAMACION	1,554
12	TECNICAS	0,000

13	RECOPILACION	0,000
14	INFORMACION	0,000
15	INVESTIGACION	0,000
16	OFIMATICA	0,000
17	CODIGO	1,855
18	GUIA	3,437
19	BASICA	1,811
20	ALGORITMOS	2,054
21	PROGRAMAS	1,772
22	INTRODUCCION	1,671
23	OBJETOS	1,791
24	CLASES	2,092

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Siguiendo el ejemplo anterior para el cálculo de TF-IDF se realiza mediante el múltiplo del TF e IDF, véase la Tabla.- 18.

Tabla.- 18 Ejemplo de resultado ponderación TF – IDF a un término.

Fundamentos	$\boldsymbol{TF - IDF = TF\ x\ IDF}$
	$\boldsymbol{TF - IDF = 2\ x\ 1,628}$
	$\boldsymbol{TF - IDF = 3,257}$

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Además la consulta del usuario también necesita el mismo tratamiento, recordando que la consulta del usuario son los términos extraídos de la carrera seguido de la materia junto a las palabras claves de la materia. Véase la Tabla.- 19.

Tabla.- 19 Ejemplo de resultado ponderación TF – IDF al Documento y consulta del usuario

		Documento 1	q = Términos consulta usuario
N	**Diccionario de Términos**	**TF-IDF**	**TF-IDF**
1	INGENIERIA	0,000	1.615
2	SISTEMAS	0,000	1.452
3	COMUNICACION	0,000	0.000
4	INTERPERSONAL	0,000	0.000
5	TECNICA	0,000	0.000
6	FUNDAMENTOS	3,257	1.628
7	INFORMATICA	1,656	0.000
8	LOGICA	0,000	0.000
9	MATEMATICA	0,000	0.000
10	ECUACIONES	0,000	0.000
11	PROGRAMACION	1,554	1.554
12	TECNICAS	0,000	0.000
13	RECOPILACION	0,000	0.000
14	INFORMACION	0,000	0.000
15	INVESTIGACION	0,000	0.000
16	OFIMATICA	0,000	0.000
17	CODIGO	1,855	1.855
18	GUIA	3,437	1.719
19	BASICA	1,811	1.811
20	ALGORITMOS	2,054	2.054
21	PROGRAMAS	1,772	1.772
22	INTRODUCCION	1,671	1.671
23	OBJETOS	1,791	1.791
24	CLASES	2,092	2.092

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Los pesos de cada término corresponden al proceso de cálculo de cada factor correspondiente del documento mediante pesos TF-IDF, de esta manera se podrá emplear como valores puntuales y ejecutar modelos y procesos que devuelvan resultados deseados.
En este paso ya se podrá usar los diferentes métodos para calcular los resultados de las consultas dadas, mediante el tratamiento, de tal manera se podrá analizar y evaluar los resultados teniendo en cuenta las características de los algoritmo de aquí en adelantes usados en el proceso de recuperación de información.

Teniendo todos estos valores, el siguiente paso es los cálculos de similitud para cada modelo expuesto a continuación.

4.1.6. Proceso de búsqueda.

Como se ha mencionado en apartados anteriores, un sistema de recomendaciones basado en contenido proporciona información al usuario en base a su necesidad, en este caso la consulta es obtenida sin que el usuario intervenga directamente en el sistema. El sistema transformará la consulta dada a una representación que compara los términos de la consulta con el conjunto de documentos indexados que anteriormente fue filtrado por medio de la consulta brindada, esto permite la más rápida preparación y ejecución de los procesos empleados al no utilizar todo el conjunto de documentos que puedan haber en una tabla de datos.

4.1.7. Proceso de similitud de resultados.

En esta fase se da en funcionamiento los distintos métodos para la obtención de los resultados, cada proceso utiliza los cálculos efectuados en el apartado anterior, de esta manera se considera varias alternativas de cálculos de similitud (Producto Escalar, Coseno, Dice, Jaccard) cada uno tiene procesos y cálculos diferentes, de lo cual se tendrá varias alternativas que determinarán cual o cuales procesos brindará los requerimientos de la consulta del usuario.

4.1.8. Proceso de similitud mediante producto escalar.

En el modelo de espacio vectorial el proceso de similitud mediante producto escalar es el más básico debido a su simplicidad y por ende el más utilizado [26].
Utilizando los resultados finales en la Tabla.- 19, se podrá calcular los valores del producto escalar del documento. Siendo la sumatoria del producto de peso del documento y de la consulta del usuario.

$$Sim(doc1, q) = \sum_{n1} (P_{(n,d)}\ x\ P_{(n,q)})$$

Tabla.- 20 Ejemplo de resultado ponderación TF – IDF al Documento 1 y consulta del usuario.

Sim(doc1,q) = INGENIERIA(0 * 1.615) + SISTEMAS(0*1.452) + COMUNICACIÓN(0*0) + INTERPERSONAL(0*0) + TECNICA(0*0) + FUNDAMENTOS(3,257*1.628) + INFORMATICA(1,656*0) + LOGICA(0*0) + MATEMATICA(0*0) + ECUACIONES(0*0) + PROGRAMACION(1,554*1,554) + TECNICAS(0*0) + RECOPILACION(0*0) + INFORMACION(0*0) + INVESTIGACION(0*0) + OFIMATICA(0*0) + CODIGO(1,855*1,855) + GUIA(3,437*1.719) + BASICA(1.811*1.811) + ALGORITMOS(2,054*2,054) + PROGRAMAS(1,772*1,772) + INTRODUCCION(1,671*1,671) + OBJETOS(1,791*1,791) + CLASES(2,092*2,092) = **38,082**

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO
ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

En la Tabla.- 20 se muestra el proceso mediante la fórmula de similitud del producto escalar, este determina la similitud del documento con relación a los términos de la consulta y de la consulta del usuario en base a los términos de la consulta, dado que se resuelve mediante la sumatoria del producto sus pesos de los términos del documento y la consulta del usuario.

4.1.9. Proceso de similitud mediante formula de coseno.

En este proceso se utiliza una forma más compleja para realizar los cálculos de similitud, como el la formula anterior este recibe cambios como los siguientes.

$$Sim(doc1, q) = \frac{\sum_{n1}(P_{(n,d)}\ x\ P_{(n,q)})}{\sqrt{\sum_{n1}(P_{(n,d)})^2\ x\ \sum_{n1}(P_{(n,q)})^2}}$$

$$Sim(doc1, q) = \frac{\mathbf{38,082}}{\sqrt{\begin{matrix}\sum (3{,}257)^2 + (1{,}656)^2 + (1{,}554)^2 + (1{,}855)^2 + (1{,}811)^2 + (2{,}054)^2 + \\ (1{,}772)^2 + (1{,}671)^2 + (1{,}791)^2 + (2{,}092)^2 \\ x \\ \sum (1.615)^2 + (1.452)^2 + (1.628)^2 + (1.554)^2 + (1.855)^2 + (1.719)^2 + \\ (1.811)^2 + (2.054)^2 + (1.772)^2 + (1.671)^2 + (1.791)^2 + (2.092)^2\end{matrix}}}$$

$$Sim(doc1, q) = \frac{\mathbf{38,082}}{\sqrt{52{,}037\ x\ 37{,}192}}$$

$$Sim(doc1, q) = \frac{\mathbf{38,082}}{43{,}993}$$

$$Sim(doc1, q) = \mathbf{0,865}$$

Los resultados obtenidos una alta similitud debido que se acerca más al máximo valor del coeficiente del coseno siendo 1.

4.1.10. Proceso de similitud mediante el coeficiente de dice.

Este es un adaptación de la formula anterior cambiando por el múltiplo de 2 de su numerador.

$$Sim(doc1, q) = \frac{2\ x \sum_{n1}(P_{(n,d)}\ x\ P_{(n,q)})}{\sqrt{\sum_{n1}(P_{(n,d)})^2\ x\ \sum_{n1}(P_{(n,q)})^2}}$$

$$Sim(doc1, q) = \frac{\mathbf{2\ x\ 38,082}}{\sqrt{\begin{matrix}\sum (3{,}257)^2 + (1{,}656)^2 + (1{,}554)^2 + (1{,}855)^2 + (1{,}811)^2 + (2{,}054)^2 + \\ (1{,}772)^2 + (1{,}671)^2 + (1{,}791)^2 + (2{,}092)^2 \\ x \\ \sum (1.615)^2 + (1.452)^2 + (1.628)^2 + (1.554)^2 + (1.855)^2 + (1.719)^2 + \\ (1.811)^2 + (2.054)^2 + (1.772)^2 + (1.671)^2 + (1.791)^2 + (2.092)^2\end{matrix}}}$$

$$Sim(doc1, q) = \frac{\mathbf{76,164}}{\sqrt{52{,}037\ x\ 37{,}192}}$$

$$Sim(doc1, q) = \frac{\mathbf{76,164}}{43{,}993}$$

$$Sim(doc1, q) = \mathbf{1,731}$$

Para los resultados del coeficiente de Dice se realiza un orden en el conjunto de documentos siendo los más altos los más susceptible a recomendación por este método.

4.1.11. Proceso de similitud mediante el coeficiente de Jaccard.

Al igual que los dos modelos anteriores para el cálculo del coeficiente de Jaccard se basa de igual manera en sus fórmulas pero en este caso eliminado el denominador la raíz cuadrada y añadiendo la resta de su nominador.

$$Sim(doc1, q) = \frac{\sum_{n1}(P_{(n,d)}\ x\ P_{(n,q)})}{\sum_{n1}(P_{(n,d)})^2 + \sum_{n1}(P_{(n,q)})^2 - \sum_{n1}(P_{(n,d)}\ x\ P_{(n,q)})}$$

$$Sim(doc1, q) = \frac{\mathbf{38,082}}{\sum (3{,}257)^2 + (1{,}656)^2 + (1{,}554)^2 + (1{,}855)^2 + (1{,}811)^2 + (2{,}054)^2 + (1{,}772)^2 + (1{,}671)^2 + (1{,}791)^2 + (2{,}092)^2 \; + \; \sum (1.615)^2 + (1.452)^2 + (1.628)^2 + (1.554)^2 + (1.855)^2 + (1.719)^2 + (1.811)^2 + (2.054)^2 + (1.772)^2 + (1.671)^2 + (1.791)^2 + (2.092)^2 \; - \; 38{,}082}$$

$$Sim(doc1, q) = \frac{\mathbf{38,082}}{52{,}037 + 37{,}192 - 38{,}082}$$

$$Sim(doc1, q) = \frac{\mathbf{38,082}}{51{,}147}$$

$$Sim(doc1, q) = \mathbf{0,744}$$

El resultado al igual que en el coeficiente de Dice los valores más altos en el conjunto de documentos son los más relevantes.

4.1.12. Resultados esperados.

En la realización de este proyecto y demostración de resultados, se utiliza la consulta del usuario únicamente mediante la selección de carrera, semestre o periodo y materia, cuyos resultados serán explicados en base a cada una de las materias de la malla curricular. Como

trabajo de investigación se tomó como referencia el desarrollo de un sistema que ejecute los procesos anteriores teniendo un motor capaz de resolver la problemática de este proyecto, el cual puede ser adaptado y mejorar la forma de brindar las recomendaciones a los usuarios, en el apartado de recomendaciones se explica las ideas y entornos aplicable a este proyecto de investigación.

Habiendo aplicado los procesos detallado en apartados anteriores se procede a mostrar los resultados obtenidos en base a Tres consultas diferentes enumerados a continuación.

Tabla.- 21 Muestra las consultas a usar en los resultados esperados

Consulta Uno
Ingeniería en Sistemas Fundamentos de programación código guía básica algoritmos programas introducción objetos clases
Consulta Dos
Ingeniería en Sistemas Lógica Matemática recursión análisis funciones naturales disyunción negación
Consulta Tres
Ingeniería Sistemas Ofimática Word Excel hojas Access administración índices tablas imágenes formulas cálculos diagramas graficas

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO
ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

4.1.13. Resultados consulta Uno.

Para el resultado mediante el proceso de similitud de Producto Escalar, se puede observar los cálculos obtenidos para cada documento de un conjunto de 10 registros.

Tabla.- 22 Muestra los resultados de la consulta uno mediante el producto escalar

N	ID	Documento	Producto Escalar
1	3122	FUNDAMENTOS DE PROGRAMACION, Algoritmos, estructura de datos y objetos	60.416
2	15	80386 GUIA DE PROGRAMACION	54.091
3	3725	INTRODUCCION A LA PROGRAMACION CON C++	51.543
4	3120	FUNDAMENTOS DE PROGRAMACION ALGORITMOS, ESTRUCTURAS DE DATOS Y OBJETOS	47.839
5	3373	GUIA DE PROGRAMACION 80386	46.324
6	4572	MANUAL DE VISUAL BASIC 5	43.968
7	464	APRENDA PRACTICANDO VISUAL BASIC 2005 USANDO VISUAL STUDIO 2005	43.287
8	2984	FUNDAMENOS DE PROGRAMACION	43.287
9	463	APRENDA PRACTICANDO VISUAL BASIC 2005 USANDO VISUAL STUDIO 2005	38.216
10	921	C/C++ COMO PROGRAMAR	38.216

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

En la Tabla.- 22 puede observarse, que el orden de los documentos esta forma de mayor a menos en el campo Producto Escalar, resultando los de mayor valor más susceptible a ser recomendado.

Tabla.- 23 Muestra los resultados de la consulta uno mediante la fórmula de coseno

N	ID	Documento	Coseno
1	15	80386 GUIA DE PROGRAMACION	0.399
2	4572	MANUAL DE VISUAL BASIC 5	0.388
3	3373	GUIA DE PROGRAMACION 80386	0.379
4	3122	FUNDAMENTOS DE PROGRAMACION, Algoritmos, estructura de datos y objetos	0.377
5	464	APRENDA PRACTICANDO VISUAL BASIC 2005 USANDO VISUAL STUDIO 2005	0.374
6	2984	FUNDAMENOS DE PROGRAMACION	0.374
7	3725	INTRODUCCION A LA PROGRAMACION CON C++	0.367
8	463	APRENDA PRACTICANDO VISUAL BASIC 2005 USANDO VISUAL STUDIO 2005	0.350
9	921	C/C++ COMO PROGRAMAR	0.350
10	4493	MANUAL DE JAVA	0.338

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Para los resultados de la similitud del coseno en Tabla.- 23, se puede observar un cambio en la posición de los documentos en relación con el producto escalar.

Tabla.- 24 Muestra los resultados de la consulta uno coeficiente de Dice

N	ID	Documento	Dice
1	15	80386 GUIA DE PROGRAMACION	4.239
2	3122	FUNDAMENTOS DE PROGRAMACION, Algoritmos, estructura de datos y objetos	4.196
3	3725	INTRODUCCION A LA PROGRAMACION CON C++	3.936
4	3373	GUIA DE PROGRAMACION 80386	3.900
5	4572	MANUAL DE VISUAL BASIC 5	3.894
6	2984	FUNDAMENOS DE PROGRAMACION	3.784
7	464	APRENDA PRACTICANDO VISUAL BASIC 2005 USANDO VISUAL STUDIO 2005	3.784
8	3120	FUNDAMENTOS DE PROGRAMACION ALGORITMOS, ESTRUCTURAS DE DATOS Y OBJETOS	3.534

9	463	APRENDA PRACTICANDO VISUAL BASIC 2005 USANDO VISUAL STUDIO 2005	3.471
10	921	C/C++ COMO PROGRAMAR	3.471

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Para el caso de la Tabla.- 24 sucede algo similar a los anteriores, pero siendo estos documentos en su mayoría cambiados de posición con respectos a sus anteriores demostraciones, pero la diferencia aes significativa.

Tabla.- 25 Muestra los resultados de la consulta uno coeficiente de Jaccard

N	ID	Documento	Jaccard
1	4493	MANUAL DE JAVA	0.191
2	3931	JAVA COMO PROGRAMAR	0.191
3	4572	MANUAL DE VISUAL BASIC 5	0.184
4	922	C/C++ EDICION REVISADA Y ACTUALIZADA 2012	0.177
5	4492	MANUAL DE JAVA	0.177
6	3935	JAVA COMO PROGRAMAR	0.177
7	3141	FUNDAMENTOS DESARROLLO WEB CON PHP, APACHE Y MYSQL	0.175
8	2984	FUNDAMENOS DE PROGRAMACION	0.174
9	464	APRENDA PRACTICANDO VISUAL BASIC 2005 USANDO VISUAL STUDIO 2005	0.174
10	3373	GUIA DE PROGRAMACION 80386	0.169

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

El resultado de Jaccard en la Tabla.- 25 arroja diferentes valores cambiando su posición en relación a las anteriores técnicas.

4.1.13.1. Comparación de resultados consulta uno

Tabla.- 26 Muestra una comparativa de la consulta uno con los métodos Producto Escalar, Coseno Dice y Jaccard.

N	ID	Producto Escalar	ID	Coseno	ID	Dice	ID	Jaccard
1	3122	**60.416**	15	**0.399**	15	**4.239**	4493	**0.191**
2	15	**54.091**	4572	**0.388**	3122	**4.196**	3931	**0.191**
3	3725	**51.543**	3373	**0.379**	3725	**3.936**	4572	**0.184**
4	3120	**47.839**	3122	**0.377**	3373	**3.900**	922	**0.177**
5	3373	**46.324**	464	**0.374**	4572	**3.894**	4492	**0.177**
6	4572	**43.968**	2984	**0.374**	2984	**3.784**	3935	**0.177**
7	464	**43.287**	3725	**0.367**	464	**3.784**	3141	**0.175**
8	2984	**43.287**	463	**0.350**	3120	**3.534**	2984	**0.174**
9	463	**38.216**	921	**0.350**	463	**3.471**	464	**0.174**
10	921	**38.216**	4493	**0.338**	921	**3.471**	3373	**0.169**

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

En la Tabla.- 26 se observa la comparación de los 4 resultados obtenidos, se puede analizar, que en la similitud del producto escalar los primeros 5 resultados corresponden a la consulta del usuario “Fundamentos de programación” siendo seleccionado usuarios expertos como el método que se adapta mejor a la consulta, además que se puede notar una diferencia de documentos entre los tres primeros resultados con los de Jaccard resultando este último un poco diferente de los demás.

4.1.14. Resultados consulta dos

Para el resultado mediante el proceso de similitud de Producto Escalar, se puede observar los cálculos obtenidos para cada documento de un conjunto de 10 registros.

Tabla.- 27 Muestra los resultados de la consulta dos mediante el producto escalar

N	ID	Documento	Producto Escalar
1	1601	CURSO DE MATEMATICAS SUPERIORES "FUNCIONES USUALES"	14.751
2	4803	MATEMATICA I	10.324
3	3108	FUNDAMENTOS DE MATEMATICAS	10.324
4	2981	FUNCIONES DE PERFIL PARA LA CUBICACION DE ARBOLES EN PIE CON CLASIFICACION DE PRODUCTOS	10.050
5	2773	FÍSICA PARA CIENCIAS E INGENIERA	10.050
6	4050	LA INGENIERA EN LOS PROCESOS DE DESERTIFICACION	10.050
7	3145	FUNDAMENTOS MATEMÁTICOS DE LA INGENIERA	10.050
8	3590	INFORME DEL EJERCICIO DE FUNCIONES	10.050
9	3379	GUIA DEL PATRIMO DE AREAS NATURALES PROTEGIDAS DEL ECUADOR	9.548
10	4226	LEY FORESTAL Y DE CONSERVACION DE AREAS NATURALES Y VIDA SILVESTRE	9.548

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Se puede observar la diferencia entre la cantidad de valor que tiene el máximo en la Tabla.- 22 del primer resultado con lo devuelto de la Tabla.- 27 esto se debe a la cantidad de palabras claves existen tanto en la tabla de Malla o Materias como también en los documentos.

Tabla.- 28 Muestra los resultados de la consulta dos mediante la fórmula de coseno

N	ID	Documento	Coseno
1	4286	LOGICA MATEMATICA PARA INFORMATICOS (EJERCICIO RESUELTOS)	0.276
2	1618	CURSOS DE MATEMATICAS SUPERIORES "ALGEBRA"	0.228
3	1620	CURSOS DE MATEMATICAS SUPERIORES "GEOMETRIA"	0.228
4	379	ANALISIS ECONOMETRICO	0.228
5	408	ANALISIS Y DISEÑOS DE SISTEMAS	0.228
6	5138	MICROSOFT EXCEL Y MICROSOFT WORD PARA ABODADOS	0.228

7	360	ANALISIS DE GENERO EN LA INVESTIGACION Y TRANSFERENCIA DE TECNOLOGIAS MEJORADAS EN LOS SISTEMAS DE PRODUCCION AGRICOLA	0.228
8	374	ANALISIS DE SISTEMAS ELECTRICOS DE POTENCIA	0.228
9	375	ANALISIS DINAMICO DE SISTEMAS INDUSTRIALES	0.228
10	385	ANALISIS ESTADISTICO SIMPLIFICADO	0.228

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

En la Tabla.- 28 se puede observar una diferencia en la obtención de estos resultados, debido a la poca información por parte de palabras claves, además el método del coseno, dice y jaccard dependen de los resultados del productor escalar, por lo que los resultados pueden coincidir directamente con la primer técnica.

Tabla.- 29 Muestra los resultados de la consulta dos coeficiente de Dice

N	ID	Documento	Dice
1	1601	CURSO DE MATEMATICAS SUPERIORES "FUNCIONES USUALES"	1.707
2	4803	MATEMATICA I	1.478
3	3108	FUNDAMENTOS DE MATEMATICAS	1.478
4	4286	LOGICA MATEMATICA PARA INFORMATICOS (EJERCICIO RESUELTOS)	1.470
5	3106	FUNDAMENTOS DE MATEMATICA	1.338
6	5366	PARQUES NACIONALES Y OTRAS AREAS NATURALES PROTEGIDAS DEL ECUADOR	1.255
7	563	AUDITORIA DE TECNOLOGIAS Y SISTEMAS DE INFORMACION	1.255
8	2760	EXCEL ANALISIS DE DATOS EMPRESARIALES	1.251
9	2773	FÍSICA PARA CIENCIAS E INGENIERA	1.237
10	4050	LA INGENIERA EN LOS PROCESOS DE DESERTIFICACION	1.237

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

En estos resultados se puede observar diferencias entre la posición y diferencia de los documentos en relación con los modelos anteriores.

Tabla.- 30 Muestra los resultados de la consulta dos coeficientes de Jaccard

N	ID	Documento	Jaccard
1	4286	LOGICA MATEMATICA PARA INFORMATICOS (EJERCICIO RESUELTOS)	0.153
2	3106	FUNDAMENTOS DE MATEMATICA	0.126
3	5366	PARQUES NACIONALES Y OTRAS AREAS NATURALES PROTEGIDAS DEL ECUADOR	0.121
4	2760	EXCEL ANALISIS DE DATOS EMPRESARIALES	0.121
5	563	AUDITORIA DE TECNOLOGIAS Y SISTEMAS DE INFORMACION	0.120
6	4856	MATEMATICAS PARA EL ANALISIS ECONOMICO	0.120
7	4860	MATEMATICAS PREVIAS AL CALCULO ANALISIS FUNCIONAL Y GEOMETRIA ANALITICA	0.120
8	4796	MATEMATICA COMERCIAL Y FINANCIERA ANALISIS PRACTICO PARA PROFESIONES JURIDICAS Y ECONOMICAS	0.120
9	4803	MATEMATICA I	0.117
10	3108	FUNDAMENTOS DE MATEMATICAS	0.117

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Los resultados de la Tabla.- 30 se puede observar una mejor coincidencia de los primeros 4 documentos con la consulta del usuario brindando buenas posibilidades de aceptación.

4.1.14.1. Comparación de resultados consulta dos

Tabla.- 31 Muestra una comparativa de la consulta uno con los métodos Producto Escalar, Coseno

N	ID	Producto Escalar	ID	Coseno	ID	Dice	ID	Jaccard
1	1601	**14.751**	4286	**0.276**	1601	**1.707**	4286	**0.153**
2	4803	**10.324**	1618	**0.228**	4803	**1.478**	3106	**0.126**
3	3108	**10.324**	1620	**0.228**	3108	**1.478**	5366	**0.121**
4	2981	**10.050**	379	**0.228**	4286	**1.470**	2760	**0.121**
5	2773	**10.050**	408	**0.228**	3106	**1.338**	563	**0.120**
6	4050	**10.050**	5138	**0.228**	5366	**1.255**	4856	**0.120**
7	3145	**10.050**	360	**0.228**	563	**1.255**	4860	**0.120**
8	3590	**10.050**	374	**0.228**	2760	**1.251**	4796	**0.120**
9	3379	**9.548**	375	**0.228**	2773	**1.237**	4803	**0.117**
10	4226	**9.548**	385	**0.228**	4050	**1.237**	3108	**0.117**

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Se puede analizar en la Tabla.- 31 que hubo un cambio de posición de documentos de los valores de coseno, dice y jaccard en relación con el producto escalar, corroborando el análisis de un experto se puede observar que los resultados lingüísticos arrojados por el producto escalar podrían satisfacer los términos de la consulta, siendo altamente susceptible a ser recomendados.

4.1.15. Resultados consulta tres.

Tabla.- 32 Muestra los resultados de la consulta tres mediante el producto escalar

N	ID	Documento	Producto Escalar
1	5146	MICROSOFT WORD 6 PARA WINDOWS	38.200
2	5138	MICROSOFT EXCEL Y MICROSOFT WORD PARA ABODADOS	27.552
3	6547	TABLAS DINAMICAS EN EXCEL 2007	23.869
4	367	ANALISIS DE LOS NEGOCIOS CON EXCEL XP	22.503
5	368	ANALISIS DE LOS NEGOCIOS CON EXCEL XP	22.503

6	2760	EXCEL ANALISIS DE DATOS EMPRESARIALES	20.910
7	2758	EXCEL 4 PARA WINDOWS	19.458
8	2759	EXCEL 5 PARA WINDOWS	19.458
9	2762	EXCEL APLICACIONES EN ALGEBRA, ESTADISTICA, PROBABILIDAD Y FISICA	19.458
10	89	Administracion de la empresa con Microsoft Excel	19.458

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Los resultados del producto escalar en la Tabla.- 32 arrojados demuestran que los resultados textuales de los documentos brindan posibilidades altas de coincidencia con la consulta del usuario.

Tabla.- 33 Muestra los resultados de la consulta tres mediante la fórmula de coseno

N	ID	Documento	Coseno
1	5138	MICROSOFT EXCEL Y MICROSOFT WORD PARA ABODADOS	0.294
2	5146	MICROSOFT WORD 6 PARA WINDOWS	0.288
3	6547	TABLAS DINAMICAS EN EXCEL 2007	0.288
4	2760	EXCEL ANALISIS DE DATOS EMPRESARIALES	0.279
5	2758	EXCEL 4 PARA WINDOWS	0.263
6	2759	EXCEL 5 PARA WINDOWS	0.263
7	89	Administracion de la empresa con Microsoft Excel	0.263
8	357	ANALISIS DE DATOS CON MICROSOFT EXCEL ACTUALIZADO PARA OFFICE 2000	0.263
9	2762	EXCEL APLICACIONES EN ALGEBRA, ESTADISTICA, PROBABILIDAD Y FISICA	0.263
10	4155	LAS MACROS EN MICROSOFT EXCEL 2010	0.263

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Al igual que en el anterior proceso la posición y ordenación de los resultados en la Tabla.- 32 no disminuyen el análisis favorable en los resultados arrojados en la Tabla.- 33 en el proceso de equiparación del coseno.

Tabla.- 34 Muestra los resultados de la consulta tres coeficiente de Dice

N	ID	Documento	Dice
1	5146	MICROSOFT WORD 6 PARA WINDOWS	3.164
2	5138	MICROSOFT EXCEL Y MICROSOFT WORD PARA ABODADOS	2.820
3	6547	TABLAS DINAMICAS EN EXCEL 2007	2.614
4	2760	EXCEL ANALISIS DE DATOS EMPRESARIALES	2.412
5	367	ANALISIS DE LOS NEGOCIOS CON EXCEL XP	2.409
6	368	ANALISIS DE LOS NEGOCIOS CON EXCEL XP	2.409
7	89	Administracion de la empresa con Microsoft Excel	2.261
8	357	ANALISIS DE DATOS CON MICROSOFT EXCEL ACTUALIZADO PARA OFFICE 2000	2.261
9	2758	EXCEL 4 PARA WINDOWS	2.261
10	2759	EXCEL 5 PARA WINDOWS	2.261

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Los resultados en este proceso véase Tabla.- 33 demuestran una buena correlación de la información brindada del el motor de recomendaciones, procesando resultados que mediante el análisis visual se puede determinar relación directa con la consulta del usuario.

Tabla.- 35 Muestra los resultados de la consulta dos coeficientes de Jaccard

N	ID	Documento	Jaccard
1	6547	TABLAS DINAMICAS EN EXCEL 2007	0.166
2	5138	MICROSOFT EXCEL Y MICROSOFT WORD PARA ABODADOS	0.165
3	2760	EXCEL ANALISIS DE DATOS EMPRESARIALES	0.162
4	2758	EXCEL 4 PARA WINDOWS	0.151
5	2759	EXCEL 5 PARA WINDOWS	0.151
6	2762	EXCEL APLICACIONES EN ALGEBRA, ESTADISTICA, PROBABILIDAD Y FISICA	0.151
7	89	Administracion de la empresa con Microsoft Excel	0.151
8	357	ANALISIS DE DATOS CON MICROSOFT EXCEL ACTUALIZADO PARA OFFICE 2000	0.151
9	4155	LAS MACROS EN MICROSOFT EXCEL 2010	0.151

10	5763	PROGRAMACION CON MICROSOFT EXCEL 2002 MACROS Y VISUAL BASIC PARA APLICACIONES	0.151

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Para este caso sucede algo similar véase la Tabla.- 35 para los resultados anteriores, esto es debido a gran coincidencia de palabras claves entre la consulta del usuario y los documentos.

4.1.15.1. Comparación de resultados consulta tres.

Tabla.- 36 Muestra una comparativa de la consulta uno con los métodos Producto Escalar, Coseno

N	ID	Producto Escalar	ID	Coseno	ID	Dice	ID	Jaccard
1	5146	**38.200**	5138	**0.294**	5146	**3.164**	6547	**0.166**
2	5138	**27.552**	5146	**0.288**	5138	**2.820**	5138	**0.165**
3	6547	**23.869**	6547	**0.288**	6547	**2.614**	2760	**0.162**
4	367	**22.503**	2760	**0.279**	2760	**2.412**	2758	**0.151**
5	368	**22.503**	2758	**0.263**	367	**2.409**	2759	**0.151**
6	2760	**20.910**	2759	**0.263**	368	**2.409**	2762	**0.151**
7	2758	**19.458**	89	**0.263**	89	**2.261**	89	**0.151**
8	2759	**19.458**	357	**0.263**	357	**2.261**	357	**0.151**
9	2762	**19.458**	2762	**0.263**	2758	**2.261**	4155	**0.151**
10	89	**19.458**	4155	**0.263**	2759	**2.261**	5763	**0.151**

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Como podemos observar en la Tabla.- 36 la posición de los documentos no cambia lo suficiente como para desestimar los resultados de cualquiera de los procesos realizados, siendo ya sea producto escalar, coseno, dice o jaccard procesos que aportan resultados susceptibles a ser recomendados.

4.2. Discusión

Mediante el análisis de las tablas de resultados se puede observar los diversos métodos utilizados en la realización del presente estudio de investigación, y tal como se muestra además de la aprobación de usuarios expertos es evidente que el método de Producto Escalar brinda mejores resultados.

Permitiendo realizar un análisis de los procesos de búsqueda de información se determina que el uso de los procesos de similitud brinda la posibilidad de determinar los mejores resultados como lo demuestra Manuel Blázquez Ochando en su trabajo expuesto, siguiendo y evaluando los distintos modelos.

Cada modelo de recomendación expuesto en los resultados muestra valores calculados de una serie de fórmulas que determinan que mientras mayor sea el valor más susceptible es el documento para ser recomendado a los estudiantes

Mediante los resultados expuestos se determina que este proyecto de investigación da lugar a un alto índice de certeza a los documentos obtenidos como resultados siendo así que este sistema puede ser usado para su ampliación y utilización en la Universidad técnica Estatal de Quevedo, y así brindar a los estudiantes recomendaciones acertadas de libros de la biblioteca de la institución.

CAPITULO V

CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones.

El trabajo desarrollado ha dado lugar a una propuesta de un sistema de recomendaciones de materiales bibliográficos para la Universidad Técnica Estatal de Quevedo, que inicialmente utiliza una base de datos de los documentos de materiales bibliográficos de la biblioteca de la universidad, el cual está elaborado como un prototipo que posee las características para un sistema completo, a continuación se explica las conclusiones alcanzadas.

- Se logró analizar la fuente de información brindada por la institución, además se determinó realizar cambios en la estructura de dicha información como se explica en la sección 4.1.1.1. logrando así su utilización en los procesos de búsqueda de información.

- Mediante el análisis y estudio de las técnicas y modelos dentro del área de los tipos de sistemas de recomendación, se encontró que una de las técnicas más utilizadas y que mejores resultados brinda en la actualidad es el modelo vectorial además de su adaptabilidad con los requerimientos de este proyecto más detalles en la sección 2.2.4 , y por lo tanto fue motivo de su implementación de los modelos antes mencionados para el proceso de obtención de los resultados esperados.

- Por medio de los resultados obtenidos por cada técnica usada, se pudo analizar, que dentro de un conjunto de diez respuestas y la ordenación de los mismos corresponden a una posible recomendación para alguna consulta en especial, dependiendo del análisis antes previsto de la información lingüística de los documentos de la colección obtenida.

- Realizando la comparación de cada proceso de obtención de resultados mediante los términos de la consulta, se puede concluir que los valores arrojados por los procesos de similitud del Producto Escalar, Coseno, Dice y Jaccard, determinan que cada uno muestra resultados congruentes entre sí, pero comparando los resultados de cada uno en las tres consultas, se puede sugerir la utilización del proceso de similitud del Producto Escalar ya que brinda resultados más acorde a los términos de la consulta además de la aceptación por parte de usuarios expertos.

5.2. Recomendaciones.

Mediante el análisis al realizar cada proceso en el presente proyecto de investigación, se presentan las siguientes recomendaciones del autor.

- Analizar las bases de datos de los documentos de la biblioteca y poblarlos con información acorde a cada documento mediante las palabras claves o stopwords, esto proporcionara mejores términos de filtrado para cada proceso de similitud.

- Completar con información relevante a cada unidad de aprendizaje, por ejemplo palabras claves como *"índices, tablas, numeración, formulas, graficas, base de datos"* para la materia de *"Ofimática"*, suficiente para brindar la extracción necesaria de información eficaz en los procesos de similitud.

- Ampliar el desarrollo del sistema creando procesos de aprendizajes, como por ejemplo recopilar los resultados de libros recomendados a otros grupos de estudiantes de otros periodos académicos y reorganizar esos resultados con los nuevos obtenidos.

- Realizar un análisis para la continuidad y finalización de un sistema completo que reúna las características suficientes ejecutar la emisión de las recomendaciones a los estudiantes de cada carrera de la Universidad.

- Analizar la posibilidad de desarrollar un sistema de procesos automáticos de envió de correos, en el cual se ejecute el prototipo expuesto en este proyecto y de esta manera enviar un numero definido de documentos obtenidos por cada unidad de aprendizaje, mediante esta herramienta a cada estudiante dependiendo de su perfil de estudios, de esta misma forma ejecutar un proceso de búsqueda de libros en la base de datos bibliográfica cada cierto tiempo para actualizar los resultados de los últimos libros insertados en dicho repositorio.

CAPITULO VI
BIBLIOGRAFÍA

6.1. Bibliografía

[1] D. Wiley., Connecting Learning Objects to Instructional Design Theory:A definition, a metaphor, and a taxonomy, Association for Instructional Technology, 2002.

[2] G. Huecas, «Slideshare,» 14 04 2010. [En línea]. Available: http://es.slideshare.net/ghuecas/filtros-colaborativos-y-sistemas-de-recomendacin. [Último acceso: 20 02 2015].

[3] Universidad Técnica Estatal de Quevedo, «Biblioteca General,» [En línea]. Available: http://biblioteca.uteq.edu.ec/biblioteca/BibliotecaUTEQ/. [Último acceso: 20 8 2015].

[4] L. g. T. a. J. t. R. Jonathan l. Herlocker Joseph a. Konstan, «Evaluating Collaborative Filtering Recommender Systems,» vol. 22, pp. 5-53, 2004.

[5] J. C. Gallardo, «Un nuevo modelo ponderado para Sistemas de Recomendación Basados en Contenido con medidas de contingencia y entropía,» JAÉN, 2012.

[6] S. M. Galán Nieto, «Filtrado Colaborativo y Sistemas de Recomendación,» Madrid, 2007.

[7] O. Velez-Langs y C. Santos, «Sistemas Recomendadores: Un enfoque desde los algoritmos genéticos,» *Industrias Data*, vol. 9, nº 1, p. 25, 6 2006.

[8] F. J. Martíınez Mimbrera, «Sistema de Recomendación Actualizable y con Gestión de usuarios,» Jaén, 2008.

[9] M. J. Barranco, L. Pérez y L. Martínez, «Un Sistema de Recomendación Basado en Conocimiento con Información Lingüística Multigranular,» Jaén, 2006.

[10] Wikipedia, «wikipedia,» 10 04 2013. [En línea]. Available: http://es.wikipedia.org/wiki/Filtrado_colaborativo. [Último acceso: 20 02 2015].

[11] D. J. Pinho Lucas, «Métodos de clasificación basados en asociación aplicados a sistemas de recomendación,» Salamanca, 2010.

[12] J. García García, J. López Puga, C. J. Cano Guillén, A. B. Gea Segura y Leticia de la Fuente Sánchez, «Aplicación de las redes bayesianas al modelado de las actitudes emprendedoras,» 2006.

[13] G. Carrillo y X. Ochoa, «Recomendación de objetos de aprendizaje basado en el perfil del usuario y la información de atención contextualizada,» Guayaquil, 2013.

[14] E. R. Núñez Valdéz, «Sistemas de Recomendación de Contenidos para Libros Inteligentes,» Oviedo, 2012.

[15] M. Seguido Font, «Sistemas de recomendación para webs de información sobre la salud,» Talalunya, 2009.

[16] C. M. González, «La recuperación de información en el siglo XX : Revisión y aplicación de aspectos de la lingüística cuantitativa y la modelización matemática de la información,» La Plata, 2008.

[17] E. Abadal y L. Codina, «Recuperación de Información,» Madrid, 2005.

[18] G. H. Tolosa y F. R. Bordignon, «Introducción a la Recuperación de Información Conceptos, modelos y algoritmos básicos,» Buenos Aires, 2008.

[19] M. Blázquez Ochando, «Técnicas avanzadas de recuperación de información: procesos,técnicas y métodos,» Madrid, 2013.

[20] A. F. Zazo Rodríguez, C. G. Figuerola Paniagua, J. L. Alonso Berrocal y R. Gómez Díaz, «Recuperación de información utilizando el modelo vectorial. Participación en el taller CLEF−2001,» Salamanca, 2002.

[21] J. A. M. Comeche, «Los modelos clásicos de Recuperación de información y su vigencia,» 2006.

[22] C. m. wijewickrema, «Una ontología basada en un sistema de clasificación de documentos totalmente automático utilizando un sistema semiautomático existente,» 2013.

[23] J. V. Ferro, «Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español,» Coruña, 2005.

[24] J. C. L. M. Manuel J. Barranco, «Evaluación de un método de ponderación de atributos multivaluados en sistemas de recomendación basados en contenido,» Jaén, 2012.

[25] P. A. Álvarez Carrillo, I. F. Vega López y E. Fernández González, «Ana´lisis Comparativo de las Medidas de Semejanza Aplicadas al Contenido de Documentos Web,» Sinaloa, 2007.

[26] L. C. A. B. J. L. R. V. d. A. E. R. García Figuerola, «El sistema de recuperación Karpanta: estudio de usuarios a través del archivo de registro,» de *El sistema de*

recuperación Karpanta: estudio de usuarios a través del archivo de registro, Zaragoza, 2004, p. 2.

[27] X. F. R. M. Johnny Javier Ávila Montalvo, «Sistema de Recomendacio´n de Contenido para TV Digital basado en Ontolog´ıas,» Cuenca, 2014.

[28] R. A. Española, 2005. [En línea]. Available: http://lema.rae.es/dpd/srv/search?id=emvx6ClOkD6qE3p6df.

[29] N. La Serna Palomino, U. Román Concha y O. Norberto, «Implementación de un Sistema de Recuperación de Información,» *Revista de investigación de Sistemas e Informática*, vol. 6, nº 1, pp. 3-4, 2009.

CAPITULO VII

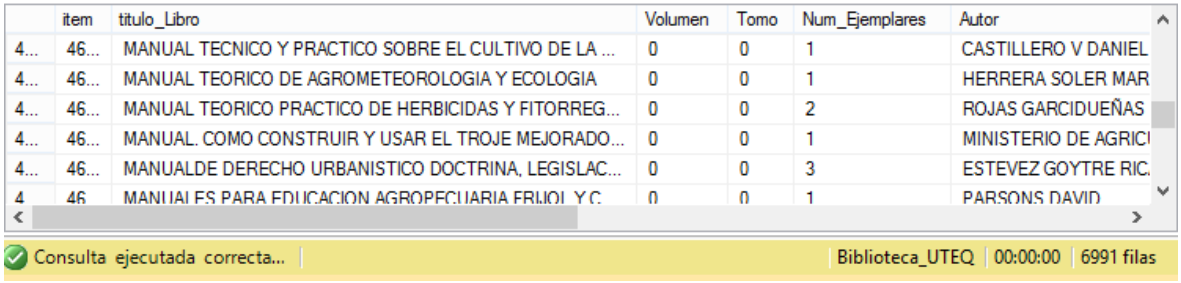
ANEXOS

7.1. Anexos.

En este apartado se muestra los procesos que estuvieron involucrados en la obtención de resultados, como se ha mencionado anteriormente en la sección 3.8.2 las herramientas utilizadas en el desarrollo de este proyecto de investigación, se realizó los procesos de indexación, normalización, cálculos, y procesos de búsqueda mediante sentencias sql ejecutadas mediante procedimientos almacenados. Estos resultados son visualizados en una aplicación de escritorio realizada en el lenguaje de programación C Sharp.

Mediante una colección de 6991 documentos donde constas todos los recursos bibliográficos de la Universidad.

Figura.- 40 Muestra la cantidad de recursos bibliográficos usados en este proyecto



	item	titulo_Libro	Volumen	Tomo	Num_Ejemplares	Autor
4...	46...	MANUAL TECNICO Y PRACTICO SOBRE EL CULTIVO DE LA ...	0	0	1	CASTILLERO V DANIEL
4...	46...	MANUAL TEORICO DE AGROMETEOROLOGIA Y ECOLOGIA	0	0	1	HERRERA SOLER MAR
4...	46...	MANUAL TEORICO PRACTICO DE HERBICIDAS Y FITORREG...	0	0	2	ROJAS GARCIDUEÑAS
4...	46...	MANUAL. COMO CONSTRUIR Y USAR EL TROJE MEJORADO...	0	0	1	MINISTERIO DE AGRICI
4...	46...	MANUALDE DERECHO URBANISTICO DOCTRINA, LEGISLAC...	0	0	3	ESTEVEZ GOYTRE RIC.
4	46	MANUALES PARA EDUCACION AGROPECUARIA FRIJOL Y C	0	0	1	PARSONS DAVID

Consulta ejecutada correcta... | Biblioteca_UTEQ | 00:00:00 | 6991 filas

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Realizando los pasos que fueron expuestos anteriormente, se realiza la distribución de los términos de la consulta.

Figura.- 41 Muestra el diccionario de términos

iden	Palabra
1	INGENIERA
2	SISTEMAS
3	COMUNICACION
4	INTERPERSONAL
5	TECNICA
6	FUNDAMENTOS
7	INFORMATICA
8	LOGICA
9	MATEMATICA
10	ECUACIONES
11	PROGRAMACION
12	TECNICAS
13	RECOPILACION
14	INFORMACION
15	INVESTIGACION
16	OFIMATICA

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

De igual manera se realiza el proceso anterior con la consulta del usuario o explicando de mejor manera la colección de información de su unidad de aprendizaje.

Figura.- 42 Muestra los términos de la consulta

1	INGENIERA
2	SISTEMAS
3	OFIMATICA
4	WORD
5	EXCEL
6	HOJAS
7	ACCESS
8	ADMINISTRACCION
9	INDICES
10	TABLAS
11	IMAGENES
12	FORMULAS
13	CALCULOS
14	DIAGRAMAS
15	GRAFICAS

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

En la base de datos de la colección de documentos se realiza el proceso de filtrado de acuerdo a los términos, los resultados de la Figura.- 42 que arroja la "query" muestran una colección filtrada del conjunto completo documentos, de esta manera mejoramos la ejecución de los procesos de las técnicas empleadas, ya que solo estarán interviniendo en información de contenido específico a la los términos de la consulta ignorando documento que no contienen dichos términos.

Figura.- 43 Muestra inicialmente la colección de documentos filtrados.

89	Administracion de la empresa con Microsoft Excel Administracion d...
104	ADMINISTRACION DE LOS SISTEMAS DE INFORMACION Len...
156	ADMINISTRACION DE SISTEMAS OPERATIVOS WINDOWS Y ...
357	ANALISIS DE DATOS CON MICROSOFT EXCEL ACTUALIZADO...
360	ANALISIS DE GENERO EN LA INVESTIGACION Y TRANSFERE...
367	ANALISIS DE LOS NEGOCIOS CON EXCEL XP ANALISIS DE L...
368	ANALISIS DE LOS NEGOCIOS CON EXCEL XP ANALISIS DE L...
374	ANALISIS DE SISTEMAS ELECTRICOS DE POTENCIA ELECTR...
375	ANALISIS DINAMICO DE SISTEMAS INDUSTRIALES Industria y...
408	ANALISIS Y DISEÑOS DE SISTEMAS DISEÑO Y ADMINISTRA...
454	APLICACION DE SISTEMAS MULTIAGENTE EN EL MODELAD...
557	AUDITORIA BASADA EN RIESGOS perspectiva estrategica de si...
562	AUDITORIA DE TECNOLOGIAS Y SISTEMAS DE INFORMACIO...

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

En el siguiente proceso realiza la búsqueda avanzada de cada término con respecto la colección de documentos, obteniendo el valor DF es decir la frecuencia del término en la colección.

Figura.- 44 Muestra parte del "query" en la ejecución del proceso DF

```
WHILE @u <= @termN
BEGIN
    SELECT @termino = p.Palabra from @TT_TERMINOS_PERFIL p WHERE p.iden = @u
    SELECT top 1 @coinci = count(*) from @tbdocument WHERE word LIKE '%'+@termino+'%'
    option (fast 1)
    IF @coinci > 0
    BEGIN
        UPDATE @TT_TERMINOS_PERFIL
        SET
            DF = DF +1 WHERE iden = @u
    END
    set @u = @u +1;
END
```

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Figura.- 45 Muestra valores calculados de DF de los términos de la consulta

Palabra	DF
INGENIERA	3
SISTEMAS	118
COMUNICACION	6
INTERPERSONAL	0
TECNICA	3
FUNDAMENTOS	13
INFORMATICA	13
LOGICA	4
MATEMATICA	9
ECUACIONES	0
PROGRAMACION	13

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Este proceso sirve para poder realizar los valores de ponderación TF-IDF de los términos de la consulta con fin de comprobar su relevancia de aquellos términos de la consulta con el diccionario de términos.

Figura.- 46 Muestra parte del "query" para el conteo de los terminos

```
while @countTermino <= @Nterminos
BEGIN
    SELECT @PalTermino = p.Palabra from @TT_TERMINOS_USUARIO p WHERE p.iden = @countTermino
    IF @CampoPalabra LIKE '%'+@PalTermino+'%'
    BEGIN
    UPDATE @TT_TERMINOS_PERFIL
        SET ConsultaUsuario = ConsultaUsuario + 1 WHERE iden = @countPalabra
    END
    SET @countTermino = @countTermino +1
END
```

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Figura.- 47 Muestra los resultados del conteo de los términos del diccionario con los terminos de la consulta

Palabra	DF	ConsultaUser
INGENIERA	3	1.000
SISTEMAS	118	1.000
COMUNICACION	6	0.000
INTERPERSONAL	0	0.000
TECNICA	3	0.000
FUNDAMENTOS	13	0.000
INFORMATICA	13	0.000
LOGICA	4	0.000
MATEMATICA	9	0.000
ECUACIONES	0	0.000
PROGRAMACION	13	0.000

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

En la Figura.- 47 se observa los primeros valores calculados que corresponde a los términos del usuario junto a los términos de la consulta.

Figura.- 48 Muestra parte del "script" para los cálculos de los valores TF-IDF

```
set @valorCampo = 0;
SELECT top 1 @valorCampo = p.ConsultaUsuario from @TT_TERMINOS_PERFIL p WHERE p.iden = @countPalabra
IF @valorCampo > 0 and @df > 0
BEGIN
    SET @ValueUser = log10(@Ndocumentos /  @df) + 1; -- valor IDF
        SET @ValueUser = @valorCampo * @ValueUser;
        UPDATE @TT_TERMINOS_PERFIL SET ConsultaUsuario = @ValueUser WHERE iden = @countPalabra
END
SELECT @valorCampo = p.ConsultaUsuario from @TT_TERMINOS_PERFIL p WHERE p.iden = @countPalabra
SET @SumatoriaUser = @SumatoriaUser + power(@valorCampo,2)
----------------------
SET @countPalabra = @countPalabra +1
SET @countTermino = 1
```

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Figura.- 49 Muestra los valores calculados de TF-IDF

Palabra	DF	ConsultaUser
EXCEL	34	1.745
HOJAS	0	1.000
ACCESS	1	3.276
ADMINISTRACC...	0	1.000
INDICES	1	3.276
TABLAS	15	2.100
IMAGENES	4	2.674
FORMULAS	22	1.934
CALCULOS	16	2.072

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

El mismo proceso se lleva a cabo para cada documento, separando cada palabra y realizando los cálculos correspondientes entre cada termino.

Figura.- 50 Muestra parte del "script" para los calculos de cada termino de la consulta en relacion con los documentos

```
if(@df != 0 AND @ValueDocu != 0)
BEGIN
    SET @ValueTFIDF = log10(@Ndocumentos / @df) + 1; -- valor IDF
    SET @ValueTFIDF = (@ValueDocu * @ValueTFIDF) * @ValueUser;
END
```

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG

Finalmente se muestra los resultados de los cuatro métodos para la recuperación de información que despliegan los descuentos posibles susceptibles a recomendación para las unidades de aprendizajes correspondientes a cada carrera de sus respectivas facultades de la Universidad Técnica Estatal de Quevedo.

Figura.- 51 Muestra resultados finales de las recomendaciones obtenidas por el sistema

codigoDoc	document	ProductoEscalar	Coseno	Dice	Jaccard
5146	MICROSOFT WORD 6 PARA WINDOWS PROGRAMACION DE CO...	38.200	0.288	3.164	0.137
5138	MICROSOFT EXCEL Y MICROSOFT WORD PARA ABODADOS Leng...	27.552	0.294	2.820	0.165
6547	TABLAS DINAMICAS EN EXCEL 2007 Informática. dispersion, circular,...	23.869	0.288	2.614	0.166
367	ANALISIS DE LOS NEGOCIOS CON EXCEL XP ANALISIS DE LOS N...	22.503	0.260	2.409	0.147
368	ANALISIS DE LOS NEGOCIOS CON EXCEL XP ANALISIS DE LOS N...	22.503	0.260	2.409	0.147
2760	EXCEL ANALISIS DE DATOS EMPRESARIALES Lenguajes y Sistema...	20.910	0.279	2.412	0.162
2758	EXCEL 4 PARA WINDOWS PROGRAMACION DE COMPUTADORA...	19.458	0.263	2.261	0.151
2759	EXCEL 5 PARA WINDOWS PROGRAMACION DE COMPUTADORA...	19.458	0.263	2.261	0.151
2762	EXCEL APLICACIONES EN ALGEBRA, ESTADISTICA, PROBABILID...	19.458	0.263	2.261	0.151
89	Administracion de la empresa con Microsoft Excel Administracion de la ...	19.458	0.263	2.261	0.151

FUENTE: UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

ELABORADO: JUSEH ROGER ALCIVAR CANSIONG