UNIVERSIDAD NACIONAL MICAELA BASTIDAS DE APURÍMAC

FACULTAD DE INGENIERÍA

ESCUELA ACADÉMICO PROFESIONAL DE INGENIERÍA INFORMÁTICA Y SISTEMAS

"RECONOCIMIENTO DE PATRONES EN IMÁGENES DIGITALES PARA LA BÚSQUEDA DE TEXTOS POR ÍNDICE DE CONTENIDOS DIGITALIZADOS, PARA EL SICGA DE LA BIBLIOTECA CENTRAL – UNAMBA, 2011"

TESIS

PRESENTADO POR:

BACHILLER YHON FUENTES HUAMÁN

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO INFORMÁTICO Y SISTEMAS

ABANCAY-APURÍMAC

2013

UNIVERSIDAD NACIONAL MICAELA BASTIDAS DE APURÍMAC

FACULTAD DE INGENIERÍA

ESCUELA ACADÉMICO PROFESIONAL DE INGENIERÍA INFORMÁTICA Y SISTEMAS

TESIS

"RECONOCIMIENTO DE PATRONES EN IMÁGENES DIGITALES PARA LA BÚSQUEDA DE TEXTOS POR ÍNDICE DE CONTENIDOS DIGITALIZADOS, PARA EL SICGA DE LA BIBLIOTECA CENTRAL – UNAMBA, 2011"

Presentado por el Bachiller **YHON FUENTES HUAMÁN** a la Escuela Académica Profesional de Ingeniería Informática y Sistemas, para optar el Título profesional de:

INGENIERO INFORMÁTICO Y SISTEMAS

Sustentado y aprobado ante el jurado integrado por:

Presidente:

Ing. Francisco Cari Incahuanaco

Primer miembro:

Ing. Marleny Peralta Ascue

Segundo miembro:

M.Sc. Ecler Mamani Vilca

Acto que dedico a:

- ✓ ***A dios*** *por su guía en mi vida.*
- ✓ ***A mis padres*** *por su apoyo y guía.*
- ✓ ***A la universidad nacional Micaela Bastidas de Apurímac*** *por albergarme en sus aulas durante 5 años y formarme un profesional digno y con conciencia social.*
- ✓ ***A la universidad nacional federicoVillareal – lima*** *por albergarme en sus aulas durante 1 año y enseñarme en aspectos de motivación y superación.*

ÍNDICE

Pág.

CAPÍTULO III

CAPÍTULO IV

ÍNDICE DE FIGURAS

Pág.

ÍNDICE DE TABLAS

Pág.

RESUMEN

Este informe de tesis se realizó con el objetivo de determinar el incremento del número de respuestas válidas en la búsqueda de textos por índice de contenido digitalizado, usando el algoritmo de reconocimiento de patrones en imágenes digitales, pero los índices de los libros se encontraban en imágenes digitalizados, para ello se desarrolló un algoritmo basado en Reconocimiento de Patrones en imágenes digitales utilizando el algoritmo de Reconocimiento Óptico de Caracteres OCR, para extraer los textos de las imágenes digitales que son los índices de los libros, entonces el motivo específico de la investigación nace de la idea de poder extraer el texto de las imágenes digitales que son los índices de los libros de la biblioteca central de nuestra Universidad Nacional Micaela Bastidas de Apurímac con la finalidad de guardar los textos extraídos en formato digital en la base de datos de la biblioteca central UNAMBA y compararlos con los temas buscados por el usuario así mostrarle a los usuarios que libros son más recomendados para su préstamo en el portal "Sistema Integral de Control, Gestión y Administración de Textos Bibliográficos-SICGA" ya que aún no existen software o aplicaciones orientas a la web que contemplen este tipo de herramientas, a su vez este tipo de búsqueda en el contenido de las imágenes digitales.

Se ha concluido que al usar el algoritmo de reconocimiento de patrones en imágenes digitales incorporando el algoritmo de reconocimiento óptico de caracteres OCR nos facilitó en el proceso de extracción del texto de las imágenes que son los índices de los libros y así incremento el número de respuestas validas en la búsqueda de texto por índice de contenidos digitalizados. Además se llegaron a comprobar los siguientes objetivos; se afirmó que el incremento promedio porcentual del número de respuesta válidas en la búsqueda de textos por índice de contenidos digitalizados, aplicando el

algoritmo de reconocimiento de patrones en imágenes digitales es superior a cero; llegando a un incremento máximo de 70.0% y un promedio de 25,5688% de respuestas válidas a temas buscados. tambien se llegó a comprobar que el porcentaje promedio de las palabras extraídas es superior al 98.8%, por lo tanto se afirmó que la extracción del total de caracteres de una imagen se extrajo un equivalente al 100% de los caracteres existentes en una imagen, las bondades a los que se llegó fue que la extracción de textos de los indicies digitales sólo se elabora por una única vez y se guardan en la base de datos estos textos extraídos, para evitar consumir recursos en el servidor, y así la rapidez en la búsqueda de un material bibliográfico se hace más eficiente, El algoritmo de reconocimiento de patrones en imágenes digitales es reutilizable en diferentes aplicaciones ya sea aplicaciones orientadas a la web o aplicación desktop y el uso de este algoritmo es estandarizado porque la extracción del texto de la imagen se desarrolló en un servicio web.

INTRODUCCIÓN

En esta tesis se presentan los resultados de la investigación sobre el proyecto denominado "**RECONOCIMIENTO DE PATRONES EN IMÁGENES DIGITALES PARA LA BÚSQUEDA DE TEXTOS POR ÍNDICE DE CONTENIDOS DIGITALIZADOS, PARA EL SICGA DE LA BIBLIOTECA CENTRAL- UNAMBA, 2011**". Con el objetivo de determinar el incremento de número de respuestas válidas en la búsqueda de textos por índice de contenido digitalizado, aplicando el algoritmo de reconocimiento de patrones en imágenes digitales a partir de este objetivo general, se determinó los siguientes objetivos específicos; determinar la cantidad de caracteres extraídos de una imagen usando el algoritmo de reconocimiento óptico de caracteres OCR y determinar el incremento del número de caracteres existentes en la base de datos aplicando el algoritmo de reconocimiento de patrones en imágenes digitales.

En esta parte de la investigación titulada Reconocimiento de Patrones en imágenes digitales para la búsqueda de textos por índice de contenidos digitalizados, para el SICGA de la biblioteca central – UNAMBA, 2011, se evaluó el incremento de respuestas válidas referentes al tema buscado que proporciona la aplicación web SICGA en el módulo de consultas al aplicar el algoritmo de reconocimiento de patrones en imágenes digitales a esta, las pruebas se realizaron ingresando temas de contenido del libros y así ubicar temas ingresados por el usuario en la base de datos de la biblioteca en la tabla índice previamente extraídos y guardados en la base de datos BDBiblioteca usando el módulo administrador de la aplicación SIGCA que se incorporó el algoritmo de reconocimiento de patrones en imágenes digitales el cual usa el algoritmo de OCR

para la extracción de los textos de las imágenes de estos índices e insertarlos en la base de datos estos textos extraídos en un campo índice de la tabla índice.

La información básica para desarrollar primero la vista lógica de funcionamiento del ingreso de los índices de un libro en la interfaz de registro del módulo portal administrador del proyecto SICGA se obtuvo mediante el análisis y requerimientos del área de procesos técnicos de la biblioteca central UNAMBA, haciendo uso de técnicas de recolección de información como entrevistas orales, recolección de documentos y formatos físicos de ingreso de registro de libros. Segundo la vista lógica de funcionamiento de módulo de búsqueda del portal consulta de proyecto SICGA también se elaboró de la misma manera.

El algoritmo de reconocimiento de patrones en imágenes digitales se elaboró en un servicio web para poder reutilizado en diferentes plataformas y diferentes lenguajes de programación, así poder integrar en diferentes aplicaciones sin importar el lenguajes de programación en el cual este desarrollada, la interfaz de búsqueda y registro de los índices de los libros es sólo una de las grandes ventajas que contempla este algoritmo.

Los índices de algunos libros se encontraron escaneados y guardados en imagen digital y sin ser usados.

Una de las limitantes más importantes fue que se tuvo que instalar correctamente el paquete de MICROSOFT OFFICE DOCUMENT IMAGING 2007 para la utilización de algunos de sus funcionalidades que proporciona este paquete.

Investigaciones sobre este tema son referentes a inteligencia artificial con los temas de reconocimiento de patrones esto implica utilizar tecnologías como son: redes neuronales, algoritmos genéticos u otro para grandes investigaciones científicas.

CAPÍTULO I

I. PLANTEAMIENTO DEL PROBLEMA

1.1. Descripción del Problema

La biblioteca central UNAMBA aun no contaba con un aplicativo que tenga en sus funciones internas algoritmos que tengan la capacidad de extraer los textos en una imagen y estos textos extraídos ser guardados en una base.

En la biblioteca central anteriormente por motivos de que no se podía extraer los caracteres de una imagen que estas son el índice de los libros, solo se podía hacer la búsqueda por título u autor de los libros más no por el contenido del libro.

En consecuencia de no tener muchos datos en la base de datos de la biblioteca central UNAMBA, no generaba muchas referencias con los temas buscados.

Actualmente los índices de los libros se encuentran escaneados y guardados en imagen digital y sin ser usados.

El proceso de búsqueda de un libro es repetitivo por un mismo usuario por la razón de que un usuario que busca un tema requerido en el libro físico y no encuentra el tema buscado entonces vuelve a realizar su consulta.

De seguir así estos índices en digital sin dar el uso entonces seguirá usando el texto físico que se encuentran en las estanterías, entonces el trabajo realizado de escaneo será en vano.

Si no se extrae el texto de estas imágenes que son los índices de los libros entonces el proceso de búsqueda seguirá siendo repetitivo y el proceso se demorara más aun por la misma razón de que cada año se aumenta el número

de usuario de la biblioteca entre alumnos, profesores, administradores y personas externas entonces se requerirá de más personal.

Si el proceso de consulta sigue siendo repetitivo y en consecuencia deficiente a la hora de realizar una búsqueda y realizar el préstamo se seguirá utilizando las fichas y mas utilitarios para estos procesos y tendrá que invertirse más en mas papeles que en realizar todo de forma automática.

1.2. Formulación del problema

1.2.1. Problema general

¿En cuánto incrementa el número de respuestas válidas en la búsqueda de textos por índice de contenido digitalizado, aplicando el algoritmo de reconocimiento de patrones en imágenes digitales UNAMBA 2011?

1.2.2. Problemas específicos

- ¿El algoritmo de reconocimiento óptico de caracteres OCR, extrae un número equivalente al total de los caracteres en una imagen?

- ¿En cuánto incrementa el número de caracteres en la base de datos aplicando el algoritmo de reconocimiento de patrones en imágenes digitales?

1.3. Justificación

El presente trabajo de investigación trata, sobre el desarrollo de algunos algoritmos que tengan la capacidad de extraer los caracteres de una imagen y estas puedan ser guardados por única vez en una base de datos.

En la aplicación web SICGA se aplicó algoritmos que ayudó a extraer los caracteres en una imágenes digitales que son los índices de cada unos de los libros, porque no existía software que extraiga los caracteres de una imagen y almacenarlas en la base de datos de la biblioteca central y así incrementar el número de palabras en la base de datos referentes a un libro en consecuencia el número de respuestas frente a un tema búsqueda se tendría como resultado más libros referentes a ese tema buscado.

La aplicación web SICGA que contiene ya el algoritmo de reconocimiento de patrones en imágenes digitales proporciona beneficios al encargado del área de circulación, específicamente al personal que realiza las tareas de registro le proporcionara mucha ayuda, además también tendrán mucho más beneficio los trabajadores de la biblioteca por que la extracción de caracteres y el guardado de los caracteres en forma digital todo se realizara en forma interna y automáticamente. También la aplicación proporciona la ventaja de mostrar los índices de los libros en PDF en su módulo de consultas.

La gran mayoría de las universidades grandes hoy en día el área de biblioteca se encuentra automatizado, porque es el centro de investigación de una casa de estudios universitarios, el mismo cambio global nos fuerza a utilizar herramientas de software que nos ayuden en los procesos de este tipo de área.

El impacto social que proporciona es mostrar el avance tecnológico que está teniendo nuestra universidad frente a las demás universidades y es la primera universidad a nivel regional que tiene una biblioteca automatizada con esta magnitud.

1.4. Limitaciones

La presente investigación abarca todo sobre desarrollo web, aplicando las tecnologías web como son aspx, javascript, c#, AJAX, servicios web y tecnologías netamente web, pero no se puede utilizar la interfaz de esta aplicación en una aplicación Desktop.

La presente investigación trae la funcionalidad de extraer imágenes más no la extracción de texto de archivos PDF pero si se modificaría algunos parámetros de código sería posible.

La presente investigación tendrá la funcionalidad correcta con la instalación correcta del paquete de MICROSOFT OFFICE DOCUMENT IMAGING 2007.

1.5. Objetivos

1.5.1. Objetivo general

Determinar el incremento del número de respuestas válidas en la búsqueda de textos por índice de contenido digitalizado, aplicando el algoritmo de reconocimiento de patrones en imágenes digitales.

1.5.2. Objetivos específicos

- Determinar el número de caracteres extraídos en una imagen, aplicando el algoritmo de reconocimiento óptico de caracteres OCR.
- Determinar el incremento del número de caracteres existentes en la base de datos aplicando el algoritmo de reconocimiento de patrones en imágenes digitales.
- Desarrollar el algoritmo de reconocimiento óptico de caracteres OCR.
- Desarrollar el algoritmo de reconocimiento de patrones en imágenes digitales.
- Desarrollar el algoritmo de búsqueda de textos por índice de contenidos digitalizados

CAPÍTULO II

II. MARCO TEÓRICO

2.1. Antecedentes

2.1.1. En el exterior

a. Algoritmo para reconocimiento de patrones y búsqueda de imágenes, TE-HSIU SUN, HORNG-CHYI HORNG, CHI-SHUAN LIU, FANG-CHIN TIEN, 2009.El método propuesto está basado en el algoritmo KRA con el que extraeremos la información que queremos procesar, La búsqueda de imágenes o información a partir de otras imágenes actualmente es una disciplina en pleno auge, grandes empresas tales como Google llevan años investigando acerca de ello. En este artículo nos centraremos en dada una imagen, buscar qué conjunto de imágenes la contienen, esto podría servir para encontrar objetos o personas dentro de paisajes, en videos de seguridad, etc. El método propuesto está basado en el algoritmo KRA con el que extraeremos la información que queremos procesar. Una vez extraídas las características principales estas serán invariantes respecto a la traslación, rotación y escalado. Para validar el método usaremos una imagen de referencia que sería la que buscaremos en una biblioteca de imágenes en las que puede aparecer rotada, trasladada, escalada o, incluso, no aparecer. El algoritmo indicará en cuáles de las imágenes de la biblioteca aparecen, con un porcentaje de confianza asociado.

b. **Reconocimiento de patrones en imágenes digitales de cromosomas, FERNÁNDEZ CASTRO, FERNANDO LUIS PALABRAS, 3-May-2010, METODO CIENTÍFICO**. La clasificación de cromosomas es parte de la citogenética clínica que tiene por objetivo el estudio de los cromosomas, su estructura y herencia. El tiempo de análisis que se necesita para la clasificación (4 a 5 días) representa un problema respecto a la gran inversión de tiempo. Además, el análisis detallado de cada cromosoma presente en un cariotipo induce a cometer errores de clasificación. Este trabajo propone el análisis de cromosomas en imágenes digitales que se obtendrán por medio de una cámara digital con una resolución de 3072x2304. Luego se realizara la clasificación de cromosomas tomando como primer paso el tratamiento de las imágenes con respecto a la rotación, traslación y escalación de un cromosoma para que pueda ser evaluado de la mejor manera posible. El segundo paso será la presentación de la imagen a la red neuronal construida según el modelo de Hopfield la cual realizara la clasificación del cromosoma presentado en la imagen.

2.1.2. En el Perú

Reconocimiento de Patrones en Imágenes Digital, DR. CESAR A. BELTRÁN CASTAÑÓN, 2006, Sociedad Peruana de Computación. Selección de los huevos no infectados por parásitos infectados de las gallinas. Estrictamente Sistemas para la búsqueda de textos

bibliográficos usando este tipo de herramientas No se conocen en el Perú antecedentes relacionados al proyecto de reconocimiento de patrones aplicados a documentos digitalizados.

2.2. Bases teóricas

2.2.1. Reconocimiento de patrones

Existen varios intentos para definir al reconocimiento de patrones, algunos son:

Reconocimiento de patrones llamado también lectura de patrones, identificación de figuras y reconocimiento de formas consiste en el reconocimiento de patrones de señales. Los patrones se obtienen a partir de los procesos de segmentación, extracción de características y descripción dónde cada objeto queda representado por una colección de descriptores. El sistema de reconocimiento debe asignar a cada objeto su categoría o clase (conjunto de entidades que comparten alguna característica que las diferencia del resto). Para poder reconocer los patrones se siguen los siguientes procesos:

- Adquisición de datos
- Extracción de características
- Toma de decisiones

El punto esencial del reconocimiento de patrones es la clasificación: se quiere clasificar una señal dependiendo de sus características. Señales, características y clases pueden ser de cualquiera forma, por ejemplo se puede clasificar imágenes digitales de letras en las clases «A» a «Z»

dependiendo de sus píxeles o se puede clasificar ruidos de cantos de los pájaros en clases de órdenes aviares dependiendo de las frecuencias [CASACUBERTA FRANCISCO ENRIQUE VIDA, 1998, Reconocimiento del Habla].

Lógica de funcionamiento de OCR

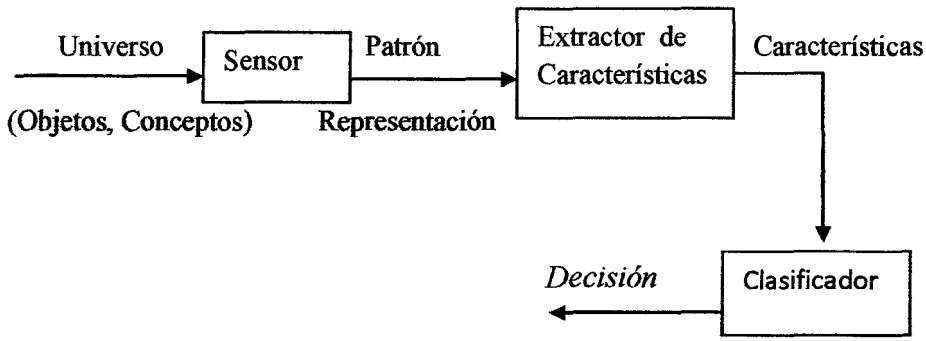


Figura Nº 1: diagrama de OCR

2.2.2. Imágenes digitales

Una imagen digital es una representación bidimensional de una imagen a partir de una matriz numérica, frecuentemente en binario (unos y ceros). Dependiendo de si la resolución de la imagen es estática o dinámica, puede tratarse de una imagen matricial (o mapa de bits) o de un gráfico vectorial. El mapa de bits es el formato más utilizado, aunque los gráficos vectoriales tienen uso amplio en la autoedición y en las artes gráficas [W. ROLSTON DAVID, 1998, Inteligencia Artificial].

2.2.3. Búsqueda de textos bibliográficos

Proceso de intentar encontrar patrones de respuestas validas en textos bibliográficos, este proceso puede de encontrar algo y ubicar ese algo puede realizarse en forma física o digital.

2.2.4. Biblioteca Central UNAMBA

La Biblioteca Central, se establece como un servicio esencial de apoyo a la comunidad universitaria y público en general en el fomento de la lectura a estudiantes, docentes y por ende la investigación, ofreciendo modernos servicios y una infraestructura favorable para el estudio.

Propicia la ciencia y la cultura a través del servicio de la información, procesando y difundiendo recursos bibliográficos y audiovisuales vinculados a todas las áreas del conocimiento y relacionadas con las escuelas académico profesionales que brinda la Universidad principalmente como apoyo a la formación profesional de los estudiantes.

En espera de que este servicio ofrecido sea en su beneficio y satisfacer algunas sugerencias y observaciones realizadas por Ustedes, nos comprometemos a continuar con este esfuerzo para el bien de la Comunidad Universitaria.

2.2.5. SICGA "Sistema Integral de Control Gestión y Administración de Textos Bibliográficos"

En una aplicación web que actualmente se encuentra en funcionamiento en la biblioteca central, esta fue evaluada por los responsables del área de biblioteca y resulto óptima en el proceso de evaluación, la aplicación

web denominada SICGA cubre las necesidades requeridas por la biblioteca central-UNAMBA con sus módulos que cuenta esta.

La aplicación ofrece portal web que tiene la capacidad de ser puesta en intranet y también en internet, lo cual está en evaluación, **anteriormente no contaba todavía con la capacidad de extracción de texto de los índices digitalizados de los libros de la biblioteca**, esta aplicación fue útil para incorporar el reconocimiento de patrones en imágenes digitalizados como medio de **interfaz**, las herramientas de hoy que ofrecen las aplicaciones web nos ayudan en el proceso de realizar una función específica en este caso la de búsqueda en el contenido digitalizado del índice de un tema requerido.

Además la aplicación SICGA tiene módulos de consulta de un libro, módulos del administrador que contiene paquetes de préstamos, devoluciones, registro de libros y usuarios el sistema en completo además contiene los paquetes de reportes y administración del Backup's de los datos de la base de datos de la biblioteca central UNAMBA. [Fuentes **Huamán Yhon y Pedro David Amao, Resolución Nº 505-2012-CU-COG-**UNAMBA]

2.2.6. Extracción de características

Es el proceso de generar características que puedan ser usadas en el proceso de clasificación de los datos. En ocasiones viene precedido por un preprocesado de la señal, necesario para corregir posibles deficiencias en los datos debido a errores del sensor, o bien para preparar los datos de

cara a posteriores procesos en las etapas de extracción de características o clasificación.

Las características elementales están explícitamente presentes en los datos adquiridos y pueden ser pasados directamente a la etapa de clasificación. Las características de alto orden son derivadas de las elementales y son generadas por manipulaciones o transformaciones en los datos [W. ROLSTON DAVID, 2001, Inteligencia Artificial y Sistemas Expertos].

2.2.7. Análisis de imágenes

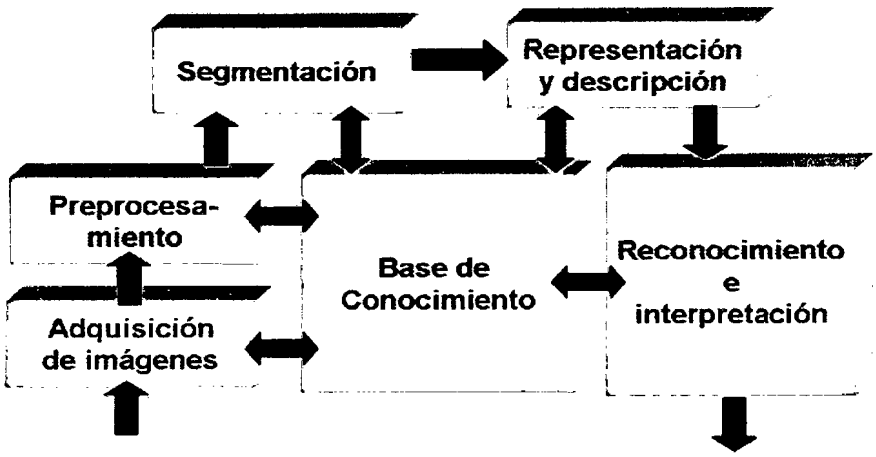


Figura Nº 2: Esquema de Análisis de imágenes [JOSÉ LUIS ALBA, 2006, texto español]

2.2.8. Reconocimiento de caracteres

Conjunto de técnicas informáticas cuyo objetivo es reconstituir los caracteres de un documento a partir de su propia imagen.

En la actualidad esta disciplina científica no sólo engloba la reconstrucción de caracteres, sino la estructuración de los documentos (títulos, subtítulos, bloques de texto, etc.) [ANTONIO BLASCO LÓPEZ, 2007, texto en español].

2.2.9. Reconocimiento óptico de caracteres (OCR)

El proceso básico que se lleva a cabo en el Reconocimiento Óptico de Caracteres es convertir el texto que aparece en una imagen en un archivo de texto que podrá ser editado y utilizado como tal por cualquier otro programa o aplicación que lo necesite.

Partiendo de una imagen perfecta, es decir, una imagen con sólo dos niveles de gris, el reconocimiento de estos caracteres se realizará básicamente comparándolos con unos patrones o plantillas que contienen todos los posibles caracteres. Ahora bien, las imágenes reales no son perfectas, por lo tanto el Reconocimiento Óptico de Caracteres se encuentra con varios problemas:

- El dispositivo que obtiene la imagen puede introducir niveles de grises al fondo que no pertenecen a la imagen original.
- La resolución de estos dispositivos puede introducir ruido en la imagen, afectando los píxeles que han de ser procesados.
- La distancia que separa a unos caracteres de otros, al no ser siempre la misma, puede producir errores de reconocimiento.

- La conexión de dos o más caracteres por píxeles comunes también puede producir errores.

2.2.10. Esquema básico de un algoritmo de reconocimiento óptico de caracteres

Todos los algoritmos de Reconocimiento Óptico de Caracteres tienen la finalidad de poder diferenciar un texto de una imagen cualquiera. Para hacerlo se basan en 4 etapas:

- Binarización,
- Fragmentación o segmentación de la imagen,
- Adelgazamiento de los componentes
- Comparación con patrones.

a. Binarización

La mayor parte de algoritmos de OCR parten como base de una imagen binaria (dos colores) por lo tanto es conveniente convertir una imagen de escala de grises, o una de color, en una imagen en blanco y negro, de tal forma que se preserven las propiedades esenciales de la imagen. Una forma de hacerlo es mediante el histograma de la imagen donde se muestra el número de pixeles para cada nivel de grises que aparece a la imagen. Para binarizarla tenemos que escoger un umbral adecuado, a partir del cual todos los pixeles que no lo superen se convertirán en negro y el resto en blanco.

Mediante este proceso obtenemos una imagen en blanco y negro donde quedan claramente marcados los contornos de los caracteres y símbolos que contiene la imagen. A partir de aquí podemos aislar las partes de la imagen que contienen texto (más transiciones entre blanco y negro).

b. Fragmentación o segmentación de la imagen

Este es el proceso más costoso y necesario para el posterior reconocimiento de caracteres. La segmentación de una imagen implica la detección mediante procedimientos de etiquetado determinista o estocástico de los contornos o regiones de la imagen, basándose en la información de intensidad o información espacial.

Permite la descomposición de un texto en diferentes entidades lógicas, que han de ser suficientemente invariables, para ser independientes del escritor, y suficientemente significativas para su reconocimiento.

No existe un método genérico para llevar a cabo esta segmentación de la imagen que sea lo suficientemente eficaz para el análisis de un texto. Aunque, las técnicas más utilizadas son variaciones de los métodos basados en proyecciones lineales.

Una de las técnicas más clásicas y simples para imágenes de niveles de grises consiste en la determinación de los modos o agrupamientos

("clusters") a partir del histograma, de tal forma que permitan una clasificación o umbralización de los pixeles en regiones homogéneas.

c. Adelgazamiento de las componentes

Una vez aisladas las componentes conexas de la imagen, se les tendrá que aplicar un proceso de adelgazamiento para cada una de ellas. Este procedimiento consiste en ir borrando sucesivamente los puntos de los contornos de cada componente de forma que se conserve su tipología.

La eliminación de los puntos ha de seguir un esquema de barridos sucesivos para que la imagen continúe teniendo las mismas proporciones que la original y así conseguir que no quede deforme.

Se tiene que hacer un barrido en paralelo, es decir, señalar los píxeles borrables para eliminarlos todos a la vez. Este proceso se lleva a cabo para hacer posible la clasificación y reconocimiento, simplificando la forma de las componentes.

d. Comparación con patrones

En esta etapa se comparan los caracteres obtenidos anteriormente con unos teóricos (patrones) almacenados en una base de datos. El buen funcionamiento del OCR se basa en gran medida a una buena definición de esta etapa. Existen diferentes métodos para llevar a cabo la comparación. Uno de ellos es el Método de Proyección, en el cual se obtienen proyecciones verticales y horizontales del carácter por

reconocer y se comparan con el alfabeto de caracteres posibles hasta encontrar la máxima coincidencia.

Existen otros métodos como por ejemplo: Métodos geométricos o estadísticos, Métodos estructurales, Métodos Neuro-miméticos, Métodos Markovianos o Métodos de Zadeh.

2.3. Definición de términos

- **Objeto**

  Es un concepto con el cual representamos los elementos sujetos a estudio. Pueden ser concretos o abstractos.

- **Patrón**

  Es sinónimo de objeto. En ocasiones se le llama así a los objetos ya clasificados.

- **Rasgo**

  Propiedad, factor, característica, etc. que se toma en cuenta para estudiar los objetos.

- **Reconocimiento**

  Proceso de clasificación de un objeto en una o más clases.

- **Filtración**

  Consiste en quitar información o datos indeseados de entrada. Dependiendo del uso, el algoritmo o método de filtrado cambia.

- **Un archivo DLL**

  Es una biblioteca que contiene el código y datos que se pueden utilizar por más de un programa a la vez. Por ejemplo, en sistemas operativos Windows, la DLL archivo Comdlg32 realiza común funciones relacionadas con el cuadro de diálogo. Por lo tanto, cada programa puede utilizar la funcionalidad contenida en este archivo DLL para implementar un cuadro de diálogo **Abrir**. Esto ayuda a promover la reutilización de código y el uso eficaz de la memoria.

  Mediante el uso de un archivo DLL, un programa puede ser modularizado en componentes separados. Por ejemplo, un programa de contabilidad podrá venderse por módulo. Cada módulo puede cargarse en el programa principal en tiempo de ejecución, si está instalado ese módulo. Dado que los módulos son independientes, el tiempo de carga del programa es más rápido y sólo se carga un módulo cuando se solicita esa funcionalidad.

  Además, las actualizaciones son más fáciles de aplicar a cada módulo sin afectar a otras partes del programa. Por ejemplo, puede que un programa de nóminas y los tipos impositivos cambian cada año. Cuando estos cambios se aíslan a un archivo DLL, puede aplicar una actualización sin necesidad de generar o vuelva a instalar todo el programa [Miguel Ángel Gómez Vicente, Manuel Martín Mohedano, **2005,** Introducción A C#].

- **Microsoft Office Document Imagen 2007-MODI**

  Microsoft Office Document Imaging (MODI) es un Microsoft Office que admita la edición de documentos escaneados por Microsoft Office

Document Scanning . Fue introducido por primera vez en Microsoft Office XP y se incluye en las versiones de Office posteriores, tales como Office 2007 . Ya no es disponible en Office 2010 . MODI permite a los usuarios Escanear documentos, reconocer imágenes con OCR , ver un documento digitalizado y Anotar los documentos escaneados incluyendo el uso de la tinta en un Tablet PC.

Mientras que el formato de archivo nativo de MODI parece ser MDI , MODI puede leer y escribir una pequeña variedad de TIFF archivos. También puede guardar mensajes de texto OCR en el archivo TIFF original. Sin embargo, MODI produce. Tif que violan la norma TIFF y pueden ser utilizados sólo por los productos de Microsoft Office Document Imaging . Imágenes en formato JPEG se pueden recuperar de estos archivos con los datos de talla de las herramientas de recuperación diseñadas para los archivos de desecho intactos a partir de imágenes de discos duros dañados, tales como lugar. El texto OCR en estos archivos se puede ver en un editor binario.

En su modo predeterminado, el motor de OCR que enderezar y reorientar la página cuando sea necesario. Si el objectname. Save () se llama al método que se guardarán las imágenes Endereza reorientó de nuevo en el archivo de imagen original.

CAPÍTULO III

III. HIPÓTESIS Y VARIABLES

3.1. Hipótesis general

Aplicando el algoritmo de reconocimiento de patrones en imágenes digitales, se incrementa el número de respuestas válidas en la búsqueda de textos por índice de contenido digitalizado.

3.2. Hipótesis específica

- Aplicando el algoritmo de reconocimiento óptico de caracteres OCR, se extrae un número equivalente al total 100% de los caracteres de una imagen.
- Aplicando el algoritmo de reconocimiento de patrones en imágenes digitales, se incrementa el número de caracteres existentes en la base de datos.

3.3. Sistemas de variables

VARIABLES	DIMENSIÓN	INDICES	ESCALA DE MEDICIÓN
Independiente X Reconocimiento de patrones en imágenes Digitales	X1: Algoritmo de reconocimiento óptico de caracteres OCR para la extracción de caracteres	Número de caracteres extraídos	Número de caracteres
	X2: Algoritmo de Reconocimiento de Patrones en Imágenes Digitales	Número de caracteres existentes en la base de datos	Número de caracteres
Dependiente Y Búsqueda de textos Por índice de contenido digitalizado	Y1: Patrones de respuestas válidas Referentes al tema buscado	Número de libros válidos encontrados referentes al tema	Número libros encontrados

Tabla N° 1: Descripción de Sistema de Variable

CAPÍTULO IV

IV. DISEÑO METODOLÓGICO

4.1. Tipo y nivel de investigación

Tipo de Investigación : Aplicada

Nivel de Investigación : Relacional

4.2. Método y diseño de investigación

Método de Investigación : Experimental

Diseño de Investigación : Cuasi-Experimental

4.3. Población y muestra

a) Población

La población motivo de esta investigación está conformada por 139 libros registrados en la base de datos biblioteca que contiene sus índices escaneados de la biblioteca central- UNAMBA.

b) Muestra

Se toma una muestra de 16 libros; esta cantidad se halla mediante una toma de muestras estratificadas el cual se muestra en la tabla Nº 2, estos libros tienen el 100% de sus índices ya escaneados en la biblioteca central – UNAMBA.

4.4. Técnica e instrumentos de recolección de datos

- Observación
- **Cuadro de codificación de datos.**
- Trabajo en la biblioteca, se elaboró los procesos de escaneado de los índices de los libros en el área de procesos de automatización de la biblioteca central UNAMBA.

4.5. Procedimiento de recolección de datos

- Entrevista con el jefe de biblioteca para autorización de la investigación
- **Exploración y conteo del número de libros registrados en la base de datos** SQL Server en general
- Exploración y conteo del número de libros registrados en la base de datos **SQL Server por categoría de acuerdo a la calificación Dewey**
- Clasificación de los libros de acuerdo la clasificación DEWEY.

4.6. Plan de tratamiento de datos

4.6.1. Diseño estadístico para el objetivo general. Aplicando el algoritmo de reconocimiento de patrones en imágenes digitales, para incrementar el **número de respuestas validas en la búsqueda de textos por índice de** contenido digitalizado

$$OG = x_2 \rightarrow y_1$$

Donde:

$OG = x_2 \rightarrow y_1$: Objetivo General

X2: Cantidad de caracteres existentes en la base de datos aplicando el algoritmo de reconocimiento de patrones en imágenes digitales.

Y1: Número de libros validos encontrados referentes al tema buscado.

n: Muestra

$\bar{x} = \sum_{i=0}^{n} \frac{X_i}{n}$: Media

$\delta^2 = \frac{\sum_{i=0}^{N}(X_i{}^2 - n\bar{x}^2)}{n-1}$: Desviación estándar

a. Datos estadísticos, para el objetivo general

n=16

$\bar{x} = 25{,}5688$

$\delta = 26{,}1710$

Sea:

μ : Incremento porcentual de número de respuesta válidas en la búsqueda de textos por índice de contenidos digitalizados aplicando el algoritmo de reconocimiento de patrones en imágenes digitales.

b. Hipótesis estadística para el objetivo general. Aplicando el algoritmo de reconocimiento de patrones en imágenes digitales, **incrementa el número de respuestas válidas en la búsqueda de textos** por índice de contenido digitalizado.

H_o: $\mu = 0$ { Incremento porcentual de número de respuesta válidas **en la búsqueda de textos por índice de contenidos digitalizados** aplicando el algoritmo de reconocimiento de patrones en **imágenes digitales es igual a cero** }

H_1: $\mu > 0$ { Incremento porcentual de número de respuesta válidas en la búsqueda de textos por índice de contenidos digitalizados **aplicando el algoritmo de reconocimiento de patrones en** imágenes digitales es mayor a cero }

c. Nivel de Significancia $\alpha = 5\% \cong 0{,}05$

d. Estadístico

Como las muestras son pequeñas, n=m=16, n, m < 30 entonces se **usa la distribución t-student con "n-1" grados de libertad**

N: tamaño de la población

n, m: muestras

$\bar{x} = \sum_{i=0}^{n} \frac{X_i}{n}$: Media

$\delta^2 = \frac{\sum_{i=0}^{N}(X_i{}^2 - n\bar{x}^2)}{n-1}$: Desviación estándar

$T_c = \frac{\bar{x} - u_0}{\delta/\sqrt{n}}$

- Hallando el punto crítico

$p(T > t_0) = \propto$ con n-1 g-l

$1 - p(T > t_0) = \propto$

$1 - \propto/2 = p(T < t_0)$ con n-1 g-l

$1\text{-}0{,}05 = p(T < t_0)$ con 15 g-l

$0{,}95 = p(T < t_0)$ con 15 g-l

$0{,}95 = p(T < 1{,}753)$ con 15 g-l

- Calculado Tc

$$T_c = \frac{25{,}5688 - 0}{26{,}171/\sqrt{16}}$$

$$T_c = 3{,}91$$

e. Región Crítica

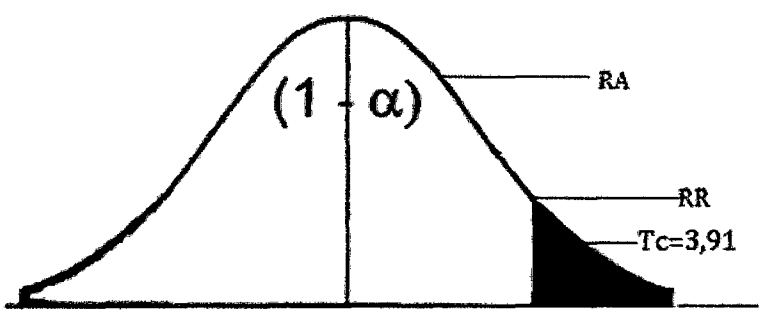
- Forma matemática de la hipótesis

$$OG = x_2 \rightarrow y_1$$

$$y_1 = cx_2 + d$$

$$y_1 = f(x_2)$$

- Gráfica de región crítica



u $t_0 = 1{,}753$ $+\infty$

$RC: < +t_0 ; +\infty >$; $RC: < +1{,}753; +\infty >$

f. Interpretación de resultados y decisión

Como $T_c = 3{,}91 \quad \in \quad < +1{,}753; +\infty >$ entonces se rechaza la hipótesis H_o.

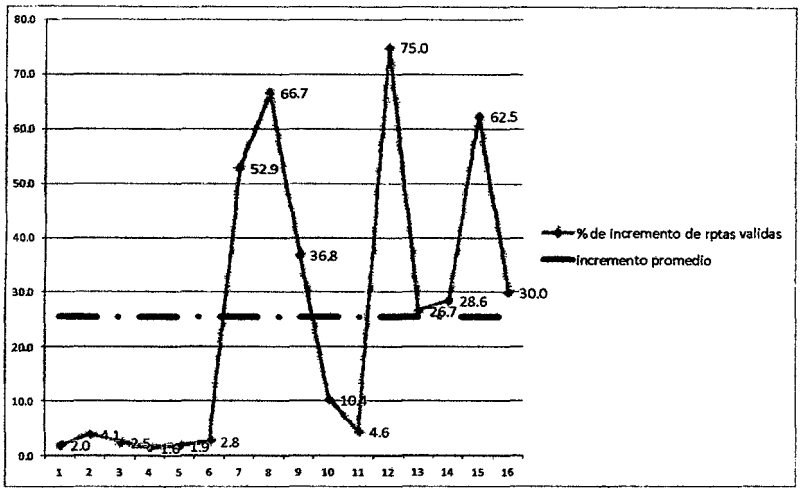


Figura N° 3: Histograma de extracción de los caracteres de una imagen

Como nos muestra en la Figura N° 3, el promedio de incremento de respuestas validas en la búsqueda de textos por índice de contenido digitalizado es de 25,5688 y como máximo arrojo un incremento de respuestas validas en la búsqueda de textos igual a 75,0, esto nos indica que existe un incremento en la búsqueda de textos por índice de contenido digitalizado.

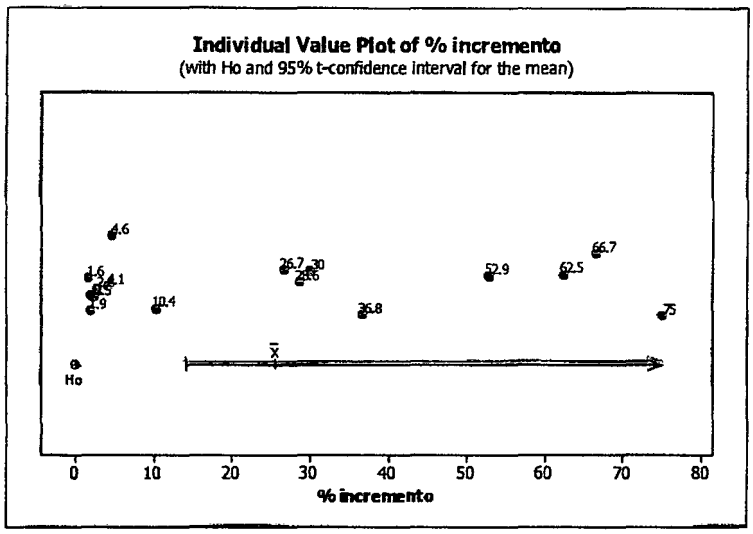


Figura Nº 4: Muestra de resultados en Minitab del promedio porcentual del incremento de respuestas validas.

Como en la *figura Nº 3 y 4,* se observa el "p-value" es 0.00 menor 0.05 nivel de significancia entonces se rechaza la hipótesis nula, por lo que podemos afirmar que el incremento promedio porcentual del número de respuesta validas en la búsqueda de textos por índice de contenidos digitalizados aplicando el algoritmo de reconocimiento de patrones en imágenes digitales es superior a cero; llegando a un incremento máximo de 75,0 y con un promedio porcentual de 25,5688

4.6.2. Diseño estadístico para la hipótesis específica 1. Aplicando el algoritmo de reconocimiento óptico de caracteres, para la extracción de una cantidad equivalente al del total de los caracteres de una imagen digital

$$\boldsymbol{OE_1 = I \rightarrow x_1}$$

Donde:

$OE_1 = I \rightarrow x_1$: Objetivo especifico 1

I: Matriz de pixeles de una imagen

x_1: Variable o indicador del objetivo específico 1

Además:

n, m : Muestras para el grupo A y B

$\delta_A{}^2 = \frac{\sum_{i=0}^{N}(X_i{}^2 - n\overline{x_A}^2)}{n-1}$: Desviación estándar para el grupo A

$\delta_B{}^2 = \frac{\sum_{i=0}^{N}(X_i{}^2 - n\overline{x_B}^2)}{m-1}$: Desviación estándar para el grupo B

$\overline{x_A}, \overline{x_B} = \sum_{i=0}^{n} \frac{X_i}{n}$: Media

$\delta_A{}^2 = \frac{\sum_{i=0}^{N}(X_i{}^2 - n\overline{x_A}^2)}{n-1}$: Desviación estándar para el grupo A

$\delta_B{}^2 = \frac{\sum_{i=0}^{N}(X_i{}^2 - n\overline{x_B}^2)}{m-1}$: Desviación estándar para el grupo B

a. **Datos Estadísticos, para el objetivo específico 1**

n=m=16

$\overline{x_A} = 1183$

$\overline{x_B} = 1170$

$\delta_A = 281$

$\delta_B = 278$

b. **Hipótesis para la hipótesis específica 1.** Aplicando el algoritmo de reconocimiento óptico de caracteres OCR, se extrae un número equivalente al total de los caracteres de una imagen

H_o: $\mu_{total} = \mu_{extra}$ (el promedio de caracteres que contiene una imagen es igual al promedio de caracteres extraídos por el OCR).

H_1: $\mu_{total} \neq \mu_{extra}$ (el promedio de caracteres que contiene una imagen es diferencia al promedio de caracteres extraídos por el algoritmo OCR).

c. **Nivel de Significancia α =5 % ≅0,05**

d. **Estadístico**

Como las muestras son pequeñas, n=m=16, n,m < 30 entonces se usa la distribución t-student con "n+m-2" grados de libertad

N: tamaño de la población

n, m: muestras

$\overline{x_A} = \sum_{i=0}^{n} \frac{X_i}{n}$: Media

$\delta_A{}^2 = \frac{\sum_{i=0}^{N}(X_i{}^2 - n\overline{x_A}^2)}{n-1}$: Desviación estándar para el grupo A

$\delta_B{}^2 = \frac{\sum_{i=0}^{N}(X_i{}^2 - n\overline{x_B}^2)}{m-1}$: Desviación estándar para el grupo B

$$T_c = \frac{\overline{x_A} - \overline{x_B}}{\sqrt{(n-1)\delta_A{}^2 - (m-1)\delta_B{}^2}} * \sqrt{\frac{nm(n+m-2)}{n+m}}$$

- Calculando el T_c

$$T_c = \frac{1183 - 1170}{\sqrt{(16-1)281^2 + (16-1)278^2}} * \sqrt{\frac{16(16)(16+16-2)}{16+16}}$$

$T_c = 0{,}1315$

- Desarrollo matemático para hallar valor del punto critico

$p(T > t_0) = \propto/2$ Con n+m-2 g-l

$1 - p(T > t_0) = \propto/2$

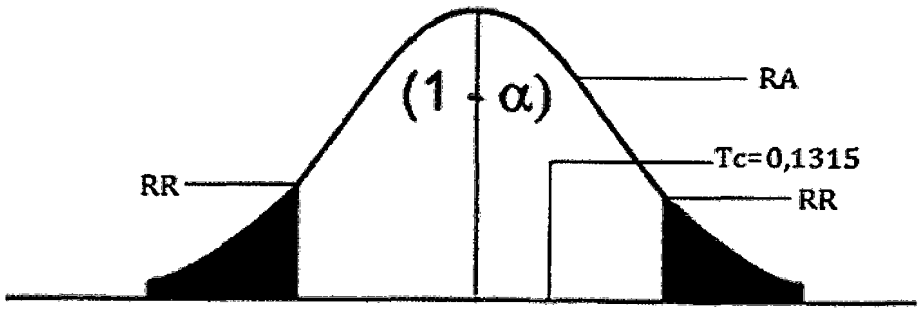
$1 - \propto/2 = p(T < t_0)$ Con n+m-2 g-l

1-0,05= $p(T < t_0)$ con 30 g-l

$0{,}975 = p(T < t_0)$ Con 30 g-l

$0{,}975 = p(T < 2{,}042)$ Con 30 g-l

e. Región critica



$$-\infty \quad -t_0 = -2{,}042 \quad u \quad tc \quad t_0 = 2{,}042 \quad +\infty$$

$$RC: <-\infty; -t_0> U <+t_0; +\infty>$$

f. Interpretación de resultados

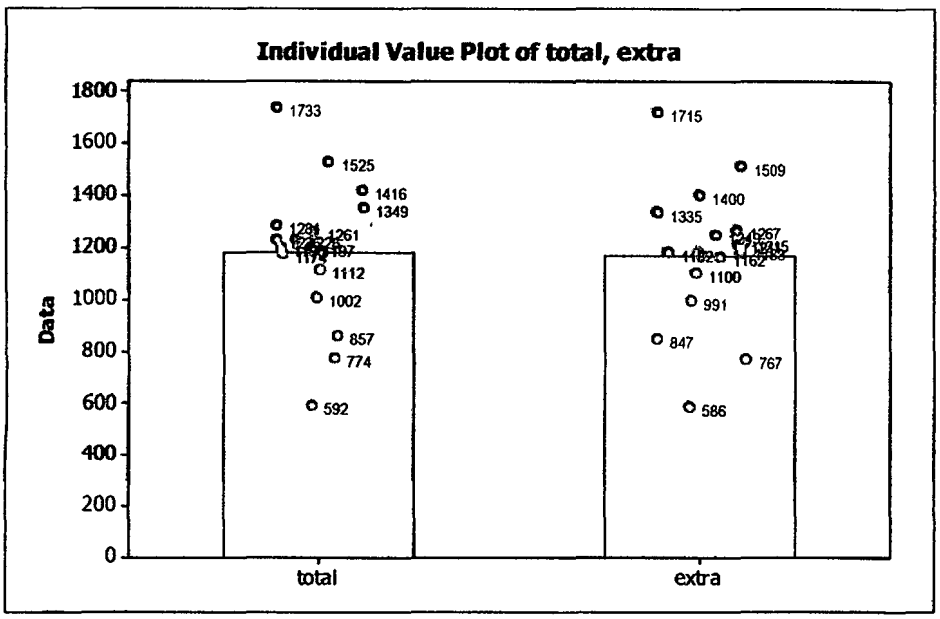


Figura N° 5: Histograma de verificación del promedio de extracción de caracteres en una muestra de 16 libros.

Como $T_c = 0{,}1315 \notin <-\infty; -2{,}042> U <+2{,}042; +\infty>$ entonces se acepta la hipótesis H_o esto nos índice que el promedio de caracteres que contiene una imagen es igual al promedio de caracteres extraídos por el OCR

Esto nos indica que aplicando del algoritmo de OCR para la extracción de los caracteres de una imagen, extrae el 100% de los caracteres de una imagen.

4.6.3. Diseño estadístico para la hipótesis especifica 2. Aplicando el algoritmo de reconocimiento de patrones en imágenes digitales, para incrementar el número de caracteres existentes en la base de datos

$$OE_2 = x_1 \rightarrow x_2$$

Donde:

$OE_2 = x_1 \rightarrow x_2$: Objetivo Específico 2

x_1: Número de caracteres extraídos aplicando el algoritmo de OCR.

x_2: Cantidad de caracteres existentes en la base de datos aplicando el algoritmo de reconocimiento de patrones en imágenes digitales.

a. Datos estadísticos para el objetivo especifico 2

$\overline{x_A} = 1228$

$\overline{x_B} = 57{,}9$

$\delta_A = 269$

$\delta_B = 27$

Sea:

μ_{final_bd}: Total de caracteres existentes en la base de datos aplicando el algoritmo de reconocimiento de patrones en imágenes digitales

μ_{inicio_bd}: Total de caracteres existentes en la base de datos sin aplicar el algoritmo de reconocimiento de patrones en imágenes digitales

b. Hipótesis para la hipótesis específica 2. Aplicando el algoritmo de reconocimiento de patrones en imágenes digitales, para incrementar el número de caracteres existentes en la base de datos

H_o: $\mu_{final\_bd} = \mu_{inicio\_bd}$

H_1: $\mu_{final\_bd} > \mu_{inicio\_bd}$

c. Nivel de Significancia $\alpha = 5\ \% \cong 0{,}05$

d. Estadístico

- Como las muestras son pequeñas, n=m=16, n,m < 30 entonces se usa la distribución t-student con "n+m-2" grados de libertad

N: tamaño de la población

n, m: muestras

$\overline{x_A} = \sum_{i=0}^{n} \frac{x_i}{n}$: Media

$\delta_A{}^2 = \frac{\sum_{i=0}^{N}(X_i{}^2 - n\overline{x_A}^2)}{n-1}$: Desviación estándar para el grupo A

$\delta_B{}^2 = \frac{\sum_{i=0}^{N}(X_i{}^2 - n\overline{x_B}^2)}{m-1}$: Desviación estándar para el grupo B

$$T_c = \frac{\overline{x_A} - \overline{x_B}}{\sqrt{(n-1)\delta_A{}^2 - (m-1)\delta_B{}^2}} * \sqrt{\frac{nm(n+m-2)}{n+m}}$$

- Desarrollo matemático para hallar valor del punto critico

$p(T > t_0) = \propto$ Con n+m-2 g-l

$1 - p(T > t_0) = \propto$

$1 - \propto/2 = p(T < t_0)$ Con n+m-2 g-l

1-0,05= $p(T < t_0)$ con 30 g-l

0,95 = $p(T < t_0)$ Con 30 g-l

0,95 = $p(T < 1{,}697)$ Con 30 g-l

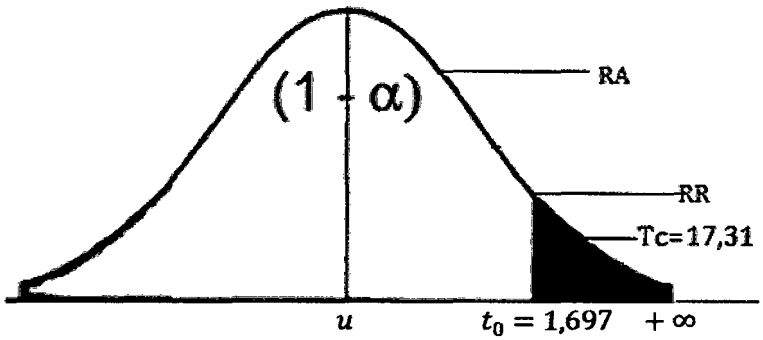
$$T_c = \frac{1228 - 57{,}9}{\sqrt{(16-1)269^2 + (16-1)27^2}} * \sqrt{\frac{16(16)(16+16-2)}{16+16}}$$

$T_c = 17{,}31226$

e. Región Crítica

Forma del objetivo especifico 2

$$OE_2 = x_1 \rightarrow x_2 ; x_2 = bx_1 + a; x_2 = f(x_1)$$



$RC: < +t_0 ; +\infty >;$ $RC: < +1,697; +\infty >$

f. Interpretación de resultados y discusiones

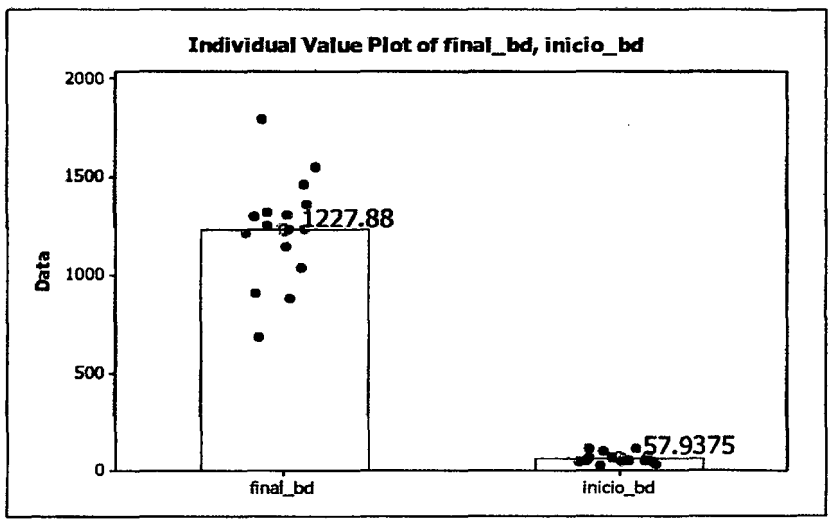


Figura N°6: Muestra de resultados de la diferencia de medias para determinar el objetivo 2

Como se observa en la figura N° 5, se tiene que "p-value" es 0.00 menor a 0.05 nivel de significancia, entonces se rechaza la hipótesis nula (Ho), por lo que podemos afirmar que ($\mu_{final\_bd} > \mu_{inicio\_bd}$) total de caracteres existentes en la base de datos aplicando el algoritmo de reconocimiento de patrones es mayor al total de caracteres existentes en la base de datos sin aplicar el algoritmo de reconocimiento de patrones en imágenes digitales.

CAPÍTULO V

V. RESULTADOS

5.1. DESARROLLO DE LOS ALGORITMOS

5.1.1. Desarrollo del algoritmo aplicando el algoritmo de reconocimiento óptico de caracteres OCR, para la extracción de los caracteres de una imagen

$$OE_1 = I \rightarrow x_1$$

a. Análisis lógico

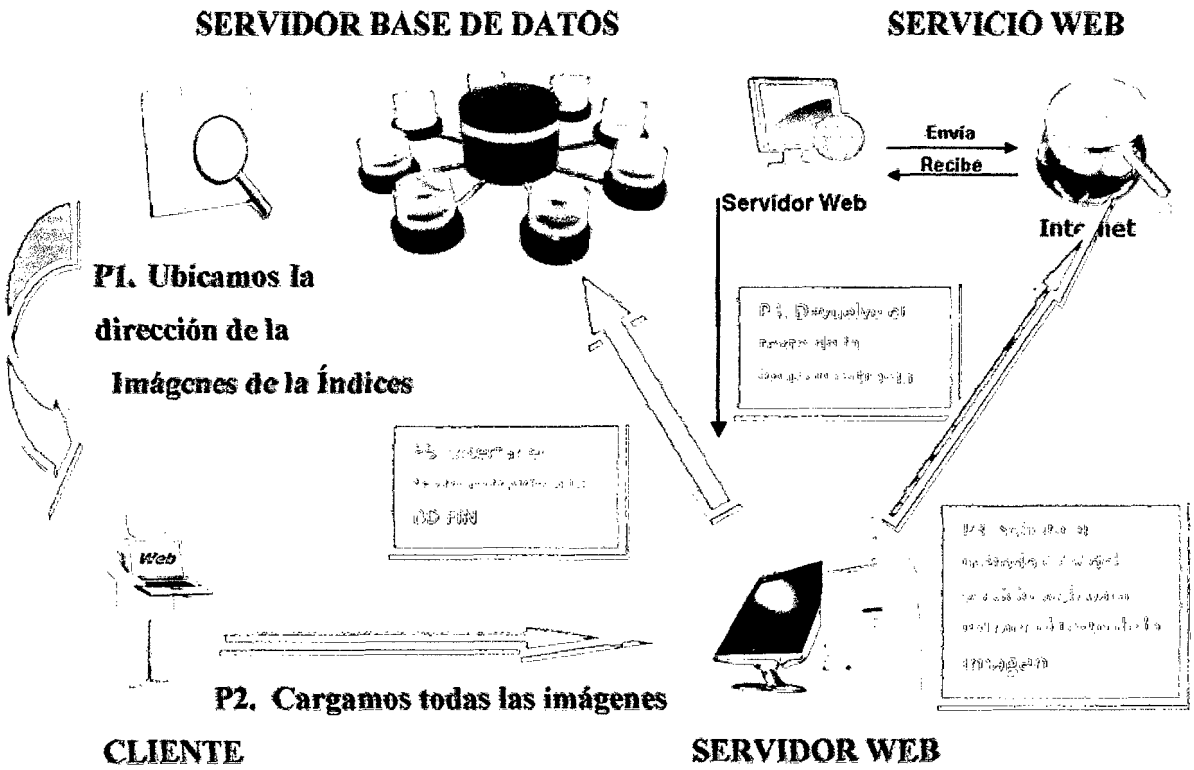


Figura N° 7: Análisis lógico de la extracción del número de caracteres extraídos de una imagen.

i. **Explicación de cada proceso**

P1→ Ubicamos la dirección de la Imágenes de la Índices.

- **Reconocemos la IP de la maquina**
- carga de los disco duros existentes en el Servidor web.

P2→cargamos todas la imágenes: una vez seleccionado la IP y el disco duro donde esta la carptea de las imágenes de los índices, seleccionamos la carpeta por descripcion principal las cuales seran insertardas en la base de datos una vez extraidas el texto de las imagnes de los índices del libro .

P3→ Solicita al método OCR del servicio web para extraer el texto de la imagen

P4→ Devuelve el texto de la imagen extraída y ese mismo texto lo inserta en base de datos

ii. **Proceso P3 extracción de caracteres de la imagen con OCR**

- binarización
- Fragmentación o segmentación de la imagen
- **Adelgazamiento de las componentes**
- comparación con patrones

Desarrollo de la 4 Etapas:

- **Binarización**

La binarización de una imagen consiste en un proceso de reducción de la información de la misma, en la que sólo persisten

dos valores: verdadero y falso. En una imagen digital, estos valores, verdadero y falso, pueden representarse por los valores 0 y 1, o más frecuentemente, por los colores negro (valor de gris 0) y blanco (valor de gris 255)

- ✓ Análisis histogramico de colores evaluación de la tonalidad RGB

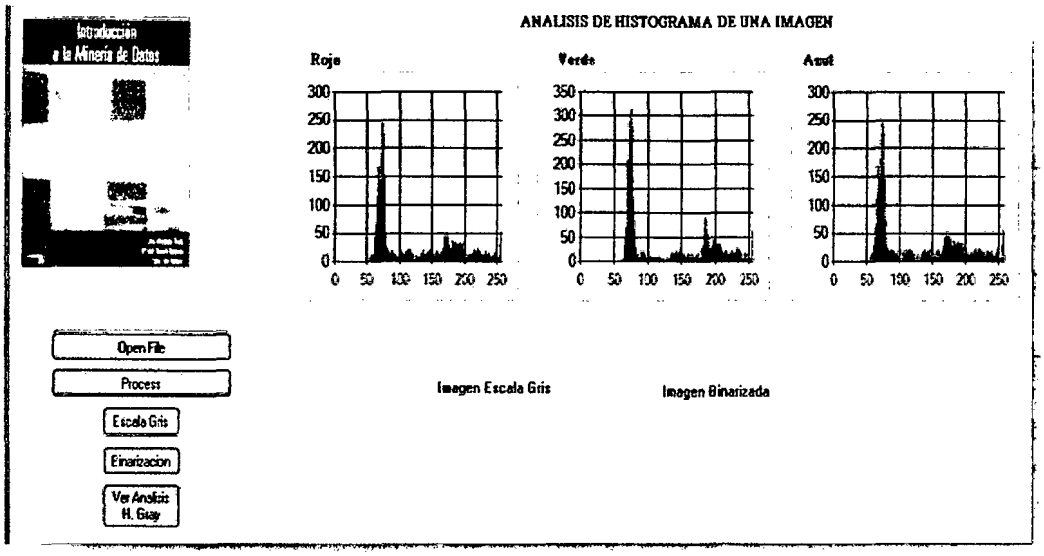


Figura Nº 8: Análisis histográmico de una imagen digitales.

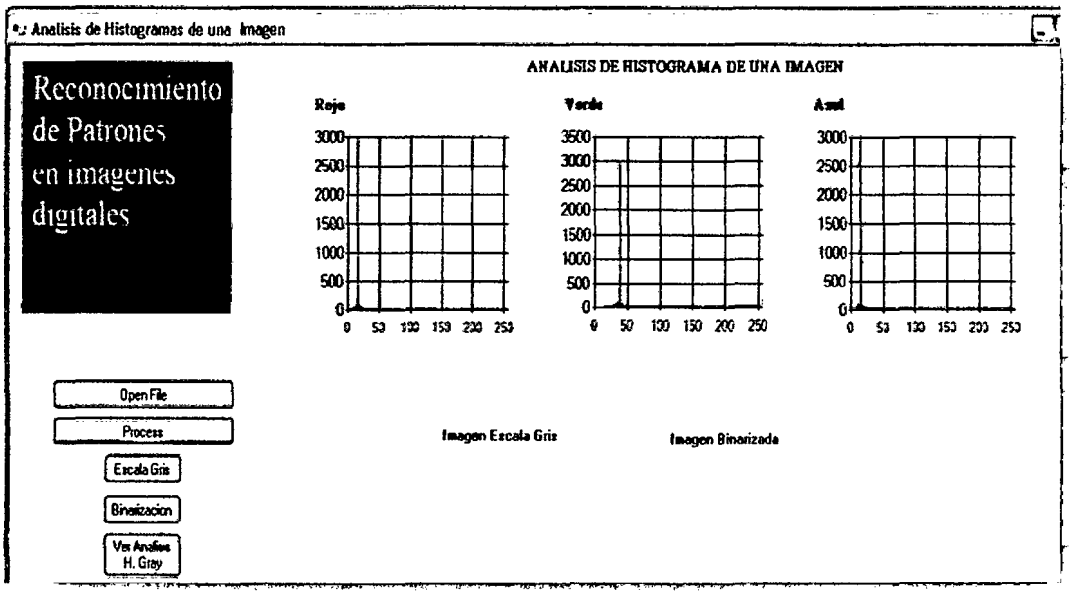


Figura Nº 9: Análisis histográmico de una imagen digitales (letras)

- ✓ Se observa en las figuras la conversión a escala de gris y binaria de la imagen digital procesada

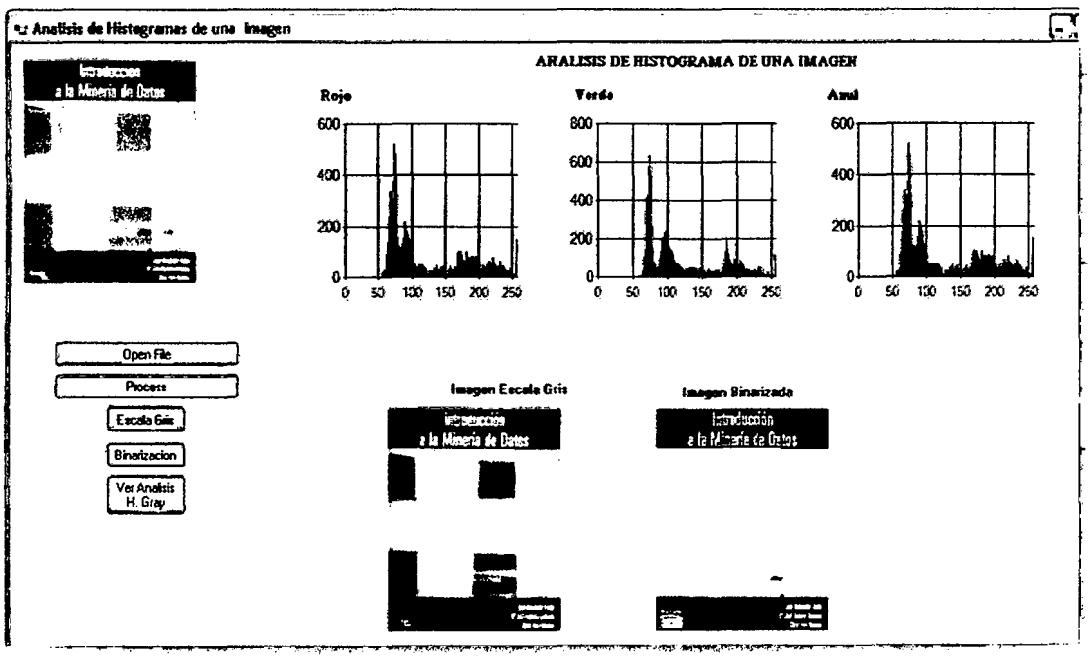


Figura Nº 10: Conversión a escala gris y binaria de una imagen digital

✓ Se observa en las figuras la conversión a escala de gris y binaria de la imagen digital procesada (letras)

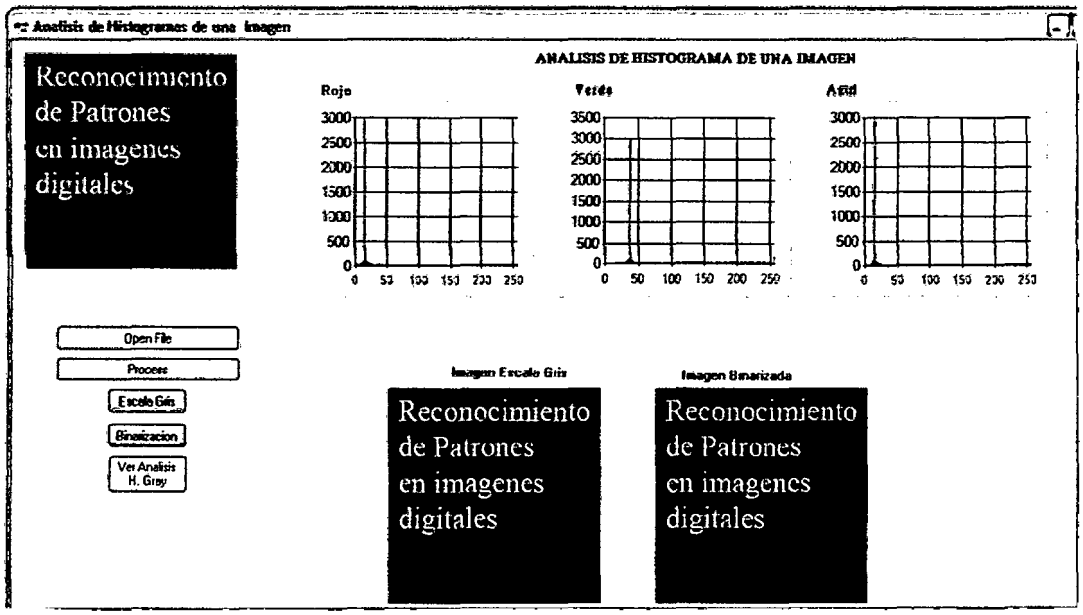


Figura N° 11: Conversión a escala gris y binaria de una imagen (letras)

- **Fragmentación o segmentación de la imagen**

  Mediante la segmentación vamos a dividir la imagen en las partes u objetos que la forman. El nivel al que se realiza esta subdivisión depende de la aplicación en particular, es decir, la segmentación terminará cuando se hayan detectado todos los objetos de interés para la aplicación.

- ✓ **Utilizaremos la técnica de histogramas**
  El histograma que se observa según la figura N° 12 nos indica el número de pixeles con los colores blanco y negro

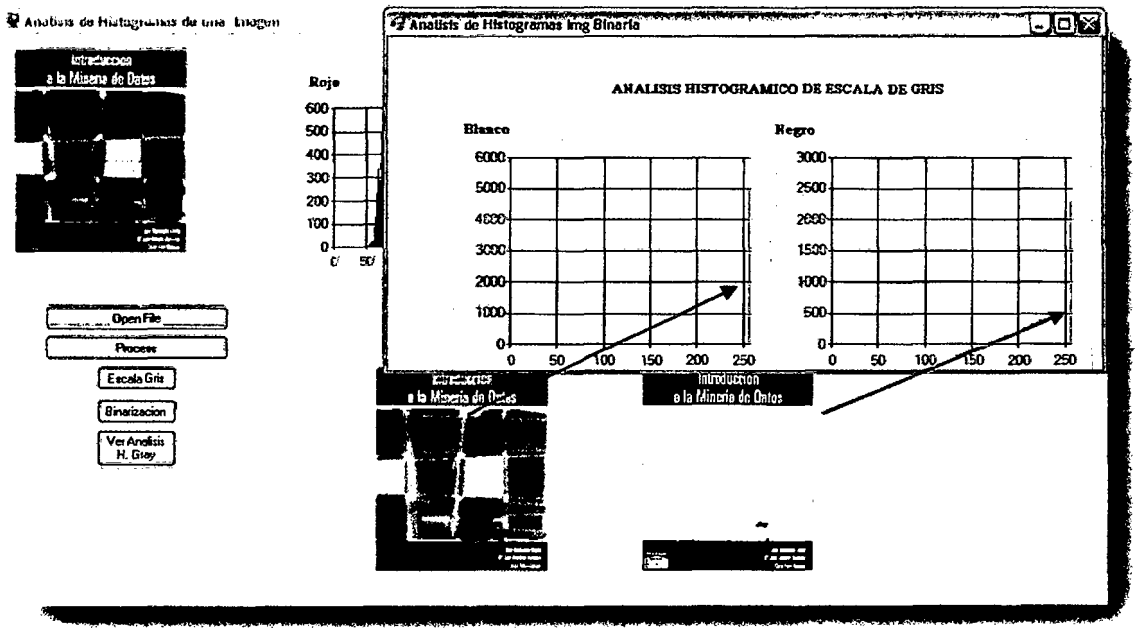


Figura N° 12: Histograma de binarización para la segmentación

✓ La imagen binaria la segmentaremos en submatrices de valores con ceros y unos para luego compararlas con patrones ya existentes

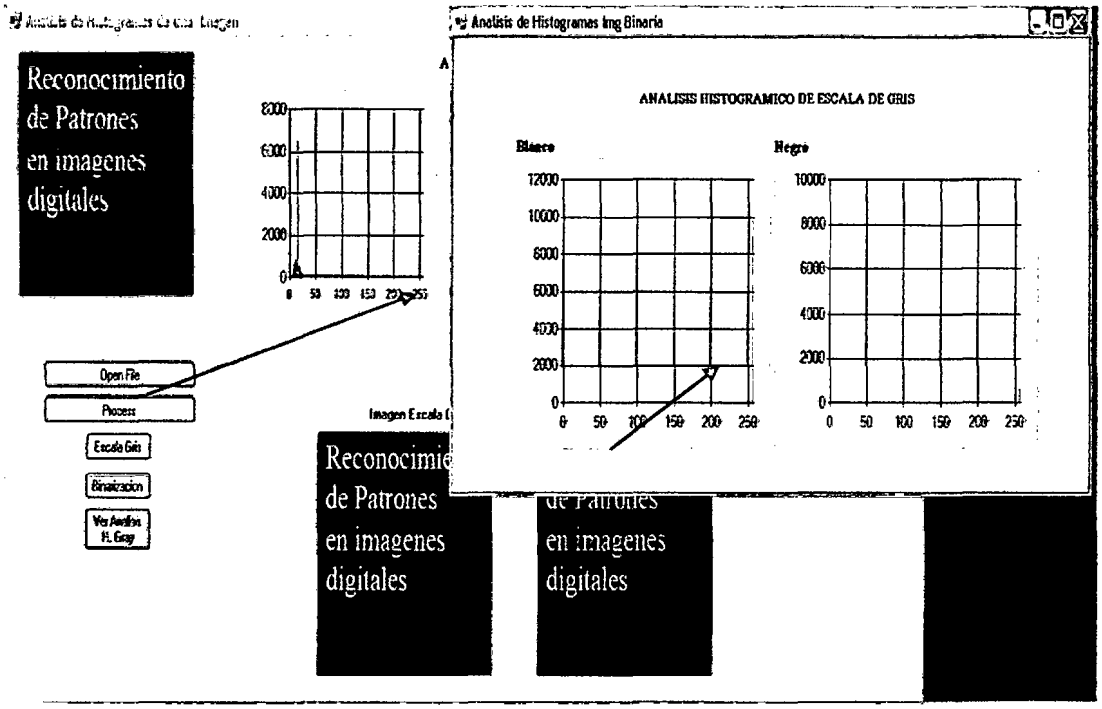


Figura N° 13: Histograma de binarización para la segmentación (letras)

✓ Proceso siguiente segmentar la imagen binaria convertida en trozos de submatrices que contengas los caracteres de la imagen

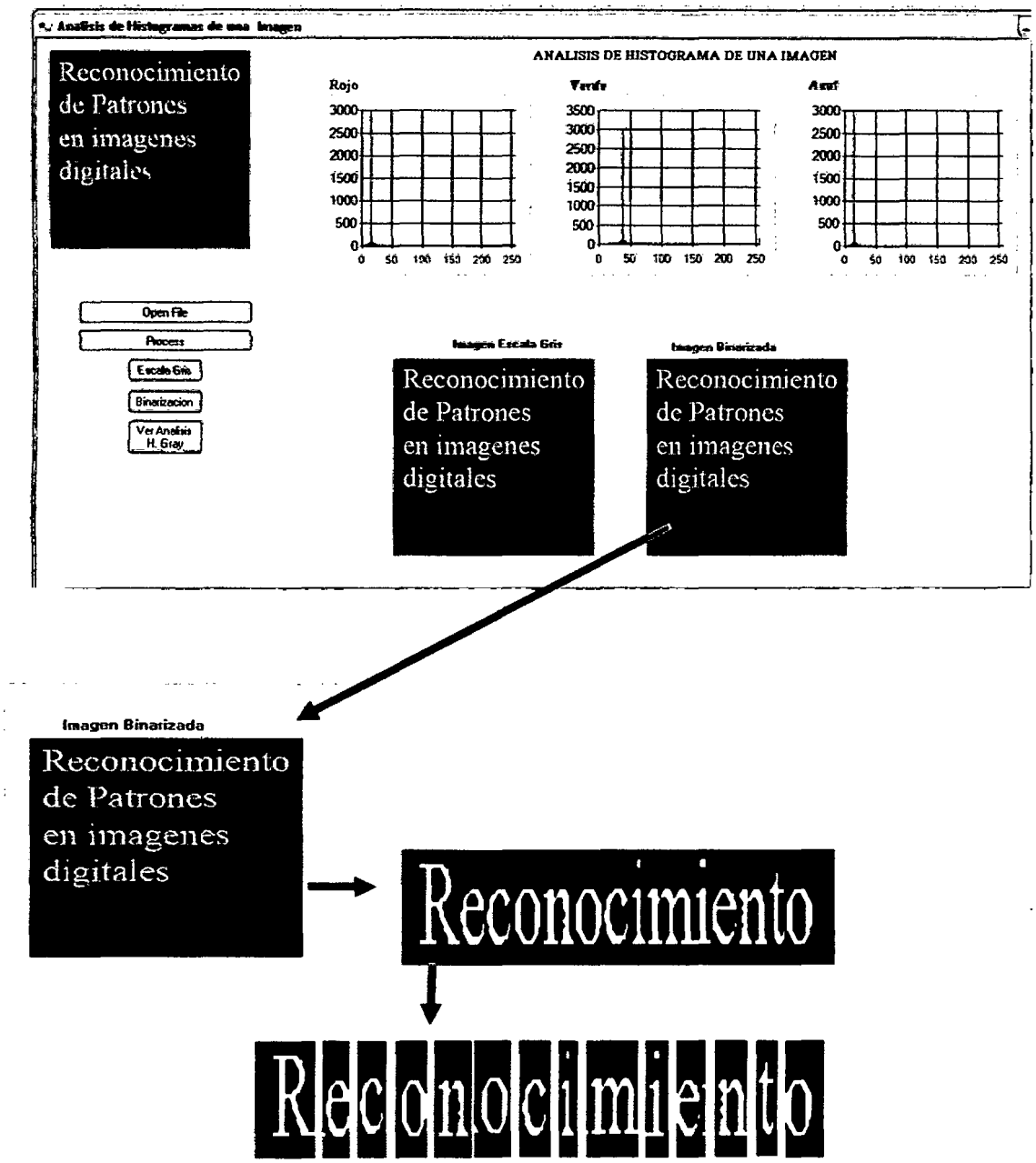
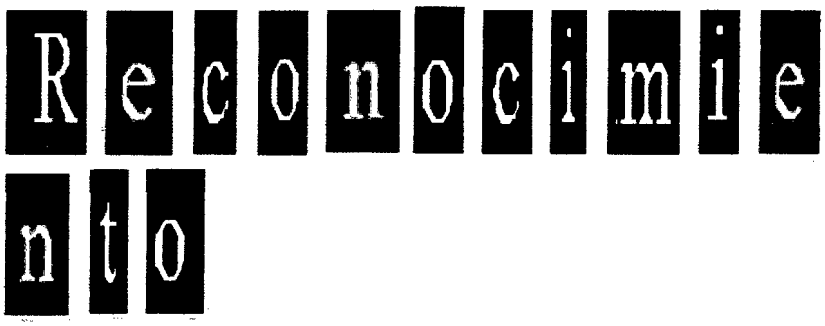


Figura Nº 14: Histograma de la segmentación de una imagen (letras)

- **Adelgazamiento de las componentes**

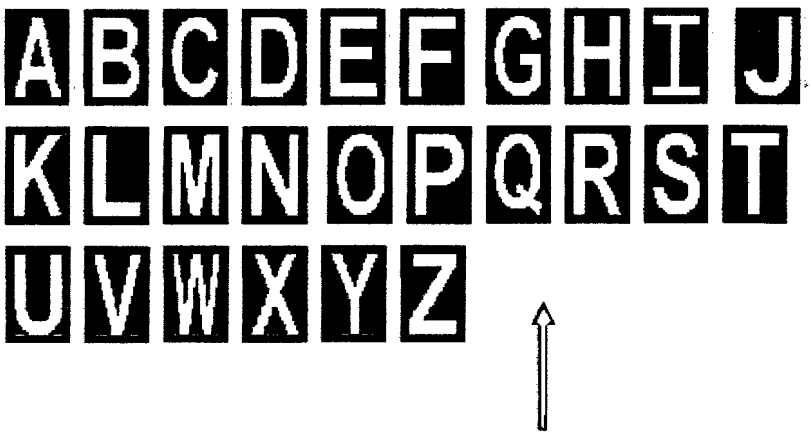
  Este procedimiento consiste en ir borrando sucesivamente los puntos de los contornos de cada componente de forma que se conserve su tipología. Este proceso se lleva a cabo para hacer posible la clasificación y reconocimiento, simplificando la forma de las componentes



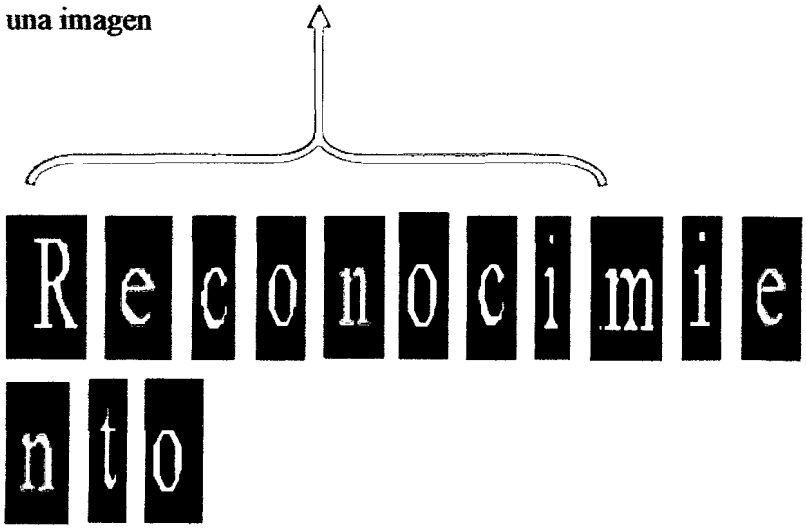
- **Comparación con patrones**

  Base de conocimiento de números

  Base de conocimiento de letras



Patrón de caracteres segmentados de la imagen binaria, listos hacer comparados con la base de conocimiento y extraidos como una imagen



Fin del proceso la extracción de los caracteres es:

La cadenas de caracteres extraídos es: {'R'

'e','c','o','n','o','c','i','m','i','e','n','t','o'}

b. Análisis Matemático

$OE_1 = I \rightarrow x_1$ Número de caracteres extraídos de una imagen usando el algoritmo de reconocimiento óptico de caracteres OCR.

Nos indica que tenemos que demostrar que:

Si tenemos que $CE \in I$ y $CA \in X_1$ **además que** $I \in CE$ **;** $A \in CA$ **y**

$I_y \in I$; $A_{x,y} \in A$ Además

$k_i \in I_y$ y $c_j \in A_{x,y}$ entonces $k_1 = c_1$, Quiere

Decir que $k_i \in CE$ y $c_1 \in CA \rightarrow k_i \in CA$..... (α)

Por tanto

$OE_1 = I \rightarrow x_1$ **seria correcto.**

Donde:

$OE_1 = I \rightarrow x_1$: Variable del objetivo específico 1

I y X_1 : Indicadores

I: matriz que contiene los pixeles de la imagen y además está matriz contiene submatrices que son caracteres de la imagen

CA: Conjunto que contiene los caracteres alfabéticos

CE: Conjunto de caracteres extraídos de la imagen

A: matrices generales que contienen las submatrices de las letras

I_y y $A_{x,y}$: son submatrices de la matrices I,A

k_i : Caracteres Extraídos de la imagen

c_j : Caracteres existentes para comparar

Demostrando sea:

I: sea la matriz de pixeles de una imagen cualquiera

Forma de la matriz M (I)

$$I=\begin{bmatrix} p_{11} & p_{12} & p_{13}\,.. & p_{1k} & \cdots & p_{1m} \\ p_{21} & p_{22} & p_{23} & p_{2k} & & p_{2m} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ p_{r1} & p_{r2} & p_{r3} & p_{rk} & & p_{rm} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ p_{n1} & p_{n2} & p_{n3}\,.. & p_{nk} & \cdots & p_{nm} \end{bmatrix}_{nxm}$$

Dónde:

$p_{i,j} = Es\ la\ representacion\ de\ cada\ pixel\ de\ la\ imagen$

$$Dim(I) = nxm; i = \overline{1,n}, j = \overline{1,m}$$

$$I = \{I_0, I_1, I_2, \ldots\ldots\ldots, I_k\}$$

Dónde: $I_i \in I; I_i = son\ submatrices\ de\ la\ matriz\ I$

$i = \overline{0,k}$ Además: $Dim(I_i) = n'xm'; i = \overline{1,n'}, j = \overline{1,m'}$

$$n', m' > 1 \ \ y\ m' \leq m\ y\ n' \leq n$$

Forma de la matriz M (I_i)

$$I_i = \begin{bmatrix} p'_{11} & p'_{12} & p'_{13}\ldots & p'_{1k} & \cdots & p'_{1m'} \\ p_{r21} & p'_{22} & p'_{23} & p'_{2k} & & p'_{2m'} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ p'_{r21} & p'_{r2} & p'_{r3} & p_{rk} & & p'_{rm'} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ p'_{n1} & p'_{n2} & p'_{n3}\ldots & p'_{nk} & \cdots & p'_{nm'} \end{bmatrix}_{n'xm'}$$

Dónde: $p'_{xy} \in I_i$

$x = \overline{1,n'}; y = \overline{1,m'}$ Además: $Dim(I_i) = n'xm'$ y

$x \leq n'y\ \ y \leq m'$ Entonces $k_i \in CE$

Entonces los caracteres extraídos son:

$$CE = \{k_1, k_2, k_3, \ldots\ldots k_{i,\ldots\ldots}, k_j \ \ldots\ldots k_x\} \ldots\ldots \textbf{(1)}$$

Sea:

A: **Matriz que contiene todo el conjunto de tipos de letras del alfabeto**

$A_{k,28}$ Submatrices que son los caracteres y diferentes tipos de letras además de caracteres numéricos que pertenecen a la matriz A

Entonces: $A_{k,28} \in A$

Donde: $A_{k,28}$ tienen las formas de la siguiente manera

- **Binarización**

$$A_{0,0} = \begin{bmatrix} 000000111000000 \\ 000001101100000 \\ 000110001100000 \\ 001111111111000 \\ 011000000001100 \\ 110000000000110 \end{bmatrix}$$

$$, \quad A_{0,1} = \begin{bmatrix} 111111110000000 \\ 110000011000000 \\ 111111100000000 \\ 110000011000000 \\ 110000011000000 \\ 111111100000000 \end{bmatrix},$$

$$A_{0,3} = \begin{bmatrix} 111111111100000 \\ 110000000000000 \\ 110000000000000 \\ 110000000000000 \\ 110000000000000 \\ 111111111100000 \end{bmatrix} \ldots\ldots A_{0,28} = \begin{bmatrix} 111111111100000 \\ 000001100000000 \\ 000110000000000 \\ 001100000000000 \\ 011000000000000 \\ 111111111100000 \end{bmatrix}$$

$$A_{1,0} = \begin{bmatrix} 111111111100000 \\ 000000011000000 \\ 111111111100000 \\ 110000001100000 \\ 110000001100000 \\ 111111111100000 \end{bmatrix}$$

$$, \quad A_{1,2} = \begin{bmatrix} 110000000000000 \\ 110000000000000 \\ 111111100000000 \\ 110000011000000 \\ 110000011000000 \\ 111111100000000 \end{bmatrix},$$

$$A_{1,3} = \begin{bmatrix} 011111111100000 \\ 110000000000000 \\ 110000000000000 \\ 110000000000000 \\ 110000000000000 \\ 011111111100000 \end{bmatrix} \ldots, A_{1,28} = \begin{bmatrix} 111111111111100 \\ 000000001100000 \\ 000000110000000 \\ 000011000000000 \\ 001100000000000 \\ 111111111111100 \end{bmatrix}$$

$$\vdots$$

$$\vdots$$

$$A_{k,28} = \begin{bmatrix} 1111111111100 \\ 1000001111000 \\ 0000011110000 \\ 0001111000000 \\ 0011110000 1000 \\ 1111111111000 \end{bmatrix}$$ Letra wide Latin

$$A_{k,r} = \begin{bmatrix} a_{11} & a_{12} & a_{13}\,.. & a_{1k} & \cdots & a_{1m'} \\ a_{21} & a_{22} & a_{23} & a_{2k} & & a_{2m'} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ a_{r1} & a_{r2} & a_{r3} & a_{rk} & & a_{rm'} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ a_{n'1} & a_{n'2} & p'_{n'3}\,.. & p'_{n'k} & \cdots & p'_{n'm'} \end{bmatrix}_{n'xm'}$$

Donde: k: es el número de tipos de letras del contiene las submatrices $A_{k,r}$

r: representa las letras del alfabeto

$i = \overline{1,k}$; j= $\overline{1,r}$; máximo de r=28

w= $\overline{1,9}$

- **<u>Comparación con patrones</u>**

Letra 'A' Arial

Sea $Dim(A_{0,0}) = n'xm'$, $i = \overline{1,n'}$; j= $\overline{1,m'}$ y

$a_{i,j} \in A_{0,0}$ Valores del vector $A_{0,0}$

Además $A_{0,0} \in A$ entonces

Demostramos que $C_1 = 'A' \leftrightarrow$ cumpla que:

Si: $i + j = n'$ entonces

$temp1 = \frac{m'}{2} + i$, $temp2 = \frac{n'}{2} + j$ y $temp3 = \frac{n'}{2} + 1$

$a_{i,j} = 1$

$a_{i,temp1} = 1$

$a_{temp2,temp2} = 1$

Entonces como $A_{0,0}$ forma la letra 'A' con los valores de las posiciones $a_{i,j} = 1$; $a_{i,temp1} = 1$; $a_{temp2,temp2} = 1$ y cumple con la restricciones anteriores por tanto $C_{1=}$ **'A'**

Letra 'E' Arial

Sea $Dim(A_{0,5}) = n'xm'$, $i = \overline{1,n'}$; j= $\overline{1,m'}$ y

$a_{i,j} \in A_{0,5}$ Valores del vector $A_{0,5}$

Además $A_{0,5} \in A$ entonces

Demostramos que C_2 = 'E ' ↔ cumpla que:

Si: $n' \geq 4$ entonces

$j > 1$ Entonces mientras $j \leq n'$ entonces

$a_{i,1} = 1$; $a_{1,j} = 1$; $k = \frac{n'}{2}$

$a_{k+1,j-1} = 1$; $a_{n,j} = 1$

Entonces como $A_{0,5}$ forma la letra 'E' con los valores de las posiciones

$a_{i,1} = 1$; $a_{1,j} = 1$; $k = \frac{n'}{2}$; $a_{k+1,j-1} = 1$; $a_{n',j} = 1$ y cumple con la restricciones anteriores por tanto $C_{2=}$ **'E'**

Letra 'I' Arial

Sea $Dim(A_{0,r}) = n'xm'$, $i = \overline{1,n'}$; j= $\overline{1,m'}$ y

$a_{i,j} \in A_{0,r}$ Valores del vector $A_{0,r}$

Además $A_{0,r} \in A$ entonces

Demostramos que C_r = 'I' ↔ cumpla que:

Si: $n' \geq 4$; $m' \geq 4$ y $j > 1$ y

Mientras $j < n'$ entonces

$a_{1,j} = 1$;

Si $i > 1$ mientras $i < n'$ entonces

$a_{i,1} = 1$; $a_{i,n'} = 1$

Entonces como $A_{0,5}$ forma la letra 'E' con los valores de las posiciones

$a_{1,j} = 1$; $a_{i,1} = 1$; $a_{i,n'} = 1$ y cumple con la restricciones anteriores por tanto $C_{r=}$ **'I'**

Letra 'O' Arial

Sea $Dim(A_{0,w}) = n'xm'$, $i = \overline{1,n'}$; j= $\overline{1,m'}$ y

$a_{i,j} \in A_{0,w}$ Valores del vector $A_{0,r}$

Además $A_{0,w} \in A$ entonces

Demostramos que $C_w = 'O' \leftrightarrow$ cumpla que:

Si: $n' \geq 4$; $m' \geq 4$ además que $j > 4$ entonces

Mientras $j < n'$ entonces

$a_{1,j} = 1$;

Mientras $i < n'$ entonces

$a_{i,1} = 1$; $a_{i,n'} = 1$

Entonces como $A_{0,w}$ forma la letra 'O' con los valores de las posiciones $a_{1,j} = 1$; $a_{i,1} = 1$; $a_{i,n'} = 1$ y cumple con la restricciones anteriores por tanto $C_{w=}$ **'O'**

Letra 'U' Arial

Sea $Dim(A_{0,z}) = n'xm'$, $i = \overline{1,n'}$; j= $\overline{1,m'}$ y

$a_{i,j} \in A_{0,z}$ Valores del vector $A_{0,z}$

Además $A_{0,z} \in A$ entonces

Demostramos que $C_z = 'U' \leftrightarrow$ cumpla que:

Si: $n' \geq 2$; $m' \geq 3$ además que si las posiciones

$a_{i,1} = 1$; $a_{i,m'} = 1$; $a_{n',j} = 1$;

Entonces como $A_{0,w}$ forma la letra 'U' con los valores de las posiciones

$a_{i,1} = 1$; $a_{i,m'} = 1$; $a_{n',j} = 1$; y cumple con la restricciones anteriores por tanto $C_{z=}$ 'U'

En general:

Sea $Dim(A_{x,y}) = n'xm'$, $i = \overline{1,n'}$; j= $\overline{1,m'}$ y

$a_{i,j} \in A_{x,y}$ Valores del vector $A_{x,y}$

Además $A_{x,y} \in A$ entonces

Demostramos que $C_y = '$alguna letra del alfabeto x$'$

$\leftrightarrow$ cumpla que los valores de $a_{i,j} = 1$ bajo las condiciones de condiciones de la letra $A_{x,y}$

Entonces como $A_{x,y}$ forma la letra ' letra del alfabeto x ' con los valores de las posiciones $a_{i,j} = 1$; y cumple con la restricciones anteriores por tanto $C_{y=}$ 'letra del alfabeto x '

Entonces el conjunto de caracteres alfabéticos tiene la forma de

$CA = \{c_1, c_2, c_3, \dots \dots c_{i,\dots\dots}, c_j \ \dots\dots c_y\}$ ..... (2) Valores estáticos y ya planteados bajo las restricciones dadas donde

$c_1 = ' \text{A}', c_2 = ' \text{B}', c_3 = ' \text{C}', c_4 = ' \text{D}', \dots \dots \dots, c_w = ' O', \dots \dots$

$\dots., c_y =$ ' alguna letra del alfabeto $A_{x,y}$ donde

$a_{i,j} = 1$ si $a_{i,j}$

$\in A_{x,y}$ que cumpla sus restricciones de acuerdo al tipo de letra

$CA =$

$\{'A', 'B', 'C', 'D', \dots \dots \dots 'Z', '1', '2', '4', '5', '6', \dots \dots . '9'\} OCA =$

$\{A_{0,0}, A_{0,1}, A_{0,2,} \dots \dots . A_{0,28,} . . , A_{k,0} , \dots . A_{k,28}, A_{w,0} , \dots . . A_{w,8}\}$

Donde

$A_{x,y}$: *son submatrice de la matriz A* y forman con sus valores $a_{i,j}$=1 una letra determinada bajo las restricciones de cada letra.

$w = \overline{1,9}$: indica el número de caracteres que representa los números enteros

$i = \overline{1,k}$: Índica el número de tipos de letras

Entonces demostraremos (α): $OE_1 = I \rightarrow x_1$

Si tenemos que $CE \in I$ y $CA \in X_1$ además que $I \in CE$; A $\in CA$ y

$I_y \in I$; $A_{x,y} \in A$ Además

$k_i \in I_y$ y $c_j \in A_{x,y}$ entonces demostrar :

Si $k_i \in CE$ y $c_j \in CA \rightarrow k_i \in CA$? $\leftrightarrow k_i = c_j$?... **(α)**

Sea $Dim(A_{x,y}) = n'xm'$, $i = \overline{1,n'}$; j= $\overline{1,m'}$ y

$a_{i,j} \in A_{x,y}$ Valores del vector $A_{x,y}$ Además $A_{x,y} \in A$ entonces

Forma con los valores $a_{i,j} = 1$ una letra del X del tipo Y bajo las condiciones del $c_j \rightarrow k_i = c_j$,

Quiere decir entonces que $k_i \in CA$ ***indica que es válida la ecuación (1) y Sustentamos que***

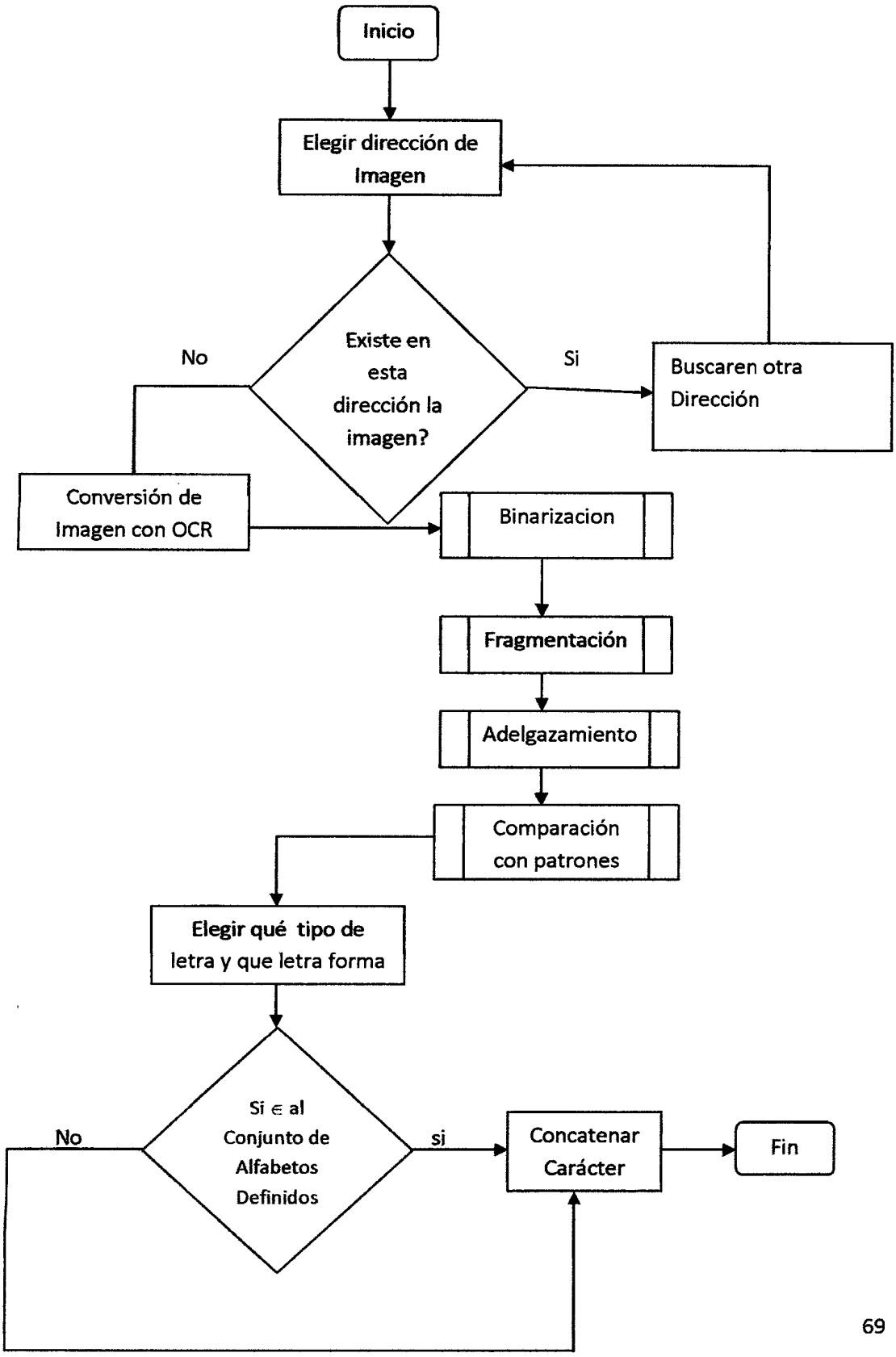
$OE_1 = I \rightarrow x_1$ ***es correcto, esto indica que se extrajo un número significativo de caracteres de cualquier imagen***

c. **Declaración de Variables**

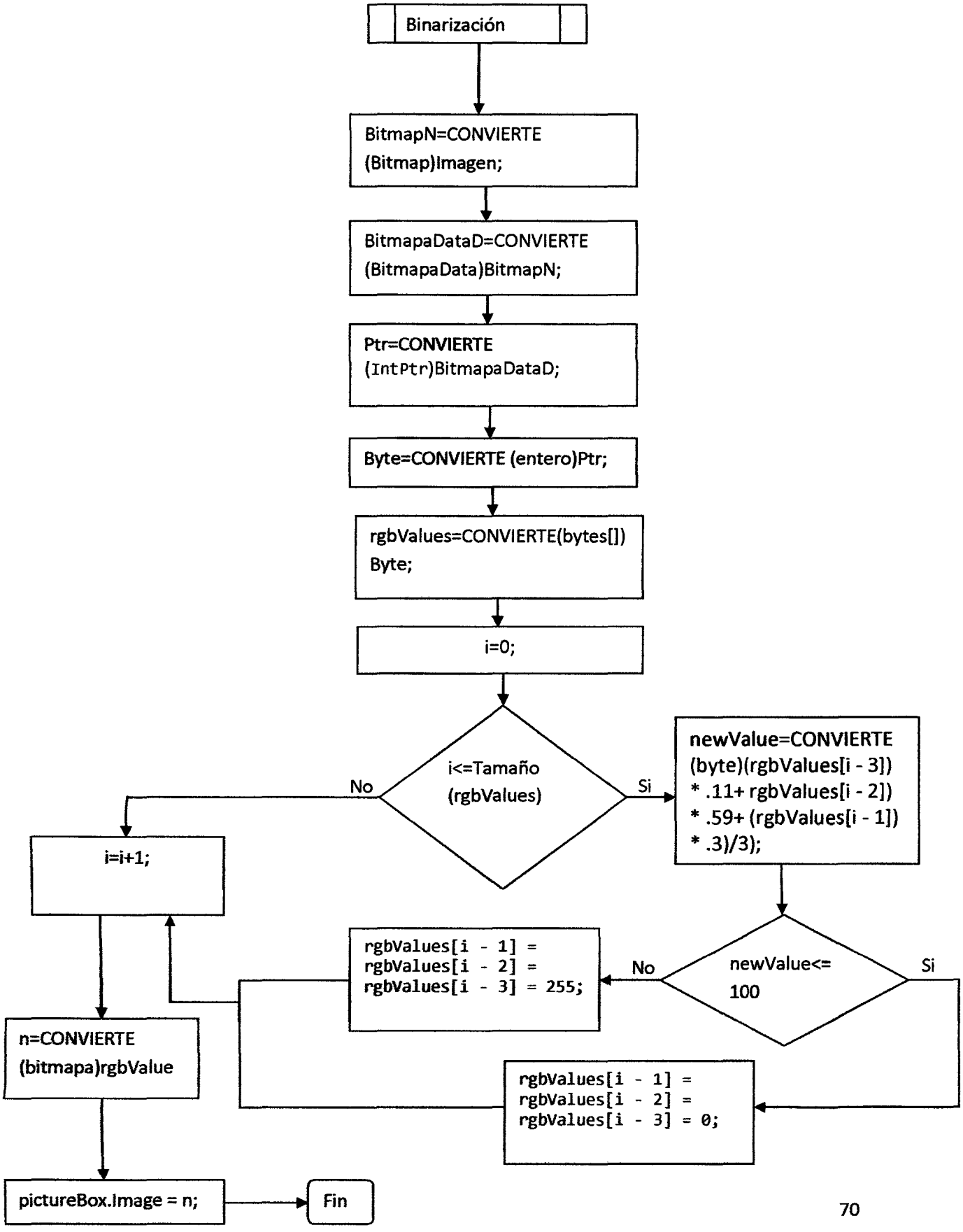
- Variables de Entrada
  - ✓ **Files: direccion de archivos de imagen**
  - ✓ Categoria: categoria a la que pertenece el libro de acuerdo a la clasificacion DEWEY
  - ✓ **Dir: direccion logica de la imagene donde esta guardado** la imagen, en que disco duro y en Ip de la maquina en donde esta
- **Variables de Salida**
  - ✓ Estado: Devuelve el texto extraido de la imagen

d. Diagrama de flujo

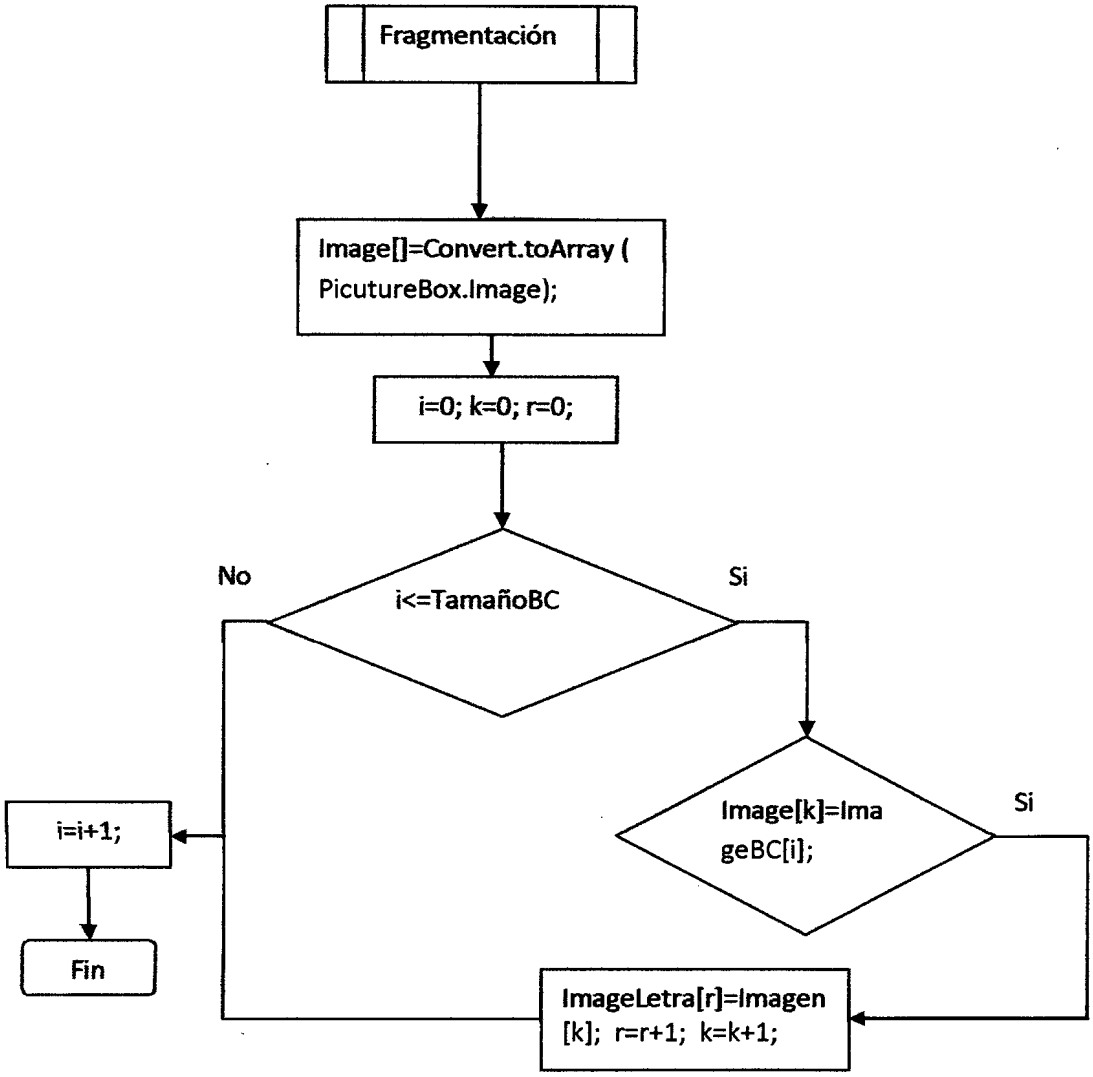
- **Diagrama de flujo del algoritmo del objetivo especifico 1**



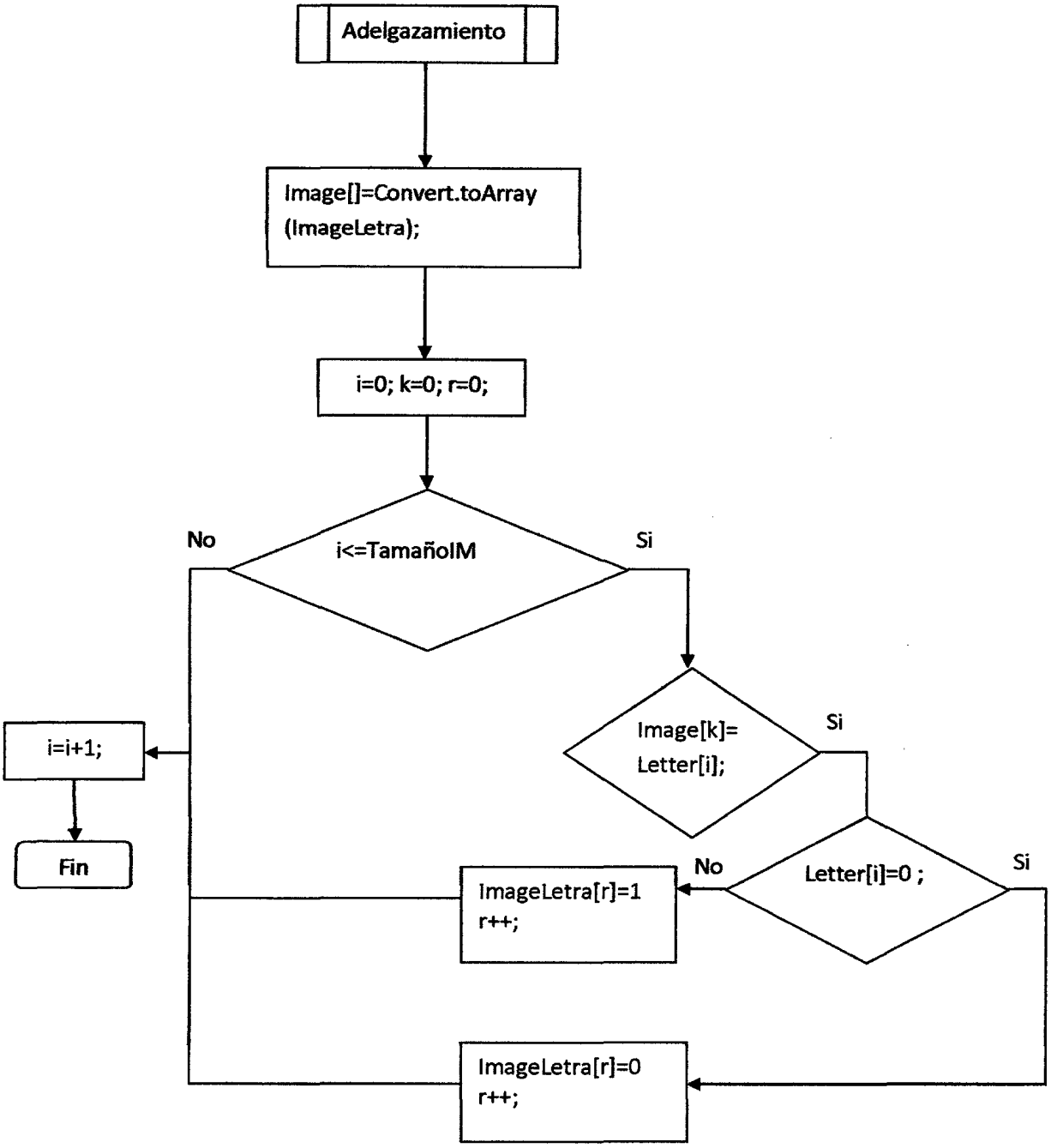
- Diagrama d flujo del algoritmo de **Binarización**



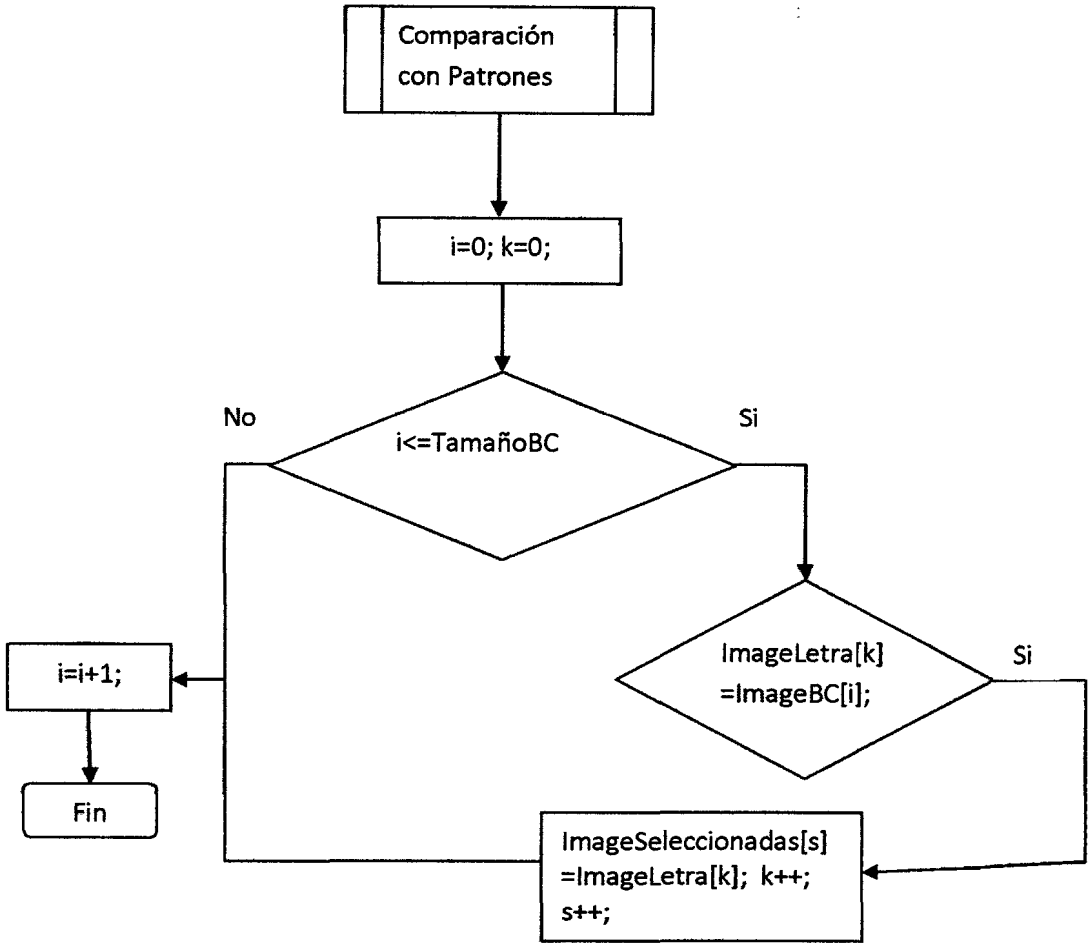
- Diagrama de flujo de fragmentación



- Diagrama de flujo adelgazamiento



- Diagrama de flujo comparación con patrones



e. Codificación

```
[WebMethod]
public void MODI_OCR_AvanzaJFH(string files, string
codigo, int categoria,ref string estado,string dir)
{
//cogemos el tipo de formato que tiene la imagen
string fileExtension =
Path.GetExtension(Convert.ToString(files));
//reemplazamos el formato de la imagen en con una cadena
vacia
string fileName =
Convert.ToString(files).Replace(fileExtension,
string.Empty);
//Verficamos que formatos es lo que tienen las imagenes
en cada carpeta
if (fileExtension == ".jpg" || fileExtension == ".JPG"
|| fileExtension == ".bmp" ||
fileExtension == ".BMP" || fileExtension == ".tif" ||
fileExtension == ".TIF" ||
fileExtension == ".gif" || fileExtension == ".GIF" ||
fileExtension == ".png" ||
fileExtension == ".PNG" || fileExtension == ".tiff" ||
fileExtension == ".TIFF")
{
try{
//OPREACION DE EXTRAER LOS TEXTO DE LAS IMAGENS
MODI.Document md = new MODI.Document(); //instanciamos
el objeto un tipo de documento
md.Create(Convert.ToString(files));//creamos un tipo de
documento
md.OCR(MODI.MiLANGUAGES.miLANG_ENGLISH, true,
true);//elegimos el tipo de lenguaje que utilizaremos
MODI.Image image = (MODI.Image)md.Images[0];
//DEVUELVE AL MODELO NEGOCIO DEL SERVIDOR PARA LUEGO
INSERTARLO
estado=image.Layout.Text;
con.InsertarIndice(codigo,categoria,dir,image.Layout.Tex
t);}
catch (Exception exc)
{
estado=exc.Message;}}}
```

f. Compilación e interpretación

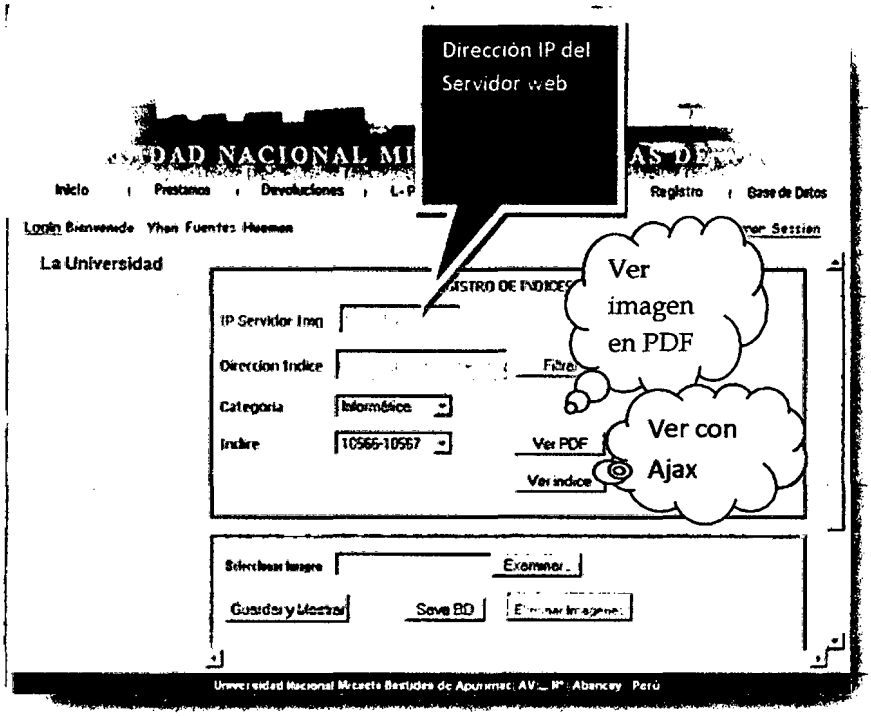


Figura Nº 15: Extraccion del texto de una imagen usando OCR

- En la figura se muestra la extracción de caracteres de una imagen usando el algoritmo OCR, para eso lo primero que hacemos es ubicar la imagen en el servidor donde está guardado, seguidamente filtramos todas las carpetas que están guardadas en el servidor de imágenes y así extraer una por una cada imagen

- Opcion ver en PDF

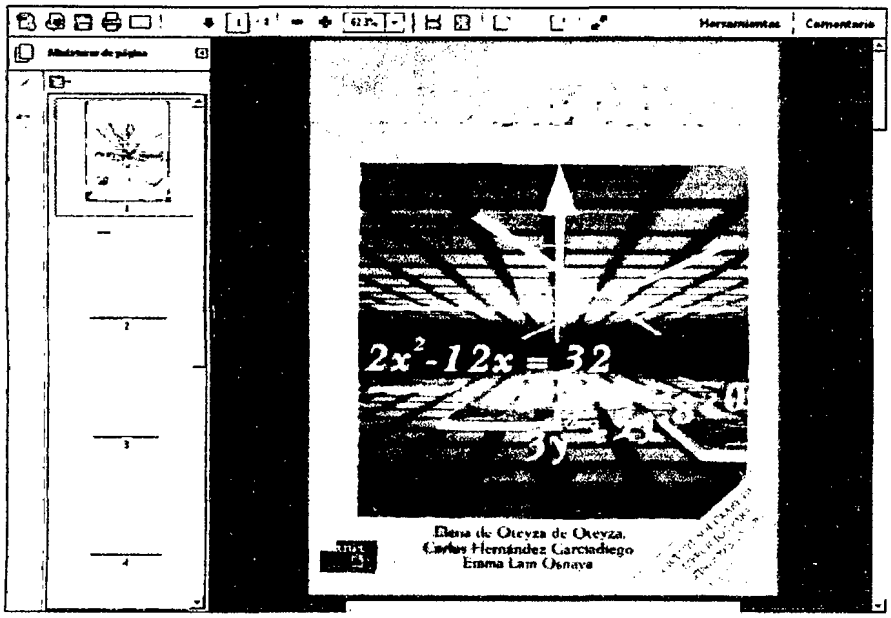


Figura N° 16 : Vista de las imágenes a ser extraidas sus caracteres en PDF
Extracción de caracteres imagen por imagen

- *Opcion ver imágenes a extraer con AJAX*

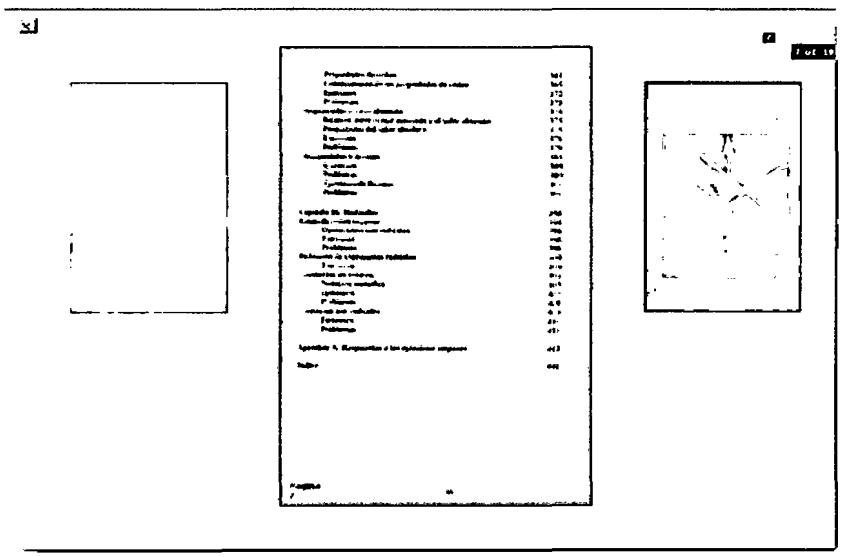


Figura N° 17: Vista de las imágenes a ser extraidas sus caracteres con AJAX

Vista

- Opción extracción de caracteres de la imagen e inserción en la base de datos

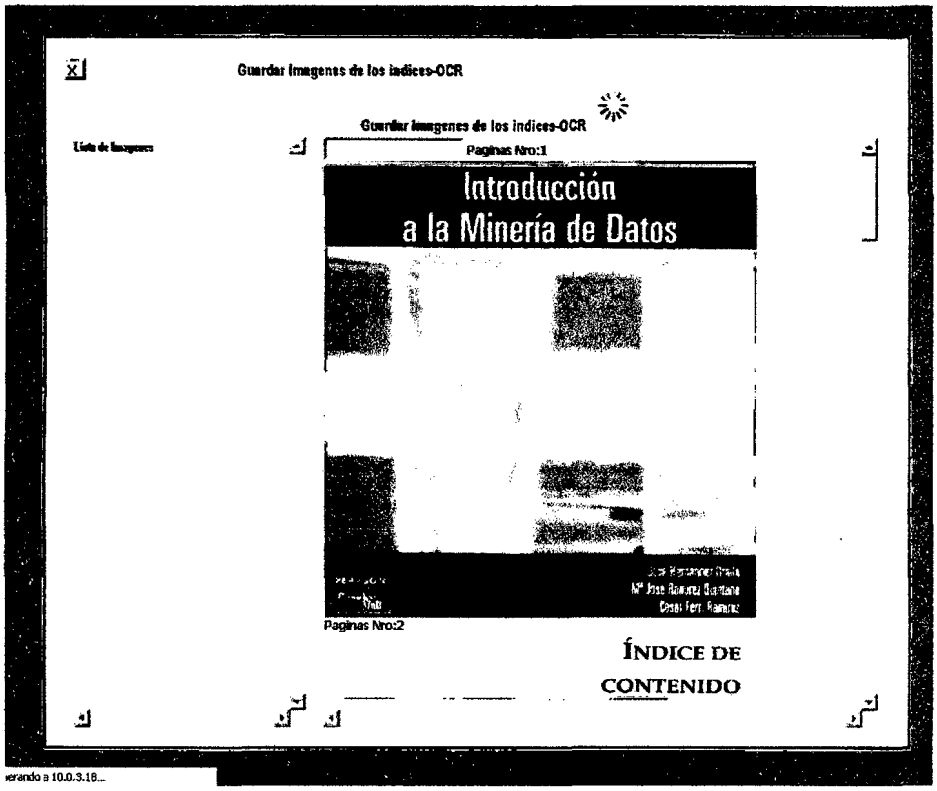


Figura N° 18:muestra de forma de extracción de caracteres de una imagen.

5.1.2. Desarrollo del algoritmo de reconocimiento de patrones en imágenes digitales RPI, para el incremento del número de palabras existentes en la base de datos

$$OE_2 = x_1 \rightarrow x_2$$

a. Análisis lógico

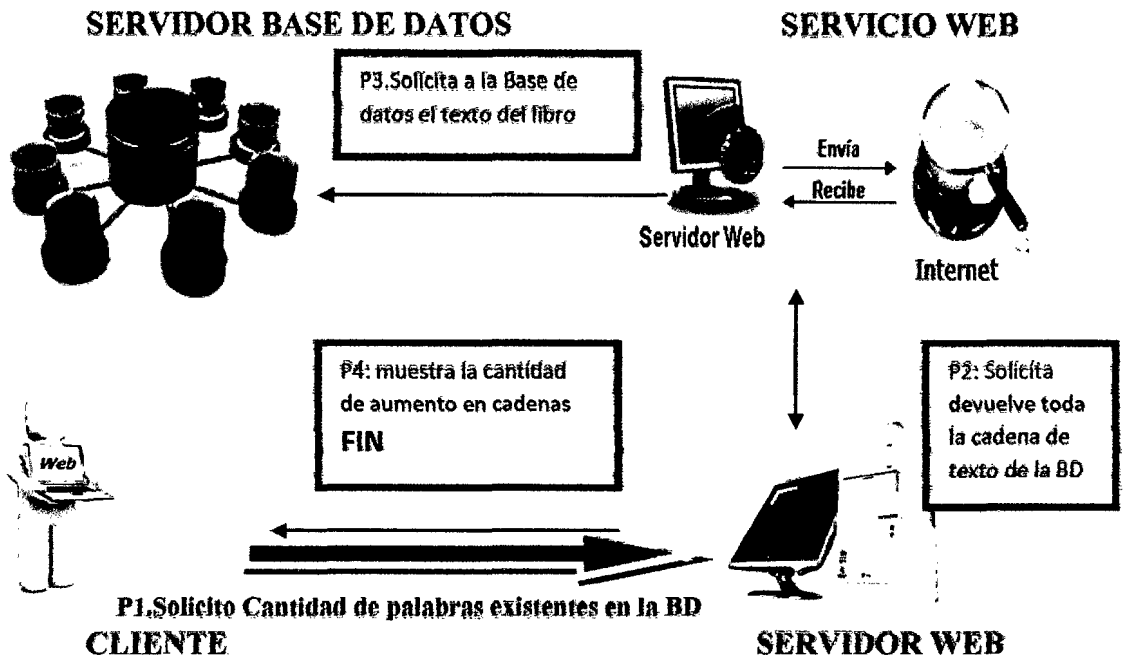


Figura N° 19 :Arquitectura de proceso de conteo del incremento de palabras existentes en la base de dato.

Explicación de cada proceso

P1→ Ubicamos la dirección de la Imágenes de la Índices.

Reconocemos la IP de la maquina

carga de los disco duros existentes en el Servidor web.

P2→cargamos todas la imágenes: una vez seleccionado la IP y el disco duro donde esta la carptea de las imágenes de los índices, **seleccionamos la carpeta por descripcion principal** las cuales seran insertardas en la base de datos una vez extraídas el texto de las imágenes de los índices del libro .

P3→Solicita al método OCR del servicio web para extraer el texto de la imagen.

P4→ Devuelve el texto de la imagen extraída y ese mismo texto lo inserta en base de datos.

b. Analisis matemático

$OE_2 = x_1 \rightarrow x_2$

Determinar el incremento del número de caracteres existentes en la base de datos aplicando el algoritmo de reconocimiento de patrones en imágenes digitales.

Demostrar que: NPE_bd= NPE_bd+C $\leftrightarrow C > 0$?..... **(β)**

Si cumple lo definido entonces $OE_2 = x_1 \rightarrow x_2$ *sería correcto y* $x_1 \rightarrow x_2$ *Serian variable correlacionales*

Porque: NPE_bd $\in x_1$ y NPE_bd+C $\in x_2$

Donde:

NPE_bd: número de palabras existentes en la base de datos

C: constante de incremento de palabras existentes en la base de Datos

Sea:

$$CE = \{k_1, k_2, k_3, \dots \dots k_{i,\dots\dots}, k_j \ \dots\dots k_x\}\ CA =$$

$$\{'A','B','C','D', \dots \dots \dots 'Z','1','2','4','5','6', \dots \dots .'9'\}\ o$$

$$CA = \{A_{0,0}, A_{0,1}, A_{0,2,} \dots \dots . A_{0,28,} . . , A_{k,0} , \dots . A_{k,28},$$

$$A_{w,0} , \dots . . A_{w,8}\}$$

Donde sabemos que:

CE: Conjunto de caracteres extraídos anteriormente

$i = \overline{1, n'}$; j= $\overline{1, m'}$ además $i \leq j$;

CA: Conjunto de caracteres alfabéticos y numéricos ya definidos

$$CV = c_i \in CA \ \wedge \ c_i \in CEr$$

Donde:

CV: Conjunto de caracteres validos

$$i = \overline{1, n'};$$

$$CEr = c_i \notin CA$$

Donde:

CEr: Caracteres de error

$i = \overline{1, n'}$;

Entonces: $\mathbf{Dim(CEr) + Dim(CV) = n'}$

Forma de CV:

$$CV = \{c_1, c_2, c_3, \dots \dots c_{i,\dots\dots}, c_j \ \dots . . c_m\}$$

$$CEr = \{e_1, e_2, e_3, \dots \dots e_{i,\dots\dots} \ \dots . . e_k\}$$

Sea:

P : Conjunto de palabras validas

$p_i \in P$: Donde p_i *son subconjuntos de palabras*

$p_0 = \{k_1, k_2, k_3\}$ $\qquad$ $e_0 = \{k_4, k_5\}$

$p_1 = \{k_6, k_7, k_8\}$ $\quad e_1 = \{k_9\}$

$p_u = \{k_i, \dots \dots, k_j\}$ $\quad e_s = \{k_r\}$

En general

$p_{w-j} = \{k_x, \dots \dots, k_m\}$

Entonces:

$$P = \{p_1, p_2, p_3, \dots \dots p_{i,\dots\dots}, p_u \ \dots\dots p_w\}$$

$$p_i \in P \ ; \ i = \overline{1, w};$$

$$p_u = \{k_i, \dots \dots, k_j\} \ = \sum_{x=i}^{j} k_x$$

Donde:

$$k_i \in CE \ y \in k_i \in CV$$

$$i \leq j$$

$$p_u = \sum_{x=i}^{j} k_x \dots\dots\dots\dots (1)$$

Sea:

CT: conjunto de caracteres existentes en la base de datos del campo titulo, Autores, Descripción Principal de la tabla libro de la base de datos (BDBiblioteca)

PT: conjunto de palabras formadas con los caracteres existentes en el conjunto CT

Entonces:

$$p'_0 = \{ce_1, ce_2, ce_3\}$$

$$p'_1 = \{ce_4, ce_5, ce_6\}$$

$$p'_v = \{ce_i, \dots \dots, ce_j\}$$

Donde:

$$PT = \{p'_1, p'_2, p'_3, \dots \dots p'_{i,\dots\dots}, p'_j \ \dots \dots p'_v\}; p'_i \in PT \quad ;$$

$$i = \overline{1, v};$$

$$ce_i \in CT \wedge p'_i \ \in PT$$

$$i \leq j$$

$$p'_v = \sum_{y=i}^{j} ce_y \ \dots\dots\dots\dots\dots\dots\dots(2)$$

<u>Entonces analizando (1) y (2):</u>

p'_v: Indica el 100% de palabras existentes en la base de datos sin usar el algoritmo de reconocimiento de patrones en imágenes digitales para extraer más caracteres del índice de las imágenes e incrementar de datos la base de datos.

p_u: Indica un X% de incremento de palabras en base de datos de la biblioteca en el campo índice de los caracteres validos.

Entonces

NPE_bd=p'_v ..... que índica el 100% de palabras en la base de datos.

NPE_bd= NPE_bd +C …….. Indica el crecimiento en X% más del 100% de palabras en la BD... (3)

<u>En (3) demostremos el incremento:</u>

Si: NPE_bd= NPE_bd +C entonces

NPE_bd= p'_v +C; C: es el incremento y C=p_u

NPE_bd= $p'_v + p_u$

<u>Reemplazamos (1) y (2):</u>

NPE_bd= $\sum_{y=i}^{j} ce_y + \sum_{x=i}^{j} k_x$

NPE_bd=100%+X% de incremento

Entonces $OE_2 = x_1 \rightarrow x_2$ ***es correcto indica que si existe un crecimiento de palabras en la base de datos, esto implica afirmar que hay un incremento de caracteres en la base.***

c. Declaración de la variables

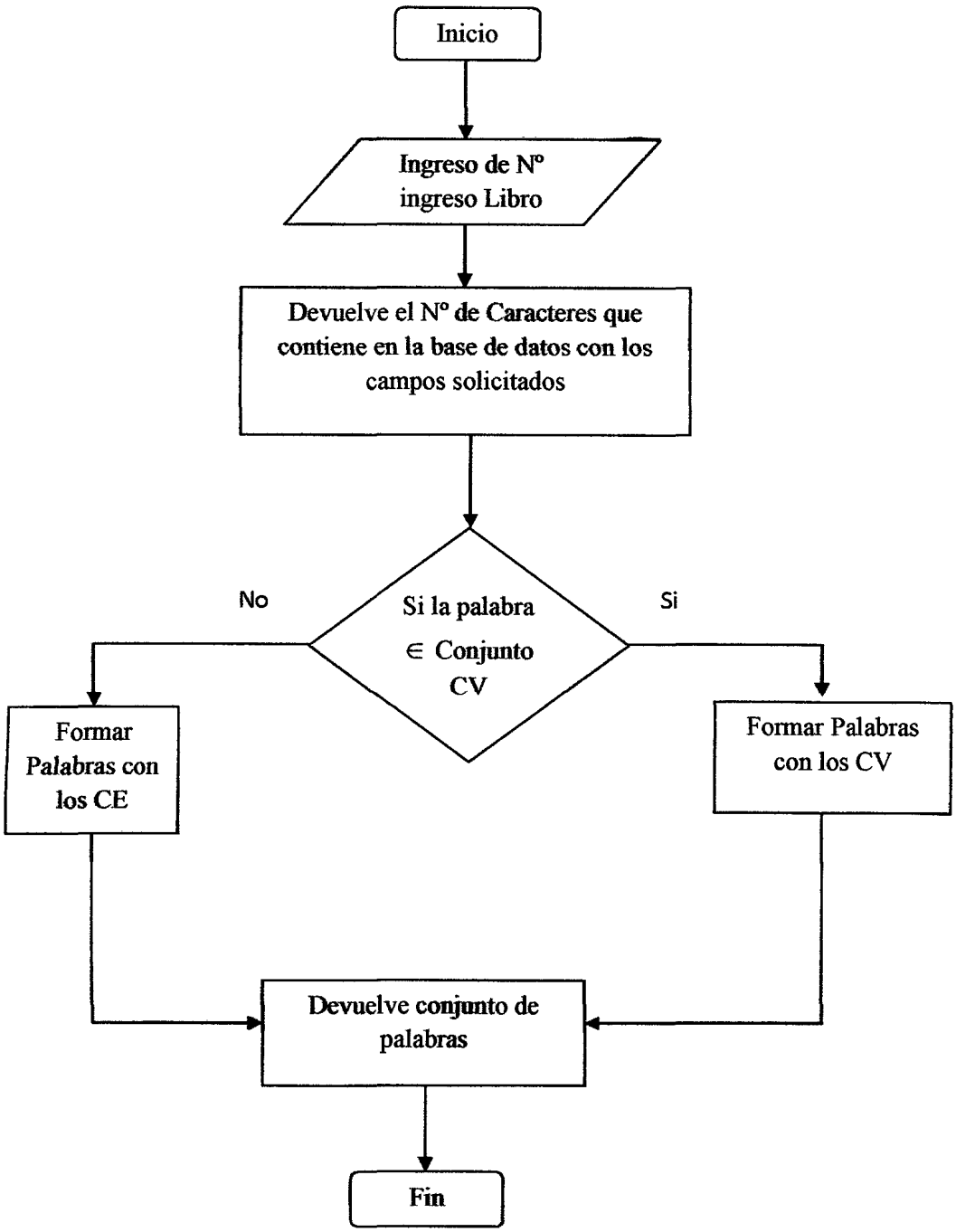
Variables de salida

- C:constante de incremento de palabras en la base de datos
- NPE_bd: número de palabras existentes en la base de datos

d. Diagrama de flujo

CV: Conjunto de caracteres validos

CE: Conjunto de caracteres Error



e. Codificación

```
//código buscar N¦ caracteres en base de datos
string CaracteresBD = con.NPEx_bd(txtN°_ingresoL.Text);
string CIndice = con.CV_(txtN°_ingresoL.Text);
int NPE_bd=0,NPex=0,CE_bd=0,CE_Ex=0;

jmodi.ORC_countCaracteres(CaracteresBD.ToCharArray(),
ref CE_bd,ref NPE_bd);
jmodi.ORC_countCaracteres(CIndice.ToCharArray(), ref
CE_Ex, ref NPex);

lblNPE_bd.Text =""+NPE_bd;
int margeErro = Convert.ToInt32(CE_Ex * 0.95);
lblCV.Text = "" + (NPex +margeErro);
lblNPExBD.Text = "" + (NPE_bd + NPex+margeErro);
//función que devuelve el numero de caracteres
existentes en la base de datos
public string NPEx_bd(string N°_ingresoL)
{
string NPE_bd="";
try
{
cmd = new SqlCommand("select Titulo+Autores+DescripcionP
"+
      "  from libro l where
      l.N°_ingresoL='"+N°_ingresoL+"' ", miconexion);
rd = cmd.ExecuteReader();

while (rd.Read())
{
NPE_bd = rd.GetString(0);
}
rd.Close();
return NPE_bd;}
catch (Exception ex)
{return "Eror " + ex.Message;}}
```

f. Compilación e interpretación

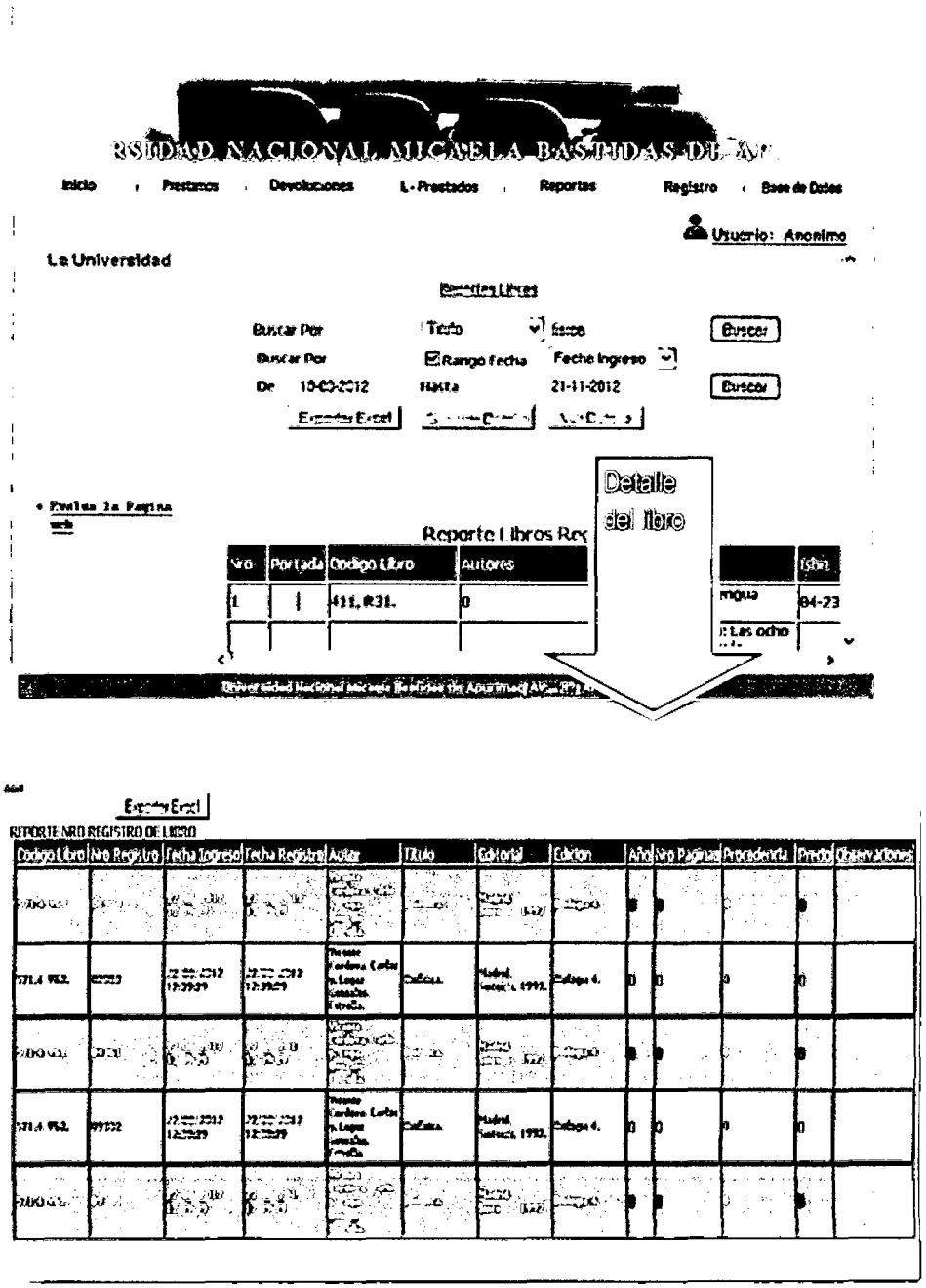


Figura Nº 20 :muestra de la interpretacion de resultados de la aplicación que contiene el algoritmo OCR

5.1.3. Desarrollo del algoritmo para determinar el incremento del número de respuestas válidas en la búsqueda de textos por índice de contenido digitalizado, aplicando el algoritmo de reconocimiento de patrones en imágenes digitales.

$$OG = x_2 \rightarrow y_1$$

a. Análisis lógico

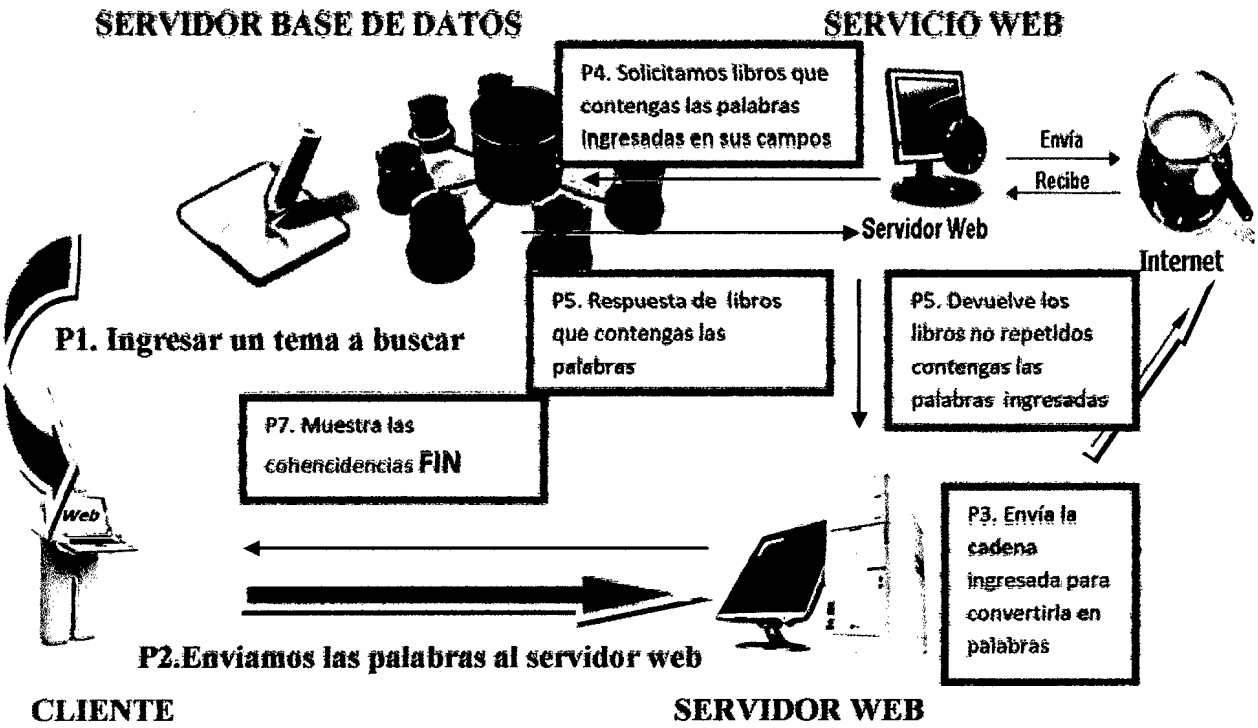


Figura N° 21:Arquitectura de proceso de búsqueda de un tema requerido

b. Análisis Matemático

Objetivo General: $OG = x_2 \rightarrow y_1$ determinar el incremento del número de respuestas validas en la búsqueda de textos por índice de contenido digitalizado, aplicando el algoritmo de reconocimiento de patrones en imágenes digitales

Demostrar que: NLE=NLE+k; $k > 0 \leftrightarrow$?..... (ϕ)

Entonces afirmaremos que: $OG = x_2 \rightarrow y_1$ *es correcto y* $OE_3 = x_2 \rightarrow y_1$ *estas 2 variables son dependientes.*

Porque: NLE $\in x_2$ y NLE'= NLE+$Q \in y_1$

Donde:

NLE: número de libros encontrados referentes al tema buscado

Q: constante de incremento del número de libros más encontrados referentes al tema buscado.

Supongamos que:

CI: conjunto de caracteres ingresados por los usuarios a buscar

PI_i : Conjunto de palabras formadas por los caracteres CI

PI: conjunto que contiene las palabras PI_i

Si:

$$CI = \{c'_1, c'_2, c'_3, \ldots\ldots c'_{k,\ldots\ldots}, c'_j \ \ldots\ldots c'_M\}$$

$$PI_0 = \{c'_1, c'_2, c'_3\}$$

$$PI_1 = \{c'_4, c'_5, c'_6\}$$

⋮

Forma general

$$PI_w = \{c'_k, \ldots\ldots, c'_j\}$$

Entonces

$$PI = \{PI_1, PI_2, PI_3, \ldots\ldots PI_{i,\ldots\ldots}, PI_u \ \ldots\ldots PI_w\}$$

$$PI_i \in PI \ ; \ i = \overline{1, w};$$

$$Dim(PI_i) = M$$

$$\boldsymbol{PI_i} = \sum_{x=k}^{j} c'_x \quad \ldots\ldots\ldots (1) \quad k \leq j$$

Donde:

$$\boldsymbol{x} = \overline{0, M} \ ; \boldsymbol{i} = \overline{0, w};$$

Ahora

NLE=NLE+ Q ; $Q > 0$ donde Q es el valor del incremento en un y% de respuestas más, referentes a la búsqueda de las palabras ingresadas

n(NLE): representa el 100% de libros encontrados referentes a las palabras ingresados o tema a buscar

n(Q): representa el Y% de incremento de libros encontrados más en la base de datos.

Sea NPE=representa el número de libros encontrados referente a las palabras coincididas en el conjunto de NPE_bd;

Donde:

$$NLE_i = p'_v \quad \leftrightarrow \quad PI_i = p'_v$$

$$\sum_{x=k}^{j} c'_x = \sum_{y=i}^{j} ce_y$$

NPE=$\{NLE_1, NLE_2, NLE_3, \ldots\ldots, NLE_u \ldots\ldots NLE_t\}$... (1)

NPE: Representa el 100% de los libros encontrados referentes al tema buscado

K= representa el número de libros encontrados referente a las palabras coincididas en el conjunto de $Q = p_u$

$$Q_j = p_u \leftrightarrow PI_i = p_u$$

$$\sum_{x=k}^{j} c'_x = \sum_{x=i}^{j} k_x$$

Q=$\{Q_1, Q_2, Q_3, \ldots\ldots Q_{i,\ldots\ldots}, Q_u \ldots\ldots Q_F\}$..... (2)

Q: representa el incremento en un Y% de libros encontrados referentes al tema buscado.

Entonces el nuevo: NLE=NLE+Q

Reemplazando (1) + (2)

$$NLE=\sum_{x=i}^{t} NLE_x+\sum_{y=1}^{f} Q_y$$

Por tanto el supuesto $NLE \in x_2$ ***y*** $NLE'=NLE+Q \in y_1$

además que Q existe y es un valor mayor a 0

$OG = x_2 \rightarrow y_1$ ***es correcto y*** $OE_3 = x_2 \rightarrow y_1$ ***son variables que estan una en función de la otra***

b. Declaracion de Variables

- Variables de entrada

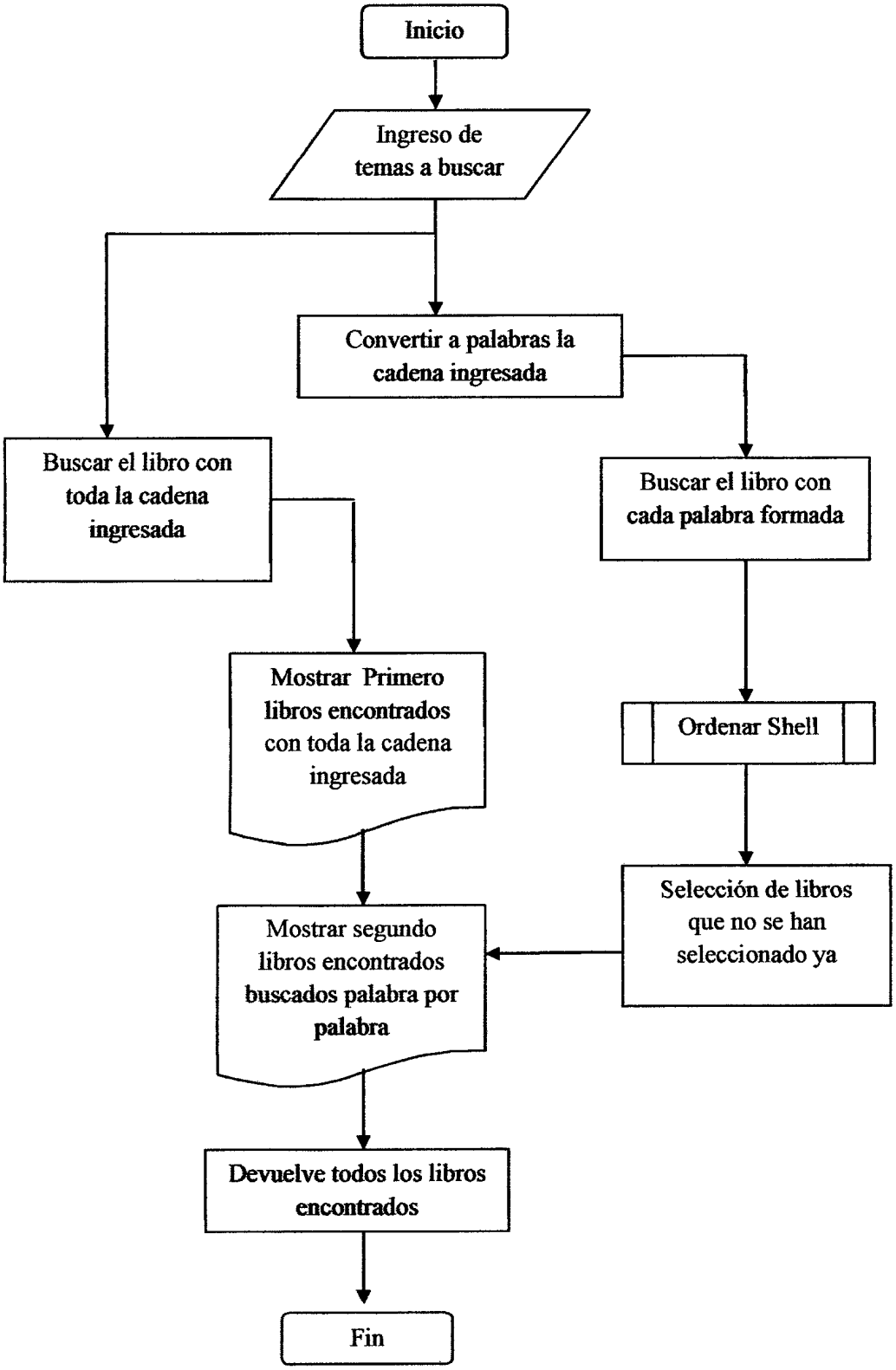
  CI: número de caracteres incresados a buscar

- variables de salida

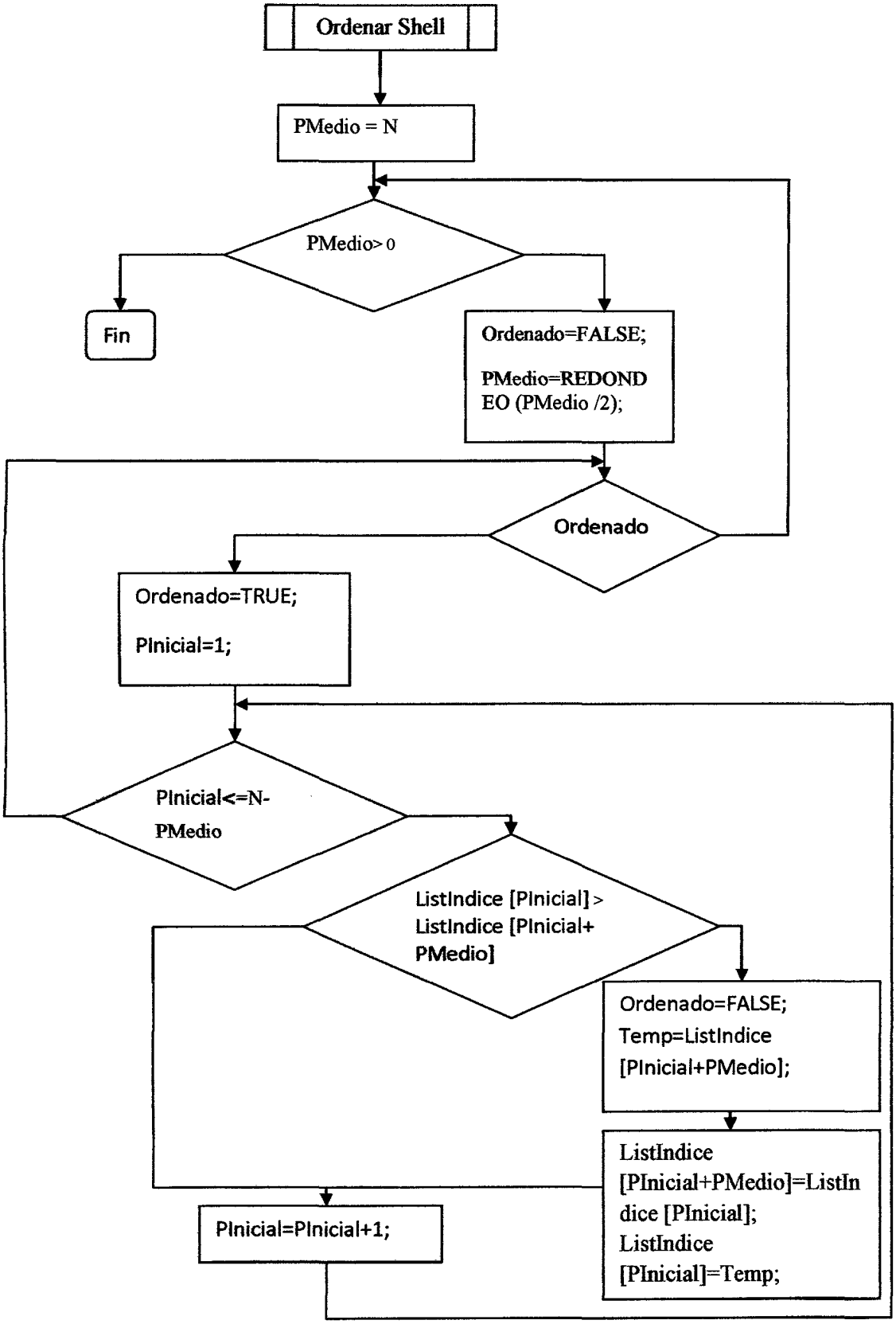
  NLE: número de respuestas encontrado encontrados

c. **Diagrama de flujo**

- Diagrama de flujo método en general del objetivo N° 2



- Diagrama de flujo de la función del método **Ordenar Shell**.



d. Codificación

```
protected void btnConsultar_Click(object sender,
EventArgs e)
{
string estad = "";
lblMensaje.Text = "";

ListboxLibros.Items.Clear();
// lo primero primer es generar los patrones de
palabras
try
{
string estado = "";
ArrayList listPatronPal =
complet.Patronpalabras(txttipobusqueda.Text, ref
estad);

//lista para los valores que contiene todas las
palabras ingresadas
ArrayList listN°_ingresoL = new ArrayList();
//lista para N° ingreso libro diferentes que
buscarems luego
ArrayList listN°_ingresoLBuscados = new
ArrayList();
//lista de arreglo de los libros encontrados por
palabra
ArrayList listtemp = new ArrayList();
lblresultConsult.Text = "";

//cuando el checkbox de busqueda especifica esta
activado buscams espeficicamente
if (ChecbxBsucarAvanzada.Checked)
{
try
{
//primero almacenamos solo con con el que
contiene toda la frase
listacursos =
con.BusquedaEspeficica(DropDowCamposLibro.Select
edItem.Value,
txttipobusqueda.Text, ref estado);


if (listacursos.Count > 0)
{
foreach (CLibro item in listacursos)
```

```
{

txtcodigoL.Text = item.CodigoL.ToString();
txtN°Reg.Text = item.N°ingresoL.ToString();
listN°_ingresoL.Add(item.N°ingresoL);
ListboxLibros.Items.Add(item.Titulo + "&" +
item.N°ingresoL + "&" + item.CodigoL +
" " + item.Autores + " " + item.Stock);

}
}
//luego buscamos palabra por palabra

if (listPatronPal != null && listPatronPal.Count
> 0)
{

for (int i = 0; i < listPatronPal.Count; i++)
{
listacursos =
con.BusquedaEspeficica(DropDowCamposLibro.Select
edItem.Value,
listPatronPal[i].ToString(), ref estado);

//llenamos toods los N°_registros encontrados en
la consulta ee
//la primera palabra
foreach (CLibro item in listacursos)
{
listN°_ingresoLBuscados.Add(item.N°ingresoL);
//lista general toods los libros con
repeticiones
}
}
//aki adjunatremos ala listo solo los codigos
seleccionados con una nueva consulta
complet.shell_Ordena(ref
listN°_ingresoLBuscados);
if ((listN°_ingresoLBuscados.Count > 0) &&
(listN°_ingresoL.Count > 0))
{

listtemp =
complet.patronseleccion(listN°_ingresoL,
listN°_ingresoLBuscados);
ArrayList listnew = new ArrayList();
for (int i = 0; i < listtemp.Count; i++)
```

```
{
listnew =
con.GetdatosN°_ingresoL(listtemp[i].ToString());

foreach (CLibro item in listnew)
{

ListboxLibros.Items.Add(item.Titulo + "&" +
item.N°ingresoL + "&" + item.CodigoL +
" " + item.Autores + " " + item.Stock);
}
}
}
olso
{
if ((listN°_ingresoL.Count <=0) &&
(listN°_ingresoLBuscados.Count > 0))
{
ArrayList listnew = new ArrayList();
for (int i = 0; i <
listN°_ingresoLBuscados.Count; i++)
{
listnew =
con.GetdatosN°_ingresoL(listN°_ingresoLBuscados[
i].ToString());

foreach (CLibro item in listnew)
{

ListboxLibros.Items.Add(item.Titulo + "&" +
item.N°ingresoL + "&" + item.CodigoL +
" " + item.Autores + " " + item.Stock);
}

}
}
else
{
ArrayList listnew = new ArrayList();
for (int i = 0; i < listN°_ingresoL.Count; i++)
{
listnew =
con.GetdatosN°_ingresoL(listN°_ingresoL[i].ToStr
ing());

foreach (CLibro item in listnew)
{
```

```
ListboxLibros.Items.Add(item.Titulo + "&" +
item.N°ingresoL + "&" + item.CodigoL +
" " + item.Autores + " " + item.Stock);
}
}
}
}
}

lblresultConsult.Text = "Resultados de la
Consultas:" + ListboxLibros.Items.Count;
lblMensaje.Text = "";
}
catch (Exception ex)
{

lblMensaje.Text = ex.Message;
}
}
//cuando buscamos en forma general
else
{
try
{
//primero almacenamos solo con con el que
contiene toda la frase
estad = "";

listacursos =
con.busquedaGeneral(txttipobusqueda.Text,ref
estad);

if (listacursos.Count > 0)
{
foreach (CLibro item in listacursos)
{
txtcodigoL.Text = item.CodigoL.ToString();
txtN°Reg.Text = item.N°ingresoL.ToString();
listN°_ingresoL.Add(item.N°ingresoL);

ListboxLibros.Items.Add(item.Titulo + "&" +
item.N°ingresoL + "&" + item.CodigoL +
" " + item.Autores + " " + item.Stock);

}
}
```

```
//luego buscamos palabra por palabra
if (listPatronPal != null && listPatronPal.Count
> 0)
{
for (int i = 0; i < listPatronPal.Count; i++)
{
listacursos =
con.busquedaGeneral(listPatronPal[i].ToString(),
ref estado);

//llenamos toods los N°_registros encontrados en
la consulta ee
//la primera palabra
foreach (CLibro item in listacursos)
{
listN°_ingresoLBuscados.Add(item.N°ingresoL);
//lista general toods los libros con
repeticiones
}
}
//aquii adjunatremos ala listo solo los codigos
complet.shell_Ordena(ref
listN°_ingresoLBuscados);
if ((listN°_ingresoLBuscados.Count > 0) &&
(listN°_ingresoL.Count > 0))
{

listtemp =
complet.patronseleccion(listN°_ingresoL,
listN°_ingresoLBuscados);
lblresultConsult.Text +=listN°_ingresoL.Count +
"!" + listN°_ingresoLBuscados.Count;
ArrayList listnew = new ArrayList();
for (int i = 0; i < listtemp.Count; i++)
{
listnew =
con.GetdatosN°_ingresoL(listtemp[i].ToString());
foreach (CLibro item in listnew)
{
ListboxLibros.Items.Add(item.Titulo + "&" +
item.N°ingresoL + "&" + item.CodigoL +
" " + item.Autores + " " + item.Stock);
}
}
}
else
{
```

```
if ((listN°_ingresoL.Count<=0) &&
(listN°_ingresoLBuscados.Count > 0))
{
ArrayList listnew = new ArrayList();
for (int i = 0; i <
listN°_ingresoLBuscados.Count; i++)
{
listnew =
con.GetdatosN°_ingresoL(listN°_ingresoLBuscados[
i].ToString());

foreach (CLibro item in listnew)
{

ListboxLibros.Items.Add(item.Titulo + "&" +
item.N°ingresoL + "&" + item.CodigoL +
" " + item.Autores + " " + item.Stock);
}
}
}
else
{
ArrayList listnew = new ArrayList();
for (int i = 0; i < listN°_ingresoL.Count; i++)
{
listnew =
con.GetdatosN°_ingresoL(listN°_ingresoL[i].ToStr
ing());

foreach (CLibro item in listnew)
{
ListboxLibros.Items.Add(item.Titulo + "&" +
item.N°ingresoL + "&" + item.CodigoL +
" " + item.Autores + " " + item.Stock);
}}
}}
}

lblresultConsult.Text = "Resultados de la
Consultas:" + ListboxLibros.Items.Count;
lblMensaje.Text = "";
}
catch (Exception ex)
{

lblMensaje.Text = ex.Message;
}}
```

```
}
catch (Exception ex)
      {lblMensaje.Text = ex.Message;}
      }
```

- **Metodo shell**

```
#region metodo de ordenamiento
[WebMethod]
public void shell_Ordena(ref String[] listIndice)
{
String listcad1, listcad2, temp;
int i, k, incremento, j, ptr;
int val1 = 0, val2 = 0;
int total = 0;
total = listIndice.Length;
incremento = total / 2;
while (incremento > 0)
{//
for (i = incremento; i <= total - 1; i++)
{///
j = i - incremento;
while (j >= 0)
{////
//instanciamos la cantidad maxima de los valores
listcad1 = listIndice[j];
val1 =
Convert.ToInt32(ExtraerNum(listcad1.ToCharArray()));
listcad2 = listIndice[j + incremento];
val2 =
Convert.ToInt32(ExtraerNum(listcad2.ToCharArray()));

if (val1 >= val2)
{/////
temp = listcad1;
//cambio de numero
listIndice[j] = listIndice[j + incremento];
listIndice[j + incremento] = temp;
}/////
else
{
j = 0;

}
j = j - incremento;
}////

/*for(int h=0;h<=total-1;h++)
{
richTextBox1.Text=listIndice[h,1].ToString();
}*/

}
incremento = incremento / 2;
}
}

#endregion
```

e. Compilación e interpretación

Módulo de búsqueda de libros referentes a temas buscados, se observa que la aplicación contiene funcionalidades como ver el libros encontrado su índece en PDF con en controles AJAX

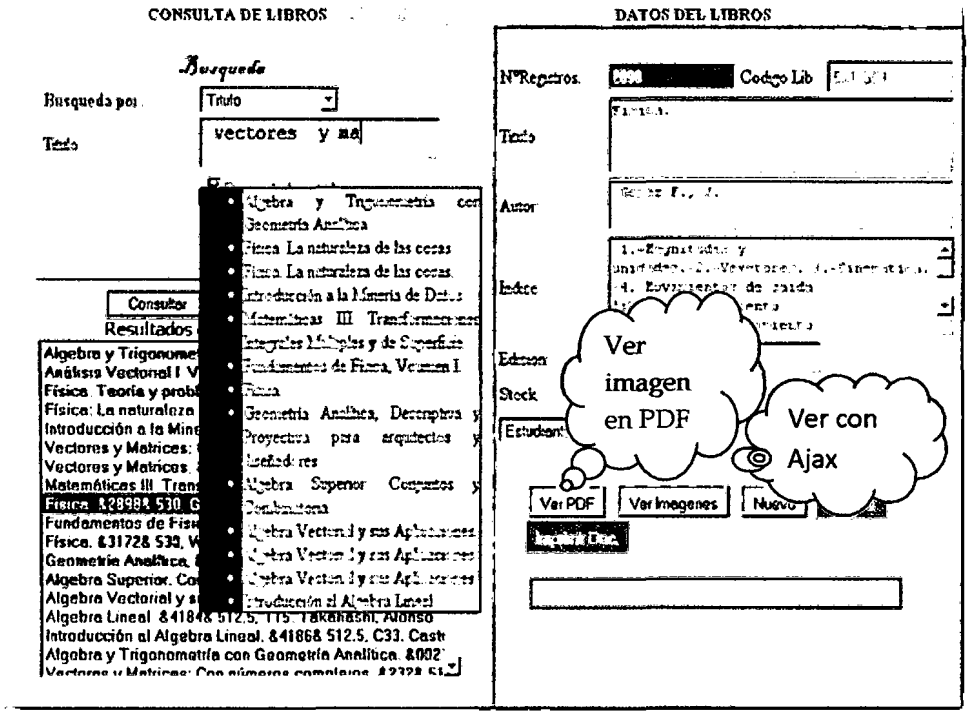


Figura N° 22 : Muestra los libros referentes a los parametros de búsqueda

Módulo consultas una de las resaltantes funciones que otorga este portal es la muestra de los índices extraido el texto es que una vez alamacenados los textos o antes de extraerlos se puede ver el índice completo del libro en PDF

Figura Nº 23:Muestra los PDF de los índice de los libros referentes a los parametros de búsqueda

5.2. Resultados de tablas

a. Cuadro de resultado de conteo de número de libros registrados que tienen sus índices ya escaneados en la biblioteca central UNAMBA.

Título Libro	Número de Libros	Tamaño total MB	Número de Libros a Tomar	N° de Libros por Categoría
Administración	27	50	2,7	2
Agroindustrial	1	7,7	0,1	1
Física	12	16	1,2	1
Informática	5	40,5	0,5	1
Ingeniería de Máquinas	1	1,26	0,1	1
Matemática	79	75,6	7,9	7
Mineralogía	2	6,37	0,2	1
Química	9	231	0,9	1
Zootecnia	3	9,9	0,3	1
TOTAL LIBROS *(Registrados)*	139	438,33	13,9	16

Tabla N° 2: Número de libros registrados que tienen sus índices ya escaneados en la biblioteca central UNAMBA por categoría de clasificación DEWEY.

b. Cuadro de resultados sobre el **objetivo general,** Incremento del número de libros válidos encontrados en la búsqueda de textos por índice de contenido digitalizado aplicando el algoritmo de reconocimiento de patrones en imágenes digitales.

N°	Descripción	N° Ingreso Libro	N° Hoja	Patrones de palabras	Y1 con X2	Y1 sin X2	% de incremento de rptas validas
1	Administración	1750	1	marketing estrategias programas aplicaciones	100	98	2,0
2	Administración	9023	1	recursos humanos empresa funcion	154	148	4,1
3	Agroindustrial	4380	2	operacion unitaria sistemas de magnitudes	81	79	2,5
4	Física	2898	2	fisica de gomez caida libre	129	127	1,6
5	Informática	03753-09211-09469-09470	2	fundamentos de la inteligencia artificial aplicaciones	108	106	1,9
6	Ing. Maquinas	09223-10615	2	teoria de la bomba centrifuga proyecto	112	109	2,8
7	Matemáticas	207-209-217-4127-4128	1	ejercicios desarrolados de graficas especiales	26	17	52,9
8	Matemáticas	0216-02765	1	vectores en el espacio,combinaciones lineales	10	6	66,7
9	Matemáticas	231	1	funciones reales de	26	19	36,8

				variable reales			
10	Matemáticas	238	2	funciones constantes, calculo de figueroa	53	48	10,4
11	Matemáticas	239	1	aplicaciones de la derivada calculo figueroa	91	87	4,6
12	Matemáticas	2093	1	longitud de arco area de una superficie	7	4	75,0
13	Matemáticas	2098	2	funciones polinomiales douglas faire	19	15	26,7
14	Mineralogía	4278	3	tratamiento de sondeos captacion y alumbramien to	9	7	28,6
15	Química	3166	4	formulacion y nomenclatura de series homologas	13	8	62,5
16	Zootecnia	9684	2	embriologia veterinario ayuda para comprender	13	10	30,0

Tabla N° 3: Datos estadísticos del incremento de número de respuestas validas de libros referentes a un tema buscado

c. Cuadros de resultados sobre el **objetivo 1**, Número de palabras extraídas aplicando el algoritmo de OCR.

N°	Descripción	Especie Folio	N° Hoja	Exactitud en Caracteres	Tamaño Total en KB	Tamaño en KB Extraído	Tamaño en KB Final Limpio	N° de Palabras Extraídas	N° de Palabras Válidas	N° Error
1	Administración	1750	1	98,9172%	835,00	824,15	10,85	**1002**	**991**	**11**
2	Administración	9023	1	98,8304%	1002,00	987,80	14,20	**1197**	**1183**	**14**
3	Agroindustrial	4380	2	98,9622%	1166,00	1151,45	14,55	**1349**	**1335**	**14**
4	Física	2898	2	99,1028%	1055,00	1043,40	11,60	**1226**	**1215**	**11**
5	Informática	03753-09211-09469-09470	2	98,9613%	1470,00	1452,1	17,90	**1733**	**1715**	**18**
6	Ing. Maquinas	09223-10615	2	98,9865%	502,00	496,15	5,85	**592**	**586**	**6**
7	Matemáticas	207-209-217-4127-4128	1	98,8294%	1023,00	1009,35	13,65	**1196**	**1182**	**14**
8	Matemáticas	0216-02765	1	98,8681%	717,00	707,30	9,70	**857**	**847**	**10**
9	Matemáticas	231	1	98,9344%	1293,00	1277	16,25	**1525**	**1509**	**16**
10	Matemáticas	238	2	98,9254%	944,00	932,05	11,95	**1112**	**1100**	**12**
11	Matemáticas	239	1	98,9149%	1087,00	1073,1	13,90	**1281**	**1267**	**14**
12	Matemáticas	2093	1	98,9906%	994,00	982,15	11,85	**1174**	**1162**	**12**
13	Matemáticas	2098	2	99,0523%	1083,00	1071,05	11,95	**1261**	**1249**	**12**
14	Mineralogía	4278	3	98,7765%	1036,00	1021	14,95	**1226**	**1211**	**15**
15	Química	3166	4	98,8806%	1231,00	1215	15,85	**1416**	**1400**	**16**
16	Zootecnia	9684	2	99,0956%	664,00	657	7,25	774	767	7

Tabla N° 4: resultado del número de caracteres extraídos usando el algoritmo OCR

d. Cuadro de resultados del **objetivo 2**, Incremento del número de caracteres existentes en la base de datos aplicando el algoritmo de reconocimiento de patrones en imágenes digitales.

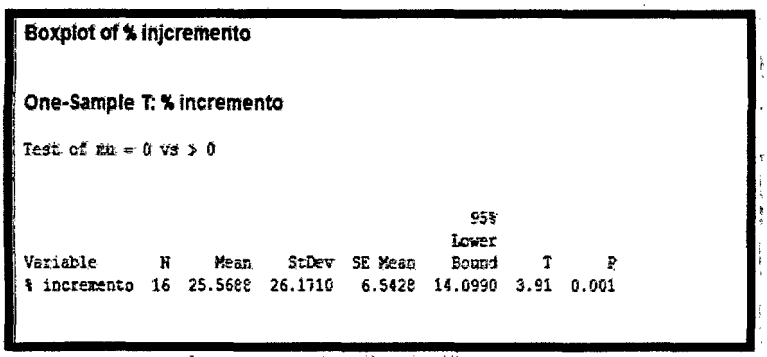
N°	Descripción	N° Ingreso Libro	N° Hoja	% de Caracteres en Aumento en la BD	N° Total de Caracteres con incremento	N° de caracteres validos extraidos	N° de caracteres Existentes en la BD
1	**Administración**	1750	1	2517,439%	1032	991	41
2	Administración	9023	1	1892,424%	1249	1183	66
3	Agroindustrial	4380	2	133600,0%	1336	1335	1
4	Física	2898	2	6494,736%	1234	1215	19
5	**Informática**	03753-09211-09469-09470	2	2417,567%	1789	1715	74
6	Ing. Maquinas	09223-10615	2	723,4043%	680	586	94
7	Matemáticas	207-209-217-4127-4128	1	2669,565%	1228	1182	46
8	Matemáticas	0216-02765	1	1536,101%	906	847	59
9	Matemáticas	231	1	3871,875%	1549	1509	40
10	Matemáticas	238	2	2600,113%	1144	1100	44
11	Matemáticas	239	1	3348,974%	1306	1267	39
12	Matemáticas	2093	1	2424,300%	1212	1162	50
13	Matemáticas	2098	2	2702,187%	1297	1249	48
14	Mineralogía	4278	3	1211,009%	1320	1211	109
15	Química	3166	4	2287,734%	1464	1400	64
16	Zootecnia	9684	2	803,669%	876	767	109

Tabla N° 5: Datos del número caracteres existentes en la base de datos por libro extraídos e insertados en base de datos biblioteca aplicando el algoritmo de reconocimiento de patrones en imágenes digitales.

5.3. Resultado de las pruebas estadísticas

- Según los resultados del promedio del incremento promedio porcentual del número de respuesta válidas en la búsqueda de textos por índice de contenidos digitalizados aplicando el algoritmo de reconocimiento de patrones en imágenes digitales es 25,5688; llegando a un incremento máximo de 75,0.

  Cuadro de resultados de los valores arrojados en el Minitab del objetivo general



Boxplot of % injcremento

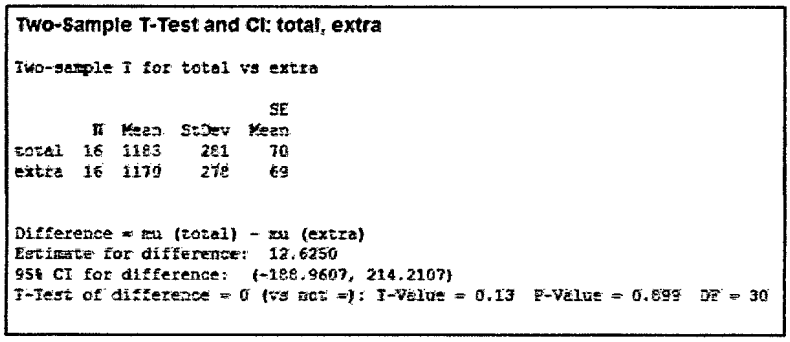
One-Sample T: % incremento

Test of mu = 0 vs > 0

Variable	N	Mean	StDev	SE Mean	95% Lower Bound	T	P
% incremento	16	25.5688	26.1710	6.5428	14.0990	3.91	0.001

Figura N° 24: Muestra de resultados de T –student para el objetivo general

- Según los resultados estadísticos se afirmó que el promedio de caracteres extraídos menos el promedio de caracteres totales de una imagen es igual a 0.899 por lo tanto se afirmó con un nivel de confianza estadístico del 95% que la extracción de caracteres en una imagen se extrajo un equivalente al total de los caracteres en una imagen.

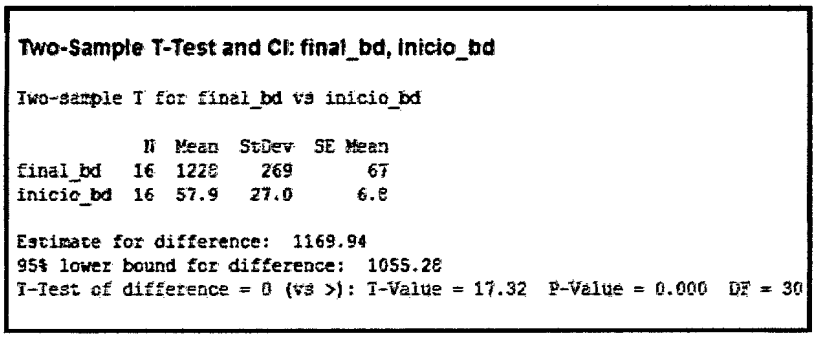


```
Two-Sample T-Test and CI: total, extra

Two-sample T for total vs extra

                      SE
       N  Mean  StDev  Mean
total  16  1183    281    70
extra  16  1170    278    69


Difference = mu (total) - mu (extra)
Estimate for difference:  12.6250
95% CI for difference:  (-186.9607, 214.2107)
T-Test of difference = 0 (vs not =): T-Value = 0.13  P-Value = 0.899  DF = 30
```

Figura N°25: resultados de datos en Minitab con distribución t-student

- Según los resultados estadístico se afirmó que ($\mu_{final\_bd} > \mu_{inicio\_bd}$) total de caracteres existentes en la base de datos aplicando el algoritmo de reconocimiento de patrones es superior al total de caracteres existentes en la base de datos sin usar el algoritmo de reconocimiento de patrones

Cuadro de resultados de la media μ_{final_bd} y $\mu_{inicio\_bd,}$ además valores de la desviación estándar usando el Minitab



```
Two-Sample T-Test and CI: final_bd, inicio_bd

Two-sample T for final_bd vs inicio_bd

             N  Mean  StDev  SE Mean
final_bd    16  1228    269       67
inicio_bd   16  57.9   27.0      6.8

Estimate for difference:  1169.94
95% lower bound for difference:  1055.28
T-Test of difference = 0 (vs >): T-Value = 17.32  P-Value = 0.000  DF = 30
```

Figura N° 26: Análisis de resultados de la diferencia de medias para determinar el objetivo 2

CONCLUSIONES

La conclusión a lo que se llegó con los experimentos desarrollados durante la ejecución del proyecto fueron, que el porcentaje promedio de palabras extraídas es equivalente al promedio de caracteres totales de la imagen.

Según los resultados del promedio del incremento promedio porcentual del número de respuesta validas en la búsqueda de textos por índice de contenidos digitalizados usando el algoritmo de reconocimiento de patrones en imágenes digitales es superior a cero; llegando a un incremento máximo de 70,0 y un promedio de 25,5688 de respuestas.

Según los resultados estadísticos se afirmaron que el promedio de caracteres totales de la imagen menos el promedio de caracteres extraídos es igual a 0,899 nos indica que se extrae el equivalente de los caracteres de la imagen

Según el resultado estadístico se afirmó que el número total de caracteres existentes en la base de datos aplicando el algoritmo de reconocimiento de patrones en imágenes digitales es 1227.55 que es superior al total de caracteres existentes en la base de datos sin usar el algoritmo de reconocimiento de patrones en imágenes digitales que es de 57.9375.

Otra de las bondades a los que se llego fue que la extracción de textos de los indicies digitales solo se elabora por una única vez y se guardan en la base de datos estos textos

extraídos, para evitar consumir recursos en el servidor, y así la rapidez en la búsqueda de un material bibliográfico se mejora.

El algoritmo de reconocimiento de patrones en imágenes digitales es reutilizable en diferentes aplicaciones ya sea aplicaciones orientadas a la web o aplicación desktop y el uso de este algoritmo es estandarizado porque la extracción del texto de la imagen se desarrollo en un servicio web.

Las imágenes digítales que son los índices de los libros son guardadas en un disco duro del servidor web mas no en la base de datos, solo los textos extraídos son guardados más su dirección de ubicación en el disco de estas imágenes. Esto nos permitió no almacenar mucha data en el servidor de la base de datos.

La generación de la aplicación web desarrollada en ASPX genero códigos DLL para su protección de código fuente total y esto proporciono eficiencia en la interpretación de estos códigos fuentes porque las DLL no consumen mucho recurso del sistema a la hora de ser interpretadas por el servidor.

RECOMENDACIONES

- Para el Uso de esta Herramienta de Reconocimiento de Patrones en imágenes Digitales es necesario crear un software donde pueda ser usado, porque esto es una técnica una herramienta más que ayuda a este tipo de funciones, que comprende algoritmos genéticos, redes neuronales u otro, que este vinculado a este tipo de algoritmo.

- Se recomienda usar uno de los 3 algoritmos: algoritmos genéticos, Redes neuronales porque son los efectivos u OCR, dependiendo del tipo de investigación que se está elaborando.

- Si se usa uno de cuales quiera de estos 3 algoritmos entonces también es recomendable utilizar la herramienta de MATLAB o NETBEANS porque estas ya traen este tipo de algoritmos en sus paquetes, pero si utilizamos paquetes como Microsoft visual Studio .net ASPX C#, Visual Basic nos será un poco más complejo.

- Utilizar algoritmo de reconocimiento óptico de caracteres para extraer patrones de una imagen por ser muy potentes en las funciones que esta proporciona.

BIBLIOGRAFÍA

- ANTONIO BLASCO LÓPEZ Y FRANCISCO FÉLEZ ESTEBAN: [Diapositiva] Texto español [2007].
- CASACUBERTA FRANCISCO ENRIQUE VIDA, 1998, Reconocimiento del Habla
- CESAR A. BELTRÁN CASTAÑÓN, 2006, Reconocimiento de Patrones en Imágenes Digital.
- DAVID GARCIA PEREZ, Visión Artificial-Manejo de Imágenes Digitales.
- FERNÁNDEZ CASTRO, FERNANDO LUIS PALABRAS, 3-May-2010, Reconocimiento de patrones en imágenes digitales de cromosomas.
- HUGO SANCHEZ CARLESSI, CARLOS REYES MEZA, Metodología y Diseño en la Investigación Científica, ISBN: 9972-885-250.
- JOHN MCCARTHY (1956), DARTMOUTH COLLEGE, IA (teoría de autómatas, redes neuronales e inteligencia).
- JOSÉ LUIS ALBA y JESÚS CID. [Diapositiva] Universidad de Vigo, Universidad Carlos III; texto en español. [Mayo de 2006].
- KENNET C. LAUDON, JANE P. LAUDON, (México, 2008), Sistemas de información general.
- OSCAR GONZALES MORENO (ANAYA-2008), Guía práctica para usuarios ASP.NET.

- PEARSON EDUCACION, Administración de la empresa digital-décima edición ISBN: 978-970-26-1191-2.
- RAÚL PINO GOTUZZO, Metodología de la Investigación, San Marcos.
- ROBERTO HERNÁNDEZ SAMPIERI, CARLOS FERNANDEZ COLLADO, PILAR BAPTISTI LUCIO, Metodología de la Investigación.
- ROGER S. PRESSMAN (2002) adaptado por DARREL INCE, Ingeniería del software Un enfoque práctico, Quinta edición.
- ROGER S. PRESSMAN, Ing. Web Ingeniería de software- 6th EdMcGraw-Hill.DERECHOS RESERVADOS, respecto a la 6ta edición en español, porMcGRAW-HILLDNTERAMERICANA DE ESPANA, *S.* A. U. Edificio Valrealty, l.a planta Basauri, 17 ,28023 Aravaca (Madrid).
- SANCHEZ CAMPO EDGAR NELSON, 2001, Redes Neuronales.
- TE-HSIU SUN, HORNG-CHYI HORNG, CHI-SHUAN LIU, FANG-CHIN TIEN,2009, Algoritmo para reconocimiento de patrones y búsqueda de imágenes.
- W. ROLSTON DAVID, 2001, Inteligencia Artificial y Sistemas Expertos.

ANEXOS

CUADRO DE PRESUPUESTO DE BIENES

CUADRO DE RESUMEN DE COSTOS	
Descripción	Precio Total (nuevos soles)
Costo por Equipo de Desarrollo	14,520.00
Costos en el Proceso de Desarrollo. (89,932% costo en Licencia de SW)	4,271.00
Costos de Implantación del Sistema Integral (100% costo en licencia de SW)	4,557.15
TOTAL INVERSIÓN S/.	23348.15
IGV (18%)	4202.667
COSTO TOTAL DE LA APLICACIÓN	27550.817

Tabla N° 6: Cuadro de resumen de costo en bienes

DESCRIPCIÓN DE GASTOS EN ELABORACIÓN DEL PROYECTO DE TESIS

Descripción	*Unidad*	*Cantidad*	*costo/unid*	*Total*
Equipos y materiales				
Bibliografía e información		1	100	100
Discos flexibles			10	0
CDs	Caja	1	10	10
Plumones de pizarra	Unidad	10	9	90
Papel Bond 60 gr.	Unidad	3	44	132
Papel Bond 80 gr.	Millar	2	66	132
Cuaderno de 100 h.	Millar	3	3	9
Revistas	Unidad	2	50	100
Lapiceros	Unidad	10	0,5	5
Computadora	Unidad	1	2400	2400
USB 4GB	Unidad	1		0
Impresora	Unidad	1	350	350
Imprevistos (10% Materiales)				822,8
Servicios				
Digitación			200	200
Impresión			300	300
Empastado			100	100
Anillados			50	50
Imprevistos (10%Servicios)				65
Otros				
Movilidad			100	100
Investigación en Internet			300	500
Total				3065,8

Tabla N° 7: Descripción de gastos en formulación del proyecto de tesis