Capítulo 6

Evaluación de algoritmos

La evaluación de los algoritmos propuestos en el capítulo 2 (Soundex, Similarex, Clases de Caracteres), se realizó en base a las características del texto previamente reconocido con OCR. De ahí que se utilizaran algoritmos que trabajaran bajo este contexto, en donde la información presentaría cierto grado de error, para el cual era necesario establecer criterios y parámetros de evaluación, aspecto al cual se dedica este capítulo.

El material utilizado como indicamos en capítulos anteriores, fue material bibliográfico perteneciente al acervo franciscano, colección que está constituida por libros que datan aproximadamente del siglo XVI a XIX y cuyas características iban a propiciar la existencia de errores en cuanto al reconocimiento de caracteres se refiere.

Como muestra se tomaron 6 libros de esta colección, de los cuales se digitalizaron aproximadamente 25 páginas por unidad(libro). El tamaño de dicha muestra pudiera parecer pequeño, sin embargo, fue suficiente para definir la eficiencia de cada algoritmo, ya que estos trabajan a nivel de error en la información, no a nivel de volumen o cantidad, debido a sus características.

A su vez, se analizó el texto generado por OCRde cada página para obtener una lista de palabras (150) que presentaban errores por reconocimiento y que nos permitirían realizar las pruebas de estos algoritmos. También fue necesario corregir cada una de las páginas de la muestra para así poder calcular el número de ocurrencias reales de este conjunto de palabras en la misma, valor que nos serviría posteriormente para el cálculo de los porcentajes de precisión del apartado 6.1.

6.1 Parámetros de evaluación

Para poderdeterminar la eficiencia de dichosalgoritmos,fue necesario establecerciertos parámetros de evaluación. En este caso nuestro parámetro fue la precisión, que es la exactitud en la localización y recuperación del texto buscado en la información existente.

Debido a la escasez en la información con respecto a estos algoritmos, no fue posible localizar parámetros de evaluación preestablecidos por lo cual se proponen los siguientes.

-porcentaje de información correcta recuperada

-porcentaje de información relevante en resultados

-promedio general.

6.1.1 Porcentaje de información correcta recuperada

Por porcentaje de información correcta recuperada nos referimos a que parte del total de ocurrencias existentes de una palabra en la muestra fueron obtenidas por el algoritmo.

Para calcular este porcentaje tomamos como referencia el número de ocurrencias correctas localizadas por el algoritmoy lo dividimos entre el numero total de ocurrencias reales existentes en la muestra.

De esta manera tenemos que:

$$PICR = \frac{OCL}{NTOR} X100$$

donde:

PICR: porcentaje de información correcta recuperada

OCL: ocurrencias correctas localizadas por algoritmo

NTOR: número total de ocurrencias reales

Ejemplo:

Palabra buscada: puntos

Ubicación real:

libro 1 página 5

libro 3 página 163

libro 4 página 145

Tabla 6.1 Ejemplo de ocurrencias recuperadas poralgoritmo

Palabra encontrada	Libro en la que se encontró	Número de página
puntos	libro 4	145
apunto	libro 1	201
puntos	libro 1	5
punto	libro 3	144

En este ejemplo la información correcta recuperada serán solamente las ocurrencias que correspondan exactamente a la palabra puntos (Tabla 6.1), así nuestro PICR para el mismo queda de la siguiente manera:

OCL = 2

NTOR = 3

Por lo tantoPICR = (2/3) * 100 =66.66%.

Esto significaría que de 100 ocurrencias reales que existan de una palabra en el acervo, este algoritmo nos estaría encontrando solamente 66 .

6.1.2 Porcentaje de información relevante en resultados

Entiéndase por información relevante aquella que presenta una correspondencia exacta con la palabra buscada con respecto al total de la información recuperada por el

algoritmo. Para calcular este porcentajese consideró el número de ocurrencias correctamente localizadas por el algoritmo, dividido entre el número total de ocurrencias recuperadas por el mismo,así:

$$PIR = \frac{OCL}{NTO} X100$$

Donde:

PIR: porcentaje de información relevante

NTO: número total de ocurrencias recuperadas en la búsqueda.

OCL: ocurrencias correctas localizadas por algoritmo

Continuando con el ejemplo de la sección6.1.1la información relevante se refiere alas ocurrencias que correspondan exactamente a la palabra puntos (Tabla 6.1). De esta manera el PIR sería:

OCL = 2

NTO = 4

Por lo tanto PIR = (2/4) * 100 = 50%

Este resultado implicaría entonces que de 100 ocurrencias devueltas por la búsqueda, solamente 50 corresponde exactamente a la palabra buscada.

6.1.3 Promedio general

Debido a que el porcentaje deinformación recuperada representa qué tanto de la información existente en la muestra se va a recuperar, y a que el porcentaje de información relevante nos dice de los resultados obtenidos por el algoritmo qué porcentaje corresponde a la búsqueda original, ambos se encuentran ligados debido a que un algoritmo que obtenga un porcentaje alto de información recuperada, pero un bajo porcentaje de información relevante no resulta práctico, ya que aunque se localice la mayor parte de la información

existente, esta se encontraría dentro de una gran cantidad de información incorrecta, por ejemplo, si un algoritmo obtiene un porcentaje de información recuperada de 100% y un porcentaje de información relevante recuperada de 10%significaría que de 100 ocurrencias existentes de una palabra se recuperarían las 100,pero habría 900 ocurrencias incorrectas más,por lo que la eficiencia general del algoritmo no seria muy buena. En base a lo anterior se promediaran ambos porcentajes para obtener un porcentaje general de eficiencia.

6.2Metodología de evaluación

Para poder evaluar los parámetros anteriormente mencionados necesitábamos establecer una metodología acorde a la naturaleza de loserrores en la información almacenada, por lo cual se estimó necesario realizar las pruebas en base a los siguientes criterios

- ubicación del error en una palabra
- longitud de palabra buscada
- localización de palabra con error específico

6.2.1 Ubicación del error en una palabra

Uno de los criterios de evaluación establecidos se refiere a la ubicación que presenta el error dentro de una palabra.

Para esto fue necesario realizar un análisis del texto reconocido para obtener palabras claves que sirvieran para las pruebas en la búsqueda de información. Dichas palabras se eligieron y clasificaron en las siguientes categorías:

- palabras que presentaron errores al inicio
- palabras que tuvieron errores en su punto medio
- palabras que presentaron errores al final

Entiéndase como error al inicio de una palabra cuando este se ubica dentro de su primera sílaba. Un error al final por lo tanto corresponderá al que se ubica en la última sílaba. Por último el error en el punto medio es el que no se encuentra en ninguno de los dos casos anteriores.

Este criterio se consideró importante debido aque es muy probable que una palabra que tenga error al inicio sea más difícil de localizar por la forma de codificación de los algoritmos.

6.2.2 Longitud de palabras buscadas

Para poder efectuar la prueba en base al tamaño de una palabra respecto a su longitud, fue necesario primeramente establecer el número mínimo y máximo permisible en la longitud de una cadena. Esto con la finalidad de aislar aquellaspalabras que no tuvieran relevancia en la búsqueda,así como de considerar la mayor cantidad de palabras existentes en nuestro lenguaje, para lo cual se generó la siguiente tabla.

Tabla 6.2 Longitud de palabra establecidos

Longitud de palabra
5
6
7
8
9
10
11

Debido a que una palabra de longitud pequeña puede presentar una probabilidad mayor de pérdida de información en comparación con una de tamañomás grande por la confusión durante el reconocimiento, seconsideró importante basar parte de las pruebas en este criterio

6.2.3 Localización de palabra con error específico

Para esta prueba se seleccionaron palabras del texto reconocido que presentabanerror en más de una sílaba. A diferencia de las dos evaluaciones anteriores, en esta sólo se tomó en cuenta si se encontraba o no la palabraen la ubicación original de donde se seleccionó, no tomando en cuenta cualquier otra ocurrencia de la misma dentro del acervo.

Finalmente, las pruebas se realizaron empleando el software desarrollado y la base de datos que contenía la información de los libros utilizados,que consistieron básicamente en la búsqueda de información a nivel de contenido.

En la siguiente sección se mostrarán los resultados de esta evaluación así como las conclusiones obtenidas de la misma.

6.3 Evaluación de los algoritmos propuestos

En base a las pruebasy parámetros establecidos en los apartados anteriores, se llegó a los resultados que se describirán a continuación. Los parámetros de evaluación PICR, PIR,sobre los cuales se obtuvieron los promedios generales mostrados en las Tablas 6.3 y 6.4 se muestran en el Apéndice B.

6.3.1 Resultados de pruebas por longitud de palabra buscada

En estos resultados se puede apreciar que mientras más larga es la palabra buscada, existe mayor probabilidad de localizarla. Estodebido aque el porcentaje depérdida de informacióna causa del reconocimiento óptico, en una palabra deeste tipo, es menor que una de longitud más corta. En la Tabla 6.3 se muestran los promedios generales de cada algoritmo para dicha prueba, los cuales muestran una relación directa entre la longitud de palabra y el valor porcentual, por lo que podemos concluir de esta pruebaque los algoritmos podrán recuperar mejor aquellas palabras que presenten una menor pérdida de información por reconocimiento, teniendo mayor probabilidad aquellas que sean de mayor longitud.

Tabla 6.3 Promedios generales(longitud de palabra)

Longitud de palabra	Clases de Caracteres	Similarex	Soundex
5	39.16	22.96	24.39
6	46.98	20.60	22.32
7	52.99	18.71	31.07
8	48.36	20.86	27.53
9	58.53	20.15	33.33
10	48.56	16.21	30.35
11	59.03	36.81	48.50
Promedio	**50.51**	**22.33**	**31.07**

6.3.2 Resultados de pruebas por ubicación del error en una palabra

A diferencia de la prueba 6.3.1, en los resultados mostrados en la Tabla 6.4 notamos que no hay una diferencia significativa entre la ubicación del error y los resultados, a pesar de quepor la forma de codificación de los algoritmos podría ser factible que los porcentajes obtenidos para las palabras con error al inicio fueran menores, tal como se supuso en secciones anteriores al establecer este criterio de evaluación. Por lo tanto de esta prueba podemos concluir que la ubicación del error en una palabra no aumenta o disminuye la probabilidad de que los algoritmos encuentren este tipo de información.

Tabla 6.4 Promedios generales (ubicación del error)

Ubicación del error	Clases de caracteres	Similarex	Soundex
Inicio	45.81	21.66	25.66
Medio	51.86	20.58	25.77
Fin	44.29	19.83	25.86
Promedio	**47.32**	**20.69**	**25.76**

Finalmente, como se muestra en la tabla 6.5, en base a los resultados obtenido de los porcentajes PICR y PIR que involucran tanto la longitud como la ubicación del error en una palabra, se concluye que el ?Clases de Caracteres? tiene el mejor desempeño promedio. Cabe hacer mención que el ?Similarex?, obtuvo un valor alto en el porcentaje de información recuperada y un valor muy bajo en el porcentaje de información relevante (ver

apéndice B) lo cual se debe a su codificación, ya que aunque obtiene una cantidad alta de la información buscada también devuelve una gran cantidad de información errónea.

Tabla 6.5 Promedio de resultados prueba 6.3, 6.4

Algoritmo	Soundex	Similarex	Clases de Caracteres
Eficiencia general	28.41	21.50	48.91

6.3.3 Resultados de pruebasen localización de palabra con error específico

En cuanto a la prueba de localización de palabra con error específico, podemos concluir, según los resultados mostrados en la Tabla 6.6 , que el ?Similarex? presenta mayor capacidad para recuperar información de este tipo con respecto a los otros dos algoritmos, en esta prueba únicamente, donde no se aplica el PICR y el PIR, debido a queesta prueba se basa exclusivamente en encontrar la palabra en una ubicación específica sin considerar cualquier otra ocurrencia de la misma. Algunos ejemplos de palabras utilizadas para esta prueba se pueden ver en el Apéndice C.

Tabla 6.6 Porcentajes de eficiencia según error específico

Algoritmo	Porcentaje palabra con error
Soundex	31.79
Similarex	41.06
Clases de Caracteres	27.15

6.4 Conclusionesen la evaluación de los algoritmos

Finalmente, para concluir, podemos decir que los tres algoritmos se encuentran relativamente cercanos en su eficiencia, sin embargo, en base a los resultados generales,el "Clases de Caracteres" presenta los resultados más altos por lo que se concluye que de estos algoritmos y para este tipo de información , el que tiene mejor desempeño es el "Clases de Caracteres".

Independientemente de la forma en que se evaluó, se pudo observar que el Soundex puede ser útil cuando los errores por parte del usuario fueran ortográficos así como en el caso en que el autor deun libro hubiese escrito cierta palabra de una manera distinta más no incorrecta como es el caso del los libros del acervo franciscano, en el que se pudo observar que muchas palabras se escribían de tal formaque en la actualidad pudieran parecer faltas de ortografía. Por ejemplo, palabras que normalmente escribiríamos con g, j, en la época en la que se ubican los libros de esta colección se escribían con "j" y ?x? respectivamente . Algunos ejemplos se pueden observar en la Tabla 6.7.

Tabla 6.7 Ejemplos de palabras

Palabra con ortografía actual	**Palabra con ortografía siglo XVI-XIX**
tragedia	trajedia
ejemplo	exemplo
juicios	juizios
hacían	hazian
aire	ayre
vela	bela

El ?Similarex?, a pesar de que recupera mucha información que no corresponde uno uno con la palabra buscada, gran parte de esa información presenta relación con la misma, por lo cual podría ser útil en el caso en que un usuario estuviese interesado en buscar datos que presentaran cierta similitud con el dato buscado.