Capítulo 2

Metodologías de consulta y preservación

En este capítulo mostramos algunas iniciativas y proyectos que se han realizado en bibliotecas digitales, abordando aspectos de preservación y consulta de material bibliográfico antiguo principalmente.A su vez, se revisan algunos de los métodos existentes en cuanto a la búsqueda y recuperación de información y se analizan aquellos que son de utilidad para este proyecto.

2.1 Trabajo relacionado

Existen diversas asociaciones e institucionesnacionales e internacionales que se han dado a la labor de desarrollar proyectos dentro del área de preservaciónde documentos, que tienen relevancia dentro de diversos contextos de trabajo.

La tecnología digital ha jugado un papel muy importante en esta labor, característica que destaca en algunos delos proyectos que describiremos brevemente en esta sección.

2. 1.1 Preservación en bibliotecas digitales

Las bibliotecas digitales han ofrecido en los últimos años grandes ventajas en diversos aspectos, principalmente en lo que se refiere a preservación de documentos de gran valor cultural e histórico.

La sociedad está profundamente interesada en conservar de alguna u otra manera materiales que documentan hechos, ideas, acontecimientos que han pasado de generación en generación a formar parte de la herencia cultural de cada país, permitiéndole comprender no solamente la época actual en la que se desenvuelve, sino el pasado sobre el cual están sentadas sus bases.

El problema existente entre la preservación y acceso a la información contenida en documentos principalmente en papel, se ve disminuido con la utilización de esta tecnología, ya que esta permite realizar operaciones diversas sobre réplicas representativas del documento original, evitando así el desgaste o la pérdida del mismo[Lesk1997].

Las necesidades en cuanto a preservación de documentos se refiere, varían de acuerdo a diversos factores como pueden ser la antigüedad, la composición de los materiales que los constituyen y el contexto en el que se encuentran, es decir, las condiciones ambientales que los afectan como puede ser la humedad, el calor, las malas condiciones de almacenamiento,así como el inadecuado manejo por parte de los usuarios [Reed-Scott 1999] .

Uno de los medios de documentación más comunes y más propensos al desgaste y pérdida es el papel, y por lo mismo es el centro de atención en cuanto a necesidades de preservación se refiere.

Por ejemplo, en recientes valoraciones en La Biblioteca del Congreso, se estimó que cerca de 77,000 libros anualmente pasan al estado de "frágil", mientras que en las colecciones de la Universidad de Yale 12% ya requieren reparación inmediata así como el 87% ya se encuentran en estado delicado [Reed-Scott 1999] .

Diversas asociaciones y organizaciones nacionales, instituciones y bibliotecas tanto nacionales como internacionales, se han dado a la labor de conjuntar esfuerzos para lograr la preservación de material bibliográfico, así como en otros formatos como publicaciones periódicas o de investigación.

Sin embargo, dichos esfuerzos se han enfocado más aún en libros que presentan un alto grado de fragilidad, muchos de los cuales son piezas únicas.

Dentro de lasasociaciones y organizaciones internacionales mencionaremos a la ARL (Association of Research Libraries) asociación cuya finalidad es abogar y fomentar programas de preservación principalmente en bibliotecas de Norteamérica (http://www.arl.org).

La NHA (National Humanities Alliance), organización que se enfoca a los avances de la humanidad y sus intereses comunes con respecto a políticas nacionales, programas y legislaciones que la afectan(http://www.nhalliance.org/).

En cuanto a bibliotecas se refiere tenemos aLa Biblioteca del Congreso, la cual ha destacado en preservación de material antiguo diverso entre los que se cuentan libros, manuscritos, enciclopedias, mapas, documentos gubernamentales entre otros, especialmente ricos en documentación de la historiaamericana
(http://lcweb2.loc.gov/ammem).

Se han generado a su vez, diversos proyectos que involucran la preservación de acervos de gran valor cultural e histórico a través de bibliotecas digitales. Uno de ellos es el proyecto denominado AGIcuyas iniciales hacen referencia al Archivo General de Indias, que constituye el más grande depósito de documentación española de más de tres siglos y cuyos orígenes datan delaño de 1785(ver http://www.mcu.es/lab/archivos/AGI.html).

Dicho Archivo está formado por 43,000 legajos concerca de80 millones de páginas originales que incluyen una gran diversidad de temascomo descubrimiento, exploración y conquista del nuevo mundo, la expansión misionera, aspectos inquisitoriales, entre otros [Gonzales 1998].

Una de las características principales en este proyecto fue el empleo de la tecnología digital comomedio para lograr en cierta medida la preservación del Archivo, así como de metodologías para el acceso y consulta al mismo.

En cuanto a instituciones nacionales involucrados en proyectos de esta naturaleza mencionaremoslas siguientes:

UNAM: (Universidad Nacional Autónoma de México) que ha desarrollado colecciones digitales de libros de índole general, principalmente para contribuir enactividades académicas y de investigación (http://www.dgbiblio.unam.mx/).

BNAH: (Biblioteca Nacional de Antropología e Historia) que apoya actividades de investigación, conservación y restauración, docencia y difusión de la cultura en conjunto con el INAH.

UDLA: (Universidad de las Américas, Puebla) en la que actualmente se desarrollan proyectos que involucran la utilización de la tecnología digital en bibliotecas y la preservación de materialimpreso de gran valor, dentro de los cuales se ubica esta tesis.

Finalmente, como podemos observar, la preservación de información contenida en papel es de gran importancia para la humanidad a diversos niveles y en diversos contextos. Para lograrlo, es indispensable la utilización de la tecnología digitalque permita no solo conservar datos o información, sino que ofrezca a los usuarios de este mediolas capacidades necesarias para descubrir la riqueza contenidaen la expresión escrita.

Una de esas capacidades es la provisión de mecanismos que permitan recuperar la información de un documento digital a través del empleo de metodologías ytécnicas de un sistema de búsqueda, aspecto que se analiza a continuación.

2.2 Recuperación de información en formato digital

Para lograr la preservación de medios de documentación como el papel, es necesario pasar de un formato físico a un formato digital que permita posteriormente extraer la información contenida en este para su utilización posterior. Esta información puede ser de dos tipos principalmente: gráfica, textualo una combinación de las dos.

Sin embargo, como mencionamos al inicio de esta sección, el contenido de documentos digitales puede ser puramente textual, a lo que algunos autores denominanimágenes textuales[Witten 1999].

Para poder manipular de alguna manera la información contenida en este tipo de imágenes, es necesario haberlarecuperado previamente con algún método o técnica específica. Una tecnología que permite la obtención automática del texto es la del reconocimiento óptico de caracteres, mejor conocida como OCR,herramientaque normalmente ya viene integrada en los dispositivos de digitalización comerciales.

La utilización de tecnología de OCR es necesaria y crucial, principalmente cuando se tiene una inmensa cantidad de documentos que es prácticamente imposible transcribir manualmente, y sobre los cuales ya se requiere la recuperación de información [Gonzales 1988].

El hecho de aplicar tecnología de reconocimiento de caracteres (OCR), implica enfrentarse a rangos de error que pueden variar según la calidad del documento digitalizado. Muchas veces dichos documentos no presentan las características ideales como pueden ser buena calidad del papel, de la impresión, considerando a su vez que existen diversos formatos de letras para los cuales no hay un OCR específico.

El empleo de revisores de ortografía sobre el texto reconocido puede sonar tentador, sin embargo, estos se basan en diccionarios de palabras válidas.Muchas colecciones de documentos contienen un extenso vocabulario que no se encuentra en diccionarios comunes como nombres propios, acrónimos, palabras de lenguas extranjeras, términos esotéricos y técnicos entre otros [Hawking 1996].

Para este proyecto,emplearemos precisamenteun digitalizador comercial con OCR integrado para poderrecuperar el texto de una imagen digital, que posteriormente

utilizaremos para realizar búsquedasde información a nivel de contenido de dichos documentos.

Partiendo de este hecho, debemos considerar que la información recuperada presentará cierto grado de error en cuanto a reconocimiento se refiere. Es por ello que nos enfocaremos en la siguiente sección a analizar métodos que contemplan dentro de su funcionamiento este tipo de errores.

2.3. Métodos de búsqueda en texto degradado

Para la búsqueda y recuperación de información de un documento, existen diversos métodos, aquellos que se basan en la búsqueda de cadenas por similitud y que no contemplan errores de escritura y los que aparte de la primera característica si consideran este tipo de errores, principalmente aquellos generados por OCR.

Dentro de las técnicas que no contemplan errores de escritura tenemos latradicional, la cual se caracteriza por la búsqueda de palabras completas con variantes como:

- al inicio de una cadena
- al final de una cadena
- dentro de una cadena
- correspondencia exacta

Sin embargo, como ya mencionamos, este método requiere que la información este escritacorrectamente.

Ahora bien, la contraparte son losmétodos que si contemplan errores de escritura. El empleo de *n-grams* , es un ejemplo característico de búsquedas basadas en la similitud de

cadenas tomando en cuenta errores de escritura, que consisteen el particionamiento de cada palabra del texto en secuencias de caracteres de diversa longitud.

Ejemplo:

escuela:"escuela" y cuyos *n-grams* de longitud dos son:

"e", "es", "sc", "cu", "ue", "el", "la", "a"

Las longitudes de particionamiento pueden variar, sin embargo, algunas presentan ciertasdesventajas. Por ejemplo, en el caso de *n-grams* de longitud "uno", implica tener una gran cantidad de comparaciones entre palabras, las de longitud "cuatro" muchas veces no permitendetectar las raíces comunes de palabras cortas. Es por ello que normalmente se emplean de dos y de tres*n-grams* para calcular las similitudes entrepalabras[Salton 1989].

Sin embargo, en comparación con otros métodos, los *n-grams* son calculados en tiempo de ejecución, lo cual implica mayor tiempo de procesamiento para recuperar información en sistemas de búsqueda.

Otros métodos son aquellosbasados en formas canónicasque consisten en empatarlas diferentes formas erróneas de una palabra determinadaa una forma común [Myka y Güntzer1995], de tal manera que tanto la palabra original como sus diversas "versiones" son agrupadas dentro del mismoconjunto, permitiendo con ello aumentar en cierto grado la probabilidad de recuperar la información que el usuario busca, a pesar de que el dato solicitado presente errores de escritura.

Estos métodos representan grandes ventajas debido a la eficiencia en cuanto a velocidad se refiere,ya que las formas canónicas se van generando al momento de llenar la base de datosque las contendrá, por lo tanto el único proceso en tiempo de ejecución de un sistema de consulta, se reduce a generar únicamente la forma canónica de la solicitud del

usuario. La complejidad aproximada de algunos de estos métodoses de O(logN) [Myka y Güntzer1995].

Existen otros métodos canónicos que no precisamente están basados en recuperación de texto con errores de escritura por reconocimiento, como es el caso del ?Soundex?, algoritmo que se basa en la similitud fonética de las palabras. Sin embargo es considerado dentro de este bloque ya que pueden manejar erroresfonéticos que implican en cierto modo la escritura incorrecta de una palabra.

Finalmente, cabe mencionar que los métodos canónicos aquí descritos producen una codificación intermedia que posteriormente es utilizada para la búsqueda y recuperación de información.

A continuación describiremos tres de estos métodos los cuales serán utilizados en este proyecto de tesis.

2.4 Soundex

El algoritmo ?Soundex? fue desarrollado por Margaret K. Odell y Robert C. Russell. Su principal característica es la codificación basada en la similitud fonética de las palabras más que en su ortografía para reducirlas a una forma común. Fue utilizado principalmente en aplicaciones que involucraban búsquedas de nombres de personas como en sistemas de reservación aérea, censos, y otras que presentaban problemas en cuanto a errores en la escrituradebido a la similitud fonética. [Knuth 1975; Myka y Güntzer1995]

El método ?Soundex?, permite reducir a una forma común aquellas palabras que son similares en cuanto a su pronunciación,haciendo más sencilla la comparación de una palabra con otra ya que lo que se almacena es el código común generado, en vez de todas las palabras de un texto.

Este algoritmo es dependiente del lenguaje utilizado, originalmente fue desarrollado para el idioma inglés, lo que implica que es necesario realizar modificaciones de acuerdo al idioma que se va a emplear, ya que es preciso agrupar por similitud fonética ciertas letras en diferentes grupos o clases [Pfeifer et al.1995].

Por sus características y las de nuestro idioma se consideró relevante su uso en este proyecto.

2.4.1 Codificación

El código soundex[Myka y Güntzer1995; Salton 1989; Rosen 1994] está formado por la primera letra de la palabra a codificar, concatenada con un valor numérico que se genera a partir del reemplazo de las tres letras siguientes por algún valorpreestablecido.

Las ocurrencias de vocales (a, e, i, o, u) y las letras W, H, Y en la palabra son eliminadas, con excepción de la primera letra.

Como ejemplo tomemos la palabra Smith, que presenta similitud con la palabra Smyth.

Se retiene la primera letra de cada palabra y se eliminan todas las ocurrencias de vocales y las consonantes w, h, y.

Smith → Smt Smyth → Smt

Se asignan a las siguientes 3 consonantes un valor representativo del grupo al que pertenecen dichas letras,como se muestra enla Tabla 2.1, con algunas excepciones que se listarán más adelante, y se elimina el resto de los caracteres después de dicha asignación.

Tabla 2.1 Equivalencias fonéticas de caracteres (inglés)[Myka y Güntzer1995]

Valor	**Términos similares fonéticamente**						
1	b	f	P	v			
2	c	g	J	k	q	s	x

3	d	t					
4	l						
5	m	n					
6	r						

Después del reemplazo obtenemos el código soundexS-530 para ambas palabras.

2.4.2 Excepciones

Algunas excepciones a estas reglas se listan a continuación:

Si no hay 3 consonantes seguidas después de la primera letra utilizar ceros para completar el código de tres dígitos.

Si la palabra tienedos o más letras consecutivas con el mismo valor numérico,la codificación se debe realizar sobre una sola de esas letras, la otra se elimina.

Ejemplo: Gutiérrez:G-362 (la segunda r no se considera dentro del código).

Si la palabra está constituida por letras consecutivas que reciben el mismo código soundex como en Jackson donde k y s =2, estas deberán ser tratadas como una sola letra, entonces únicamente se codificará 2 una sola vez.

Esta regla también aplica a la primera letra de la palabra, por ejemplo en Pfister, p y f reciben el mismo código (1) sin embargo la f es descartada quedando únicamente P-236 como código.

Una vez hecha la codificación para cada palabra en el texto, la búsqueda de información se realiza sobre estos valores, que finalmente arrojará resultados con las siguientes características:

- Las palabras son idénticas.
- Las palabras comparadas tienen el mismo código soundex pero son diferentes.
- Las palabras son totalmente diferentes [Pfeifer et al.1995].

2.5 Clases de Caracteres

El método denominado "Clases de Caracteres", se basa en la codificación de palabras parcialmente reconocidas o que presentan errores que fueron generados al momento de realizar el reconocimiento con tecnología OCR.

Se utiliza cuando existe confusión de ciertos caracteres con otros. Existen clases o grupos de caracteres que tienen asociado un elemento denominado canónico y que es representativo de cada clase o grupo.

Para generar los grupos es necesario estableceruna relación (x, y) que puede ser definida de acuerdo a una tabla de equivalencias o matriz de confusión, generada a partir del análisis del texto reconocido para determinar los caracteres que presentan una probabilidad alta de confundirse con otros[Myka y Güntzer1995].

Finalmente se define un elemento representativo de cada conjunto o elemento canónico para posteriormente utilizarlo en la generación del código

2.5.1 Codificación

Para cada palabra en el texto se reemplazan todas las ocurrencias de los caracteres que están listados en la Tabla 2.2 por su elemento canónico, lo cual dará origen a los códigos que serán utilizados para la búsqueda de información.

Tabla 2.2 Tabla de confusión original(Clases de Caracteres)
[Myka y Güntzer1995]

Elemento canónico	**Caracteres de confusión**				
e	e	c	a	s	
1	1	1	I	I	
O	O	C	o	0	D
f	f	t			
y	y	v	V		

M	M	N	H	M	
S	S	5			
g	g	q			
h	h	b			
u	u	n			
F	F	E			
K	K	k			

2.6 Similarex

El método denominado "Similarex", contempla dentro de su codificación aquellos errores generados por el uso detecnología OCR. A diferencia del ?Clases de Caracteres? considera otro tipo de erroresen base a patrones de caracteres que pueden aparentar o confundirse con ciertas letras.

Al igual que el ?Soundex?, el ?Similarex? elimina información innecesaria en su codificación como las vocales. Por otro lado, trata de empatar aquellas partes de la información más confusas con respecto a una forma común [Myka y Güntzer1995] .

2.6.1 Codificación

Para poder generar la codificación correspondiente a cada palabra en un texto utilizando el método ?Similarex?, es necesario primeramente analizar las secuencias de caracteres de izquierda a derecha haciendo un seguimiento de secuencias de longitud tres, esto con la finalidad de reemplazar aquellas ocurrencias que presenten similitud con algunos caracteres como se muestra en el ejemplo siguiente, donde tres letras ?i? normalmente son confundidas con la letra ?M? .

Ejemplo:

iii, iin, nii--> M

Posteriormente realizar el paso anterior pero para secuencias de caracteres de longitud dos:

Ejemplo:

in, iu, ri, rI, rn, ni, ui, tn -- > M

ti-- > n

Se hace el reemplazo de caracteres individuales en base a la Tabla 2.3, la cual previamente fue generada a través del análisis del texto reconocido para definir los grupos, de acuerdo a la confusión existente entre caracteres.

Al igual que los dos algoritmos anteriores es necesario establecer un elemento canónico para poder generar el código final.

Tabla 2.3 Tabla de confusiónoriginal (Similarex)
[Myka y Güntzer1995]

Elemento canónico	**Caracteres de confusión**			
n	U	h	b	
M	M	N	H	
I	l	l	i	r
O	o	0	D	C
Y	v	V	y	
f	t			
S	5			
g	q			
F	E			
K	k			

Finalmente , omitir todas las ocurrencias restantes de e, c, a, s.

A diferencia del ?Soundex?, el ?Similarex? se basa en la similitud en apariencia de los caracteres o patrones de caracteres. A su vez el similarex distingue mayúsculas, minúsculas y números[Myka y Güntzer1995].

La ventaja del ?Similarex? sobre el uso de ?Clases de caracteres? es que integra la confusión con respecto a una secuencia de caracteres en vez de considerar un carácter individual solamente.

De la misma forma que el de ?Clases de Caracteres?, entre más se conozca acerca de las características de confusión específicas de un dispositivo OCR, los métodos pueden ser modificados para su utilización en un ambiente específico.

Finalmente hagamos notar que la recuperación de información se hace a través de la búsqueda en los códigos generados más que en el texto mismo, una vez que ha sido codificada la solicitud de un usuario, al momento de realizar búsquedas de información.

2.7 Resumen

La preservación de material bibliográfico antiguo se ha facilitado gracias al desarrollo de la tecnología digital y su aplicación en bibliotecas.

Muchos han sido los esfuerzos realizados por diversas instituciones, organizaciones y asociaciones para proteger y salvaguardar la información valiosa impresa, considerando lagran relevancia que tiene como base histórica y como legado para futuras generaciones.

Para la búsqueda y recuperación de información tanto de imágenes como en texto, existen diversos métodos y técnicas. Para este proyecto fijaremos nuestra atención en los métodos canónicos descritos en el apartado 2.4, 2.5 y 2.6, que se basan en la codificación del texto degradado, generado previamente con tecnologíaOCR.

En el siguiente capítulo se revisan aspectos de implementación de los algoritmos pertenecientes a las secciones listadas en el párrafo anterior, su aplicación en la búsqueda y recuperación de información de colecciones especiales y diseño general del sistema propuesto.