Capítulo 1.

Introducción

La mayor parte de la información disponible a nuestros días está en papel, fotografías o videos. Para integrarlo a Bibliotecas Digitales , este gran volumen de información debe ser digitalizado en imágenes y el texto contenido en ellas convertido a ASCII, para su almacenamiento, recuperación y fácil manipulación [Wu et al. 1997] .

Dentro del proyecto Flora de China (FOC), a cargo de instituciones como el Jardín Botánico de Missouri, existe una gran colección de información en texto impreso que se pretende digitalizar. El objetivo de esto es hacer aprovechable toda esta información a mayor número de investigadores. Para hacer esto posible es necesario emplear técnicas de reconocimiento de texto sobre el papel y así facilitar al usuario la exploración de esta biblioteca digital con imágenes textuales de manera eficiente.

Este trabajo aborda el estudio de las diversas técnicas y soluciones de reconocimiento de texto en sus diferentes formas, así como sistemas de construcción de Bibliotecas Digitales a partir de documentos en texto.

Además de la investigación se plantea una solución al reconocimiento y exploración del contenido de las Tarjetas de Hu, una colección de datos en papel del proyecto Flora de China.

1.1 Bibliotecas Digitales.

Múltiples concepciones de Biblioteca Digital han surgido de acuerdo al punto de vista de quien las plantea. Para algunos, el término puede sugerir computarizar una biblioteca; probablemente la más apropiada definición es llevar todas aquellas funciones en una biblioteca de una nueva manera, añadiendo nuevos tipos de recursos de información; nuevas formas de adquisición de material, métodos de almacenamiento, conservación, catalogación, y consulta del mismo

Hay cuatro campos de estudio relacionados con las Bibliotecas Digitales [Fox et al. 1995]:

El primero, llamado *publicación electrónica* se refiere a la conversión de documentos en papel a versiones electrónicas. Se puede usar software comercial para este propósito, como Acrobat Capture de Adobe (http://www.adobe.com) , que a partir de una imagen digitalizada genera un documento en el Formato de Documentos Portable (PDF), detectando secciones, párrafos y reconociendo texto para reproducir fielmente el documento original. En este campo

entran además los estándares derivados de *Standard Generalized Markup Language* (SGML), como HTML, que permiten definir documentos con sus cualidades esenciales para su publicación tanto electrónica como en papel. También se incluye la edición de documentos electrónicos en aplicaciones especiales de esta área, como lo son los procesadores de palabras y las hojas de cálculo.

El segundo campo son los *hipermedios*. Estos dan el poder de ligar el conocimiento usando recursos como el hipertexto y así emular el concepto de conocimiento interconectado en el cerebro [Bush 1959]. Además del texto ligado son los demás medios de representación de datos y comunicación con el usuario los que hacen útil este campo. Entre estos se encuentran los modelos en dos y tres dimensiones y el sonido.

El tercero es el campo de la *educación*. La información en una biblioteca digital es más que un almacenamiento masivo. La forma de hacer accesible ésta información compilada y ordenada y las facilidades de una biblioteca digital para la comunicación entre usuarios debe fomentar un ámbito de aprendizaje.

El cuarto campo es el *manejo de datos e información*. La información almacenada en una biblioteca digital puede incluir datos geográficos, datos comprimidos e información en múltiples medios (sonido, video). En este campo se incluyen las técnicas de manejo de bases de datos, ya sean relacionales u orientadas a objetos. Es en ésta última área en la que el análisis de texto y la recuperación de información son cruciales para la conversión, indexación, representación, búsqueda y presentación de la información.

Los beneficios de la construción de bibliotecas digitales son muchos, entre ellos está el de promover el aprendizaje informal, es decir, si la información está más cerca del que quiere aprenderla no es necesario acudir a bibliotecas físicamente ni tomar cátedras [Marchionini & Maurer 1995]. El aprendizaje así se convertiría cada vez menos en un privilegio de algunas clases sociales.

Es fácil suponer que en un futuro cercano las diferentes bibliotecas digitales desarrolladas en distintas partes del mundo convergerán para aprovechar la información y servicios entre ellas. Por ello, es también necesario crear estándares de interoperabilidad entre los servicios de las mismas sin quitarles su tinte de heterogeneidad [Paeckpe et al. 1998].

1.1.1 Biblioteca Digital Florística

La Biblioteca Digital Florística (FDL) es un proyecto auspiciado por la Fundación Nacional para la Ciencia (NSF) de Estados Unidos, cuyo objetivo principal es ofrecer un vasto número de servicios para facilitar el uso y distribución de un amplio acervo de descripciones, discusiones, imágenes, ilustraciones y mapas relacionados con plantas

[Sánchez & Ayala 1998].

La FDL comprende la información generada por los proyectos Flora de China y Flora de Norteamérica. En el desarrollo de éste proyecto participan también otros proyectos de investigación florística. Una comunidad considerable de investigadores están incluidos dentro del proyecto. Sólo en FNA, por ejemplo, participan más de 800 científicos, entre autores, editores y coordinadores.

El diseño de la FDL propuesto por Sánchez & Schnase [1998] (figura 1.1), incluye los componentes necesarios para la construcción de la Biblioteca y para el uso de la misma. Se puede ver un conjunto de servicios construídos sobre un repositorio distribuido de objetos en forma electrónica como descripciones textuales, referencias bibliográficas, mapas e ilustraciones.

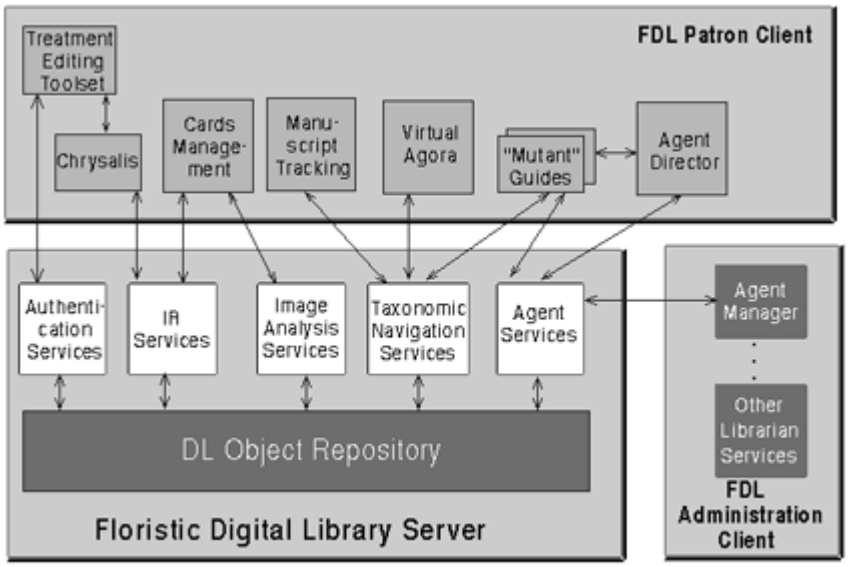


Figura 1.1: Arquitectura de la Biblioteca Digital Florística.

1.2 Proyecto Flora de China

El proyecto Flora de China (FOC), es un proyecto colaborativo entre diversas instituciones para hacer disponible a través de Internet y otros medios electrónicos toda la información ahora existente acerca de las plantas de China dentro de la FDL.

La palabra ”flora” se refiere tanto a las plantas que existen en cierta región como a la publicación de descripciones científicas de tales plantas. Para distinguir entre uno y otro significado, se usa la primera letra mayúscula (Flora) para denominar a las publicaciones. Las Floras difieren de los manuales populares en que las primeras tratan de cubrir todas las plantas, en vez de sólamente cubrir las más comúnes o representativas.

Una Flora casi siempre contendrá nombres científicos, nombres comúnes, referencias literarias, descripciones, habitats, distribución geográfica, ilustraciones, y otras características [Herbario de Harvard 1998].

Toda esta información que constituye la Flora se presentará en el idioma inglés y se constituirá de descripciones de aproximadamente 30,000 especies de plantas vasculares de China.

Esta recolección de datos describirá y documentará las especies antes mencionadas. Todas las plantas vasculares de China serán cubiertas, incluyendo descripciones breves, llaves identificadoras, distribución geográfica y otras características.

La colaboración de diversos organismos internacionales con la que se cuenta para la investigación, publicación, revisión, y edición es el sello de calidad de la producción de la Flora. El proyecto tiene cinco centros fuera de China, en el Herbario de la Universidad de Harvard, la Academia de Ciencias de California, el Instituto Smithsoniano, el Real Jardín Botánico de Edinburgo y el Jardín Botánico de Missouri. Además se integran cuatro centros en China: el Instituto de Botánica de Beijing, el Instituto Kunming de Botánica, el Instituto Jiangsu de Botánica, y el Instituto de Botánica del Sur de China. Más de 600 científicos en el mundo cooperan en la preparación de tratamientos individuales de la Flora [Herbario de Harvard 1998].

El diseño del sistema que contendrá toda esta información consta de un manejador de bases de datos relacional que contiene tratamientos florísticos, rutinas de entrada de datos, rutinas de extracción y traductores e interfaces al sistema (Figura 1). La información podrá ser extraída en diferentes formatos electrónicos. El resultado es una representación altamente estructurada de tratamientos taxonómicos de plantas. Por tratamientos taxonómicos entendemos textos que incluyen descripciones morfológicas de las plantas clasificadas, referencias a publicaciones que las discuten, así como imágenes y diagramas que las ilustran.

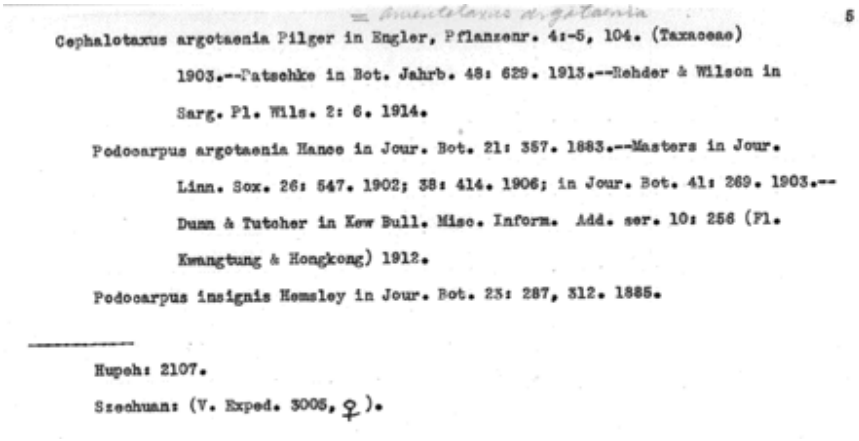
El software utilizado para el almacenaje y recuperación de la información es el manejador de base de datos Informix Universal Server. Las interfaces del usuario con la base de datos son para su uso en WWW y se emplea una combinación de CGI, Javascript, Java, ActiveX y Herramientas "Datablade" de Informix [Ihsan & Schnase 1996].

1.3 Las Tarjetas de Hu

Dentro del proyecto Flora de China se encuentra una colección de tarjetas que constituye la única base de datos a nivel infra-específico de 30,000 plantas vasculares de China, las cuales representan una octava parte de la flora del mundo. Esta compilación es denominada las Tarjetas Hu, y consta de un conjunto de tarjetas de cartón de 5x8

pulgadas de diferentes colores que contienen texto manuscrito y mecanografiado además de recortes de hojas pegadas, compiladas por la Dra. Hu Shiu-Ying [Herbario de Harvard 1998]. La figura 1.2 es un ejemplo de estas tarjetas.

La colección se inició en los primeros años de los 50´s cuando en el Arnold Arboretum de la Universidad de Harvard se planteó un proyecto para preparar una recolección de flora de China. La Dra. Hu, junto con otras cuatro o cinco personas investigaron en toda la literatura botánica entre los años de 1773 y 1955 para recolectar todos los nombres que se han usado para las plantas de China. Esta compilación hasta ahora se encuentra físicamente en archivos del herbario de la Universidad de Harvard, y su uso es limitado, puesto que cualquier interesado en recuperar información de la colección debe físicamente hurgar entre las tarjetas, o solicitar copias de ellas, lo cual no permite aprovechar al máximo la información en las mismas.


5

Cephalotaxus argotaenia Pilger in Engler, Pflanzenr. 4:-5, 104. (Taxaceae)
1903.--Patschke in Bot. Jahrb. 48: 629. 1913.--Rehder & Wilson in
Sarg. Pl. Wils. 2: 6. 1914.

Podocarpus argotaenia Hance in Jour. Bot. 21: 357. 1883.--Masters in Jour.
Linn. Soc. 26: 547. 1902; 38: 414. 1906; in Jour. Bot. 41: 269. 1903.--
Dunn & Tutcher in Kew Bull. Misc. Inform. Add. ser. 10: 256 (Fl.
Kwangtung & Hongkong) 1912.

Podocarpus insignis Hemsley in Jour. Bot. 23: 287, 312. 1885.

Hupeh: 2107.

Szechuan: (V. Exped. 3006, ♀).

Figura 1.2: Una tarjeta de Hu

Según la información contenida en las tarjetas, éstas se dividen en tres tipos:

Tarjetas escritas a máquina con el nombre de la planta y referencia de la publicación en donde fue usada.

Recortes de artículos que contienen el nombre de la planta y citas bibliográficas.

Tarjetas escritas a mano con el nombre de la planta y citas a bibliografías.

Las tarjetas contienen además información acerca de plantas sinónimas a ellas y su localización geográfica.

Esta información es vital para investigadores en las diferentes áreas relacionadas con la Botánica: preservación de la flora, hallazgo de las propiedades medicinales de dichas plantas, y mantenimiento de un registro que sea accesible a cualquier usuario. Por ello, se ha decidido abrir esta información hacia la comunidad investigadora, haciéndola parte de la biblioteca digital florística del Jardín Botánico de Missouri bajo el nombre de "Annotated Flora of China Checklist" (AFCC). Para que esto suceda, estas tarjetas deberán ser digitalizadas y se deberá aplicar un método que facilite la explotación de la información visible en las mismas, es decir, el texto.

Para facilitar al usuario la exploración de la información dentro del conjunto de las tarjetas se pueden usar diversas técnicas, entre las que se encuentran la asociación de patrones e indexado [Witten et. al, 1994] y el reconocimiento de patrones. La primera morfológicamente hace una asociación entre un símbolo preestablecido y un patrón dado mediante heurísticas y sirve primariamente para comprimir la imagen textual.

Este trabajo se enfoca en la segunda técnica, la cual analiza la imagen dando como resultado el significado que para el ser humano tienen los símbolos incluidos en ella. [Bishop 1995].

El objetivo principal de este trabajo es proveer al usuario herramientas de consulta y de almacenamiento de las tarjetas de Hu. Esto se logra por medio del reconocimiento de los caracteres en cada tarjeta y su almacenamiento en un manejador de bases de datos, para su posterior consulta y mejoramiento por parte del usuario.

El sistema realizado en este proyecto para solucionar el problema se plantea en dos partes: una interfaz de consultas de las tarjetas de Hu, y un sistema de interpretación de tarjetas. En conhunto, estos sistemas forman el HuSystem.

Adicionalmente a la obtención de datos dentro de las tarjetas se tiene proyectado a futuro aplicar técnicas de compilación de lenguaje natural y así extraer sólo la información relevante de las tarjetas.

1.4 Organización del documento

En el capítulo II se revisarán las diferentes técnicas asociadas a la detección y reconocimiento de texto en imágenes digitalizadas. En el capítulo III se mencionan las soluciones evaluadas para reconocimiento de caracteres. Se procede después, en el capítulo IV a plantear un modelo de reconocimiento ajustado a las imágenes del campo de estudio, revisando su implementación en el capítulo V. Finalmente, el capítulo VI provee las conclusiones acerca del presente trabajo.

Dircio Palacios Macedo, R. 1998. Reconocimiento y Consulta de Imágenes Textuales en Bibliotecas Digitales. Tesis Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas-Puebla. Diciembre.