

En este trabajo se demuestra el potencial del uso de una gramática para determinar la estructura de la información en una biblioteca digital botánica. En particular se muestra su utilidad para enfrentar problemas en la introducción de información de forma manual a la base de datos. A continuación se resumen los logros de X-tract, los problemas que se encontraron en su desarrollo, así como el trabajo a futuro que puede aumentar su utilidad.

Para sustentar la solución propuesta se analizaron diferentes trabajos relacionados con la extracción de información. X-tract, principal resultado del trabajo desarrollado, es un sistema que permite el análisis de descripciones morfológicas para su almacenamiento en la base de datos FDL. Una de las características más importantes de X-tract radica en su interfaz con el usuario. La interfaz le permite al usuario participar en la corrección de las palabras analizadas antes de actualizar la base de datos.

La operación de X-tract es muy sencilla. Primeramente el usuario introduce el nombre del archivo de texto que debe encontrarse en HTML o introduce texto en formato libre para su análisis. X-tract analiza en ambos casos el archivo dado o crea un archivo con el texto introducido si se refiere a la segunda opción. Con la ayuda de un analizador sintáctico y un analizador gramatical despliega una tabla con la organización propuesta y el nombre de las características a las que se refieren los valores. Para actualizar la base de datos, X-tract recorre toda la tabla creada para saber el orden en el que se encuentra cada palabra y saber si se refiere a una estructura, una subestructura, una característica o un valor.

### 5.1 Logros

### 5.2 Problemas

### 5.3 Trabajo a futuro

### 5.4 Comentarios finales

#### 5.1 LOGROS

Algunas de los logros de X-tract a diferencia de otros sistemas se encuentran los siguientes:

interfaz con el usuario, lo que permite verificar el resultado del análisis de la descripción antes de actualizar la base de datos,

sistema fácil de manejar ya que sólo requiere de dar el nombre del archivo o el texto a analizar,

X-tract permite el análisis de textos en HTML o en formato libre lo que ayuda al usuario a analizar una descripción completa o sólo parte de ella o incluso buscar una palabra en el glosario para saber a que característica se refiere,

X-tract cuenta con una gramática capaz de localizar descripciones morfológicas dentro de un archivo de texto dado,

X-tract muestra el resultado de haber actualizado la base de datos.

## 5.2 PROBLEMAS

Un problema que fue hallado al analizar los resultados de X-tract se presenta cuando aparece la estructura después del valor. Por ejemplo, "subterranean axes". En este caso *axes* corresponde a una estructura y *subterranean* es el valor de dicha estructura. Esto se debe a que la gramática no está enfocada al orden gramatical que debe de llevar una oración. Generalmente las estructuras vienen primero y por lo tanto existirían varios casos por analizar, además del número de ocurrencias de este problema.

## 5.3 TRABAJO A FUTURO

Como se mencionó anteriormente las pruebas y el desarrollo del trabajo se basaron en la creación de un sistema capaz de analizar los tratamientos taxonómicos localizados en la página electrónica de FNA. Estos tratamientos están formados por descripciones precedidas y seguidas de código HTML. Además, las estructuras principales se encuentran entre "<B>" y "</B>". Es así como se desarrolló una gramática capaz de analizar textos con esta organización. Es decir, siempre una descripción morfológica comienza con una palabra entre "<B>" y "</B>".

Como trabajo a futuro queda la programación del caso de descripciones que no comienzan de esta forma.

X-tract sólo actualiza las tablas correspondientes a la descripción ya que la gramática sólo analiza esta parte. Queda pendiente entonces, el análisis del nombre de la descripción como de los autores y demás partes que componen un tratamiento taxonómico.

## 5.4 COMENTARIOS FINALES

La construcción de una biblioteca digital es un esfuerzo de miles de autores trabajando juntos con un amplio rango de herramientas para capturar sus pensamientos. X-tract es una herramienta que además de ayudar en el análisis de descripciones morfológicas y a la actualización de la base de datos es una herramienta que auxilia al autor en la construcción de las descripciones.

Una contribución adicional de X-tract es que hace posible que los autores distribuidos por todo Norteamérica puedan enviar sus tratamientos para el análisis de descripciones de manera electrónica, gracias a la disponibilidad de X-tract vía WWW.

Al finalizar el sistema se logró demostrar que es posible mejorar notablemente la rapidez y facilidad para introducir las descripciones morfológicas a la biblioteca digital.

X-tract forma parte de un sistema, más general, de apoyo a los autores de descripciones morfológicas. Con su uso se espera agilizar la construcción de la biblioteca digital además de constituir un paso en la investigación en el área de extracción de información en la Biblioteca Digital Florística.

índice   resumen   1   2   3   4   5   referencias

Abascal Mena, M. R. 1998. [Extracción de estructura a partir de descripciones textuales botánicas](#). Tesis Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas-Puebla. Diciembre.

Derechos Reservados © 1998, Universidad de las Américas-Puebla.