Capítulo 1. Introducción

1.1 Antecedentes

Actualmente existen grandes volúmenes de información en Internet y por medio de los motores de búsqueda podemos hacer consultas para buscar información específica. Al realizar una consulta, los motores de búsqueda pueden arrojar miles o millones de resultados; si se presenta este caso, el usuario tendrá que invertir mucho tiempo para poder encontrar la información que busca; en caso de no encontrarla, volverá a repetir el proceso anteriormente mencionado.

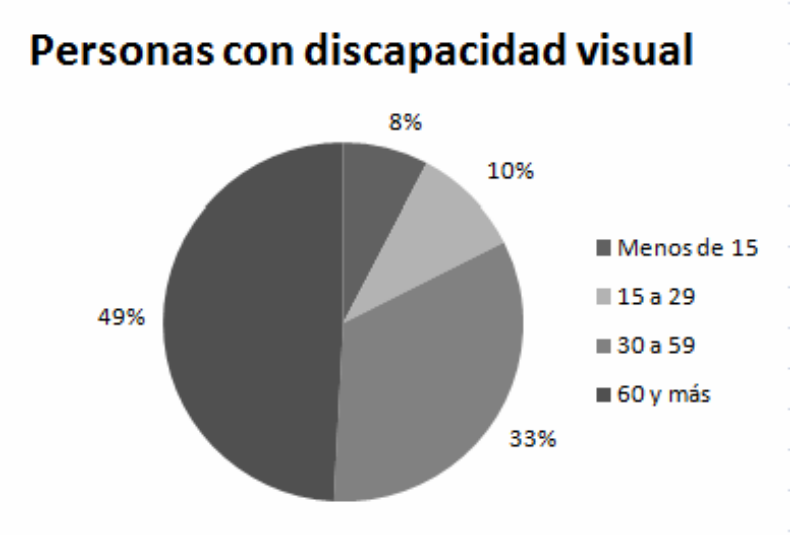
Como podemos observar, la búsqueda de información puede convertirse en un proceso muy lento; además, debemos tomar en cuenta los diferentes usuarios de Internet, entre algunos tenemos:

- Expertos
- Principiantes
- Personas con capacidades diferentes

Dentro del grupo de personas con capacidades diferentes podemos encontrar discapacidades visuales, motrices y auditivas. Las personas invidentes hacen uso de diversas herramientas para poder usar la computadora; entre estas podemos mencionar a los lectores de pantalla, impresoras braille, sistemas de ampliación de la pantalla, programas de reconocimiento de texto OCR parlantes, etc. [Alvarez, 2006].

Un censo de 1999 en Estados Unidos reportó que aproximadamente 1.5 millones de personas con diferentes tipos de condiciones visuales hacían uso de las computadoras[1].

El XII Censo General de Población y Vivienda 2000 de México reportó que existe un promedio de 26% de personas con alguna discapacidad visual[2], de las cuales 230, 862 son hombres y 236, 178 son mujeres. En la siguiente gráfica se muestra el porcentaje de personas con discapacidad visual agrupadas por edad, según el INEGI.



Fuente INEGI. XII Censo General de Población y Vivienda 2000

[1] Información tomada de la página AFB American Foundation for the Blind, http://www.afb.org/Section.asp?SectionID=15&DocumentID=1367#comp

[2] INEGI. XII Censo General de Población y Vivienda 2000. Base de datos. http://www.inegi.gob.mx/prod_serv/contenidos/espanol/bvinegi/productos/censos/poblacion/2000/discapacidad/visual_i.pdf

La Organización Mundial de la Salud publicó que … “en 2002 el número de personas con deficiencia visual en todo el mundo superó los 161 millones, y de ellos 37 millones sufrían ceguera”[3]

1.2 Definición del problema

Anteriormente mencionamos, que para encontrar información específica en Internet debemos invertir mucho tiempo; una forma de resolver este problema sería revisar el resumen de la página en lugar de revisar todo el documento, de esta forma se agilizaría este proceso y por consiguiente aumentaría la productividad de los usuarios. Esta solución beneficiaría a todos los usuarios, pero especialmente simplificaría el trabajo de las personas invidentes, porque al contar con el resumen de la página conocerían de forma más rápida el contenido de ésta.

La solución a este problema es crear un sistema que genere de forma automática el resumen de un documento en español, con el propósito de agilizar la búsqueda de información en Internet por parte de los usuarios que hablan este idioma, sin importar el tipo de usuario o su área de desarrollo; pero debemos mencionar que esta solución tendrá un valor agregado para los usuarios con discapacidades visuales. Estas personas harán uso de un hardware especial para poder escuchar el resumen; en la tesis “Linter-Vox: Control Interactivo para Escuchar Documentos de Internet” de Gustavo Elizalde Garrido se hizo una aplicación para que las personas con discapacidad visual trabajen cómodamente en Internet [Elizalde, 2006].

[3] Información tomada de la página Global data on visual impairment in the year 2002 http://whqlibdoc.who.int/bulletin/2004/Vol82-No11/bulletin_2004_82(11)_844-851.pdf

1.3 Objetivo General

Crear un sistema de alto desempeño que genere el resumen de una página Web en español, utilizando la técnica basada en la selección de las oraciones más relevantes del documento; esta técnica se puede aplicar a documentos de cualquier género y dominio; otra característica primordial de esta técnica es que emplea muy pocos recursos, y por consiguiente el sistema podrá arrojar rápidamente los resultados.

1.4 Objetivos Específicos

Para poder crear el Sistema de generación automática del resumen debemos tomar en cuenta los siguientes objetivos:

- Evaluar algunas técnicas para generar resúmenes de forma automática.
- Investigar las técnicas para generar el extracto de un documento.
- Implementar una técnica que arroje, rápidamente, buenos resultados.
- Colocar los datos de salida en un formato legible (archivo de texto plano) para el sintetizador de voz.
- Diseñar un sistema extensible.
- Realizar pruebas al sistema.

1.5 Alcances

El sistema que se va a crear va a tomar en cuenta los siguientes puntos:

- Los documentos que se van a procesar están en formato HTML.
- Crear una lista de las etiquetas HTML que se tomarán en cuenta para formar el documento.
- Crear un parser.

• El sistema dará dos opciones al momento de ejecutarse; la primera opción toma en cuenta las palabras clave, los términos multipalabra, la tipografía y el título del documento; la segunda opción, solamente toma las palabras clave y los términos multipalabra.

• Los documentos que se van a procesar están escritos en español.

• El sistema va a ser diseñado de forma modular, con el objetivo de poder agregar en un futuro nuevos módulos para resolver otros problemas, por ejemplo: el uso de otros idiomas.

1.6 Limitaciones

Hay varias técnicas que generan resúmenes automáticos con muy buenos resultados; sin embargo, no se eligieron porque el sistema va a correr en tiempo real y por lo tanto el programa debe arrojar rápidamente los resultados. Las técnicas mencionadas anteriormente, usan varios recursos que aumentan el tiempo de espera para obtener los resultados, y también no se pueden aplicar a documentos con diferentes géneros y diferentes dominios como los que se van a usar en el sistema (páginas Web); no obstante, se realizará una investigación de estas técnicas.

1.7 Organización del documento

El documento de tesis está organizado de la siguiente manera, en el capítulo 2 presentamos el marco teórico relacionado con la Generación del Resumen Automático, el capítulo 3 muestra el Diseño del Sistema, el capítulo 4 muestra el uso del Sistema, el capítulo 5 informa los resultados de las pruebas realizadas al sistema, y finalmente se detallan las conclusiones de este trabajo.