Informática
Año 2013/2014
Facultad de CC.
Matemáticas

Práctica 1

Crea un índice de páginas web

28 de marzo de 2014

Realizar búsquedas en la red es muy fácil: se va a la página de google y se busca lo que se desea. Pero, ¿qué hace google para darnos una respuesta? ¿cómo es posible que sepa las páginas en las que aparecen las palabras que estamos buscando?

En este ejercicio vamos a resolver esta última pregunta y vamos a ver que es relativamente fácil. Algo más complicado es decidir un orden en el que mostrar las páginas en las que aparecen los términos que buscamos... pero ese problema lo resolveremos otro día. Mientras tanto, vamos a ir preparando nuestro *crawler*[1] por si en algún momento se da la situación que anuncia el siguiente titular:

1 Análisis de una página

Una parte esencial del problema que tenemos que resolver es hacer que nuestro programa analice el contenido de una página web para sacar todas sus palabras. Es mucho más fácil de lo que parece: una página web no es más que un fichero.

Como entrenamiento, escribe un programa que lea un fichero (en local, en tu propia máquina) y que cuente el número de veces que aparece cada palabra.

2 Estructura de Datos

Una parte esencial de la solución es encontrar la estructura de datos adecuada en la que guardar los datos. Siempre que se diseña una estructura de datos hay que tener en cuenta no sólo los datos que queremos guardar, sino el uso que haremos de ellos.

En nuestro caso, lo que queremos hacer es asociar a una palabra las direcciones web (que conocemos) en las que aparece dicha palabra. Qué conjunto de palabras consideramos

[1] Si no sabes qué significa... ¡busca en google que todavía es gratis!

(castellano, asturiano, inglés, sustantivos, no artículos...) es una decisión importante que hay que plantearse y decidir.

Una vez creada la estructura, la utilizaremos para hacer búsquedas como lo hacemos con google. Es decir, si pongo "pera" y "agua" me gustaría que, de forma fácil, pudiese encontrar en mi estructura el conjunto de páginas web (conocidas) en las que aparecen ambas palabras.

Es muy interesante pensar en el lenguaje para hacer las consultas a la estructura de datos. Podríamos considerar otros operadores lógicos como la disyunción o la negación...

Define una estructura de datos para almacenar las palabras y las páginas y define las funciones de almacenamiento y búsqueda en la estructura.

3 Recorriendo internet

La parte del problema que puede parecer más *sorprendente* es la de que nuestro programa vaya *recorriendo* páginas de internet. No es tan complicado.

Como comentábamos en el primer apartado, acceder a una página web es esencialmente igual que acceder a un fichero. En cada página tenemos que buscar los hiperenlaces (enlaces, links...) que nos llevan a otras páginas web y guardarlos. Cuando acabamos de procesar una página saltamos a otra de las que tenemos guardadas para visitar. Te sorprenderás de la cantidad de páginas que puede "leer" tu programa en el tiempo que pasas viendo un capítulo de tu serie favorita.

El punto con el que hay que tener cuidado es no repetir páginas visitadas, además de la pérdida de tiempo que supone, podríamos entrar en ciclos en los que procesaríamos una y otra vez páginas visitadas. Resolver este problema es fácil: basta tener una lista de páginas visitadas y otra de páginas por visitar... eso sí, ¡hay que saber qué hacer con estas listas!

Fecha límite de entrega: Semana del 14 al 20 de Abril.