Universidad de Chile
Departamento de Ingeniería Industrial
IN831 – Web Mining
Prof.: Juan Velásquez – Victor Rebolledo
Prof.Aux: Ivan Videla C

Tarea#1 Web Structure Mining

GOOGLITO
Web Crawler y Page Rank

Una importante empresa de marketing le ha confiado a UD un sistema de obtención de información sobre juegos que permita obtener **las páginas Web más** relevantes aparecidas en Wikipedia (en.wikipedia.org) de acuerdo a una palabra **clave**. La información relevante debe encontrarse en las siguientes categorías:

Sports: Soccer, Basketball, Baseball....
Players: Player-Game....

Basándose en estos datos se requiere:

1. Construir un Crawler que recorra en.wikipedia.org y obtenga las páginas solo pertenecientes al sitio wikipedia que se encuentren relacionadas por link con los temas anteriores. Con esto se dispondrá de una base de datos donde se almacenará el grafo de las páginas en cuestión y sus links asociados. Limite el número de páginas a un máximo de 300.

2. Obtener las palabras claves de las páginas y almacenar una matriz M [palabra x página] cuyas entradas es un número entero que indica la cantidad de veces que la palabra se encuentra mencionada en la página. Esta matriz será almacenada como una tabla en una base de datos. Incorpore también el URL para fines de las tareas siguientes.

3. Implementar el algoritmo PageRank (TM) de Google que entregue el ranking de páginas extraídas en el punto 1 usando el grafo obtenido, el cual será almacenado en la base de datos.

4. Implementar un buscador que dado una palabra clave busque entregando un listado de páginas usando la matriz V y entregue las n páginas coincidentes ordenadas por el ranking.

5. Los programas deben ser implementados en JAVA y el administrador de base de datos es MySQL.

Universidad de Chile
Departamento de Ingeniería Industrial
IN831 – Web Mining
Prof.: Juan Velásquez – Victor Rebolledo
Prof.Aux: Ivan Videla C

Los entregables son:

Informe del proyecto, que incluye además instalación, descripción de programas y bases de datos (archivo pdf).

Dump de base de datos (mysql -u root --opt mibasedatos> grupo_XX.sql).
Códigos fuentes de Programas (no ejecutables grupo_XX.zip).

Se evaluarán las siguientes partes:

- **Informe**: De acuerdo al estándar en proyectos reales (**2pts**).
- **Instalación**: Programa compila y corre se seguirán estrictamente las instrucciones del informe.
- **Base de datos se instala sin errores**, se inspecciona que el código tenga la "intención" buscada en este proyecto. (**1 pts**)
- **Funcionamiento**: Solo si se cumple el punto anterior se verificará que el programa hace lo que se solicitó que haga (**3 pts**).
  Se desglosa en:

- **Crawler y grafo**: (**0.8 pt**).
- **Palabras Claves**: (**0.8 pt**).
- **Page Rank y Buscador**: (**1.4 pt** solo si lo anterior funciona).

Se requiere que UD (la tarea es en grupos de a 2 personas). Se descontarán $\mathbf{2^i}$ **pts**, por dia de atraso, donde i es la cantidad de días desde que expiró el plazo.

BONUS: (**1pts**) Si Funciona al menos el Crawler y la búsqueda de palabras claves, adicionalmente podrá mejorar heurísticamente el buscador es decir la relevancia de las páginas. Ya sabemos que el ámbito son los deportes y los deportistas, entonces las búsquedas a realizar son justamente sobre esos temas (ejemplo: buscar "Real Madrid").
Incluya entonces su heurística en el informe e impleméntela.

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

Universidad de Chile
Departamento de Ingeniería Industrial
IN831 – Web Mining
Prof.: Juan Velásquez – Victor Rebolledo
Prof.Aux: Ivan Videla C

Como lo vimos hoy en clases, el crawler y buscador tienen funcionalidad completa pero requiere de ciertos ajustes:

1) Cargar links en tabla de links -> crear metodos en GooglitoRepository y usarlo en el for del metodo visit

2) Sincronizar metodos de bd para que no ocurra null pointer exception

3) Mejorar la busqueda de palabrar (filtrar mas)

4) Mejorar el search.php, para que busque mas de 1 match en keyword, porque ahora solo toma el primero.

Por lo que acordamos que la tarea se entrega en 2 partes, la parte 1 seria el Lunes 7 de Septiembre a las 23:59 hrs.
Esta consiste en:
- un informe preliminar que al menos indique como compilar, instalar y correr el sistema.

- mejorar el sistema actual completando la parte 1

- la parte 2, 3 y 4 son bonus

Entonces la nota de la tarea 1 se calculara como (Parte1 + Parte2)/2 donde la Parte2 consistirá en calcular el ranking y el informe final del proyecto.

Los archivos ya se encuentran publicados.