UNIVERSIDAD DE BUENOS AIRES

Introducción el Modelo de Hipergrafos para técnicas de reputación web

75.39 – Aplicaciones Informáticas

Facultad de Ingeniería

2° CUATRIMESTRE

11

Índice

Capítulo 1:

Introducción

Desde hace mucho tiempo, el orden de los resultados arrojados por un buscador web sobre una determinada consulta, es muy importante, debido a que el usuario espera encontrar entre las primeras páginas sugeridas por el buscador exactamente lo que está buscando, reduciendo de esta manera el tiempo necesario para obtener los resultados esperados al mínimo posible. Motivo por el cual los buscadores web emplean diversas técnicas para asignar relevancia y establecer un orden a cada una de las páginas resultantes de la ejecución de una consulta que se realiza a través de la web sobre el ingreso de un determinado conjunto de términos. Como efecto final, cada resultado es clasificado y calificado de acuerdo a una serie de algoritmos de relevancia y reputación diseñados para aplicarse sobre todas las páginas y documentos previamente indexados por el buscador web utilizado.

1.1 Técnicas de relevancia de páginas y documentos

Nos centraremos en el estudio de los *sistemas de recuperación de información en la web* debido a su interés técnico y comercial, su gran popularidad y fácil acceso. En primer lugar, enunciaremos las características más importantes de la colección más grande jamás creada de documentos y páginas: **Internet.**

Las técnicas utilizadas en la actualidad para asignar relevancia a páginas y documentos web son diversas pero pueden dividirse en dos grandes grupos:

1. **Análisis del contenido** *(Búsqueda Semántica).*
2. **Análisis de Links.**

Análisis del Contenido

Este tipo de estudios tiene como objetivo clasificar las páginas según su contenido, teniendo en cuenta los términos buscados en la consulta. Dados los términos de la misma, puede haber muchos documentos en los cuales aparezcan. Sin embargo esto no es suficiente para afirmar que la información contenida en los mismos sea la buscada por el usuario. Para dejar más claro el concepto introducimos el siguiente ejemplo: La consulta "*Jorge Borges*" pude realizarse por un usuario que busca información bibliográfica sobre el autor. Que una página contenga ambas palabras no asegura tener información precisa del mismo, puesto que el tema que se trate sea otro y solo se vinculó a la persona con el contenido, en un solo

Ejemplo de consulta: Jorge Borges

párrafo. Precisamente, la tarea de estos métodos, es darle a dichos documentos un puntaje bajo, de manera que estén a lo último en la lista de resultados y el usuario no pierda tiempo revisando contenidos que no tratan el tema que es motivo de su consulta.

Análisis de Links

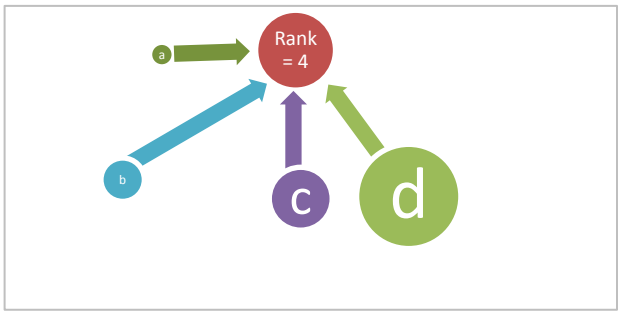
Estos métodos de ranking son el principal objetivo de estudio del presente trabajo. En este tipo de análisis se asigna un puntaje a cada página al igual que en el método antes descripto, pero este puntaje es independiente de los términos de búsqueda consultados.

La principal característica de este tipo de metodologías consiste en que para cada página analizada se realiza el cálculo de su reputación, utilizando como información, la cantidad de links existentes que referencian esa página. La manera más simple de llevar a cabo este método es, por ejemplo, tomar la cantidad de links que referencian a un documento y proponer como reputación del mismo dicho numero. La justificación para la aplicación de estos métodos consiste en que un documento web que posee una cantidad significativamente grande de links referenciándolo, es por lo general, un sitio que contiene cierta información muy relevante para muchos sitios razón por la cual este debiera obtener una mejor reputación que otros sitios de menor referencia.

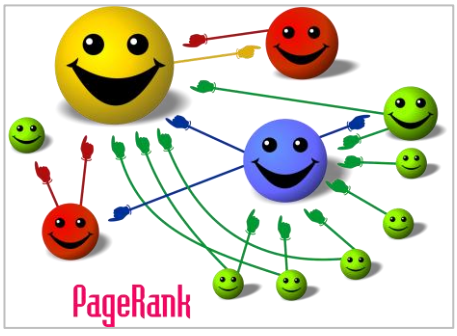
En general, los métodos de reputación se utilizan en combinación con otros métodos cuyo objetivo es el de buscar los sitios y documentos que cumplen con un determinado criterio de búsqueda. Una vez obtenidos los sitios que mejor se ajustan al criterio de la consulta del usuario, estos son ordenados según su reputación, la cual dependerá del tipo de método utilizado (Los detalles de dichos métodos serán abarcados en el apartado X). De esta manera se busca lograr que los sitios que mejor responden a las consultas de un usuario en cuanto a su contenido y reputación, sean los primeros en la lista de resultados a fin de reducir los tiempos de búsqueda del usuario lo máximo posible.

A continuación se exponen las dos técnicas de análisis de link que se utilizaran en el presente trabajo expuesto.

1. **Indegree**
2. **PageRank**



Funcionamiento del Indegree



Funcionamiento del PageRank

Capítulo 2:

Desarrollo del Estado del Arte

Método Indegree

Esta técnica es la más simple en el análisis de links. Básicamente consiste en contar la cantidad de links que hay hacia una página y proponer dicho número como su reputación. Es decir:

Reputación(Pagina X) = Cantidad de páginas con links hacia X

Es esperable que un método de fácil cálculos tenga muchas desventajas y este caso no es la excepción. El primer inconveniente es que un usuario, con la atención de que su sitio obtenga buena reputación puede llegar a crear muchos otros que hagan referencia al primero sin que el método nos advierta ni evite el posible fraude. Más allá de que es posible y muy a menudo que las páginas pertenecientes a un mismo autor se referencien entre sí, sin mala intención, es necesario poder valorar con más fuerza a los links provenientes de hosts o dominios diferentes a los cuales pertenece el documento analizado en ese instante. Una situación similar pude darse con las páginas de publicidad que son referenciadas por numerosos sitios, a través de los banners que poseen los mismos. Como resultado de lo expuesto en el párrafo anterior, se deduce que es posible encontrar documentos en los cuales solo se hallan avisos o promociones con una alta relevancia sin merecerlo.

Método PageRank

La cantidad de problemas que trajo acarreado el método **Indegree**, generó que se buscaran alternativas al mismo. La técnica llamada **PageRank**, fue la propuesta más satisfactoria que se ha encontrado. La clave de su éxito, estuvo basada en que para que una página tenga buena reputación no solo se considera la cantidad de links entrantes hacia la misma, sino que además estos links deben pertenecer a su vez, a páginas con alta reputación. De esta manera el camino para que se den los fraudes vistos en el punto anterior es más complicado, ya que ahora, un sitio no solo le basta tener varios links entrantes para un buen puntaje, sino que además, cada uno de estos deben tener a su vez, una buena calificación.

A continuación se detalla el cálculo desarrollado para este método:

$$PR(p) = (1 - c)\, x \sum_{q \in I(p)} \frac{\mathrm{PR(q)}}{||\mathrm{O(q)}||} + \frac{\mathrm{c}}{||\mathrm{r}||}$$

c: Factor de ajuste

I(p): Set de páginas que apuntan a p

||O(q)||: Número de páginas apuntadas por q

||r||: Numero de páginas en la colección

Un punto cuestionable de este método es la tendencia que genera a que un documento web pueda obtener una alta calificación a pesar de ser apuntado por un número pequeño de páginas, debido a que estas tienen una alta reputación.

Ambas técnicas toman como base para sus cálculos un grafo, previamente armado, el cual representa a la web entera y sus vínculos. Los vértices de dicho grafo representan a cada una de las páginas web, mientras que los arcos indican los links presentes. A modo de ejemplo, podemos decir que si tenemos dos vértices A y B conectados por un arco dirigido de A hacia B, podemos concluir que la pagina A hace referencia mediante un link a la página B. Este es el tema central por el cual es motivado este estudio. Si se analiza con cuidado esta representación, se puede afirmar que la manera de encarar el armado del grafo puede generar ruido en los resultados. Es necesario considerar que varias páginas web pueden pertenecer al mismo dominio o host. A partir de aquí nos hacemos el siguiente cuestionamiento:

"*Cuando se tiene una página X apuntada por varias páginas pertenecientes a un mismo Dominio o Host, ¿Todas ellas deben contribuir con el mismo peso para el cálculo de la calificación de X?*"

La respuesta es **NO**. Es bueno para un documento ser apuntado por otros de buen prestigio, pero además hay que pedir variedad en los orígenes desde los cuales parten estas conexiones o links. Así nos asegurarnos que la comunidad que considera a la pagina como útil, es amplio y no se trata de un grupo minoritario que mas allá de su reputación, no deja de ser un pequeño punto en una extensa red.

Es a partir de este momento en donde proponemos modificar el armado del grafo considerando la situación anteriormente expuesta. La idea es armar un Hipergrafo en donde se utilice un criterio específico de agrupamiento, el cual nos represente de manera más efectiva la comunidad web, para generar el menor ruido posible en los resultados.

El Modelo Hipergrafo

Básicamente se trata de un grafo dirigido como el anterior, representado de la siguiente manera:

$$H = (V,E)$$

siendo "V" el conjunto de vértices y "E" un set de Hiperarcos.

Al igual que en el anterior modelo de Grafos, cada vértice "v" representa a una página, mientras que cada hiperarco "e" cumple con la siguiente definición:

$$e = (B,v) \quad v \text{ no pertenece a } B$$

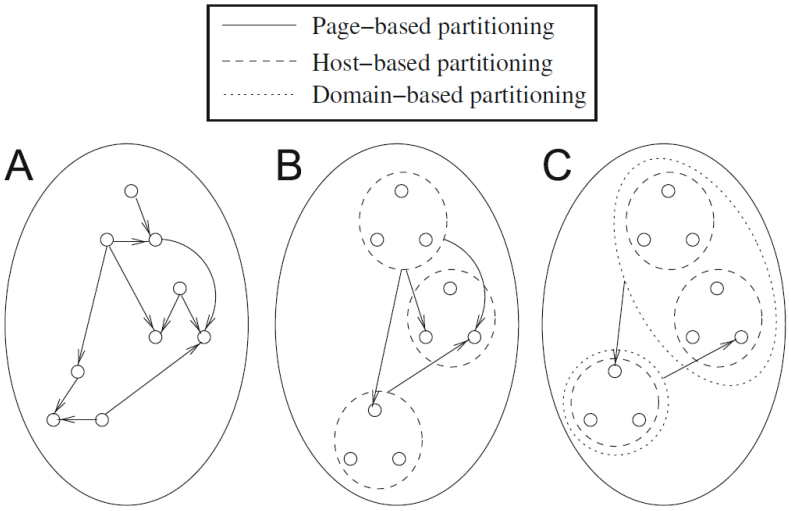
Siendo "B" un bloque de paginas agrupadas mediante un determinado criterio, la idea de esta representación es que existe un hiperarco $e(B,v)$ si y solo si hay como mínimo una página del bloque "B" que tiene un link a la pagina "v" siendo "v" no perteneciente al conjunto "B".

El eje central de la idea expuesta esta en el criterio de agrupamiento que se defina para su implementación. De esta elección, depende el éxito de los resultados.

En nuestro caso, vamos a utilizar tres criterios de agrupamiento:

1) **Partición a base de páginas**: Un bloque se compone de una única página web. En este caso llegamos al modelo de grafo tradicional.
2) **Partición basada en Dominios**: Todas las páginas de un bloque pertenecen al mismo dominio web.
3) **Partición basada en Host**: Todas las páginas del bloque pertenecen al mismo host web.

Cada uno de los hipergrafos que surjan de estos tres criterios serán utilizados como base para los métodos *Indegree* y *PageRank*.



Tal como se puede observar en la figura anterior el hipergrafo **A** considera cada pagina como criterio de partición, el hipergrafo **B** considera el Host como criterio de partición y finalmente el hipergrafo **C** considera el Dominio como criterio de partición.

La aplicación del hipergrafo en un método sencillo como el Indegree, nos ayuda a comprobar más fácilmente como la construcción del hipergrafo contribuye a otorgar calificaciones más justas a las páginas. Además adoptamos el método de agrupamiento 1, que es equivalente al modelo de grafo tradicional, para comprobar si las teorías expuestas son ciertas.

Por último se aplica el hipergrafo al método de Page Rank, dado que es uno de los más utilizados en la realidad por los algoritmos de análisis de links, con el fin de proponer una mejora en los mismos, sustentada por los resultados que se obtengan.

La utilización del Hipergrafo, implica una modificación a las ecuaciones que describen las técnicas de Page Rank e Indegree. Dichas modificaciones se analizarán mas adelante como parte de la presentación del problema a resolver.

Capítulo 3:

Presentación del Problema a resolver

Aplicación del Hipergrafo al Método Indegree

Dada la simpleza matematica del método Indegree, las modificaciones necesarias para la utilización del Hipergrafo no son demasiado complejas. La utilización de bloques que agrupan páginas, señala que no se tendrán en cuenta los links internos, dado que en principio, el principal motivo de este estudio es darle más prioridad a la variedad de fuentes de donde provienen los links. De esta manera la puntuación de cada página según el método Indegree con la aplicación del Hipergrafo, queda determinada por la siguiente fórmula:

$$HI(p) - \sum_{B \in I(p)} 1$$

Básicamente la calificación de cada página es igual a la sumatoria de la cantidad de hiperarcos entrantes a la misma.

Aplicación del Hipergrafo al método PageRank

El primer paso es obtener la manera de calcular la reputación de cada bloque. Una estrategia efectiva, es considerar que la reputación de un bloque es la sumatoria de la reputación de las páginas, de las cuales se compone:

(1) $$GR(B) = \sum_{peB} \mathrm{PR(p)}$$

Dado que este es el primer calculo que debe realizarse, se asigna a cada página un puntaje inicial:

Si la pagina p tiene hiperarcos entrantes ➔ $PR(p) = \frac{1}{||V||}$

Si la pagina p no tiene ningún hiperarco entrante ➔ $PR(p) = 0$

donde ||V|| es el número de páginas con hiperarcos entrantes en la colección.

La razón por la cual le damos puntaje 0 a las paginas sin hiperarcos entrantes es para no favorecer de manera injusta con un PR alto a los bloques con muchos documentos, entre los cuales, varios no poseen conexiones entrantes.

El segundo paso es definir el cálculo de PR de cada página utilizando los valores iníciales del bloque y la pagina misma, de la siguiente manera:

(2) $$PR(p) = (1-c)\,x \sum_{BeI(p)} \frac{\mathrm{GR(B)}}{||\mathrm{O(B)}||} + \frac{\mathrm{c}}{||\mathrm{V}||}$$

c: factor de ajuste

$I(p)$: Set de bloques de páginas que apuntan a la pagina p

$||O(B)||$: Número de páginas apuntadas por el bloque B

De aquí en más los pasos (1) y (2) deben repetirse de manera iterativa hasta que los valores correspondientes a cada página converjan. Cabe resaltar que del mismo modo que ocurre con el método tradicional de PageRank, la convergencia de los valores en este caso, también está asegurada.

Tipos de Consulta

Existen dos tipos de consulta que los usuarios pueden llegar a formular en un buscador web:

1) Consultas Navegacionales.
2) Consultas informacionales.

Las consultas de tipo ***Navegacionales*** buscan encontrar una página web cuya dirección sea la especificada en la consulta. Por ejemplo la consulta "*yahoo*" está buscando el sitio web www.yahoo.com.ar

Ejemplo de consultas Navegacionales

Por el contrario las consultas de tipo ***Informacionales*** son utilizadas por los usuarios que buscan información acerca de un tema específico, descripto a través de los términos que se especifican en la misma, por ejemplo, la consulta: "Bibliografía Borges" busca sitios web que contenga información bibliográfica del autor. La resolución de este tipo de consultas esta mucho mas vinculada al análisis del contenido semántico de una página web que a su ranking, con lo cual tal como se especificó anteriormente no formará parte del presente estudio.

Ejemplo de consultas Informacionales

Métricas para evaluar resultados

En este apartado vamos a exponer aquellas metodologías necesarias para evaluar la calidad de los resultados obtenidos a través de la utilización de los métodos de análisis de link. Si bien, el mejor método siempre va a ser la opinión del usuario, se necesita poder defender las técnicas de rankeo mediante resultados estadísticos, obtenidos a partir de un número extremadamente grande de pruebas, necesitando gran capacidad de procesamiento.

Las consultas **navegacionales** (tipo 1) son las más fáciles de evaluar dado que solo hay un resultado correcto, y es el sitio web buscado. Para estos casos utilizamos MRR (*Mean Reciprocal Ranking*) que es la métrica más común para la evaluación de la calidad de los resultados en consultas navegacionales. Dicha métrica presenta la siguiente formulación matemática:

$$MRR(QS) = \frac{\sum_{\forall qi \in QS} \frac{1}{\text{PosCorrectAnswer(qi)}}}{|QS|}$$

donde:

QS el set de consultas.

“PosCorrectAnswer” es la posición en la que se encuentra el sitio buscado en el ranking

El resultado arrojado por el cálculo es número entre 0 y 1, donde 1 es el mejor valor posible, obteniéndose solo si todos los sitios buscados se encuentran ubicados en el primer lugar del listado entregado como resultado de la consulta.

Capítulo 4:

Propuesta de Solución

Desarrollo

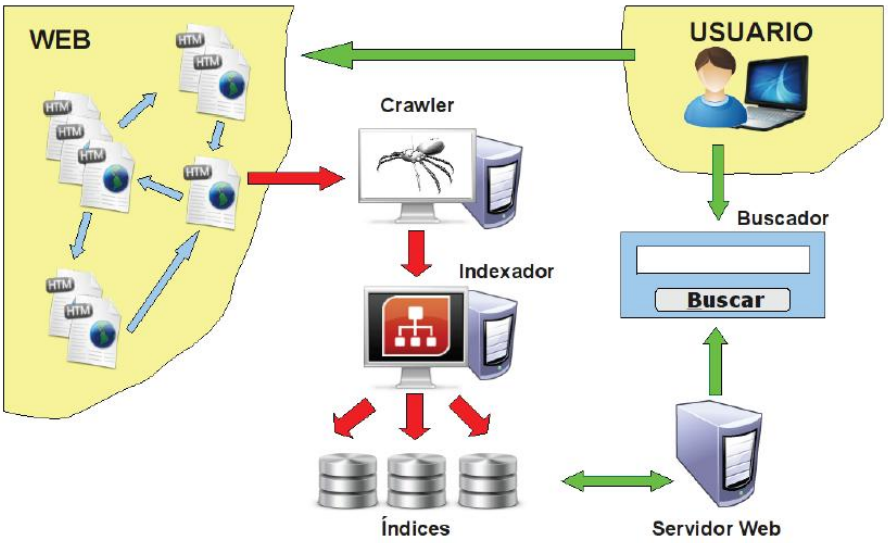
Se busca poder comprobar las mejoras que introduce la utilización del hipergrafo en los métodos de rankeo de páginas. Para poder realizar dicha tarea, se aplicara su construcción, como base del método de PageRank, el cual es frecuentemente utilizado en la realidad. También se aplicara el hipergrafo con el método Indegree, dado que su escasa complejidad de cálculo, permite visualizar más fácilmente, cómo influye el mismo en los resultados finales. También se computaran los resultados para los métodos clásicos de **PageRank** e **Indegree** utilizando un grafo común, para luego poder compararlos con los del Hipergrafo y poder sacar conclusiones.

Diseño del Software

Para llevar a cabo lo expuesto anteriormente se desarrollará un software que constara de dos modulos:

1. Modulo Indexador
2. Modulo de procesamiento de ***Consultas Navegacionales***

El **modulo indexador** será el encargado de examinar los links de cada una de las páginas de un *set de pruebas*, con el fin de construir el Hipergrafo y el grafo común, correspondientes para luego realizar la aplicación de los métodos de **PageRank** e **Indegree**. Luego de haber procesados los métodos, cada uno de los rankings finales será archivados en una Base de Datos, dando la opción de ser consultados en cualquier momento.



Arquitectura sistema de RI para la web

Tal como se lo comento en la sección de estado del arte un sistema de Recuperación de Información (RI) posee una arquitectura similar a la que se muestra en la captura anterior. Esta arquitectura cuenta básicamente con dos procesos importantes que destacaremos a continuación:

1- La ***Indexación*** es la operación que se realiza con cierta periodicidad y consiste en el análisis de los documentos de la colección, es decir las páginas web, para crear los índices de términos que permitan acceso a los mismos de la manera más reciente posible. Para alimentar al sistema de indexación se necesita de otro proceso que vaya recorriendo el grafo que representa la web en busca de nuevos nodos para analizar. A este último proceso se le conoce como <<Crawler>> o <<Araña>>, Por su complejidad de elaboración y por ser una herramienta que en general no aporta conocimientos sobre el tema tratado, se decidió omitir la creación de un web-Crawler y en su lugar las páginas serán generadas mediante un **set de pruebas**.

2- El proceso de ***Búsqueda*** comienza cuando un usuario realiza una consulta al servidor web del sistema de recuperación de información, este se encarga de transformar la consulta en una petición a la base de datos de índices donde se buscaran los nodos que conformaran el resultado. Normalmente los buscadores web presentan la lista de resultados ordenándolos según su relevancia estimada, basada en algún algoritmo puntuación como los mencionados anteriormente. Algunos buscadores también presentan sugerencias a la consulta cuando detectan que el conjunto de resultados obtenido es escaso o poco relevante, muchas veces esto se debe a una consulta mal planteada o con faltas de ortografía. De igual manera la funcionalidad de consultas no será implementada en nuestro modulo de procesamiento de consultas.

El **modulo de procesamiento de consultas navegacionales** constara de dos funcionalidades

1. **Modo Iterativo:** Poseerá una interfaz para que el usuario realice una consulta **navegacional**. Luego de realizar la consulta, el modulo pondrá en funcionamiento un motor de procesamiento que seleccionara aquellas páginas candidatas, y utilizará el ranking realizado por el modulo indexador para darles un orden y luego mostrar el resultado en pantalla.

2. **Modo Procesamiento:** Dado la necesidad de aplicar la medida **MRR** en un conjunto amplio de consultas para obtener un numero que signifique la calidad de los resultados que se obtienen, se incluirá una función adicional en la que se leerán consultas de un ***archivo de entrada***, se las procesara y se computara para cada una su **MRR**, dando como resultado final un promedio de los mismos.

Cabe resaltar que para ambas funcionalidades existirá la opción para seleccionar bajo que método se desea procesar la/las consultas:

1. **PageRank** - (grafo con partición de pagina)
2. **HyHostPR** - (hipergrafo con partición de Host)
3. **HyDomPR** - (hipergrafo con partición de Dominio)
4. **Indegree** - (grafo con partición de pagina)
5. **HyHostInd** - (hipergrafo con partición de Host)
6. **HyDomInd** - (hipergrafo con partición de Dominio)

Capítulo 5:

Implementación de la Solución

Lo primero que se decidió a la hora de comenzar a desarrollar los módulos de la aplicación fue la selección de tecnologías a utilizar. Luego de realizar un análisis sobre las tecnologías más utilizadas para la realización de buscadores e indexadores, sin perder de vista el aspecto técnico de la tecnología, la madurez y los costos asociados. ***(ver bibliografía 2)***

Análisis de Portabilidad

Un factor muy relevante a la hora de realizar este proyecto es tener en cuenta la portabilidad de la aplicación construida, por tal motivo las tecnologías a seleccionar deberían cumplir con los requerimientos de portabilidad. Lo que en líneas generales podríamos resumir como la posibilidad de implementar dicha aplicación en diferentes plataformas operativas, ampliando de esta manera las opciones a la hora de seleccionar el servidor donde alojarla.

Análisis de Costos

El análisis de costos realizado se baso exclusivamente en la determinación de las licencias a pagar por la plataforma tecnológica seleccionada, las opciones como .NET, Oracle, MS-SQL, etc. Tienen asociado un costo que en principio podría omitirse o por lo menos reducirse considerablemente con alternativas de tecnologías *Open Source*.

Selección de la tecnología

Basándonos en las apreciaciones anteriormente detalladas y en otros factores como la simplicidad de desarrollo, tiempos de respuesta, detección de errores, etc. Se seleccionaron las siguientes tecnologías:

- ✓ PHP: Lenguaje de desarrollo para el modulo de consultas
- ✓ MySQL: Motor de base de datos para la indexación de paginas
- ✓ JAVA: Lenguaje de desarrollo para el modulo indexador
- ✓ LINUX: Sistema operativo para el servidor web
- ✓ JQuery (Librería Javascript)

Dicha selección nos permite tener un costo nulo de licencias, y un nivel de seguridad y portabilidad aceptables para futuras versiones del producto.

Herramientas Utilizadas

Las herramientas que se utilizaron para la realización del proyecto como para su ejecución y mantenimiento son las que se listan a continuación:

- ✓ Eclipse (IDE PHP / IDE JAVA)
- ✓ MySQL Query Browser (Cliente MySQL Linux)
- ✓ SQLyog client (Cliente MySQL Windows)
- ✓ Firefox Browser (Web Browser)
- ✓ Ubuntu (Sistema operativo Linux)
- ✓ Windows XP / Windows 7 (Sistema operativo Microsoft)
- ✓ StarUML (Herramienta para creación de diagramas UML)

Diseño de clases: *Módulo Indexador*

A continuación se detalla el diagrama UML *(Unified Modeling Language)* de clases del modulo de Indexación realizado en lenguaje JAVA, dicho diagrama fue generado dentro de la herramienta Eclipse, de esta manera cualquier modificación sobre el código de dichas clases presentes en el modulo es actualizado al instante obteniéndose el correspondiente grafico de clases con las nuevas modificaciones tal como se muestra a continuación:

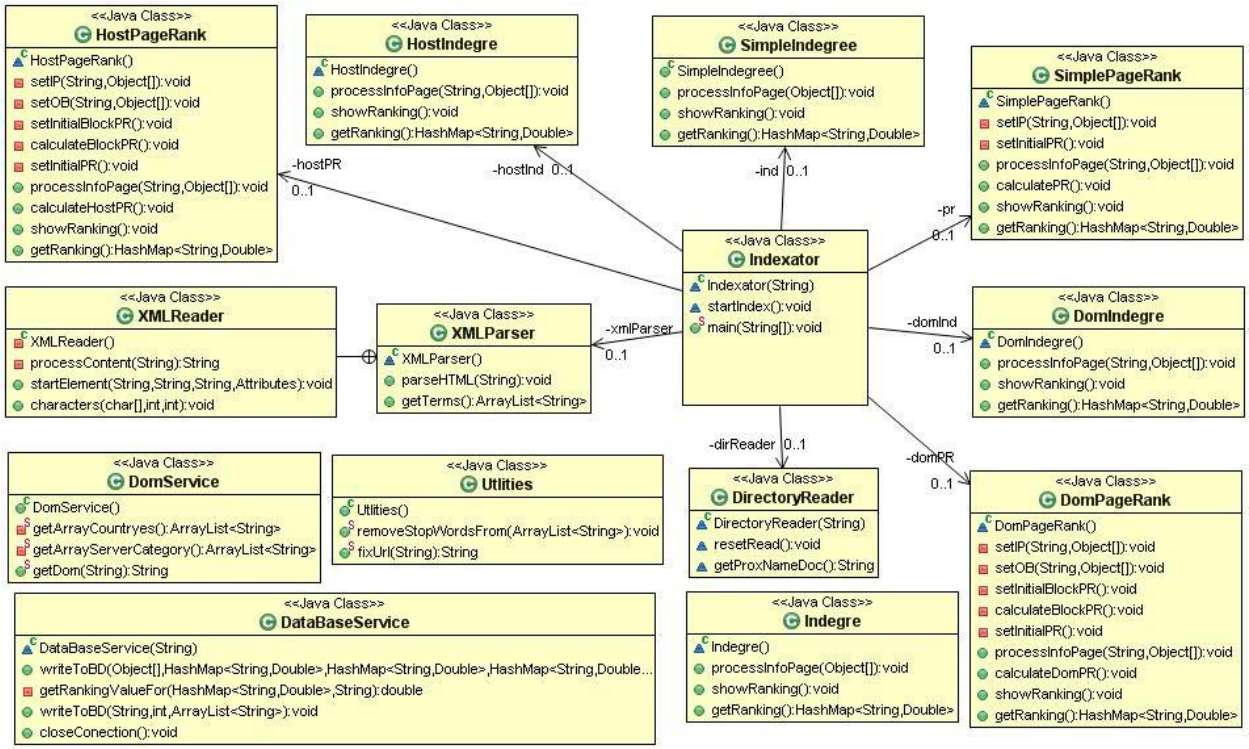


Diagrama de Clases Modulo Indexador

Tal como se puede apreciar en el diagrama anterior la estructura básica está compuesta por una clase principal llamada ***Indexador*** la que utiliza básicamente clases para manejo de XML, además de una clase dedicada a la utilización de base de datos funcionando como puente de conexión a MySQL, algunas clases de utilitarios y principalmente las clases dedicadas al procesamiento de cada método de Ranking desarrollado en el presente trabajo.

Diseño de clases: *Módulo Procesamiento de consultas navegacionales*

A continuación se detalla el diagrama UML *(Unified Modeling Language)* de relación entre los componentes físicos del modulo de procesamiento de consultas realizado en lenguaje PHP, dicho diagrama ***no es un diagrama de clases*** dado que la estructura del código fue desarrollado con metodología de ***Scripting***, es decir sin utilización de Clases, por este motivo las entidades aquí presentadas representan *archivos físicos y su interacción*. El diagrama fue generado con la herramienta open source llamada StarUML:

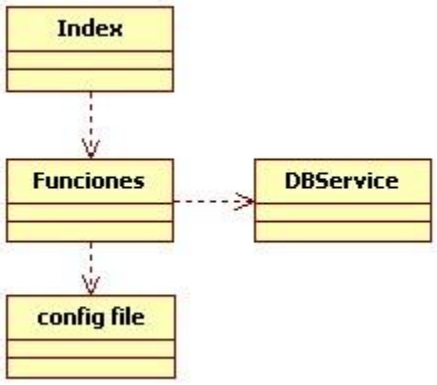


Diagrama interacción archivos físicos del modulo de procesamiento de consultas

Diseño de Base de Datos: DER (Diagrama de Entidad-Relación)

A continuación se detalla el diagrama DER de la estructura de tablas presentes en el modelo de persistencia seleccionado para el presente trabajo, dicho Diagrama fue realizado con la Herramienta SQLyog y cabe aclarar que la simplicidad del mismo se debe a la omisión de índices invertidos necesarios para realización de búsquedas informacionales de carácter optimo, las cuales aun así pueden ser realizadas con el presente diseño:

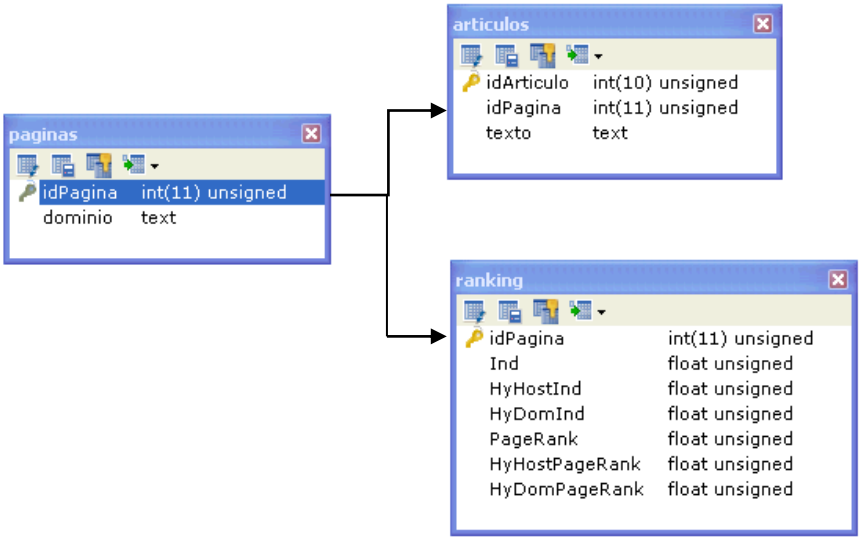


Diagrama de Entidad-Relaciones

A pesar de lo mencionado anteriormente, es decir la ausencia de índices invertidos o campos de texto de formato FULL TEXT para la realización de búsquedas informacionales, se introdujo una variación para poder llevar a cabo dichas consultas, aunque esta variación es posible solamente a el motor de MySQL que permite realizar una búsqueda booleana sobre uno o varios campos de textos de formato TEXT tal como se puede apreciar en las columnas ***dominio*** y ***texto.*** Solamente en carácter informativo se muestra a continuación la consulta implementada que resuelve dichas búsquedas:

```
SELECT
        idArticulo , MATCH ( dominio, texto ) AGAINST ( '*$busqueda*'IN BOOLEAN MODE ) AS Score
FROM
   articulos a, paginas p, ranking

WHERE
        MATCH ( dominio, texto ) AGAINST ( '*$busqueda*' IN BOOLEAN MODE )
        AND a.idPagina = p.idPagina
        AND p.idPagina = r.idPagina
GROUP BY idArticulo
ORDER BY r.$metodoRanking DESC, Score DESC
```

Donde las variables **$busqueda** y **$metodoRanking** indican el texto ingresado en el cuadro de búsqueda del modulo de consultas y el método de ranking seleccionado para la priorización de los resultados obtenidos respectivamente.

Capítulo 6:

Prueba de la solución propuesta

Set de datos para la realización de las pruebas

El set de datos que se utilizará para la realización de las pruebas del software desarrollado se compone de un directorio de páginas web previamente definidas sobre el cual se desarrollaran los procesos de indexación y las posteriores consultas. Utilizar un repositorio local de páginas en lugar de la web entera nos permite incluir páginas con el objetivo de no entorpecer los resultados, y entonces comprobar cómo reaccionan los métodos propuestos de una manera práctica y en el menor tiempo posible, al procesar una cantidad de documentos fijada, de acuerdo a nuestras necesidades.

Tamaño del set de datos

El tamaño del set de datos establecido fue de **20 páginas**, dada la particularidad de la situación y el análisis intensivo a realizar sobre las relaciones existentes entre cada componente del set, se estableció un numero de datos que presente un balance entre la necesidad de una administración manual del set y una cantidad representativa de datos para evaluar los métodos sin introducir perdidas de objetividad originadas por una mala selección del set muestral a utilizar.

Presentación del set de datos

A continuación se presenta una tabla con el set de páginas seleccionadas para la realización de las pruebas, mas adelante analizaremos su composición con respecto al grafo generado y los diferentes hipergrafos involucrados, como así también lo que denominaremos el Factor de conectividad (Fc).

IdPagina	Dominio	IdPagina	Dominio
1	www.BlogDePablo.com	11	www.GuitarraOnLine.com/Notas
2	www.Deportes.UOL.com.br	12	www.GuitarraOnLine.com/Patrocinantes
3	www.DOH.com	13	www.Juegos.UOL.com.br
4	www.Electrodom.com	14	www.Philips.com
5	www.ForoOpinion.com	15	www.Philips.com/Sedes
6	www.ForoOpinion.com/cafeteras	16	www.Philips.com/ServiceTecnico
7	www.ForoOpinion.com/DOH	17	www.Racing.com.ar
8	www.ForoOpinion.com/ForoLibre	18	www.UOL.com.br
9	www.ForoOpinion.com/philips	19	www.UOL.com.br/Destacados
10	www.GuitarraOnLine.com	20	www.ZonaNoticias.com.ar

Set de Paginas para pruebas

Factor de Conectividad

El Factor de conectividad es una medida que utilizaremos para medir el grado de relación en el grafo e hipergrafos utilizados en el siguiente modelo. El cálculo de dicho factor es la relación que existe entre la cantidad de links en dicho grafo y la cantidad de nodos utilizados como unidad de partición. Cabe aclarar que la unidad de partición varía en base a la cantidad de links que tiene en cuenta para el cálculo de los métodos de ranking tal como se describe a continuación:

$$Fc(G) = \frac{L}{U}$$

Donde los parámetros son los siguientes:

$G = Grafo\ interviniente.$
$L = Cantidad\ de\ links\ salientes\ en\ el\ grafo.$
$U = Cantidad\ de\ unidades\ de\ partición\ en\ el\ grafo.$

La idea de introducir este factor de conectividad se basa en la comparativa de los diferentes grafos e hipergrafos generados y si estos mantienen un nivel de conectividad de similar característica al modificar la unidad de partición.

Variación de la estructura de Grafos

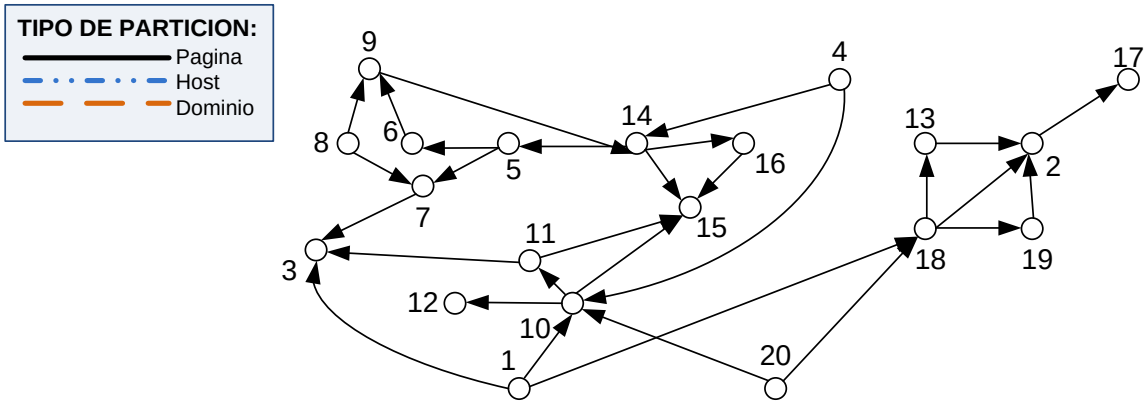
A continuación analizaremos la variación de la estructura de los grafos generados a partir de la modificación de la unidad de partición.

Unidad de Particion	*L*	*U*	*Fc*
Paginas	20	29	1.45
Host	11	16	1.4545
Dominio	9	13	1.4444

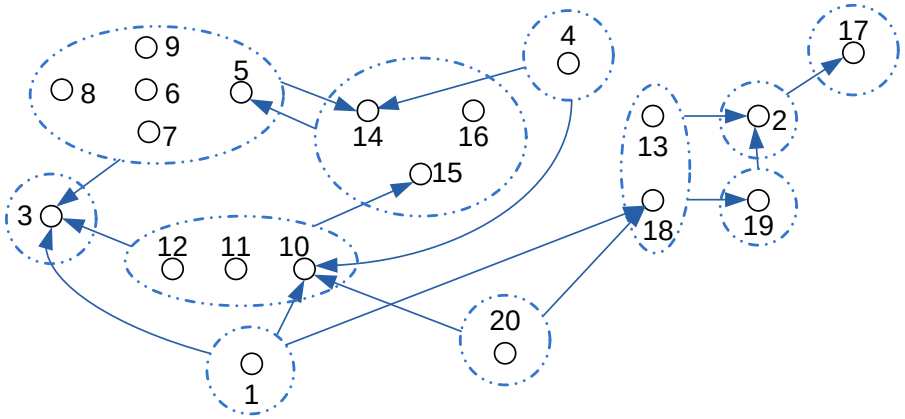
Análisis de Factor de Conectividad

Tal como se observa en la tabla anterior a pesar de las variaciones en la cantidad de links salientes y la cantidad de unidades de partición de cada grafo el factor de conectividad ronda dentro de un mismo valor y por lo tanto no se aprecia en un primer análisis perdida alguna de información.

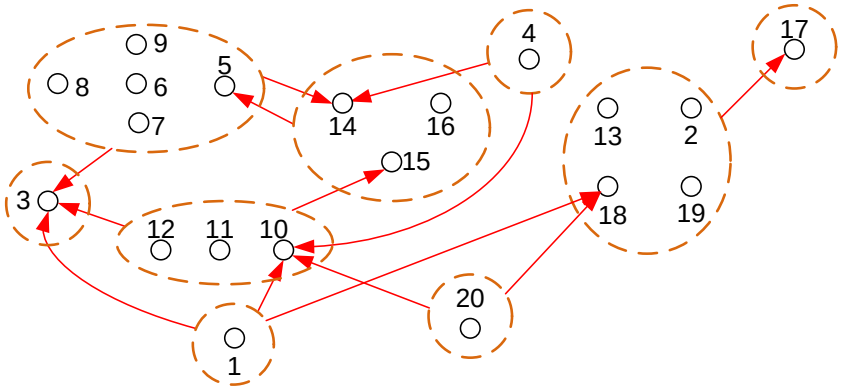
Analicemos ahora la estructura de los grafos para cada tipo de unidad de partición:



Grafo con tipo de Partición: Paginas



Grafo con tipo de Partición: Host



Grafo con tipo de Partición: Dominio

Tabla de precedencia

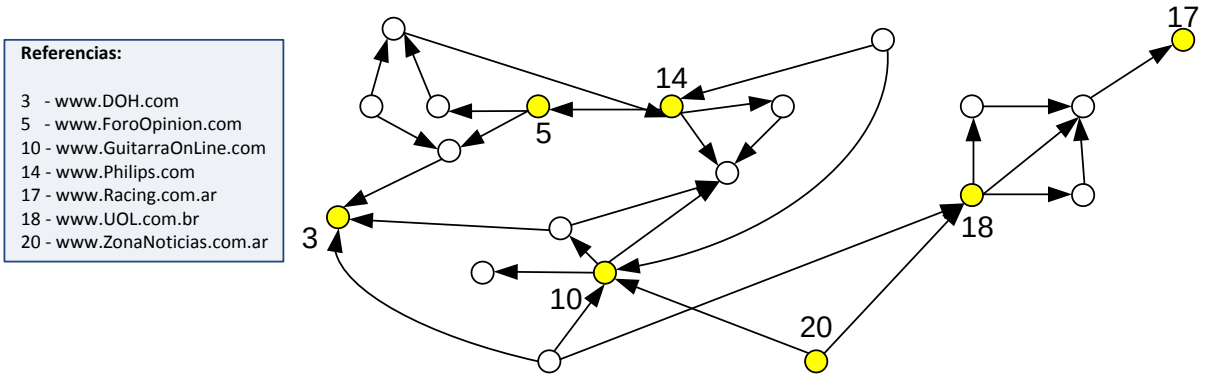
En la tabla que abajo se muestra, podemos observar las relaciones entre las páginas del set de datos desde otra perspectiva, en este caso analizándolo en el formato de una tabla de precedencias.

ID	ENVIA \ RECIBE	ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	# Links Salientes
1	www.BlogDePablo.com				✓							✓								✓			3
2	www.Deportes.UOL.com.br																		✓				1
3	www.DOH.com																						0
4	www.Electrodom.com											✓				✓							2
5	www.ForoOpinion.com							✓	✓														2
6	www.ForoOpinion.com/cafeteras										✓												1
7	www.ForoOpinion.com/DOH				✓																		1
8	www.ForoOpinion.com/ForoLibre								✓		✓												2
9	www.ForoOpinion.com/philips															✓							1
10	www.GuitarraOnLine.com												✓	✓			✓						3
11	www.GuitarraOnLine.com/Notas				✓												✓						2
12	www.GuitarraOnLine.com/Patrocinantes																						0
13	www.Juegos.UOL.com.br			✓																			1
14	www.Philips.com						✓										✓	✓					3
15	www.Philips.com/Sedes																						0
16	www.Philips.com/ServiceTecnico																✓						1
17	www.Racing.com.ar																						0
18	www.UOL.com.br			✓											✓						✓		3
19	www.UOL.com.br/Destacados			✓																			1
20	www.ZonaNoticias.com.ar											✓								✓			2
# Links entrantes			0	3	3	0	1	1	2	0	2	3	1	1	1	2	4	1	1	2	1	0	29

Tabla de precedencias

Selección de pruebas

A continuación se identificaran cuales son las páginas más relevantes a analizar y posteriormente se definirán los términos apropiados para la realización de las consultas navegacionales que se espera retorne cada una de las paginas en cuestión. Por tanto en primer lugar identificamos las páginas más representativas del grafo:



Identificación de páginas relevantes

Selección de términos de búsqueda

Una vez seleccionadas las páginas más relevantes definimos entonces los términos de búsqueda asociados al set de pruebas determinado con antelación.

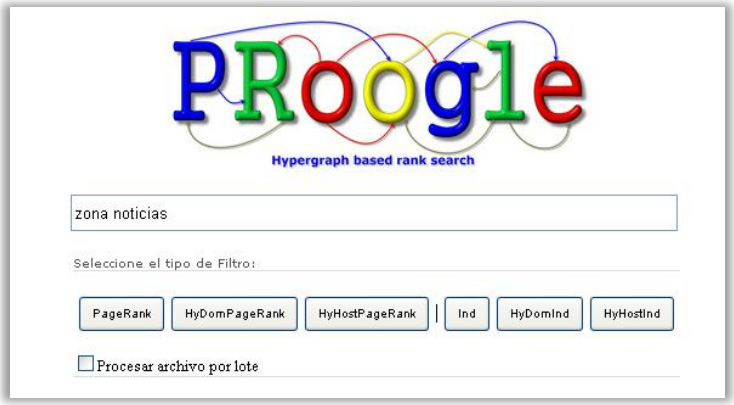
IdPagina	Dominio	Termino de Busqueda
3	www.DOH.com	*doh*
5	www.ForoOpinion.com	*foro opinion*
10	www.GuitarraOnLine.com	*guitarra on line*
14	www.Philips.com	*philips*
17	www.Racing.com.ar	*racing*
18	www.UOL.com.br	*oul*
20	www.ZonaNoticias.com.ar	*zona noticias*

Términos para el set de pruebas

Evaluando el set de pruebas

Con los términos definidos tenemos entonces completo el set de pruebas, el siguiente paso es realizar la búsqueda de los términos y registrar los resultados obtenidos para poder desarrollar las conclusiones. A continuación desarrollaremos paso a paso un ejemplo de prueba para el caso del término *"zona noticias"*.

El primer paso es ingresar en modulo de consultas navegacionales e introducir el término seleccionado en el cuadro de búsqueda tal como se muestra en la imagen siguiente:



Ejemplo de búsqueda

Una vez ingresado el término de búsqueda presionamos el tipo de método de ranking que deseamos utilizar para ordenar los resultados obtenidos, por defecto (al presionar **enter**) el método seleccionado es ***PageRank.***

Veamos ahora los resultados y el orden de estos para cada método de ranking utilizado:

PageRank*:** El método inicial obtiene la posición de la página buscada en ***3° lugar.


Cantidad de resultados: 3
Tipo de busqueda: **PageRank**

1 - www.Deportes.UOL.com.br
noticias Detalles España campeón mundo medirá México duelo norteamericanos sueñan primera victoria ibéricos Holanda viajará Ucrania tendrá chance enfrentar coanfitriones Eurocopa 2012 Estos duelos amistosos ofrecerán equipos europeos oportunidad medir nivel probar jugadores figuraron nóminas Mu ...
PageRank: 0.0218663

2 - www.UOL.com.br/Destacados
Las noticias importantes mundo Breve Racing le gano 0 all Boys hora lucir equipo ahora son promesas Finalmente Boca acuerda Riquelme Notas Deporte ...
PageRank: 0.0070125

3 - www.ZonaNoticias.com.ar
Zona noticias Lo electrodomesticos casa En seccion presentamos ultimas innovaciones respecto electrodomesticos hogar Todo necesitas hogar mejores lugares mano conocen Guitarras On line Deportes brasil Ronaldinho seguiria carrera profesional santos brasil info UOL Lo Futbol brasil ...
PageRank: 0

IdPagina	Dominio	PR
20	www.ZonaNoticias.com.ar	3°

Resultados método PageRank

HyDomPageRank: El método basado en dominio obtiene la posición de la página buscada en **2° *lugar*** es decir una mejora con respecto al método PageRank tradicional.


Cantidad de resultados: 3
Tipo de busqueda: **HyDomPageRank**

1 - www.Deportes.UOL.com.br
noticias Detalles España campeón mundo medirá México duelo norteamericanos sueñan primera victoria ibéricos Holanda viajará Ucrania tendrá chance enfrentar coanfitriones Eurocopa 2012 Estos duelos amistosos ofrecerán equipos europeos oportunidad medir nivel probar jugadores figuraron nóminas Mu ...
HyDomPageRank: 0.024

2 - www.ZonaNoticias.com.ar
Zona noticias Lo electrodomesticos casa En seccion presentamos ultimas innovaciones respecto electrodomesticos hogar Todo necesitas hogar mejores lugares mano conocen Guitarras On line Deportes brasil Ronaldinho seguiria carrera profesional santos brasil info UOL Lo Futbol brasil ...
HyDomPageRank: 0

3 - www.UOL.com.br/Destacados

IdPagina	Dominio	PR	DomPR
20	www.ZonaNoticias.com.ar	3°	2° ↑

Resultados método DomPageRank

HyHostPageRank: Finalmente el método basado en Host obtiene la posición de la pagina buscada en el ***1° lugar***, lo cual implica una mejora con respecto al **PageRank** y además una mejora sobre el método **HyDomPageRank**


Cantidad de resultados: 3
Tipo de busqueda: **HyHostPageRank**

1 - www.ZonaNoticias.com.ar
Zona noticias Lo electrodomesticos casa En seccion presentamos ultimas innovaciones respecto electrodomesticos hogar Todo necesitas hogar mejores lugares mano conocen Guitarras On line Deportes brasil Ronaldinho seguiria carrera profesional santos brasil info UOL Lo Futbol brasil ...
HyHostPageRank: 0

2 - www.Deportes.UOL.com.br
noticias Detalles España campeón mundo medirá México duelo norteamericanos sueñan

IdPagina	Dominio	PR	DomPR	HostPR
20	www.ZonaNoticias.com.ar	3°	2° ↑	1° ↑↑

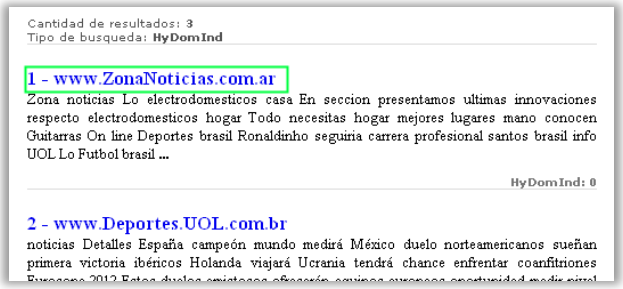
Resultados método HostPageRank

Indegree: El método Indegree ubica la pagina solicitada en la posición 3°.

IdPagina	Dominio	IND
20	www.ZonaNoticias.com.ar	3°

Resultados método Indegree

HyDomIndegree: El método basado en dominios ubica la pagina solicitada en la posición 1°, lo cual implica una mejora sobre el método Indegree.



IdPagina	Dominio	IND	DomIND
20	www.ZonaNoticias.com.ar	3°	1° ↑

Resultados método DomIndegree

HyHostIndegree: El método Indegree basado en host ubica la pagina solicitada en la posición 2°, lo cual implica una mejora sobre el método Indegree pero a su vez un resultado no tan bueno como el obtenido por el método Indegree basado en dominios.

IdPagina	Dominio	IND	DomIND	HostIND
20	www.ZonaNoticias.com.ar	3°	1° ↑	2° ↑↓

Resultados método HostIndegree

Finalmente analizaremos los mismos resultados para todo el set de datos determinado anteriormente indicando a través de símbolos para destacar la referencia de un método con sus diferentes variaciones, utilizando para tal fin las siguientes referencias:

Referencias:
→ Igual
↑ Mejor
↓ Peor

Cabe destacar que en la siguiente tabla se observa para cada página del set de pruebas escogido la posición de dicha pagina con respecto a los resultados obtenidos por la consulta navegacional correspondiente y en cuanto a las variaciones de los diferentes métodos los símbolos indican si se mantuvo igual si hubo un mejor resultado y en caso de los métodos basados en Host, indican la referencia con respecto a los otros métodos intervinientes dentro de su estructura de algoritmo respectiva.

IdPagina	Dominio	PR	DomPR	HostPR	IND	DomIND	HostIND
3	www.DOH.com	**1°**	1° →	1° →→	**1°**	1° →	1° →→
5	www.ForoOpinion.com	**3°**	1° ↑	1° ↑→	**3°**	2° ↑	2° ↑→
10	www.GuitarraOnLine.com	**2°**	2° →	2° →→	**1°**	1° →	1° →→
14	www.Philips.com	**2°**	1° ↑	1° ↑→	**3°**	1° ↑	1° ↑→
17	www.Racing.com.ar	**2°**	2° →	1° ↑↑	**2°**	1° ↑	2° →↓
18	www.UOL.com.br	**2°**	2° →	1° ↑↑	**2°**	1° ↑	2° →↓
20	www.ZonaNoticias.com.ar	**3°**	2° ↑	1° ↑↑	**3°**	1° ↑	2° ↑↓

Tabla comparativa de resultados

Tal como se pudo observar en la tabla anterior se destaca que en todos los casos se produce una mejora con respecto de la utilización de métodos basados en hypergrafos contra los métodos tradicionales basados en grafos, o en algunos casos simplemente se mantiene constante la posición de un resultado a través de los diferentes métodos. Razón por la cual se puede concluir que la utilización de hypergrafos en ninguna medida perjudica el proceso de rankeo y más aun en líneas generales lo mejora obteniéndose mejores resultados que con métodos tradicionales.

Evaluando el set de pruebas con métrica MRR

Previamente fueron definidos los detalles de la métrica MRR, pero para simplificar su comprensión se detalla a continuación el proceso de preparación del set de pruebas necesario para llevar a cabo la implementación y análisis de dicha herramienta de medición.

Como primer paso definimos un archivo al que llamaremos **MRR.txt** el cual contendrá los datos necesarios para realizar el procesamiento por lotes.

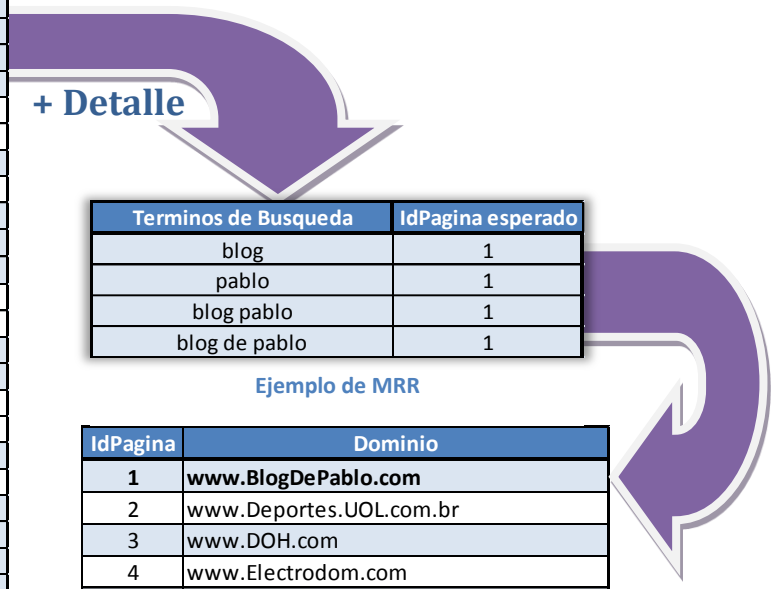
Estructura del archivo de procesamiento por lote MRR

La estructura del archivo necesario se basa en una cantidad considerable de las consultas más relevantes para cada uno de los sitios determinados en el set de pruebas, tal como se indicó al inicio del presente documento las consultas que se utilizaran son consultas navegacionales las cuales presentan como objetivo principal determinar un único sitio web como resultado esperado, ergo la estructuración del archivo de procesamiento se encuentra determinado por dos columnas, la primera indica los términos de la búsqueda actual y la segunda indica el ID numérico asociado al sitio al cual se intenta acceder mediante los términos correspondientes.

A continuación se define un ejemplo de la estructura y el presente archivo que se utilizará para llevar a cabo las pruebas de la métrica MRR.

Terminos de Busqueda	IdPagina esperado
blog	1
pablo	1
blog pablo	1
blog de pablo	1
deportes	2
uol deportes	2
doh	3
electrodom	4
opinion	5
foro opinion	5
foroopinion	5
cafeteras	6
foro cafeteras	6
foro doh	7
foro opinion doh	7
foro libre	8
libre	8
foro philips	9
guitarra on line	10
guitarra notas on line	11
notas guitarra online	11
guitarra notas	11
notas	11
guitarra patrocinantes	12
patrocinantes	12
guitarra on line patrocinantes	12
uol juegos	13
juegos	13
philips	14
philips sedes	15
sedes de philips	15
service	16
service tecnico	16
servicio tecnico	16
philips service tecnico	16
racing	17
racing club	17
uol	18
destacados	19
uol destacados	19
zona noticias	20
noticias	20

Tabla MRR



Terminos de Busqueda	IdPagina esperado
blog	1
pablo	1
blog pablo	1
blog de pablo	1

Ejemplo de MRR

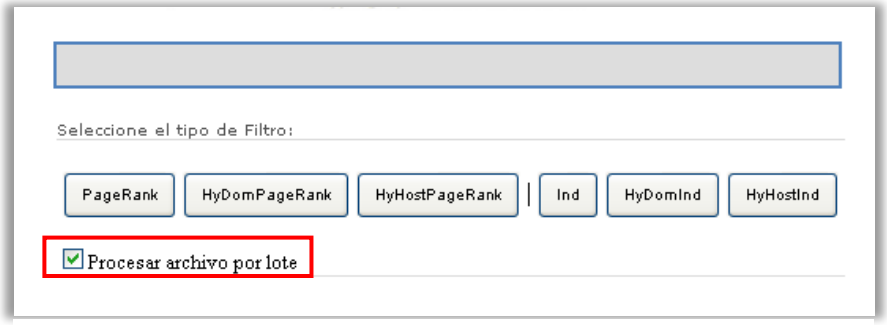
IdPagina	Dominio
1	**www.BlogDePablo.com**
2	www.Deportes.UOL.com.br
3	www.DOH.com
4	www.Electrodom.com
5	www.ForoOpinion.com
6	www.ForoOpinion.com/cafeteras
7	www.ForoOpinion.com/DOH
8	www.ForoOpinion.com/ForoLibre
9	www.ForoOpinion.com/philips
10	www.GuitarraOnLine.com
11	www.GuitarraOnLine.com/Notas
12	www.GuitarraOnLine.com/Patrocinantes
13	www.Juegos.UOL.com.br
14	www.Philips.com
15	www.Philips.com/Sedes
16	www.Philips.com/ServiceTecnico
17	www.Racing.com.ar
18	www.UOL.com.br
19	www.UOL.com.br/Destacados
20	www.ZonaNoticias.com.ar

Set de Paginas para pruebas

Dominio de referencia

Ejecutando la métrica MRR

Finalmente con el archivo de consultas definido, lo siguiente es llevar a cabo la ejecución del mismo a través de la interfaz de usuario en el modulo de consultas navegacionales, tal como lo hiciéramos con una consulta puntual esta vez simplemente seleccionamos el botón ***"Procesar archivo por lote"*** tal como se indica en la siguiente figura

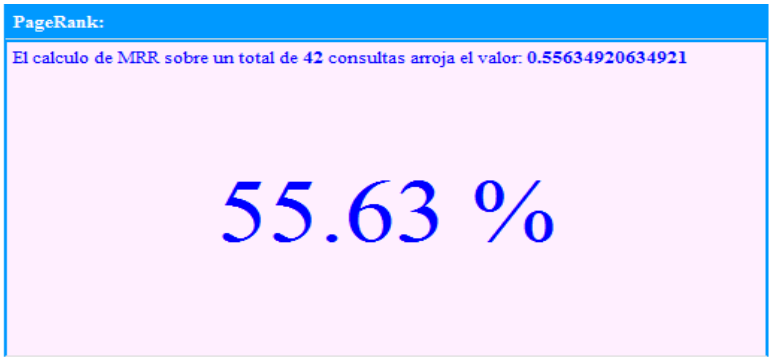


Procesamiento por lote

Tal como se puede apreciar en la imagen anterior al activar el procesamiento por lote el cuadro de búsqueda se deshabilita automáticamente, evitando así que el usuario ingrese dato alguno a la búsqueda y sirviendo además para identificar el método de procesamiento eliminando ambigüedades y cualquier subjetividad.

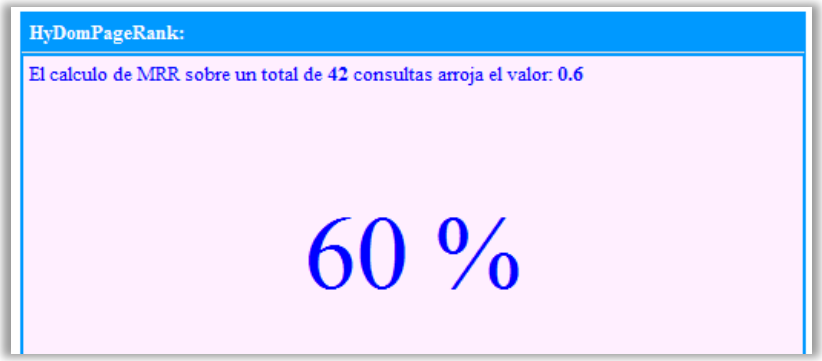
A continuación y tal como sucede para cada búsqueda normal debemos seleccionar un método de ranking y analizar los resultados obtenidos, los cuales se muestran en las siguientes capturas.

PageRank:



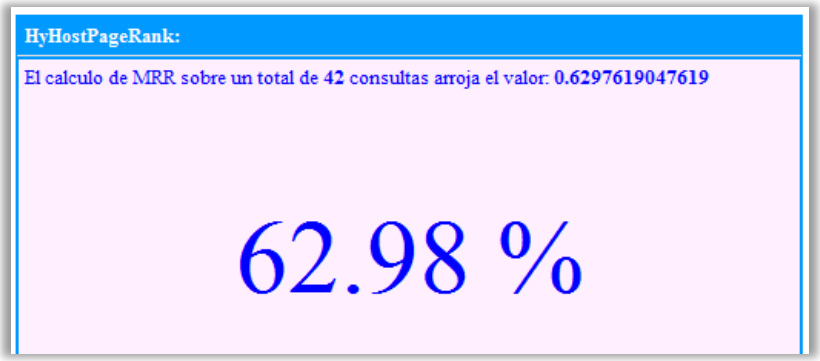
MRR PageRank

Hyper-Domain PageRank:



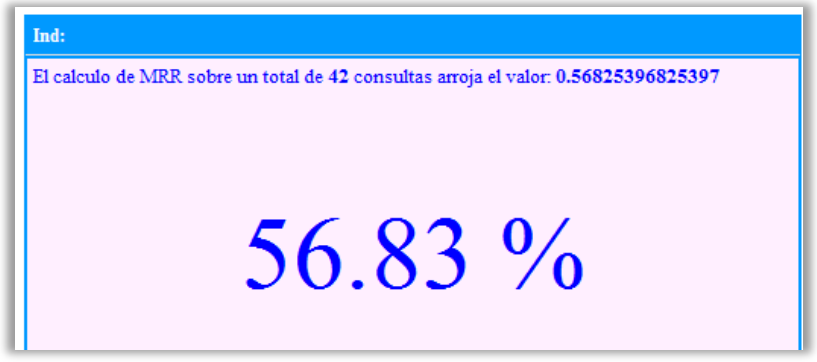
MRR Hyper-Domain PageRank

Hyper-Host Page Rank:



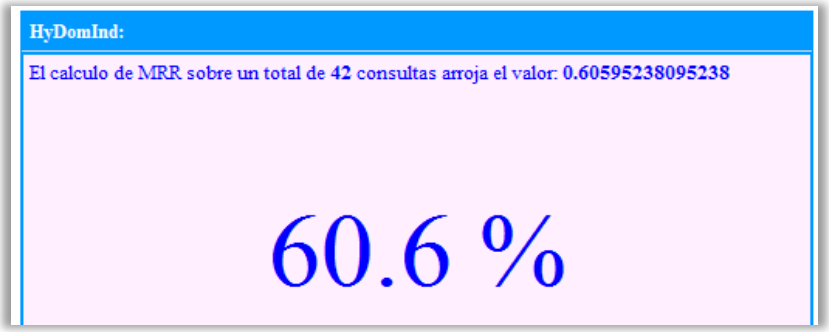
MRR Hyper-Host PageRank

Indegree:



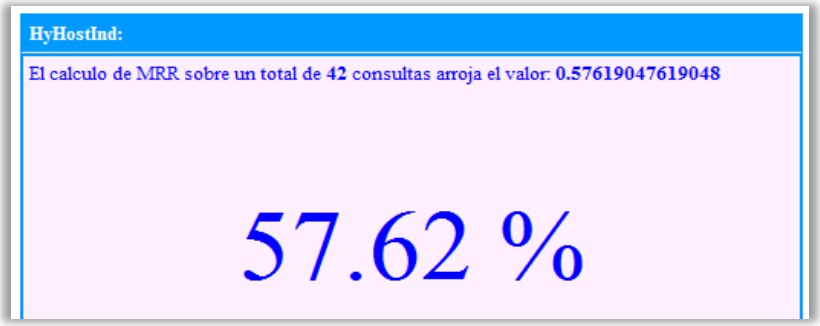
MRR Indegree

Hyper-Domain Indegree:



MRR Hyper-Domain Indegree

Hyper-Host Indegree:



MRR Hyper-Host Indegree

Como conclusión de la utilización de la métrica MRR observamos que la utilización de Hipergrafos representa una mejora con respecto a los métodos de ranking tradicionales estudiados.

Finalmente observamos un cuadro comparativo donde se aprecian las mejoras introducidas para cada método utilizado en detalle.

Metodo utilizado	Unidad de Particion	MRR		
PageRank	Paginas	0.55634920634921	-	**55,63%**
HyDomPageRank	Dominio	0.6	↑	**60,00%**
HyHostPageRank	Host	0.6297619047619	↑	**62,98%**

Comparativa métrica MRR - PageRank

Metodo utilizado	Unidad de Particion	MRR		
Indegree	Paginas	0.56825396825397	-	**56,83%**
HyDomIndegree	Dominio	0.60595238095238	↑	**60,60%**
HyHostIndegree	Host	0.57619047619048	↑	**57,62%**

Comparativa métrica MRR - Indegree

Capítulo 7:

Conclusiones y Futuras Líneas de Trabajo

La principal ventaja de utilizar un modelo hipergrafos en técnicas para la reputación de páginas web es permitir que el modelo controle la calidad de las conexiones presentes en las páginas web que estamos evaluando. Este control se logra mediante la adecuada definición del criterio de partición que se usa para crear hiperarcos. Esta flexibilidad abre la oportunidad de realizar más estudios para *determinar nuevos y mejores criterios de partición*, y permite a los diseñadores de motores de búsqueda poder elegir la mejor abstracción de hipergrafos para su colección de destino.

Los experimentos que llevamos a cabo han demostrado que el modelo de hipergrafos se puede utilizar para proporcionar una mejor estimación de la reputación de la página y mejorar el ranking de búsqueda final del motor. Hemos estudiado tres métodos de partición distinta derivada de la jerarquía URL:

- Basado en páginas
- Basado en host
- Basados en dominios

En los experimentos realizados con el motor de búsqueda de base de datos a partir de los algoritmos de análisis de enlaces utilizando el modelo hipergrafo proporcionan mejores resultados para consultas navegacionales que los obtenidos mediante el modelo de tradicional, sin pérdida de información para las consultas. Este es un ejemplo de las posibles ventajas de usar el modelo hipergrafo.

Una desventaja del uso de los modelos hipergrafo es que el número de hiperarcos tiende a ser menor que el número de arcos en la colección. En algunas situaciones el modelo hipergrafo puede causar una pérdida en la calidad de los resultados de búsqueda. Sin embargo, hay que tener en cuenta que los experimentos fueron realizados sobre una colección con un bajo grado de información por la falta de una base de datos de prueba más extensa, motivo por el cual mostraron algunas deficiencias en el modelo que podrían ser suplidas con la futura evaluación sobre una base de datos real.

En futuras investigaciones, se plantea la intención de investigar más a fondo la posibilidad de utilizar algoritmos de agrupamiento como las estrategias de partición. La idea es determinar las particiones basadas en las propiedades deseadas del hipergrafo, tales como el contenido y la independencia de la relación entre las páginas, en lugar de utilizar sólo la jerarquía de direccionamiento como una guía. Otra dirección es el estudio de las posibles correlaciones entre el mejor criterio de partición y las características de la recogida, además del uso de otras colecciones web disponibles para realizar los experimentos.

Capítulo 8:

Bibliografía

1- A Hypergraph Model for Computing Page Reputation on Web Collections

2- Uso de Información Semántica para la mejora de la Recuperación de Información en la Web

3- Análisis comparativo de las herramientas de programación Web

4- Relational link-based ranking

5- Modeling the web as a hypergraph to compute page reputation

6- Link Analysis for Private Weighted Graphs

7- Link Analysis and Web Search (Capítulo 14)

8- Exploiting PageRank at Different Block Level

9- Web Page Scoring Based on Link Analysis of Web Page Sets