UNIVERSIDAD DE BUENOS AIRES

Introducción del Modelo de Hipergráfos para técnicas de reputación web

75.39 – Aplicaciones Informáticas

Facultad de Ingeniería

2° CUATRIMESTRE

2011

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

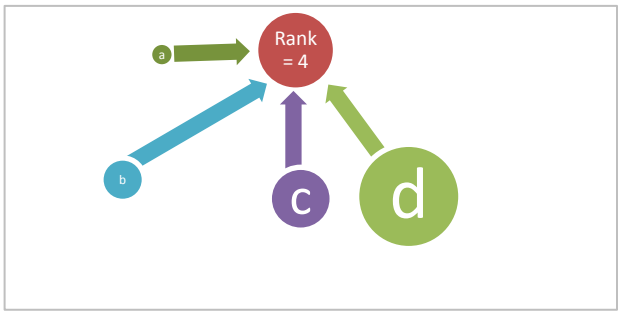
Estos métodos de ranking son el principal objetivo de estudio del presente trabajo. En este tipo de análisis se asigna un puntaje a cada página al igual que en el método antes descripto, pero este puntaje es independiente de los términos de búsqueda consultados.

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.

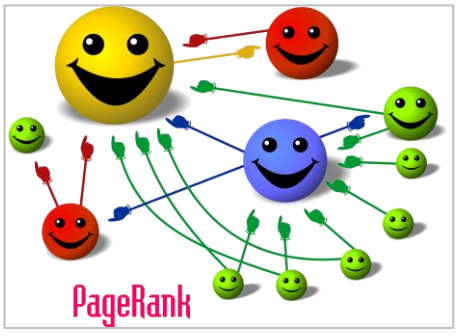
En general, los métodos de reputación se utilizan en combinación con otros métodos cuyo objetivo es el de buscar los sitios y documentos que cumplen con un determinado criterio de búsqueda. Una vez obtenidos los sitios que mejor se ajustan al criterio de la consulta del usuario, estos son ordenados según 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

$\mathrm{I(p)}$: Set de páginas que apuntan a p

$||\mathrm{O(q)}||$: Número de páginas apuntadas por q

$||\mathrm{r}||$: 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 Hipergráfo 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 Hipergráfo

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".

Capítulo 3:

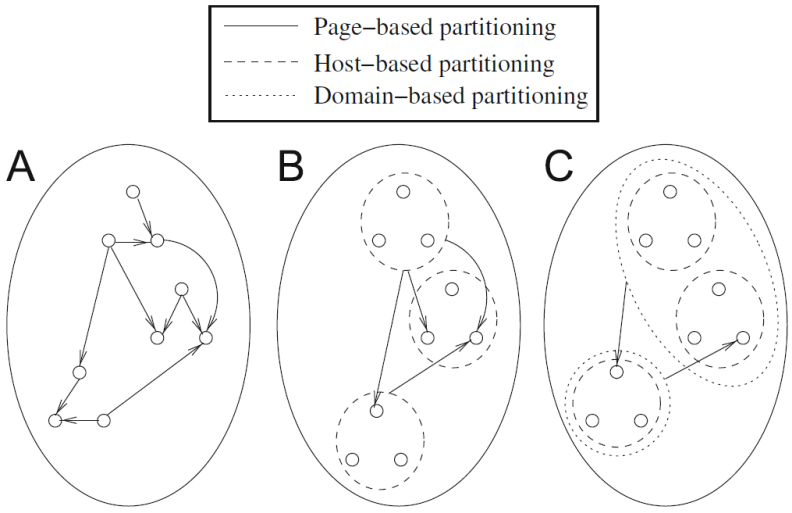
Presentación del Problema a resolver

El eje central de la idea expuesta en el presente trabajo se basa en el criterio de agrupamiento que se utiliza en los métodos de reputación web y análisis de links, por tanto el principal objetivo propuesto o problema a resolver consta en la aplicación del concepto de hipergráfos sobre los métodos tradicionales de reputación, de su implementación y de los posteriores criterio de agrupamiento seleccionados dependerá el éxito o fracaso de la introducción de sendas mejoras basadas en el análisis de los resultados obtenidos.

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 hipergráfos que surjan de estos tres criterios será utilizado como base para los métodos *Indegree* y *PageRank*.



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

La aplicación del hipergráfo en un método sencillo como el Indegree, nos ayuda a comprobar más fácilmente como la construcción del hipergráfo 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 hipergráfo 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 Hipergráfo, implica una modificación a las ecuaciones que describen las técnicas de Page Rank e Indegree. Dichas modificaciones se analizarán más adelante como parte de la presentación del problema a resolver.

Aplicación del Hipergráfo al Método Indegree

Dada la simpleza matemática del método Indegree, las modificaciones necesarias para la utilización del Hipergráfo 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 Hipergráfo, 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 Hipergráfo 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:

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

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 la siguiente fórmula para evaluar la calidad de los resultados obtenidos:

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

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

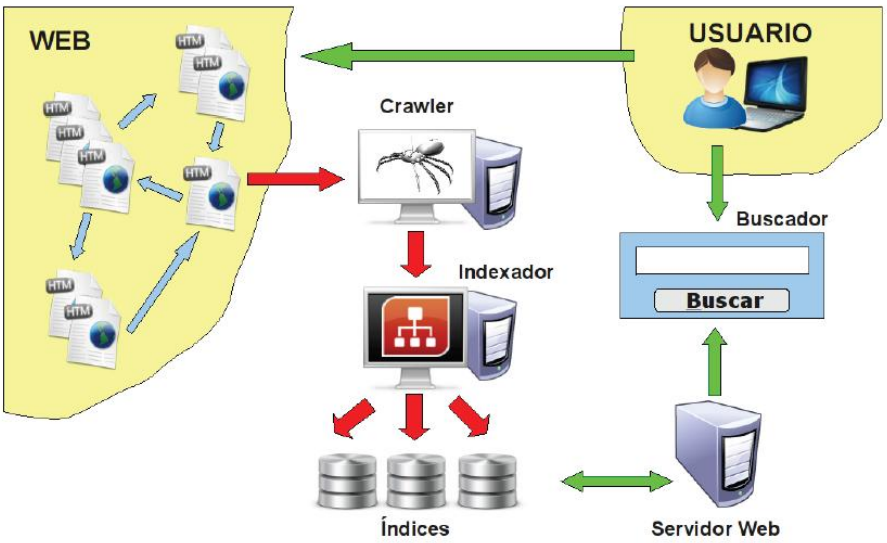
Se busca poder comprobar las mejoras que introduce la utilización del hipergráfo en los métodos de rankeo de páginas. Para poder realizar dicha tarea, se aplicara su construcción, como base del método de PageRank, el cual es frecuentemente utilizado en la realidad. También se aplicara el hipergráfo con 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 Hipergráfo y poder sacar conclusiones.

Diseño del Software

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

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 Hipergráfo 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)
2. HyHostPR (hipergráfo)
3. HyDomPR (hipergráfo)
4. Indegree (grafo)
5. HyHostInd (hipergráfo)
6. HyDomInd (hipergráfo)

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)
- ✓ Firefox Browser (Web Browser)
- ✓ Ubuntu (Sistema operativo Linux)

Capítulo 6:

Prueba de la solución propuesta

Set de Pruebas

El set de pruebas para el software desarrollado se compondrá de un directorio de páginas web sobre el cual se desarrollaran las consultas. Utilizar un repositorio local de páginas en lugar de la web entera nos permite incluir páginas que tengan como objetivo 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.

Ejemplo de prueba:

Termino consultado: ***Racing***

Metodo	Posición del sitio
PageRank	2
Dom PageRank	2
Host PageRank	1
Indegree	3
Host Indegree	2
Dom Indegree	1

Se observo que las pruebas sobre el método MRR favorecen a los métodos de *PageRank* e *Indegree* por sobre los otros métodos, esto debe estar basado en que en general ambos métodos otorgan un mayor ranking a paginas que no deberían tener en cuenta los links internos, de dominio o del mismo host.

Capítulo 7:

Conclusiones y Futuras Líneas de Trabajo

La principal ventaja de utilizar un modelo hipergráfos 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 hipergráfos para su colección de destino.

Los experimentos que llevamos a cabo han demostrado que el modelo de hipergráfos 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 hipergráfo 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 hipergráfo.

Una desventaja del uso de los modelos hipergráfo es que el número de hiperarcos tiende a ser menor que el número de arcos en la colección. En algunas situaciones el modelo hipergráfo 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 hipergráfo, 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