Universidad Autónoma Metropolitana Unidad Azcapotzalco

División de Ciencias Básicas e Ingeniería

Licenciatura en Ingeniería en Computación

Propuesta de Proyecto Terminal:

Sistema de recuperación de información de textos de investigación de la Web

Luis Yamil García Jurado

Matricula: 207330149

Trimestre: 2012 Otoño

31 de octubre de 2012

Segunda versión

Asesores

Maricela Claudia Bravo Contreras

Profesor Asociado, Departamento de Sistemas

María Lizbeth Gallardo López

Profesor Asociado, Departamento de Sistemas

Objetivo general

Diseñar un sistema basado en arañas Web focalizadas que recuperen información de contenidos de investigación públicos en la Web, el cual reciba como entrada varias palabras clave para realizar la búsqueda y recuperar los documentos relacionados mas relevantes.

Objetivos específicos

- Diseñar un sistema basado en arañas web que implemente técnicas de búsqueda inteligente para encontrar textos relacionados a ciertas palabras clave.
- Construir un modulo de minería, que se encargue de minar los datos para escoger los documentos más relevantes encontrados.
- Implementar una interfaz que permita realizar la búsqueda y visualizar la información adquirida, de una manera sencilla para el usuario.

Introducción

Arañas web

Una araña web (también conocida como *crawler*) es un pequeño software que recorre el entramado de páginas web de Internet de forma automática y sistemática.
Una araña web es un tipo especializado de robot de la web que se encarga de llevar a cabo un tipo concreto de tareas, en particular, de recorrer las páginas web, analizar su información y hacer exploraciones de las páginas que encuentre referenciadas mediante una URL para pasearlas y procesarlas. Pueden ser utilizadas con fines diversos, aunque su uso más conocido es el de agente software en los motores de búsqueda, donde su función básica es proporcionar al indizador el contenido apropiado para ser indizado, según distintos criterios e intereses.

Se tienen 2 tipos básicos de arañas web que son:

1. Arañas web generales
   Son la forma más básica de una araña web y se caracterizan por visitar todos los sitios web, sin hacer distinción entre ellos; es decir, dado un conjunto de URLs en una cola de descarga, por cada una obtiene las URL que ésta contenga y las añade a la misma cola de descarga para después procesarlas, y sigue con este procedimiento hasta un determinado punto.
   Pero al no hacer distinción entre los documentos, se tienen ciertas desventajas, como son: el saturar el ancho de banda con demasiadas descargas y obtener documentos con muy poca o nula relevancia para la búsqueda deseada.

2. Arañas web focalizados
   Estos funcionan de manera distinta, pues hacen su cola de descargas de acuerdo a una serie de prioridades; es decir, asignan prioridad a cada enlace no visitado

estimado en un valor de la página enlazada. Estas prioridades se calculan según su relación con un conjunto de palabras clave ingresadas como búsqueda.
En este tipo de arañas web se busca evitar descargas innecesarias mediante la relevancia de una página en la búsqueda, pero al mismo tiempo tienen como desventaja el tiempo de procesamiento, que resulta ser mucho más costoso.

Minería de datos

La minería de datos (conocida como Knowledge Discovery in Database o KDD) se puede definir como el proceso de descubrimiento de patrones útiles, correlaciones significativas, tendencias, además de obtener conocimiento de fuentes de datos mediante Estadística, Inteligencia Artificial, entre otros.
El análisis de los datos que se hace mediante esta disciplina aprovecha la información que se encuentra en las bases de datos, permitiendo hacer predicciones y ayudando en la toma de decisiones. El minado de datos comprende 3 etapas:

1. Pre-procesado: Puede ser que los datos no estén en condiciones adecuadas para minarse, así que esta etapa se encarga de hacer una reducción de los datos para quedarse con los atributos necesarios y remover el ruido que pudieran tener.
2. Minado de datos: Aquí se encuentra el algoritmo encargado de extraer los patrones de los datos pre-procesados.
3. Post-procesado: Aquí se determina que patrones son útiles, según la aplicación que los requiera.

El proceso de minería de datos se realiza iterativamente y debe realizarse repetidas veces para conseguir resultados útiles.

Justificación

En la actualidad se busca siempre la optimalidad. Se intenta tener siempre los mejores resultados posibles para una tarea dada y la búsqueda de información en la web es un buen ejemplo de esta continua búsqueda de eficacia.

Se tienen hoy en día, algunos buscadores web que utilizan arañas web para sus búsquedas, pero estos utilizan algoritmos muy generales para realizar su tarea, por lo que para ciertos temas de búsqueda no se obtienen los resultados óptimos que se esperaría recibir; es aquí donde este proyecto toma dirección, pues se busca combinar algunas técnicas de minería de datos para encontrar aquellos documentos que sean más relevantes en temas de investigación
Así, con la construcción de este sistema se contará con una aplicación que encuentre mejores resultados para búsquedas en la web sobre temas de investigación y aprendizaje, cosa que los algoritmos de las arañas web actuales no siempre consiguen.

Trabajos Relacionados

Proyectos terminales

1. Sistema de recuperación de información semántico. Este proyecto tiene ciertas similitudes con el propuesto en este documento ya que se encarga de recopilar, minar y descubrir información relevante sobre un conjunto dado de documentos; de la misma forma como las arañas web encuentran páginas relacionadas según ciertos criterios de búsqueda. Pero se nota fácilmente sus diferencias, pues el sistema de recuperación de información semántico busca información dentro de los documentos y es esa información la que procesa [1].

Tesis

1. Método general de Extracción de información basado en el uso de Lógica Borrosa. Aplicación en portales web. Esta tesis propone la creación de un agente inteligente que ayude al usuario en su proceso de encontrar información relevante, la manera en que este agente logra su cometido es donde radica la principal diferencia con nuestro proyecto, pues en esta tesis proponen la lógica difusa para la gestión de la información, mientras que nosotros haremos uso de las arañas web. [2].
2. Buscador basado en contenido XML de las páginas de centros y departamentos de la UV. Este es un proyecto que se encarga de buscar páginas dentro de la Universidad de Valencia, partiendo de una serie de palabras clave, obteniendo resultados precisos a lo que el usuario pide. Este proyecto es similar al propuesto debido a su búsqueda en base a ciertas palabras clave, pero este buscador basado en XML tanto solo busca dentro de una universidad y no en la web como tal, además de que lo hace mediante los ficheros XML de los documentos a recuperar [3].

Propuestas de proyectos terminales

1. Sistema de procesamiento de textos de investigación. Esta propuesta se asemeja al sistema propuesto en este documento por el uso de un texto de entrada como referencia para extraer la información de los textos fuente. Pero su diferencia radica en que el sistema de procesamiento de textos de investigación, extrae la información solo para hacer una lista de etiquetas semánticas [4].

Software

1. Googlebot. Este es el programa que usa el buscador de Google para descubrir páginas nuevas y actualizadas que añade a su índice mediante un proceso que se llama: rastreo. Este programa usa algoritmos basados en arañas web, que determinan qué sitios rastrear, qué tan seguido y cuántas páginas rastrear para cada sitio; esta técnica esta basada en los mismos fundamentos que usaremos en el proyecto, pero aplicados a documentos de investigación.
2. Yahoo! Slurp. Esta es la araña web del buscador Yahoo! Que obtiene contenido del mecanismo Yahoo! Search. El programa Slurp está basado en la tecnología de búsqueda de Yahoo! y se encarga de indexar todas las paginas recolectadas por su mecanismo de búsqueda, tal y como realizaremos en nuestro proyecto, aunque Slurp

solo se encarga de indexarlas sin hacer un mayor sistema de filtrado, cosa que si haremos en nuestro proyecto propuesto.

Descripción Técnica

El fin de este proyecto es el de proporcionar un sistema que, dada una cierta búsqueda, arroje los resultados más relevantes encontrados, utilizando para esto una araña web focalizada, que recorra las páginas; así como un minador de datos que encuentre los mejores resultados para el usuario.
En la figura 1. Se puede ver un esquema del funcionamiento del sistema.

1. Buscador
   Este módulo se encarga de capturar los datos ingresados por el usuario para proporcionarlos a la araña web focalizada.

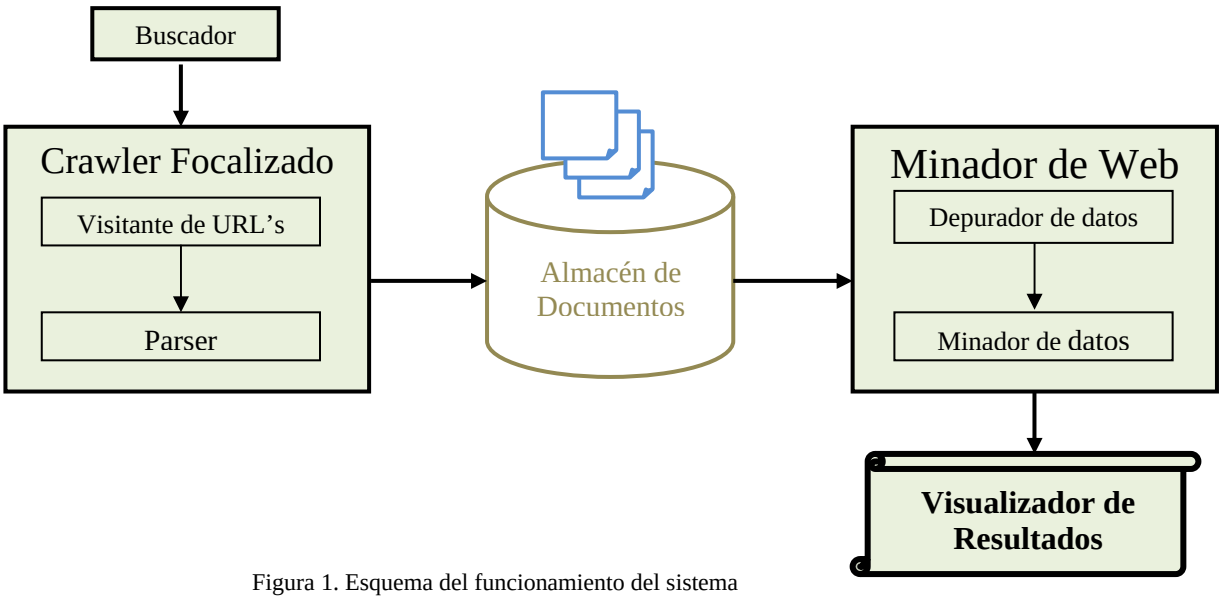

Figura 1. Esquema del funcionamiento del sistema

2. Araña web focalizada
   Este modulo se divide en 2 sub-módulos:
   - Visitante: Se encargará de revisar todas las URL's dadas como inicio para partir la búsqueda y recolectar los enlaces que tengan a otras páginas. Realizará una búsqueda en amplitud, mandando al modulo parser los enlaces para seguir siendo procesados.
   - Parser: Este módulo se encargará de visitar y abrir la página de cada URL que se encuentre en la cola de páginas a revisar y comparará el contenido de la página con las palabras clave para determinar si la página se recupera o no.

3. Almacén de documentos
   Este se divide en 2 sub-módulos:
   - Administrador de documentos: Este módulo consistirá de un conjunto de interfaces de programación para el acceso (consulta y actualizaciones) a los documentos de investigación recuperados.
   - Almacén: En este módulo se descargarán todos los documentos de investigación que se hayan descargado con la araña web focalizado.

4. Módulo de minería web
   Este módulo se divide en 2 sub-módulos:
   - Depurador de datos: Las páginas descargadas serán depuradas mediante un proceso de pre filtrado que discriminara si una pagina es relevante o no, teniendo en cuenta las palabras claves y el orden en que se escribieron, preparando las seleccionadas para su posterior proceso de minado.
   - Módulo de aprendizaje de datos: Este módulo integrará un método de aprendizaje no supervisado de tipo estadístico para realizar el agrupamiento de datos.
5. Visualizador de resultados
   En este modulo se despliegan los resultados conseguidos al usuario.

Especificación Técnica

Este proyecto propuesto se realizará usando lenguaje java en el entorno de desarrollo NetBeans IDE7.2.
Para el desarrollo de la araña web se usará la API “Java Web Crawler” que es software libre.
Para la realización de la minería de datos se usará “Bixo”, una herramienta de minería web de software libre.

Este proyecto se dará por concluido cuando el sistema regrese al menos los 10 primeros resultados con alta relevancia para la búsqueda dada.
El sistema se probará con 15 temas distintos y se realizará la evaluación de los resultados con las mediciones de “Precision and Recall” [5].
En términos de recuperación de información y reconocimiento de patrones, hay 2 estándares para conocer el grado de relevancia de un resultado: precisión y recuperación (precision and recall). La precisión es la fracción de instancias recibidas que son relevantes, es decir, nos dice cuántas de las páginas recuperadas son realmente importantes; mientras que la recuperación se refiere a la fracción de instancias relevantes que son recibidas, es decir, de todas las páginas importantes, cuantas se recuperaron. Se evaluaran los resultados midiendo estos patrones ya que es la manera más fácil de saber cuántos de los resultados fueron relevantes y cuántos de los relevantes obtuvimos.

Al concluir el proyecto terminal se entregarán tres discos compactos al Coordinador de Estudios de Ingeniería en Computación que incluirán el reporte final del proyecto terminal en un archivo PDF (sin restricciones) y el código fuente de la aplicación en un archivo comprimido (sin restricciones). El reporte final contendrá al menos: portada, resumen, tabla

de contenido, objetivos, introducción, desarrollo del proyecto, conclusiones, bibliografía y apéndices. Los apéndices contendrán al menos un listado del código fuente desarrollado.

Calendario de Trabajo

Enseguida se describe el calendario de trabajo para este proyecto dividido en 2 trimestres correspondientes a los 9 créditos (99 horas, 9 por semana) del Proyecto Terminal de Ingeniería en Computación I y el otro correspondiente a los 18 créditos del Proyecto Terminal de Ingeniería en Computación II (198 horas, 18 por semana).

Trimestre 13-I	1	2	3	4	5	6	7	8	9	10	11	Horas
Instalación y configuración de herramientas a utilizar												9
Diseño del araña web												18
Diseño e implementación del visitante												18
Diseño e implementación del parser												27
Diseño del almacén de datos												18
Primera revisión												9

Trimestre 13-P	1	2	3	4	5	6	7	8	9	10	11	Horas
Diseño del minador web												34
Diseño e implementación del depurador de datos												34
Diseño e implementación del Módulo de aprendizaje de datos												52
Diseño de la interfaz gráfica												34
Segunda revisión y pruebas												16
Ajustes finales												16
Elaboración del reporte final												22

Recursos

- Software. El ambiente de desarrollo y las APIs necesarias son software libre y se cuenta con una conexión a Internet de 2 MB, suficiente para las pruebas de conexión necesarias.
- Hardware. Se cuenta con una computadora personal con procesador AMD a 2.4 GHz, 4 GB de memoria RAM y 500 GB de disco duro, características suficientes para la terminación del proyecto.

Los asesores se responsabilizan de guiar al alumno y de que todos los recursos anteriormente citados estarán disponibles para el alumno, de modo que el proyecto terminal se pueda concluir en tiempo y forma.

Maricela Claudia Bravo Contreras	María Lizbeth Gallardo López
_____	_____

Bibliografía

[1] Ugalde Chávez, Selene M. de J., Noviembre del 2011, Sistema de recuperación de información semántico, proyecto terminal, Disponible en: http://cpti.azc.uam.mx/ProyectosTerminales/Propuestas/PropuestaSelene.pdf
[2] Ropero Rodríguez, Jorge, Noviembre del 2009, Método general de Extracción de información basado en el uso de Lógica Borrosa. Aplicación en portales web, Disponible en: https://www.educacion.gob.es/teseo/imprimirFicheroTesis.do?fichero=16217
[3] Arce Morell, D. Andrés, 2009, Buscador basado en contenido XML de las páginas de centros y departamentos de la UV, Disponible en: http://www.docstoc.com/docs/23968149/PFC-Buscador-basado-en-XML-para-UV
[4] Bravo Contreras, Maricela C., Septiembre del 2012, Sistema de procesamiento de textos de investigación, propuesta de proyecto, Disponible en: http://cpti.azc.uam.mx/ProyectosTerminales/PropuestasProyectos.html
[5] Middleton, Christian y Baeza-Yates, Ricardo, “A Comparison of Open Source Search Engines”, reporte tecnico, 2007.