Universidad Autónoma Metropolitana

Unidad Azcapotzalco

División de Ciencias Básicas e Ingeniería

Ingeniería en Computación

“Sistema configurable de Minería Web”

Alumno: Urquiza Pérez Alina Firma: _____
Matricula: 207201853
Trimestre lectivo. 11-O
Fecha de Entrega: 11 de Noviembre del 2011
Versión: Primera Versión

Asesor: Silva López Rafaela Blanca Firma: _____
Departamento: Sistemas
Categoría: Titular B -Medio Tiempo

Asesor: Maricela Claudia Bravo Contreras Firma: _____
Departamento: Sistemas
Categoría: Titular C -Tiempo Completo

Objetivo General

Implementar un sistema de minería de contenido Web configurable que permita copiar documentos de la Web a directorios locales con el propósito de descubrir patrones de información ocultos entre los documentos.

Objetivos Específicos

- Crear un sistema configurable que permita especificar una URL ó IP, dominio, el directorio local de almacenamiento de los documentos y condiciones de paro, para recopilar documentos de la Web.

- Convertir los documentos extraídos en el proceso de recuperación de información, en documentos que sean más digeribles, fáciles de leer y de analizar.

- Aplicar una técnica de minería Web con el propósito de descubrir automáticamente patrones ocultos sobre cada uno de los sitios recopilados.

- Diseñar una interfaz para entender, visualizar e interpretar los patrones.

Introducción

La World Wide Web (también conocida como la Web) es un sistema de navegación de páginas electrónicas. El crecimiento acelerado de Internet y especialmente de la Web, ha hecho cada vez más imprescindible para los usuarios utilizar herramientas para encontrar, extraer, filtrar y evaluar toda la información disponible. La Web se ha convertido en el medio de comunicación más popular entre usuarios porque facilita enormemente el intercambio de información. Entre las aplicaciones más usadas en la Web se encuentran los buscadores, por ejemplo Google. Tradicionalmente las personas utilizan los buscadores para localizar recursos de información. Sin embargo, muchas ocasiones los usuarios quedan insatisfechos con los resultados arrojados por el buscador.

Existen diferentes problemas a los que se enfrentan los usuarios debido al crecimiento exponencial de recursos y contenidos de información en la Web. Entre los más importantes se encuentra: la baja precisión en las búsquedas y la escasa cobertura. La baja precisión en los resultados de las búsquedas se refiere a que la información encontrada es irrelevante con respecto a las necesidades del usuario. La escasa cobertura se debe a que no todos los buscadores tienen la suficiente capacidad de indexar la Web, debido a varios factores; el ancho de banda, el espacio de disco duro, el costo económico, etc. La minería Web resuelve este tipo de problemas al descubrir patrones interesantes.

"La Minería de Web es el uso de las técnicas de minería de datos para el descubrimiento y extracción automática de información de documentos y servicios de la Word Wide Web."[1]

Existen 3 tipos de Minería:

a) *Minería de contenido de la Web*. Consiste en la extracción de información útil del contenido de los documentos Web. Esta información puede ser texto, imágenes, audio, vídeo, o registros estructurados, tales como listas y tablas.

b) *Minería de la estructura de la Web*. Consiste en el análisis, clasificación y categorización de las relaciones entre diferentes páginas Web. Este tipo de minería se puede dividir en dos tipos según el tipo de datos estructurales utilizados.

   - Estructura de hipervínculos: Un hipervínculo es una unidad estructural que conecta una página Web con otra, ya sea dentro de la misma página Web o un sitio Web diferente.
   - Estructura del documento: El contenido de una página Web se organiza en un formato de estructura de árbol al que le llamamos HTML.

c) *Minería del uso de la Web*. Se analizan los datos que interactúan con el usuario tales como logs de acceso a los servidores, Browsers (Buscadores), datos de sesiones, datos de Cookies, datos de registros y transacciones de los usuarios. Analizar los logs de diferentes servidores Web, puede ayudar a entender el comportamiento del usuario.

Este proyecto consiste en realizar Minería de Contenido de la Web, de las tres unidades de la UAM (Azcapotzalco, Iztapalapa y Xochimilco). Se utilizara una técnica de minería Web para obtener patrones relacionados a partir de datos HTML[1]. Se desarrollará un programa (Crawler)[2] que nos ayude a inspeccionar las páginas y recopile información sobre su contenido; es decir visita las páginas de manera recursiva a partir de un conjunto de hipervínculos de páginas iníciales. La información obtenida se almacenará en un repositorio en el cual se podrá consultar datos útiles para el usuario.

1 HTML.-Es el lenguaje de marcado predominante para la elaboración de páginas Web. Es usado para describir la estructura y el contenido en forma de texto, así como para complementar el texto con objetos tales como imágenes.

2Crawler.- Es un programa que inspecciona las páginas del World Wide Web de forma metódica y automatizada.

Antecedentes

Referencias Internas

Se buscó en la biblioteca de UAM Azcapotzalco en el área de proyectos terminales en la carpeta de Licenciatura en Ingeniería en Computación, Ingeniería en Electrónica y Maestría En Computación pero no se encontraron proyectos directamente relacionados con esta propuesta, sin embargo los siguientes trabajos tienen ciertas similitudes ya que utilizan Minería de datos. La minería Web es una metodología de recuperación de la información que usa herramientas de la minería de datos para extraer información.

1. Díaz Jiménez Cristina Alicia, "Lenguaje de manipulación y minería de datos".[2]
2. Nancy Guzmán González, "Aplicación de Distintas Técnicas de Minería de Datos para el Tratamiento de Información".[3]
3. Roberto Emmanuel Ramírez Islas," Programa Asistente para la carga de datos del libro diario de empresas en el sistema minero de datos WEKA".[4]

4. Ing. José Guadalupe Mejía Vega, "Aplicaciones de reglas de asociación para Web Mining", tesis para obtener el grado de maestro en ciencias de la computación.

Este Trabajo desarrolla una aproximación de lo que son las reglas de asociación, las cuales tienen su origen en la tecnología OLAP (on line analytical process), para ello se emplea una metodología que propone la construcción del repositorio de datos (extracción, transformación y transportación de los datos) para posteriormente implementar un algoritmo asociativo que permita minar las páginas Web explícitas e implícitas, que los usuarios han utilizado directa o indirectamente, con la finalidad de descubrir patrones típicos de uso de la red Mundial.

El proyecto que realizaré también hace uso de una técnica de minería Web que permita minar las páginas Web. La diferencia entre los dos proyectos en que el mío trabaja con Minería de contenido de la Web y el otro con Minería del uso de la Web. [5]

Referencias externas

1.- Daedalus

Es una empresa ubicada en Madrid, España. Daedalus se ha ido consolidando como un referente tecnológico en diversas áreas: las tecnologías de la lengua, la minería de datos, la tecnología Web y la inteligencia de negocio. Daedalus es una empresa fuertemente comprometida con la investigación, el desarrollo y la innovación. Participa en diversos proyectos de minería Web a nivel tanto internacional como nacional algunos de sus proyectos son:

- WMA: Web Mining Analytics es un proyecto dedicado al desarrollo de herramientas que faciliten la extracción y el análisis de información estratégica disponible en Internet. WMA será capaz de extraer automáticamente datos específicos de distintas fuentes (Internet, bases de datos, etc.) y de proporcionar información muy relevante sobre los mismos. Esta solución integra los tipos de minería:

Minería de contenido de la Web: Desarrollar un componente que permita identificar datos estructurados (precios, tarifas, etc.) en textos o en documentos no estructurados

Minería del uso de la Web: Se realizarán métodos de búsqueda en Internet de cara a obtener información lo más relevante posible según lo solicitado por un usuario.

- MOWGLI: es un proyecto dedicado a la generación de perfiles en comercio electrónico En este proyecto se utiliza la minería Web para generar perfiles de usuario. Para llevar esto a cabo utilizan Minería del uso de la Web.

El Proyecto a similitud de los proyectos de Daedalus es que ellos trabajan con Minería de contenido de la Web y Minería del uso de la Web y con datos estructurados y no estructurados. Emplearé Minería de contenido de la Web y datos HTML.[6]

2.- Julio García Seminario Y David Casanova, "Implementación De Una Web Mining sobre reconocimiento de patrones de comportamiento de usuarios para la caja municipal de santa (Caja de Ahorro). Universidad San Pedro, Escuela de Informática y Sistemas del Área Facultad de Ingeniería ubicada en Perú.

El objetivo de este proyecto fue determinar el patrón de comportamiento de los usuarios de páginas Web y desarrollaron un modelo solución en la cual se describan los procedimientos para determinar un patrón de comportamiento. Desarrollar un prototipo en base al modelo solución y que tenga como finalidad determinar un parámetro que identifique un patrón de comportamiento.

La diferencia es que ellos se basaron en la minería del uso de la Web y en mi caso se empleará minería de contenido de la Web.[7]

3.- Francisco Manuel Rangel Pardo, "Clasificación de Páginas Web en Dominios Específicos". Para obtener el grado de Maestro en Lenguajes y Sistemas Informáticos: Tecnologías de la Lengua en la Web .Universidad Nacional de Educación a Distancia.

Este trabajo consiste clasificar las páginas Web en dominios determinados para ello se centra en obtener una representación formal de la intención del autor para transmitir información acerca de la pagina que se crea y que se plasma de la meta-información de la misma en la estructura de los enlaces y en la URL a partir del dominio de teatro.

Este proyecto se basa en minería de la estructura de la Web, emplearé minería de contenido de la Web. [8]

4.- Esp. Ing. Hernán Merlino, "Ambiente de integración de herramientas para exploración de datos centrados en la Web" Tesis para obtener el grado de en Ingeniería del Software. Instituto Tecnológico de Buenos Aires**.**

En este trabajo se propone una herramienta para exploración de datos Web que permite estructurar todo el proceso de exploración. La mayor ventaja de esta herramienta es poder utilizar diversas técnicas de exploración, además de permitir la reutilización de procesos ya ejecutados con anterioridad y la combinación de los mismos para su posterior comparación; todo esto llevado a cabo sin un alto grado de complejidad. Este proyecto se basa en los tres tipos de Minería Web. [9]

Justificación

La Minería Web se encarga básicamente de excavar la Web con el objetivo de encontrar información que es interesante y que a simple vista no es evidente ni fácil de entender. Me interesó desarrollar mi proyecto en la línea de la minería Web porque la Web se ha posicionado en uno de los medios más importantes para adquirir información, pero debido a su crecimiento exponencial mucha de la información que se encuentra no es la que en verdad se requiere, la minería Web resuelve este conflicto, encuentra información relevante. Al Minar las páginas de la Universidad Autónoma Metropolitana se va a extraer información útil del contenido HTML de los documentos de la misma.

La información obtenida se almacenará en un repositorio que servirá para que el usuario no tenga que hacer toda la búsqueda sin saber por dónde empezar o que tenga que explorar todas las páginas de la UAM. Podrá consultar una especie de diccionario que le dirá en que pagina están los datos útiles que busca sobre la UAM.

Descripción Técnica.

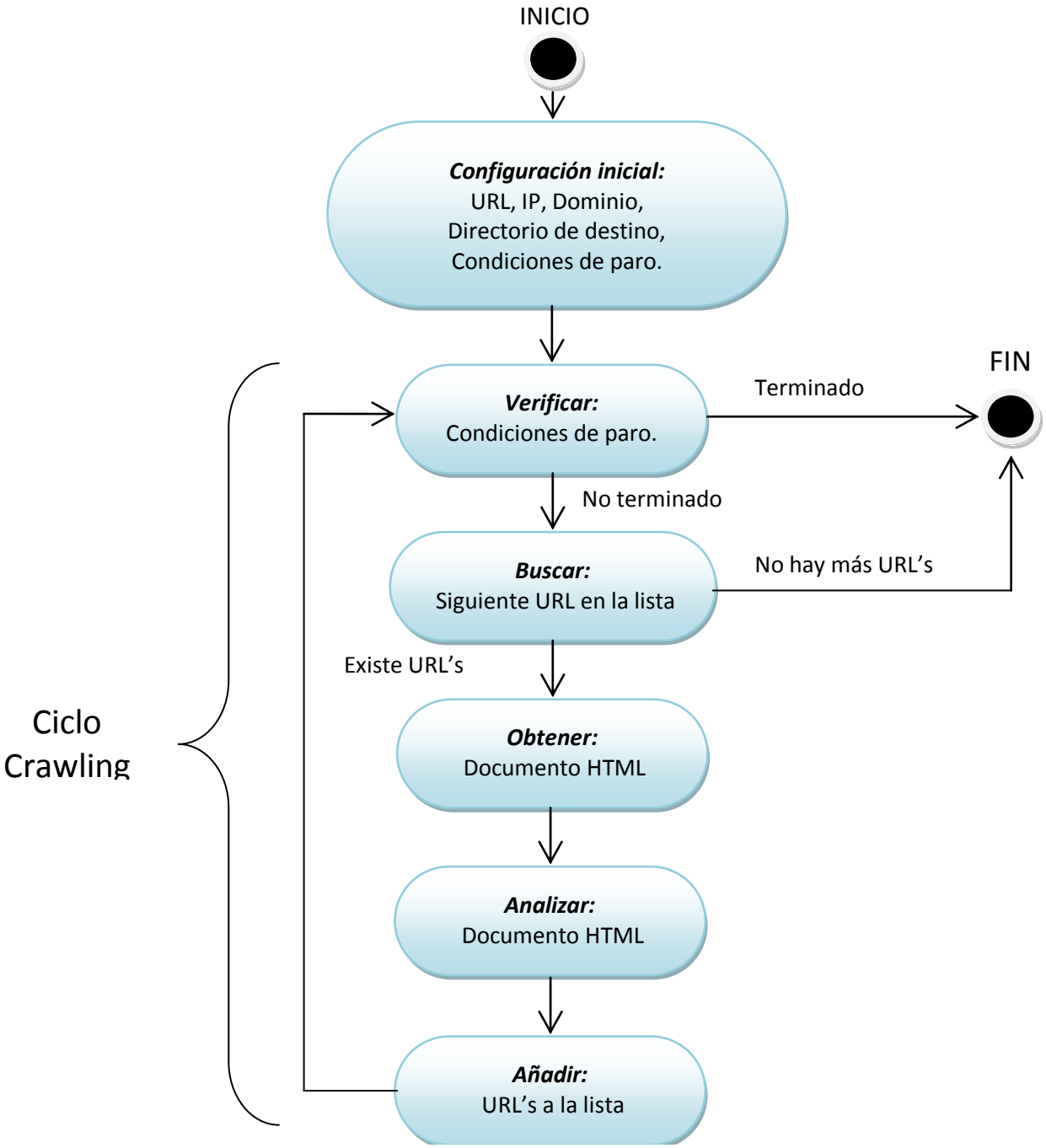
El Proyecto consiste en elaborar un sistema configurable que realice Minería de Contenido Web que permita copiar documentos HTML de la Web a directorios locales con el propósito de descubrir patrones de información ocultos entre los documentos.

El Sistema se divide en 4 etapas:

- Selección y Recopilación de datos: Lo primero es determinar qué es lo que se quiere obtener (documentos HTML), para esto se configurará un programa llamado *Crawler* que recibe los siguientes parámetros:

- URL ó IP, definir un dominio
- Directorio donde se guardaran los documentos.
- Condiciones de paros

Posteriormente el *Crawler* localiza los documentos a adquirir, los captura y almacena los datos pertinentes. El *Crawler* repite este proceso de manera finita mientras las nuevas páginas descargadas ofrezcan más enlaces que seguir ó hasta que se cumplan las condiciones de paro .El objetivo de esta etapa es recuperar automáticamente los documentos mías importantes, indexándolos para optimizar la búsqueda.



Secuencia Básica de un Crawler

- Extracción de la información**:** Consiste en filtrar y limpiar los datos recogidos. Una vez extraída una determinada información a partir de un documento HTML se eliminarán los datos erróneos o incompletos, presentando las restantes de manera ordenada y con los mismos criterios formales hasta conseguir una homogeneidad, listos para su transformación por medios automáticos. La finalidad es identificar y etiquetar el contenido esencial del documento para mapear a algún modelo de datos.

- Reconocimiento de Patrones: En esta etapa, se descubren automáticamente patrones ocultos sobre cada uno de los sitios recopilados y almacenados, esto se logra con la ayuda de una técnica de minería Web llamada "*Clustering*". Esta técnica agrupa automáticamente los datos con características similares sin tener una clasificación predefinida.

El *Clustering* asume que los datos pueden dividirse, razonablemente, en grupos que contienen datos similares. Si tal división existe, ésta puede estar oculta y debe ser descubierta. Al terminar esta etapa tenemos como resultado patrones identificados, listos para ser analizados.

- Análisis: Una vez que los patrones han sido identificados, se creará una interfaz para entender, visualizar e interpretar los patrones.

Todo esto se describe de forma más general en la Figura 2.

Figura 2: Etapas de la Minería Web

Especificación Técnica

El proyecto pretende minar únicamente una parte pequeña de la web, en un dominio establecido, en este caso se va a minar la UAM (Azcapotzalco, Iztapalapa y Xochimilco) y solamente datos Semi-estructurados (Documentos HTML). El Sistema analizará la estructura de los documentos HTML para descubrir patrones ocultos.

El sistema se desarrollará en una plataforma abierta, utilizando como IDE[3]: Eclipse[4] sobre sistema operativo Linux. Para garantizar la portabilidad del sistema, se desarrollará sobre plataforma Java, utilizando servicios web.

El sistema integra servidores de aplicaciones abiertos como Tomcat[5], así como bases de datos relacionales abiertas como Postgres[6].

Licencia de Software Libre-Creative Commons.

- Reconocimiento (Attributtion): En cualquier explotación de la obra autorizada por la licencia hará falta reconocer la autoría.

- Compartir Igual (Share alike): La explotación autorizada incluye la creación de obras derivadas siempre que mantengan la misma licencia al ser divulgadas.

- Reconocimiento - Compartir-Igual (by-sa): Se permite la distribución de las cuales se debe hacer una licencia igual a la que regula la obra original.

Entregables

Los entregables para este proyecto son:

- Código fuente y compilado de la aplicación.

- Esquema de la base de datos y diagrama entidad-relación.

- Diagramas UML de casos de uso, clases y navegación.

- Diccionario de datos.

- Manual de usuario.

- Documentación de referencia: manual de instalación y configuración, javadoc.

- Instalación y configuración en el servidor indicado por el asesor.

---

3 IDE.- (Integrated Development Environment - Entorno integrado de desarrollo). Aplicación compuesta por un conjunto de herramientas útiles para un programador.

4 Eclipse.- Es una plataforma de desarrollo integrado que puede ser usada para crear diversas aplicaciones como sitios de internet

5 Tomcat.- Es un servidor web con soporte de servlets y JSPs.

6 Postgres.- Sistema de gestión de base de datos relacional orientada a objetos y libre.

El proyecto se da por concluido cuando se entregue la documentación indicada y se instale en el servidor E-Learning Knowledge[7].

Calendario de Trabajo

A continuación se presenta el calendario de actividades Para el trimestre 12-I, cada semana equivale a 9 horas de trabajo mientras que para el trimestre 12-P, cada semana equivale a 18 horas de trabajo.

TRIMESTRE 12-I	1	2	3	4	5	6	7	8	9	10	11
Crear el sistema configurable											
Recuperación automáticamente los documentos mías importantes, indexándolos para optimizar la búsqueda. (Proceso del Crawler.)											
Extracción de la información											
Aplicación de la técnica de Clustering para agrupación los datos con características similares											
Análisis de datos											
Crear interfaz para visualizar e interpretar los patrones.											

7 E-Learning Knowledge.- Proyecto creado con la finalidad de crear entornos colaborativos de aprendizaje

TRIMESTRE 12-P	1	2	3	4	5	6	7	8	9	10	11
Implementar el repositorio											
Realizar pruebas											
Realizar Manual de Usuario											
Realizar manual de instalación y configuración											
Instalación y configuración en el servidor (E-Learning Knowledge)											
Realizar pruebas en servidor											

Recursos

Para el diseño de este proyecto se cuenta con el siguiente software:

- PostgreSQL. Manejador de Base de datos de software libre.
- JDK 6u13. Kit de desarrollo Java.
- Apache-tomcat-6.0.18. Servidor web de licencia libre.
- Eclipse 3.4.2.- Entorno de desarrollo integrado de código abierto multiplataforma.

El software que se utilizará en la elaboración del proyecto es de licencia libre, por lo tanto no se requiere la compra de las licencias de dicho software. Los recursos de hardware que se utilizarán durante la elaboración del proyecto son los siguientes:

- Laptop TOSHIBA SATELITE con las siguientes características:
- Procesador Intel Dual Core 2.0 GHz
- Memoria RAM 3 GB
- 500 GB de HD
- Sistema Operativo Linux Ubuntu 11.
- Se hará uso del equipo del aula Educativa Multimedia.

Bibliografía

[1] M.P. José E., "Estado del Arte del Web" [Online], Centro de Aplicaciones de Tecnología de Avanzada (CENATAV), Habana, Cuba, Rep. RT_001, 2007. Disponible en: http://www.cenatav.co.cu/doc/RTecnicos/RT%20SerieGris_001web.pdf

[2] D.J. Cristina Alicia, "Lenguaje de manipulación y minería de datos" Propuesta de Proyecto Terminal, Ingeniería en Computación, División de CBI, UAM Azcapotzalco, D.F., México, 2010.

[3] G.G. Nancy, "Aplicación de Distintas Técnicas de Minería de Datos para el Tratamiento de Información" Propuesta de Proyecto Terminal, Ingeniería en Computación, División de CBI, UAM Azcapotzalco, D.F., México, 2010.

[4] R.I. Roberto Emmanuel, "Programa Asistente para la carga de datos del libro diario de empresas en el sistema minero de datos WEKA" Propuesta de proyecto terminal, Ingeniería en Computación, División de CBI, UAM Azcapotzalco, D.F., México, 2010.

[5] M.V Ing. José Guadalupe "Aplicaciones de reglas de asociación para Web Mining" [Online], Tesis para obtener el grado de Maestro en Ciencias de la Computación, UAM Azcapotzalco, D.F., México, 2002.Disponible en: http://newton.azc.uam.mx/mcc/01_esp/ 11_tesis/tesis/terminada/021201_mejia_vega_jose.pdf

[6] Data, Decisions and Language, S. A.. "*Daedalus-Data*", [Online]. Disponible en: http://www.daedalus.es/

[7] G. S. Julio, C. G. David,"Implementación de una Web Mining sobre Reconocimiento de Patrones de Comportamiento de Usuarios para la Caja Municipal Del Santa" [Online], Proyecto de Tesis, Facultad de Ingeniería, Escuela de Informática y Sistemas, Univ. San Pedro, Chimbote, Perú, 2010. Disponible en: http://www.scribd.com/doc/54965076/Proyecto-de-Tesis-Web-Mining.

[8] R. P. Francisco Manuel, "Clasificación de Páginas Web en Dominios Específicos" [Online], Memoria de Proyecto, Universidad Nacional de Educación a Distancia, 2007. Disponible en: http://mavir2006.mavir.net/docs/FMRangelClasificacionPaginasWebDominiosEspecificos. pdf.

[9] M. Hernán," Ambiente de Integración de Herramientas para Exploración de Datos centrados en la Web" [Online], Tesis de Magister en Ing. del Software, Inst. Tec. de Buenos Aires, Argentina, Noviembre, 2005. Disponible en: http://www.itba.edu.ar/archivos/secciones/merlino-tesisdemagister.pdf