UNIVERSIDAD POLITÉCNICA DE PUEBLA

PROGRAMA ACADÉMICO DE

INGENIERÍA EN INFORMÁTICA

Sistema de Búsqueda de Noticias Educativas en Periódicos Digitales

Ofelia Quintero Cerón

Reporte Técnico PII-33-12-09

COMITÉ EVALUADOR

cDr. Eduardo López Domínguez (*Asesor*)

MC. Argelia Berenice Urbina Nájera (*Sinodal*)

Dr. Jorge de la Calleja Mora (*Sinodal*)

PROFESOR(A) DE PROYECTO DE INVESTIGACIÓN II

Dra. María Auxilio Medina Nieto

Juan C. Bonilla, Puebla

Diciembre de 2010

ÍNDICE

CAPÍTULO 1 PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN

1.1 Introducción

En un mundo globalizado que cambia rápidamente como es el de la actual sociedad de la información y del conocimiento, estar permanentemente informado se ha convertido en una necesidad apremiante, en fuente de conocimiento y también de dinero.

La propagación de fuentes de información, tanto en el ámbito científico, profesional e incluso doméstico, ha creado una oleada reciente de subscripciones a servicios en línea de noticias en periódicos digitales, lo cual pone de manifiesto la importancia que la sociedad tiene de estar permanentemente informada sobre temas que son de su interés [Pinto 2004][UTM 2002].

Internet, es la fuente de información más grande jamás conocida, es una de las principales fuentes de generación y transmisión de información. Uno de los problemas principales de Internet es el crecimiento constante y descontrolado de la información a la que los usuarios pueden acceder. Este crecimiento desmesurado está contribuyendo a que los usuarios tengan dificultades para encontrar la información que precisan de manera simple y eficiente. Esta situación se agrava cada día porque mejoran las características tecnológicas y la cantidad de información que se maneja se vuelve más grande [UTM 2002].

Por tal motivo, la necesidad de almacenamiento y recuperación de información se ha vuelto vital e indispensable para el Departamento Editorial de la Universidad Politécnica de Puebla (UPP), el cual está encargado de dar a conocer noticias educativas publicadas en periódicos o revistas. La recuperación de información es una tarea que requiere tiempo, dinero y esfuerzo por tal motivo, este proyecto propone realizar un sistema de recuperación de noticias que facilite las actividades de búsqueda, recuperación y clasificación de noticias relevantes de periódicos digitales de libre acceso.

1.2 Objetivo general

Implementar un prototipo de recuperación y clasificación de noticias de periódicos digitales nacionales.

1.3 Objetivos específicos

- Implementar un método para la recuperación de noticias educativas en periódicos digitales.
- Implementar un método para la clasificación de noticias.
- Crear una interface que integre ambos métodos.

1.4 Justificación

En la Universidad Politécnica de Puebla (UPP) existe la necesidad de informar a los alumnos y académicos sobre noticias educativas que ocurren dentro de la ciudad. Actualmente en la UPP, la búsqueda, recuperación y clasificación de noticias relevantes se hace de manera manual, lo que lleva a mayores demoras y dificultades para la persona encargada del departamento Editorial, generando pérdida de dinero y tiempo para la universidad. Por tal motivo, se propone hacer un sistema de búsqueda de noticias educativas en periódicos digitales de libre acceso. Este sistema automatizará la búsqueda de las noticias, obtendrá la liga de donde fue recuperada la noticia sin la necesidad de comprar periódicos. Además, el sistema clasificará las noticias en educativas y no educativas.

1.5 Metodología

El desarrollo del proyecto de investigación se realizará de acuerdo a las actividades que se muestran en el siguiente cronograma.

En los meses de septiembre a diciembre de 2008 se realizaron las siguientes actividades:

Cronograma de actividades

2008																
Meses	Septiembre				Octubre				Noviembre				Diciembre			
Semanas	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Actividades																
Investigación de recuperación de información																
Consultar bibliografía relacionada																
Identificar periódicos digitales de libre acceso																
Elaboración de la propuesta																
Revisión de literatura																
Diseño del sistema																

En la segunda parte del proyecto que se conforma de los meses de Septiembre a diciembre de 2009 se realizaron las siguientes actividades. En esta parte del proyecto estoy realizando la implementación de la matriz para la clasificación de las noticias

2009																
Meses	Septiembre				Octubre				Noviembre				Diciembre			
Semanas	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Actividades																
Diseñar interfaz del sistema de recuperación de información																

Implementar el módulo para la recuperación de información																
Implementar diccionario con las noticias recuperadas																
Definir prototipo para clasificar las noticias																
Implementar matriz para la clasificación de las noticias																
Investigar y realizar pruebas del software WEKA																
Implementar archivo para la clasificación de las noticias con WEKA																
Revisión de la literatura																
Elaboración del reporte de investigación																

Espacios rellenos de color azul ya están terminados

Espacios rellenos de color verde falta por implementarlos

1.6 Recursos de hardware y software

Recursos de hardware

- Procesador Intel Pentium 4 a 1.8.Ghz.
- Disco duro de 80 GB
- Memoria RAM de 1 GB

Recursos de software

- Sistema operativo Microsoft Windows XP
- Paquetería de Microsoft Office 2007
- JCreator Pro 4.5
- JDK 1.6
- Programa Weka Algoritmo de Vecinos más Cercanos

1.7 Alcances y limitaciones

Alcances

- Se utilizará el programa WEKA para clasificar las noticas
- Se contará con una liga de donde fue recuperada la noticia
- El sistema se presentará como un prototipo para resolver el problema de recuperación de noticias educativas.

Limitaciones

- Debido a que existen demasiados periódicos digitales, se restringe la exploración de todos los periódicos internacionales.
- El tipo de contenido buscado en los periódicos digitales para la recolección de información será únicamente en idioma Español.
- Serán sólo periódicos de libre acceso.

CAPÍTULO 2 MARCO TEÓRICO

2.1 Sistema de búsqueda de noticias educativas en periódicos digitales

Un sistema de búsqueda de noticias en un software que va a recuperar el código fuente de los periódicos disponibles y de libre acceso, la forma en que realiza la búsqueda de las noticias es por medio de etiquetas que contiene el código fuente del periódico, esto permite que el sistema realice una búsqueda sencilla permitiendo al usuario visualizar los encabezados de las noticias obtenidas.

2.2 Definición de periódico digital

El periódico digital, se define como producto interactivo y multimedia, integra diferentes recursos como el texto, la imagen, el vídeo y el sonido; y está revolucionando los conceptos básicos del periodismo impreso. El periodismo en Internet no sólo se encuentra en las páginas de periódicos, televisión o radio en línea, también está presente en otros sitios como puede ser la recepción de información de los teléfonos celulares [Hipertex 2008].

Principales características del periódico digital:

1.- PRODUCTO DIGITAL: El producto llega a la pantalla por medio de bytes, se transmite por redes telemáticas, además los costos se reducen considerablemente respecto a las ediciones impresas. Esto suele resultar una ventaja considerable y es accesible en cualquier circunstancia siempre y cuando se tenga una computadora y una conexión a internet.

2.- HIPERTEXTO: El hipertexto permite pasar de pagina a pagina, acceder a los textos, imágenes fijas o en movimiento, el hipertexto se apoya en la capacidad de la mente humana para relacionar ideas, hechos y datos diferentes, así a través de links o enlaces incluidos en el texto principal, se facilita el acceso a archivos conectados entre si.

3.- INSTANTÁNEO: El acceso es instantáneo. Se obtiene la información de manera inmediata y esta puede ser consultada casi en tiempo real comparando con otros medios como son la radio y la televisión. A diferencia del periódico de papel, el electrónico no se ve obligado a esperar la siguiente edición para poner a disposición las últimas noticias

4.- ACTUALIZABLE: El periódico digital renueva la información conforme se vaya generando la noticia, ya que un medio que no actualiza al instante los sucesos más importantes de nada sirve.

5.- GRATUITO: La mayoría de los servicios que proporcionan los periódicos en línea son gratuitos. Internet es información a un costo muy bajo e incluso gratuito. La facilidad de los periódicos en línea favorece la consulta del usuario.

2.3 Definición de noticia

Una noticia es el relato o redacción que refiere a un hecho novedoso ocurrido dentro de una comunidad o determinado ámbito específico lo cual hace que merezca su divulgación. La noticia satisface la curiosidad del lector al responder a las siguientes preguntas: ¿Qué? ¿Quién? ¿Cuándo? ¿Dónde? ¿Por qué? y ¿Cómo? [América 2008]

La noticia despierta el interés humano en cuanto a**:**

Rarezas.- Lo insólito o fuera de lo común fascina a los lectores.

Conflicto.- Los lectores quieren saber quién ganará las elecciones, guerras, competencias deportivas, etc.

Emoción.- Hay noticias que involucran emocionalmente a los lectores. Ejemplos: las relacionadas con niños, educación, animales, personas desvalidas, maltratadas o discriminadas y víctimas de desastres. Hay noticias que suscitan en el lector risa, cólera, solidaridad, entre otros [Hipertext 2008].

2.4 Definición de noticia educativa

Una noticia educativa es el relato que refiere a una institución educativa, a una redacción de un artículo educativo digital como los que se encuentran en los periódicos en línea, bibliotecas virtuales, entre otros. La noticia educativa puede contener información relevante sobre escuelas, alumnos, aprovechamiento académico, nuevas aportaciones por parte del gobierno a alguna institución, sucesos históricos, entro otros [América 2008].

Algunas de las principales características del género noticia educativa son las siguientes:

Veracidad: Los hechos o sucesos deben ser verdaderos y, por lo tanto, verificables.

Objetividad: El periodista no debe verse reflejado en ella mediante la introducción de ninguna opinión o juicio de valor. En la noticia no ha de aparecer quien la ha redactado, sólo se adivinará que tiene un autor porque en ella se da una selección de la realidad, de manera que el periodista escoge los elementos que le parecen interesantes y relevantes. Pero en ningún caso se mostrará su opinión.

Claridad: Los hechos deben ser expuestos de forma ordenada y lógica.

Brevedad: Los hechos deben ser presentados brevemente, sin reiteraciones o datos irrelevantes.

Generalidad: La noticia debe ser de interés social y no particular.

Actualidad: Los hechos deben ser actuales o recientes.

Consecuencia: Tiene interés noticioso todo lo que afecte a la vida de las estudiantes.

Oportunidad: Mientras más rápido se dé a conocer noticia de carácter educativo mayor valor posee. [América 2008].

2.5 Lenguaje natural

El Lenguaje Natural, es entendido como la herramienta que utilizan las personas para expresarse, posee propiedades que disminuyen la efectividad de los sistemas de recuperación de información textual. Estas propiedades son la variación y la ambigüedad lingüística. La variación lingüística se refiere a la posibilidad de utilizar diferentes palabras o expresiones para comunicar una misma idea. En cambio, la ambigüedad lingüística se produce cuando una palabra o frase permite más de una interpretación. [PLN 2005].

PROCESAMIENTO DE LENGUAJE NATURAL

Es el conjunto de instrucciones que una computadora recibe en un lenguaje de programación dado. El procesamiento del lenguaje natural presenta múltiples aplicaciones:

•Corrección de textos

•Recuperación de la información

•Extracción de Información y Resúmenes

•Búsqueda de documentos

En el campo de la recuperación de la información han desarrollado sistemas que permiten obtener información sobre estadísticas deportivas, información turística, geografía, entre otros.

2.5.1 El procesamiento del lenguaje natural en la recuperación de información textual

La complejidad asociada al lenguaje natural cobra especial relevancia cuando necesitamos recuperar información textual que satisfaga la necesidad de información de un usuario. Es por ello, que en el área de Recuperación de Información Textual las técnicas de Procesamiento de lenguaje natural (PLN) son muy utilizadas tanto para facilitar la descripción del contenido de los documentos, como para representar la consulta formulada

por el usuario, todo esto, con el objetivo de comparar ambas descripciones y presentar al usuario aquellos documentos que satisfagan en mayor grado su necesidad de información [UTM 2002].

Dicho de otro modo, un sistema de recuperación de información textual lleva a cabo las siguientes tareas para responder a las consultas de un usuario, ver Figura 1:

1. Indexación de la colección de documentos: en esta fase, mediante la aplicación de técnicas de PLN, se genera un índice que contiene las descripciones de los documentos. Normalmente, cada documento está descrito mediante el conjunto de términos que, hipotéticamente, mejor representa su contenido.
2. Cuando un usuario formula una consulta, el sistema la analiza y si es necesario la transforma con el fin de representar la necesidad del mismo modo que el contenido de los documentos.
3. El sistema compara la descripción de cada documento con la descripción de la consulta y presenta al usuario aquellos documentos cuyas descripciones más se asemejan a la descripción de su consulta.
4. Los resultados suelen ser mostrados en función de su relevancia, es decir, ordenados en función del grado de similitud entre las descripciones de los documentos y de la consulta [PLN 2005].

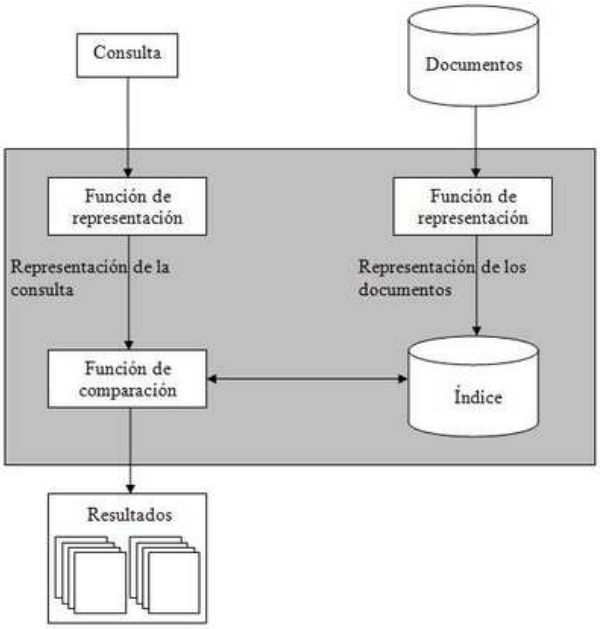


Figura 1: Arquitectura de un sistema de recuperación de información.

2.5.2 ¿Qué es la recuperación de información?

La recuperación de información es un proceso donde se accede a una información previamente almacenada, mediante herramientas informáticas que permiten establecer ecuaciones de búsqueda específicas. Dicha información ha debido de ser estructura previamente a su almacenamiento.

Gran parte de las necesidades de recuperación de información para los usuarios está resuelta con la implementación de algunos buscadores. Los cuales plantean nuevas soluciones a la organización de información en cantidades imposibles de manejar para los usuarios porque cuando un usuario se plantea la necesidad de obtener nueva información sobre un asunto o materia de su interés, está manifestando una carencia, una situación irregular de sus estructuras mentales y cognitivas. Un estado mental de incertidumbre que mueve al individuo a desarrollar una serie de acciones para salir de ese estado [Pinto 2004]. La recuperación de información engloba las acciones encaminadas a identificar, seleccionar y acceder a los recursos de información útiles al usuario. Para la recuperación de información en documentos de la Web se han desarrollado diferentes estrategias como el

uso de metadatos (los metadatos proveen la información necesaria para que los datos puedan ser empleados ágilmente en diferentes aplicaciones) o la utilización de lenguajes semánticos basados en XML para indizar documentos Web y representar el conocimiento incluido en ellos

2.5.3 Aprendizaje automático

El Aprendizaje Automático es una rama de la Inteligencia Artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. Se trata de crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento. En muchas ocasiones el campo de actuación del Aprendizaje Automático se traslapa con el de la estadística, ya que las dos disciplinas se basan en el análisis de datos. Sin embargo, el Aprendizaje Automático se centra más en el estudio de la Complejidad Computacional de los problemas.

El Aprendizaje Automático tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, reconocimiento del habla y del lenguaje escrito, juegos y robótica y para la búsqueda de noticias en periódicos digitales

2.5.4 Tipos de aprendizaje

- Aprendizaje supervisado
- Aprendizaje no supervisado

2.5.5 Aprendizaje supervisado

El aprendizaje supervisado no sólo requiere la elaboración manual del esquema de categorías, sino también, requiere un proceso de aprendizaje o entrenamiento por parte del clasificador, que debe ser supervisado manualmente en mayor o menor medida.

Para la construcción de los patrones de las categorías se utilizan documentos clasificados manualmente de antemano, que sirven como ejemplo. El proceso de formar esos patrones de cada clase a partir de esos documentos preclasificados se conoce como entrenamiento o aprendizaje.

Aunque hay gran cantidad de algoritmos capaces de hacer clasificación supervisada, la idea o enfoque básico es muy parecido, de alguna manera se debe construir un patrón representativo de cada una de las clases o categorías y aplicar alguna función que permita estimar el parecido o similitud entre el documento a clasificar y cada uno de los patrones de las categorías.

2.5.6 Aprendizaje no supervisado

Es el aprendizaje que no necesita de un profesor, supervisor o validador externo para realizar su aprendizaje. Son capaces de modificar sus parámetros internamente, adaptándose únicamente en el conjunto de entrenamiento que requieren.

Su principal fundamento se basa en la redundancia que hay en el lenguaje natural y de esta forma poder sacar relaciones semánticas, distinguir expresiones superfluas, descubrir clases, entre otros.

Por otro lado, su principal ventaja frente al aprendizaje supervisado es que no necesitan un método de clasificación manual que haría del sistema de recuperación o extracción de información demasiado costosa. Sin embargo, se necesitan grandes cantidades de información para poder sacar las relaciones y redundancias de información anteriormente mencionadas pero este hecho no es preocupante debido a la facilidad de recuperar, extraer y/o obtener información en la actualidad.

Principales características

Familiaridad de conceptos y agrupamiento. A partir de un conjunto de entrada se desea conocer si hay un cierto orden o jerarquía en la información recuperada o extraída.

Extracción y relación de características. Realizar un mapa topológico de los datos de entrada, a través del diseño de la red de tal forma que patrones de entrada parecidos, produzcan respuestas similares.

Análisis de las componentes principales. Detectar qué componentes de los datos de entrada tienen más valor para la recuperación.

Prototipo. Obtener prototipos o ejemplares del conjunto de información que se pretende buscar o encontrar.

2.5.7 Aprendizaje por refuerzo

El objetivo del aprendizaje por refuerzo es usar el premio-castigo para aprender una función, la cual permitirá tomar decisiones en el futuro de qué acción tomar a partir de una percepción del entorno. La función del aprendizaje por refuerzo es que utiliza la información contenida en él para realizar la toma de decisiones.

2.6 Algoritmos para clasificación

2.6.1 Algoritmo del vecino más próximo

El algoritmo del vecino más próximo (Nearest Neighbour, NN) es uno de los más sencillos de implementar [Haddad 2008]. La idea básica es como sigue: se calcula la similitud entre el documento a clasificar y cada uno de los documentos de entrenamiento, el más parecido estará indicando a qué clase o categoría se debe asignar.

Vecinos más cercanos (KNN, por sus siglas en ingles) es un método de aprendizaje basados en instancias. Este algoritmo no tiene una fase de entrenamiento, por lo que la clasificación de nuevas instancias se realiza en tiempo de ejecución, comparando la nueva instancia con todas las instancias del conjunto de entrenamiento. En este algoritmo se almacena todo el conjunto de entrenamiento, de tal modo que para clasificar una nueva instancia *i*, se busca en los ejemplos almacenados casos similares.

Existen varias formas de determinar a los casos más similares entre la instancia *i* y las instancias del conjunto de entrenamiento, siendo la *distancia euclidiana* la más utilizada. Una vez identificados los *k* casos más similares, le es asignada una clase a la instancia *i*, la cual es elegida de acuerdo a las clases de sus *k* ejemplos similares. La forma más común de asignar la clase de la instancia *i*, es elegir la clase más frecuente en sus vecinos [Ponce et a1].

2.6.2 Redes neuronales

Las redes neuronales en general han sido propuestas en numerosas ocasiones como instrumentos útiles para la Recuperación de Información y también para la clasificación automática. Una de las principales aplicaciones de las redes neuronales es el reconocimiento de patrones. Básicamente, una red neuronal consta de varias capas de unidades de procesamiento o neuronas interconectadas para la capa de entrada recibe términos, mientras que las unidades o neuronas de la capa de salida mapea clases o categorías [Figuerola et a1 2004].

Es posible entrenar una red para que, dada una entrada determinada, produzca la salida deseada.

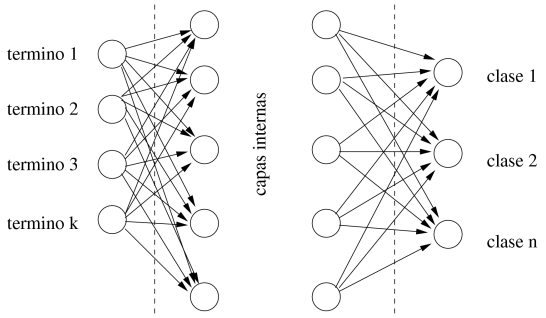


Figura 2. Red neuronal para clasificación automática

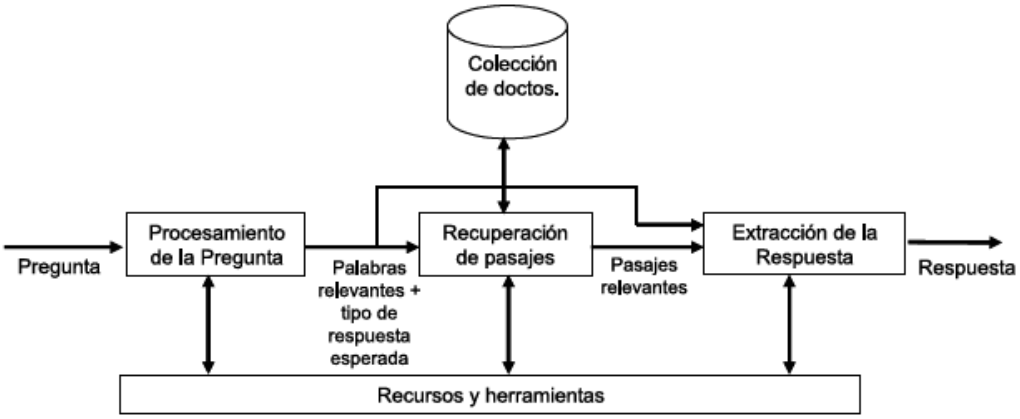
2.7 Trabajo relacionado

Arquitectura General de un Sistema de Búsqueda de Respuestas

Un sistema de BR engloba métodos de distintas disciplinas que tratan el lenguaje escrito, lo cual hace interesante, pero a la vez más complicado, el desarrollo del mismo.

Este tipo de sistemas se encuentran en la intersección de diferentes tareas de investigación, principalmente Recuperación de Información (RI) para la formulación de peticiones de información, análisis de unidades de información (documentos, párrafos, etc.), así como para el análisis y retroalimentación de relevancia de las unidades de información recuperadas; y el Procesamiento del Lenguaje Natural (PLN) para la extracción de información relevante de la pregunta, y la extracción de atributos que caractericen a las respuestas candidatas y que permitan discriminar las respuestas correctas de las incorrectas [Juárez 2007].

Un sistema de BR consta usualmente de tres módulos: el modulo de Procesamiento de la Pregunta, el modulo de Recuperación de Pasajes y el modulo de Extracción de la Respuesta [Téllez et a1]. La figura 3 muestra de manera gráfica la arquitectura general de un sistema de BR.



Procesamiento de la Pregunta

Un sistema de BR recibe como entrada una pregunta formulada en lenguaje cotidiano.
Esto permite una interacción entre el usuario y el sistema muy sencillo y natural. Sin embargo, una pregunta en lenguaje cotidiano, realizada de manera directa, contiene pocas palabras y elementos que deben explotarse al máximo para obtener un buen resultado, el cual se traduce en una respuesta correcta a la pregunta realizada.

El primer módulo de un sistema de BR, el Procesamiento de la Pregunta, es el encargado de extraer toda la información posible de la pregunta formulada. En este módulo se extraen las siguientes características: Palabras relevantes, entidades nombradas y tipo de la respuesta esperada.

Recuperación de pasajes

Este es el segundo modulo de un sistema de BR. En este modulo se utiliza la información del primer modulo, específicamente las palabras relevantes de la pregunta, para realizar la extracción de texto relevante a la pregunta. Lo anterior se logra aplicando técnicas de RI a la colección de documentos en la cual se encuentra contenida la información que puede dar respuesta a la pregunta planteada.

Cada documento o pasaje recuperado es acompañado, entre otras cosas, de un peso numérico, el cual indica la relevancia del documento o pasaje respecto a las palabras de la pregunta utilizadas para hacer la consulta.

Extracción de la Respuesta

Del módulo de Procesamiento de la Pregunta se tienen las palabras relevantes de la pregunta y el tipo de respuesta esperada; del segundo módulo se tienen varios textos donde presumiblemente se encuentra la respuesta

Por otro lado, un sistema de Búsqueda de Respuestas es dar una respuesta concreta y precisa, evitando al usuario la tediosa tarea de revisar un documento completo para encontrar información específica [Juárez 2007].

CAPÍTULO 3 DISEÑO DE LA INVESTIGACION

3.1 Introducción

En este capítulo se describe el diseño del proyecto "Sistema de búsqueda de noticias educativas en periódicos digitales". La importancia de este trabajo radica en que partir de los resultados del mismo se podrán elaborar un sistema el cual pueda recuperar las noticias educativas de forma automática.

3.2 Documento de requerimientos

Se requiere diseñar e implementar un sistema de búsqueda de noticias educativas en periódicos digitales de libre acceso para la Universidad Politécnica de Puebla, en donde permita ingresar a los periódicos y recuperar los encabezados de las noticias, para después poder clasificar cuales son educativas y desechar las que no lo son. De esta manera, se obtendrán los resultados de noticias educativas que sean de interés a los alumnos y profesores de dicha institución.

3.3 Casos de uso

Los casos de uso son una técnica para especificar el comportamiento de un sistema:
Un caso de uso es una secuencia de interacciones entre un sistema y un usuario que usa alguno de sus servicios.

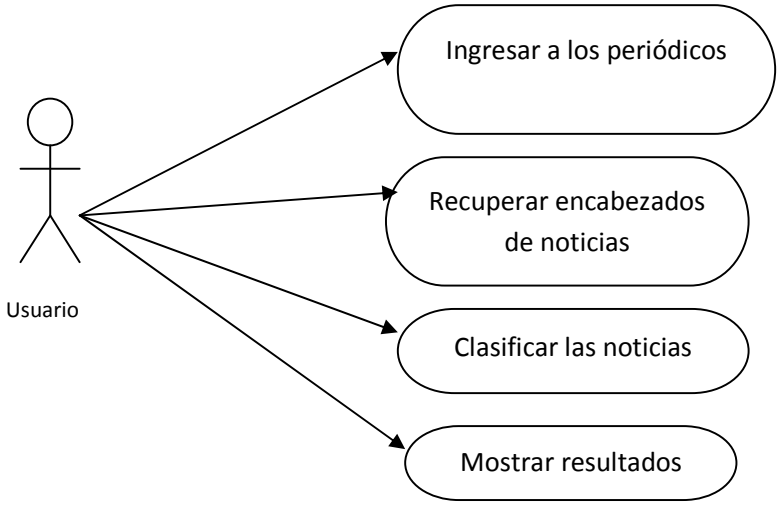


Figura 4: Caso de uso general

Caso de uso ingresar a los periódicos digitales

E n este caso de uso se ingresara a los periódicos digitales de libre acceso que estén contemplados dentro de este proyecto.

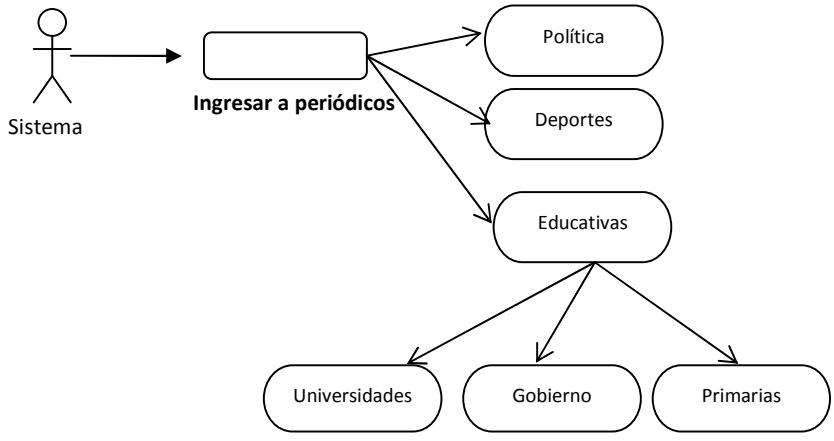


Figura 5: Caso de uso ingresar a los periódicos de libre acceso

Caso de uso recuperar encabezados de noticias

En este caso de uso se podrá obtener todos los encabezados de las noticias digitales de todos los periódicos disponibles del sistema.

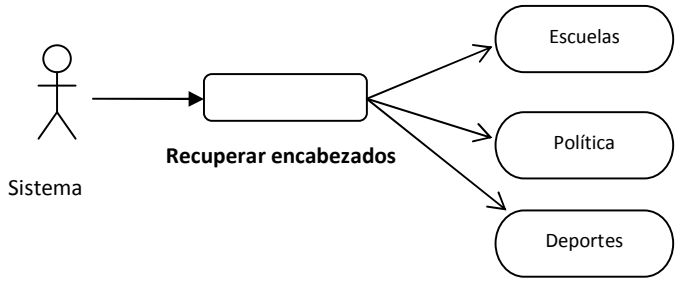


Figura 6: Caso de uso recuperar encabezados de noticias

3.4 Manejo de datos

Los datos que se utilizarán en el proyecto serán de libre acceso para los alumnos y administrativos de la Universidad Politécnica de Puebla. Las operaciones que se van a realizar serán de tipo lectura, ya que se tendrá acceso a la página de donde fue recuperada la noticia.

3.5 Requerimientos funcionales y no funcionales

Los requerimientos funcionales y no funcionales se conocen como el conjunto de características de calidad, que es necesaria tener en cuenta para el diseño e implementación del software

3.5.1 Requerimientos funcionales

El sistema permitirá la recuperación de las noticias de periódicos digitales de libre acceso, lo cual permitirá ahorro de tiempo y dinero para la UPP, también mostrara la url de la noticia y podrán ser guardadas dentro de un archivo de texto para poder clasificarlas de acuerdo al tipo que pertenecen. Para ello se utilizara del programa WEKA y se hará lo siguiente:

- Clasificar las noticias por medio del programa WEKA con el algoritmo de vecinos más cercanos
- Utilización de funciones de visualización de datos en WEKA
- Utilización de diversos métodos de aprendizaje automático desde WEKA

3.5.2 Requerimientos no funcionales

- El sistema debe estar en capacidad de dar respuesta al acceso de todos los usuarios con tiempo de respuesta aceptable y uniforme, en la medida de las posibilidades tecnológicas.
- El sistema debe estar en capacidad de permitir en el futuro el desarrollo de nuevas funcionalidades, clasificar todo tipo de noticias, entre otras funcionalidades después de su construcción.
- El prototipo debe ser fácil de instalar para el departamento de Editorial.
- Costo. Las noticias serán recuperadas de periódicos de libre acceso por lo tanto no tendrá ningún costo para la Universidad Politécnica de Puebla.

CAPÍTULO 4 IMPLEMENTACIÓN

4.1 Fuentes de recolección

Las noticias serán extraídas sólo de periódicos digitales de libre acceso y que cuenten con un formato HTML en su código fuente los periódicos disponibles son: El sol de Puebla, Esto, El financiero, e-Consulta y el Economista. Estos periódicos se tomaron en cuenta porque son de libre acceso y se puede accesar a su código fuente sin ningún problema. Además son periódicos en los cuales el número de lectores es alto.

4.3 HtmlParser

El programa se realizo bajo la plataforma de Java, el cual es un lenguaje de programación con el que podemos realizar cualquier tipo de programa. En la actualidad es un lenguaje muy extendido y cada vez cobra más importancia tanto en el ámbito de Internet como en la informática en general.

Actualmente Java se utiliza en un amplio abanico de posibilidades y muchas veces con grandes ventajas. Con Java se pueden programar sistemas que faciliten la búsqueda de noticias en periódicos digitales, para ello se utilizarán las librerías de HTML Parser.

HTML Parser es una biblioteca Java, la cual se utiliza para analizar código HTML, ya sea en una forma lineal o anidada. Principalmente utilizada para la recuperación o transformación por medio de filtros y etiquetas fáciles de usar. Para utilizar la biblioteca, tendrá que añadir el htmllexer.jar o htmlparser.jar a sus clases al compilar y ejecutar.

HTML Parser comprende la extracción de programas de recuperación de información que no están destinados a preservar la fuente de la página. Esto abarca usos como:

- La extracción de enlace para rastrear a través de páginas web o direcciones de correo electrónico de la consulta.
- Un navegador la etapa preliminar de la visualización de la página

- La comprobación de vínculos para garantizar que los enlaces son válidos

Para la recuperación de noticias el Parser proporciona acceso a los contenidos de la página, a través de un NodeIterator o un NodeList.

Clase NodeIterator se utiliza para iterar sobre los nodos de almacenamiento en el DOM. Esta aplicación bloquea el Modelo de Objetos del Documento (DOM) para leer el archivo y abren el nodo después.

El NodeList proporciona la abstracción de una colección ordenada de nodos, sin definir o limitar cómo se aplica esta colección.

Para obtener la información de la página web se hace en dos pasos los cuales se explican en las siguientes tablas:

Código para obtener el código fuente de la página

```
1. public void obtenInfo(){
2. parser = new Parser (direccionu, Parser.DEVNULL); //el sistema hace un Parser
3. NodeIterator nd=parser.elements(); // los nodos son devueltos
4. while(nd.hasMoreNodes()){  // Comprueba si se dispone de más nodos.
5. NodeIterator cp=nd;
6. NodeList b=nd.nextNode().getChildren(); // Obtener el siguiente nodo
7. SimpleNodeIterator sn=null;
```

Tabla 1. Obtener código fuente del periódico

En la tabla 1 se muestra como se obtiene en código fuente de la página del periódico en la línea 2 se hace un Parser a la página web del periódico deseado, para que con la clase NodeIterator permita que los miembros de una lista de nodos sean devueltos de forma secuencial, en la línea 4 con el método hasMoreNodes comprueba si existen más nodos

disponibles , en la línea 6 se obtiene el siguiente nodo en la secuencia de HTML o null en caso de que ya no existan más nodos como se muestra en la línea 7

Código para devolver la URL de la noticia

```
1. private String devuelveLink(String linea) {
2. private Vector e_consulta(){ //Se dirige al periódico e-Consulta
3. vector.add("<div class=\"col_fron_title\">"); //Etiqueta
4. return vector; //Va a regresar el encabezado de la noticia
5. }
6. private Vector soldepuebla(){ //Ingresa al código fuente de el periódico El Sol de Puebla
7. vector.add("<div class=\"cabezaprincesto"); //Todas las etiquetas con las cuales están definidas los encabezados de las noticias
8. vector.add("<div class=\"resumenesto");
9. vector.add("<div class=\"cabezaprincesto2");
10. vector.add("<div class=\"cabezaterciaria");
11. private String linkConsulta(){ //Después de devolver el encabezado de la noticia, va a devolver la url de la misma
12. return "http://www.e-consulta.com/";
13. }
14. if(ins.contains("</a>"))
15. ins=ins.replace("</a>", "");
16. }
```

Tabla 2. Código para devolver la URL de la noticia

En la línea 2 se crea un vector en el cual ingresara al código fuente del periódico, cada periódico define sus encabezados con una serie de etiquetas como se muestra en la línea 3 para el periódico e-Consulta, en la línea 4 va a regresar todo el texto que se haya encontrado dentro de esas etiquetas. Lo mismo sucede de las líneas 6 a la 12. Para obtener la url del periódico después de que recupero los encabezados continua con la url la cual está definida por medio de las etiquetas <a> y como s muestra en las líneas 14 y 15.

4.2 WEKA

WEKA es una herramienta de aprendizaje automático y data mining, escrita en lenguaje Java, gratuita y desarrollada en la Universidad de Waikato (WEKA = Waikato Environment for Knowledge Analysis).

El paquete Weka contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades. La versión original de Weka fue un front-end en TCL/TK para modelar algoritmos implementados en otros lenguajes de programación, más unas utilidades para pre procesamiento de datos desarrolladas en C para hacer experimentos de aprendizaje automático. Esta versión original se diseñó inicialmente como herramienta para analizar datos procedentes del dominio de la agricultura pero la versión más reciente basada en Java (WEKA 3), que empezó a desarrollarse en 1997, se utiliza en muchas y muy diferentes áreas, en particular con finalidades docentes y de investigación [Melville 2002].

4.2.1 Características de WEKA

WEKA es una herramienta de aprendizaje automático y data mining, escrita en lenguaje Java algunas de sus características son:

- Es fácil de utilizar por un principiante gracias a su interfaz gráfica de usuario.
- Es portable porque está completamente implementado en Java y puede correr en cualquier plataforma

- Contiene una extensa colección de técnicas para pre procesamiento de datos y modelado.
- Está disponible libremente bajo la licencia pública general de GNU.

Weka soporta varias tareas estándar de minería de datos, especialmente, pre procesamiento de datos, almacenamiento, clasificación, regresión, visualización, y selección. Todas las técnicas de Weka se fundamentan en que los datos están disponibles en un fichero plano o una relación, en la que cada registro de datos está descrito por un número fijo de atributos. Weka también proporciona acceso a bases de datos vía SQL gracias a la conexión JDBC (Java Database Connectivity) y puede procesar el resultado devuelto por una consulta hecha a la base de datos. No puede realizar minería de datos multi-relacional, pero existen aplicaciones que pueden convertir una colección de tablas relacionadas de una base de datos en una única tabla que ya puede ser procesada con Weka [Melville 2002].

Weka utiliza diferentes algoritmos para la clasificación de información, en el caso de este proyecto se utilizara el algoritmo de vecinos más cercanos.

CAPÍTULO 5 RESULTADOS

5.1 Pantalla Principal

En esta pantalla muestra el inicio del programa en el cual solo muestra la bienvenida al sistema.

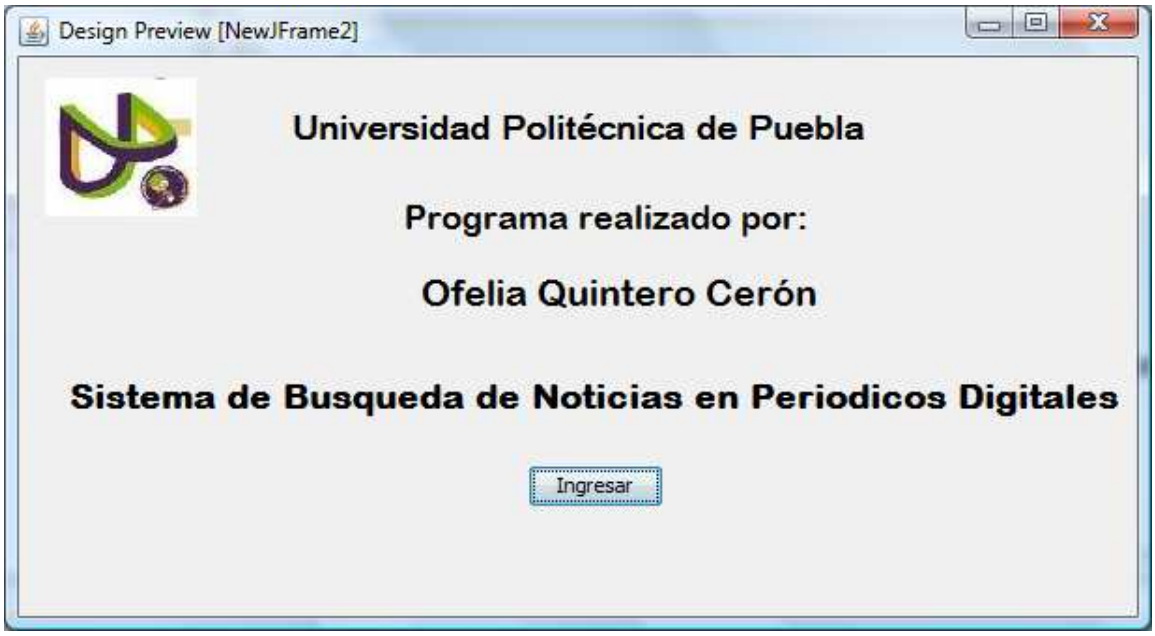


Figura 7. Pantalla principal

5.2 Pantalla de menús

En esta sección se describe el funcionamiento de la interfaz gráfica donde se manipula la información de los diferentes periódicos digitales disponibles, los cuales permiten actividades como: elegir la url del periódico, recuperar la información de dicho periódico, mostrar las noticias que fueron recuperadas y guardar dichas noticias en un archivo de texto (txt).

La Figura 8 muestra la interfaz donde se realiza opción de obtener la pagina. En la parte superior contiene el menú principal que da la opción de obtener la página del periódico deseado dónde se manipulan los datos. En este caso sólo cuenta con una única opción para después mostrar un menú desplegable con los periódicos disponibles lo cual se muestra en la Figura 9.

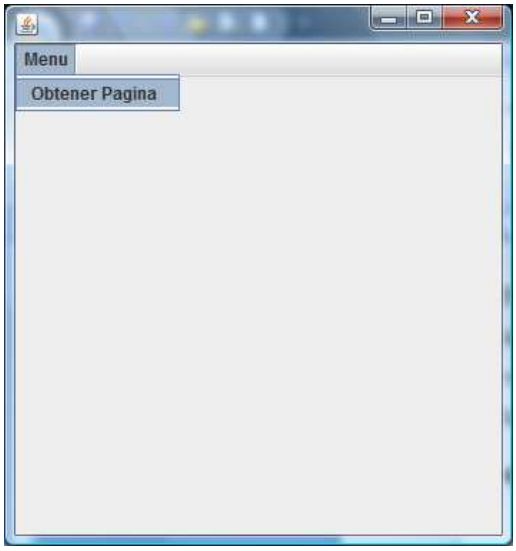


Figura 8. Pantalla de menú

5.3 Pantalla de periódicos disponibles

En esta sección se muestran las pantallas donde se permite consultar la información de dichos periódicos digitales.

Pantalla de periódicos disponibles

La Figura 9 muestra la pantalla principal en donde muestra los periódicos disponibles para la recuperación de los encabezados de dichos periódicos, aquí se puede seleccionar cualquier periódico con un solo click y esperar unos segundos para que muestre los encabezados de las noticias como se muestra en la Figura 10.

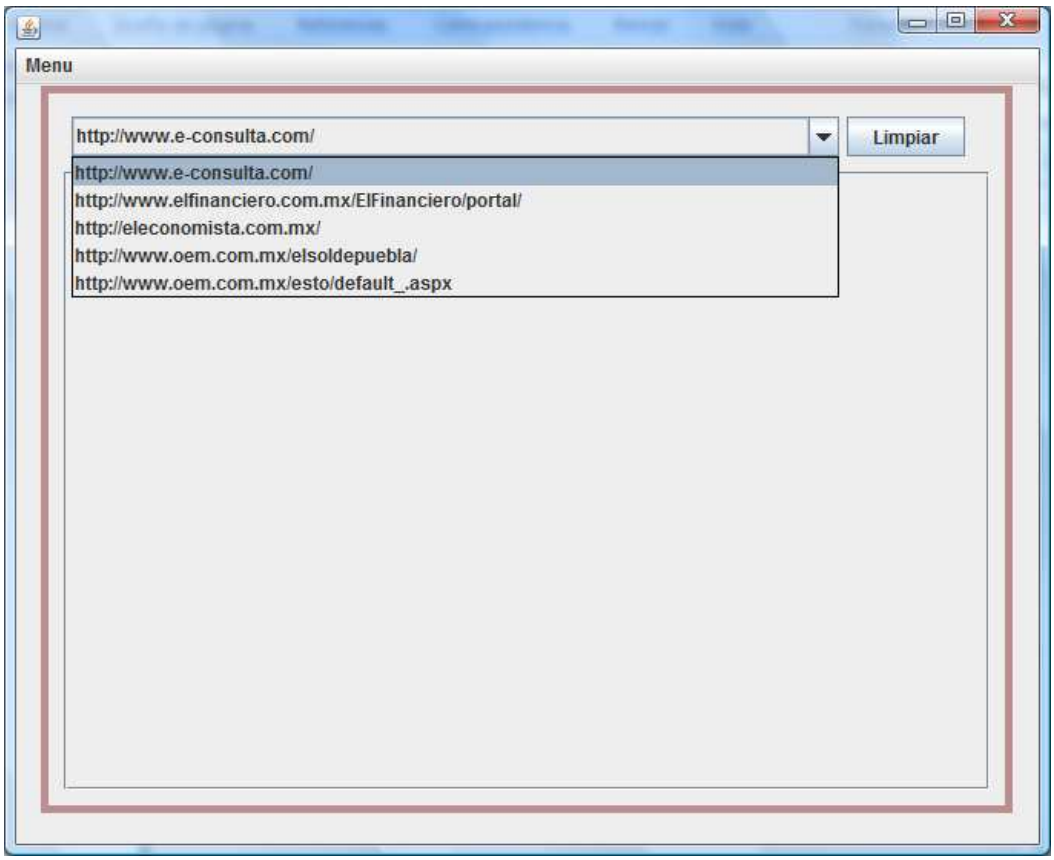


Figura 9. Periódicos disponibles

5.4 Pantalla consulta de resultados

En esta sección se muestran las pantallas donde se permite consultar la información que se obtuvo del periódico digital de libre acceso, además tiene la opción de guardarlas en un archivo txt.

Pantalla consulta de resultados El Financiero

En la Figura 10 muestra los resultados que fueron obtenidos de la consulta al periódico junto con la url en caso de que requiera visualizar la noticia completa da click en el botón que está del lado derecho de la noticia y muestra la noticia completa como se da a conocer en la Figura 11.

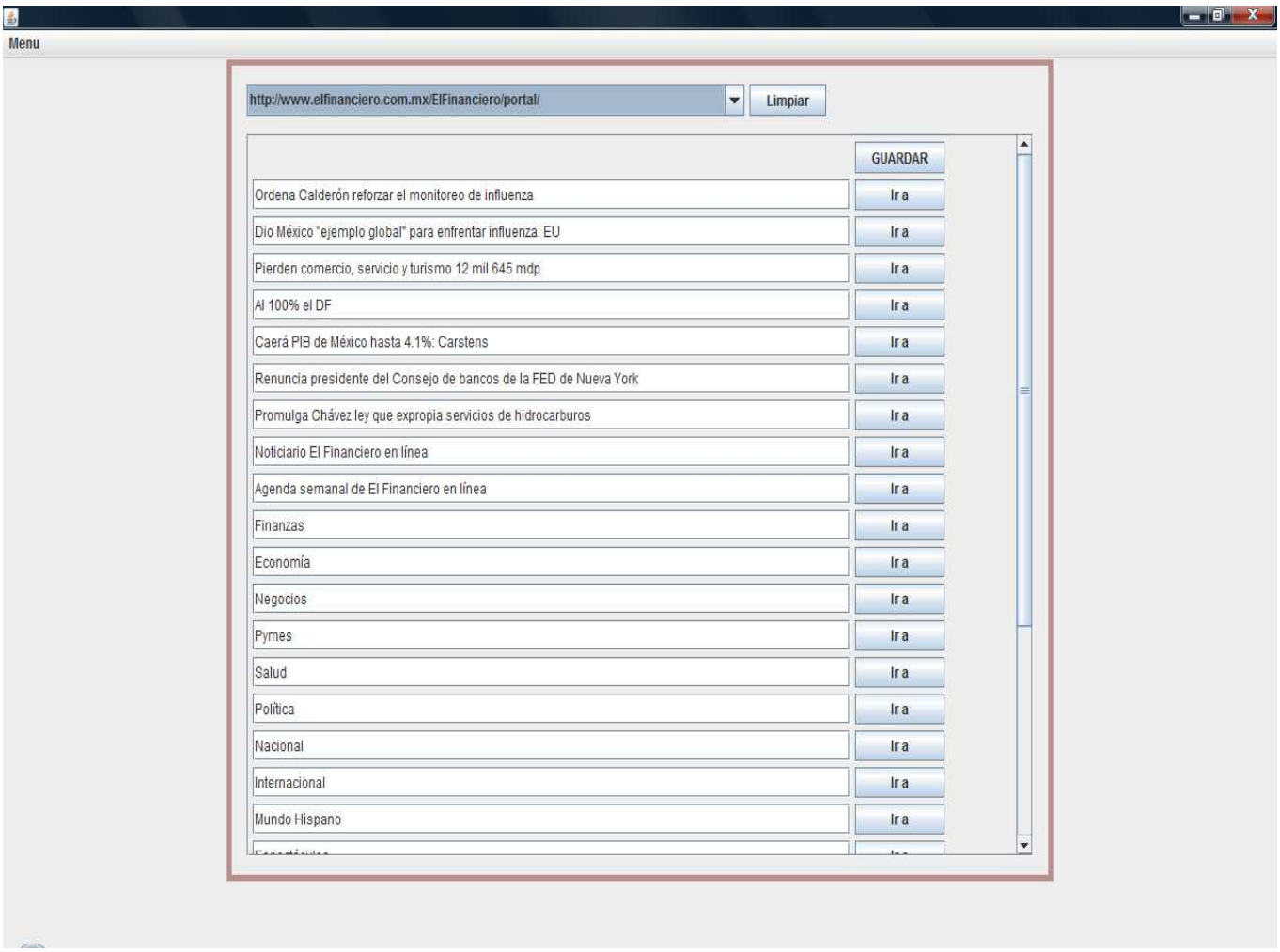


Figura 10. Consulta de resultados

5.4.1 Pantalla noticia completa

En esta sección se muestra la pantalla donde se permite consultar toda la noticia completa que se obtiene de la url del periódico, esto se muestra cuando se da clic en el botón “ir a”.

Pantalla noticia completa

En la Figura 11 muestra los resultados que fueron obtenidos de la liga de la noticia que fue recuperada.

Figura 11. Noticia completa

5.4.2 Pantalla para guardar los encabezados

En esta sección se muestra la pantalla donde se permite guardar los encabezados de las noticias que fueron recuperadas junto con la url de dichas noticias, éste lo guardará como un archivo de texto txt.

Pantalla Guardar encabezados

En la Figura 12 muestra en donde se desean guardar los encabezados de las noticias que fueron recuperadas.

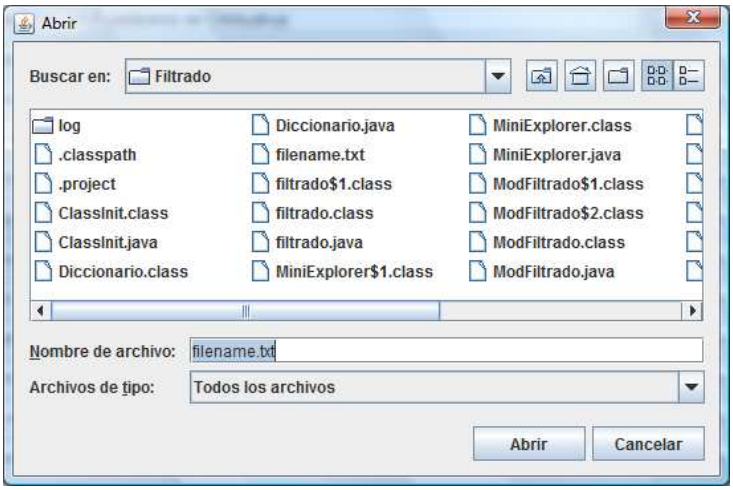


Figura 12. Guardar encabezados

5.4.3 Pantalla Encabezados y URL

En esta sección se muestra la pantalla donde se almacenaron los encabezados de los periódicos que fueron obtenidos anteriormente junto con la url de la noticia. Esto genera que se puedan almacenar las noticias más relevantes y de interés para la Universidad Politécnica de Puebla.

Pantalla Encabezados y URL

En la Figura 13 muestra en un bloc de notas todas las noticias que fueron recuperadas junto con su respectiva url de los periódicos digitales de libre acceso.

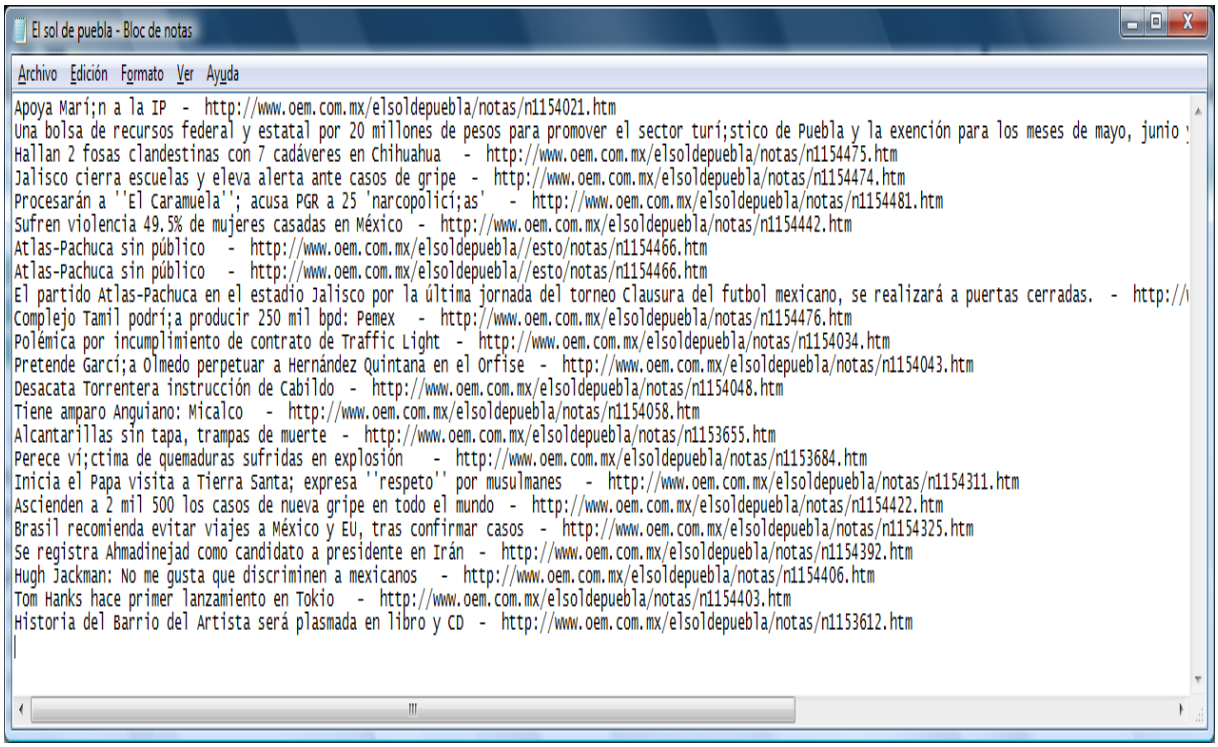


Figura 13. Encabezados y URL

Referencias

[PLN 2005] Asociación Mexicana para el Procesamiento del Lenguaje Natural (PLN). Copyright 2005.

Fecha de acceso: 12 de septiembre del 2008.

Disponible en: http://www.ampln.org/NLP.htm

[Tramullas 1997-2000] Tramullas, J. 1997-2000. *Introducción a la documática*, sección 3 Recuperación de información.

Fecha de acceso: 16 de septiembre de 2008

Disponible en: http://tramullas.com/documatica/3-1.html

[Mitchell 1997] Mitchell, T. 1997. *Machine Learning*, McGraw Hill

Disponible en: "http://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico"

[Pinto 2004] Pinto M. 2004 *Búsqueda y recuperación de información*

Fecha de acceso: 9 de octubre de 2008

Disponible en: http://www.mariapinto.es/e-coms/recu_infor.htm#ri1

[América 2008] America.gov. Copyright 2008.

Fecha de acceso: 21 de octubre del 2008.

Disponible en: http://www.america.gov/esp/

[Hipertext 2008] Departamento de Periodismo y de Comunicación Audiovisual

(Hipertext). Copyright 2008.

Fecha de acceso: 14 de noviembre de 2008

Disponible en: http://www.hipertext.net/web/pag277.htm#Referencias

[Vilares 2006] Vilares, J. 2006. Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español. Departamento de Computación Universidad de Coruña, campus de Elviña s/n, Marzo

[UTM 2002] Universidad Tecnológica de Malasya. Clasificación de noticias Web utilizando redes neuronales basadas en ACC. Copyright 2002

Fecha de acceso: 21 de mayo de 2009

Disponible en: http://eprints.utm.my/3089/1/sice02-0163.pdf

[Juárez 2007] Juárez, A. 2007 Extracción de respuestas mediante aprendizaje automático utilizando atributos léxicos. Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Enero

[Haddad 2008] Haddad, I. 2008 Algoritmo de segmentación de imágenes usando los K- vecinos más cercanos. Reporte Técnico IRI-TR 08-07. Instituto de Robótica e Informática Industrial. Agosto

[Téllez et a1] Téllez et a1. s/a. Aplicando la clasificación de texto en la extracción de información. Coordinación de Ciencias Computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE).

[Díaz 2007] Díaz, C. 2007. Clasificación no supervisada, clustering y mapas auto organizativos (Kohonen), Abril

[Ponce et a1] Ponce, et a1. s/a. Recuperación de información de páginas web mediante una ontología que es poblada usando clasificación automática de textos. Centro Nacional de Investigación y Desarrollo Tecnológico

[Figuerola et a1 2004] Figuerola et a1. 2004 Algunas técnicas de clasificación automática de documentos. Universidad de Salamanca

[Melville 2002] Melville. 2002 Weka Tutorial

Fecha de acceso: 25 de junio de 2009

Disponible en: http://www.cs.utexas.edu/users/ml/tutorials/Weka-tut/index.htm