VIRTUAL-ACT. ASISTENTE PARA LA GENERACIÓN DE CONFERENCIAS VIRTUALES.

M.Sc. Antonio Celso Fernández Orquín[1], Ing. Lisandra Miranda Santana[2]

1. Universidad de Matanzas, Autopista a Varadero km 3/12, Matanzas, Cuba.

2. Delegación Gaviota, Varadero., Matanzas, Cuba.

CD de Monografías 2008

Resumen.

Virtual-Act: Fue creada para asistir a los profesores en la elaboración conferencia en formato electrónico. Utilizando técnicas de Extracción de Información, permite crear y transformar las conferencias a una nueva forma de representación. Dichas transformación fueron hechas con empleo de recursos de Procesamiento del Lenguaje Natural, específicamente la extracción de información, la detección de patrones y la detección de entidades, así como las expresiones regulares. La investigación comprende un estudio del estado del arte, se plasman también los elementos teóricos que sirven de base para el desarrollo de la propuesta. Se refleja el proceso de construcción y pruebas de la propuesta de solución. Para el análisis de los resultados se utilizan los indicadores de precisión y cobertura, calculándose la F-medida y comparando estos valores con los referidos a nivel internacional. Como resultado se obtiene un objeto de aprendizaje que puede ser manipulado por los Entornos Virtuales de Aprendizaje.

***Palabras claves**: Generación de documentos electrónicos, Transformación de documentos electrónicos; Conferencias en formato electrónico.*

Introducción.

Apreciando el auge y la popularidad que hoy tiene la educación a distancia, así como el método de enseñanza semipresencial, y aprovechando el avance de la computación, se puede observar que, en aras de contribuir al perfeccionamiento del proceso de enseñanza, se ha difundido bastante la utilización de los entornos virtuales de aprendizaje (EVA), o también llamados Gestores de Contenidos. Éstos permiten utilizar diversos tipos de materiales y recursos educativos. Centrando la atención en la educación superior, se toma como referencia la Universidad de Matanzas "Camilo Cienfuegos" (UMCC).

Si bien estas bondades de la informática han venido a ofrecer recursos potentes para enfrentar nuevos retos en la educación, también habría que decir que han obligado a pensar en las modificaciones necesarias que deben sufrir las estructuras tradicionales y los viejos métodos de enseñanza. Algunas de las transformaciones están íntimamente ligadas a la necesidad de la explotación eficiente de las posibilidades que ha traído a la enseñanza, las TIC y la Computación.

Se puede decir que la situación problemática, se debe a dos razones fundamentales: desconocimientos de la totalidad de los recursos educativos que brindan estas herramientas informáticas y a la no existencia de un recurso capaz de asistir a los profesores, no especialistas en informática, en la transformación necesaria que tienen que aplicar a sus asignaturas, materiales y recursos virtual, para que puedan aprovechar al máximo las ventajas de estás tecnologías.

Se define entonces como problema científico a resolver: La Inexistencia de una herramienta informática que guíe a los profesores en la elaboración de actividades docente tipo conferencias metodológicamente bien estructuradas, permitiendo la transformación de las existentes. La hipótesis a demostrar sería: Una herramienta informática que haga uso de técnicas de procesamiento del lenguaje natural y un lenguaje de marcado, permitirá transformar las conferencias en formato electrónico a una forma de representación que garantice una mejor estructuración de estos documentos. Para resolver el problema planteado, se traza como objetivo: Diseñar una herramienta que sea capaz de generar documentos docentes, tipo conferencia, con una propuesta de estructuración metodológica, permitiendo convertir automáticamente las existentes a la nueva forma de representación.

En el entorno universitario cubano no se encontró un software que asista a los profesores en la confección de documentos docentes tipo conferencia, que permita además procesar los documentos existentes. Tampoco se ha tenido referencia de la utilización de un lenguaje de marcado como XML, elemento este que permitirá realizar búsquedas especializadas por las partes componentes. Para el procesamiento automático de los documentos existentes se utilizaron técnicas de Procesamiento del Lenguaje Natural (PLN), apoyando la detección de entidades y la extracción de información con el uso de expresiones regulares. Se obtiene un objeto de aprendizaje SCORM que puede ser manipulado por los EVA.

El trabajo se estructura de la siguiente forma:

- **¡Error! No se encuentra el origen de la referencia.**. Tiene en cuenta aspectos referentes a los antecedentes del trabajo y objetos de estudio. Se presentan algunos aspectos que servirán de base para la solución propuesta.
- Descripción y construcción de la solución propuesta . Describe algunos aspectos generales sobre la concepción inicial del sistema. Se hace un estudio de la factibilidad para el proyecto de software
- Análisis de los resultados obtenidos. Aborda fundamentalmente el análisis de los resultados obtenidos, teniendo en cuenta los criterios de los clientes.

1. Fundamentación teórica.

1.1. Antecedentes del trabajo

No se ha encontrado evidencia de la existencia de sistemas informáticos que usen técnicas de extracción de información (EI) con incidencia sobre los documentos tipo

conferencia en la Educación Superior Cubana. En la UMCC se desarrolló un sistema que extrae información de otro tipo de documentos docentes, los Programas de Disciplina.

Los trabajos realizados por el CREA para homogeneizar la presentación y el contenido de los documentos asociados al Proceso Docente Educativo (PDE), hasta donde se tiene conocimiento, se basan en plantillas de *Microsoft Word* entregadas a los profesores para la posterior transformación a HTML[1] y colocación en una plataforma de contenidos, por un grupo de especialistas. No se tiene referencia de publicación de esta información o de alguna mejora del tratamiento de la información con técnicas más novedosas, aunque existen otros trabajos que tienen incidencia en el PDE.

La investigación realizada es parte de un proyecto mayor conocido como SIGID-PLN encaminado a mejorar la Gestión de Información relacionada al Proceso Docente en la Educación Superior utilizando técnicas de Procesamiento de Lenguaje Natural. Actualmente participan en dicho proyecto otros estudiantes y profesores de la facultad, pertenecientes al grupo de investigación ANUBIS.

1.2. Formas de Enseñanza: la actividad Conferencia.

Para Carlos Álvarez (Álvarez de Zayas, 1988) la forma de organización como categoría es "la estructuración y el ordenamiento interno de los componentes personales de dicho proceso: docente y estudiante, y de los elementos de contenido de las disciplinas: conocimientos y habilidades". En (Álvarez de Zayas, 1988, Colectivo de Autores, 2003) aparecen diferentes clasificaciones de las formas de enseñanza, la Tabla 1.1 muestra la clasificación para el Curso Regular Diurno (CRD) atendiendo a su función.

Tabla 1.1 Clasificación de las Formas de Enseñanza en la Educación Superior.

Tipo de actividad	Forma organizativa	Tipo de forma organizativa
Académica	Clase	Conferencia, Clase Práctica, Clase Encuentro, Seminario y Práctica de Laboratorio
Laboral	Práctica laboral, Estancia, práctica docente, educación en el trabajo	Depende del tipo de profesional
Científico-Investigativo	Trabajo investigativo de estudiantes	Trabajo de Diploma o Proyecto de Curso

En la actualidad, la concepción de las formas de organización debe incorporar a esta categoría lo que las TIC traen como retos al proceso de enseñanza aprendizaje. Es

[1] Hipertext Markup Language- Lenguaje de etiquetado hipertextual de documentos para su presentación en los navegadores Web.

frecuente escuchar términos como entornos virtuales y otros que utilizados de forma creativa también son importantes formas que permiten organizar el enseñar y el aprender en estas nuevas condiciones. (Colectivo de Autores, 2003)

Teniendo en cuenta la clasificación de las formas de enseñanza que se presenta en la Tabla 1.1 se puede decir que esta investigación se centrará en el tipo de actividad académica, cuya forma organizativa es la clase y específicamente de tipo conferencia. Sin impedir que la actividad conferencia juegue su papel como trasmisora de nuevos conocimientos, éstos pueden ser proporcionados sin la necesidad de la presencia del profesor.

1.3. La actividad docente conferencia: su estructura metodológica.

El reglamento de Trabajo Docente Metodológico platea aspectos que ayudan a entender algunos términos. En el Articulo 107: "La conferencia es el tipo de clase que tiene como objetivo principal la transmisión a los estudiantes de los fundamentos científicos-técnicos más actualizados de una rama del saber con un enfoque dialéctico-materialista, mediante el uso adecuado de métodos científicos y pedagógicos, de modo que les ayude en la integración de los conocimientos adquiridos y en el desarrollo de las habilidades y valores que deberán aplicar en su vida profesional". (MES, 2007)

La estructura organizativa de la clase presenta tres momentos en su proceso: Introducción, Desarrollo y Conclusiones.

Haciendo un análisis de lo que se plantea en (Colectivo de Autores, 2003, Verrier, 2005) y las consideraciones de los pedagogos de la UMCC entrevistados, la estructura metodológica de la conferencia utilizada es:

1. Introducción.

   - Rememoración sobre la clase anterior.
   - Preguntas de control.

Observación: Estos dos momentos se pueden unir, es decir, a través de preguntas de control, realizar la rememoración de lo anterior.

2. Desarrollo

   - Motivación de la clase.
   - Objetivo(s).
   - Temática.

- Contenido.
- Trabajo Independiente.
- Bibliografía.

3. Conclusiones

- Realizar las preguntas de comprobación sobre el cumplimiento del objetivo o los objetivos formulados
- Presentar las Generalizaciones científico-teóricas más importantes. No es hacer resumen de la Conferencia, son Generalizaciones.
- Si es necesario y sigue a la Conferencia que se desarrolla otra Conferencia, se puede hacer una motivación, según la temática a desarrollar, pero que no sea una cosa forzada. Además pudiera motivarse el desarrollo de la próxima clase independiente del tipo que sea, aumentando en el estudiante la motivación por el tema.

1.4. Análisis crítico de cómo se ejecutan actualmente esos procesos, las causas que originan la situación problemática y las consecuencias.

En la UMCC están actualmente en uso dos gestores de contenido: Claroline y Moodle en los que se colocan los recursos educativos. En estas plataformas se encuentran materiales de todas las carreras de la universidad y de la mayoría de las asignaturas, por lo que cualquiera de ellos constituye una fuente para realizar un muestreo del tipo de actividad docente que se desea analizar.

Cuando un estudiante decide consultar alguno de estos materiales, la mayoría de las veces lo que hace es descargarlo de la plataforma para leerlo o imprimirlo, ya que no existe la posibilidad de algún tipo de interacción con los recursos que se suben a la plataforma.

Al surgir la necesidad de consultar alguna conferencia que aborde un tema específico, que haya sido impartida por un profesor determinado, o incluso se requiera la de un curso dado; no existe un mecanismo que facilite la obtención del resultado esperado. La implementación de un buscador como solución para esta problemática no sería del todo eficiente debido a la inexactitud de los algoritmos que se utilizan para la recuperación de información.

Para detallar más sobre el estado actual de la temática se debe centrar la atención en uno de los gestores de contenido y seleccionar una muestra para analizar los resultados arrojados.

1.5. Análisis de los documentos conferencias existentes en los gestores de contenidos en la UMCC

Inicialmente se tomó para el análisis la plataforma Moodle, ya que Claroline estaba en proceso de transformación en el momento en que se realizó el estudio. Además otro punto a su favor es que permite el manejo de paquetes SCORM.

Para determinar el tamaño adecuado de la muestra que se debe seleccionar, basado en que la variable de medición es binomial, es decir, que el objeto de muestreo: la conferencia está metodológicamente correcta o no; se usa la siguiente fórmula según Pita Fernández en (Fernández, 1996).

$$n = \frac{Z_{\alpha}^{2} * p * q}{d^{2}}$$

n es el tamaño mínimo de la muestra que se debe analizar, al tomar p y q como 0.5 por no tener proporción esperada, se garantiza además maximizar el tamaño de la muestra. Como se conoce que la población es finita, es decir, se sabe la cantidad de conferencias que existen en Moodle hasta la fecha de realización del estudio, y se desea saber cuántos del total hay que tomar como muestra para estudiar, la fórmula sería:

$$n = \frac{N * Z_{\alpha}^{2} p * q}{d^{2} * (N - 1) + Z_{\alpha}^{2} * p * q}$$

Tomando en cuenta el nivel de confianza en un 95%, sería:

Zα 2 = 1.962 (nivel de confianza 95%), p = 0.5 (proporción esperada), q = 1 – p (en este caso 1 – 0.5 = 0.5), d = 0.05 (precisión, en este caso se espera un 0.5% de error), N= 327 conferencias.

La muestra que se debe tomar es de 177 conferencias, del análisis de estas se determinó que el 48 % de ellas se pueden considerar metodológicamente correctas y el resto no lo está. De lo anterior podemos concluir lo siguiente:

- Los profesores no aprovechan al máximo las potencialidades que brindan los ordenadores, se limitan a colocar, cuando lo hacen, conferencias en formato Word.

- Algunos profesores ponen sólo las presentaciones de PowerPoint, lo que mejora en algo la interacción, pero no contienen la estructura metodológica completa ya que se utilizan para guiarse durante el desarrollo de la conferencia, lo cual; no cumple con lo que se consideró como conferencia virtual para esta investigación.
- No todos los profesores siguen la estructura metodológica propuesta por el colectivo de pedagogos de la UMCC.
- El editor de texto más utilizado para la confección de los materiales docentes es Microsoft Word. Como ya se ha mencionado, con este modelo, no existe una buena interacción y se dificultaría cualquier intento de hacer búsquedas especializadas sobre las partes componentes de estos materiales, a pesar de encontrarse en formato electrónico.

También del análisis de la plataforma Moodle se concluye que:

- De los cursos analizados una parte, aunque minoritaria, no coloca materiales docentes en el curso (a veces sólo bibliografía), se encontraron 43 cursos de un total de 108 sin las conferencias correspondientes.
- Casi la totalidad de los cursos tiene las conferencias en ficheros de *Microsoft Word*, exceptuando a 6 cursos que tienen presentaciones de *Microsoft Power Point* y de estos 3 sólo las presentaciones que utilizan los profesores para guiarse durante la clase.

2. Fundamento científico o marco teórico.

2.1. PLN: Procesamiento del lenguaje Natural.

El **Procesamiento de Lenguajes Naturales, PLN**, o **NLP** del idioma inglés *Natural Language Processing,* "es una subdisciplina de la Inteligencia Artificial y la rama ingenieril de la lingüística computacional. El PLN se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para la comunicación entre personas o entre personas y máquinas por medio de lenguajes naturales. El PLN no trata de la comunicación por medio de lenguajes naturales de una forma abstracta, sino de diseñar mecanismos para comunicarse que sean eficaces computacionalmente -que se puedan realizar por medio de programas que ejecuten o simulen la comunicación-. Los

modelos aplicados se enfocan no sólo a la comprensión del lenguaje de por sí, sino a aspectos generales cognitivos humanos y a la organización de la memoria".(Wikipedia: la enciclopedia libre)

Las principales tareas de trabajo en el PLN son (Wikipedia: la enciclopedia libre): Síntesis del discurso, Análisis del lenguaje, Comprensión del lenguaje, Reconocimiento del habla, Síntesis de voz, Generación de lenguajes naturales, Traducción automática, Respuesta a preguntas, Recuperación de la información, Extracción de la información

2.2. Detección de entidades

La detección o Reconocimiento de entidades (RE), como más comúnmente se conoce en la comunidad científica, es una parte importante de los Sistemas de Extracción de Información, porque al reconocer y clasificar las entidades existentes en un texto, se puede tener una información más clara y una mejor interpretación del documento. Algunas de las técnicas que se aplican en la detección y clasificación de entidades son las basadas en conocimiento y las basadas en aprendizaje.

En esta investigación no se utiliza ninguna de las técnicas antes mencionadas para detectar las entidades, puesto que se conocen previamente según la estructura metodológica de la conferencia. Para la detección de las entidades se emplean como recurso las expresiones regulares, que son construidas previamente según los patrones encontrados durante el análisis de las conferencias ubicadas en los gestores de contenido de la UMCC y otras obtenidas por los autores en búsquedas realizadas.

2.3. Extracción de Información

El objetivo de la Extracción de Información (EI) es identificar y extraer información de forma automática de texto escritos libremente en lenguaje natural. Horacio Rodríguez la define en (Rodríguez, 2001) como el proceso de "localizar las porciones de un texto dado que contengan información relevante para las necesidades de un usuario y proporcionar dicha información de forma adecuada a su proceso (manual o automático)", destacando que "el criterio de relevancia viene indicado por modelos predefinidos (normalmente mediante modelos Objeto/Atributo/Relación)". Típicamente un SEI extrae informaciones sobre entidades, relaciones y eventos a partir de documentos en un dominio restringido. La principal aplicación de un SEI es llenar una base de datos con información proveniente de textos sin un formato predefinido.

El SEI desarrollado realiza este proceso con algunas variaciones según lo analizado anteriormente. La información extraída es volcada a una estructura de datos con la que se genera un fichero XML en el cual la información se almacena de forma estructurada según las partes de la conferencia y con el uso de elementos definidos por la autora.

2.4. Expresiones Regulares

En el área de la programación, las expresiones regulares son un método por medio del cual se pueden realizar búsquedas en cadenas de caracteres; sin importar la longitud de la búsqueda requerida, ni la de la cadena de búsqueda. (Wikipedia: la enciclopedia libre) Si es necesario encontrar todas las apariciones de un patrón definido de caracteres en un archivo de millones de caracteres, las expresiones regulares proporcionan una solución para el problema. Una expresión regular es un modelo de texto formado por caracteres ordinarios y caracteres especiales, conocidos como meta caracteres. El modelo describe una o varias cadenas que deben coincidir al buscar texto.(MSDN Library para Microsoft VisualStudio 2005). También se puede definir el término expresión regular (patrón) como "notación para representar un conjunto de cadenas mediante una sola cadena"(Madrigal, 2005). Los metacaracteres son una importante herramienta para hacer que con una sola cadena se capturen muchas otras que cumplan con el patrón representado por medio de esta.

Se puede decir entonces que las expresiones regulares proporcionan un método eficaz y flexible para procesar texto. Las expresiones regulares de *Microsoft .NET Framework* incorporan las funciones más comunes de otras implementaciones de expresiones regulares, como las de Perl y awk.(MSDN Library for Microsoft VisualStudio 2005).

2.5. Objetos de Aprendizaje

En los últimos tiempos ha tenido un gran impacto en el e-learning[2] el concepto de objeto de aprendizaje (LO)[3]. Incluyó un nuevo paradigma de creación de contenidos que requirió un cambio drástico en el diseño instruccional, la arquitectura de las plataformas y de los sistemas de distribución de contenidos. (Pedruelo, 2004)

Los LO fueron seleccionados por la IEEE[4] para distribuir pequeños componentes instruccionales, auto-contenidos y reutilizables, que se pueden distribuir a través de Internet. Para la IEEE, un objeto de aprendizaje es "cualquier entidad, digital o no digital, que puede ser utilizada o referenciada durante el aprendizaje basado en el ordenador". También se define como "cualquier recurso digital que puede ser utilizado para soportar aprendizaje". "Un objeto de aprendizaje es un bloque que puede combinarse de infinitas formas para construir colecciones de objetos que forman lecciones, cursos, módulos, etc. (Pedruelo, 2004)

---

[2] Puede definirse como: "aquella actividad que utiliza de manera integrada y pertinente computadores y redes de comunicación, en la formación de un ambiente propicio para la construcción de la experiencia de aprendizaje".

[3] LO por sus siglas del inglés Learning Object.

[4] Institute of Electrical and Electronic Engineers por sus siglas en inglés.

En la solución presentada por esta investigación se emplean los paquetes SCORM (Sharable Content Object Reference Manual) como representación final de la información que se obtiene. SCORM es uno de los más representativos dentro de la categoría de centrados en el contenido, pero no es un estándar, es un modelo desarrollado de una colección de especificaciones.

La forma más básica es un Recurso. "Los Recursos son representación electrónica de textos, imágenes, sonidos, objetos de evaluación o cualquier otra entidad que pueda mostrarse en un navegador. Un recurso puede combinarse con otros para crear nuevos recursos. Se describen utilizando metadatos que permiten su búsqueda en repositorios de recursos y su reutilización." (Pedruelo, 2004)

Un SCO "es un una colección de uno o más Recursos que representan un recurso de aprendizaje ejecutable capaz de comunicarse y de ser lanzado por una plataforma de formación. Es la unidad más pequeña que la plataforma puede manejar." (Pedruelo, 2004)

Una vez que el contenido de aprendizaje está construido, es necesario ponerlo a disposición de los alumnos, repositorios o plataformas de contenidos. Para ello, SCORM utiliza de forma estricta la especificación IMS *Content Packaging Specification* (Pedruelo, 2004). Esto hace posible disponer de una forma estandarizada para intercambiar contenido entre distintas plataformas y una descripción de la estructura y del comportamiento de una colección de contenidos de aprendizaje.

Un Paquete de Contenidos está formado por dos componentes: un documento en XML que describe la estructura del contenido y los recursos, llamado manifiesto (imsmanifest.xml), y los ficheros físicos (o URL) con el contenido real del paquete. "Representa una unidad de aprendizaje que tiene relevancia instruccional y puede repartirse independientemente" (Pedruelo, 2004). Se decide tomar como unidad independiente la conferencia, el paquete generado por la herramienta propuesta.

En el mercado existe una variedad de plataformas que soportan SCORM que explícitamente utilizan algún estándar completo (habitualmente IMS o SCORM) o una parte de ellos (Pedruelo, 2004). Moodle es una de ellas, soporta la utilización de objetos SCORM, incluye también herramientas que facilitan la migración de cursos entre versiones distintas. En la UMCC esta plataforma es una de las que se utiliza para colocar los recursos del PDE.

Como resultado del uso de la herramienta propuesta se obtendrán paquetes de contenido SCORM. Se toma, para esta investigación, como unidad relevante de aprendizaje el documento Conferencia por su importancia dentro del PDE. Se utiliza para la creación de los paquetes la aplicación "AppHerramientaIA" (Alfonso, 2008), desarrollada por Leobel Pérez Alfonso estudiante de la Facultad de Informática.

3. Descripción y construcción de la solución propuesta

3.1. Ámbito del software.

La primera actividad de gestión de un proyecto de software es determinar el ámbito. Pressman en [20] plantea que se puede definir respondiendo algunas cuestiones sobre el contexto, objetivos de información, función y rendimiento.

Virtual-Act., Herramienta para asistir a los profesores en la confección de conferencias en un entorno virtual: Genera la conferencia como objeto de aprendizaje en un paquete SCORM. Permite confeccionar nuevas conferencias, delimitando claramente cada parte de la estructura metodológica, además de capturar las ya realizadas con otros programas y que se encuentren en formato .doc o HTML. La interfaz debe permitir el trabajo con las distintas partes del documento conferencia sin que el acceso a partes no consecutivas sea secuencial. Ofrece ayuda sobre la estructura metodológica de este tipo de documento docente y sobre el uso de la aplicación. Está dirigido a los profesores del entorno universitario cubano.

La entrada de información al sistema es cualquier documento conferencia en los formatos antes mencionados. Este se somete a las técnicas PLN para extraer la información correspondiente a cada parte. El contenido puede ser editado utilizando las bondades que ofrece la herramienta. Ya sea con la información proporcionada por el profesor, en caso de elaborar una nueva conferencia, como con la obtenida luego del proceso de extracción, se genera un fichero XML. Haciendo uso del software HerramientaOA (Alfonso, 2008) se crea un paquete SCORM que contiene, además del fichero XML, información adicional sobre su presentación futura en un gestor de contenido que pueda manejar estos objetos de aprendizaje.

Las conferencias obtenidas facilitarán la búsqueda de cualquier información específica de este dominio, una estructura física con delimitación de la información de cada parte e incorpora además elementos de navegabilidad vinculando cada epígrafe de la temática con su desarrollo dentro del contenido de la conferencia.

3.2. Descripción general de la propuesta

Al utilizar el asistente se puede elaborar conferencias nuevas, pero también se pueden cargar las existentes y transformarlas a la nueva estructura de forma automática. Cuando se le proporciona a la herramienta un documento de Word o HTML, se guarda en los archivos del sistema en formato HTML, es decir, para el caso de que sea un documento Word se convierte a HTML utilizando recursos del lenguaje y si es HTML es guardado directamente. A partir de este momento comienza el procesamiento transitando por las tres fases que se explican a continuación.

3.2.1. Fases del proceso

Fase I: Detección de entidades

La primera fase es la detección de entidades, consiste en localizar las entidades presentes en el documento del conjunto previamente fijado. El dominio es restringido por lo que antes de comenzar se definen las entidades. En la

Figura 3.1 se muestra el conjunto de entidades básicas de la conferencia, aunque para representar toda la información se definieron otras que se explicaran más adelante.

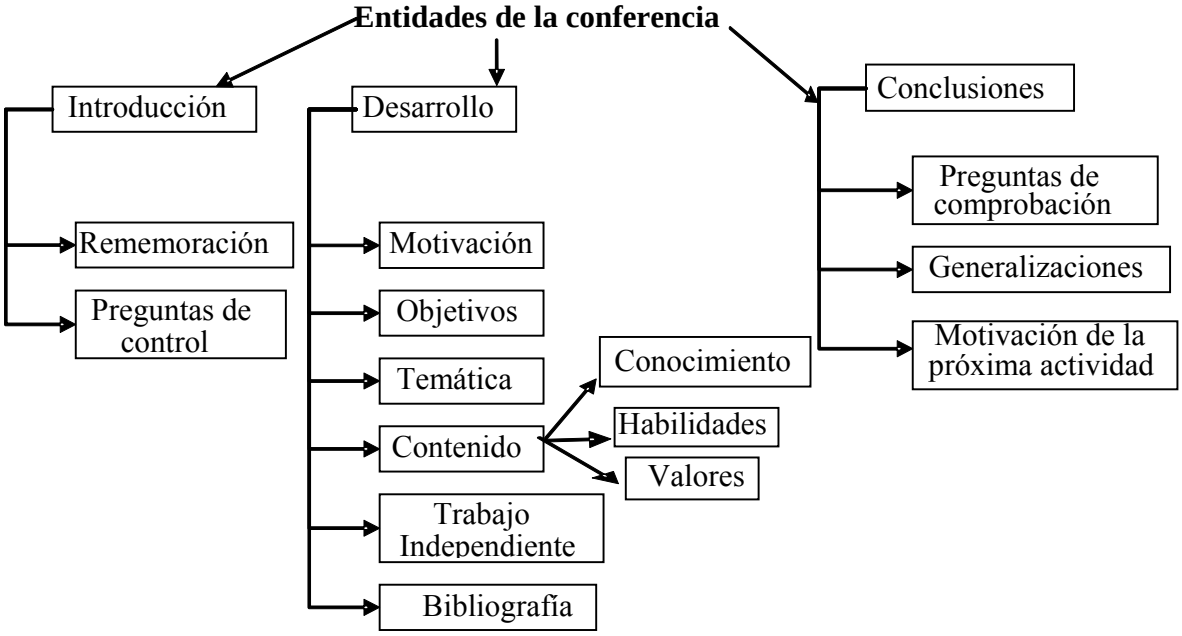

Figura 3.1 Conjunto de entidades para la conferencia

Para el proceso de detección de entidades lo primero que se hace es eliminar las etiquetas innecesarias del documento HTML, esto se realiza con el apoyo de las expresiones regulares construidas a partir de los patrones del estándar HTML para cada tipo de etiquetas. Este proceso se le denomina Limpiar HTML. Luego de obtener un documento HTML limpio, se pasa a toquenizar el documento. Esto significa llevar a átomo el texto que se va a procesar, para este caso en particular se considera átomo cada etiqueta HTML de apertura o de cierre y el texto contenido en ellas, quedando en cada línea del documento una etiqueta (de apertura o cierre) o texto. Se procede entonces a detectar cada una de las entidades especificadas en la

Figura 3.1. Las entidades encontradas se etiquetan. Debido al volumen de información que se maneja (generalmente más de 3000 líneas), el proceso de etiquetado se simula guardando los índices de posición que delimitan las entidades. Las expresiones regulares

son el recurso fundamental en esta fase, para su construcción y prueba se utilizaron algunas herramientas que facilitaron el trabajo y se probó con conferencias obtenidas de la UMCC.

Es necesario destacar que las expresiones identifican las entidades indistintamente en mayúscula o minúscula, pero sólo si aparecen como subtítulo, no si se encuentran como parte del contenido, es decir, si la línea empieza con la coincidencia de la expresión.

Fase II: Extracción de información

Después de identificadas las entidades se extrae la información de cada entidad detectada y la información se almacena. Este proceso también se sustenta en las expresiones regulares y se repiten las fases I y II, pero para las entidades específicas del marcado HTML del documento. Después que la información está contenida en la estructura de datos, la instancia de la clase conferencia es serializada generando un fichero XML que contiene el marcado según la información extraída.

Fase III: Empaquetado

La tercera y última fase es la de empaquetado de contenidos que proporciona la funcionalidad para describir y empaquetar las conferencias en XML como objetos de aprendizaje. El empaquetado de contenidos señala la descripción, estructura y localización de los materiales de aprendizaje. Se empaqueta utilizando el modelo SCORM y una herramienta externa que los genera. Es necesario crear algunos ficheros que especifican qué empaquetar y cómo hacerlo. El resultado de esta fase es la creación de un fichero comprimido que contiene, a parte de todos los ficheros que le pertenecen, un fichero XML que contiene todas las etiquetas que definen la estructura del contenido y los recursos que lo conforman.

4. Análisis de los resultados obtenidos

4.1. Evaluación de efectividad y eficiencia

Se analiza como medida de evaluación la efectividad, que determina la precisión del proceso de detección de entidades y de la extracción de la información relevante en las conferencias. Según los resultados obtenidos por Fernández (Orquín, 2005), se puede emplear para determinar la efectividad del Reconocimiento de Entidades las Fórmula 4.1 y Fórmula 4.2, se adoptará la precisión como el resultado de dividir las entidades identificadas correctamente entre el total de entidades identificadas. La cobertura o recall se define como la división entre las entidades identificadas entre el total de entidades que existen en el documento en análisis.

$$precisión = \frac{entidades\ identificadas\ correctamente}{entidades\ identificadas} \qquad recall = \frac{entidades\ identificadas\ correctamente}{entidades\ en\ el\ documento}$$

Fórmula 4.1 Precisión Fórmula 4.2 Cobertura

La fórmula 3.3 muestra otra de las medidas utilizadas es la F-medida que es una función que regula el balance entre la precisión y cobertura a través del parámetro β. Para cualquier valor de β, F_β se encontrará en el rango entre cero y uno, para β=1 precisión y cobertura tienen la misma importancia.

$$F_\beta = \frac{(\beta^2 + 1) * precisión * cobertura}{(\beta^2 * precisión + cobertura)}$$

Fórmula 4.3 F-medida

El problema que se analiza tiene un dominio restringido y las entidades fueron determinadas previamente. Para el cálculo de precisión y cobertura las entidades que se van a tener en cuenta son la correspondientes a las partes de la conferencia según la estructuración metodológica.

Se seleccionó al azar un conjunto de quince conferencias de diferentes carreras que estaban colocadas en *Moodle* y que no formaron parte del conjunto de entrenamiento del sistema. Se procesaron con la herramienta y se llegó a los resultados que se muestran en la Tabla 4.1. La cantidad de entidades a detectar en cada documento fue determinada por un humano con experiencia en el tema. El total de entidades posibles a identificar es 16.

Tabla 4.1 Resumen de las pruebas de precisión y cobertura

Conferencias	Ent. Doc.	Ent. Ident.	Ent. Indent. Correct.	Precisión	Cobertura
Conferencia 1	12	12	12	1	1
Conferencia 2	16	15	15	1	0.9375
Conferencia 3	8	5	5	1	0.625
Conferencia 4	7	5	5	1	0,714285
Conferencia 5	10	10	10	1	1
Conferencia 6	5	4	4	1	0.8
Conferencia 7	5	4	4	1	0.8
Conferencia 8	12	12	12	1	1

Conferencia 9	16	15	15	1	0.9375
Conferencia 10	5	3	3	1	0.6
Conferencia 11	7	6	6	1	0.857142
Conferencia 12	11	11	11	1	1
Conferencia 13	16	16	16	1	1
Conferencia 14	13	13	13	1	1
Conferencia 15	10	9	9	1	0.9
Promedio				1	0,87809513

Observando los resultados obtenidos se aprecia que la precisión tiene valor uno para todos los casos. Esto se debe a que no se detectan entidades incorrectas, sino que las entidades en el documento son detectadas o no. Los valores bajos de la cobertura se deben principalmente a que las entidades no cumplieron con el patrón preestablecido en la expresión regular. Algunos ejemplos son:

- Para la motivación de la próxima clase se encontraron 2 conferencias con "Motivación del próximo contenido (si es necesario)", lo cual no cumple con el patrón ya que se analiza la oración como subtítulo.

- La temática estaba indicada en un caso como asunto, esto no se tuvo en cuenta para la construcción del patrón de expresión regular.

- El tema en un caso tenía delante espacios y el subtítulo no coincidía con la expresión regular.

Se toman la precisión y cobertura promedio obtenidas en las pruebas anteriores y se calcula la F medida utilizando la Fórmula 4.3, considerando β = 1 para dar la misma importancia a la precisión que a la cobertura.

$$F_\beta = \frac{(1^2 + 1) \times 1 \times 0.87809513}{(1^2 \times 1 + 0.87809513)} = 0.935091$$

El valor de F-medida obtenido es considerablemente alto si se tiene en cuenta que las conferencias son escritas libremente por los profesores, aún guiándose por la estructura metodológica.

Conclusiones.

Una vez transcurridas todas las etapas de la investigación y reflejados los principales resultados en este documento se puede plantear que:

- El empleo de técnicas del procesamiento del lenguaje natural, específicamente la extracción de información, fueron el vehículo para obtener la información de las conferencias existentes procesadas con la herramienta.
- Las expresiones regulares, lenguaje de marcado XML, la obtención de un objeto de aprendizaje SCORM, fueron los principales recursos que apoyaron un resultado satisfactorio.
- Con el software Virtua-Act es posible procesar las conferencias existentes y crear nuevas favoreciendo una guía importante en cuanto a la estructuración metodológica de estos documentos.

Bibliografía.

Alfonso, L. P. (2008) Herramienta para Crear Paquetes de objetos de aprendizaje. *Jornada Científica Estudiantil.* Matanzas.

Álvarez de zayas, c. (1988) Fundamentos teóricos de la dirección del proceso de formación del profesional de perfil amplio.

Álvarez de Zayas, C. (2001) *Hacia una escuela de Excelencia*, Centro de Estudios de Perfeccionamiento de la Educación Superior (CEPES). Universidad de la Habana.

Colectivo de autores (2003) *Preparación pedagógica integral para profesores universitarios,* La Habana, Editorial Felix Varela.

Crea y CUJAE, C. D. I. (2004) Fundamentos del Modelo Universidad para laAutoEducación CUJAE (UAC)

Fernández, S. P. (1996) Determinación del tamaño muestral.

Horruitiner Silva, D. P. T. P. (2006) *La Universidad Cubana: el modelo de formación,* La Habana, Editorial Felix Varela.

Madrigal, V. J. D. (2005) Tratamiento de texto con Perl. Expresiones Regulares.

Mes, M. D. E. S. (2007) Reglamento de Trabajo Docente y Metodológico Resolusión No. 210/2007.

MSDN Library For Microsoft Visualstudio 2005 Expresiones Regulares.

Orquín, A. F. (2005) Un sistema para la detección de entidades basado en Modelos de Probabilidad de Máxima Entropía. Memoria para optar por la suficiencia investigadora. *Departamento de Lenguajes y Sistemas Informáticos.* España, Universidad de Alicante.

Pedruelo, M. R. (2004) El estándar SCORM para EaD. *Tesina del Máster en Enseñanza y Aprendizaje Abiertos y a Distancia.* Universidad Nacional de Educación a Distancia.

Pressman, R. S. (2002) *Ingeniería de Software. Un enfoque práctico,* España, Mc Graw-Hill Interamericana

Rodríguez, H. (2001) Extracción y Recuperación de información. España.

Santana, L. M. (2008) Virtual-Act. Herramienta para la gener. *Documentación del sistema.* Matanzas, Universidad de Matanzas "Camilo Cienfuegos".

Santana, L. M. & Reyes, R. R. (2008) Factibilidad de Software. Matanzas, UMCC.

Verrier, R. D. R. A. (2005) Las formas del proceso docente educativo. *Curso de Didáctica Universitaria.* Matanzas.

Wikipedia: La Enciclopedia Libre Expresión regular.

Wikipedia: La Enciclopedia Libre Procesamiento de lenguajes naturales.