SISTEMA PARA VERIFICAR LA AUTENTICIDAD DE LOS TRABAJOS ENTREGADOS EN FORMATO DIGITAL PARA OBTENER EL GRADO DE LICENCIATURA EN EL INSTITUTO TECNOLÓGICO DE ACAPULCO

Crisol Angelina Mendiola Piza[1], M.T.I Eloy Cadena Mendoza[2],
M.C. Francisco Javier Gutiérrez Mata[3] y M.T.I Juan Miguel Hernández Bravo[4]

Resumen—En base al historial de trabajos presentados por los alumnos del Instituto Tecnológico de Acapulco a partir del año 2007, podemos observar casos en los que se han detectado duplicidad de la información en las opciones de titulación de Tesis e Informe de Residencias. Por lo que se pretende diseñar un sistema que lleve a cabo una búsqueda comparando el nuevo trabajo presentado contra los ya existentes y alertar mediante un porcentaje de similitud este resultado.

Palabras clave—Búsqueda por similitud, Algoritmos de comparación, métrica de espacios.

Introducción

El proceso de titulación comienza una vez que el egresado cumpla con la acreditación del 100% de los créditos de su plan de estudios y acreditación de un programa de lengua extranjera. El egresado comienza su proceso de titulación una vez que entrega la documentación necesaria en el Departamento de Servicios escolares y esté le genera un juego de recibido que deberá de entregar a la coordinación de titulación junto con una opción de titulación. Las opciones de titulación pueden ser: a) Tesis Profesional, b) Libros de Texto o Prototipos Didácticos, c) Proyectos de Investigación, d) Diseño o Rediseño de Equipo, Aparato o Maquinaria, Curso Especial de Titulación, e) Examen Global por Áreas de Conocimiento, f) Memoria de Experiencia Profesional, g) Escolaridad por Promedio, h) Escolaridad por Estudios de Postgrado. Estas opciones de titulación varían respecto al plan de estudios que curso el egresado. De estas opciones antes mencionadas solo se realizarán las comparaciones de las Tesis y Memoria de Residencias profesionales, cuyo historial se encuentra disponible a partir del año 2005 en archivos electrónicos en formato PDF, almacenados en un equipo de cómputo en la Tesiteca del Instituto Tecnológico de Acapulco. Se opta trabajar con estas dos opciones de titulación ya que en un análisis realizado sobre la cantidad de alumnos titulados con respecto a las opciones que maneja la institución, se observa que estas dos tienen el mayor índice.

Planteamiento del problema

Se realizó un análisis de los egresados con sus respectivas formas de titulación del año 2011 al 2017 y se observó que la opción de titulación Informe de Residencias Profesionales mostró el mayor número de alumnos titulados con 703 mientras que la opción de Tesis Profesional está en segundo lugar 569. El estudiante en su último semestre cursa las Residencias Profesionales en una determinada empresa que tenga previamente un convenio establecido con la institución para que los alumnos puedan realizar sus Residencias trabajando con un proyecto existente en la empresa. En muchos casos las empresas no cambian los proyectos y los estudiantes trabajan con un mismo proyecto semestre tras semestre. Este inconveniente trae como consecuencia que los alumnos desde el momento en que presentan su reporte final de Residencias Profesionales presenten trabajos similares con los que habían entregado anteriormente sus compañeros que trabajaron en este mismo proyecto. Toda esta serie de eventos que se están arrastrando desde el momento en que el alumno selecciona un proyecto generará como resultado final que el alumno al momento de titularse mediante la opción X, Informe de Residencias Profesionales entregue un trabajo similar al de sus compañeros que ya se han titulado mediante esta opción. Tesis profesional: Este tipo de trabajo consta de una propuesta concreta, desarrollada mediante una metodología de investigación; dicha propuesta

[1]Ing. Crisol Angelina Mendiola Piza es estudiante de Maestría en Sistemas Computacionales en un programa PNPC en el Instituto Tecnológico de Acapulco, morrigan.yume@gmail.com (**autor corresponsal**).
[2] MTI. Eloy Cadena Mendoza es docente de Maestría del Instituto Tecnológico de Acapulco, eloy_cadena@yahoo.com .
[3] MC. Francisco Javier Gutiérrez Mata es docente y Jefe del centro de cómputo del Instituto Tecnológico de Acapulco, fcomata84@hotmail.com.
[4] MTI. Juan Miguel Hernández Bravo es docente y Jefe del departamento de Sistemas y Computación del Instituto Tecnológico de Acapulco, jhernandez@yahoo.com

deberá ser original, en este último término nos referimos a que la tesis pueda ser sometida o demostrada mediante pruebas y razonamientos apropiados para demostrar su originalidad. Retomando la primera opción de titulación que fue Informe de Residencias Profesionales, existen casos en los que el proyecto que el alumno desarrolló en sus Prácticas Profesionales lo transformen a una tesis ya que está genera un valor curricular mayor al alumno y también al asesor que será su director de tesis.

Pero con esta última observación se recae en el mismo inconveniente de titularse por Informe de Residencias Profesionales ya que los egresados siguen trabajando sobre los mismos proyectos, con la diferencia que ahora cambian de Opción de titulación de Informe de Residencias Profesionales a Tesis Profesional, pero en su desarrollo siguen trabajando sobre un proyecto ya desarrollado teniendo como única variante la opción de titulación.

Descripción del método

Hasta el momento se conoce que el primer archivo obtenido en formato PDF de titulación es del año 2005 de una tesis de la carrera de Arquitectura. A partir de ese año se obtendrán los archivos de las Tesis y Memorias de Residencias de las diferentes licenciaturas del Instituto Tecnológico de Acapulco. Los archivos electrónicos de la Tesiteca que es donde se tienen almacenas las Tesis, están solo almacenados en una computadora en carpetas, y se lleva el registro en una tabla general de la base de datos que ocupa el Centro de Información. Esta tabla llamada tesis cuenta con los siguientes atributos: Autor, EntidadRegistro, Fecha, Folio, IdCharola, Id,Estante, IdTesis, Observación, Paginas, TieneCD, Titulo, TituloEgresado. Como primer paso debemos clasificar la información y almacenarla. Para ello se necesitará crear una base de datos y con ella almacenar estos archivos electrónicos y poder manipularlos. Esta base de datos nos facilitará la clasificación y el acceso a la información, sobre todo tendríamos una información más completa y específica de dichas Tesis y Memorias de Residencias, deberá contener una información desde el plan de estudios del alumno hasta los asesores de titulación.

El siguiente paso es diseñar el sistema que manipulará el usuario final, en este caso es el encargado la coordinación de titulación de la División de Estudios Profesionales. Este sistema se encargará de hacer la búsqueda y comparación en la base de datos de la información ingresada por el coordinador, arrojándole como resultado un porcentaje de similitud donde se encuentre mayor coincidencia entre las Tesis y Memorias de Residencias. El sistema contará con diversos filtros para realizar una búsqueda de forma granular, ejemplos de los filtros que manejara son: título, antecedentes, planteamiento, objetivos, hipótesis, justificación y marco teórico. En cuanto a las acciones del usuario se contempla que él pueda subir los archivos electrónico de la tesis desde su sesión y que pueda modificar datos en el registro de ellos.

El proceso de la búsqueda consiste en buscar palabras utilizando un lenguaje de consulta. Debido a que será un sistema de búsqueda a gran escala es necesario distribuir los documentos en los nodos, para se utilizaran los dos enfoques clásicos: Particionado por documentos, que almacenan una porción de ellos en cada porción de los índices que se crearán, y con ellos sea más fácil acceder a las consultas mediante los espacios métricos generados en cada uno y particionado por términos, en el que cada nodo mantiene la información de las listas completas en solamente un subconjunto de términos. Para la resolución de esta consulta solo participan los nodos que poseen la información de los términos involucrados.

Este proceso de comparación utilizará un algoritmo de búsqueda secuencial de texto llamado algoritmo de fuerza bruta, el cual recorrerá el texto carácter a carácter, buscando coincidencias con los caracteres de las palabras buscada en cada uno de los párrafos. Se sitúa el patrón en la primera posición y se compara carácter por carácter hasta encontrar un valor o llegar al final del patrón, se pasa a continuación a la siguiente posición y se repite este proceso. Termina hasta alcanzar el final del texto, no existe un preprocesamiento del patrón.

En la figura 1 se muestra el ejemplo de cómo funciona el análisis de este algoritmo al comparar la cadena inicial contra un patrón determinado. (Busqueda secuencial de texto)

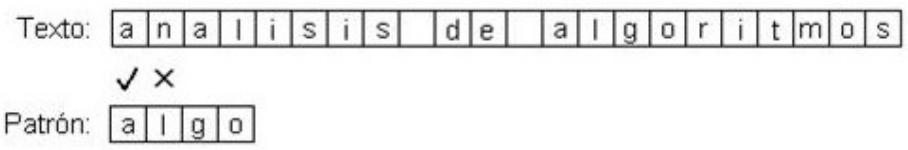

Figura 1. Alineación del patrón con la primera posición del texto (Busqueda secuencial de texto)

Se comparan los caracteres uno a uno hasta que se acabe el patrón, si se detiene la búsqueda por una discrepancia, se desliza el patrón a una posición hacia la derecha y se intenta calzar el patrón nuevamente. Esto se puede observar en la figura 2. (Busqueda secuencial de texto)

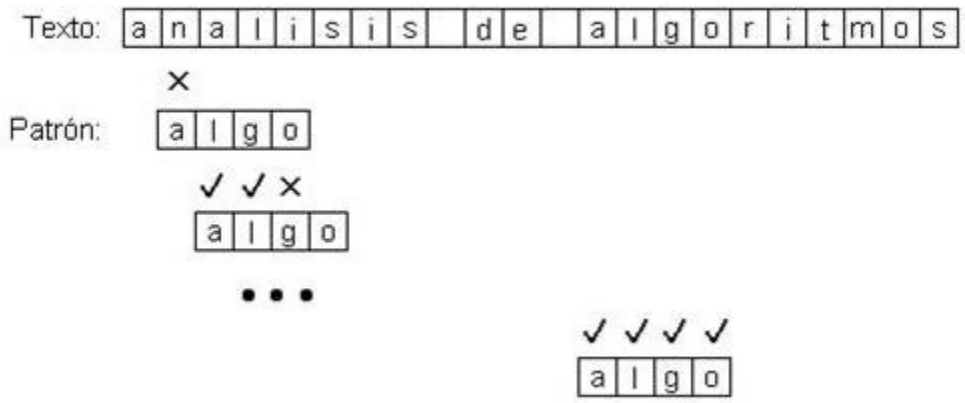

Figura 2. Se intenta seguir buscando el patrón nuevamente (Busqueda secuencial de texto)

Marco teórico

Este proyecto contemplará en su marco teórico los siguientes conceptos con los que estaremos trabajando en el desarrollo del proyecto, debido a que seleccionaremos un lenguaje de programación y gestor de base de datos que nos brinde las características óptimas para lograr un buen desarrollo, las herramientas mencionadas a continuación son de licencia libre.

Visual Studio Community 2017

Esta nueva versión cuenta con un completo IDE extensible y gratuito con todas las características para crear aplicaciones modernas para Windows, Android e iOS, aplicaciones web y servicios en la nube. (Visual Studio, 2018)

Este nuevo instalador diseñado desde cero brinda las siguientes ventajas:

- Reducir al mínimo el consumo de memoria de Visual Studio.
- Instalar más rápidamente con menos impacto en el sistema y desinstalar de una forma más limpia.
- Facilitar la selección y la instalación únicamente de las características que se necesitan.

PostgreSQL

Es un potente sistema de base de datos relacional de objetos abierto que utiliza y amplía el lenguaje SQL combinado con muchas características que almacenan y escalan de forma segura las cargas de trabajo de datos más complicadas. Viene con características destinadas a ayudar a los desarrolladores a crear aplicaciones, administradores para proteger la integridad de los datos y crear entornos tolerantes a fallas, y ayudarlo a administrar sus datos sin importar cuán grande o pequeño sea el conjunto de datos. Además de ser de código abierto y gratuito.

Este gestor intenta cumplir con el estándar SQL donde dicha conformidad no contradice las características tradicionales o podría llevar a decisiones arquitectónicas deficientes. Muchas de las funciones requeridas por el estándar SQL son compatibles, aunque a veces con una sintaxis o función ligeramente diferente. Se pueden esperar más movimientos hacia la conformidad a lo largo del tiempo. La versión 10 es con la que se trabajará en el transcurso del proyecto, cumple con al menos 160 de las 179 características obligatorias para SQL: conformidad con el estándar 2011. (The PostgreSQL Global Development Group, 1996)

Archivos electrónicos

Los archivos digitales se emplean indistintamente para indicar la reunión de documentos creados mediante medios informáticos, sea tanto un conjunto de ficheros generados por una aplicación como una colección de documentos textuales e icónicos digitalizados, en muchos casos accesibles a través de internet. El formato de documento portátil (PDF) se utiliza para presentar e intercambiar documentos de forma fiable, independiente del

software, el hardware o el sistema operativo. No requiere de mucho espacio de alojamiento, ya que generalmente los archivos en formato PDF, son divididos en secciones que no rebasan 1*megabyte*, de hecho, el peso promedio de un documento es de 200 KB a 400 KB. Esto varía dependiendo del número de páginas que tenga nuestro documento y del tipo de original. Posee las características de todo documento electrónico como: la posibilidad de incluir hipervínculos, botones, animaciones, formularios, videos o gráficos. Se menciona solo este formato de archivo electrónico debido a que es con el que se trabajará para este proyecto desde un inicio. (Navarro, 2018)

Objetivo general

Detectar similitudes en los escritos de los trabajos de titulación a nivel licenciatura del Instituto Tecnológico de Acapulco.

Objetivos específicos

a) Desarrollar un sistema que realice la comparación de trabajos electrónicos en formato PDF.

b) Alertar mediante un porcentaje de similitud a los interesados.

c) Conocer las características de las versiones Acrobat 7.0, Acrobat 8.0, Acrobat 9.0, Acrobat 9.1 y Acrobat X (10) de los archivos electrónicos en formato PDF.

Hipótesis

Con la intensión de disminuir la copia de reporte de los trabajos derivados de los proyectos afines se desarrollará un sistema que realice la comparación del nuevo trabajo entregado contra los ya existentes. Dicho sistema comenzara a analizar el texto desde el título hasta el marco teórico; limitando un porcentaje de similitud tolerable hasta un valor determinado, en cuanto se detecte que el porcentaje es mayor a esto se sigue realizando la comparación del demás contenido del archivo contra el existente para poder obtener un porcentaje final de similitud de las partes que conforman dicho documento.

Alcance

Este proyecto realizará la comparación de los trabajos de titulación de los egresados de licenciatura del Instituto Tecnológico de Acapulco, obtenidos del año 2005 al 2017 en formato electrónico PDF. Se desarrollará un sistema informático para buscar la similitud de los trabajos que presentan los egresados, para obtener el grado académico de licenciatura. Como resultado esta búsqueda se obtendrá un porcentaje de similitud, detectada mediante la comparación de este nuevo trabajo con los existentes del periodo ya mencionado.

Limitaciones

- Se iniciará con el año 2005 a realizar la comparación de trabajos nuevos contra esta fecha de inicio.
- Se trabajará con archivos electrónicos en formato PDF de las siguientes versiones: Acrobat 7.0, Acrobat 8.0, Acrobat 9.0, Acrobat 9.1 y Acrobat X (10).
- Estos archivos en formato electrónico PDF provienen de las siguientes opciones de titulación de licenciatura, que generan un archivo electrónico; a) Memoria de Residencia Profesional y b) Tesis.

Referencias

Abraham Silberschatz, H. F. (2002). *Fundamentos de base de datos.* España: Mc Graw Hill.

Google, S. (s.f.). *Busqueda secuencial de texto*. Recuperado el 27 de Septiembre de 2018, de Sites Google: https://sites.google.com/site/busquedasecuencialdetexto/

Navarro, M. Á. (2018). Los archivos de documentos electrónicos. *Revista internacional de Información y comunicación*, 5.

Seco, J. A. (2002). *El lenguaje de programación C#*.

Sommerville, I. (2011). *Ingeniería de Software.* México : Pearsón Educación

Texto, B. s. (s.f.). Recuperado el 27 de Septiembre de 2018, de Sites Google: https://sites.google.com/site/busquedasecuencialdetexto/algoritmo-fuerza-bruta

The PostgreSQL Global Development Group. (1996). Recuperado el 20 de Septiembre de 2018, de https://www.postgresql.org

Visual Studio . (Julio de 2018). Recuperado el 22 de Agosto de 2018, de https://visualstudio.microsoft.com/es/