ISSN 2410-3993

Volumen 2, Número 5 – Octubre – Diciembre -2015

# Revista de Tecnología e Innovación

ECORFAN®

**Bases de datos**

**Google scholar**.

**ECORFAN®**

Revista de Tecnología e Innovación, Volumen 2, Número 5, de Octubre a Diciembre 2015, es una revista editada trimestralmente por ECORFAN-Bolivia. Santa Lucía N-21, Barrio Libertadores, Cd. Sucre. Chuquisaca, Bolivia. WEB: www.ecorfan.org,revista@ecorfan.org. Editora en Jefe: RAMOS-ESCAMILLA-María, Co-Editor: SERRUDO-GONZÁLEZ, Javier. ISSN-2410-3993.Responsables de la última actualización de este número de la Unidad de Informática ECORFAN. ESCAMILLA-BOUCHÁN, Imelda, LUNA-SOTO, Vladimir, actualizado al 31 de Diciembre 2015.

# Presentación

ECORFAN, es una revista de investigación que pública artículos en las áreas de: Revista de Tecnología e Innovación

En Pro de la Investigación, Docencia, y Formación de los recursos humanos comprometidos con la Ciencia. El contenido de los artículos y opiniones que aparecen en cada número son de los autores y no necesariamente la opinión de la Editora en Jefe.

Como primer artículo está *Sistema Recomendador Orientado a la Educación Basado en la Distancia entre Likes de Facebook y Conceptos* por MORALES, Alejandro, LÓPEZ-CHAU, Asdrúbal y REYES, Luis, como siguiente artículo está *El algoritmo de agrupamiento K-Modas: Un caso de estudio* por RENDÓN, Eréndira, ZEPEDA, Ricardo, BARRUETA, Elizabeth y ITZEL-MARÍA, Abundez con adscripción Departamento de Sistema y Computación, Instituto Tecnológico de Toluca, como siguiente artículo está *Una versión modificada del algoritmo de agrupamiento Isodata* por RENDON, Eréndira, MENDOZA, Marcos, CISNIEGA, Roció y CARBAJAL, Guillermo, como siguiente artículo está *Desarrollo de un software para la simulación y control de un robot industrial* por LAZARO-ARVIZU, Y, MORALES-CAPORAL, R, ORDOÑEZ-FLORES, R, QUINTERO-FLORES, P y LEAL-LÓPEZ, M, como siguiente artículo está *Adaptación del MMPI Mediante un Sistema Experto en Base a Probabilidades para el Diagnóstico de Desviaciones Psicopáticas en el Instituto Tecnológico de Pachuca* por RAMÍREZ-MEJIA J., MAGGI-NATALE C., ARRIETA-ZUÑIGA J., HERNANDEZ-RAMÍREZ A. y GONZÁLEZ-MARRON D. con adscripción Instituto Tecnológico de Pachuca, como siguiente artículo está *Metodologías actuales de desarrollo de software* por RIVAS, Carlos Ignacio, CORONA, Verónica Paola, GUTIÉRREZ, José Fructuoso y HERNÁNDEZ, Lizeth, como siguiente artículo está *Publicación en Internet del inventario de infraestructura física del I.T.P mediante Bases de Datos Geoespaciales y Sistema de Información Geográfica* por HERNÁNDEZ, Javier, ARRIAGA, Sergio y RERGIS, Raúl, como siguiente artículo está *Sistema de monitoreo del LOBOBUS* por REYES, Cecilia`, BARRETO, Aldrin`` y BAUTISTA, Verónica Edith`` con adscripción `Instituto Tecnológico de Pachuca, ``Benemérita Universidad Autónoma de Puebla.

## Contenido

Artículo **Revista Tecnología e Innovación**
Diciembre 2015 Vol.2 No.5 921-928

# Sistema Recomendador Orientado a la Educación Basado en la Distancia entre Likes de Facebook y Conceptos

MORALES, Alejandro*†, LÓPEZ-CHAU, Asdrúbal y REYES, Luis

Recibido 5 de Julio, 2015; Aceptado 24 de Noviembre, 2015

## Resumen

Hoy en dia las redes sociales otorgan un área de oportunidad para el análisis de la información que sus usuarios proporcionan en ellas. Facebook es la red social más importante debido al gran número de usuarios con los que cuenta. Este artículo presenta el desarrollo de un método para la identificación de relaciones entre los Likes de usuarios de Facebook, y una serie de conceptos. Para demostrar la utilidad del método propuesto, se aplicó éste a cada uno los distintos programas de posgrado ofertados dentro del Instituto Tecnológico de Orizaba, en Veracruz, México. Los resultados con usuarios reales demuestran la efectividad de la propuesta, y brindan un escenario prometedor para poder aplicarse a otros casos. La aplicación desarrollada actualmente se encuentra en etapa de revisión por parte Facebook para su liberación y uso público.

**Facebook, Graph API, Orientación Vocacional, Recomendación Automática.**

## Abstract

Today's social networks provide an opportunity area for the analysis of information that users provide in them. Facebook is the largest social networks due to the large number of users are there. This paper presents the development of a method for identifying relationships between the extracted Likes of Facebook users, and a series of concepts. To demonstrate the utility of the proposed method, it is applied to each individual graduate programs offered within the Technological Institute of Orizaba, Veracruz, Mexico. The results with real-world users demonstrate the effectiveness of the proposal and provide a promising scenario to be applied to other cases. The developed application is currently being reviewed by Facebook for their release and public use.

**Automatic Recommendation, Facebook, Graph API, Vocational Orientation.**

**Citación:** MORALES, Alejandro, LÓPEZ-CHAU, Asdrúbal y REYES, Luis. Sistema Recomendador Orientado a la Educación Basado en la Distancia entre Likes de Facebook y Conceptos. Revista de Tecnología e Innovación 2015, 2-5: 921-928

* Correspondencia al Autor (Correo Electrónico: ing.alejandromd@gmail.com)
† Investigador contribuyendo como primer autor.

## Introducción

Las redes sociales han tenido un gran impacto a nivel mundial durante los años recientes. Los tipos de datos que estas redes almacenan pueden ser utilizados para mediante un análisis descubrir relaciones, comportamientos o tendencias en ellos. Hoy en día la red social más popular es Facebook, con alrededor de 1,350 millones de usuarios según Lucia Sanjaime (2012). Esta red social permite, a través del uso de su biblioteca Graph API, acceder a datos de los usuarios atraves de una aplicación, la cual puede extraer sus datos, previa autorización, a cambio de obtener algún servicio.

De acuerdo a las políticas de la plataforma se pueden extraer todos los datos que el usuario autorice. Algunos de los datos que se pueden extraer son los siguientes: dirección de correo electrónico, edad, Likes ("me gusta"), ciudad de origen, ente otros. Los Likes representan las páginas de Facebook a las cuales un usuario ha indicado que son de su agrado o interés.

Dada la gran actividad que los usuarios de Facebook usualmente tienen, este artículo propone un método para identificar posibles relaciones los Likes de usuarios dentro de Facebook, y los programas de posgrado ofertados dentro del Instituto Tecnológico de Orizaba (ITO), que son Maestría en Sistemas Computacionales, Maestría en Ingeniería Administrativa, Maestría en Ingeniería Electrónica, Maestría en Ingeniera Industrial y Maestría en Ingeniería Química, para proporcionar al usuario una recomendación del programa de posgrado más adecuado de acuerdo a sus Likes. Esto se realiza mediante un análisis de la proximidad entre los términos obtenidos de los datos de usuario (Likes) y los conceptos asignados a los programas educativos de posgrado ofertados en el ITO.

El método propuesto se implementó en una aplicación Web, y actualmente se encuentra en etapa de validación por parte de la red social, para una vez autorizada liberarse al público. De acuerdo con los resultados preliminares, el método propuesto realiza recomendaciones acertadas, orientando asi a los usuarios sobre su afinidad con cada programa educativo del ITO.

El resto del artículo está dividido en 4 secciones. En la Sección 2 "Preliminares" se presentan los principales métodos para calcular similitud entre cadenas, el cual es un concepto clave en la propuesta realizada en este trabajo. En la Sección 3 "Sistema Propuesto" se presenta la arquitectura para el sistema propuesto, la cual, se encarga de realizar un filtro mediante la proximidad de los Likes de usuarios extraídos y los conceptos de cada especialidad. En la sección 4 "Resultados" se muestran los resultados obtenidos con las pruebas realizadas a la aplicación con los usuarios de prueba. Por último, en la sección 5 se presentan las conclusiones.

## Preliminares

El contenido que se encuentra en Facebook es de diversos tipos, este puede ser publicaciones o páginas con información, las cuales pueden contener imágenes, video, audio, y/o texto. La forma más eficiente computacionalmente para buscar conocimiento de la actividad de los usuarios dentro de la red social, consiste en analizar secuencias de caracteres.

Datos tales como el nombre de usuario, correo electrónico y los Likes, pueden ser obtenidos como texto usando la Graph API de Facebook. Sin embargo, estos datos por sí mismos, no brindan un conocimiento sobre la relación que existe entre la actividad de un usuario con otros aspectos.

ISSN-2410-3993

Para descubrir posibles relaciones entre usuarios y programas educativos del ITO, en este trabajo se analizan los datos usando el concepto de proximidad o similitud entre cadenas para comparar Likes extraídos de Facebook con distinos conceptos asociados a los programas de posgrado ofertados dentro del ITO. A continuación, se describen los métodos más importantes para calcular la similitud entre cadenas.

**Distancia entre cadenas**

Hoy en día se tienen distintos métodos para el cálculo de similitud entre cadenas, esta similitud puede considerarse como una distancia tomando en cuenta las tres propiedades fundamentales de este concepto. Algunos de los métodos más importantes para calcular la distancia entre cadenas son Levenshtein o edición, Brecha Afín, Smith-Waterman, Jaro, y q-Grams según Iván Amón, Francisco Moreno & Jaime Echeverri (2012).

La distancia de Levenshtein entre dos cadenas de texto A y B, se basa en el conjunto mínimo de operaciones de edición necesarias para transformar A en B (o viceversa). Las operaciones de edición permitidas son eliminación, inserción y sustitución de un carácter y cada una tiene un costo unitario siendo referido como distancia de Levenshtein (1966). Un problema con la distancia de Levenshtein, es que tiende a fallar cuando intentan identificar cadenas equivalentes que han sido demasiado "truncadas", ya sea mediante el uso de abreviaturas o la omisión de tokens.

La distancia de brecha afín ofrece una solución a lo anterior, al penalizar la inserción/eliminación de k caracteres consecutivos (brecha) con bajo costo.

Para ello, usa una función afín $R(k) = g + h \bullet (k-1)$, donde g es el costo de iniciar una brecha, h el costo de extenderla un carácter y $h << g$. Gotoh (1982) describe un modelo para entrenar automáticamente esta función de similitud a partir de un conjunto de datos.

La similitud de Smith-Waterman entre dos cadenas A y B según Smith (1982) es la máxima similitud entre una pareja (A', B'), sobre todas las posibles, tal que A' es sub cadena de A y B' es sub cadena de B. El modelo original define las mismas operaciones de la distancia de edición y, además, permite omitir cualquier número de caracteres al principio o al final de ambas cadenas.

Jaro (1976) desarrolló una función de similitud que define la trasposición de dos caracteres como la única operación de edición permitida. Los caracteres no necesitan ser adyacentes y pueden estar alejados cierta distancia d que depende de la longitud de ambas cadenas.

Un q-gram, también llamado n-gram, es una subcadena de longitud q. Según Yancey (2006) el principio tras esta función de similitud es que, cuando dos cadenas son muy similares tienen muchos q-grams en común. Es común usar uni-grams (q = 1), bi-grams o di-grams (q = 2) y tri-grams (q = 3). Es posible agregar q – 1 ocurrencias de un carácter especial (no definido en el alfabeto Σ original) al principio y final de ambas cadenas. Esto llevará a un puntaje de similitud mayor entre cadenas que compartan algún prefijo o sufijo, aunque presenten diferencias en el medio.

**Sistema Propuesto**

La Graph API de Facebook, permite extraer, datos previa autorización de los usuarios. Como se mencionó anteriormente, los Likes, son las páginas de Facebook que son de interés para los usuarios.

ISSN-2410-3993

Para cada Like por parte de los usuarios, se puede extraer su el nombre de la página asi como algunos otros datos. Entre los datos adicionales que se pueden extraer se encuentra la "Categoría", dato que brinda una idea acerca del contenido de cada página.

La arquitectura propuesta en este trabajo para la identificación de posibles relaciones entre usuarios y programas educativos del ITO, se presenta en la figura 1.

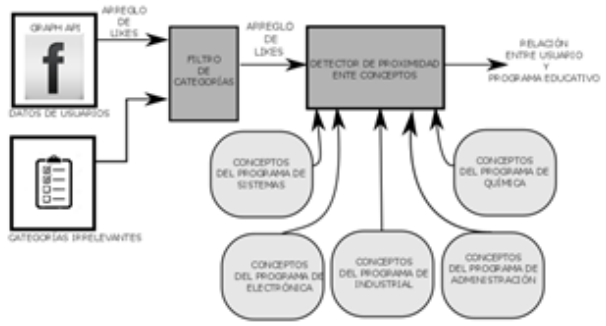



**Figura 1** Arquitectura del sistema propuesto

El bloque de Filtro de Categorías que se observa en la figura anterior, se encarga de eliminar el ruido existente en los Likes del usuario. Es considerado ruido aquellos Likes irrelevantes para las relaciones que se intentan detectar, es decir, aquellas categorías de páginas que no aportan nada en el descubrimiento de conocimiento. La hipótesis que se planeta en este punto es que existen algunos tipos de páginas que no permiten diferenciar las preferencias de los usuarios hacia algún programa educativo. La Tabla 1 contiene las categorías que son consideradas irrelevantes, y que son desechadas en el primer bloque.

| Categorías Eliminadas | Categorías Eliminadas |
|---|---|
| Movie Theater | Artist |
| Musician/Band | Clothing |
| Tours/Sightseeing | Sports/Recreation/Activities |
| Public Figure | Sports League |
| Musician/Band | Public Figure |
| on-Governmental Organization (NG | Actor/Director |
| Arts/Entertainment/Nightlife | TV Channel |
| Concert Venue | Movie |
| Food/Grocery | Recreation/Sports Website |
| Sports Team | Bar |
| Athlete | Real Estate |
| TV Channel | Pet |
| Sports League | Arts/Humanities Website |
| Consulting/Business Services | Non-Profit Organization |
| Attractions | |

**Tabla 1** Categorías irrelevantes

El bloque recibe como entrada un arreglo de Likes de usuario, el cual, es extraído con la API Graph de Facebook. La salida del bloque, es otro arreglo de Likes que no contiene las páginas que estén categorizadas dentro de la lista presentada de acuerdo con la hipótesis mencionada.

El segundo bloque, denominado Detector de Proximidad entre Conceptos, es usado para determinar la relación entre los programas educativos y los Likes de los usuarios. Esto se realiza determinando la distancia entre el nombre o títulos de las páginas a las cuales el usuario les ha dado Like, y una lista de conceptos de cada programa educativo. La función de distancia empleada es la proporcionada por PHP Simlar Text (PHP, 2015), que es una función que toma como base la distancia de Levenshtein para calcular la similitud entre dos cadenas.

Con el objetivo de determinar un nivel de ajuste para la similitud entre los conceptos comparados, se emplea un umbral de comparación como parámetro de usuario. Como se observa en el algoritmo 1 el valor del umbral para este ejemplo debe ser mayor a 50, esto indica que se requiere minimo del 51% de similitud ente dos cadenas, para considerar que existe una similtud entre ellas.

ISSN-2410-3993

```
$contadorIndustrial = 0;
    foreach ($arrayIndustrial as $valor) {
        foreach ($arrayLikes as $valor1) {
            similar_text($valor,$valor1,$porcentaje);
                if((int)$umbral>50){
                $contadorIndustrial++;
                }
        }
    }
```

**Figura 2** Filtrado de Likes para la especialidad de industrial

Una vez realizado el filtro de los Likes de los usuarios, se emplea el dato User_Location, que es la ciudad actual en la que se encuentra el usuario según sus datos de Facebook, extraído igualmente de Facebook, para realizar una selección de usuarios que se encuentren cercanos a la región de la institución en la que se ofertan los programas.

## Resultados

A continuación se presentan los resultados obtenidos al aplicar el método propuesto a los datos de cuatro distintos usuarios de Facebook. Es importante mencionar que los cuatro usuarios cuyos datos fueron usados en los experimentos pertenecen al área de sistemas, debido a que la aplicación todavía no ha sido validada por Facebook, y por esta razón no se encuentra disponible para el público en general. Para poder usar una aplicación que todavía no ha sido aprobada, fue necesario registrar a los usuarios como Tester dentro de la aplicación en la plataforma de Facebook.

Para la primera prueba (usuario 1), se utilizó un perfil de un usuario con profesión de ingeniero en sistemas computacionales. La lista completa de los Likes extraídos para este usuario contiene un total de 68 páginas. Esta lista es pasada a través del primer bloque denominado Filtro de Categorías, el cual reduce la lista a la sublista mostrada en la Tabla 2.

| User | Page | Category |
|---|---|---|
| User 1 | EnicPanda | Computers |
| User 1 | Geek Site | Community |
| User 1 | Universidad Tecnológica de la Riviera Maya | School |
| User 1 | Chistes de los políticos Locos | Community |
| User 1 | P&G México - Reclutamiento | Company |
| User 1 | EXITO - Instituto Tecnológico de Orizaba | Education |
| User 1 | GE Careers México | Industrials |
| User 1 | Marketing_BetaV1.0 Community | Education |
| User 1 | Facebook Developers | Product/Service |
| User 1 | Apalabrados | App Page |
| User 1 | Grammarly | Education Website |
| User 1 | Vacantes Ayp Reclutamiento | Small Business |
| User 1 | Muy Interesante | Science Website |
| User 1 | Programación con Java | Community |
| User 1 | Programador de Palo | Community |
| User 1 | Mochileros.com.mx | Website |
| User 1 | Jardín del Edén | Farming/Agriculture |
| User 1 | Córdoba Coworking | Organization |
| User 1 | Museo Mendoza | Community Organization |
| User 1 | Unilever Careers | Company |
| User 1 | Umiro México | Company |
| User 1 | The Institute Orizaba | Education |
| User 1 | INGLÉS UNIVERSAL® | Organization |
| User 1 | The Film Zone | Interest |
| User 1 | Blockbuster Mexico | Company |
| User 1 | Memes ITO Orizaba | Community |
| User 1 | Lupita | Company |
| User 1 | Instituto Tecnológico de Orizaba | University |

**Tabla 2** Likes después del primer filtro

Los Likes filtrados previamente, se introducen al bloque Detector de Proximidad de Conceptos, que produce como paso intermedio la Tabla 3. El umbral usado en todos los experimentos fue establecido en mayor a 50.

| SISTEMAS | | |
|---|---|---|
| **Concepto** | **Nombre página** | **Porcentaje de similitud** |
| Programación | Programación con Java | 74% |
| Programa | Programación con Java | 53% |
| Programa | Programador de Palo | 59% |
| Geek | Geek Site | 61% |
| Programador | Programación con Java | 54% |
| Programador | Programador de Palo | 73% |
| **ELECTRÓNICA** | | |
| **Concepto** | **Nombre página** | **Porcentaje de similitud** |
| Alternador | Apalabrados | 57% |
| Digitales | Cerebro Digital | 58% |
| Programación | Programación con Java | 74% |

**Tabla 3** Resultado de aplicar el detector de proximidad

Para la presentación del resultado que se muestra al usuario final, se realiza un conteo de los Likes más próximos a los conceptos de cada programa educativo. La Figura 2 muestra la forma en que se da el resultado para el usuario 1. Es importante mencionar que el usuario 1 es uno de los autores de este trabajo, y su formación profesional es del área de sistemas.

ISSN-2410-3993


**Figura 3** Resultado para el usuario 1

Una vez explicada la forma en que funciona el algoritmo, para los siguientes usuarios se presentan directamente los resultados obtenidos mediante la gráfica generada por la aplicación.

Para la segunda prueba (usuario 2), se utilizó el perfil de un usuario con formación profesional también orientada a la computación.

En la figura 3 se muestra el resultado de la aplicación del método propuesto a los datos del usuario. Como se puede observar, la precisión no fue del 100% adecuada debido a que según el resultado el usuario tiene relación con la electrónica y con la química. En lo que respecta a la electrónica, de acuerdo a las características de la carrera, es similar con respecto a la computación pero en el caso de química, la relación no es demasiada como el resultado lo indica.

Una vez explicada la forma en que funciona el algoritmo, para los siguientes usuarios se presentan directamente los resultados obtenidos mediante la gráfica generada por la aplicación.
Para la segunda prueba (usuario 2), se utilizó el perfil de un usuario con formación profesional también orientada a la computación. En la figura 3 se muestra el resultado de la aplicación del método propuesto a los datos del usuario.

Como se puede observar, la precisión no fue del 100% adecuada debido a que según el resultado el usuario tiene relación con la electrónica y con la química. En lo que respecta a la electrónica, de acuerdo a las características de la carrera, es similar con respecto a la computación pero en el caso de química, la relación no es demasiada como el resultado lo indica.

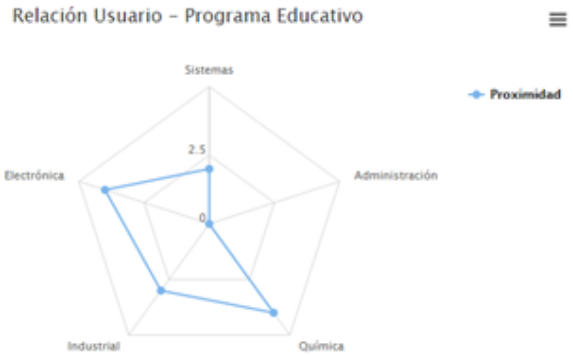



**Figura 4** Resultado para el usuario 2

Para la tercera prueba (usuario 3) se utilizó el perfil de un usuario con profesión de ingeniero en sistemas computacionales. En la figura 4 se muestra el resultado obtenido.

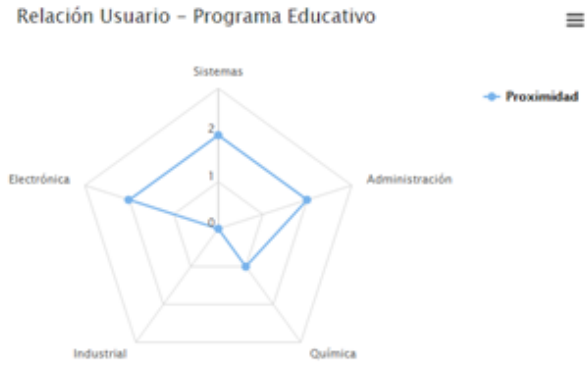



**Figura 5** Resultado para el usuario 3

Con base en la información de este usuario, su perfil es de ingeniería en computación, por lo que el resultado obtenido se apega en gran medida a su profesión, pues como se observa la tendencia es sistemas, electrónica, aunque de acuerdo a sus Likes dentro de la red social, se encontraron conceptos relacionados con la carrera de administración.

ISSN-2410-3993

Para la cuarta prueba (usuario 4), se utilizó el perfil de un usuario con formación profesional también orientada a la computación.

En la figura 5 se muestra el resultado. En este caso el resultado dado por el sistema se consideró como no exitoso, debido a que el sistema arroja un resultado del programa de química. Analizando la causa de esto, se encontró que el funcionamiento del método de similitud entre cadenas con el umbral de similitud utilizado, determinaba la cercanía entre algunos conceptos que no estaban correctos en su totalidad. Por esta razón el sistema en ocasiones tiende a fallar en sus recomendaciones.

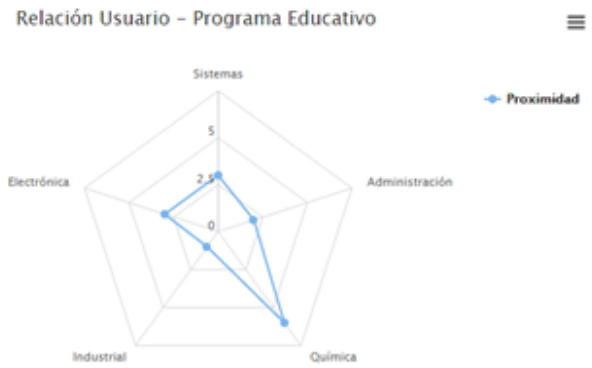



**Figura 5** Resultado para el usuario 4

## Conclusiones

Facebook, con más de 1,350 millones de usuarios, es la red social más importante a nivel mundial. El análisis de datos dentro de las redes sociales se ha incrementado recientemente, debido que en estos datos pueden encontrarse relaciones interesantes para ser empleadas por las organizaciones para distinas actividades.

Este trabajo propuso un proceso automatico para encontrar relaciones entre Likes de usuarios de Facebook y su posible interés en estudiar un programa de posgrado en el Instituto Tecnológico de Orizaba en Veracruz, México.

El sistema utiliza el concepto de similitud entre cadenas, para detectar páginas a las que los usuarios han manifestado ser de su interés, y su proximidad con una lista de conceptos creada para cada programa educativo. Cabe mencionar que la arquitectura del método propuesto es fácilmente adaptable para aplicarse a diversas situaciones en donde se requiere realizar alguna recomendación al usuario, basándose en su actividad en Facebook, primordialmente en sus Likes.

El método propuesto fue probado con cuatro usuarios, y como resultado se encontró que la propuesta ofrece resultados satisfactorios, aunque en algunos casos la recomendación que ofrece tiene cierta divergencia con respecto a lo esperado. Actualmente, se está trabando en una mejora de la versión presentada del método, asi como en la autorización por parte de Facebook para que la aplicación pueda ser liberada y utilizada por todo el público.

## Referencias

Amón, Iván, Moreno, Francisco, & Echeverri, Jaime. (2012). Algoritmo fonético para detección de cadenas de texto duplicadas en el idioma español. Revista Ingenierías Universidad de Medellín, 11(20), 127-138. Retrieved August 19, 2015, from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1692-33242012000100011&lng=en&tlng=es.

Gotoh, O. 1982. An Improved Algorithm for Matching Biological Sequences, Journal of Molecular Biology, 162, 3, 705-708.

Jaro, M. A. 1976. Unimatch: A Record Linkage System User's Manual, technical report, Washington, D. C.: US Bureau of the Census.

Levenshtein, V. I. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Soviet Physics Doklady, 10, 8, 707-710.

Lucia Sanjaime Calvet (2012). Redes Sociales y Marketing. Escola Tècnica Superior d'Enginyeria Informàtica Universitat Politècnica de València. (Doctoral dissertation).
PHP. (2015). PHP Documentation. 19/08/2015, de PHP Sitio web: http://php.net/docs.php

Smith, T. F. y Waterman, M. S. 1981. Identification of Common Molecular Subsequences, Journal of Molecular Biology, 147, 1, 195-197.

Yancey, W. E. 2006. Evaluating String Comparator Performance for Record Linkage. En Proceedings of the Fifth Australasian Conference on Data mining and Analytics, 23-21.

ISSN-2410-3993

Artículo **Revista Tecnología e Innovación**
Diciembre 2015 Vol.2 No.5 929-941

# El algoritmo de agrupamiento K-Modas: Un caso de estudio

RENDÓN, Eréndira†, ZEPEDA, Ricardo, BARRUETA, Elizabeth y ITZEL-MARÍA, Abundez

*Departamento de Sistema y Computación, Instituto Tecnológico de Toluca*

Recibido 5 de Julio, 2015; Aceptado 24 de Noviembre, 2015

## Resumen

En este trabajo se desarrolló un software que utiliza el algoritmo K- modas para realizar agrupamiento con bases de datos descritas en datos categóricos, para probar el software se presenta un caso estudio, donde se encontrarán las caracterizas de los estudiantes que terminaron su carrera con un título. Las pruebas se realizaron con una base de datos del Instituto Tecnológico de Toluca de la carrera de Ingeniería en Sistemas Computacionales.

**Algoritmos de agrupamiento, algoritmo K-Modas, datos categóricos.**

## Abstract

In this paper we developed a software that uses K-modas algorithm in order to cluster with databases described as categorical data. To test the software we present a study case, where the K-modas algorithm was used in order to find the students' features that finished their carrier with a degree. We worked with a data base of Instituto Tecnológico de Toluca, from Computational System Engineering carrier.

**Clustering Algorithm, K-Modas algorithm, categorical data.**

**Citación:** RENDÓN, Eréndira, ZEPEDA, Ricardo, BARRUETA, Elizabeth y ITZEL-MARÍA, Abundez. El algoritmo de agrupamiento K-Modas: Un caso de estudio. Revista de Tecnología e Innovación 2015, 2-5: 929-941

† Investigador contribuyendo como primer autor.

## Introducción

El descubrimiento del conocimiento en bases de datos (KDD) es el proceso global de búsqueda de nuevo conocimiento a partir de los datos almacenados en las bases de datos. Este proceso incluye: filtrado, procesamiento, transformación, técnicas de minería de datos, interpretación y validación del conocimiento extraído (Fayyad U.M., 1996), ver figura 1.


**Figura 1** Proceso KDD

La minería de datos es un paso importante en el proceso KDD. La minería de datos tiene dos tareas principales: las predictivas y las descriptivas. En las tareas descriptivas existen varias técnicas, tales como el agrupamiento (clustering), sumarización, modelado de dependencias. El agrupamiento es una técnica muy utilizada en las tareas de minería de datos, por esta razón, ha sido ampliamente estudiado debido a la gran variedad de aplicaciones donde se puede trabajar esta técnica. Se puede encontrar en la literatura una gran variedad de algoritmos de agrupamiento (Kaufman L, 1989), los cuales pueden ser utilizados en función del tipo de datos que trabajen, es decir si la base de datos está descrita en datos de tipo numérico o categórico. El algoritmo K-Modas es un algoritmo de agrupamiento (Zhexue, 1998) que trabaja con datos categóricos. En esta investigación, la base de datos que se utilizó está descrita con este tipo de datos, dicha base de datos contiene la descripción de los estudiantes del Tecnológico de Toluca de la carrera de ingeniería en Sistemas Computacionales.

De esta manera en este trabajo se utilizó el algoritmo K-modas para encontrar las características de los buenos estudiantes, es decir aquellos que terminan titulados.
Tratando con estudiantes, existen ciertos factores que influyen con el rendimiento y éxito académico que pertenecen al grupo de datos categóricos. (Tinto, 1992) se postula que los estudiantes ingresan a la universidad con diversas habilidades y patrones de características personales, familiares y académicas, incluidas metas y predisposiciones iniciales para asistir a la universidad. Estas últimas se modifican y reformulan continuamente a través de una serie de interacciones entre el individuo, las estructuras y miembros de los sistemas sociales y académicos de la institución.

Así nuestra investigación se centró en desarrollar un software que utiliza el algoritmo de agrupamiento K-modas para determinar los factores o características que influyen en el éxito o no de un estudiante (obtención del título) en una base de datos de estudiantes de ingeniería en sistemas computacionales del Instituto Tecnológico de Toluca. Es importante resaltar que el software desarrollado puede trabajar con otros tipos de base de datos.

El resto de este trabajo se encuentra organizado de la siguiente manera en la sección 1 se decriben los trabajos relacionados con la solución que se presenta, en la sección 2 se describen algunas definiciones necesarias para un mejor entendimiento del algoritmo del algoritmo K-modas, así como la descripción de éste, en la sección 3 se proporciona la metodología que se utilizó para la programación del software “K-modas7”, en la sección 4 se describen los resultados obtenidos, finalemnete en la última sección se presentan las conclusiones a las que se llegaron.

**Trabajos relacionados**

Dentro del sector educativo se encuentran diversos elementos que permiten identificar el rendimiento y éxito académico de los estudiantes. En la actualidad existe un significativo interés por el estudio de las variables relacionadas con el éxito académico y la manera en que se comportan los resultados que se generan a través de diferentes técnicas y métodos. Existen investigaciones que han sido realizadas por expertos en el tema, aportando conocimiento para mejorar y analizar estas variables o factores, donde establecen que las condiciones académicas, la adaptación a la institución, las estrategias de aprendizaje y la situación socioeconómica son algunos de los elementos decisivos en el éxito escolar. Algunas de las investigaciones que se han realizado al respecto son:

En (Navarro, 2003) se menciona que existen diversas variables que pueden identificarse de la siguiente forma, en relación con los individuos, una de ellas son las características que son susceptibles de modificarse a través del proceso educativo y aquellas que no pueden modificarse, como las características genéticas y las experiencias previas. También establece que siempre que se pretende encontrar el fracaso escolar se apunta hacia los programas de estudio, la falta de recursos de las instituciones y rara vez se piensa en el papel que los padres juegan.

Se realizó una investigación por parte de (Martínez, 2003) acerca del perfil de éxito de un estudiante de posgrado, donde se indica que la obtención del grado a nivel posgrado es baja y repercute tanto en el ámbito social como educativo. Las variables que se relacionan dentro del estudio son el nivel de conocimientos previos, una mayor capacidad intelectual, características psicológicas, hábitos académicos positivos y algunas otras variables, tiene como resultado un mayor éxito académico.

En (Gómez, 2003) se tiene como objetivo investigar las características motivacionales, cognitivas y autorreguladoras, así como las actividades de aprendizaje que llevan durante la carrera de Química en la Universidad Nacional Autónoma de México. En este estudio se observa que los aciertos, razonamientos, estrategias y concepciones alternativas han contribuido a perfeccionar las áreas sobre el proceso de aprendizaje y la identificación del éxito en los estudiantes. Otros autores coinciden en que los factores personales y académicos determinan si un estudiante es exitoso o no al final de su carrera profesional (Acosta, 2004).

En (Belvis, 2009) se desarrolló un estudio que pretende determinar cuáles son los factores que afectan al rendimiento académico de los estudiantes universitarios en España. Se realizó una encuesta a una muestra de estudiantes de siete Facultades de Educación españolas, con lo cual se detectaron los factores que más inciden en el éxito o fracaso del estudiante son: la situación laboral; la dedicación y motivación por los estudios; las becas de estudio; las condiciones de acceso a la titulación y la preparación académica previa, así como el rendimiento académico que se consigue en los primeros semestres de estudio en la universidad. En este estudio se analizan e interpretan los resultados obtenidos y se realizan propuestas para mejorar las intervenciones y los servicios de apoyo para estudiantes.

En (Gatica, 2010) se menciona que “Los estudios universitarios representan demandas, compromisos, metas de mayor dificultad y exigencia. Se ha observado en la Facultad de Medicina un alto índice de reprobación y abandono durante los 2 primeros años de la licenciatura, el cual disminuye de manera importante en el área clínica”. Por tal motivo se propone analizar las variables que intervienen en el rendimiento y éxito académico durante los primeros años de la carrera.

Ya que durante este periodo puede estar definida la continuidad de los estudios universitarios. En este estudio se dividen las variables en factores académicos, personales y socioeconómicos, tomando en cuenta el éxito académico como la acreditación oportuna de las asignaturas, exámenes departamentales y una puntuación determinada durante los primeros 2 años de la carrera Médico Cirujano de la Facultad de Medicina de la UNAM en la Ciudad de México.

El éxito académico del estudiante de licenciatura proporciona ciertos beneficios a la sociedad por su contribución al desarrollo económico, cultural y social del país, que se manifiesta en la productividad de sus actividades docentes, de investigación y difusión de la cultura.

**Definiciones preliminares**

**Algoritmo de agrupamiento**

El objetivo de los algoritmos de agrupamiento es encontar particiones disjuntas de un conjunto de datos o base de datos, de tal manera que los objetos en el misno grupos sean lo más similares que los objetos de los otros grupos (Jain, 1988).

**Descripción del algoritmo k-modas**

El algoritmo k-modas (Zhexue., 1998), fue diseñado para agrupar grandes conjuntos de datos categóricos, y tiene como objetivo obtener las k modas que representan al conjunto

**Dominios y atributos categóricos**

Zhexue en (Zhexue., 1998), describe los datos categóricos como objetos descritos únicamente por atributos categóricos o como una versión simplificada de los objetos simbólicos definidos en (Godwa, 1992).

Considera a todos los atributos numéricos (cuantitativos) al categorizarlos y no considera los atributos categóricos que están contenidos por una combinación de valores determinados. Los objetos y atributos categóricos aceptados por el algoritmo k-modas son definidos en (Zhexue., 1998).

Suponga que $A_1, A_2, \ldots, A_m$ son los m atributos que describen a un objeto en un espacio $\Omega$ y dominio$DOM(A_1), DOM(A_2), \ldots, DOM(A_m)$. Un dominio $DOM(A_j)$ es definido como categórico si es un conjunto finito y no ordenado. $\Omega$ Es un espacio categórico si todo $A_1, A_2, \ldots, A_m$ es categórico.

**Objetos categóricos**

Como en (Godwa K.C., 1991), un objeto categórico $X \varepsilon\, \Omega$ es representado como la conjunción lógica de pares atributo-valor $[A_1 = X_1] \wedge [A_2 = X_2] \wedge \ldots \wedge [A_m = X_m]$, donde $X_j\, \varepsilon\, DOM(A_j)$, para 1 $1 \leq j \leq m$ mismo para atributo-valor $[A_j = X_j]$ es llamado selector. X es un vector de la forma $[X_1, X_2, \ldots, X_m]$ y cada objeto en $\Omega$ tiene exactamente m valores atributos y si el valor para el atributo $A_j$ no está disponible para un objeto $X$ , entonces $A_j = \varepsilon$ donde $\varepsilon$ representa al valor de un atributo no disponible.

Sea $X = \{X_1, X_2, \ldots, X_n\}$ un conjunto de $n$ objetos categóricos $X \subseteq \Omega$ .El objeto $X_i$ es representado como $[X_{i1}, X_{i2}, \ldots, X_{im}]$. Dos objetos $X_i, X_k$ son iguales $X_i = X_k$ si $x_{ij} = x_{kj}$ para todo$1 \leq j \leq m$. La relación $X_i = X_k$ no quiere decir que $X_i, X_k$ sean algunos objetos en las bases de datos del mundo real. Esto implica que dos objetos tienen igual valor categórico en sus atributos $A_1, A_2, \ldots, A_m$.

Asuma que *X* consiste de *n* objetos en donde *p* objetos son distintos. Sea *N* la cardinalidad del producto cartesiano $DOM(A_1)x\ DOM(A_2)x\ \dots x\ DOM(A_m)$.
Tenemos que $p \leq N$. De cualquier modo, *n* puede ser tan grande como ***N.***

**Medidas de disimilaridad utilizadas**

Sean *X, Y* dos objetos categóricos descritos por *m* atributos categóricos. La medida de disimilaridad entre *X* y *Y* se define por el total de las no coincidencias de los atributos categóricos de los objetos. El número más pequeño de las diferencias significa que los objetos son similares (Zhexue., 1998).

Formalmente:

$$d(X,Y) = \sum_{j=1}^{m} \delta(X_j, Y_i) \quad (1)$$

Donde:

$$\delta(X_j, Y_i) = \begin{cases} 0\ (x_j = y_i) \\ 1\ (x_j \neq y_i) \end{cases} \quad (2)$$

*d(X,Y)* da igual importancia a cada categoría del atributo. Si se toma en cuenta las frecuencias de las categorías en el conjunto de datos, se define la medida de disimilaridad,

Como:

$$d_{x^2}(X,Y) = \sum_{j=1}^{m} \frac{n_{xj} + n_{yj}}{n_{xj} n_{yj}} \delta(X_j, Y_j) \quad (3)$$

Donde $n_{xj}$ y $n_{yj}$ son el número de objetos en el conjunto de datos, que tienen las categorías $x_j$ y $y_j$ para el atributo *j*. Zhexue denomina a la ecuación 3, distancia *xi-cuadrada* y la propone para descubrir grupos de objetos con baja representación en la base de datos.

**Modas de un conjunto**

Sea $X$ un conjunto de objetos descritos por atributos categóricos. Una moda de $X$ es un vector $Q = [q_1, q_2, \dots., q_m] \in \Omega$ que minimiza a $D(Q,X) = \sum_{i=1}^{n} d(X_i, Q)$ donde $X = (X_1, X_1, \dots\dots, X_n,)$ y $d$ pueden ser calculadas con la ecuación 2 o la ecución 3.

**Función criterio**

Suponga que $\{S_1, S_2, \dots, S_k\}$ es una partición de $X$ donde $S_1 \neq \emptyset$ (conjunto vacío), para $1 \leq l \leq k$ y $\{Q_1, Q_2, \dots, Q_k,\}$ las modas de $\{S_1, S_2, \dots, S_k\}$ . El costo total de la partición es definido por:

$$E = \sum_{l=1}^{k} . \sum_{i=1}^{n} y_{i,l} d(X_i, Q_l) \quad (4)$$

Donde $y_{i,l}$ es un elemento de la matriz de la partición $Y_{nxl}$ como en (Godwa, 1991) y *d* puede ser definida como la ecuación 1 o la ecuación 3. Similar al algoritmo *k-medias*, el objetivo de agrupar el conjunto *X* es encontrar un conjunto $\{Q_1, Q_2, \dots, Q_k,\}$ que puede minimizar E. La ecuación 4, puede ser minimizada por el algoritmo k-modas.

**El algoritmo K-Modas**

El algoritmo k-modas es una versión del k-medias para datos categóricos.

En k-modas se hacen 3 modificaciones a k-medias:

- Uso de diferentes medidas de disimilaridad.
- Sustitución de k medias por k modas para formar los centros.
- El método basado en las frecuencias de los datos para actualizar las modas.

La actualización de las modas se realiza en cada asignación de un objeto a su grupo, mientras que en k-medias es al final de cada iteración del algoritmo. El algoritmo k-modas al igual que el algoritmo k-medias produce soluciones óptimas locales, que dependen del conjunto de modas iniciales y el orden de los objetos en el conjunto de datos.

**Descripción del algoritmo K-modas**

Paso 1: Seleccionar k modas iniciales, una para cada grupo.

Paso 2: Asignar cada objeto a la moda más cercana utilizando la distancia d. Actualizar la moda del grupo después de cada asignación.

Paso 3: Después que todos los objetos han sido asignados a un grupo, volver a examinar la disimilaridad de los objetos con las modas actuales. Si un objeto es encontrado tal que su moda más cercana corresponde a otro grupo, asignar el objeto a su nueva moda y actualizar la moda de ambos grupos.

Paso 4: Repetir el paso 3 hasta que no existan objetos cambiados de grupo.
Metodología

La investigación es de tipo descriptiva y experimental, la cual consta de 3 etapas (descriptiva, iterativa y resultante), que representan la recolección y procesamiento de los datos, así como los resultados obtenidos.

Estas etapas se encuentran definidas a continuación.

**Etapa descriptiva**

La información se obtuvo a partir de la herramienta de análisis de documentos a través de las oficinas de Servicios Escolares y Desarrollo Académico del Instituto Tecnológico de Toluca.

Tomando como muestra a los alumnos de la carrera de Ingeniería en Sistemas Computacionales de las generaciones 2000 a 2003.

De acuerdo a los datos obtenidos se admitirán en el estudio a todos los alumnos que cumplan con los siguientes criterios:

- Contar con expediente individual en el Instituto Tecnológico de Toluca.
- Haber cursado la carrera sin ser provenientes de otra institución.
- Contar con la información completa de las variables estudiadas.

Las variables empleadas en el estudio han sido asignadas a partir de investigaciones dirigidas al análisis y comportamiento de los factores que influyen en el proceso académico del estudiante a nivel licenciatura, dichas investigaciones realizan procesos diferentes al momento de evaluar los factores, sin embargo, regularmente se encuentran dentro de una clasificación conformada por tres grupos:

1. Factores académicos.
2. Factores personales.
3. Factores socioeconómicos.

De acuerdo con la clasificación anterior se han elegido las variables que intervendrán de manera trascendental en el desarrollo del estudio, son definidas como variables independientes y señaladas a continuación:

a) Estado Civil.

b) Edad.

c) Trabajo.

d) Dependientes Económicos.

e) Institución de procedencia.

f) Tiempo de egreso.

g) Periodo de ingreso.

h) Promedio.

Se determinó como variable dependiente al éxito académico (obtención del título a nivel licenciatura) considerado 7 años a partir de la última generación evaluada.

**Etapa iterativa**

El paradigma de construcción de prototipos inicia con la comunicación. El ingeniero de software y el cliente encuentran y definen los objetivos globales para el software, identifican los requisitos conocidos y las áreas del esquema en donde es necesaria más definición. Entonces se plantea con rapidez una iteración de construcción de prototipos y se presenta el modelado (en la forma de un diseño rápido). El diseño rápido conduce a la construcción de un prototipo. Después, el prototipo lo evalúa el cliente/usuario y con la retroalimentación se refinan los requisitos del software que se desarrollará. (Pressman, 2005).

Siguiendo el modelo anterior, se plantearon diversos apartados para llevar a cabo la construcción de prototipos, validarlos y continuar con el desarrollo de la aplicación. A continuación se describen de manera práctica, dichos apartados.

**Pantalla principal**

De manera inicial se determinó el requerimiento de áreas de texto para visualizar los resultados.

Las opciones para elegir los parámetros de entrada (número de grupos a formar, tipo de ecuación y selección de modas iniciales). También contar con los botones para realizar las acciones de agrupamiento y las frecuencias de dominios. Así como la lógica principal del algoritmo de agrupamiento k-modas.

Posteriormente se identificó que se debería contar con ciertas validaciones, de acuerdo a las opciones elegidas como parámetros de entrada, ya que las variantes no son aplicables en todos los casos.

Se presentó el prototipo y se agregó la validación del número de grupos a formar, para que sea mayor o igual a 2, y menor al número total de elementos. A su vez se colocó la barra donde aparece el nombre y la ruta del archivo que se está utilizando para el agrupamiento.

Se colocó una barra de menú en la parte superior de la interfaz, originalmente con el apartado de “abrir” en la sección de archivo. Ya que con esta opción, se carga el archivo para ser analizado y agrupado.

Una vez identificado el agrupamiento de datos, se solicitó la creación de una rutina que permita guardar archivos de texto, con los resultados que genera la aplicación. Definiendo 3 tipos de archivos: 1. Resultados con etiqueta de grupo. 2. Resultados con etiqueta de grupo y los parámetros de entrada ocupados. 3. Resultados ordenados de acuerdo con las etiquetas de grupo.

De manera final se valoró y se integró la opción cerrar, para complementar el menú. Y comenzar con la asignación de teclas rápidas, así como el inicio de generar otros apartados dentro de la barra de menú.

ISSN-2410-3993

**Barra de menú complementaria**

Los demás apartados añadidos en la barra de menú (editar, herramientas, ayuda), se determinó mediante un cambio de color en editar, la creación de archivos a través de una consulta a la base de datos con la opción de herramientas y contar con una guía rápida e información del software.

Se redefinió la parte de generación de archivos, debido a que varía de acuerdo a los parámetros de la base de datos y las opciones que pueden desarrollarse al crear los archivos de texto, que serán utilizados para realizar el agrupamiento.

Se presentaron las diversas iniciativas y con los cambios requeridos, se validaron los apartados mencionados en los párrafos anteriores, para finiquitar el proceso en la creación de la aplicación. Tomando en cuenta que se encuentra abierta la posibilidad de futuras mejoras o modificaciones, en caso de ser requeridas.

**Etapa resultante**

Los datos que proporciona la aplicación k-modas7, serán representados en forma de grupos, etiquetando cada uno de sus elementos, para validar y determinar los factores que influyen en el desarrollo del estudiante para lograr la obtención del título y perfil de éxito académico.

Estos resultados podrán ser observados en la aplicación o también generar un archivo de texto, con los datos correspondientes. Los cuáles serán analizados por expertos del Departamento de Desarrollo Académico del Tecnológico de Toluca.

Finalmente en la Figura 2 pueden observarse las etapas del procesamiento de información en forma gráfica y simplificada.

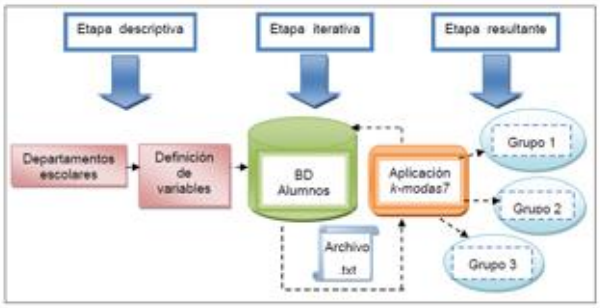



**Figura 2** Etapas del procesamiento de información

La etapa descriptiva efectúa la recolección de datos y la definición de las variables que se utilizan en la investigación.

La parte intermedia se forma a partir de la base de datos y la interacción con la aplicación k-modas7 mediante archivos de texto.

Los resultados son identificados por círculos que representan agrupaciones de alumnos con características similares.

**Aplicación K-Modas7**

La aplicación k-modas7 es una herramienta que permite agrupar grandes cantidades de datos, mediante parámetros de entrada y archivos de texto. En primera instancia se presenta la pantalla inicial, para describir los elementos que la contienen, que puede observarse en la Figura 3.

**Figura 3** Pantalla inicial

Existen 3 elementos principales en los que se compone la interfaz principal, hay una barra de menú en la parte superior, en la cual se efectúan las diversas acciones para iniciar el proceso de agrupamiento, así como opciones de edición y ayuda.

En el menú Archivo se elige el documento de texto (.txt) que se va agrupar, contando con una serie de datos identificados por un separador y con el mismo número de elementos por cada registro.

Si no se cuenta con un archivo de datos elaborado, se utiliza el menú Herramientas para generar un archivo, haciendo una consulta a la base de datos para obtener la información necesaria, para ser agrupada.

El menú Editar permite cambiar de color los datos resultantes en pantalla, y en el menú de Ayuda vienen una serie de instrucciones que sirven de apoyo para el uso de la aplicación. Para la segunda parte, puede verse una serie de opciones a elegir. En las que se encuentra el número de grupos a formar, el tipo de ecuación y la elección de las modas iniciales. Esta sección se representa por el número 2, se debe llenar el cuadro de texto con la cantidad de grupos que deseamos formar, posteriormente las ecuaciones con las que cuenta el algoritmo es la ecuación binaria y xi-cuadrada. Para finalizar la elección de parámetros, seleccionar entre primeros k elementos o modas ficticias.

En la última zona de la interfaz, se pueden ver los resultados que generan el archivo elegido, dominios de frecuencias y el agrupamiento. Todo de acuerdo a los parámetros seleccionados y que se mencionaron anteriormente.

A continuación se presenta una serie de pasos, para hacer uso correcto de la aplicación.
Ejecutar la aplicación k-modas7 para iniciar el proceso.

Seleccionar el menú Archivo-Abrir, ubicado en la Figura 4 y elegir un documento de texto almacenado en el equipo.


**Figura 4** Barra de menú

Generar un archivo en el menú Herramientas, identificado en la Figura 5 a través de una consulta a la base de datos, para crear el documento de texto (.txt) que se estará agrupando con el uso de la aplicación. Hay que determinar los parámetros de conexión a la base de datos, así como la tabla y las condiciones necesarias para hacer uso de este apartado. Finalizando con el nombre del archivo a crear.

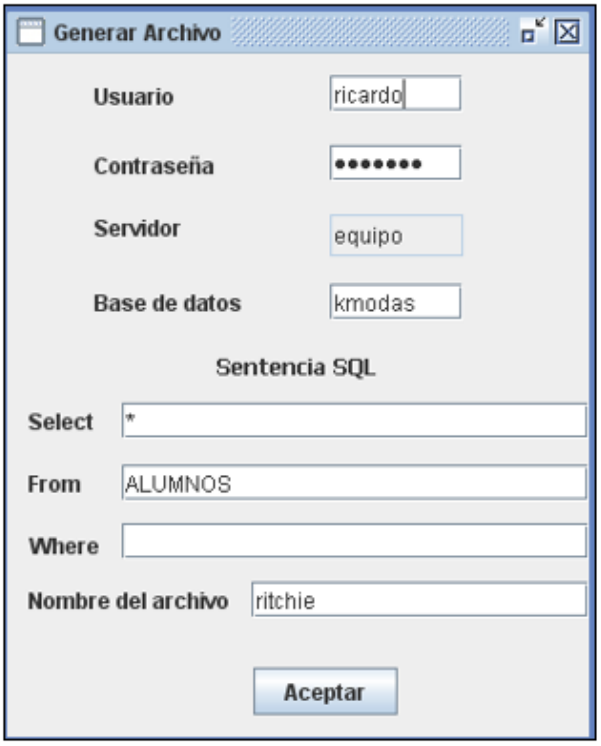



**Figura 5** Pantalla generar archivo

Una vez seleccionado el archivo que se desea agrupar, se deberá escribir el número de grupos a formar (k), con el cual se determina las particiones con las que contaran los resultados finales.

Se debe tomar en cuenta que estos grupos deben ser mayores a 1 y menores al total de elementos para analizar.

Después se tiene que elegir el tipo de ecuación que utilizará el algoritmo k-modas, dentro de la aplicación, contando con las opciones de ecuación 1 (binaria) o ecuación 2 (xi-cuadrada).

En caso de seleccionar la ecuación 2, se deberá obtener las frecuencias de dominios para poder continuar el proceso, dando clic en el botón de Obtener frecuencias.

Ahora se tendrá que seleccionar el método de elección de modas iniciales, ya que deben crearse una serie de modas para que a partir de ellas se genere el agrupamiento. Se cuenta con las opciones de primeros k elementos y modas ficticias. A continuación se muestran los parámetros mencionados en la Figura 6.

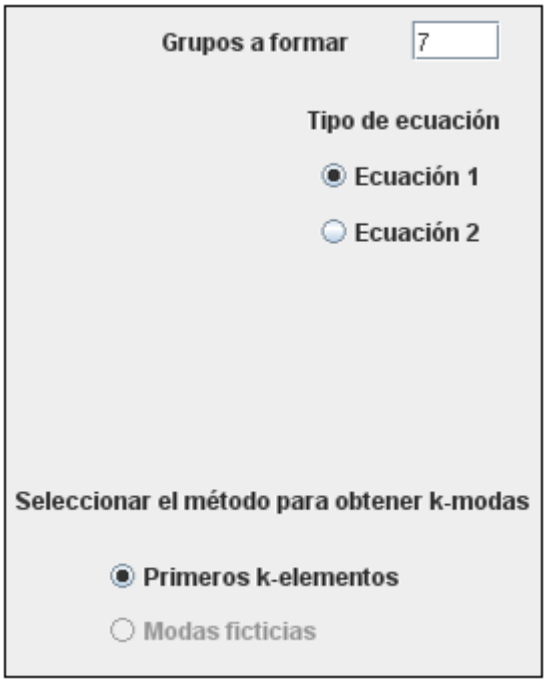



**Figura 6** Sección de parámetros de entrada

Presionar el botón efectuar agrupamiento, para que se generen los resultados correspondientes, de acuerdo a los parámetros seleccionados. En esta parte termina el proceso que genera los grupos, para hacer uso de estos hay que hacer clic en el menú Archivo-Guardar, con lo que se va a generar una carpeta que contiene 3 archivos de texto para utilizar los resultados agrupados.

## Resultados

## Diseño de pruebas

Para desarrollar las pruebas de la investigación, fue necesario utilizar diversos parámetros que determinan el rumbo del proceso y de los resultados.

Se tendrá que elegir inicialmente un archivo de texto que contenga los datos para analizar, después se debe asignar el número de grupos a formar (k), seleccionar el tipo de ecuación (binaria o xi-cuadrada) y finalmente el método para determinar las modas iniciales (primeros k-elementos o modas ficticias).

De acuerdo a las opciones antes mencionadas, se presentará de manera estructurada, las posibles combinaciones para realizar las pruebas necesarias en el estudio, de acuerdo a la figura 7.

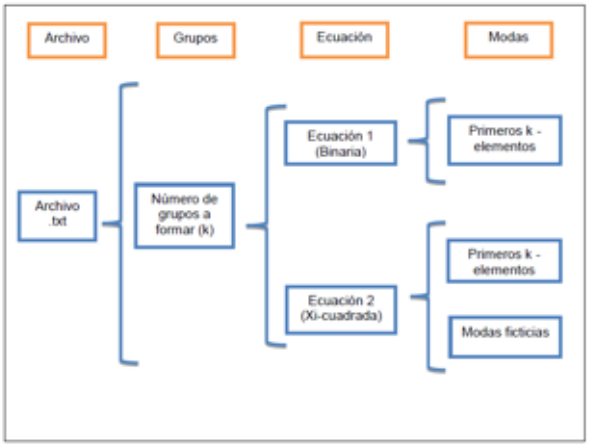



**Figura 7** Representación estructurada del diseño de las pruebas.

**Descripción de las pruebas realizadas**

En esta sección se muestran los resultados obtenidos al agrupar el archivo alumnos_145.txt con la aplicación k-modas7, empleando diferentes opciones de configuración, al elegir el número de grupos a formar, el tipo de ecuación y la forma de elección de modas iniciales.

Para llevar a cabo las diversas pruebas, se utilizó principalmente un equipo Dell Inspiron 1420, con procesador Core2 Duo, una memoria RAM de 2GB y un disco duro de 80 GB.

**Resultados de las pruebas**

En este apartado se describen los resultados de las pruebas desarrolladas, presentando una parte de la tabla que contiene los parámetros utilizados en la aplicación, así como los datos finales del perfil del estudiante y la pureza de la agrupación. Ver Tabla 1.

| Parámetros / Pruebas | Prueba 1 | | Prueba 2 | | Prueba 3 | |
|---|---|---|---|---|---|---|
| Grupos a formar | k=2 | | | | | |
| Tipo de ecuación | Ecuación 1 | | Ecuación 2 | | | |
| Modas iniciales | Primeros k elementos | | Primeros k elementos | | Modas Ficticias | |
| G0 | (S,27,SI,0,PREPARATORIA ESTATAL,4.5,AGOSTO - DICIEMBRE 2001,81,SI) | 82.00% | NA | NA | (S,26,NO,0,PREPARATORIA ESTATAL,5.5,AGOSTO - DICIEMBRE 2003,84,SI) | 60.42% |
| G1 | (S,26,NO,0,PREPARATORIA ESTATAL,5.5,AGOSTO - DICIEMBRE 2003,84,NO) | 51.58% | (S,26,NO,0,PREPARATORIA ESTATAL,5.5,AGOSTO - DICIEMBRE 2003,84,SI) | 60.14% | NA | NA |
| G2 | | | | | | |
| G3 | | | | | | |
| G4 | | | | | | |
| G5 | | | | | | |
| G6 | | | | | | |
| G7 | | | | | | |
| G8 | | | | | | |
| G9 | | | | | | |

**Tabla 1** Resultados de las pruebas

Los resultados finales se generan con 15 pruebas y diferentes variantes en la elección de parámetros de entrada.

Para la aplicación que emplea el algoritmo de agrupamiento k-modas. En la tabla 1 se presentan los resultados de las tres mejores pruebas, las cuales obtuvieron la mejor pureza de grupos.

A partir de los experimentos desarrollados puede verse que en la prueba donde k=4, la ecuación empleada es la binaria (ecuación 1) y el método de elección de modas es primeros k elementos, que se encuentran tiene un porcentaje de los más altos en el estudio para los alumnos que si logran obtener el título de Ingeniería en Sistemas Computacionales (66 de 69 elementos), contando con una pureza de 95.65%. Para esta prueba el perfil de los estudiantes se describe con las siguientes características {Estado Civil = Soltero, Edad = 26, Trabajo = No, Dependientes Económicos = 0, Institución de procedencia = Preparatoria Estatal, Tiempo de Egreso = 5, Periodo de Ingreso =Agosto – Diciembre 2003, Promedio = 84, Titulo = Si}.

Para la prueba donde el número de grupos a formar (k)=5, ecuación 1 y primeros k elementos como modas iniciales, tiene una pureza del 100% para un grupo de 11 elementos. Esta prueba contiene las características principales de {Estado Civil = Soltero, Edad = 27, Trabajo = No, Dependientes Económicos = 0, Institución de procedencia = Preparatoria Estatal, Tiempo de Egreso = 6, Periodo de Ingreso = Agosto – Diciembre 2001, Promedio = 81, Titulo = Si}.

Como se puede ver el resultado en dos de las pruebas es igual, haciendo referencia a la moda planteada anteriormente {Estado Civil = Soltero, Edad = 27, Trabajo = No, Dependientes Económicos = 0, Institución de procedencia = Preparatoria Estatal, Tiempo de Egreso = 6, Periodo de Ingreso = Agosto – Diciembre 2001, Promedio =81, Titulo = Si}.

Esto significa que el algoritmo, proporciona resultados confiables independientemente del número de grupos a formar (k), ya que se puede encontrar similitudes en los datos finales al momento de hacer diferentes pruebas, con diversos parámetros.

## Conclusiones

Se diseñó un software que realiza agrupamiento de una base de datos con el algoritmo K-Modas. Además se conformó una base de datos con información de 145 alumnos pertenecientes a la carrera de Ingeniería en Sistemas Computacionales, tomando en cuenta las generaciones del año 2000 a 2003 en los diversos periodos escolares. Así la base de datos de pueba estuvo conformada por 150 registros descritos en los siguientes campos:

a) Estado Civil.

b) Edad.

c) Trabajo.

d) Dependientes Económicos.

e) Institución de procedencia.

f) Tiempo de egreso.

g) Periodo de ingreso.

h) Promedio.

Se utilizaron 2 medidas de disimilaridad en el estudio, la primera es la ecuación binaria y la segunda es la ecuación xi-cuadrada. El algoritmo k-modas fue evaluado para emplearlo en esta investigación, ya que puede ser utilizado con datos no numéricos.

De acuerdo a las 15 pruebas realizadas con el algoritmo de agrupamiento k-modas, pueden observarse diferentes resultados conforme a los parámetros de entrada que requiere para su funcionamiento. Los mejores resultados obtenidos encontraron que, aquellos estudiantes que podrán lograr la obtención del título de Ingeniería en Sistemas Computacionales deberán contar con las siguientes características:

Estado Civil: Soltero.
Edad: 27 años.
Trabajo: No.
Dependientes Económicos: 0.
Institución de Procedencia: Preparatoria Estatal.
Tiempo de Egreso: 6 años.
Periodo de Ingreso: Agosto – Diciembre 2001.
Promedio: 81.
Título: Si.

## Referencias

Acosta E., Cortés MT., y Vélez I. (2004). Seguimiento de egresados de la Facultad de Medicina de la UNAM. Revista de Educación Superior, 7-20.

Navarro Rubén Edel (2003). «Factores asociados al rendimiento académico.»Revista Iberoamericana de Educación.

Belvis Pons Esther, Andrés Moreno Ma. Victoria, y Ferrán Ferrer Julia. (20009). «Los factores explicativos del éxito y fracaso académico en las universidades españolas, en los años del cambio hacia la convergencia Europea.» Revista Española de Educación comparada, no 15, 61-92.

Fayyad U.M., Piatesky-Shapiro G., y Smyth P. (1996) «From Data Mining To Knowledge Discovery: An Overview.» Editado por G. Piatesky-Shapiro, P. Smyth and R. Uthurusamy, U.M. Fayyad. In Knowledge Discovery and Data Mining (AAAI Press/The MIT Press), Menlo Park, CA.

Gatica Lara Florina, Méndez Ramírez Ignacio, Sánchez Mendiola Melchor, y Martínez González Adrían. (2010). «Variables asociadas al éxito académico en los estudiantes de la Licenciatura en Medicina de la UNAM.» Revista de la Facultad de Medicina de la UNAM 53, no 5, 9-11.

Godwa K.C., y Diday E. (March/April 1992). «Simbolic Clustering Using a new Similarity Measure.» IEEE Transaction on Systems, Man and Cybernetic 22, no 2, 368-378.

Gómez Moliné Margarita. (2003). «Algunos factores que influyen en el éxito académico de los estudiantes universitarios en el área de química.» Tesis doctoral, Barcelona.

Zhexue Huang (1998). Extendions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery 2,3, Kluwer Academic Publisher, no 3, 283-304, 1384-5810.

Martínez González A., Urrutia Aguilar M.E., Martínez Franco A.I., Ponce Rosas R., y Gil Miguel A. (2003) «"Perfil del estudiante de posgrado con éxito académico en la UNAM".» Revista de investigación e innovación educativa, no 32, 133-145.

Pressman, Roger S. (2005). Ingeniería del Software: Un Enfoque Práctico. España: McGraw-Hill.

Tinto Vincent. (1992). «El abandono de los estudios superiores: una perspectiva de las causas del abandono y su tratamiento.» Cuadernos de planeación universitaria, México: UNAM (ANUIES) 6, no 2,9-37.

Godwa K.C., y Diday E. (1991). «Symbolic Clustering Using a New Disimilarity Measure.» Pattern Recognition, 567-578.

Hand D.J. (1981),«Discrimination and Classification.» John Wiley & Soon.

Kaufman L., Rousseeuw P. J. (1989), Finding Groups in Data “An Introducction to Cluster Analysis, Wiley series in probability and Mathematical Statistics.

Jain A.J., Dubes R. C. (1988), Algorithms for Clustering Data, Prentice Hall.

ISSN-2410-3993

Artículo **Revista Tecnología e Innovación**
Diciembre 2015 Vol.2 No.5 942-957

# Una versión modificada del algoritmo de agrupamiento Isodata

RENDON, Eréndira*†, MENDOZA, Marcos, CISNIEGA, Roció y CARBAJAL, Guillermo

Recibido 5 de Julio, 2015; Aceptado 24 de Noviembre, 2015

## Resumen

El algoritmo de agrupamiento Isodata es uno de los más utilizados por la comunidad de minería de datos, aunque cuenta con algunas desvestajas. En este artículo se presentan dos versiones modificadas del algoritmo de agrupamiento Isodata, que calcula automáticamente los parámetros de entrada θ_c y θ_s. Las pruebas realizadas sugieren que se obtienen los mismos resultados de acuerdo a la medida SSE.

**Agrupamiento, Isodata, Minería de datos**

## Abstract

Isodata algorithm is one of the most used by the data mining community, even though it has some disadvantages. In this paper we present two modified versions of Isodata clustering algorithm where θ_c and θ_s input parameters are automatically calculate. Results show similar performance to the original algorithm according to SSE measure.

**Clustering, Isodata, Data mining.**

**Citación:** RENDON, Eréndira, MENDOZA, Marcos, CISNIEGA, Roció y CARBAJAL, Guillermo. Una versión modificada del algoritmo de agrupamiento Isodata. Revista de Tecnología e Innovación 2015, 2-5: 942-957

* Correspondencia al Autor (Correo Electrónico: erendon @ittoluca.edu.mx)
† Investigador contribuyendo como primer autor.

## Introducción

El aprendizaje automático ha proporcionado técnicas básicas para la minería de datos, para extraer conocimiento de las bases de datos. El aprendizaje automático es dividido en dos áreas: el aprendizaje supervisado y el aprendizaje no supervisado; dentro del aprendizaje no supervisado existe una herramienta denominada agrupamiento o clustering. Por otro lado el agrupamiento es una técnica muy utilizada en la minería de datos.

El agrupamiento encuentra grupos o particiones en un conjunto de datos o base de datos, de tal manera que los objetos que queden en el mismo grupo sean similares entre si y disimilares de los objetos de los otros grupos.

Dentro del agrupamiento se cuentan con técnicas de agrupamiento básicas, las jerárquicas y las de partición. Las técnicas jerárquicas organizan los datos en una secuencia anidada de grupos, pueden iniciar considerando un objeto como un grupo y de esta forma ir mezclándolos, la mezcla continúa hasta que todos los objetos pertenecen a un solo grupo o cuando el usuario decide escoger un nivel de agrupamiento en la jerarquía; por otro lado se puede optar por el método inverso, considerando todos los objetos como un grupo e ir dividiendo el grupo en grupos más pequeños, hasta que un objeto sea considerado un grupo o el usuario decida la jerarquía o nivel de agrupamiento.

Así mismo las técnicas de agrupamiento basadas en partición van obteniendo un número k de particiones de los datos, optimizan una función objetivo en donde k es el número de grupos deseados del conjunto de datos, la forma de representar los grupos es por centros de gravedad o por objetos asignados al centro más cercano (centroides), buscando obtener grupos naturales presentes en los datos empleando ajuste en los centros.

Dentro de los algoritmos más comunes de esta familia tenemos el k-Means (Jain 1988), (Kaufman L, 1989), PAM (Partitional Around Medoid) (Kaufman L., 1989), CLARA (Clustering Large Applications) (Kaufman L., 1989), ISODATA (Iterative Self-Organizing Data Analysis Techniques) (Ball G., 1965), todos estos algoritmos funcionan con datos de tipo numérico.

El algoritmo de agrupamiento ISODATA, el cual tiene como base el algoritmo k-Means, incluye una serie de operaciones heurísticas e involucra un conjunto de parámetros extra, el algoritmo ISODATA emplea iteraciones en las cuales incorporan la eliminación de grupos poco numerosos, la fusión de grupos cercanos y la división de grupos dispersos.

El algoritmo ISODATA es considerado un excelente algoritmo de agrupamiento, si y sólo si los parámetros que requiere de entrada están correctamente definidos, ya que al ser un algoritmo iterativo depende en gran medida del conocimiento a priori del conjunto de datos y su experiencia para poder proporcionar eficientemente los parámetros que necesita el algoritmo.

La eficiencia de algoritmo ISODATA depende de estimación correcta de los parámetros de entrada, de tal forma que puede crear grupos perfectamente establecidos y diferenciados, o en caso contrario generar grupos débiles que no aportarán conocimiento significativo a la persona que lo emplea, ya que el objetivo del algoritmo es encontrar información interesante y relevante dentro del conjunto de datos.

El algoritmo ISODATA posee grandes ventajas sobre otros algoritmos de agrupamiento al introducir la división y fusión de grupos, buscando grupos naturales presentes en el conjunto de datos; por otro lado cabe

ISSN-2410-3993

mencionar que el algoritmo ISODATA al igual que muchos de los algoritmos de partición presentan sensibilidad debido a los parámetros de entrada que requieren para funcionar, es aquí donde se encuentran los parámetros que determinan la fusión ($\theta_C$) y división de grupos ($\theta_S$). Sin embargo en muchas aplicaciones reales es difícil calcular correctamente estos parámetros, entonces una forma de eliminar esta desventaja es calcular automáticamente los parámetros tanto de fusión como de división de grupos, realizando tal estimación sin contar con información a priori y considerando la forma en cómo se encuentran distribuidos los objetos previamente o en los primeros pasos de la aplicación del algoritmo.

Actualmente el algoritmo requiere de un conocimiento a priori del conjunto de datos para poder establecer por el usuario los parámetros antes mencionados, entonces el problema a resolver es estimar los parámetros $\theta_C$ y $\theta_S$ sin contar con información a priori.

En este trabajo se presenta dos versiones del algoritmo Isodata, donde se incluyen un método que estima de manera adecuada el parámetro de entrada de fusión de grupos θC y así mismo el parámetro de división de grupos θS del algoritmo de agrupamiento ISODATA.

Este trabajo se encuentra organizado de la siguiente manera, en la primera sección se presentan algunos trabajos que se han realizados sobre el algoritmo Isodata, en la sección se describe el algoritmo Isodata, en la sección 3 se presenta los algoritmos de las modificaciones propuestas, en la sección cuatro se presntan las pruebas y los resultados obtenidos y finalmente en sección cinco las conclusiones a las cuales se llegaron.

**Trabajos relacionados**

En (Kohei A., 2007) se presenta un nuevo método donde se emplean algoritmos genéticos para obtener los parámetros de fusión y división de grupos. Según los resultados obtenidos el uso de algoritmos genéticos para la obtención de los parámetros θC y θS genera una mejor selección de los grupos. En este nuevo método los algoritmos genéticos proporcionaron un método alternativo para determinar el umbral en la separación e integración de la variedad de grupos formados por el algoritmo ISODATA, los resultados obtenidos muestran mejoría notable el resultado, debido a que el método típico ejecutado en el ISODATA distribuye el grupo suponiendo que es una función convexa y cuando la distribución del grupo es una función cóncava éste puede responder en cierta medida por la fusión y división, pero si el procedimiento convencional del algoritmo es seguido entonces el grupo clasificado correctamente puede ser destruido, mediante lo descrito anteriormente el método propuesto en (Kohei A. ,2007) obtiene grupos mejor distribuidos y definidos.

En (El-Zaart., 2010) se expone la aplicación del algoritmo ISODATA en la segmentación de imágenes, fundamental en diversas vertientes del procesamiento de imágenes. En esta investigación se asume que los datos de la imagen son modelados por la distribución Gamma en combinación con el algoritmo ISODATA se desarrolla un nuevo método útil en la fase de segmentación de imágenes. La aplicación del ISODATA en (El-Zaart., 2010) es calcular los umbrales y segmentar la imagen, el objetivo perseguido es dividir la imagen en una región no homogénea (histograma) en dos sub-regiones (modo), de esta forma un histograma de una imagen puede ser en modo simétrico o asimétrico.

ISSN-2410-3993

La distribución Gamma es empleada para modelar formas simétricas y no simétricas, por lo tanto se emplea esta distribución para aproximar el histograma de una imagen por una mezcla de distribuciones y así los parámetros estadísticos extraídos de la imagen pueden ser lo más exactos posibles. El propósito es usar la distribución Gamma para estimar los parámetros necesarios y aplicar el ISODATA al segmentar la imagen. El algoritmo propuesto en (El-Zaart., 2010) pretende mejorar la división y fusión de las clases, si la clase no es homogénea los parámetros iniciales de la clase son requeridos para dividir en dos subclases diferentes. Las clases se combinarán si bien el número de miembros (píxeles) es menor que el valor para los miembros mínimos de una clase ó por otro lado si los centros de dos clases están más cerca que el valor de distancia mínima media entre dos clases. En conclusión la división y los pasos de la fusión en el ISODATA de (El-Zaart., 2010) requieren una estimación de medias y umbrales, y mediante la distribución Gamma se realiza el cálculo de parámetros de fusión y división de clases.

En (Pavan K., 2008) se expone un método para la generación del factor de mezcla o fusión empleado en el ISODATA. Como se describe en (Pavan K. 2008) aplicar la inteligencia artificial en asuntos de genética es cada vez más común, en específico en los microarrays cuyo objetivo es identificar genes co-expresados y patrones de coherencia además del análisis de las expresiones genéticas. En esta investigación se propone un algoritmo de generación automática del factor de mezcla para el ISODATA (AGMFI), de esta forma agrupar los datos de microarrays sobre la base de ISODATA, en AGMFI se generan valores iniciales para el factor de mezcla en vez de seleccionar valores heurísticos como en el ISODATA tradicional.

Se desarrollaron pruebas para medir el desempeño del AGMFI mediante la aplicación de un conjunto de datos conocido y a disposición del público y por otro lado datos sintéticos, los experimentos indican que el algoritmo aumento el enriquecimiento de los genes de función similar en el grupo. En el algoritmo expuesto solo se emplean 4 parámetros de entrada (número de grupos, número mínimo de elementos, parámetro de la división y el número máximo de iteraciones) con los cuales se puede seguir la secuencia normal del ISODATA hasta la parte de la posible fusión de grupos. Para generar el factor de mezcla se debe calcular la matriz de distancias entre grupos, encontrar la mínima distancia entre dos grupos y hallar la distancia promedio entre todos los grupos del conjunto, posterior se obtiene un promedio de las distancias anteriormente mencionadas y se procede con el calculó del factor de mezcla. Las conclusiones obtenidas muestran mejores resultados en comparación con el ISODATA tradicional y el K-Medias, pero los resultados continúan teniendo gran dependencia de los centroides iniciales.

**Algoritmo de agrupamiento Isodata**

Los parámetros de entrada que maneja el algoritmo ISODATA (Ball G., 1965) son los siguientes:

$N_C$: Número actual de grupos que han sido formados.

$k$: Número deseado o estimado a priori de grupos.

$\theta_N$: Número mínimo de elementos o miembros de un grupo para constituirlo como tal.

$\theta_S$: Desviación típica máxima, servirá para aplicar el criterio de división de un grupo o clase en dos, la división se realiza si la desviación típica del grupo es superior a $\theta_s$.

$\theta_C$: es un parámetro de unión de dos grupos, se emplea para comprobar si la distancia euclídea entre dos grupos es menor que $\theta_C$ en cuyo caso son dos grupos a fusionar.

$L$: Cuando en una iteración genérica del algoritmo existe más de una pareja de grupos susceptibles a unirse, este parámetro limita el número de fusiones que pueden llevarse a cabo en esa iteración.

$I$: Número máximo de iteraciones que puede ejecutar el algoritmo.

**Pasos del algoritmo ISODATA**

El algoritmo ISODATA se describirá a continuación mediante una serie de pasos para su fácil comprensión.

**Inicialización**

Se comienza con darle valor a los parámetros, recomendando asignar $k$ ha $N_C$, se eligen $k$ elementos entre los $P$ elementos a clasificar: $X_1, X_2, \dots, X_P$ formando con cada uno de ellos un grupo inicial. Se tienen entonces los $k = N_C$ centroides $Z_1, Z_2, \dots, Z_{Nc}$.

**Distribuir los elementos entre los distintos grupos.**

Se agrupan los elementos $x_1, x_2, \dots, x_P$ entre los $N_C$ grupos ya formados, siguiendo el principio de la mínima distancia euclídea, empleando la siguiente ecuación:

$$x_j \in \alpha_i \;\; si \;\; \|x_i - Z_i\| mínima$$
$$\forall j = 1,2 \dots p; \quad \forall i = 1,2 \dots N_c \qquad (1)$$

**Eliminar los grupos con un número insuficiente de miembros.**

Se procede con la eliminación de grupos que tengan un número de elementos inferior a $\theta_N$, actualizando el parámetro $N_C$, si la eliminación de grupos procede posterior a ésta se debe volver a agrupar esos elementos entre los centroides existentes.

**Actualizar los centroides de los grupos.**

La actualización se lleva a cabo calculando la media muestral de cada grupo, empleando la siguiente ecuación:

$$Z_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_j; \qquad i = 1,2 \dots N_c \qquad (2)$$

Donde $N_i$ es el número de elementos de la clase $\alpha_i$.

**Cálculo de la distancia euclídea media de cada grupo**

Para cada grupo se debe obtener la distancia euclídea media de sus elementos con respecto a su centroide, empleando la siguiente ecuación:

$$\overline{D_i} = \frac{1}{N_i}\sum_{j=1}^{N_i}\|x_j - Z_i\|; \qquad i = 1,2 \dots N_c \qquad (3)$$

Lo que devuelve este parámetro es una medida de la dispersión de los elementos de cada grupo con respecto a su media, y se utilizará posteriormente para la posible división de un grupo.

**Cálculo de la distancia media de todos los grupos**

De las distancias obtenidas en el paso anterior se obtiene el promedio:

$$\overline{D} = \frac{1}{N_c}\sum_{i=1}^{N_c} N_i \overline{D}_i \qquad (4)$$

**Comprobación de bifurcaciones**

- Se comprueba en primer lugar si se trata de la última iteración, si es así entonces se hace $\theta_C = 0$ y se avanza al paso 11 (unión de grupos).

- Por otro lado se verifica si es posible unir grupos, considerando si $N_C \geq 2k$, si es así se avanza al paso 11 (unión de grupos).

- Si no se cumple alguna de las condiciones anteriores se prosigue con la secuencia natural que se describe a continuación.

**Cálculo del vector de desviaciones típicas de cada grupo**

Al trabajar con un vector de características n-dimensional, los grupos presentan un vector n-dimensional de desviaciones típicas como se muestra a continuación:

$$\sigma_i = \begin{pmatrix} \sigma_{i1} \\ \sigma_{i2} \\ \cdots \\ \cdots \\ \sigma_{in} \end{pmatrix};\ \ \sigma_{ij} = \sqrt{\frac{1}{N_i}\sum_{K=1}^{N_i}\left(X_{kj} - Z_{ij}\right)^2} \tag{5}$$

De la fórmula anterior donde:

- $i$= 1,2,…, $N_c$ (grupos actuales);
- $j$= 1,2,…, $n$ (características);
- $K$= 1,2,…, $N_i$ (elementos de la clase $\alpha_i$);

**Obtener desviaciones típicas máximas de cada grupo**

De cada grupo se selecciona el componente mayor del correspondiente vector de desviaciones típicas, entonces se forma el conjunto:

$$\{\sigma_{1\,max}, \sigma_{2\,max} \cdots\ \sigma_{Nc\,max}\}$$

Posible división de grupos

Para una clase, $\alpha_j$ en que se cumple que $\sigma_{jmax} > \theta_S$ y cumple con alguna de las siguientes condiciones:

- $D_j > D$ y $N_j > 2(\theta_N + 1)$
- $N_C \leq \frac{K}{2}$; $N_C$ es el número de elementos del grupo

La primera condición significa que la dispersión media del grupo $\sigma_j$ candidato a dividirse en dos, es superior a la media de las dispersiones de todos los grupos; y la segunda condición significa que el número de sus elementos es al menos superior al doble del número mínimo para formar un grupo.

Si se cumple entonces se divide el grupo en dos, siguiendo alguno de los procedimientos que se plantean a continuación:

1. Una posibilidad para el proceso de división es crear dos nuevos centroides, $Zj^+$ y $Zj^-$ a partir de $Zj$, de tal forma que las componentes de los nuevos centroides coincidan con los de $Zj$, excepto la componente con la máxima dispersión, es decir la $Z_k$, siendo la dispersión $\sigma_{jmax}$, entonces los componentes de $Zj^+$ y $Zj^-$ serán:

$$Z_j k^+ = Z_j k + \gamma\sigma_{j\,max} \tag{6}$$

$$Z_j k^- = Z_j k - \gamma\sigma_{j\,max}\ , con\ 0 < \gamma < 1 \tag{7}$$

*Para este trabajo el valor de $\gamma$ se tomará como* 0.5

- Lo que se pretende con esta división es distribuir adecuadamente las muestras originales del grupo antes de la división entre los dos nuevos grupos.

2. Por otro lado se tiene la alternativa de división basada en obtener las dos muestras del grupo $\alpha_j$ más alejadas entre sí y con respecto a su centroide, si las muestras obtenidas se representan como $Zj^+$ y $Zj^-$ los dos nuevos centroides se calcularan de la manera siguiente:

$$Z_{j1} = \frac{(Zj^+ + Zj)}{2} \quad (8)$$

$$Z_{j2} = \frac{(Zj^- + Zj)}{2} \quad (9)$$

**Cálculo de distancias entre grupos**

Para la posible unión de grupos se debe calcular previamente todas las distancias entre parejas de grupos, empleando:

$$D_{ij} = D_{ji} = \|Z_i - Z_j\| \quad (10)$$
$$i = 1,2 \dots N_c - 1; j = i + 1, i + 2 \dots N_c$$

**Posible unión**

Se comparan las distancias $D_{ij}$ con el parámetro $\theta_C$ de forma que se toman, si existen, las $L$ más pequeñas en orden creciente, teniendo:

$$\{D_1, D_2 \dots D_L\} \; con \; D_1 < D_2 < \cdots < D_L$$

**Proceso de unión**

Se comienza con los pares de grupos con las distancias menores, supóngase que se unirán los grupos $i, j$ cuya distancia es $D_{ij}$ encontrada dentro del conjunto $\{D_1, D_2, \dots, D_L\}$ con $D_1 < D_2 < \cdots < D_L$. Sí y sólo sí ninguno de estos dos grupos ha sido fusionado previamente con otro en esta misma iteración, se forma un grupo único cuyo centroide es:

$$Z_{ij} = \frac{1}{N_i + N_j} * \left(N_i Z_i + N_j Z_j\right) \quad (11)$$

Siendo $N_i$ y $N_j$ el número de elementos de los grupos $\alpha_i$ y $\alpha_j$ respectivamente antes de la fusión. En cada unión se actualiza el parámetro $N_C$ ya que el grupo se puede unir una sola vez en cada iteración, generalmente no se obtendrán $L$ uniones en cada iteración.

**Comprobar última iteración**

Se comprueba si se ha llegado a la última iteración, comparando con el parámetro $I$, el caso negativo se vuelve al paso 2 iniciando una nueva iteración.

Para una fácil comprensión se muestra en la Figura 3.1 el diagrama de flujo del algoritmo ISODATA.

**Método propuesto**

**Modificación M1 del algoritmo Isodata**

El algoritmo ISODATA (Ball G., 1965) con la Modificación 1 se describirá a continuación.

**Inicialización**

Se empieza asignando valores a los parámetros, se recomienda asignar $k$ ha $N_C$, se eligen $k$ elementos entre los $P$ elementos a clasificar: $X_1, X_2, \dots, X_P$ formando con cada uno de ellos un grupo inicial. Se tienen entonces los $k = N_C$ centroides $Z_1, Z_2, \dots, Z_{Nc}$.

**Distribuir los elementos entre los distintos grupos**

Se agrupan los elementos $x_1, x_2, \dots, x_P$ entre los $N_C$ grupos ya formados, siguiendo el principio de la mínima distancia euclidiana, empleando la siguiente ecuación:

$$x_j \in \alpha_i \;\; si \;\; \|x_i - Z_i\| mínima \quad (12)$$
$$\forall j = 1,2 \dots p; \quad \forall i = 1,2 \dots N_c$$

**Eliminar los grupos con un número insuficiente de miembros**

Se procede con la eliminación de grupos que tengan un número de elementos inferior a $\theta_N$, actualizando el parámetro $N_C$, si la eliminación de grupos procede posterior a ésta se debe volver a agrupar esos elementos entre los centroides existentes.

**Actualizar los centroides de los grupos**

La actualización se lleva a cabo calculando la media muestral de cada grupo, empleando la siguiente ecuación:

$$Z_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_j; \qquad i = 1,2 \dots N_c \tag{13}$$

Donde $N_i$ es el número de elementos de la clase $\alpha_i$.

**Comprobación de bifurcaciones**

- Se comprueba en primer lugar si se trata de la última iteración, si es así entonces se hace $\theta_C = 0$ y se avanza al paso 11(unión de grupos).

- Por otro lado se verifica si es posible unir grupos, considerando si $N_C \geq 2k$, si es así se avanza al paso 11 (unión de grupos).

- Si no se cumple alguna de las condiciones anteriores se prosigue con la secuencia natural que se describe a continuación.

**Cálculo del vector de desviaciones típicas de cada grupo**

Al trabajar con un vector de características *n*-dimensional, los grupos presentan un vector *n*-dimensional de desviaciones típicas como se muestra a continuación:

$$\sigma_i = \begin{pmatrix} \sigma_{i1} \\ \sigma_{i2} \\ \cdots \\ \dots \\ \sigma_{in} \end{pmatrix};\ \sigma_{ij} = \sqrt{\frac{1}{N_i}\sum_{K=1}^{N_i}\left(X_{kj} - Z_{ij}\right)^2} \tag{14}$$

Donde:

$i$= 1,2,…, $N_c$ (grupos actuales);
$j$= 1,2,…, $n$ (características);
$K$= 1,2,…, $N_i$ (elementos de la clase $\alpha_i$);

La desviación típica de cada grupo ($\sigma_i = (\sigma_{i1}, \sigma_{i2}, \dots, \sigma_{in})$) se almacena de acuerdo a las características empleadas, mas adelante se empleará junto con otros componentes en la fase de división de un grupo.

**Cálculo de la matriz de distancias entre grupos**

En este paso se calculan las distancias entre grupos, es decir obtener las distancias entre todos los grupos actuales, para esto se emplea la siguiente fórmula:

$$D_{ij} = D_{ji} = \left\|Z_i - Z_j\right\| \tag{15}$$
$$i = 1,2 \dots N_c - 1; j = i+1, i+2 \dots N_c$$

La matriz que ejemplifica el escenario se muestra en la Tabla 1:

| Grupo/Grupo | 1 | 2 | 3 | 4 | i |
|---|---|---|---|---|---|
| 1 | 0 | $D_{1,2}$ | $D_{1,3}$ | $D_{1,4}$ | $D_{1,i}$ |
| 2 | | 0 | $D_{2,3}$ | $D_{2,4}$ | $D_{2,i}$ |
| 3 | | | 0 | $D_{3,4}$ | $D_{3,i}$ |
| 4 | | | | 0 | $D_{4,i}$ |
| i | | | | | 0 |

**Tabla 1** Ejemplo de matriz de distancia entre grupos.

$Con\ i = 1,2 \dots N_c - 1;$

De los resultados obtenidos en la matriz se selecciona la mínima distancia entre dos grupos ($D_{min}$) y además se debe calcular el promedio de todas las distancias obtenidas como a continuación se muestra:

$$\overline{D} = \frac{D_{12} + D_{13} + \cdots + D_{(i-1)\,i}}{i} \quad (16)$$

Con $i$= número de distancias entre centroides del conjunto de datos.

**Cálculo del factor de mezcla $\theta_C$.**

Una vez obtenido $D_{min}$ y $\overline{D}$ se procede a calcular el factor de mezcla, el cual se obtiene de la siguiente forma:

$$\theta_C = \frac{D_{min} + \overline{D}}{2} \quad (17)$$

**Obtención de rangos por clase**

Haciendo uso de las desviaciones típicas (calculadas en el paso 6) de cada grupo ($\sigma_i = (\sigma_{i1}, \sigma_{i2}, \ldots, \sigma_{in})$) y junto con los centroides actuales $(Z_1, Z_2, \ldots, Z_{Nc})$ se procede a calcular rangos por cada grupo y por cada característica como se muestra a continuación:

$$R_i = \begin{pmatrix} R_{i1} \\ R_{i2} \\ \cdots \\ \ldots \\ R_{in} \end{pmatrix};\ R_{ij} = \left[\mu_{ij} - \sigma_{ij}\ , \mu_{ij} + \ \mu_{ij}\right] \quad (18)$$

Donde:

$i$= 1,2,…, $N_c$ (grupos actuales);
$j$= 1,2,…, $n$ (características);
$\mu$= es la media centroide $i$ de la característica $j$
$\sigma$= La desviación típica del grupo $i$ con la característica $j$ (previamente calculado en el paso 6)

De lo anterior se obtiene un rango por clase, este rango está en función del número de características de los objetos del conjunto analizado.

Posible división de grupos

Con los rangos por clase se comienza a evaluar si cada característica del objeto se encuentra dentro del rango establecido, es decir para un objeto $X_{ij}$ que pertenece a la clase $\alpha_i$ , la característica $j$ de dicho objeto debe encontrar entre los rangos calculados por la Ec. 18 para la característica $j$ de la clase $i$, se denota como sigue: $X_{ij} \in R_{ij} = \left[\mu_{ij} - \sigma_{ij}\ , \mu_{ij} + \ \mu_{ij}\right]$ en donde:

$i$= 1,2,…, $N_c$ (grupos actuales);
$j$= 1,2,…, $n$ (características);
$\mu$== es la media centroide $i$ de la característica $j$
$\sigma$= La desviación típica del grupo $i$ con con la característica $j$ (previamente calculado).

Si todas las características del objeto $X_{ij}$ se encuentran dentro de los rangos establecidos para la clase $\alpha_{ij}$ el objeto es considerado como parte de dicha clase, de lo contrario el objeto no es considerado parte de la clase. Esta comparación se realiza para todos los objetos de una clase establecida, y al final se debe obtener un porcentaje de los objetos que quedaron dentro y fuera, a continuación se enuncian las reglas para una posible separación de grupos:

- Si el porcentaje de objetos dentro es igual al 60% o más del total de objetos del grupo, entonces no se divide dicha clase y se avanza al paso 11

- De lo contrario si el porcentaje de objetos dentro del rango es menor que el 60 % del total de ellos para esa clase, si se cumple entonces se divide el grupo en dos, siguiendo alguno de los procedimientos que se plantean a continuación:

1. Una posibilidad para el proceso de división es crear dos nuevos centroides, $Zj^+$ y $Zj^-$ a partir de $Zj$, de tal forma que las componentes de los nuevos centroides coincidan con los de $Zj$, excepto la componente con la máxima dispersión, es decir la $Z_k$, siendo la dispersión $\sigma_{jmax}$, entonces los componentes de $Zj^+$ y $Zj^-$ serán:

$$Z_j k^+ = Z_j k + \gamma \sigma_{j\,max} \quad (19)$$

$$Z_j k^- = Z_j k - \gamma \sigma_{j\,max}\,, con\ 0 < \gamma < 1 \quad (20)$$

Lo que se pretende con esta división es distribuir adecuadamente las muestras originales del grupo antes de la división entre los dos nuevos grupos.

2. Por otro lado se tiene la alternativa de división basada en obtener las dos muestras del grupo $\alpha_j$ más alejadas entre sí y con respecto a su centroide, si las muestras obtenidas se representan como $Zj^+$ y $Zj^-$ los dos nuevos centroides se calcularan de la manera siguiente:

$$Z_{j1} = \frac{(Zj^+ + Zj)}{2} \quad (20)$$

$$Z_{j2} = \frac{(Zj^- + Zj)}{2} \quad (21)$$

**Posible unión**

De la matriz calculada en el paso 7 generamos una lista con las distancia entre parejas de grupos. Se comparan las distancias $D_{ij}$ con factor de mezcla $\theta_C$ de forma que se toman, si existen, las $L$ más pequeñas en orden creciente, teniendo:

$$\{D_1, D_2 \dots D_L\}\ con\ D_1 < D_2 < \cdots < D_L$$

**Proceso de unión**

Se comienza con los pares de grupos con las distancias menores, supóngase que se unirán los grupos $i, j$ cuya distancia es $D_{ij}$ encontrada dentro del conjunto $\{D_1, D_2, \dots, D_L\}$ con $D_1 < D_2 < \cdots < D_L$. Sí y sólo sí ninguno de estos dos grupos ha sido fusionado previamente con otro en esta misma iteración, se forma un grupo único cuyo centroide es:

$$Z_{ij} = \frac{1}{N_i + N_j} * \left(N_i Z_i + N_j Z_j\right) \quad (22)$$

Siendo $N_i$ y $N_j$ el número de elementos de los grupos $\alpha_i$ y $\alpha_j$ respectivamente antes de la fusión. En cada unión se actualiza el parámetro $N_C$ ya que el grupo se puede unir una sola vez en cada iteración, generalmente no se obtendrán $L$ uniones en cada iteración.

**Comprobar última iteración**

Se comprueba si se ha llegado a la última iteración, comparando con el parámetro $I$, el caso negativo se vuelve al paso 2 iniciando una nueva iteración.

**Modificación M2 del algoritmo Isodata**

Los parámetros de entrada que son manejados por ésta modificación del algoritmo son:

$N_C$: Número actual de grupos que han sido formados.

$k$: Número deseado o estimado a priori de grupos.

$\theta_N$: Número mínimo de elementos o miembros de un grupo para constituirlo como tal.

ISSN-2410-3993

$L$: Cuando en una iteración genérica del algoritmo existe más de una pareja de grupos susceptibles a unirse, este parámetro limita el número de fusiones que pueden llevarse a cabo en esa iteración.

$I$: Número máximo de iteraciones que puede ejecutar el algoritmo.

El algoritmo ISODATA (Ball G., 1965) con la Modificación 2 se describirá a continuación.

**Inicialización**

Se establecen los valores para los parámetros previamente mencionados, se recomienda asignar $k$ ha $N_C$, se eligen $k$ elementos entre los $P$ elementos a clasificar: $X_1, X_2, \dots, X_P$ formando con cada uno de ellos un grupo inicial. Se tienen entonces los $k = N_C$ centroides $Z_1, Z_2, \dots, Z_{Nc}$.

**Distribuir los elementos entre los distintos grupos**

Se agrupan los elementos $x_1, x_2, \dots, x_P$ entre los $N_C$ grupos ya formados, siguiendo el principio de la mínima distancia euclidiana, empleando la siguiente ecuación:

$$x_j \in \alpha_i \;\; si \;\; \|x_i - Z_i\| mínima \qquad (23)$$
$$\forall j = 1,2 \dots p; \quad \forall i = 1,2 \dots N_c$$

**Eliminar los grupos con un número insuficiente de miembros**

Se procede con la eliminación de grupos que tengan un número de elementos inferior a $\theta_N$, actualizando el parámetro $N_C$, si la eliminación de grupos procede posterior a ésta se debe volver a agrupar esos elementos entre los centroides existentes.

**Actualizar los centroides de los grupos**

La actualización se lleva a cabo calculando la media muestral de cada grupo, empleando la siguiente ecuación:

$$Z_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_j; \qquad i = 1,2 \dots N_c \qquad (24)$$

Donde $N_i$ es el número de elementos de la clase $\alpha_i$.

**Comprobación de bifurcaciones**

- Se comprueba en primer lugar si se trata de la ultima iteración, si es así entonces se hace $\theta_C = 0$ y se avanza al paso 12 (unión de grupos).

- Por otro lado se verifica si es posible unir grupos, considerando si $N_C \geq 2k$, si es así se avanza al paso 12 (unión de grupos).

- Si no se cumple alguna de las condiciones anteriores se prosigue con la secuencia natural que se describe a continuación.

**Cálculo del vector de desviaciones típicas de cada grupo**

Al trabajar con un vector de características n-dimensional, los grupos presentan un vector n-dimensional de desviaciones típicas como se muestra a continuación:

$$\sigma_i = \begin{pmatrix} \sigma_{i1} \\ \sigma_{i2} \\ \cdots \\ \cdots \\ \sigma_{in} \end{pmatrix}; \; \sigma_{ij} = \sqrt{\frac{1}{N_i}\sum_{K=1}^{N_i}\left(X_{kj} - Z_{ij}\right)^2} \qquad (25)$$

De la fórmula anterior donde:

$i$= 1,2,…, $N_c$ (grupos actuales);
$j$= 1,2,…, $n$ (características);
$K$= 1,2,…, $N_i$ (elementos de la clase $\alpha_i$);

La desviación típica de cada grupo ($\sigma_i = (\sigma_{i1}, \sigma_{i2}, \dots, \sigma_{in})$) se almacena de acuerdo a las características empleadas, mas adelante se empleará junto con otros componentes en la fase de división de un grupo.

**Cálculo de la matriz de distancias entre grupos**

En este paso se calculan las distancias entre grupos, es decir obtener las distancias entre todos los grupos actuales, para esto se emplea la siguiente fórmula:

$$D_{ij} = D_{ji} = \|Z_i - Z_j\| \quad (26)$$
$$Con\ i = 1,2 \dots N_c - 1; j = i+1, i+2 \dots N_c$$

De los resultados obtenidos en la matriz se selecciona la mínima distancia entre dos grupos ($D_{min}$) y además se debe calcular el promedio de todas las distancias obtenidas ($\overline{D}$).

**Cálculo del factor de mezcla $\theta_C$**

Una vez obtenido $D_{min}$ y $\overline{D}$ se procede a calcular el factor de mezcla, el cual se obtiene de la siguiente forma:

$$\theta_C = \frac{D_{min} + \overline{D}}{2} \quad (26)$$

**Cálculo de la matriz de distancias entre elementos de un grupo**

Se procede a calcular las distancias entre elementos de un grupo, se obtienen las distancias entre todos los elementos de cierto grupo haciendo uso de la distancia euclidiana, para esto se emplea la siguiente fórmula:

$$D_{ij} = D_{ji} = \|, Z_i - Z_j\| \quad (27)$$

Al obtener las distancias se generara una matriz de distancias entre objetos, la matriz será por cada grupo del conjunto de datos, en la Tabla 2 se muestra la plantilla de la matriz antes mencionada:

| Objeto | 1 | 2 | 3 | 4 | j |
|---|---|---|---|---|---|
| 1 | 0 | $D_{12}$ | $D_{13}$ | $D_{14}$ | $D_{1j}$ |
| 2 | | 0 | $D_{23}$ | $D_{23}$ | $D_{2j}$ |
| 3 | | | 0 | $D_{34}$ | $D_{3j}$ |
| 4 | | | | 0 | $D_{ij}$ |
| i | | | | | 0 |

**Tabla 2** Ejemplo de Matriz de distancia entre objetos

$$\text{Con i} = 1,2 \dots \text{O}_\text{c} - 1; \text{j} = \text{i} + 1, \text{i} + 2 \dots \text{O}_\text{c}$$
$$\text{con O}_\text{C} = \text{numero de elementos del grupo analizado}$$

**Obtención del promedio de distancias entre elementos de un grupo**

Con las distancias obtenidas en el paso 9 se procede a calcular un promedio entre éstas, es decir después de los cálculos se tendrá un promedio por cada grupo del conjunto de datos, para esto empleamos la siguiente ecuación:

$$P_g = \frac{D_{i,j} + D_{i,j+1} + \cdots + D_{i+1,j+2} + \cdots + D_{Oc-2,\ Oc-1}}{2} \quad (28)$$
$$Con\ i = 1,2 \dots O_c - 1; j = i+1, i+2 \dots O_c$$

$$con\ O_C = número\ de\ elementos\ del\ grupo\ analizado\ y\ g = número\ de\ grupo$$

**Posible división de grupos**

Una vez obtenidos los promedios entre objetos por cada grupo se procede a realizar la evaluación para saber si existe la posibilidad de división o no. Por cada grupo se obtendrá la distancia de cada uno de sus objetos a su centroide correspondiente.

Una vez obtenidos estos valores se comienza a evaluar por grupo si la distancia que se obtuvo de cada objeto a su centroide es menor o mayor que el promedio de distancia $P_g$(se obtuvo en el paso 10) para el grupo correspondiente.

A continuación, se debe obtener el porcentaje de los objetos cuya distancia a su centro es menor que $P_g$, una vez calculados estos porcentajes de cada grupo se procede según las siguientes reglas:

- Si la cifra obtenida es 60 por ciento o más el grupo no es propenso a dividirse

- De lo contrario si la cifra es menor al 60 por ciento, el grupo debe dividirse de mediante alguno de los procedimientos que se plantean enseguida:

- Una posibilidad para el proceso de división es crear dos nuevos centroides, $Zj^+$ y $Zj^-$ a partir de $Zj$, de tal forma que las componentes de los nuevos centroides coincidan con los de $Zj$, excepto la componente con la máxima dispersión, es decir la $Z_k$, siendo la dispersión $\sigma_{jmax}$, entonces los componentes de $Zj^+$ y $Zj^-$ serán:

$$Z_j k^+ = Z_j k + \gamma \sigma_{j\,max} \tag{29}$$

$$Z_j k^- = Z_j k - \gamma, con\ 0 < \gamma < 1 \tag{30}$$

Lo que se pretende con esta división es distribuir adecuadamente las muestras originales del grupo antes de la división entre los dos nuevosgrupos.

Por otro lado se tiene la alternativa de división basada en obtener las dos muestras del grupo $\alpha_j$ más alejadas entre sí y con respecto a su centroide, si las muestras obtenidas se representan como $Zj^+$ y $Zj^-$ los dos nuevos centroides se calcularan de la manera siguiente:

$$Z_{j1} = \frac{(Zj^+ + Zj)}{2} \tag{31}$$

$$Z_{j2} = \frac{(Zj^- + Zj)}{2} \tag{32}$$

**Posible unión**

De la matriz calculada en el paso 7 generamos una lista con las distancia entre parejas de grupos. Se comparan las distancias $D_{ij}$ con factor de mezcla $\theta_C$ de forma que se toman, si existen, las $L$ más pequeñas en orden creciente, teniendo:

$$\{D_1, D_2 \dots D_L\}\ con\ D_1 < D_2 < \cdots < D_L$$

**Proceso de unión**

Se comienza con los pares de grupos con las distancias menores, supóngase que se unirán los grupos $i, j$ cuya distancia es $D_{ij}$ encontrada dentro del conjunto $\{D_1, D_2, \dots, D_L\}$ con $D_1 < D_2 < \cdots < D_L$. Sí y sólo sí ninguno de estos dos grupos ha sido fusionado previamente con otro en esta misma iteración, se forma un grupo único cuyo centroide es:

$$Z_{ij} = \frac{1}{N_i + N_j} * \left(N_i Z_i + N_j Z_j\right)$$

Siendo $N_i$ y $N_j$ el número de elementos de los grupos $\alpha_i$ y $\alpha_j$ respectivamente antes de la fusión. En cada unión se actualiza el parámetro $N_C$ ya que el grupo se puede unir una sola vez en cada iteración, generalmente no se obtendrán $L$ uniones en cada iteración.

ISSN-2410-3993

### Comprobar última iteración

Se comprueba si se ha llegado a la última iteración, comparando con el parámetro $I$, el caso negativo se vuelve al paso 2 iniciando una nueva iteración.

### Resultados

### Datos utilizados

En el presente trabajo se emplearon un total de 12 conjuntos de datos en los que se aplicó el algoritmo ISODATA tradicional, la modificación 1 (M1) y modificación 2 (M2) del mismo. Los datos utilizados son descritos por dos características y el número de objetos en cada conjunto varía desde unas pocas decenas hasta miles.

Los distintos conjuntos de datos se sometieron a los tres algoritmos como se ha mencionado.

### Diseño de pruebas

A continuación se explicará cómo se emplearon los conjuntos de datos en cada algoritmo y las modificaciones realizadas en los parámetros, cabe mencionar que los parámetros de número de iteraciones ($I$) y número de fusiones en cada iteración ($L$) se colocaran como valores fijos en 50 y 3 respectivamente.

Adicionalmente los parámetros $\theta_S$ y $\theta_C$ solo aplican para el algoritmo ISODATA tradicional y solo en éste se definirán los valores para estos parámetros; en el caso de $\theta_N$ se establecerá utilizando el 10% y 15 % del total de cada conjunto de datos.

### Evaluación de resultados

Los agrupamientos obtenidos con cada uno de los algoritmos de agrupamiento (ISODATA, ISODATA M1 e ISODATA M2), fueron evaluados utilizando la suma de cuadrados del error (SSE). Ver Tablas 3- 14

| Algoritmo/ Modificación | $N_C=3$, $\theta_N$=60 | $N_C=3$, $\theta_N$=90 | $N_C=5$, $\theta_N$=60 | $N_C=5$, $\theta_N$=90 |
|---|---|---|---|---|
| ISODATA | 31.93394 | 34.717712 | 31.4551357 | 34.717712 |
| ISODATA M1 | 57.96550 | 34.708398 | 57.994967 | 34.708398 |
| ISODATA M2 | 57.9949 | 57.99496 | 56.5691353 | 56.5691353 |

**Tabla 3** Dataset1.txt (599)

| Algoritmo/ Modificación | $N_C=3$, $\theta_N$=90 | $N_C=3$, $\theta_N$=128 | $N_C=5$, $\theta_N$=90 | $N_C=5$, $\theta_N$=128 |
|---|---|---|---|---|
| ISODATA | 29.45267 | 38.07237 | 29.4526789 | 38.0723713 |
| ISODATA M1 | 39.43961 | 37.86860 | 37.88588236 | 37.8686060 |
| ISODATA M2 | 64.7807 | 64.78071 | 64.78071844 | 64.7807184 |

**Tabla 4** Dataset2.txt (849)

| Algoritmo/ Modificación | $N_C=3$, $\theta_N$=60 | $N_C=3$, $\theta_N$=90 | $N_C=5$, $\theta_N$=60 | $N_C=5$, $\theta_N$=90 |
|---|---|---|---|---|
| ISODATA | 17.484 | 22.579314 | 15.988139 | 17.4840136 |
| ISODATA M1 | 31.820 | 31.402071 | 31.402071 | 31.4020715 |
| ISODATA M2 | 31.820 | 31.4020715 | 31.402071 | 31.4020715 |

**Tabla 5** Dataset3.txt (599)

| Algoritmo/ Modificación | $N_C=3$, $\theta_N$=45 | $N_C=3$, $\theta_N$=68 | $N_C=5$, $\theta_N$=45 | $N_C=5$, $\theta_N$=68 |
|---|---|---|---|---|
| ISODATA | 7.087268 | 38.5480 | 7.087268 | 39.008392 |
| ISODATA M1 | 39.00839 | 38.7854 | 38.785422 | 38.785422 |
| ISODATA M2 | 38.7854223 | 38.78542 | 38.7854223 | 38.78542 |

**Tabla 6** Dataset4.txt (450)

| Algoritmo/ Modificación | $N_C=3$, $\theta_N$=80 | $N_C=3$, $\theta_N$=120 | $N_C=5$, $\theta_N$=80 | $N_C=5$, $\theta_N$=120 |
|---|---|---|---|---|
| ISODATA | 27.84606 | 27.84606 | 27.84606 | 27.846066 |
| ISODATA M1 | 45.5001105 | 47.982240 | 45.500 | 45.50011 |
| ISODATA M2 | 45.500110 | 45.50011 | 45.500110 | 47.98224 |

**Tabla 7** Dataset6.txt (800)

ISSN-2410-3993

| Algoritmo/ Modificación | $N_C = 3$, $\theta_N$=4 | $N_C = 3$, $\theta_N$=6 | $N_C = 5$, $\theta_N$=4 | $N_C = 5$, $\theta_N$=6 |
|---|---|---|---|---|
| ISODATA | 28.19670 | 113.95089 | 28.19670 | 45.96554 |
| ISODATA M1 | 42.8228 | 42.82285 | 28.1967030 | 42.82285 |
| ISODATA M2 | 45.96554 | 45.9655498 | 45.965549 | 45.965549 |

**Tabla 8** Dataset5.txt (36)

| Algoritmo/ Modificación | $N_C = 3$, $\theta_N$=220 | $N_C = 3$, $\theta_N$=330 | $N_C = 5$, $\theta_N$=220 | $N_C = 5$, $\theta_N$=330 |
|---|---|---|---|---|
| ISODATA | 19.749312 | 22.180236 | 18.2654703 | 22.25704 |
| ISODATA M1 | 31.55348 | 31.5534 | 31.553486 | 31.55348 |
| ISODATA M2 | 31.5534868 | 31.553486 | 31.553486 | 31.5534 |

**Tabla 9** Dataset7.txt (2200)

| Algoritmo/ Modificación | $N_C = 3$, $\theta_N$=50 | $N_C = 3$, $\theta_N$=75 | $N_C = 5$, $\theta_N$=50 | $N_C = 5$, $\theta_N$=75 |
|---|---|---|---|---|
| ISODATA | 12.95964 | 12.9596 | 12.9596417 | 12.95964 |
| ISODATA M1 | 63.734338 | 18.559544 | 18.559544 | 18.559544 |
| ISODATA M2 | 81.684252 | 68.799705 | 81.68425 | 68.799705 |

**Tabla 10** Dataset8.txt (500)

| Algoritmo/ Modificación | $N_C = 3$, $\theta_N$=16 | $N_C = 3$, $\theta_N$=24 | $N_C = 5$, $\theta_N$=16 | $N_C = 5$, $\theta_N$=24 |
|---|---|---|---|---|
| ISODATA | 2.9853545 | 2.98535 | 2.9853545 | 2.98535 |
| ISODATA M1 | 2.9853545 | 2.985354 | 2.985354 | 2.985354 |
| ISODATA M2 | 9.0381 | 9.754494 | 9.754494 | 9.03817257 |

**Tabla 11** Dataset9.txt (155)

| Algoritmo/ Modificación | $N_C = 3$, $\theta_N$=13 | $N_C = 3$, $\theta_N$=20 | $N_C = 5$, $\theta_N$=13 | $N_C = 5$, $\theta_N$=20 |
|---|---|---|---|---|
| ISODATA | 16.203494 | 16.2034946 | 18.543878 | 16.2034946 |
| ISODATA M1 | 14.58531 | 14.609812 | 14.071601 | 13.863110 |
| ISODATA M2 | 18.5438 | 18.5438789 | 125.126418 | 18.54387 |

**Tabla 12** Dataset10.txt (399)

| Algoritmo/ Modificación | $N_C = 3$, $\theta_N$=40 | $N_C = 3$, $\theta_N$=60 | $N_C = 5$, $\theta_N$=40 | $N_C = 5$, $\theta_N$=60 |
|---|---|---|---|---|
| ISODATA | 35.66549 | 35.66549 | 35.66549 | 35.66549 |
| ISODATA M1 | 28.3235 | 35.6654 | 26.34472 | 31.174204 |
| ISODATA M2 | 35.665492 | 35.665492 | 35.665492 | 35.6654923 |

**Tabla 13** Dataset11.txt (128)

| Algoritmo/ Modificación | $N_C = 3$, $\theta_N$=13 | $N_C = 3$, $\theta_N$=20 | $N_C = 5$, $\theta_N$=13 | $N_C = 5$, $\theta_N$=20 |
|---|---|---|---|---|
| ISODATA | 25.8259 | 25.82594 | 25.825948 | 25.825948 |
| ISODATA M1 | 25.8259 | 25.825948 | 25.82594 | 25.825948 |
| ISODATA M2 | 25.82594 | 25.8259489 | 25.8259489 | 25.825948 |

**Tabla 14** Dataset12.txt (128)

## Conclusiones y trabajos futuros

En este trabajo se presentan dos versiones del algoritmo de agrupamiento Isodata, las cuales no requieren como parámetros de entrada θ_c y θ_s, parámetro de unión de grupos y la desviación estándar respectivamente, las pruebas se realizaron con conjuntos de datos sintéticos de los cuales se conoce el número exacto de grupos que los forman. Se utilizó la suma de cuadrados del error para evaluar la eficiencia de las modificaciones propuestas. Los experimentos realizados con los 12 conjuntos de datos, indican que los resultados son al menos iguales que el algoritmo Isodata original, ya que solo en un caso el algoritmo Isodata obtuvo mejores resultados. También es importante resaltar que la modificación M1 dio mejores resultados que la modificación M2. Aunque pensamos que es necesario realizar más pruebas con conjuntos de datos reales, para contar asegurarmos que las modificaciones propuestas son confiables, para ellos continuaremos realizando pruebas con otro tipo de conjuntos de datos.

## Referencias

Ball G. H., Hall D. J. (1965), Isodata: a method of data analysis and pattern classification, Stanford Research Institute, Menlo Park,United States. Office of Naval Research. Information Sciences Branch.

Ali El-Zaart, (2010), Expectation-maximization technique for fibro-glandular discs detection in ammography images. Comp. in Bio. and Med. 40(4):392-401.

Kohei A., XianQiang Bu. (2007). ISODATA clusteringwith parameter (threshold for merge and split) estimation based on GA: Genetic lgorithm. Reports of the Faculty of Science and Enginneering, Saga University, 36, No. 1, 17-23.

Kaufman L., Rousseeuw P. J. (1989), Finding Groups in Data " An Introducction to Cluster Analysis, Wiley series in probability and Mathematical Statistics.

Jain A.J., Dubes R. C. (1988), Algorithms for Clustering Data, Prentice Hall.

Pavan K., Rao D., Sridhar, Gr.(2008), Automatic Genetation of Merge Factor for Clustering Microarray. IJCSNS International Journal of Computer Science and Network Security Vol. 8, No. 9, 127-131.

Artículo **Revista Tecnología e Innovación**
Diciembre 2015 Vol.2 No.5 958-967

# Desarrollo de un software para la simulación y control de un robot industrial

LAZARO-ARVIZU, Y†, MORALES-CAPORAL, R, ORDOÑEZ-FLORES, R, QUINTERO-FLORES, P y LEAL-LÓPEZ, M

Recibido 5 de Julio, 2015; Aceptado 24 de Noviembre, 2015

## Resumen

Hoy en día la programación de robots hace uso de herramientas de simulación que reproducen la dinámica del robot, tanto para capacitar al personal que lo utiliza previo a su operación, así como para eliminar movimientos erróneos antes de su implementación en el robot real. En algunas aplicaciones académicas, MATLAB® que es un software de desarrollo integrado (IDE) permite cálculos y visualizaciones (graficas) de alta dimensión y excelente resolución, por lo que lo hace una herramienta muy poderosa para poder desarrollar aplicaciones orientadas a la robótica, como es el caso del Robotic Toolbox® con el cual se pueden simular los movimientos de un robot en un entorno gráfico. En este artículo se presenta un conjunto de aplicaciones utilizando esta herramienta con fines didácticos, para diseñar y controlar las trayectorias de un robot KUKA IRB-2600 previo a su implementación y que se visualiza en tres dimensiones. Las aplicaciones propuestas permiten a los alumnos realizar prácticas, para aprender el manejo de un robot y visualizar los movimientos sin necesidad de tener el robot real. Para comprobar el software propuesto, se utilizan comunicaciones OPC (OLE for Process Control) entre el servidor de ABB y el cliente de MATLAB® desarrollado para mover un robot real, en este caso un KUKA IRB-2600. El software desarrollado dispone de un entorno gráfico que facilita la interacción con el usuario.

**Robótica, Simulación, Programación, Docencia.**

## Abstract

Today robot programming makes use of simulation tools that reproduce the dynamics of the robot, both to train personnel using prior to operation and to eliminate erroneous movements prior to implementation in real robot. In some academic applications, MATLAB software is Integrated Development Environment (IDE) enables calculations and visualizations (graphic) High dimension and excellent resolution, making it a very powerful tool to develop robotics applications-oriented, such as Robotic is the case with Toolbox® which can simulate the movements of a robot in a graphical environment. In this article a set of applications is presented using this tool for teaching purposes, to design and control the trajectories of a KUKA robot IRB-2600 prior to implementation and displayed in three dimensions. The proposed applications allow students to do internships to learn how to use a robot and visualize movements without having the real robot. To test the proposed software, communications OPC (OLE for Process Control) between the server and client ABB developed MATLAB to move a real robot, in this case a KUKA IRB-2600 are used. The developed software has a graphical interface that facilitates user interaction.

**Robotics, Simulation, Programming, Teaching.**

**Citación:** LAZARO-ARVIZU, Y, MORALES-CAPORAL, R, ORDOÑEZ-FLORES, R, QUINTERO-FLORES, P y LEAL-LÓPEZ, M. Desarrollo de un software para la simulación y control de un robot industrial. Revista de Tecnología e Innovación 2015, 2-5: 958-967

† Investigador contribuyendo como primer autor.

## Introducción

Durante los estudios de algunas ingenierías, las asignaturas relacionadas con robótica tienen dos objetivos que no siempre son fáciles de combinar. Por un lado, se pretende proporcionar al alumno las bases teóricas de la robótica, entre ellas se explican los sistemas de coordenadas, las trayectorias, la cinemática y dinámica del robot y por otro el control del mismo. [1, 2, 3, 4].

Además, de ser posible se desea que el alumno aprenda a programar un robot en su entorno real, con sus particularidades en sus lenguajes. Una solución aquí propuesta para observar ambos objetivos es el uso de algunas aplicaciones de MATLAB®/ SIMULINK® y otras desarrolladas para robótica, donde de una forma práctica se puede aplicar y visualizar la teoría de robots simulados y reales.

MATLAB® tiene el "Robotic Toolbox®" donde se pueden desarrollar algoritmos teóricos para el modelado de robots, estudio de su cinemática, dinámica y diseño de trayectorias sobre un robot genérico [5, 6].

La aplicación "ARTE®" ("A Robotic Toolbox for Education") [7] añade a la anterior aplicación el modelo en tres dimensiones (3D) de un gran número de robot industriales, con funciones para obtener sus cinemáticas inversas, de forma que los conceptos teóricos pueden ser aplicados a robots industriales.

En el Instituto Tecnológico de Apizaco (ITA) se dispone de dos robots industriales KUKA [8], uno de ellos es un antropomórfico de seis grados de libertad con una capacidad de 16 kg, ubicado dentro del tamaño de los de serie mediano, empleado principalmente en la industria de la automatización para tareas de soldadura, recubrimientos, manipulación de objetos, etc. Y en este caso para fines de investigación y docencia.

Las aplicaciones desarrolladas conseguirán que el alumno pueda comprender la teoría con la herramienta de trabajo que luego va a emplear. Se han realizado pantallas para la visualización del robot en 3D, una aplicación grafica basada en GUIDE® de MATLAB® [9], y simulaciones del robot en 3D basadas en SimMechanic® [10, 11, 12].

La aplicación propuesta está basada en el servidor OPC que dispone ABB [13] y el OPC cliente de MATLAB® [14].

Un servidor OPC es el método de conectividad de datos basado en los estándares más populares del mundo. Es utilizado para responder a uno de los mayores retos de la industria de la automatización: cómo comunicar dispositivos, controladores y/o aplicaciones sin caer en los problemas habituales de las conexiones basadas en protocolos propietarios.

Dicha aplicación permite, para un período de muestreo dado, mostrar y medir las variables de salida y permanentes del robot desde una aplicación OPC cliente.

Se consideran seis variables de salida en la estación robotizada, una para cada eje del robot. Estas variables serán leídas por el OPC servidor y modificadas desde MATLAB® por una aplicación OPC cliente. En la función de visualización en 3D del robot se ha añadido la opción de modificar de forma simultánea el valor de estas variables para definir la posición de los ejes. El procedimiento que el robot ejecuta en su propio lenguaje RAPID lee de forma ininterrumpida las seis variables de los ejes, y se mueve a la posición correspondiente. De esta forma se consigue que la trayectoria del robot sea controlada por la aplicación de MATLAB®, y que el robot y su modelo en 3D sigan la misma trayectoria de forma simultánea.

ISSN-2410-3993

El alumno podrá aplicar esta herramienta tanto en el robot virtual representado con RobotStudio® [15] como el real, ya que el servidor OPC es válido para ambos casos. El alumno puede analizar y validar su programa de control en el robot virtual, y una vez verificado su funcionamiento ejecutarlo y verificarlo en el robot real. En cualquier caso, la aplicación es válida para poder ser aplicada con el robot en modo manual en condiciones de máxima seguridad, ya que si se suelta la tecla de hombre muerto, el robot se para automáticamente.

El resto del artículo tiene la siguiente estructura, en la sección 2 se analizan las aplicaciones en MATLAB® / SIMULINK® relacionadas con robótica, además de la cinemática del robot.

En la sección 3, se analiza las posibles comunicaciones entre un robot y una computadora. En la sección 4, se detallan las aplicaciones realizadas en MATLAB® / SIMULINK®, y en la sección 5 se citan las conclusiones del artículo.

## Sistemas robotizados en MATLAB® / SIMULINK®

MATLAB® / SIMULINK® son dos entornos de cálculo numérico y visualización de datos de la compañía Matworks® [10, 16], con grandes posibilidades para el diseño de sistemas dinámicos y su simulación. Dentro de las aplicaciones comerciales de ambos entornos no existe uno específico para el diseño y simulación de sistemas robotizados, pero diversos autores han desarrollado aplicaciones de software libre usando MATLAB® / SIMULINK® para sistemas robotizados. [17, 18, 19] La aplicación de robótica más conocida para MATLAB® / SIMULINK® es “Robotic Toolbox®" realizada por Peter Corke [6] que desarrolla los contenidos del libro del propio autor [5]. Dicha aplicación dispone de software para robótica fija y móvil.

En la parte de robótica fija propone algoritmos para modelar un robot usando el método de Denavit-Hartenberg, con los que se puede obtener la cinemática directa del robot. La aplicación propone un método general de cinemática inversa para un robot con muñeca esférica, que puede ser aplicada a robots PUMA. Los modelos gráficos de los robots son muy simples, pero genéricos, ya que se representan únicamente los ejes y las uniones entre ellos. En la última versión se ha desarrollado un modelo 3D para el robot PUMA. La dinámica de los robots se representa mediante bloques convencionales de Simulink®.

La mayoría de los fabricantes de robots industriales han desarrollado modelos gráficos en 3D utilizando programas de diseño asistido por ordenador (CAD), y estos modelos en muchos casos son libres y se pueden obtener de internet. Del archivo del robot en los principales formatos de CAD se puede pasar a una estructura compuesta por un fichero “*.xml" con los datos del robot y ficheros “*.stl" con la información gráfica de cada modelo de brazo del robot. Existen aplicaciones de MATLAB® para leer la información de los ficheros “*.stl" y obtener graficas en 3D de los brazos del robot en MATLAB®.

La aplicación “ARTE®" (“A Robotic Toolbox for Education") [7] aprovecha estos ficheros para generar modelos de los principales brazos robóticos del mercado.

Esta aplicación dispone de la cinemática directa, basada en Denavit-Hartenberg, e inversa de un conjunto de robots. También dispone de datos dinámicos de alguno de estos robots. Con ello se puede estudiar los movimientos y trayectorias de un robot industrial viendo su evolución sobre modelos gráficos de 3D.

Además, esta aplicación dispone de algoritmos similares a las instrucciones de movimiento del lenguaje RAPID, para comprender la forma en la que se diseñan estas funciones y su aplicación a robot industriales.

La aplicación grafica ROKISIM (Robot Kinematic Simulation) [20] muestra un modelo grafico en 3D de muchos robots industriales y simula sus movimientos básicos.

La aplicación comercial SimMechanic® [11] para SIMULINK® dispone de una herramienta especial, SimMechanic® Link [12], que permite importar modelos de robot realizados con programas de CAD donde la información de su estructura esta en ficheros "*.xml" y la información sobre sus brazos en ficheros "*.stl". A partir de esta información, la aplicación desarrolla un modelo SIMULINK® (SimMechanic) con la estructura en serie articulación-brazo del robot y un modelo grafico en 3D con la figura del robot. Sobre el diagrama de SIMULINK® se pueden hacer cambios para simular los lazos de control de cada brazo del robot y de esta forma, simular la dinámica del mismo.

Se concluye indicando que existen muchas aplicaciones para modelar y simular un robot fijo usando MATLAB® / SIMULINK®. Algunas son genéricas para cualquier tipo de robot definido con el método de Denavit-Hartenberg y otras aprovechan los modelos de CAD de robot industriales para estudiar la cinemática y dinámica de robots concretos.

## Cinemática del robot

La cinemática del robot estudia el movimiento del mismo con respecto a un sistema de referencia. Así la cinemática se interesa por la descripción analítica del movimiento espacial del robot como una función del tiempo. Pero primordialmente por la orientación y posición del extremo final del robot con los valores que toman sus coordenadas articulares.

La cinemática estudia dos problemas particulares en el robot. El primero de ellos, se conoce como el problema cinemático directo, y consiste en determinar, los valores conocidos de las articulaciones, cual es la posición del extremo final del robot con respeto al sistema de coordenadas que se toman como referencia. [1]

El segundo es el denominado problema cinemático inverso, este resuelve la configuración que debe adoptar el robot para que su extremo alcance una posición y orientación conocida. [1]

## Cinemática directa del robot IRB-2600

Un robot manipulador puede considerarse una cadena cinemática formada por objetos rígidos o eslabones unidos entre sí mediante articulaciones.

Se puede establecer un sistema de referencia fijo situado en la base del robot y describir la localización de cada uno de los eslabones con respecto a dicho sistema de referencia.

Las herramientas matemáticas que se han utilizado para describir la configuración del robot son las matrices de transformación homogénea y el algoritmo de Denavit-Hartenberg [1]. En la tabla I se muestran los parámetros del algoritmo de Denavit-Hartenberg correspondientes a la cinemática directa del robot IRB-2600.

| i | $\theta_i$ | $d_i$ | $a_i$ | $\alpha_i$ |
|---|---|---|---|---|
| 1 | $\theta_i$ -90° | 445 | 150 | -90 |
| 2 | $\theta_2$ | 0 | 700 | 0 |
| 3 | $\theta_3$ | 0 | 0 | 90 |
| 4 | $\theta_4$ | -795 | 0 | -90 |
| 5 | $\theta_5$ | 0 | 0 | 90 |
| 6 | $\theta_6$ | 0 | -85 | 180 |

**Tabla 1** Valores de Denavit-Hartenberg correspondiente al robot IRB-2600.

## Comunicación de la aplicación propuesta con el robot real

La comunicación entre la aplicación de MATLAB® / SIMULINK® y el robot se podrá haber realizado usando múltiples protocolos. Las ventajas e inconvenientes de dos de ellos se resumen a continuación:

Comunicación por OPC: ABB dispone de un servidor OPC que puede ser ejecutado desde una computadora. Esta aplicación detecta los robots conectados a la computadora, ya sean reales o virtuales, y extrae de ellos, en cada período de muestreo fijado, la información del sistema, sus variables de entradas y salidas, y las variables persistentes del programa Rapid del robot. Desde una aplicación OPC® cliente se puede leer y modificar dichas variables. No es preciso que la lectura y escritura estén sincronizadas, cuando se modifique una variable desde el exterior, dicha variable queda modificada en la estación del robot.

La comunicación entre la aplicación de Matlab y el robot de ABB ha sido realizada haciendo uso del servidor OPC proporcionado por ABB [13], y del cliente OPC proporcionado por la aplicación "OPC Toolbox®" de MATLAB® [14]. La comunicación se ha realizado usando variables de salida analógicas virtuales por una razón de seguridad.

La escritura en variables de salida en RAPID es posible en estado manual y automático, mientras que la escritura en variables persistente solo es posible en automático. Con fines educativos y por seguridad, el robot IRB-2600 se suele usar en modo manual, ya que en este modo se puede parar el robot ante cualquier eventualidad dejando de pulsar el botón de hombre muerto. Se han detenido seis variables de salida virtuales normalizadas entre [-100; 100] para transmitir la información de los seis ejes del robot.

La forma de comunicarse entre MATLAB® y el robot IRB-2600 es la siguiente:

MATLAB®: Cuando la aplicación de MATLAB® quiera cambiar la posición del robot debe modificar el valor de las variables del servidor OPC correspondientes a los ejes del robot.

Robot IRB-2600 (RAPID): El programa en Rapid consiste en un bucle infinito que lee las salidas analógicas correspondientes a los ejes y desplaza al robot a la posición definida por los valores de las variables.

El protocolo que se han empleado proporciona independencia frente al tiempo de lectura y escritura.

Si se envía una trayectoria definida por muchas posiciones de ejes con un período de muestreo muy bajo, puede que el programa Rapid no sea capaz de leer algún punto pero llegara al destino sin quedarse bloqueado. La sincronización entre MATLAB® y Rapid depende de la velocidad con que envía los datos a las variables de salida.

ISSN-2410-3993

## Descripción de las aplicaciones

En este trabajo se han desarrollado tres aplicaciones para modelar y simular el robot IRB-2600 y sus movimientos respectivamente con MATLAB® y después poder mover el robot real, o el virtual si no se cuenta con el robot real.

Las dos primeras aplicaciones hacen uso de las funciones de "Robotic Toolbox®" [6] y "ARTE®" [7]. La tercera aplicación, hace uso de las herramientas de SimMechanic® [11, 12].

## Aplicación grafica de simulación

Primero se modela el robot IRB-2600, con su cinemática directa e inversa y se programan una función plot irb2600() para visualizar las posiciones del robot en 3D. La cinemática directa y las trayectorias del robot se han diseñado utilizando la aplicación de [6], mientras que la cinemática inversa ha sido obtenida de la aplicación de [7].

A partir de los ficheros "*.stl" del robot, se ha construido una función plot irb2600() con las siguientes objetivos (Ver anexo). La fig. 1 muestra el robot simulado.

Dibujar el robot en 3D con la posición de ejes deseada. Se puede dibujar una secuencia de posiciones si la entrada es una matriz de posiciones de ejes. Se puede definir un tiempo de parada entre posiciones que va a ser útil para sincronizar el movimiento de la gráfica y el del robot real.

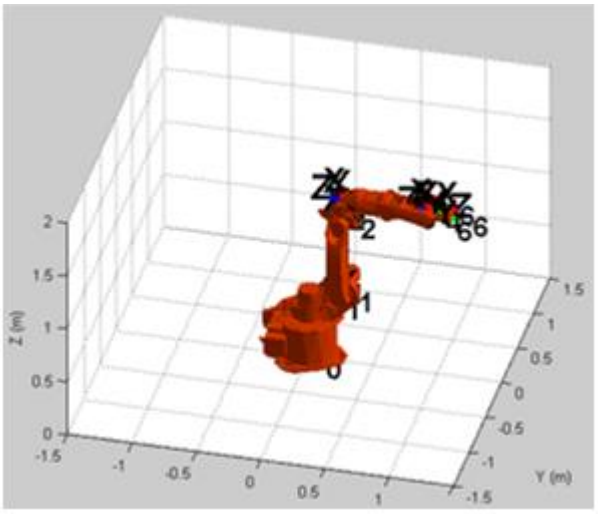



**Figura 1** Robot simulado en 3D.

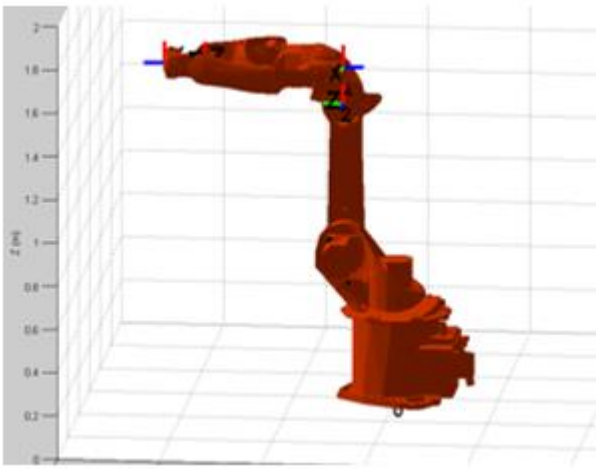


**Figura 2** Posición inicial del robot simulado

**Figura 3** Posición inicial del robot real.

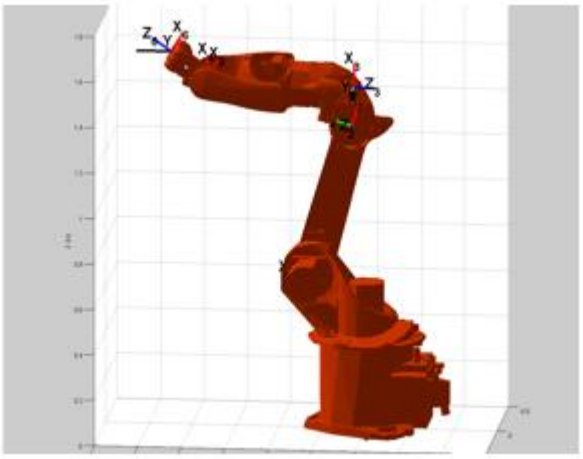


**Figura 4** Posición final del robot simulado.

**Figura 5** Posición robot real.

El usuario puede programar la posición en las variables de entrada del robot correspondiente, usando el servidor OPC de ABB.

Cuando el programa de Rapid se activa, el robot se moverá a las mismas posiciones que el robot simulado. El tiempo de espera entre posiciones es un parámetro de sintonización entre el robot y la simulación.

Las siguientes imágenes muestran las pruebas que se hicieron en la fig. 2 y 3 muestran la posición inicial del robot real y el robot simulado.

Las posiciones finales se muestran en la fig. 4 y 5.

ISSN-2410-3993

## Aplicación grafica (GUIDE®) para el movimiento del robot

A partir de la función plot irb2600() se ha realizado una aplicación grafica en GUIDE® de MATLAB® [9] que permite al usuario realizar de forma segura los movimientos de ejes por ángulos y posiciones cartesianas. Dichos movimientos se visualizan con la función plot irb2600() y podrán ser enviadas a las variables de los ejes del robot a través del servidor OPC. Con ello, el usuario podrá mover el robot con una aplicación grafica de MATLAB®, de forma similar a como se mueve con la aplicación FlexPendant de ABB [21].

Las principales características de esta aplicación son las siguientes:

- Movimiento a una posición en coordenadas articulares. Se puede grabar dicha posición para simular una trayectoria.
- Movimiento a una posición del punto de trabajo.
- Se puede grabar dicha posición para simular una trayectoria
- Movimiento del robot entre las posiciones grabadas anteriormente, ya sean posiciones articulares o del punto de trabajo.
- Movimientos de demostración del robot.
- Conexión con OPC para comunicar datos.

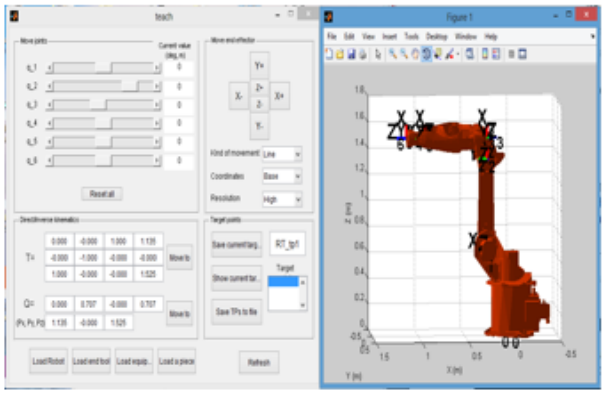


**Figura 6** Pantalla inicial de la aplicación gráfica.

La fig. 6 muestra la pantalla inicial de la aplicación gráfica. Se aprecian los diferentes botones para mover el robot mediante coordenadas articulares, coordenadas de punto de trabajo, y resto de elementos mencionados.

## Aplicaciones didácticas y de investigación

El objetivo de las aplicaciones realizadas en este artículo es servir de puente entre el estudio matemático y simulado de los sistemas robotizados y el práctico de movimiento real de un robot comercial. Se desea que el alumno pueda realizar algoritmos semejantes a los implementados en lenguaje RAPID y otros algoritmos nuevos que el alumno desarrolle.

Como ejemplos de ejercicios realizables con estas aplicaciones se pueden mencionar:

Modelar trayectorias interpolando puntos en el espacio de coordenadas auriculares.

Modelar trayectorias interpolando puntos en el espacio de coordenadas espaciales del punto de trabajo.

Mover el punto de trabajo del robot para minimizar el error entre la posición de dicho punto y una referencia dada, usando un algoritmo de optimización local. Usar para ello incrementos en las coordenadas auriculares.

ISSN-2410-3993

Mover el punto de trabajo del robot en la dirección medida por un sensor de esfuerzos, que el robot IRB-2600 (que actualmente el robot no tiene).

Mover el punto de trabajo del robot hacia una posición móvil detectada por una cámara de visión (proyecto futuro) asociada a una aplicación de MATLAB®.

Programar las trayectorias punto a punto, adicionando al robot una pistola para soldadura y que este pueda realizar las trayectorias repetidamente (En proceso).

## Conclusiones

Se han presentado las propuestas de aplicaciones usando el "Robotic Toolbox®" de MATLAB® / SIMULINK® y su comunicación con el software de ABB para controlar, simular, diseñar e implementar los movimiento de un robot industrial. Esto es posible gracias a que la herramienta de comunicación entre el robot y las aplicaciones de MATLAB® la cual se lleva a cabo por medio de un intercambio de información usando el servidor OPC de ABB y el cliente OPC de MATLAB®. Con esta herramienta los alumnos pueden realizar en un entorno visual la programación previa de los movimientos de cada eslabón del robot para una aplicación específica. Una vez verificado el diseño de los movimientos por simulación, es posible implementarlos de manera rápida en el robot real, en caso de contar con él. Eliminando movimientos riesgosos tanto para el usuario como para el equipo.

## Referencias

Barrientos A., Peñin L.F., Balaguer C. y Aracil P. (2007) Fundamentos de Robótica (2 edición), MacGraw-Hill.

Ollero A. (2001) Robótica, manipuladores y robots móviles, Editorial Macombo.

Pérez Cisneros M.A., Cuevas Jiménez E.V., Zaldívar Navarro D. (2014) Fundamentos de Robótica y Mecatrónica con MATLAB y SIMULINK.
Editorial RA-MA.

Spong M., Hutchinson S y Vidyasagar (2006) M., Robot, Modeling & Control, Wiley.

Corke, P. (2013) Robotics, Vision & Control, Springer.

Peter Corke (2014) Robotics Toolbox. http://petercorke.com/Robotics Toolbox.html.

Gil Aparicio A. (2014) ARTE (A Robotics Toolbox for Education) Universidad Miguel Hernández (Elche, España), http://arvc.umh.es/arte/index en.html

www.kuka-robotics.com/Robotics.

MathWorks (2014). GUIDE Toolbox User's Guide.

MathWorks (2014). SIMULINK User's Guide.

MathWorks (2014). SimMechanic User's Guide.

MathWorks (2014). SimMechanic Link User's Guide (2014).

ABB robotic. Application manual: IRC5 OPC Server. Doc. ID: 3HAC023113-001.

MathWorks (2014). OPC Toolbox User's Guide.

ABB robotic. Manual del operador: RobotStudio. Doc. ID: 3HAC032104-005.

MathWorks (2014). MATLAB® User's Guide.

ISSN-2410-3993

Hassine Belhadj, Saber Ben Hassen, Khaled Kaániche and Hassen Mekki, "KUKA Robot control based Kinect image analysis", in IEEE International Conference on Individual, and Collective Behaviors in Robotics, pp. 21-26.

Rene González Rodriguez and Luis Hernández Santana, "Platform to develop real time visual servoing control in kinematics systems", in revista Ingenieria Mecánica, vol. 15, no. 3, pp. 233-241, 2012. ISSN 1815-5944.

Velásquez Lobo, Ramirez Cortés and Rangel Magdaleno, "Modeling a Biped Robot on Matlab/SimMechanics", in IEEE International Conference on Individual, and Collective Behaviors in Robotics, pp. 203-206.

Bonev I. (2013) ROKISIM (Robot Kinematic Simulation), Escuela de Superior de Tecnología (Quebec, Canadá) http://www.parallemic.org/RoKiSim.html.

ABB robotic. Manual del operador: Introducción a RAPID (RobotWare 5). Doc. ID: 3HAC029364-005

ABB robotic. Manual del operador IRC5 con FlexPendant. Doc. ID: 3HAC16590-5.

Artículo **Revista Tecnología e Innovación**
Diciembre 2015 Vol.2 No.5 968-979

# Adaptación del MMPI Mediante un Sistema Experto en Base a Probabilidades para el Diagnóstico de Desviaciones Psicopáticas en el Instituto Tecnológico de Pachuca

RAMÍREZ-MEJIA J.†, MAGGI-NATALE C., ARRIETA-ZUÑIGA J., HERNANDEZ-RAMÍREZ A. y GONZÁLEZ-MARRON D.

*Instituto Tecnológico de Pachuca*

Recibido 5 de Julio, 2015; Aceptado 24 de Noviembre, 2015

## Resumen

Se presenta un Sistema Experto Probabilístico para determinar y ayudar en el diagnóstico de la desviación psicopática con el fin de apoyar en el proceso de orientación de los estudiantes para complementar su formación. El trabajo hace uso de la técnica del Inventario Multifásico de la Personalidad de Minnesota (SEAD MMPI) en sus diez escalas, evaluando de manera más especifica la desviación psicopática, la cual contiene a su vez cinco subescalas propias, este texto se centra en el estudio y análisis de la alineación social. El sistema Experto utiliza la tecnología de los sistemas expertos clásicos bivalentes, teoremas probabilísticos, incertidumbre, el teorema de Bayes para probabilidades condicionales y la revisión Bayesiana.

**Adaptación del MMPI Mediante un Sistema Experto en Base a Probabilidades para el Diagnóstico de Desviaciones Psicopáticas en el Instituto Tecnológico de Pachuca.**

## Abstract

Is Presented a Probabilistic Expert System to determine and assist in the psychopathic deviation diagnosis to support the process of guiding students to complement their training. The work uses the technique of Multiphasic Personality Inventory Minnesota (MMPI) in its ten scales, evaluating more specific psychopathic deviation that contains itself five subscales, this text talk about the study and social alignment analysis. The expert system uses bivalent classic expert's systems, probabilistic theorems, uncertainty, Bayes´ theorems for conditional probabilities and Bayesian review.

**Adaptation of the MMPI through an Expert System based in Probabilities to Diagnose the Psychopathic Deviations in the Technological Institute of Pachuca.**

**Citación:** RAMÍREZ-MEJIA J., MAGGI-NATALE C., ARRIETA-ZUÑIGA J., HERNANDEZ-RAMÍREZ A. y GONZÁLEZ-MARRON D. Adaptación del MMPI Mediante un Sistema Experto en Base a Probabilidades para el Diagnóstico de Desviaciones Psicopáticas en el Instituto Tecnológico de Pachuca. Revista de Tecnología e Innovación 2015, 2-5: 968-979

† Investigador contribuyendo como primer autor.

## Introducción

El Inventario Multifásico de la Personalidad de Minnesota (MMPI) es una de las pruebas psicológicas (dentro de la categoría de los Inventarios Descriptivos de Personalidad) que se aplica en la práctica clínica para obtener un perfil de personalidad, tanto de los elementos sanos, como de la alteración de un sujeto, y sirve como apoyo para el diagnóstico, pronóstico y tratamiento de las características psicológicas que el MMPI clasifica en 3 escalas (L, F y K) que son de validez del test, y clínicas, que son:

- Escala de Hipocondriasis. (Hs)
- Escala Depresión. (D)
- Escala de Histeria. (Hi)
- Escala de Desviación Psicopática. (Dp)
- Escala de Intereses Masculino-Femenino. (Mf)
- Escala Paranoica. (Pa)
- Escala de Psicastenia. (Pt)
- Escala de Esquizofrenia. (Es)
- Escala de Hipomanía. (Ma)
- Escala Introversión-Extroversión. (Si)

El MMPI consta de una lista de 566 frases expuestas con oraciones declarativas las cuales en su mayoría son afirmativas; la persona a examinar lee las frases y contesta cierto o falso si se aplican en su caso en una hoja de respuestas para su evaluación.

El MMPI evalúa la personalidad basándose en 10 trastornos distintos, de los cuales cada uno puede tener un mínimo de 5 interpretaciones, por lo que podrían darse por lo menos 50 posibles diagnósticos bien definidos, además de existir muchas combinaciones entre los trastornos principales y secundarios que podrían considerarse para el crecimiento de un análisis de éste tipo.

## Antecedentes

El MMPI fue construido en el contexto del Hospital de la Universidad de Minnesota (EEUU), en grupos de pacientes psiquiátricos y no pacientes. Fue publicado por primera vez en 1942. Proporcionaba al usuario datos sobre las llamadas Escalas Clínicas. Así como tres indicadores de la validez de las respuestas de un sujeto: la cantidad de preguntas no respondidas, una estimación de un estilo de respuestas "defensivo" (escala de Mentiras) y una medida de las respuestas extremadamente desviadas o azarosas (escala F). El extendido uso de la prueba y los cambios sociales y culturales producidos enlos 60´s y 70´s hicieron necesario plantearse la necesidad de una reestandarización del Inventario y una adecuación semántica de sus ítems. Esta tarea fue iniciada en 1983 por un equipo de trabajo integrado por los psicólogos James Butcher, W. Grant Dahlstrom, John Graham y Auke Tellegen y culminó con la publicación, en 1989, del nuevo Manual del MMPI-II, editado por la Universidad de Minnesota.

En México, Rafael Núñez y Ofelia Rivera, realizaron la investigación del Inventario Multifásico de la Personalidad de Minnesota (MMPI) en él han hecho adaptaciones con relación a la población mexicana, con lo que han obtenido una herramienta que ha llegado a tener validez, confiabilidad, objetividad y estandarización en nuestro País.

ISSN-2410-3993

## Planteamiento del problema

El MMPI se aplica a muy pocas personas, por lo que no se valora la personalidad de todos los aspirantes a ingresar o del personal que presta sus servicios en el Sistema Nacional de Institutos Tecnológicos, esto debido a que el proceso manual es lento y tedioso (2 a 3 horas) cuando lo realiza un solo Psicólogo, cayendo en la poca aplicación del MMPI, el tiempo de aplicación disminuye con el uso de la tecnología computacional creando una base de conocimientos que contenga las 566 preguntas que conforman el test, y realizando los procesos y algoritmos computacionales necesarios para determinar un resultado rápido y confiable y tener al final un sistema experto clásico y probabilístico que contemple la bivalencia y las ponderaciones probabilísticas de ocurrencia de cada carácter de personalidad y de cada respuesta asignada a las cuestiones que la computadora realizará a los usuarios.

## Objetivo

Disminuir el tiempo de aplicación del Inventario Multifásico de la Personalidad de Minnesota (MMPI) para aumentar el número de personas (alumnos y maestros) a los que se les pueda aplicar el test, de una forma interactiva computacional para tener información histórica involucrada en la toma de decisiones.

## Hipótesis

Es factible desarrollar un sistema experto, rápido y eficiente para interpretar la Alineación Social en su escala de Desviación Psicopática del MMPI, por medio de una base de conocimientos y cálculos probabilísticos.

## Justificación

La evaluación de la personalidad de un individuo es un tema muy especial en el ámbito de la psicología.

La realización de un sistema experto probabilístico constituye una herramienta muy útil en el diagnóstico que un experto en el área pueda determinar hacia una persona, ya que por medio de la tecnología se evitan tiempos de aplicación y tiempos de análisis de los datos entregados por el usuario, y garantizan un mejor estudio de parte del personal experto en la evaluación de la personalidad.

## Metodología

La metodología para la creación de un Sistema Experto se representa como un modelo de "ciclo de vida", donde se reconoce la naturaleza evolutiva del desarrollo del sistema.

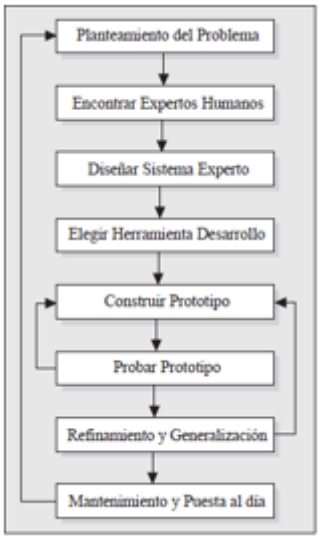



**Figura 1** Diagrama del desarrollo de ciclo de vida

## Planteamiento del Problema

Mediante la investigación de campo, análisis documental, observación, entrevistas y sesiones de "Lluvia de ideas" se procede a identificar los tipos de problema a resolver.

## Encontrar expertos humanos

Reduce la lista de las aplicaciones que recibirán consideración seria.

ISSN-2410-3993

Para seleccionar aquellos candidatos, cada uno de los elementos en la lista original se evalúa con relación a un conjunto de criterios de filtración.

**Diseñar sistema experto**

Una vez que se haya seleccionado un problema, la próxima tarea es diseñar un prototipo que represente una pequeña parte del sistema final.

**Elegir herramientas de desarrollo**

Con el diseño del prototipo correcto se procede a elegir la herramienta mejor adecuada para construir el sistema, es recomendable hacer un análisis de las herramientas actuales de programación y elegir junto con el experto en el área la herramienta que se conozca mejor.

**Construir prototipo**

El ingeniero de conocimiento y el grupo de desarrolladores interpretan las especificaciones funcionales y del diseño del producto, para crear los programas y editar las bases de conocimiento y de datos necesarias en el funcionamiento del Sistema Experto.

De acuerdo con las políticas, estándares y técnicas de programación, se codifican y prueban cada uno de los módulos, depurando las fallas y errores que se detecten.

**Probar prototipo**

Es responsabilidad del grupo de trabajo entregar un producto que cumpla eficazmente con las especificaciones del producto y que procure aprovechar eficientemente los recursos.

Las pruebas se realizan de manera ordenada y con datos preferentemente reales para poder revisar los resultados y detectar posibles fallas en la etapa de construcción.

Si es necesario se debe regresar a la etapa de codificación para solucionar problemas en tiempos de ejecución.

**Refinamiento y Generalización**

Se realizan pruebas de forma más específica y en algunas ocasiones de forma real, puesto en marcha con el cliente final, en esta etapa los errores deben ser mínimos y si es el caso se debe regresar a la etapa de codificación para solucionar contratiempos en la implementación o tiempo de ejecución del sistema.

**Mantenimiento y puesta al día**

Una vez creado el Sistema Experto y habiendo realizado las pruebas pertinentes, se procede a la implementación dentro del sector requerido.

Se elabora la documentación técnica, operativa y promocional del producto y se hace su presentación ante los usuarios e interesados.

Cualquier software mantiene una etapa de mantenimiento donde se pueden solucionar detalles al momento de la codificación o incluso mejorar algunas partes o realizar mejoras a nivel general.

**Factores de certidumbre**

Un factor de certidumbre (F.C.) es un mecanismo relativamente informal para cuantificar el grado al cual, fundamentado en la presencia de un conjunto da¬do de evidencias, se cree o no en una conclusión dada. Los factores de certi¬dumbre se han aplicado ampliamente en dominios donde las evidencias se van recogiendo en forma incremental.

Un F.C. es un valor numérico que expresa el punto al que, basados en un conjunto de evidencias, debemos aceptar una conclusión determinada. Un F.C. con el valor de 1 implica la creencia total.

ISSN-2410-3993

Mientras que un F.C. de 0 indica totalmente lo contrario, la no creencia. Para cada regla en el sistema, el Experto Humano en el dominio asigna un F.C. Un F.C. es una cuantificación subjetiva del juicio y la intuición de un Experto Humano.

En un sistema que emplea factores de certidumbre, existe el principio de que las reglas deben ser estructuradas de manera que dada cualquier regla, o bien aumenta la creencia en una conclusión dada o incrementa la no creencia, es decir las probabilidades de ocurrencia varían conforme la muestra analizada de personas aumenta.

Una medida de creencia MC [c,e] es un número que señala el grado al cual nuestra creencia en una conclusión c se incrementa, fundamentada en la presencia de la evidencia e.

Por definición:

$$0 \leq \mathrm{MC}[c, e] \leq 1$$

En forma semejante, una medida de no creencia, MD[*c*,*e*] es un número que señala el grado al cual se aumenta la no creencia en *c* con base en la presencia de *e*.

En razón de la restricción descrita en el Principio anterior, para cualquier regla dada si:

$$\text{Si } \mathrm{MC}[c, e] = 1, \text{entonces } \mathrm{MD}[c, e] = 0$$

$$\text{Si } \mathrm{MD}[c, e] = 1, \text{entonces } \mathrm{MC}[c, e] = 0$$

El factor de certidumbre acumulativa, que ofrece un medio de establecer la certidumbre de una conclusión desde un punto de vista global, se forma por la combinación de los grados de creencia y no creencia representados por la medida acumulativa de creencia y la medida acumulativa de no creencia para la conclusión.

Específicamente, un factor de certidumbre acumulativo se define, para un punto específico durante la ejecución del sistema, como sigue:

$$\mathrm{FC}[c, e_c] = \mathrm{MC}\,[c, e_f] - \mathrm{MD}\,[c, e_a] \qquad (1)$$

Dónde:

FC*[c,e$_c$]* = el factor de certidumbre acumulativo para *e*, dado *e$_c$* (la certidumbre neta en la conclusión, dada la evidencia actual).
*c* = la conclusión en consideración
*e$_c$* = todas las evidencias relativas a *c*, que se han considerado hasta el momento especificado de la ejecución.
MC *[c,e$_f$]* = la medida acumulativa de creencia para *c*, dado *e$_f$*.
*e$_f$* = toda la evidencia a favor de *c* que se ha considerado.
MD *[c,e$_a$]* = la medida acumulativa de no creencia para *c*, dado *e$_a$*.
*e$_a$* = todas las evidencias contra *c* que se han considerado.

La definición anterior implica la necesidad de calcular MC y MD para cada posible conclusión en el sistema. Este cálculo se realiza primeramente, iniciando ambos términos en cero y luego incluyendo en forma incremental el efecto de cada regla aplicable. Cada vez que se considera una regla adicional, se calcula una nueva MC y una nueva MD sobre la base del efecto de una nueva regla combinada con las actuales MC y MD.

La medida de creencia que resulta de la consideración de dos fuentes de evidencia se puede calcular mediante el empleo de la siguiente fórmula:

$$\text{si: } \mathrm{MD}\,[c, s1\ \&\ s2] = 1$$
$$\text{entonces } \mathrm{MC}\,[c, s1\ \&\ s2] = 0 \qquad (2)$$

si no: $\text{MC}\,[c, s1\ \&\ s2] = \text{MC}\,[c, s1] + \text{MC}\,[c, s2]\,(1 - \text{MC}\,[c, s1])$

Donde:

MC $[c,s_1 \,\&\, s_2]$ = la medida de creencia basada en un par de fuentes.

En el caso elemental, $s_1$ y $s_2$ simplemente son dos reglas individuales $r_1$ y $r_2$. En general, $s_1$ puede representar un conjunto de reglas cuyos efectos acumulativos se han considerado previamente y $s_2$ representa una nueva regla cuyos efectos han de ser agregados a la creencia acumulativa previamente existente. (*MD* $[c,s_1 \,\&\, s_2]$ es la medida de no creencia para el mismo par de fuentes y es igual a 1 si y solamente si la conclusión se conoce como falsa con seguridad absoluta). En forma semejante, MD se define mediante:

$$\text{si: } \text{MC}\,[c, s1\ \&\ s2] = 1 \quad \text{entonces } \text{MD}\,[c, s1\ \&\ s2] = 0 \qquad (3)$$

si no: $\text{MD}\,[c, s1\ \&\ s2] = \text{MD}\,[c, s1] + \text{MD}\,[c, s2]\,(1 - \text{MD}\,[c, s1])$

La razonabilidad de esta función es clara cuando reconocemos que la adición de nueva evidencia que apoya la creencia en una conclusión aumenta la credibilidad de la conclusión, pero no la establece absolutamente. A medida que se combine un gran número de elementos que soporten la evidencia, la MC total crece en forma asintótica hacia la unidad.

El factor: $\text{MC}\,[c, s2]\,(1 - \text{MC}\,[c, s1])$ (4)

De la ecuación 2 describe la contribución a MC que ofrece la nueva pieza de evidencia. Este factor se puede ver como la medida hasta el punto al cual la nueva evidencia mitiga la duda que permanecía después de haber considerado la evidencia previa.

Este grado de atenuación, en forma muy razonable, es proporcional a la fortaleza de la nueva evidencia. Para MD vale el mismo argumento.

Como ejemplo, tómese las cuatro reglas de la tabla 1 que sugieren la conclusión *c* de que una persona tenga Alineación Social para calcular el factor de certidumbre acumulativo, dados los factores de certidumbre componentes:

| Regla | Factor de Certidumbre Componente |
|---|---|
| Estoy seguro de que la vida es cruel conmigo | 1.00 |
| Nadie parece comprenderme | 0.90 |
| Si la gente no la tomara contra mi tendría mas éxito | 0.87 |
| Alguien me tiene mala voluntad | 0.45 |

**Tabla 1** Base de Conocimientos parcial para Alineación Social

Para la regla 1:

MC = MCcomp = 1 MD = MDcomp = 0

La ecuación 2 se emplea para incluir el efecto de la regla 2:

MC = 1 + 0.9(1 – 1) = 1 MD = 0

Después de considerar la regla 3:

MC = 1 + 0.87(1 – 1) = 1 MD = 0

Finalmente se incluye el efecto de la regla 4:

MC = 1 MD = 0.45

(Las respuestas a las tres primeras preguntas son afirmativas y la última es negativa)

Al desarrollar el factor final de certidumbre se tiene:

FC = 1 – 0.45 = 0.55

La figura 2 es una muestra de la interface que arroja los resultados de un Test Diagnóstico y el cálculo del factor de certidumbre (F.C.), el cual se usa en el Teorema de Bayes.

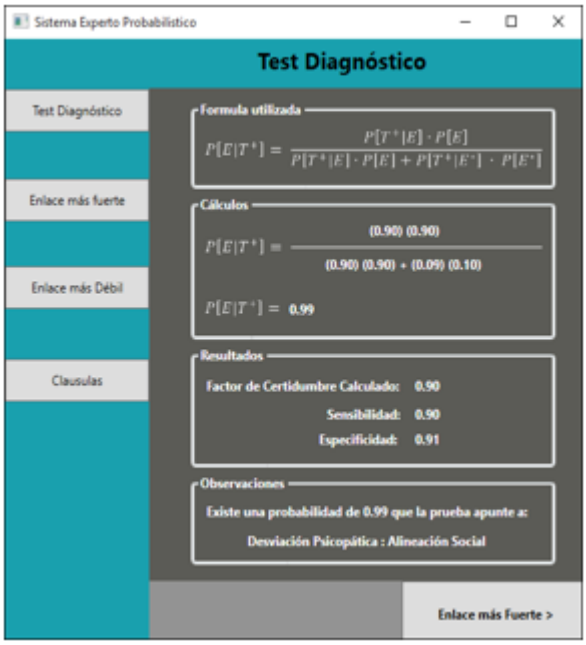



**Figura 2** Pantalla del Sistema Experto Probabilístico con un F.C. calculado

## Ejemplo de Test de Diagnóstico

Con el objeto de diagnosticar una Desviación Psicopática (Alineación Social) se usa un test que consiste de una serie de preguntas de falso y verdadero para determinar:

1. Si a una persona de un determinado núcleo se le aplica el test y da positivo, ¿Cuál es la probabilidad de que tenga una Desviación Psicopática?

2. Si el resultado del test diera negativo ¿Cuál es la probabilidad de que no tenga una Deviación Psicopática?

Solución.

Utilizando la notación:

$$E \equiv Padece\ la\ enfermedad$$
$$E^- \equiv No\ padece\ la\ enfermedad$$
$$T^+ \equiv El\ resultado\ del\ test\ es\ positivo$$
$$T^- \equiv El\ resultado\ del\ test\ es\ negativo$$

Y sabiendo los porcentajes estimados correspondientes resumidos en la tabla 1 al tomar un conjunto de 100 personas sanas y 100 personas enfermas a las que se les aplicó el test.

| | $E$ | $E^-$ |
|---|---|---|
| | | |
| $T^+$ | 91 | 2 |
| | | |
| $T^-$ | 9 | 98 |
| | | |
| | 100 | 100 |

**Tabla 2** Porcentajes estimados de un conjunto de 100 personas sanas y 100 enfermas.

Se tiene:

$$Sensibilidad\ o\ Tasa\ de\ Verdaderos\ Positivos \equiv P[T^-|E] = 0.91$$
$$Especificidad\ o\ Tasa\ de\ Verdaderos\ Negativos \equiv P[T^-|E^-] = 0.98$$
$$Tasa\ de\ Falsos\ Positivos \equiv P[T^+|E^-] = 0.02$$
$$Tasa\ de\ Falsos\ Negativos \equiv P[T^-|E] = 0.09$$
$$Incidencia\ de\ la\ Enfermedad\ en\ el\ Núcleo \equiv P[E] = 0.20$$

## Índice Predictivo de Verdaderos Positivos

$P[E|T^+]$, por el teorema de Bayes es:

$$P[E|T^+] = \frac{P[T^+|E]\cdot P[E]}{P[T^+|E]\cdot P[E]+P[T^+|E^-]\cdot P[E^-]} \qquad (5)$$

$1 - P[T^-|E^-]$ $\qquad$ $1 - P[E]$

$$P[E|T^+] = \frac{(0.91)(0.20)}{(0.91)(0.20) + (0.02)(0.80)}$$

$$P[E|T^+] = 0.9192$$

**Índice Predictivo de Verdaderos Negativos**

$P[E^-|T^-]$, por el teorema de Bayes es:

$$P[E^-|T^-] = \frac{P[T^-|E^-]\cdot P[E^-]}{P[T^-|E^-]\cdot P[E^-]+P[T^-|E]\cdot P[E]}$$

$$1 - P[T^+|E^-]$$

$$P[E^-|T^-] = \frac{(0.98)(0.80)}{(0.98)(0.80) + (0.09)(0.20)}$$

$$P[E^-|T^-] = 0.9775$$

**Revisión Bayesiana**

Cuando se interpreta el resultado de una prueba, el Psiquiatra convierte una probabilidad previa de enfermedad en la probabilidad posterior revisada para la prueba siguiente. El Psiquiatra debe emplear su mejor juicio, con todos los datos disponibles para asignar una valoración razonable a estas probabilidades.

Para determinar la probabilidad de que una persona tenga una Desviación Psicopática, el Psiquiatra debe conocer la sensibilidad y especificidad de la prueba, que en este caso es del 91% y el 98%, respectivamente. Y de la misma forma, se consideran a las 200 personas de las cuales el 50% presenta la enfermedad y el otro 50% no.

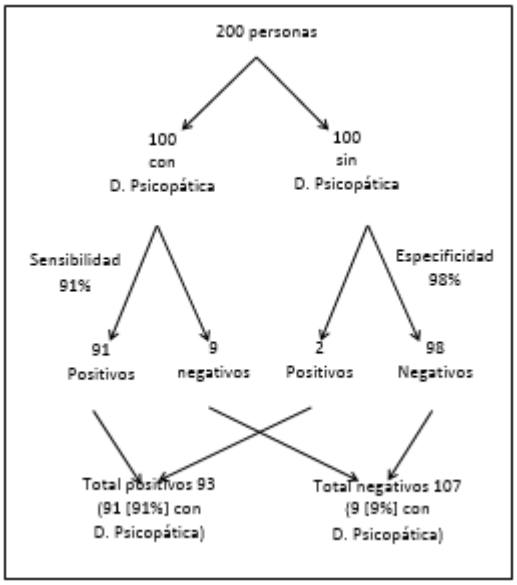



**Figura 3** Interpretación de los resultados de la prueba de D. Psicopática de 200 personas

De acuerdo a la figura 3, los resultados de la prueba son positivos en 91 personas (91% la sensibilidad de la prueba) y en 2 personas (2% la tasa de positivos falsos de la prueba). De las 93 personas del grupo inicial con resultado positivo de la prueba (verdaderos y falsos), 91 (91%) tendrán realmente una desviación psicopática. Así, la probabilidad posterior o revisada tras la prueba de una desviación psicopática positiva es del 91%, haciendo el diagnóstico más probable que improbable.

Si el resultado de la prueba fuera negativo, de las 107 personas con resultados de la prueba negativo (verdaderos y falsos), 9 (9%) tendrían realmente una desviación psicopática.

Así, la probabilidad posterior o revisada de desviación psicopática tras un resultado negativo en la prueba es del 9%, lo que hace que el diagnóstico sea improbable pero no imposible.

| A Diagnóstico | B Probabilidad Previa (%) | C Probabilidad condicional del resultado (%) | D Producto (B x C) | E Probabilidad Posterior (%) |
|---|---|---|---|---|
| | | Resultado Positivo | | |
| D.Psicopática. | 20 | 91 | 1820 | 92 |
| NO Desv. Psicopática. | 80 | 2 | 160 | 8 |
| | | | 1980 total | |
| | | Resultado Negativo | | |
| D.Psicopática. | 20 | 9 | 180 | 2 |
| NO Desv. Psicopática. | 80 | 98 | 7840 | 98 |
| | | | 8020 Total | |

**Tabla 2** Representación de cálculos de probabilidad para pruebas positivas y negativas.

Para demostrar cómo se usa la tabla 2 para revisar las probabilidades, se considera una segunda persona. Se acepta que la probabilidad previa es de aproximadamente el 20%. La mitad superior de la tabla 2 interpreta un resultado positivo en la prueba de la Desviación Psicopática; la mitad inferior interpreta un resultado negativo.

Aunque la sensibilidad y especificidad de esta prueba permanecen invariables (es decir, el 91% y el 98% respectivamente), un resultado positivo en la prueba aumenta la probabilidad de una Desviación Psicopática hasta el 92% o casi la certeza.

Y un resultado negativo la disminuye hasta el 2% o incluso a más probable que al contrario. El proceso de uso de la probabilidad de enfermedad previa a la prueba y las características de la prueba para calcular la probabilidad posterior a la prueba se denomina revisión Bayesiana o teorema de Bayes.

Cuando se deben interpretar varias pruebas, se puede aplicar el teorema de Bayes de modo secuencial, empleando la probabilidad posterior de una prueba como probabilidad previa para la siguiente. Las probabilidades condicionales usadas para interpretar los resultados de la prueba siguiente deben estar basadas en la referencia diagnóstica aceptada y en los resultados observados en la prueba precedente.

**Resultados de las pruebas**

Los datos iniciales que se utilizaron en estas pruebas son:

Desviación Psicopática de Alineación Social con los factores:

Sensibilidad = 0.65
Especificidad = 0.68

| # de pregunta | factor de verosimilitud | no. de pregunta | factor de verosimilitud |
|---|---|---|---|
| 16 | 0.0470 | 202 | 0.0700 |
| 24 | 0.0470 | 275 | 0.0470 |
| 35 | 0.0700 | 284 | 0.3950 |
| 110 | 0.3020 | 291 | 0.0230 |
| 121 | 0.0930 | 293 | 0.1400 |
| 123 | 0.0230 | 338 | 0.6280 |
| 127 | 0.0907 | 347 | 0.5350 |
| 151 | 0.0230 | 364 | 0.0930 |
| 157 | 0.0470 | | |

**Tabla 3** Factores de verosimilitud o probabilidad para Alineación social.

ISSN-2410-3993

El porcentaje de Sensibilidad y Especificidad se calculó por el Experto Humano y el Ingeniero de Conocimiento aplicando la revisión Bayesiana con 40 casos específicos de D.P. y No D.P. en su subescala de Alineación Social, dando como resultado los valores indicados para Sensibilidad y Especificidad.

El factor de verosimilitud se calculó por muestreo y frecuencia de ocurrencia en respuestas afirmativas de 50 casos de estudio con Desviación Psicopática en su subescala de Alineación Social con empatamiento exacto y aproximación.

La tabla 4 muestralos resultados de los 20 casos de estudio probados con el Sistema Experto en la subescala de Alineación Social de la escala 4 de Desviación Psicopática del MMPI.

**Funcionamiento del Sistema Experto Probabilístico**

La primera de las cuatro opciones del sistema es la interface “Adquisición del Conocimiento”, en donde se crea la base de conocimientos proporcionando datos como son la Cláusula (Enfermedad) o carácter de personalidad, su “Sensibilidad” y “Especificidad”; se continúa con los argumentos de la Cláusula o síntomas (Argumentos de la regla), y la Probabilidad de Verosimilitud. Si se desea hacer cambios en alguna Cláusula, sus Argumentos o los parámetros, se escribe el nombre de la Cláusula en el campo correspondiente y el sistema al encontrar que ya existe, cambia la interface para hacer las actualizaciones (Altas, Bajas o Modificaciones).

Para hacer la inferencia a la B. C. Se selecciona la segunda opción del menú y en la interface “Representación del Conocimiento” que aparece se van contestando las preguntas del test (Desviación Social) en forma afirmativa o negativa hasta que se realiza un empatamiento con encadenamiento hacia atrás, y se muestra el diagnóstico; Si no llegara a existir algún empatamiento, como es el caso de contestar ´no´ a por lo menos una pregunta de cada cláusula, el sistema da opción de ingresar una nueva cláusula con los argumentos que se contestaron positivamente. Al ir contestando las frases, se van calculando automáticamente las probabilidades de aceptación/rechazo de los argumentos y cláusulas del test para llegar a utilizarlos en la tercera opción del menú “Incertidumbre y Probabilidad”, en donde se calcula el factor de certidumbre de que se presente la enfermedad o característica de personalidad del individuo utilizando la revisión Bayesiana; en esta misma opción se encuentran los resultados de los métodos probabilísticos de enlace más fuerte y más débil para el diagnóstico como complemento. En la opción cuatro y última del S.E. es donde se ingresan los parámetros iniciales del mismo para su ejecución.

ISSN-2410-3993

| # de prueba | s. e. bivalente | factor de certidumbre | teorema de bayes | observaciones |
|---|---|---|---|---|
| 1 | ENCA-DENA | 0.7809 | 0.88 | Todas las frases afirmativas |
| 2 | | 0.5233 | | |
| 3 | | 0.5286 | | |
| 4 | | 0.5337 | | |
| 5 | | 0.5288 | | |
| 6 | ENCA-DENA | 0.5831 | 0.74 | Todas las frases afirmativas |
| 7 | | 0.6123 | | |
| 8 | | 0.6180 | | Todas las frases negativas |
| 9 | | 0.6222 | | Abajo de rechazo todas positiv. |
| 10 | X | 0.7367 | 0.85 | Arriba de rechazo todas positiv. |
| 11 | | 0.7417 | | |
| 12 | | 0.8053 | | Todas las frases negativas |
| 13 | ENCA-DENA | 0.7349 | 0.85 | Todas las frases positivas |
| 14 | | 0.7704 | | |
| 15 | | 0.7735 | | |
| 16 | | 0.7764 | | |
| 17 | | 0.7794 | | |
| 18 | | 0.7823 | | Todas las frases negativas |
| 19 | | 0.7845 | | |
| 20 | ENCA-DENA | 0.7631 | 0.87 | |

**Tabla 4** Resultados de los 20 casos de estudio para Desviación Psicopática: Alineación social

## Conclusiones

Es importante hacer notar que este Sistema Experto Probabilistico sirve como un apoyo al Experto Humano en la determinación del carácter de personalidad, mas no es determinante, puesto que es necesario dar una interpretación al resultado.

De acuerdo al Experto Humano, el sistema experto funciona, y con base en la tabla de resultados se concluye que:

Si existe un empatamiento exacto en las respuestas de las frases de alguna subescala de Desviación Psicopática, de acuerdo al MMPI y como es el caso de la prueba 6, el individuo tiene Alineación Social.

Si el punto anterior se cumple se refuerza con el Teorema de Bayes para tener una probabilidad más certera de que, si la prueba le resulta positiva, tenga el perfil.

En el caso de Alineación Social, el Factor de Certidumbre P(E) fue aumentando en forma constante, así como la probabilidad por Bayes.

Se hace la observación, por lo tanto, de que el Factor de Certidumbre se considera fiable como la probabilidad a priori de que suceda un evento, en lugar de lo que propone la revisión Bayesiana que dice que se puede considerar la probabilidad calculada (a posteriori) de una prueba como la probabilidad previa de la prueba siguiente. Esta consideración se realizó y se observó que funciona para un número limitado de pruebas; conforme la revisión de pruebas va aumentando junto con sus probabilidades de suceder o no suceder, estas tienden a un 100% de probabilidad y no disminuye. Para que no suceda lo antes mencionado, se recomienda realizar otra revisión Bayesiana como la de la figura 3 conforme las probabilidades aumentan a un 100% y se mantengan constantes por causa de la revisión de un número considerable de pruebas y en consecuencia estar utilizando los mismos factores de Sensibilidad y Especificidad, es fuertemente recomendable que estos dos factores se calculen proporcionalmente al número de pruebas aplicadas.

Como mejora del sistema se remarca la existencia de la versión 2 del MMPI, el cual contiene mejoras en los reactivos a contestar, se añaden mas escalas y el rango de edad aumenta.

ISSN-2410-3993

**Referencias**

Bratko, Ivan. (2011). "PROLOG Programming for Artificial Intelligence". Cuarta Edicion. Addison-Wesley.

Canavos, George C. (2003). "Probabilidad y Estadística Aplicaciones y Métodos". McGraw-Hill.

Chajewska, Urszula. Halpern, Joseph Y. (1997). "Defining Explanations in Probabilistic Systems". Stanford University, Department of Computer Science. Stanford, CA,.

Drakopoulos, John. (1994). "Probabilities, Possibilities, and Fuzzy Sets". Stanford University, Department of Computer Science, Knowledge Systems Laboratory. Palo Alto, C.A.

Gleason, Howard Terrance. (1995). "Probabilistic Knowledge Base Validation", Faculty of the School of Engineering of the Air Force Institute of Technology, Air University In Partial Fulfillment of the Requirements for the Degree of Master of Science.

Hashim, Saffa H. Seyer,Philip. (1998). "Turbo Prolog Advanced Programming Techniques". Tab Books Inc., U.S.A.

Jankowski, Norbert; Gomula, Jerzy; (1997), "Simultaneous Differential Diagnoses Basing on MMPI Inventory Using Neuronal Networks and Decision Trees Methods", Department of Computer Methods & Psychology Outpatient Clinic, Nicholas Copernicus University, ul. Grudziadzka 5,87100 Torún, Teléfono: +4856 6113307,

Luo, Chengjie; Yu, Clement; Lobo, Jorge; (1996), "Computation of Best Bounds of Probabilities from Uncertain Data", Department of Electrical Engineering and Computer Science, University of Illinois at Chicago.

Núñez, Rafael;(1994), "Aplicación del MMPI a la Psicopatología", Tercera Edición, El Manual Moderno, S.A. de C.V., México.

Rodríguez, Alfredo; (1993), "Fundamentos y Práctica de la Construcción de Sistemas Expertos Versión 4.01", Editorial Academia La Habana.

Rolston, David W.; (1994), "Principios de Inteligencia Artificial y Sistemas Expertos", McGraw Hill, USA.

Rosis1, Fiorella de; Grasso2, Floriana; Berry3, Dianne C.; (1997). "Strengthening Argumentation in Medical Explanations by Text Plan Revision", 1 Departimento di Informática, Universita di Bari, Italy, e-mail: derosis@gauss.uniba.it, 2 Department of Computing & Electrical Engineering, Heriot-Watt University Edinburgh, UK., e-mail: floriana@cee.hw.ac.uk, 3 Department of Psichology, University of Reading, UK, e-mail: d.c.berry@reading.ac.uk.

Santos, Eugene Jr.; (1990), "Unifying Time and Uncertainty for Diagnosis", Department of Electrical and Computer Engineering, Air Force Institute of Technology, Wright-Patterson AFB, OH.

Schildt, Herbert; (1990), "Turbo Prolog Programación Avanzada", Primera Edición, McGraw Hill, Mexico.

Siemens, Nixdorf; (1991), "Sistemas Expertos Volumen I y II", Marcombo S.A.

Thurstone, L.L.; (1990), "Inventario de Rasgos Temperamentales", Laboratorio Psicométrico de la Universidad de Carolina del Norte U.S.A.

Peña, Alejandro; (2006), Sistemas Basados en Conocimiento: Una Base para su Concepción y desarrollo., Instituto Politécnico Nacional.

ISSN-2410-3993

Artículo **Revista Tecnología e Innovación**
Diciembre 2015 Vol.2 No.5 980-986

# Metodologías actuales de desarrollo de software

RIVAS, Carlos Ignacio*†, CORONA, Verónica Paola, GUTIÉRREZ, José Fructuoso y HERNÁNDEZ, Lizeth

*Instituto Tecnológico de Pachuca. Felipe Angeles Km. 84.5, Venta Prieta, 42083 Pachuca de Soto, Hgo., México*

Recibido 5 de Julio, 2015; Aceptado 24 de Noviembre, 2015

## Resumen

Las metodologías de desarrollo de software son indispensables para crear o actualizar software de calidad que cumpla con los requisitos de los usuarios; son una parte fundamental de la Ingeniería de software la cual denomina metodología a un conjunto de métodos coherentes y relacionados por unos principios comunes. El objetivo del artículo es brindarle al lector un panorama general de las que existen agrupándolas, de acuerdo a su evolución, al tipo de software por desarrollar, a la forma de generarlo y a su agilidad y prontitud para adaptarse a los cambios tecnológicos. Una contribución del artículo al conocimiento de los desarrolladores de software, se presenta en los resultados, donde se dan recomendaciones para seleccionar la metodología más apropiada. El artículo lo integran tres partes; la primera es la introducción donde se plantea la rápida evolución del software, la enorme demanda de este y la justificación de emplear metodologías de desarrollo del software. La segunda muestra un panorama general de las metodologías existentes y en la tercera están los resultados donde se dan recomendaciones para seleccionar la adecuada.

**Ingeniería de software, metodologías de desarrollo de software**

## Abstract

The software development methodologies are essential to create or update quality software that meets the requirements of users; they are an essential part of software engineering methodology which called a coherent set of methods and related by common principles. The objective of this article is to give the reader an overview of the existing grouping them according to their evolution, the type of software to develop, and how to create agility and readiness to adapt to technological changes. Item contribution to the knowledge of software developers is presented in the results, where recommendations are given for selecting the most appropriate methodology. The article was composed of three parts; The first is the introduction where the rapidly changing software arises, the huge demand for and the justification of using software development methodologies. The second shows an overview of existing methodologies and the third are the results where recommendations are given for selecting the right.

**Software engineering, software development methodologies.**

**Citación:** RIVAS, Carlos Ignacio, CORONA, Verónica Paola, GUTIÉRREZ, José Fructuoso y HERNÁNDEZ, Lizeth. Metodologías actuales de desarrollo de software. Revista de Tecnología e Innovación 2015, 2-5: 980-986

* Correspondencia al Autor (Correo Electrónico: crivaspalacios@yahoo.com.mx)
† Investigador contribuyendo como primer autor.

## Introducción

Desde hace cinco décadas (principios de los 60), la tecnología computacional e informática ha evolucionado a pasos agigantados en el hardware, que son los componentes fisicos y tangibles de los sistemas de cómputo (procesador, memoria RAM, monitor, teclado, disco duro, etcétera), y más aún en el software, que es el conjunto de programas, procedimientos y documentación relacionada que asocia un sistema computacional, específicamente la parte lógica de la computadora (Mclver, 2011).

Por lo que se refiere al hardware, los avances son palpables; nos damos cuenta al observar y operar las máquinas y dispositivos computacionales que están a nuestro alcance, como computadoras, tabletas, teléfonos móviles, televisores, cámaras fotográficas, computadoras de automóviles, computadoras de videojuegos y muchos otros. En 1965, y de acuerdo al vertiginoso desarrollo del hardware, el ejecutivo Gordon Moore, de la empresa fabricante de chips y microprocesadores Intel, observó que cada nuevo chip (monocristal semiconductor que contiene un circuito integrado) de procesador tenía aproximadamente el doble de la capacidad de su predecesor, y que cada nuevo chip, salía al mercado en un plazo de 18 a 24 meses; esto ahora se conoce como la ley de Moore. La tendencia de duplicar la capacidad de procesamiento de cómputo cada dos años continúa en nuestros días y es extraordinariamente precisa, además de que constituye la base para predicciones en la industria de fabricación de procesadores y sistemas computacionales (Mclver, 2011).

Si la evolución del hardware es muy acelerada y la cantidad que se fabrica en la actualidad es enorme, mayor aún es la demanda de software, ya que, cada computadora y dispositivo computacional requiere muchos programas para funcionar.

Asimismo una computadora, ya sea multiusuario, de red o personal, puede atender muchos usuarios, al mismo tiempo que utilizan múltiples aplicaciones de software. De igual forma, hay infinidad de usuarios de empresas, fábricas, instituciones, negocios, gobierno y otros, que emplean software de todo tipo, ya sea empresarial, de propósito general, de propósito específico y particular según sus necesidades. También, la demanda de software crece enormente porque se requiere para profesionistas de diversas disciplinas tales como ingenieros, actuarios, matemáticos, abogados, contadores, comunicadores, médicos, estudiantes de diferentes carreras y para actividades como la educación, la aviación, las ciencias, las finanzas, la cultura y las artes, la medicina, la astronomía, la gastronomía, la hotelería, el gobierno, el transporte, en fín la demanda de software es mucha y diversa.

Ahora bien, ¿quién desarrolla o fabrica el software? Los ingenieros en Sistemas Computacionales, los licenciados en Informática y en general los profesionales de desarrollo de software. Estos deben desarrollar software de calidad que atienda las necesidades y cumpla con los requisitos que los usuarios demandan, y que además que sea amigable, es decir, fácil de usar. Pero crear software es algo muy complejo, sobre todo cumplir con los atributos de calidad que los usuarios (personas, empresas, instituciones) requieren; debido a ello, desde que se inició la fabricación de computadoras y hubo la necesidad de programas para que funcionaran (década de los 60), surgieron también las metodologías de desarrollo de software (MDS), que es el tema central de este artículo.

Las MDS son parte esencial de la ingeniería de software (IS), que es la disciplina profesional que trata fundamentalmente de las actividades llevadas a cabo por personas que producen.

ISSN-2410-3993

Usan o modifican artefactos de software (un artefacto es algo tangible creado con un propósito práctico) (Sánchez, 2012).

Las MDS son indispensables para crear, o modificar software de calidad que cumpla con los requisitos de los usuarios, ya que si no se utiliza la metodología apropiada, seguramente no se alcanzará el objetivo.

El problema actual es que de las diversas MDS que existen no se selecciona la adecuada, y en el peor de los casos no se emplea ninguna, para desarrollar el software que se requiere. Para dar una solución a lo anterior, en este artículo se presenta un panorama general de las metodologías que se pueden utilizar, agrupándolas por tipos de aplicaciones particulares; asimismo, a manera de resultados, se proporcionan algunas sugerencias para seleccionar la adecuada.

La metodología de investigación para obtener los resultados y conclusiones de este artículo consistió en: seleccionar un tema interesante y útil para los profesionistas, académicos y estudiantes del desarrollo de software, investigar cuáles existen, y se agruparon de acuerdo a su evolución, tipos de aplicaciones, prontitud y adaptabilidad de desarrollo, y al final se presentan los resultados de un análisis sencillo para elegir una MDS apropiada.

## ¿Qué son las metodologías de desarrollo de software?

Inicialmente, es importante conocer la definición de metodología y desarrollo. Metodología es una palabra compuesta por tres vocablos griegos: metá (“más allá”), odós (“camino”) y logos (“estudio”); considerando lo anterior, la definición de metodología son los métodos para luego determinar cuál es el más adecuado.

El concepto de metodología es “conjunto de métodos coherentes y relacionados por unos principios comunes”. El concepto de desarrollo, está vinculado a la acción de desarrollar o a las consecuencias de este accionar, por lo tanto es necesario, rastrear el significado del verbo desarrollar: se trata de incrementar, agrandar, extender, ampliar o aumentar alguna característica de algo físico (concreto) o intelectual (abstracto) [1]. Por lo anterior, se concluye que metodología de desarrollo es: el estudio y determinación de cuál es el método más adecuado para dar incremento a algo en este caso al software.

Actualmente el término desarrollo es el más utilizado para referirse a las actividades que involucran la creación, fabricación, actualización o modificación de software.

## ¿Cuáles metodologías existen y cómo se pueden agrupar?

Con base en la información de los cursos de IS impartidos por el autor, en el Instituto Tecnológico de Pachuca, la recopilada por alumnos de la materia, los textos de ingenieria de software y la investigación en internet sobre el tema, a continuación se presentan los nombres de las metodologías que existen y una forma de agruparlas.

### Metodologías clásicas

De acuerdo con Pressman (2010), las MDS clásicas son llamadas también modelos de proceso prescriptivo, y fueron propuestas originalmente para poner orden en el caos del desarrollo de software que existía cuando se empezó a generar masivamente. La historia indica que estos modelos tradicionales, propuestos en la década del 60, han dado cierta estructura útil al trabajo de IS y constituyen un mapa razonablemente eficaz para los equipos de software. Estas MDS son:

- Ciclo de Vida o Cascada
- Incremental
- Evolutivo
- Espiral
- Prototipos
- Desarrollo basado en componentes

En la Figura 1 se muestra un diagrama de la metodología de Ciclo de vida o Cascada

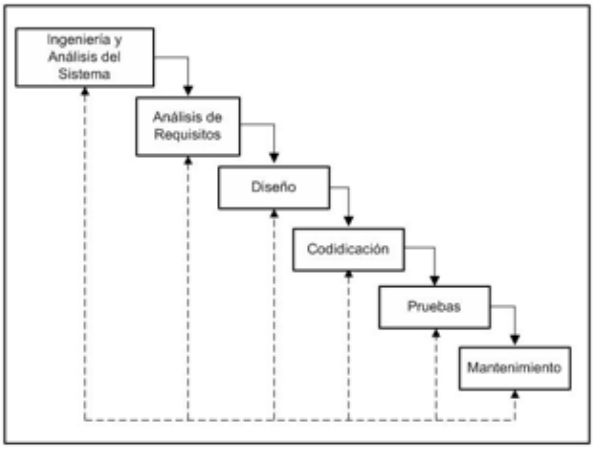



**Figura 1** Diagrama de metodología Ciclo de vida o Cascada

## Metodologías orientadas a objetos

En los años finales del siglo XX se publicaron centeneres de artículos y libros que proponían distintas metodologías, técnicas y notaciones para el desarrollo orientado al objeto. En cuanto al proceso de desarrollo de software, se distinguen tres grandes corrientes:

Metodologías dirigidas por los datos (data- driven), que se basan en la parte estructural de los objetos y son una extensión del modelo conceptual en el modelo Entidad/Relación. Estas son:

- Fusión
- Object Modelling Technique (OMT)

Metodologías dirigidas por las responsabilidades (responsability-driven), que representan el enfoque más purista de la orientación al objeto centrándose en las "responsabilidades" de los objetos, esto es, las acciones que puede llevar a cabo un objeto. Dos de estas son:

- Object Management Facility (OMF)
- Object Management System (OMS)

Proceso de unificado de desarrollo de software (USDP Unified Software Develpment Process): se deriva de la metodología Objectory, de Jacobson; la metodología de Booch; y la técnica de modelado de objetos, de Rumbaugh. (Piattini, 2000).

- Unified Process (UP)

En la figura 2 se muestra un diagrama del Proceso unificado de desarrollo de software.

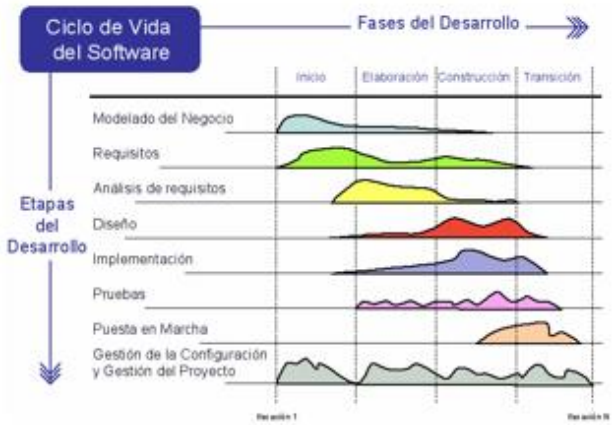



**Figura 3** Diagrama de metodología Proceso Unificado de desarrollo de software

## Metodologías ágiles

Actualmente, las empresas operan en un entorno global que cambia rápidamente; en ese sentido, deben responder a nuevas oportunidades y mercados, al cambio de las condiciones económicas así, como al surgimiento de productos y servicios nuevos y competitivos. Para ello es necesario emplear computadoras y dispositivos computacionales, por lo que el software es partícipe de casi todas las operaciones empresariales, de modo que debe desarrollarse de manera ágil para responder con oportunidad y calidad a todo lo necesario. Estas MDS son:

- Programación extrema (XP), es de las más exitosas y se considera también emergente
- Mobile-D (ágil y extrema para móviles)
- Scrum
- Crystal
- Evolutionary Project Management (Evo)
- Feature Driven Development (FDD)
- Adaptive Software Development (ASD)
- Lean Development

En la Figura 3 se muestra un diagrama de procesos de metodología ágil de desarrollo.


**Figura 3** Diagrama de metodología Ágil de desarrollo

## Metodologías formales

Los métodos formales son soluciones matemáticas para resolver problemas de software y hardware a nivel de requisitos, especificación y diseño.

Generalmente, se puede utilizar la teoría de autómatas para aumentar y validar el comportamiento de la aplicación diseñando un sistema de autómata finito. Los métodos formales suelen aplicarse en software de aviación, especialmente si es progralógica de seguridad crítico (Pressman, 1997).

- Red de Pettri
- RAISE
- Vienna Development Method (VDM)

## Metodologías para la web

El crecimiento desenfrenado que está teniendo la web está ocasionando un impacto en la sociedad, y el nuevo manejo de información en las diferentes áreas ha hecho que las personas tiendan a realizar sus actividades por esta vía. La ingeniería y las metodologías web están relacionadas con el establecimiento y utilización de principios científicos, de ingeniería y gestión, y con enfoques sistemáticos y disciplinados del éxito y desarrollo.

ISSN-2410-3993

Empleo y mantenimiento de sistemas y aplicaciones basados en la World Wide Web de alta calidad. (Pressman, 2010). A continuación se presentan algunas MDS para web:

- Ingeniería web
- Diseño de webapps
- Método de diseño de hipermedios orientados a objetos (MDHOO)

En la figura 4 se muestra un diagrama de metodología de desarrollo para sistemas web.

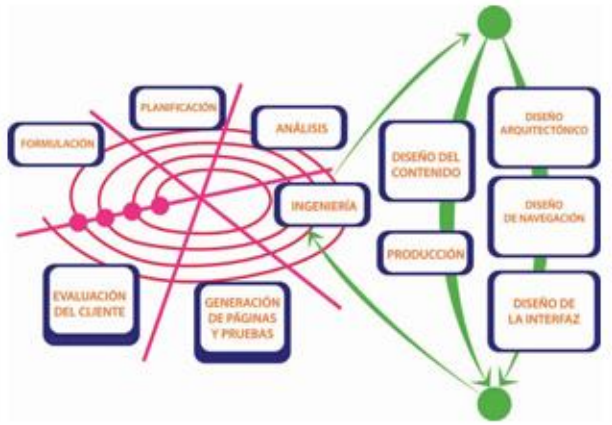



**Figura 5** Metodología Ingeniería Web

## Otras metodologías

Con base en las diferentes apliaciones y tipos de software por desarrollar, otras metodologías son:

- Reingeniería
- Ganar-ganar
- Ingeniería de software distribuido
- Ingeniería para software educativo

En la figura 5 se muestra un diagrama de metodología de Reingeniería de software.

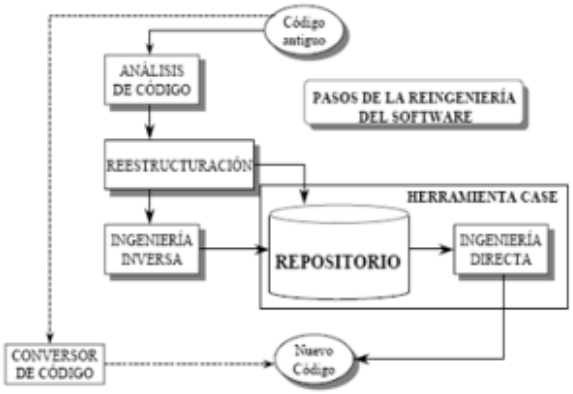



**Figura 6** Diagrama de metodología de Reingeniería de software

## Resultados

Los resultados de la investigación son seleccionar cuál MDS es la apropiada para desarrollar un producto de software. En el apartado anterior se observa que hay diversidad de MDS, por lo que para elegir la adecuada, debe considerarse el tipo de software a desarrollar (que puede ser de propósito general, propósito específico, educativo, tiempo real, etcétera), la premura y el entorno de globalización, los estándares que se emplean en el sitio de desarrollo (empresa consultora de software o lugar de trabajo), los dispositivos donde correrá el software (móviles) o sí el software será para la web.

Por ejemplo, si el desarrollador trabaja por su cuenta, podría escoger la del Ciclo de Vida, que es la más sencilla y sigue todos los pasos formales para obtener un producto de calidad. Para el caso de productos muy grandes y complejos que requieren revisiones por etapas y que el cliente o usuario las apruebe progresivamente, se recomienda utilizar las MDS de Espiral o Evolutiva. Si es el caso de una empresa que ya cuenta con sistemas funcionales pero es necesario actualizarlos debido a nuevas tecnologías computacionales de hardware y software, entonces se recomienda la Reingeniería o el Desarrollo basado en componentes.

ISSN-2410-3993

Si es necesario entregar los productos de software en tiempo breve y que se adapten de inmediato a los cambios de tecnologías, para lograrlo es necesario un equipo de trabajo conformado por individuos muy comprometidos, cuya capacidad y habilidad para colaborar es el fundamento para el éxito del proyecto, entonces es conveniente seleccionar MDS ágiles y extremas. Finalmente, si el objetivo es desarrollar un sistema sencillo o complejo para la web se cuenta, con metodologías altamente especializadas y apropiadas para este fín.

## Conclusiones

Todas las metodologías tienen ventajas que se pueden aprovechar, dependiendo de las condiciones del software que se pretende desarrollar; de igual forma presentan desventajas cuando no se consideran todos los factores que intervienen al realizar el trabajo. Lo importante es utilizar siempre una MDS apropiada, para lo cual, si es la primera vez que se empleará, es necesario conseguir información y documentación sobre ella. En el caso de que ya se haya empleado, lo recomendable es actualizarse y adaptarla lo mejor posible para obtener un producto de calidad que cumpla con los requerimientos funcionales y no funcionales.

## Referencias

Mclver McHoes Ann y Flynn Ida (2011). Sistemas Operativos. México, CENAGE Learning. (6ª. ed.).

Piattini Mario, Calvo-Manzano José y Cervera Joaquín (2000). Análisis y diseño detallado de aplicaciones informáticas de gestión, México, Alfaomega Grupo Editor.

Pressman Roger S. (2010). Ingeniería del software. Un enfoque práctico (7ª. ed.). México: McGraw-Hill Interamericana

Sánchez Salvador, Sicilia Miguel Ángel y Rodriguez Daniel (2012). Ingeniería del Software. Un enfoque desde la guía SWEBOK, México, Alfaomega Grupo Editor.

Pressman, R. S. (1997). Ingeniería del Software: Un enfoque práctico. Mikel Angoar.
Disponible en:
http://books.google.es

http://www.google.com.mx/definición(Real Academia Española RDA)

ISSN-2410-3993

Artículo **Revista Tecnología e Innovación**
Diciembre 2015 Vol.2 No.5 987-997

# Publicación en Internet del inventario de infraestructura física del I.T.P mediante Bases de Datos Geoespaciales y Sistema de Información Geográfica

HERNÁNDEZ, Javier*†, ARRIAGA, Sergio y RERGIS, Raúl

Recibido 5 de Julio, 2015; Aceptado 24 de Noviembre, 2015

## Resumen

El objetivo principal de este trabajo es demostrar que con el uso de la geomática se puede optimizar la manipulación de inventarios de la infraestructura física como mobiliario (sillas, butacas, pizarrones, escritorios) y equipamiento contenido en un inmueble, mediante la integración de Geografía e Informática, denotando la ubicación y capacidad de la infraestructura física con la que cuenta la Institución, además de ayudar en la toma de decisiones para hacer nuevas adquisiciones, renovaciones de material y el mantenimiento de dichos bienes. El sistema desarrollado es una herramienta construida con software libre basado en la geomatica, que le permite al Instituto Tecnològico de Pachuca, la gestión rápida y eficaz del inventario de infraestructura física mediante consultas geoespaciales por parte de los usuarios, a través de bases de datos que describen las propiedades y atributos de los objetos y el uso de las coordenadas para la localización espacial, lo hace un modelo que asemeja el escenario real.

**Infra estructura Fisica, Sistema de Información Geografica, SIG, Aplicaciones de la Ingenieria**

## Abstract

The main objective of this dissertation is demonstrate that the use of geomatics can be optimized the handling of physical inventories of infrastructure such as furniture (chairs, chairs, blackboards, desks) and equipment contained in a building, through the integration of Geography and Informatics denoting the location and capacity of the physical infrastructure of the institution, and help in the take descicions to make new acquisitions, renovations and maintenance material for such property. The developed system is a tool constructed with free software based in the geomatica, which it allows him the Institute Tecnològico of Flashily dressed, the rapid and effective management of the inventory of physical infrastructure by means of consultations geoespaciales on the part of the users, across databases that describe the properties and attributes of the objects and the use of the coordinates for the spatial location, it is done by a model who makes alike the royal scene.

**Infra structures Physics, system of Geographical information, SIG, Applications of the Engineering**

**Citación:** HERNÁNDEZ, Javier, ARRIAGA, Sergio y RERGIS, Raúl. Publicación en Internet del inventario de infraestructura física del I.T.P mediante Bases de Datos Geoespaciales y Sistema de Información Geográfica. Revista de Tecnología e Innovación 2015, 2-5: 987-997

* Correspondencia al Autor (Correo Electrónico: planeación@itpachuca.edu.mx)
† Investigador contribuyendo como primer autor.

## Introducción

En el Instituto Tecnológico de Pachuca, no se tiene un sistema informático para la gestión del inventario de la Infraestructura física, el cual consiste en saber cuántos bienes muebles e inmuebles existen, que superficie, dimensiones, uso, que status tienen. Este inventario en un sistema informático, deberá dar respuestas a preguntas como, ¿cuál es la superficie del área verde?, ¿cuántas aulas están disponibles en ciertos edificios académicos?, ¿cantidad de cubículos de los profesores?, entre otras.

Para lograr estas respuestas se hace uso de los Sistemas de Información Geográfica, los cuales nunca han sido utilizados en objetos de Infraestructura física como son los edificios, banquetas, áreas verdes, líneas hidráulicas, aulas, etc.

Nuestro objetivo principal, es demostrar que por medio de bases de datos geoespaciales en un Sistema de Información Geográfica (SIG) se puede obtener de manera eficiente consultas espaciales a infraestructuras físicas. Y que ayude al manejo de inventarios y/o la toma de decisiones.

Este inventario es una necesidad administrativa ya que sirve para planear y controlar los bienes que se tienen, estos requieren mantenimiento, ampliaciones y su uso es constante.

Es por ello que surge la necesidad de investigar y desarrollar un SIG para inventarios de infraestructuras físicas que nos permita realizar análisis espaciales, tales como mediciones de áreas, longitud o distancia, creación de buffers. Todo esto para saber ¿qué edificios son los más cercanos a un lugar?, ¿Cuál es la ruta más corta para llegar a un aula determinada?, ¿Cuántos pizarrones se encuentran en un edificio?, ¿Cuál sería el lugar indicado para construir un nuevo edificio?

H1: Aplicando las herramientas geomáticas de software libre se puede publicar la base de datos geoespacial, obteniendo una mejor administración de los datos.

H2: Utilizando el análisis espacial en el Sistema de Información Geográfica, el área administrativa puede conocer a detalle la infraestructura física del Instituto Tecnológico de Pachuca, realizando eficientemente la toma de decisiones.

H3: Aplicando las bases de datos geoespaciales, el Instituto Tecnológico de Pachuca podrá conocer su infraestructura en cuanto a aulas, mobiliario, áreas deportivas, áreas verdes, etc.

## Desarrollo

El sistema es desarrollado a base de tecnología de software libre esto implica un menor costo de implementación, lo que ayuda a obtener un producto de buena calidad, sin tener que hacer gastos excesivos en comparación con software patentado. Beneficia a los usuarios en la elaboración de inventarios, alojándolo en una plataforma web, obteniendo beneficios de accesibilidad en diferentes lugares y dispositivos.


**Figura 1** Partes de un SIG

Los SIG son importantes porque integran información espacial y no espacial en un sistema simple, ofreciendo un marco consistente para el análisis de los datos geográficos.

El objetivo general de los SIG es generar información válida para la toma de decisiones. Los objetivos específicos son manejar bases de datos grandes y heterogéneas referenciadas geográficamente, interrogar a las bases de datos sobre la existencia de ciertos fenómenos (qué sucede, en dónde y cuándo), permitir la interacción en forma flexible del sistema y el intérprete, incrementar el conocimiento sobre el fenómeno estudiado e implementar modelos sobre su comportamiento.

Probablemente la parte más importante de un sistema de información geográfica son sus datos. Los datos geográficos y tabulares pueden ser adquiridos por quien implementa el sistema de información, así como por terceros que ya los tienen disponibles.

El sistema de información geográfica integra los datos espaciales con otros recursos de datos y puede incluso utilizar los manejadores de base de datos más comunes para manejar la información geográfica.

Los datos geográficos son entidades espacio–temporales que cuantifican la distribución, el estado y los vínculos de los distintos fenómenos u objetos naturales y sociales. Un dato se caracteriza por tener:

- Posición absoluta: sobre un sistema de coordenadas (x, y, z).
- Posición relativa: frente a otros elementos del paisaje (topología, incluido, adyacente, cruzado, entre otros).
- Figura geométrica que lo representa (punto, línea, polígono).
- Atributos que lo describen (características del elemento o fenómeno).

Los datos geográficos son la clave para diferenciar un SIG de otro sistema de información. Además, antes de discutir operaciones SIG, se debe comprender la naturaleza de los datos geográficos; por ejemplo, si tomamos el elemento vías, podemos referirnos a su ubicación con la pregunta ¿dónde está? Y a sus características, como longitud, nombre, límite de velocidad y dirección.

**Componentes de los datos geográficos**

Los datos geográficos cuentan con tres componentes que hacen referencia a su localización, atributos y la variable tiempo, conozcamos sobre cada una de ellas:

Componente espacial: Hace referencia a la localización geográfica, las propiedades espaciales de los objetos y las relaciones espaciales que existen entre ellos (Gutiérrez y Gould, 1994). En la tabla 1 se muestran los elementos de los componentes espaciales.

ISSN-2410-3993

| Elementos del componente espacial | Descripción |
|---|---|
| Localización geográfica | La localización geográfica o posición de los objetos en el espacio se expresa mediante un sistema de coordenadas, que debe ser el mismo para las distintas capas o "estratos de la información", como se está presentando la realidad del área de estudio. |
| Propiedades espaciales | Los objetos que representan la realidad tienen ciertas propiedades espaciales; por ejemplo, para una línea, longitud, forma, pendiente y orientación. |
| Relaciones espaciales | Los objetos espaciales mantienen relaciones entre sí basadas en el espacio, como conectividad, contigüidad, proximidad, etc. |

**Tabla 1** Elementos de los componentes espaciales.

Componente temática: Son las características que se conocen como atributos de los objetos con los que representamos el mundo real. Cada objeto puede registrar un determinado valor para sus atributos (variables), los cuales pueden presentar cierta regularidad en el espacio y en el tiempo y, además, pueden ser de distinto tipo y escala de medida (Gutiérrez y Gould, 1994).

Los atributos se expresan como variables, que pueden ser:

- Continuas: Es decir, que admiten cualquier valor en un rango.
- Discretas: Son aquellas que sólo admiten valores en números enteros.
- Fundamentales: Se obtienen directamente del proceso de medición. Por ejemplo, población.
- Derivadas: Se obtienen al relacionar dos o más variables fundamentales. Por ejemplo, densidad de la población.

Para que las variables (atributos) puedan ser almacenadas en un SIG, deben ser descritas mediante categorías.

Componente temporal: La consideración de la dimensión temporal en un SIG supone la necesidad de almacenar y tratar grandes volúmenes de datos, ya que cada estrato, capa o nivel de información se debe almacenar tantas veces como momentos temporales se consideren para el análisis del área de estudio (Gutiérrez y Gould, 1994).

Para proceder a aplicar con éxito los diferentes procesos de análisis espacial, se requiere que la información georreferenciada presente calidad en términos de accesibilidad, integridad, precisión, actualidad, consistencia, fuentes de información y procesos de producción.

Una de las herramientas elementales de los SIG`S es el hardware ya que con el se permite la obtención de coordenadas geográficas de un modo inmediato, con las consecuencias que esto tiene para su uso en actividades como la elaboración de cartográfia. Un ejemplo muy común para la obtención de coordenadas geográficas son los Sistemas Globales de Navegación por Satélite (GNSS), es un sistema que permite conocer en todo momento y en cualquier punto del globo la localización exacta de dicho punto con un margen de error del orden de unos pocos metros o menos. Para ello, se basan en él envió de señales entre un dispositivo situado en el punto concreto y una red de satélites, pudiendo establecerse la posición exacta mediante las características de dicha transmisión. El ejemplo más extendido de un GNSS es el Sistema de Posicionamiento Global (GPS). Que se divide en 3 subsistemas o segmentos.

ISSN-2410-3993

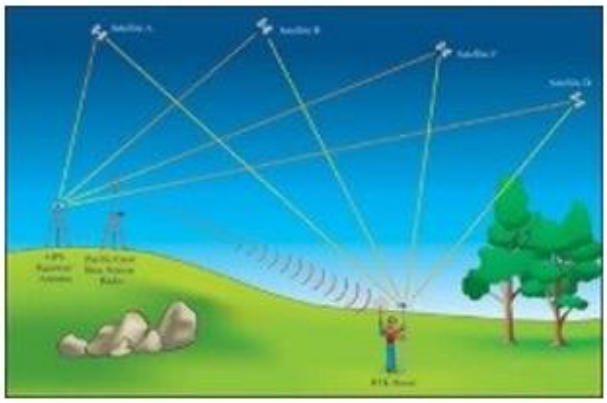


**Figura 2** Triangulacion GPS

Segmento espacial: Lo componen los satélites de la constelación GPS (un total de 27, siendo 24 de ellos operativos y 3 de reserva), con los cuales se comunican las unidades receptoras, y en función de los cuales puede triangularse la posición actual de estas. En la figura 2 se ilustra la triangulación que realizan los GPS.

Segmento de control: Lo forman un conjunto de estaciones terrestres que controlan el funcionamiento de los satélites, pudiendo enviar señales a estos para modificar su comportamiento. La figura 3 ilustra el segmento de control.


**Figura 3** Segmento de control GPS

Segmento de usuarios: Lo conforman los receptores GPS y todos los dispositivos que hacen uso de la señal de los satélites para el cálculo de posiciones. En la figura 4 observamos algunos de los receptores GPS

**Figura 4** Receptores GPS

El funcionamiento del sistema se basa en la triangulación de la posición mediante las señales procedentes de un cierto número de los satélites. Esta posición se calcula no únicamente en sus coordenadas x e y, sino también en z, es decir en elevación. El sistema GPS emplea como sistema geodésico de referencia el WGS84. La precisión en el cálculo de la elevación es menor que la correspondiente a las restantes coordenadas, aunque también es de utilidad y puede emplearse en aplicaciones que van desde levantamientos y replanteos a usos en tiempo real como el cálculo de elevación en vuelos.

## Modelos de Sistemas de Información Geográfica.

En un principio los SIG usaron estructuras de almacenamiento vectorial muy simples como la spaguetti y el diccionario de vértices que no lograban manejar relaciones topológicas.

## Spaguetti

Para cada objeto espacial se registra su identificador, seguido por una lista d coordenadas de los vértices (puntos) que definen su posición en el espacio. Posee desventajas como: El sistema almacena información sobre la localización de los elementos, pero no sobre las relaciones que existen entre los elementos; es decir se registra la geometría pero no la topología.

ISSN-2410-3993

También esta estructura de datos genera mucha información redundante (ej. registra dos veces las coordenadas de un lado común de dos polígonos).

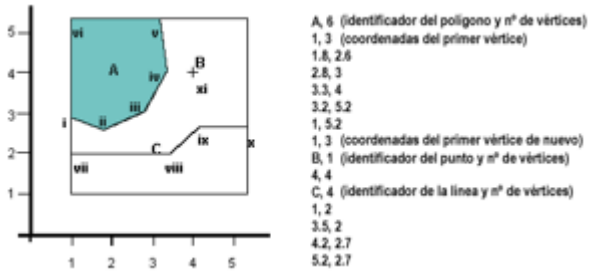



**Figura 5** Modelo Spaguetti

## Diccionario de Vértices

Un mapa se representa mediante dos archivos de datos: Un archivo está constituido por una relación de vértices, en la que constan las coordenadas X, Y, y otro archivo con los vértices que definen cada objeto. Esta estructura resuelve los problemas de repetición de coordenadas de los puntos, de la estructura Spaghetti; pero es muy pobre desde el punto de vista topológico.

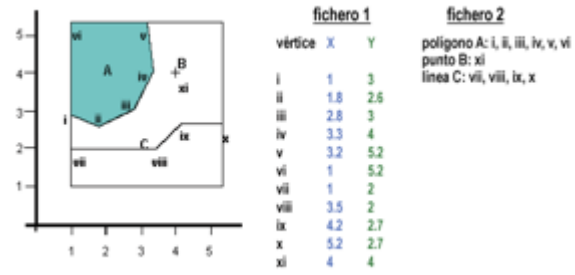



**Figura 6** Modelo Diccionario de Vértices

Hoy en día se usa la estructura arco-nodo, en la cual el sistema puede identificar relaciones como la inclusión, adyacencia, etc.

## Vectoriales

La mayoría de los elementos que existen en la naturaleza pueden ser representados mediante formas geométricas e información los primeros se representan por medio de (puntos, líneas o polígonos, o bien conocidos como vectoriales) mientras que los otros utilizan celdas o pixeles que contienen la información de dicho elemento (raster). Ambas formas nos ayudan a ilustrar de una manera más versátil el espacio y con ellos podemos comprender y analizar mejor los elementos estudiados.

La diferencia entre ambos es clara, mientras que en el tipo vectorial se trabaja con líneas que crean polígonos, con el tipo raster se trabaja con una matriz para representar el terreno.

El modelo vectorial es una estructura de datos utilizada para almacenar datos geográficos. Los datos vectoriales constan de líneas o arcos, definidos por sus puntos de inicio y fin, y puntos donde se cruzan varios arcos, los nodos. La localización de los nodos y la estructura topológica se almacena de forma explícita. Las entidades quedan definidas por sus límites solamente y los segmentos curvos se representan como una serie de arcos conectados. El almacenamiento de los vectores implica el almacenamiento explícito de la topología, sin embargo solo almacena aquellos puntos que definen las entidades y todo el espacio fuera de éstas no está considerado, a estos datos se les pueden asignar diversas propiedades, cualitativas o cuantitativas.

Un SIG vectorial se define por la representación vectorial de sus datos geográficos. De acuerdo a las peculiaridades de este modelo de datos, los objetos geográficos se representan explícitamente y, junto a sus características espaciales, se asocian sus valores temáticos.

ISSN-2410-3993

Dicha estructura se basa en puntos elementales. Se pueden representar de muchas maneras, pero por lo general se suele hacer en una estructura de arco-nodo. En una estructura de datos de arco-nodo, los objetos en la base de datos se estructuran jerárquicamente. En este sistema, los puntos son los elementos básicos elementales. Los arcos son los segmentos lineales individuales que se definen mediante una serie de pares coordenados x-y. Los nodos se encuentran en los extremos de los arcos y forman los puntos de intersección entre los arcos. Los polígonos son áreas completamente limitadas por una serie de arcos. Los nodos son compartidos por los arcos y los polígonos contiguos. Las estructuras arco-nodo permiten la codificación de la geometría de los datos sin redundancia. Contrariamente a lo que sucede con la estructura total del polígono, los puntos se almacenan sólo una vez (figura 3).

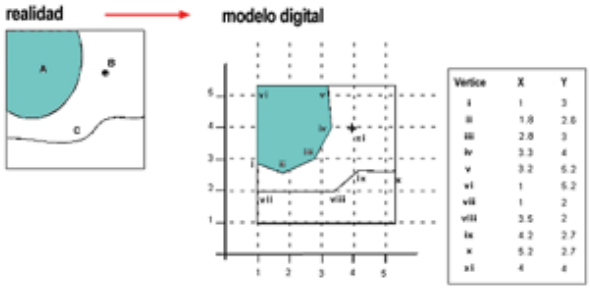



**Figura 7** Modelo de datos vectorial

Las unidades básicas de información geográfica en los datos vectoriales son puntos, líneas (arcos) y polígonos. Cada una de éstas se compone de uno o más pares de coordenadas, por ejemplo, una línea es una colección de puntos interconectados, y un polígono es un conjunto de líneas interconectadas.

**Coordenada**

Pares de números que expresan las distancias horizontales a lo largo de ejes ortogonales, o tríos de números que miden distancias horizontales y verticales, o n-números a lo largo de n-ejes que expresan una localización concreta en el espacio n-dimensional. Las coordenadas generalmente representan localizaciones de la superficie terrestre relativas a otras localizaciones.

**Punto**

Abstracción de un objeto de cero dimensiones representado por un par de coordenadas X, Y. Normalmente un punto representa una entidad geográfica demasiado pequeña para ser representada como una línea o como una superficie; por ejemplo, la localización de un edificio en una escala de mapa pequeña, o la localización de un área a la que una instalación da servicio en una escala de mapa media.

**Línea**

Conjunto de pares de coordenadas ordenados que representan la forma de entidades geográficas demasiado finas para ser visualizadas como superficies a la escala dada (curvas de nivel, ejes de calles, o ríos), o entidades lineales sin área (límites administrativos). Una línea es sinónimo de arco.

**Polígono**

Entidad utilizada para representar superficies. Un polígono se define por las líneas que forman su contorno y por un punto interno que lo identifica. Los polígonos tienen atributos que describen al elemento geográfico que representan.

ISSN-2410-3993

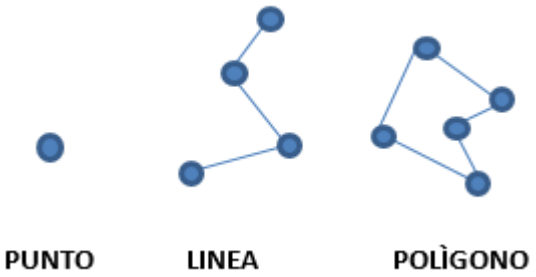



**Figura 8** Unidades básicas vectoriales

## Raster

Este tipo de SIG por su parte, se caracteriza porque la representación de la información no se realiza por medio de puntos, líneas o polígonos, sino por celdas o píxeles.

Los SIG tipo raster consisten en un conjunto de mapas individuales, todos referidos a la misma zona del espacio y todos ellos representados digitalmente en forma raster, es de decir, utilizando una rejilla de rectángulos de igual tamaño. En cada uno de estos rectángulos o posiciones un número codifica el valor que alcanza en ese punto (pixel) del espacio la variable cartografiada en el mapa. Considera la realidad como un continuo en el que las fronteras son la excepción y la regla la variación continua. La representación se realiza dividiendo ese continuo en una serie de celdas o pixeles y asignándole a cada una un valor para cada una de las variables consideradas. Cada píxel contendrá una información única. Los cambios de escala se reflejan en el tamaño de las celdas ya que el tamaño o resolución de la celda o cuadrícula variará dependiendo de la precisión de los datos y los requerimientos del estudio. En general, cuanto más pequeña sea la resolución, mayor será la exactitud de los datos, pero a su vez mayores serán los requerimientos de memoria.

Una serie de celdas raster se llama tessela. Un conjunto de celdas de igual valor se llama zona. Un conjunto de zonas se llama clase.

Conclusión de la elección de un modelo de inforcion geográfica. La elección de un modelo u otro dependerá de si las propiedades topológicas son importantes para el análisis. Sí es así, el modelo de datos vectorial es la mejor opción, pero su estructura de datos, aunque muy precisa, es mucho más compleja y esto puede ralentizar el proceso. Por ello, si el análisis que nos interesa no requiere acudir a las propiedades topológicas, es mucho más rápido, sencillo y eficaz el uso del formato raster.

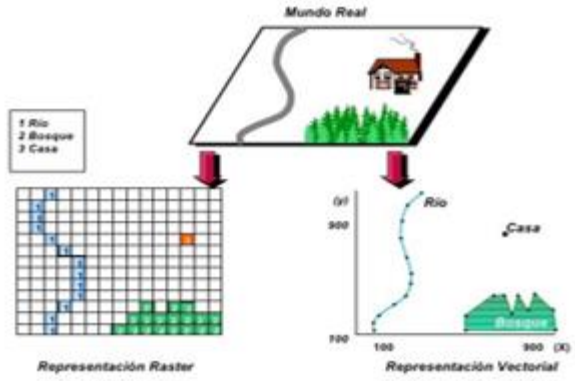



**Figura 9** Representación raster y vectorial del mundo real.

## Metodología a emplear

Es importante contar con un método bien fundamentado para llevar por buen camino el desarrollo del sistema. Por lo cual, para el desarrollo de éste proyecto se decidió utilizar una metodología ágil, pues es importante realizar un buen trabajo pero además se debe estar consciente de que en todo momento ocurrirán cambios y que debe desarrollarse en el menor tiempo posible y con un mínimo de errores.

ISSN-2410-3993

Para nuestro desarrollo utilizaremos la metodología de Programación Extrema (XP) ya que se diferencia de las metodologías tradicionales principalmente en que pone más énfasis en la adaptabilidad que en la previsibilidad, las fases de XP son:

1. Planeación
2. Diseño
3. Codificación
4. Pruebas

Se utiliza un método ágil en lugar de un tradicional ya que estos responden rápidamente a los cambios que puedan surgir durante el desarrollo de proyecto, además de que estos métodos son incrementales, es decir se construye en pequeños y frecuentes avances guiados por pruebas, son rápidos y no se concentran en realizar una documentación exhaustiva del proyecto.

La estrategia a emplear para el desarrollo del proyecto consiste en:

La recopilación de la información tanto analógica como digital existente, como pueden ser los planos, inventarios en hojas de Cálculo y cualquier otra base de datos existente así como la generación de la información geométrica relacionada al área de infraestructura física.

Realizar el análisis y estructuración de la información combinándola de acuerdo a las características que se utilizan en los sistemas de información geográfica.

Realizar cada una de las capas que integraran el sistema como por ejemplo: edificios, áreas verdes, laboratorios, aulas, etc. Incorporando a cada una de ellas la información recopilada a una base de datos.

Instalar un Mapserver como también un servidor de base de datos de Postgres y PostGIS.

Integrar cada una de las capas con Mapserver y Postgres.

Se desarrollaron las interfaces con las que el usuario realizara consultas geoespaciales mediante un sistema pmapper.

Implementar el sistema en el servidor local y verificar su correcto funcionamiento.

Elaborar la tesis con la información y resultados obtenidos del sistema.

Dar el significado de las variables en redacción lineal y es importante la comparación de los criterios usados

**Resultados**

Se entrega funcionando operativamente el sistema para que el Instituto Tecnológico de Pachuca a través del departamento de planeación, utilice las consultas en la toma de decisiones que sobre los temas de infraestructura física, requieren en su quehacer diario.

La experiencia que nos deja este proyecto en nuestro curriculum de vida, es satisfactorio para el emprendimiento de nuevos proyectos donde los conociemientos adquieridos durante nuestra carrera profesionas son aplicables en nuestra vida diaria.

ISSN-2410-3993

**Imágenes que muestran la estructura y el funcionamiento del sistema.**

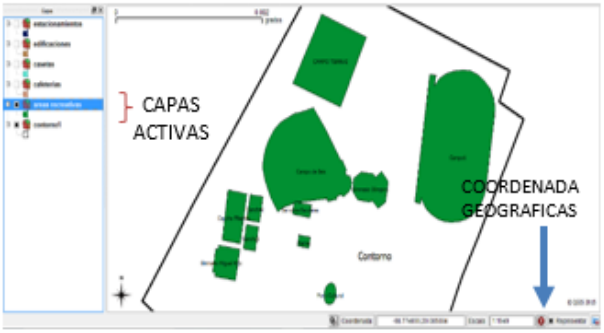



**Figura 10** Capa que contiene áreas recreativas

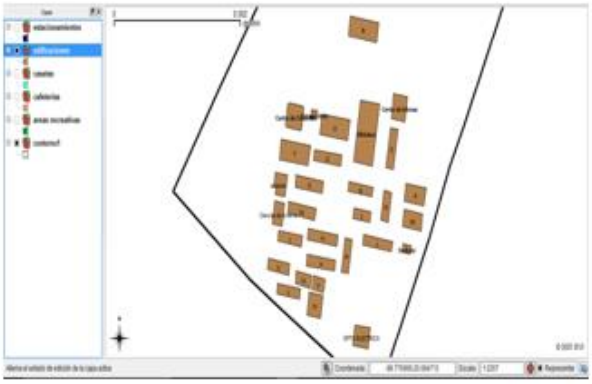


**Figura 11** Edificaciones del ITP

**Figura 12** Capas activas (representación del ITP)

## Agradecimiento

Se expresa el agradecimiento al M. en C. Javier Hernández Orozco por su motivación y apoyo incondicional en proyectos relacionados con este tema, así como su amable invitación y participación en este 1er. Congreso Nacional.
Al Instituto Tecnológico de Pachuca por compartir la información para la realización de esta investigación y confiar plenamente en nosotros como sus estudiantes.

## Conclusiones

La introducción de la tecnología de sistemas nos encamina a que los diversos sistemas de información se conviertan en elementos de importancia en la organización. El amplio espectro de aplicaciones de un sistema basado en SIG impulsa a una mejor calidad en la toma de decisiones, así como una mejor administración de bienes (muebles e inmuebles), ya que se obtiene un amplio panorama visual y analítico de la infraestructura del ITP.

El SIG permite generar información digital, procesarla, administrarla, analizarla y cruzar distintos niveles de dicha información, permitiendo así, modelar escenarios probables y, sobre ellos, planificar. Este proceso, a su vez, genera nuevos datos y nuevos puntos de vista.
Se obtienen grandes mejoras para el departamento de planeación programación y presupuestario ya que se cuenta con medidas reales con puntos exactos en los que físicamente se encuentra la infraestructura por tal motivo se obtienen resultados verdaderos que podrán ser analizados por los usuarios para sus fines laborales.

## Referencias

Victor Oyala. (2011). Sistemas de Informacion Geografica. Creative Common Atribucion.

CEA. (2010). Sistemas de Informacion Geografica. 2010, de Confederacion de Empresarios de Andalucia Sitio web: http://sig.cea.es/SIG

Tripod. (2006). Fase XP. 2006, de tripod.com Sitio web: http://programacionextrema.tripod.com/fases.htm

Gemini. (2005). Modelos de SIG´s Sitio Web: http://gemini.udistrital.edu.co/comunidad/profesores

Miliarium.(2009).Modelos y tipos de Sistemas de Información Geográfica. Sitio Web: http://www.miliarium.com/Proyectos/Nitratos/Modelos/SIG/TiposSIG.asp

ISSN-2410-3993

Artículo **Revista Tecnología e Innovación**
Diciembre 2015 Vol.2 No.5 998-1006

# Sistema de monitoreo del LOBOBUS

REYES, Cecilia*†`, BARRETO, Aldrin`` y BAUTISTA, Verónica Edith``

*`Instituto Tecnológico de Pachuca*
*``Benemérita Universidad Autónoma de Puebla*

Recibido 5 de Julio, 2015; Aceptado 24 de Noviembre, 2015

## Resumen

Este trabajo presenta una propuesta al monitoreo del LOBOBUS haciendo uso del Internet de las cosas (IoC). LOBOBUS es el sistema de transporte público gratuito usado en las instalaciones de la Universidad Autónoma de Puebla. Los autobuses tienen varias paradas establecidas en donde los usuarios deben esperar a que los autobuses lleguen. El sistema fue implementado usando un módulo GPS junto con el microcontrolador con WiFi llamado Electric Imp, el cual incluye de manera gratuita servicios en la nube. También se incluye una base de datos conectada a la plataforma Carriots la cual está diseñada para el IoT. Los datos se almacenan en Carriots y se usan para que a través de una página web que incluye mapas de google se pueda monitorear y administrar el LOBOBUS. Así mismo se desarrolló una aplicación Android con el propósito de que el usuario a partir de su ubicación conozca el tiempo estimado en el cual el transporte llegará a la parada. El sistema fue probado varias veces en diferentes rutas y funcionó de manera adecuada, esto permite evaluar la posibilidad de implementarse en el transporte público. A diferencia de otros trabajos reportados el sistema hace uso de una red WiFi para su operación.

**Electric Imp, GPS, Transporte**

## Abstract

This paper presents a proposal to the monitoring of Lobobus using the Internet of Things (IoT). Lobobus is a free bus transportation system implemented at the Autonomous University of Puebla in which every unit have many bus stop where users wait until the bus arrives. It was developed using a GPS shield along with a 32bits microcontroller with Wi-Fi connection called Electric Imp, which includes freecloud services. The system contains a database connected to the Carriots Platform designed for the Internet of Things. Data are stored and used in a web page including Google Maps for monitoring and managing the Lobobus. An app was developed in order to know an estimated time in which the bus would take to our GPS position. The system was tested several times in different routes and in all of them it worked properly. This allows us to evaluate the possibility for extending the system to other public transportation systems. This system uses only Wi-Fi connection for its operation instead of GSM cellular data as reported in others works.

**Electric Imp, GPS, Transportation**

**Citación:** REYES, Cecilia, BARRETO, Aldrin y BAUTISTA, Verónica Edith. Sistema de monitoreo del LOBOBUS. Revista de Tecnología e Innovación 2015, 2-5: 998-1006

* Correspondencia al Autor (Correo Electrónico: ce_908@hotmail.com)
† Investigador contribuyendo como primer autor.

## Introducción

Cada día es más común la implementación de dispositivos que vinculen y transfieran información del entorno a nuestros dispositivos móviles gracias a sensores y otros componentes electrónicos que realizan mediciones en distintos aspectos. Uno de estos aspectos es la geolocalización que es el conocimiento de la propia ubicación geográfica de modo automático, el cual puede ser proporcionado comúnmente por dispositivos receptores GPS, los cuales y gracias a la red de satélites que rodea al planeta podrán ubicarnos en cualquier parte del globo terráqueo en el cual nos ubiquemos.

El tema de la geolocalización anteriormente se había desarrollado solo en el entorno industrial, pero con la integración de dispositivos GPS en los automóviles y en los dispositivos móviles se ha convertido en un punto atractivo para el desarrollo de aplicaciones.

Por ejemplo en Moedano (2013) se presenta un sistema que permite monitorear el flujo vehicular y analiza los elementos que influyen en la problemática de la acumulación del tráfico en diversos puntos de la Ciudad de México. Se reportan recorridos de reconocimiento de carreteras y se presenta una aplicación que recababa datos provenientes de un GPS y tras analizarlos mostraba en un mapa la afluencia en dicha carretera tomando como parámetros los términos: flujo adecuado, flujo moderado y congestión. En Flores (2013) se propone una aplicación móvil capaz de tomar la posición actual del GPS de un smartphone y buscar en un rango menor a 3 km las estaciones más cercanas del transporte colectivo Metro de la Ciudad de México para posteriormente mostrarlas en un mapa de Google Maps. En caso de no encontrar estaciones en ese rango, se extiende el rango hasta encontrar tres estaciones del transporte.

En ambos trabajos no hay posibilidad de conocer de manera precisa los tiempos de recorrido.

Otro trabajo reportado en Antolines (2013), describe un prototipo que permite georeferenciar dispositivos con tecnología GPS por medio de la recolección en tiempo real de la latitud y la longitud. La información es almacenada en una base de datos que se encuentra dentro de una memoria SD. Para la visualización de los datos, se creó una aplicación kml que los vincula a la plataforma Google Earth para obtener un recorrido gráfico. Al conectar la memoria a una computadora y ejecutar la aplicación, los datos son puestos en Internet para su manipulación directa con Google Earth. En la implementación del hardware se hizo uso de módulos compatibles con la plataforma Arduino UNO. Una ventaja del proyecto es la proyección de los datos en el mapa en forma de puntos y en texto. Por otra parte, tiene limitantes puesto que los datos no se pueden ver en tiempo real y habría que esperar a que el administrador los ponga en línea; para acceder a ellos se necesita forzosamente la aplicación y la interacción con un usuario está muy acotada.

Stahl (2013) propone un monitoreo de los autobuses internos de la Universidad haciendo uso de hardware de la familia Arduino: Arduino Mega, GPS módulo 1.1 y módulo GSM/GPRS IComsat versión 1.1. El módulo GSM/GPRS envía los paquetes de información por medio del formato JSON y son recibidos en un servidor PHP, posteriormente son mostrados en una página web por medio de un mapa de Google Maps.

Una característica de este sistema, es el manejo de estados del autobús, pues el icono del autobús cambia de color en función de su estado actual: verde moviéndose, azul detenido, rojo error y negro desconectado.

ISSN-2410-3993

A pesar de esto, el mapa no cuenta con iconos de paradas a pesar de que si se tienen contempladas en las coordenadas y carece de textos que contengan información de la posición del autobús. Otra desventaja es que el usar diferentes módulos de Arduino, a pesar de ser hardware libre, incrementa el costo de manera importante.

A diferencia de estos trabajos se propone una solución a la problemática que enfrentamos a diario a la hora de transportarnos, de la cual surge la principal cuestión: ¿Cuánto tardará en pasar el camión? Para esto, se considera que en la actualidad la mayoría de la gente cuenta con un dispositivo móvil, ya sea teléfono inteligente, tableta o laptop, con conexión a Internet y a su vez interactúan por medio de ellos con aplicaciones de diversas categorías que facilitan muchas de sus actividades. En base a esto se propone una aplicación que sea capaz de mostrar el tiempo que tardará en pasar el camión en el sentido de ayudarnos a reducir los tiempos de espera o de incertidumbre con respecto a una ruta en particular.

Así mismo se desarrolló un sistema basado en un microcontrolador con WiFi que fuera capaz de monitorear los recorridos de los autobuses, entregar información respecto a su ubicación y tiempos de llegada a las paradas en tiempo real por medio de una aplicación móvil la cual esté disponible para los usuarios del servicio. Además de dar mediciones exactas con respecto a la ubicación de los autobuses y proporcionar datos en tiempo real. También cuenta con un panel de control para la apropiada administración del servicio en general.

## Desarrollo

Esta propuesta surge partir de la situación actual que vive la comunidad universitaria de la Benemérita Universidad Autónoma de Puebla.

Ya que cuenta con transporte público que brinda servicio gratuito dentro de todo el campus de la Universidad conocido como Ciudad Universitaria (CU).

El transporte se denomina LOBOBUS e incluye tres rutas diferentes. Los camiones pueden o no cambiar de ruta según se les asigne en la base, la cual está ubicada a un costado de la Facultad de Filosofía y Letras dentro de la misma Universidad. Además cuenta con una parada de inicio/terminación de ruta en común o base, ubicada en una lateral del estacionamiento central.

El sistema muestra el tiempo que tarda en pasar el transporte LOBOBUS en un punto predeterminado, con el fin de que el usuario pueda decidir entre esperarlo o buscar un medio de transporte alterno dentro de la Universidad. La información del LOBOBUS deberá estar disponible en tiempo real y legible para cualquier usuario. Dicha información debe estar disponible para todos los usuarios con respecto a su posición o por lo menos a la de las paradas oficiales.

## Metodología

En el desarrollo se usó la metodología en cascada que se aprecia en la figura 1, sus iteraciones permiten implementar la construcción del sistema y corregir errores en etapas tempranas, además de facilitar la identificación de las actividades gracias a sus etapas.

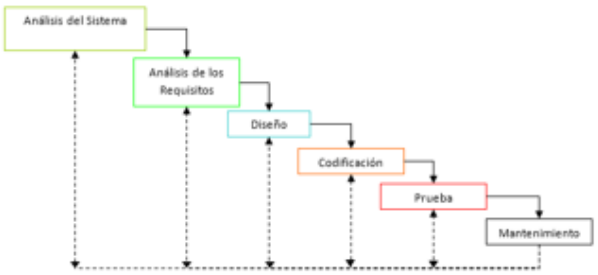



**Figura 1** Metodología en cascada.

Los principales elementos que integran el sistema propuesto son:

- GPS
- Electric Imp
- Conexión web

## GPS

El GPS es el dispositivo encargado de recibir datos provenientes de los satélites que rodean el planeta, dichos datos se les conoce como efemérides y se presentan en formatos NMEA. Dichas efemérides contienen información como la latitud, longitud entre otros dependiendo de cuál de las diferentes decodificaciones (GGA, RMC, GSV, GSA, etc.) se seleccione. Físicamente es una tarjeta de 25.5mm x 35mm x 6,5 mm y un peso de 8.5gr (sin batería), cuenta con una antena Path de 15 mm x 15 mm x 4 mm capaz de vincular hasta 22 satélites de rastreo y 66 de búsqueda y se muestra en la figura 2. Su frecuencia de actualización es de 1 a 10 Hz y su exactitud en la posición es menor a 3 metros.

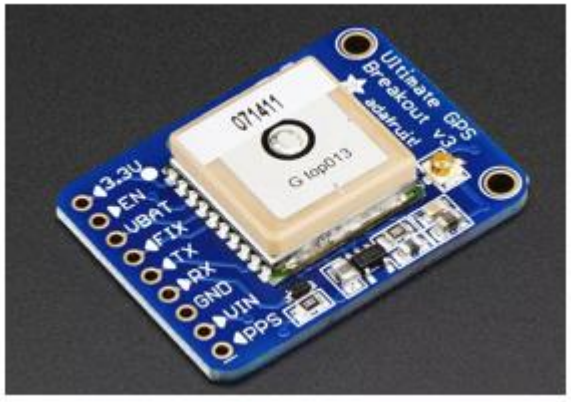



**Figura 2** GPS AdafruitUltimate

La conexión del GPS a Internet se realizó a través del dispositivo Electric Imp.

## Electric Imp

Electric Imp es un microcontrolador de 32 bits con WiFi, que actúa como puerta de entrada para conectar a un servicio de Internet. El hardware Imp está disponible en varias formas; módulos con o sin antena integrada, y como un solo chip para aplicaciones de alto volumen.

Electric Imp cuenta con un micro-servidor alojado en la nube llamado agente, el cual es programable bajo el lenguaje Squirrel con el fin de ajustar el dispositivo a las necesidades de comunicación según sea su uso.

Cada dispositivo tiene su propio y único agente. Además, el agente gestiona la conexión con otros servidores por medio de Internet bajo protocolos HTTPS. También protege el dispositivo de daños, autentica cada petición, asegura los datos que pasan hacia y desde el dispositivo y protege su producto de accesos no permitidos.

Teniendo en claro la importancia de los dispositivos mencionados, se prosigue a la conexión física de ambos que se muestra en la figura 3.

**Figura 3** Conexión GPS y Electric Imp

Un punto importante es el almacenamiento de los datos que va proporcionando el GPS con el fin de obtener estadísticas posteriores a cualquier recorrido realizado.

ISSN-2410-3993

Esto requirió comunicar el GPS a Internet e incluir una base de datos en la que sólo se tenga acceso autentificado del administrador del sistema y que pueda dar de alta alguna de las rutas.

Por otra parte, una característica principal de este sistema es la visualización de diversas entidades, tales como autobuses, paradas y ubicación propia. Todo esto con el fin de utilizar los datos de tal forma que sean fáciles de comprender y manipular, a lo cual se ha establecido el uso de mapas digitales con puntos o marcas que nos indiquen de que entidad se trata.

Para vincular los datos de Electric Imp con un usuario final es necesario tener un punto donde se puedan conjuntar con los elementos gráficos y a su vez sea un punto básico de visualización y manejo de toda la información. Lo anterior nos llevó a la creación de una página web que incluya todos los elementos ya sea de usuario normal y de administrador.

Para tener la correcta conexión del dispositivo a la página web es necesario un servidor intermedio que pueda facilitar dicha conexión, para lo cual se ha seleccionado Carriots.

## Carriots

Carriots es un alojamiento de aplicaciones y plataforma de desarrollo especialmente diseñado para proyectos relacionados con el Internet de las Cosas (IoT) y Máquina a Máquina (M2M). Esto permitió recolectar los datos de los dispositivos conectados, para almacenar y crear aplicaciones potentes con pocas líneas de código de Groovy. Carriots proporciona un entorno sencillo de desarrollo, APIs robustas y hospedaje.

La aplicación puede integrarse fácilmente con sistemas de TI externo a través de APIs potentes, servicios web y un hosting que contiene un ambiente que se ajusta automáticamente para satisfacer cualquier demanda, ya sea de uno o varios dispositivos.

## Diseño del sistema

El diagrama de flujo de la figura 4 muestra el recorrido de los datos del GPS hacia la página web, al igual que otorga los elementos para la generación del código del programa de la figura 5.

Una forma óptima de dar uso a toda esta información es por medio de una aplicación móvil que simplifique la búsqueda dentro de la web de todos estos datos y a su vez proporcione información de la ubicación de dicho móvil para que sea mostrada por medio de una marca en el mapa digital a la vez que se muestran los demás elementos.

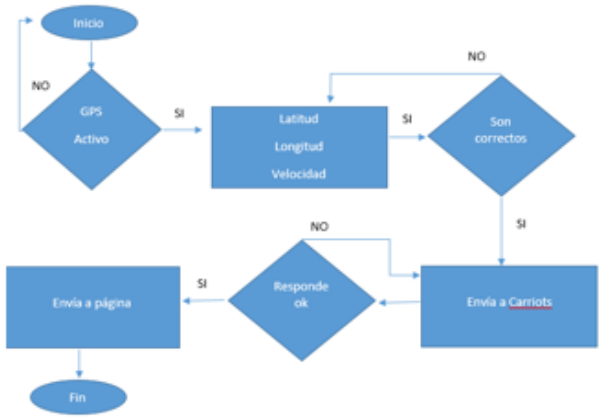



**Figura 4** Diagrama de flujo del envío de datos

```
programa conecta
 {
tiempo = ahora;
        funcion tomar_datos
        {
        tramas=total_tramas;
        mientras tramas dif 0
                {
                si trama=tiempo
                        {
                        enviar trama;
                        tramas=0;
                        }
                si no
                        {
                        tiempo -1;
                        }
                }
        }
}
```

**Figura 5** Pseudocódigo de la conexión de Carriots a la página www.lobobus.hol.es

En base a lo anterior se puede observar la importancia de los gráficos, ya que proporcionarán información valiosa del funcionamiento del sistema y de sus componentes.

Se puede describir el funcionamiento del sistema comenzando por el dispositivo GPS Adafruit que vincula la red de satélites públicos que orbitan el planeta, al recibir datos los envía al dispositivo Electric Imp que previamente necesita estar conectado a una red Wifi. El dispositivo GPS y Electric Imp fueron colocados dentro del autobús, este último posteriormente envían los datos de latitud y longitud al servidor Carriots, donde se permite la extracción de los datos para el uso y visualización de los mismos dentro de la página web www.lobobus.hol.es (figura 6). Cabe mencionar que la página está dentro de un dominio gratuito y el servicio de alojamiento es brindado por www.hostinger.mx.

**Figura 6** Diseño del sistema

En la base de datos, hay que tener presente el registro y baja, autobuses, operadores, recorridos y administradores. Así como tener datos sobre las rutas para verificar que los recorridos se estén cumpliendo. En la figura 7 se muestra el modelo de la base de datos.

Dentro de la página de la figura 8 se necesitan formularios de acceso al panel de control habilitado solo para administradores autorizados por medio de un usuario y contraseña, para dar de alta un nuevo administrador es necesario que otro administrador lo registre.

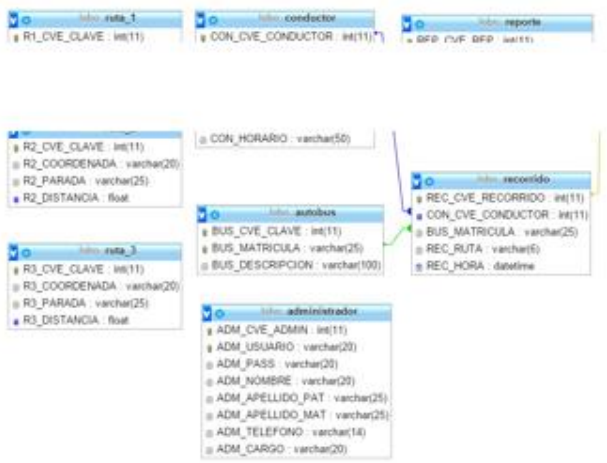



**Figura 7** Modelo de la base de datos

En el panel de control de la figura 9 se encuentra el menú con las opciones principales: inicio, recorrido, reporte, autobús, operador, administrador y cerrar sesión.

ISSN-2410-3993

Las opciones autobús, operador y administrador, cuentan con interfaces para dar de alta un nuevo elemento, buscar todos los elementos, búsqueda individual y borrar.


**Figura 8** Página www.lobobus.hol.es

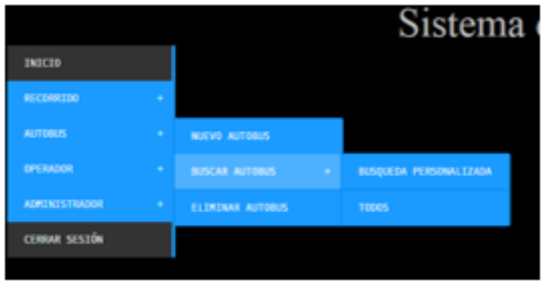



**Figura 9** Panel de control

Otra función de la página es localizar los autobuses gráficamente dentro de un mapa tomado de Google Maps, así como trazar sus recorridos desde el punto donde se encuentra el autobús hasta la parada indicada. Las paradas y el autobús deberán ser identificadas por medio de iconos que contendrán información de sí mismos y que se muestran en la figuras 10 y 11 respectivamente.

**Figura 10** Icono del LOBOBUS

**Figura 11** Icono de las paradas

## Resultados

Los resultados obtenidos se presentan a continuación:

La conexión informática de los dispositivos GPS y Electric Imp con el servidor Carriots fue un éxito ya que los datos están disponibles en la red Internet como se aprecia en la figura 12.

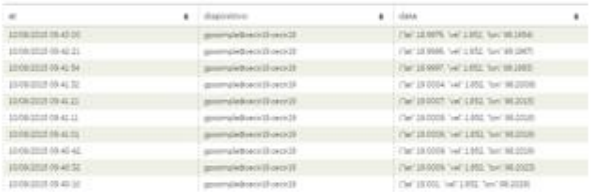


**Figura 12** Datos del GPS en Carriots.com

Otra conexión primordial es la de la página web con el servidor Carriots quien por medio del uso de peticiones CORS de Java Script envía los datos a la página web del LOBOBUS. En base a lo anterior se pudo visualizar el LOBOBUS y las paradas por medio de un mapa incluido en la página por medio del icono que de igual manera proporciona información al hacer click sobre el LOBOBUS como se observa en la figura 13.


**Figura 13** Mapa de las paradas del LOBOBUS

ISSN-2410-3993

Al dar click sobre una parada se mostrará el recorrido que debe realizar el LOBOBUS para llegar a ella y el tiempo que tardará como se muestra en la figura14.


**Figura 14** Información del recorrido del LOBOBUS a una parada

El panel de control funcionó adecuadamente incluyendo acciones administrativas, tales como las generaciones de los recorridos, para que la información sea visible, ya que de lo contrario el sistema no proporcionará información del autobús sino está activo en algún recorrido.

La aplicación es compatible con sistemas operativos ANDROID, es ligera y fácil de utilizar. La primera pantalla ilustrada en la figura 15 proporciona información de coordenadas en formato DEGREES (latitud y longitud) y cuenta con un botón de etiqueta "Busca LOBOBUS", que al dar click, redirecciona el móvil a la página web del sistema (ver figura16) y agrega la marca personal (ver figura17) al mapa.

**Figura 15** Pantalla de la aplicación

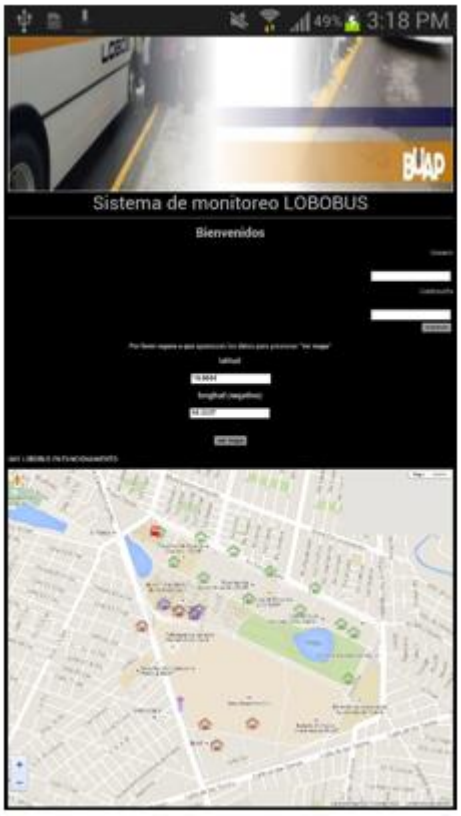



**Figura 16** Vista de la página web desde la aplicación

**Figura 17** Icono de la marca personal.

ISSN-2410-3993

## Conclusiones

La propuesta desarrollada presenta una alternativa al monitoreo del transporte, en este caso se realizó dentro de Ciudad Universitaria de la Benemérita Universidad Autónoma de Puebla, donde se hace uso principalmente del Internet de las Cosas, pero podría extenderse a otras rutas de transporte público.

La aplicación permite tener un seguimiento del movimiento de cada autobús que incluya el sistema, lo cual permite conocer su localización vía GPS dentro de la ruta asignada. Esto junto con el microcontrolador con WiFi Electric Imp permitió que los datos se muestren en páginas web, así como el desarrollo de una app en Android que permita conocer a partir de la geolocalización en cuanto tiempo llegará la siguiente unidad del transporte LOBOBUS a ubicación de la parada.

A diferencia de otras alternativas donde se hace uso de envío de datos vía celular y que implica un costo, la propuesta sólo requiere que durante todo el trayecto de la ruta se tenga cobertura de Internet, con lo cual no existe un costo adicional por el uso del sistema. Cómo parte del trabajo futuro se mejorará la interface del sistema y se desarrollará una aplicación para dispositivos móviles (tabletas, teléfonos inteligentes, etc) donde se pueda disponer de la información del teléfono (ubicación), para integrar un icono con la finalidad de mostrar la posición del dispositivo y como llegar a las paradas oficiales del LOBOBUS.

## Referencias

Moedano Cardiel M. A., Moreno Ibarra E. A., Torres Ruiz M. J. (2013). Análisis del Comportamiento del Tránsito Vehicular con Base en el Sensado de Dispositivos Móviles. Research in Computing Science, 63, pp.131-137

Flores Mendoza, Y., Moran Flores, M A. Moreno Cervantes, A. E. (2013). Sistema Auxiliar Basado en Android para el Tránsito de Usuarios del Sistema de Transporte Colectivo Metro de la Ciudad de México. Research in Computing Science, 63, pp.151-159.

Antolines Estupiñan J.A, Mantilla Gelvez Y. (2013), Implementación de un prototipo de comunicaciones remoto que permita georeferenciar dispositivos con tecnología GPS en las unidades tecnológicas de Santander. Unidades Tecnológicas de Santander.

Stahl Leiton A. G. (2013). Diseño e implementación de un prototipo de sistema de geolocalización para buses. Universidad de Costa Rica.

Facebook oficial del LOBOBUS. www.facebook.com/pages/Lobobus-BUAP/166397513563263, junio 2015.

Roger S. Pressman. (1998). Ingeniería del Software: Un enfoque práctico. México: Mc Graw Hill.

Dispositivo GPS Adafruit. www.adafruit.com/products/746, junio 2015

Dispositivo Electric Imp. www.electricimp.com/product, mayo 2015

Servidor Carriots www.carriots.com/, mayo 2015

ISSN-2410-3993

# Instrucciones para Autores

A. Envió de artículos con las áreas de Tecnología e Innovación.

B. La edición del artículo debe cumplir las siguientes características:

- Redactados en español o en inglés (preferentemente). Sin embargo, es obligatorio presentar el título y el resumen en ambos idiomas, así como las palabras clave.

- Tipografía de texto en Times New Roman #12 (en títulos- Negritas) y con cursiva (subtítulos- Negritas) #12 (en texto) y # 9 (en citas al pie de página), justificado en formato Word. Con Márgenes Estándar y espaciado sencillo.

- Usar tipografía Calibre Math (en ecuaciones), con numeración subsecuente y alineación derecha: Ejemplo;

$$\sigma\epsilon\sum:H\sigma=\bigcap\_(s<\sigma)Hs \quad (1)$$

- Comenzar con una introducción que explique el tema y terminar con una sección de conclusiones.

- Los artículos son revisados por los miembros del Comité Editorial y por dos dictaminadores anónimos. El dictamen será inapelable en todos los casos. Una vez notificada la aceptación o rechazo de un trabajo, su aceptación final estará condicionada al cumplimiento de las modificaciones de estilo, forma y contenido que el editor haya comunicado a los autores. Los autores son responsables del contenido del trabajo y el correcto uso de las referencias que en ellos se citen. La revista se reserva el derecho de hacer los cambios editoriales requeridos para adecuar los textos a nuestra política editorial.

C. Los artículos pueden ser elaborados por cuenta propia o patrocinados por instituciones educativas ó empresariales. El proceso de evaluación del manuscrito no comprenderá más de veinte días hábiles a partir de la fecha de su recepción.

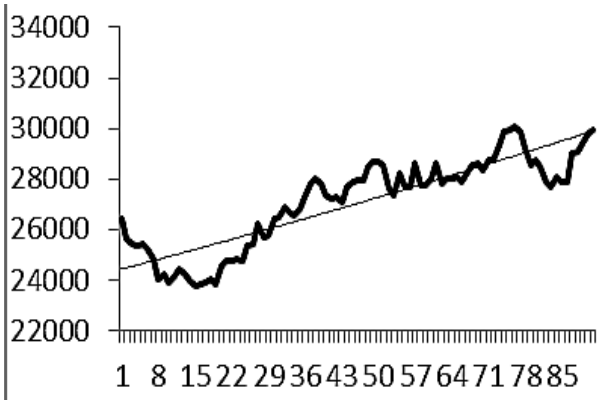
D. La identificación de la autoría deberá aparecer únicamente en una primera página eliminable, con el objeto de asegurar que el proceso de selección sea anónimo.

E. Los cuadros, gráficos y figuras de apoyo deberán cumplir lo siguiente:

- Deberán explicarse por sí mismos (sin necesidad de recurrir al texto para su comprensión), sin incluir abreviaturas, indicando claramente el título y fuente de consulta con referencia abajo con alineación izquierda en tipografía número 9 con negritas.

- Todo el material de apoyo será en escala de grises y con tamaño máximo de 8cm de anchura por 23cm de altura o menos dimensión, además de contener todo el contenido editable

- Las tablas deberán ser simples y exponer información relevante. Prototipo;




**Gráfica 1. Tendencia determinista versus estocástica**

F. Las referencias bibliográficas se incorporarán al final del documento con estilo APA.

La lista de referencias bibliográficas debe corresponder con las citas en el documento.

G. Las notas a pie de página, que deberán ser usadas sólo excepcionalmente para proveer información esencial.

H. Una vez aceptado el artículo en su versión final, la revista enviará al autor las pruebas para su revisión. ECORFAN-Bolivia únicamente aceptará la corrección de erratas y errores u omisiones provenientes del proceso de edición de la revista reservándose en su totalidad los derechos de autor y difusión de contenido. No se aceptarán supresiones, sustituciones o añadidos que alteren la formación del artículo. El autor tendrá un plazo máximo de 10 días naturales para dicha revisión. De otra forma, se considera que el (los) autor(es) está(n) de acuerdo con las modificaciones hechas.

I. Anexar los Formatos de Originalidad y Autorización, con identificación del Artículo, autor (s) y firma autógrafa, de esta manera se entiende que dicho artículo no está postulado para publicación simultáneamente en otras revistas u órganos editoriales.

**Revista de Tecnología e Innovación**

# Formato de Originalidad

**Sucre, Chuquisaca a ____ de ____ del 20_____**

Entiendo y acepto que los resultados de la dictaminación son inapelables por lo que deberán firmar los autores antes de iniciar el proceso de revisión por pares con la reivindicación de ORIGINALIDAD de la siguiente Obra.

Artículo (Article):

_____

Firma (Signature):

_____

Nombre (Name)

# Revista de Tecnología e Innovación

## Formato de Autorización

**Sucre, Chuquisaca a ____ de ____ del 20_____**

Entiendo y acepto que los resultados de la dictaminación son inapelables. En caso de ser aceptado para su publicación, autorizo a ECORFAN-Bolivia a difundir mi trabajo en las redes electrónicas, reimpresiones, colecciones de artículos, antologías y cualquier otro medio utilizado por él para alcanzar un mayor auditorio.

I understand and accept that the results of evaluation are inappealable. If my article is accepted for publication, I authorize ECORFAN-Bolivia to reproduce it in electronic data bases, reprints, anthologies or any other media in order to reach a wider audience.

Artículo (Article):

_____

Firma (Signature)

_____

Nombre (Name)

ISSN 2410-3993