Detección Automática de problemas académicos vía # tag

Meliza Contreras González
Benemérita Universidad Autónoma de Puebla
mcontreras@cs.buap.mx

Pedro Bello López
Benemérita Universidad Autónoma de Puebla
pbello@cs.buap.mx

Resumen

El presente proyecto surge ante la necesidad de contar con un sistema computacional que permita detectar de forma automática problemas escolares asociados al avance académico de los estudiantes, el modelo presentado se basa en el uso básico de técnicas de aprendizaje automático, minería de datos y las tecnologías de la información a través de la web. La extracción de datos de forma automática utilizando el API publica de twitter, su almacenamiento en una base de datos relacional y su posterior análisis permitirá detectar que estudiantes tienen problemas en sus estudios. Como resultado de este trabajo de investigación se presenta el modelo a seguir y una propuesta de desarrollo del sistema vía web.

Abstract

This project was born by the need for having a computer system that it can automatically detect school problems related to academic achieve of the students, this model is based on basic technics of machine learning, data mining concepts and information technologies in the web. The information extraction is automatically done by the Twitter API, it is saved in a relational data base and later it is analyzed for detecting the problems the student have in the school. Like a result of this research work the solution model and the proposal of the web information system are shown.

Palabras clave/ Key Words: Aprendizaje Automático, Tecnologías de la información, Minería de Datos, Desarrollo Web. Machine Learning, Information Technology, Data Mining, Web Development.

Introducción

Aprender es una actividad realizada por el ser humano desde antes que nace y continúa desarrollándola durante toda su vida, por lo que involucra aspectos como: la adquisición de nuevo conocimiento[7], su clasificación, el desarrollo de habilidades a través de la práctica, el descubrir nuevos hechos con base en conocimiento previo, así como la búsqueda de facilitar el proceso de aprendizaje de modo que sea más rápido y eficiente. En el ser humano este proceso se da de forma natural y siempre es posible mejorarlo.

Ahora con el gran auge de las Tecnologías de la Información[9][11], los estudiantes han adoptado nuevas y variadas formas de aprender donde la computadora es un medio para acceder a la información principalmente en internet, como se muestra en la Figura 1 la mayoría de los usuarios se conectan a internet para el uso del correo electrónico, entrar a las redes sociales y realizar búsquedas[1].

[1] Fuente de consulta: Asociación mexicana de internet https://www.amipci.org.mx/es/

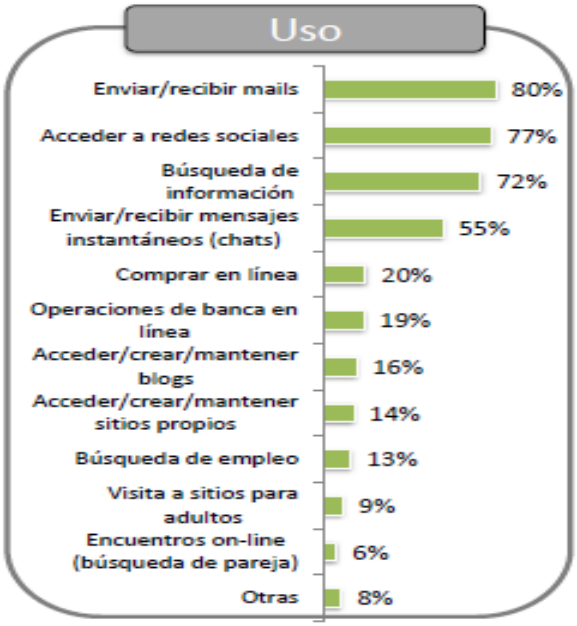

Figura 1. Uso de internet en el 2014

A partir de esta tendencia se ha desarrollado una nueva disciplina llamada Pensamiento Computacional [2] cuyos objetivos son:

- Formular problemas de manera que permitan usar computadoras y otras herramientas para solucionarlos.
- Organizar datos de manera lógica y analizarlos.
- Representar datos mediante abstracciones, como modelos y simulaciones.
- Automatizar soluciones mediante pensamiento algorítmico (una serie de pasos ordenados).
- Identificar, analizar e implementar posibles soluciones con el objeto de encontrar la combinación de pasos y recursos más eficiente y efectiva.
- Generalizar y transferir ese proceso de solución de problemas a una gran diversidad de estos.

Estas habilidades se apoyan y acrecientan mediante una serie de disposiciones o actitudes que son dimensiones esenciales del Pensamiento Computacional. Estas disposiciones o actitudes incluyen:

- Confianza en el manejo de la complejidad

- Persistencia al trabajar con problemas difíciles
- Tolerancia a la ambigüedad
- Habilidad para lidiar con problemas no estructurados (open-ended)
- Habilidad para comunicarse y trabajar con otros para alcanzar una meta o solución común.

Considerando estos elementos resulta indispensable establecer estrategias basadas en estas tecnologías para desarrollar competencias y habilidades, así como contextualizar los problemas que tienen los estudiantes, puesto que estos aprenden con el uso de la tecnología ya que son nativos digitales, donde cada vez más se sustituye el lápiz y el cuaderno por dispositivos electrónicos, y es justo aquí donde se propone capturar los posibles problemas académicos mediante el uso #tags en las redes sociales. En la Figura 2 se muestra que la población joven de entre 13 y 24 años son los que más usan internet por lo que indica que justo los estudiantes en su mayoría utilizan este medio como fuente de consulta y aprendizaje, además 9 de cada 10 usuarios de internet tienen acceso a una red social y según se estima los usuarios de internet están en promedio más de 5 horas por día (Asociación mexicana de internet https://www.amipci.org.mx/es/), por lo que se les hace natural exponer de forma pública cada una de sus vivencias y problemáticas mediante las redes sociales a todos sus amigos y conocidos, en particular nuestro modelo está interesado en abordar problemas de índole académico.

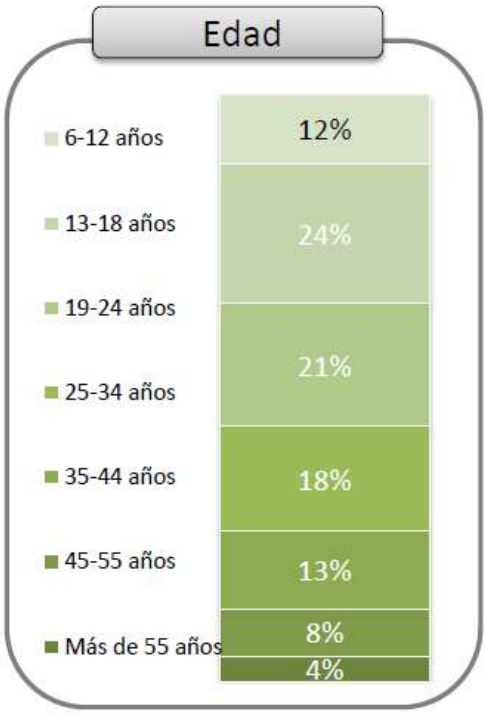

Figura 2. Distribución de edades del uso de internet[2]

El Aprendizaje Automático[1], que pertenece al campo de la Inteligencia Artificial, permite resolver problemas de manejo y clasificación de la información, en particular la Minería de Datos nos permite extraer información para ser analizada y obtener resultados para prevenir o aplicar alguna propuesta de solución.

La Minería de Datos también llamada exploración de datos es un área de las Ciencias Computacionales referido al proceso que intenta descubrir patrones en grandes conjuntos de datos para lo cual utiliza métodos de la inteligencia artificial, el aprendizaje automático, la estadística y los sistemas de bases de datos, además de estos conceptos integrados con las tecnologías de la información podemos obtener características importantes de la población estudiada. Esta disciplina intenta extraer información de un conjunto de datos, en nuestro caso de las redes sociales, y almacenarla en una estructura portable y comprensible (base de datos) para su uso posterior [3][4][5].

[2] Fuente de consulta: Asociación mexicana de internet https://www.amipci.org.mx/es/

Modelo del sistema

Para la detección de problemas académicos se propone integrar un conjunto de técnicas y herramientas de tecnologías de la información junto con una metodología de desarrollo de aplicaciones rápidas para generar un sistema web. La propuesta del sistema esta modelada en forma gráfica en la Figura 3 donde se integran diversos elementos que se describirán a continuación.

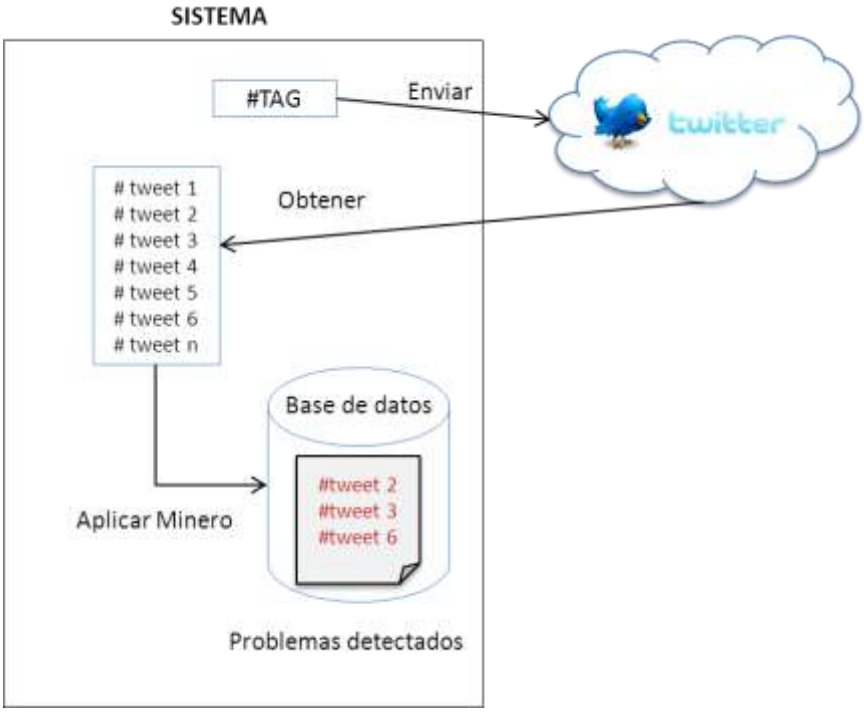

Figura 3. Modelo del sistema para la detección de problemas académicos

El proceso inicia con la publicación de un #Tag en twitter desde el sistema propio o directamente desde una cuenta de twitter. Después de algún tiempo (días) el sistema busca en twitter los resultados de los #tags publicados por el sistema o bien se puede hacer una búsqueda de todos los #tags de acuerdo a alguna palabra clave, por ejemplo “BUAP”, “FCC”, entre otros. Los resultados son filtrados por medio de un minero de datos donde se buscaran palabras clave o frases como “me fue mal”, “no voy a pasar”, “me quiere reprobar”, “me odia el profe”, “me cae mal el maestro(a)”, etc. Todo los tweets obtenidos serán almacenados en la base de datos con MySQL[10] para generar posteriormente un reporte para su análisis.

- **Twitter**

Twitter es un servicio en red de microblogging que se ofrece desde 2006 y que actualmente ha ganado mucha popularidad, por medio de internet se permite enviar mensajes de texto plano de corta longitud llamados tweets, que se muestran en la página principal del usuario y otros usuarios los pueden seguir y participar comentando el tweet[6].

Los mensajes del twitter se mantienen en un servidor y se dispone de una API (Application Programming Interface) abierta para todo tipo de desarrolladores. Los usuarios pueden agrupar mensajes sobre un mismo tema mediante el uso de etiquetas de almohadilla – palabras o frases iniciadas mediante el uso de una "#" (almohadilla) conocidas como hashtag.

Twitter es una red de que permite conocer lo que se habla y lo que interesa a la gente. Así mismo, en cierta medida, es la noticia en vivo que está superando a los medios tradicionales de comunicación, por su inmediatez y novedad. Existen diversas aplicaciones para extraer información, los servicios gratuitos ofrecen informaciones básicas, mientras que los sitios web de pago permiten obtener una información mucho más elaborada. Sin embargo, cualquier persona interesada dispone de los recursos necesarios para explorar, por su cuenta, esa inmensa mina de datos que es hoy Twitter y extraer conocimiento, tanto predictivo como explicativo, en diversos campos como en el educativo, marketing o el sociológico y antropológico.

Para extraer la información de twitter, el entorno de programación PHP[10] suministra extensiones o paquetes, como twitter-API-PHP que permite, entre otras opciones, extraer tweets públicos aplicando diversos criterios tales como rango de fechas, usuario, topicos, hashtag, y palabras o frases claves. Una vez que hemos obtenido los datos en bruto, es decir, la colección de tweets que cumplen unas determinadas condiciones, podemos trasladar la información de los tweets en tablas normalizadas que permitirán realizar una exploración analítica de los datos y su representación gráfica. Asimismo, al margen de los datos estructurados que podemos extraer: usuario, conexiones con otros usuarios, fecha y hora de publicación del tweet, etcétera, el aspecto más interesante es extraer información significativa del propio texto del tweet, una información no estructurada que plantea importantes desafíos en la búsqueda

semántica. Asimismo, tenemos la posibilidad de realizar análisis, en el texto del tweet, de actitudes positivas o negativas hacia un determinado acontecimiento, producto o servicio.

- **Funciones principales del sistema**

Para el desarrollo del sistema se propone utilizar una metodología de desarrollo rápido de aplicaciones debido a que se utilizan las etapas básicas del ciclo de vida del software y se obtiene un producto en poco tiempo de desarrollo que puede ser probado. En la Figura 4 se muestra un Diagrama de Casos de Uso donde se representan las principales funciones del sistema y el actor que participa, a continuación se muestra la descripción de las funcionalidades.

Figura 4. Principales funciones del sistema para la detección de problemas académicos

Login: El sistema solo será utilizado por un administrador que tendrá acceso a todas las funciones por medio de un login y un password.

Publicar #Tag: Esta función sirve para enviar un #tag a twitter en internet con el fin de capturar las ideas de algún tema específico por ejemplo *#ReprobeUnaMateria*, lo que se espera es que los usuarios de twitter sigan el #Tag y posteriormente se recuperae información al respecto.

Registrar usuario Twitter: Con esta función se registrará la población que nos interesa saber sus comentarios a través de la red social de twitter, en particular utilizamos los usuarios registrados para aplicarles diversas consultas con el fin de detectar si estos usuarios tienen algún problema académico.

Recuperar datos de #Tag: Esta función permite obtener la información de los #Tag que se hayan publicado por el sistema o bien obtenerlos de acuerdo a alguna palabra clave. Estos hashtag son procesados y solo se almacenan los que se detecte que pertenecen a los usuarios registrados en el sistema para posteriormente ser analizados.

Generar reportes de #Tag: Esta opción permite generar la ocurrencia del número de participantes por cada #Tag en una gráfica simple de barras.

Reporte de problemas: Esta función permite listar los problemas detectados de los usuarios e indica si ya fueron atendidos o no, se toma de la base de datos los #Tag almacenados y se aplica minería de datos para obtener aquellos que tengan algún tipo de problema académico.

- **Almacenamiento de la información**

Para concentrar la información necesaria para la detección de problemas académicos se utiliza una base de datos relacional para obtener de forma ágil los aspectos a revisar y almacenar solo los datos que nos interesa. En la Figura 5 se muestra el esquema conceptual para el diseño de la base de datos, donde se utilizan las tablas:

Twiteros: En esta tabla se almacenan los principales datos de los usuarios del twitter que nos interesa analizar y seguir en sus conversaciones.

#Tags: Esta tabla se utiliza para almacenar los #Tags publicados o los que se obtienen de twitter para ser analizados.

Filtrados: En esta tabla se almacenan solo los comentarios de los usuarios registrados en el sistema para posteriormente obtener un reporte de los problemas detectados.

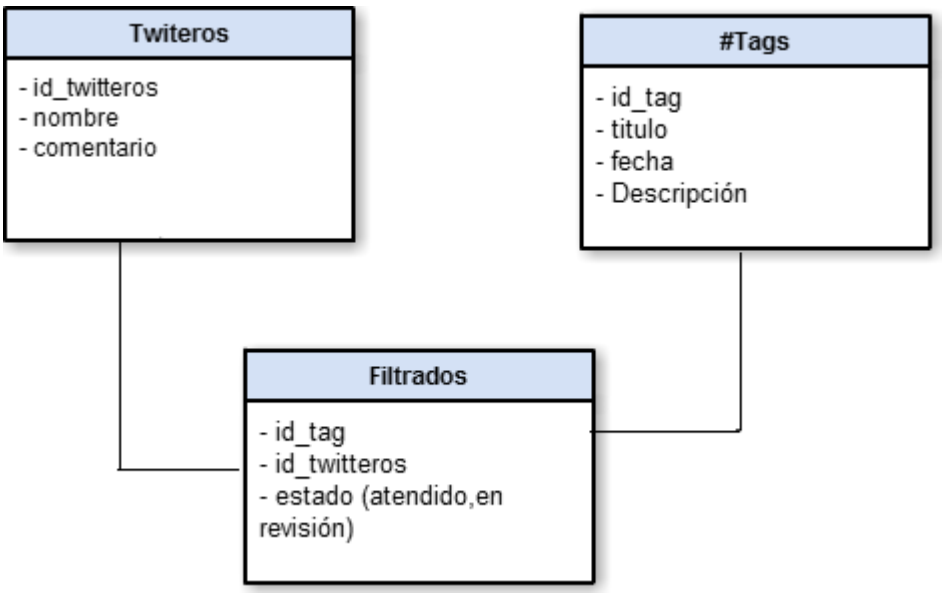

Figura 5. Tablas con atributos para el diseño de la base de datos del sistema para la detección de problemas académicos

Resultados preliminares

El modelo de detección automática de problemas académicos se basa en el uso de las Tecnologías de la Información[8] para lo cual se diseñó un prototipo de una Aplicación Web inicial con el que se han realizado algunas pruebas, en la aplicación desarrollada se cuenta con las funciones descritas en el Diagrama de Casos de Uso. En la Figura 6 se muestra la página inicial del sistema con el usuario administrador que debió acceder con usuario y contraseña.

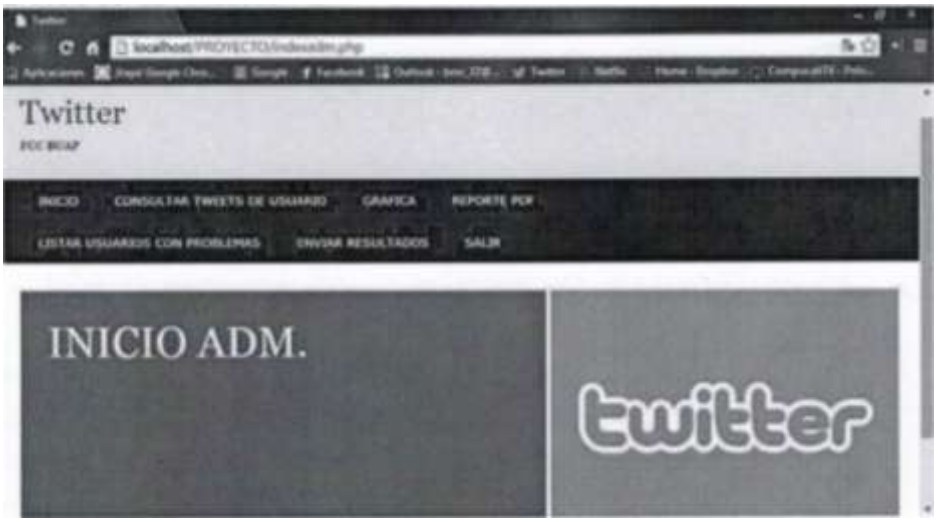

Figura 6. Pagina Inicial del usuario administrador

Una de las principales funciones del sistema es realizar búsquedas de los #Tags que se han publicado o de los que se ha obtenido información en twitter, esto se muestra en la Figura 7. Donde en particular se busca los tweets almacenados en la base de datos de un determinado usuario, además de obtener una consulta en pantalla, se puede exportar a pdf directamente desde la aplicación.

Figura 7. Página de consulta de tweets por usuario

El sistema permite contabilizar el número de tweets por cada uno de los twiteros, en la Figura 8 se muestra en un ejemplo de cómo se verá una gráfica de los usuarios registrados como twiteros y el número de #Tag en que han participado.

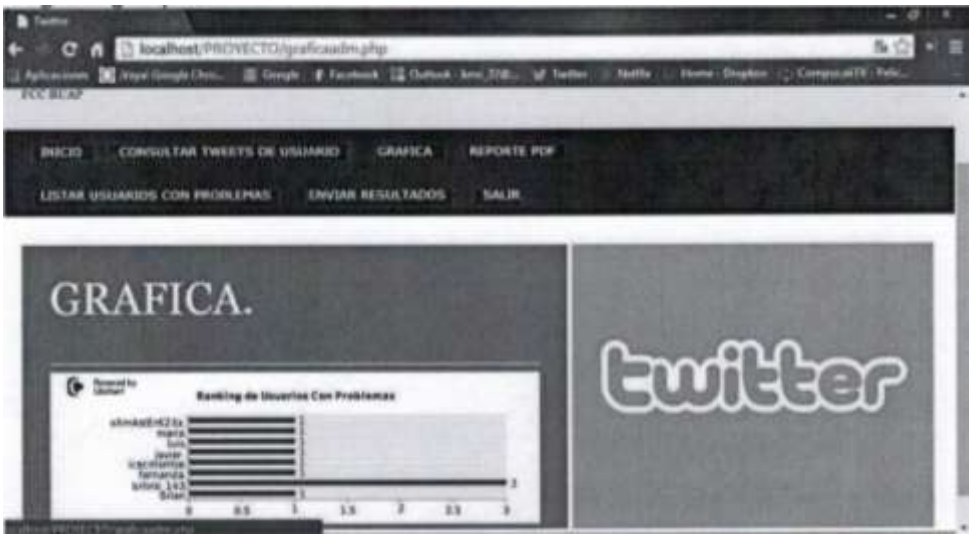

Figura 8. Ejemplo de Grafica de twiteros

Al momento se han almacenado cerca de 50 tweets, de acuerdo a la Tabla 1 estos tweets se han obtenido de los hashtag públicos de twitter y a los resultados obtenidos se les aplicaron consultas simples con MySql obteniendo que los principales problemas de esta muestra son: “mal”, “tuve problemas”, “me enferme”.

Tabla I. Principales hashtags utilizados

.

#Hashtag	tweets
#OrgullosamenteBUAP	9
#OpinionBuap	8
#BUAP	12
#Hackathon	6
#FirefoxOS	5
#FCC	8
#Puebla http://bit.ly/1ihAIa4	3
#DrupalDay	5

Conclusión

En el presente trabajo se muestra un modelo computacional para detectar problemas principalmente académicos utilizando como base el análisis de la información por medio de las tecnologías de la información. En la propuesta del modelo utilizamos algunos conceptos básicos del aprendizaje automático, la minería de datos y el diseño de una aplicación web que integra el modelo presentado.

Las tecnologías de la información nos presentan un conjunto de métodos, técnicas y sobre todo herramientas para la extracción del conocimiento y análisis de la información, al utilizar las redes sociales como medio de estudio hace que podamos acceder a lo que nuestros alumnos están escribiendo, diciendo y tal vez pensando. Nuestro modelo trata de alguna manera de lanzar una especie de anzuelo para captar la información de nuestro entorno, procesar esta información y obtener algún parámetro que sea utilizado para detectar problemas académicos y de esta forma proponer algún tipo de ayuda a los twiteros. Como se mencionó antes para reducir el rango de aplicación y almacenamiento de tanta información solo guardamos los comentarios de los usuarios registrados.

Como resultados preliminares se obtuvo una propuesta de un sistema en web que permite al administrador del sitio web el manejo de las principales funciones para la detección de problemas académicos, este primer prototipo es necesario mejorarlo e integrarle más herramientas de análisis y despliegue grafico de la información utilizando por ejemplo el lenguaje R que además proporciona un amplio abanico de herramientas estadísticas con modelos lineales y no lineales, tests estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento.

Este modelo puede resultar de utilidad para las disciplinas sociales y el sector educativo pues permitirá graficar tendencias y detectar necesidades o demanda de productos y servicios, conocimientos para ser clasificados y asi proponer mejoras en donde existan deficiencias.

Bibliografía

- Carbonell, J.G. Michalski, R.G. y Mitchell, T.M.(1983). An Overview of Machine Learning. Michalski, R.S. et al.(eds).Machine Learning:Symbolic Computation. Berlin Heidelberg:Springer-Verlag
- Thornton C.J. (1993). Techniques in Computational Learning. An Introduction. Chapman & Hall.
- Wenger, E. (1987). Artificial intelligence and Tutoring Systems. Morgan Kaufmann En Computational approaches to the communication of knowledge. Los Altos.
- Kodratoff, Y. Michalski, R.G.(1990). Machine Learning, An Artificial Intelligence Approach, Volumen III, Morgan Kaufmann Publishers, Inc. CA. USA.
- Ponce, P.C. (2010). Inteligencia Artificial con aplicaciones a la ingeniería. México: Alfaomega.
- Eccher, C. (2011). Diseño Web Profesional. España: Anaya
- Schunk, D.H. (2012). Teorías del aprendizaje, una perspectiva educativa. Sexta edición. México: Pearson.
- Law, E. Von Ahn, L. (2012). Human Computation. Synthesis Lectures on Artificial Intelligence and Machine Learning. 2011.
- Quinn, A. J, Bederson, B. (2011). Human Computation: A Survey and Taxonomy of a Growing Field. Vancouver: *CHI '11 Proceedings of the 2011 annual conference on Human factors in computing systems,* 403-1412.
- Anderson, S.P. (2012). Diseño que seduce, Cómo crear webs y aplicaciones atractivas al usuario. México: Grupo Anaya, S.A.
- Bello, P. Contreras, M. Rodríguez , M. Ángel N. (2012). Computación humana en el aprendizaje del Inglés, Congreso internacional de investigación AcademiaJournals.com Celaya 2012. ISSN 1946-5351.