Sistema de gestión de encuestas turísticas en línea para el análisis de segmento meta en la Riviera Maya

Online tourism survey management system for target segment analysis in the Riviera Maya

Artículo de Investigación

Meliza Contreras González*, Pedro Bello López, Mario Rossainz López, Brenda Karen Miranda Hernández

Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla,
Av San Claudio y 14 Sur, Cd Universitaria, Puebla, Puebla, México, C.P. 72590.

mel_22281@hotmail.com, pb5pbello@gmail.com, mrossainzl@gmail.com, brendamiranda19@outlook.com

PALABRAS CLAVE:

Aplicación web, algoritmos de clasificación, mercadotecnia.

RESUMEN

Actualmente hay una gran competencia para la generación de encuestas en línea, existen formatos diversos. Sin embargo, en algunos casos se requiere realizar una encuesta a la medida, como en el caso del sector turismo, donde en cada zona se aplican distintos criterios de acuerdo a la ubicación geográfica para posteriormente aplicar algoritmos de clasificación para detectar las clases que reflejen el segmento de mercado. Por lo que en este trabajo se propone un sistema que permite generar encuestas para posteriormente transformar los datos resultantes a un esquema compatible con Weka para la realización de análisis de datos.

KEYWORDS:

Web application, clustering algorithms, marketing.

ABSTRACT

Currently there is great competition for the generation of online surveys, there are various formats. However, in some cases it is necessary to carry out a customized survey, as in the case of the tourism sector, where different criteria are applied in each area according to the geographical location, and then classification algorithms are applied to detect the classes that reflect the market segment. Therefore, in this work we propose a system that allows the generation of surveys to later transform the resulting data into a Weka compatible scheme for data analysis.

Recibido: 31 de agosto 2020 • **Aceptado:** 5 de abril 2021 • **Publicado en línea:**

1 INTRODUCCIÓN

La situación actual de la pandemia requiere generar estrategias para mejorar las condiciones del comercio e incentivar el turismo. Por otro lado las generaciones actuales requieren diferentes tipos de turismo de acuerdo a su edad, condición económica y preferencias. Hoy más que nunca se hace necesario promover un turismo sustentable y de calidad que ofrezca productos y servicios innovadores, con mayor valor agregado y con una adecuada articulación de la cadena de valor.

Los principales centros turísticos: Cancún, Los Cabos, Loreto, Ixtapa Zihuatanejo y Huatulco requieren incrementar su economía con un acertado estudio de marketing de acuerdo al segmento meta buscado.

En la actualidad, en el contexto de los procesos modernizadores de la sociedad el turismo se manifiesta como un sector económico internacionalizado y competitivo, con fuertes lazos de dependencia externa, impulsando al gobierno federal para clasificar al turismo en: turismo cultural, turismo de reuniones, turismo deportivo, turismo de salud y bienestar, turismo de sol y playa, turismo de naturaleza, turismo gastronómico.

De ahí la importancia de generar un corpus de información de los turistas que contemplen:

- El intercambio de información con el turista.
- Consultar tarifas
- Hacer reservas
- Realizar pagos
- Consultar e informarse acerca del destino y sus atracciones turísticas.
- Consultar los niveles de ocupación
- Consultar las características de las posibles opciones de alojamiento
- Mayor segmentación o propuestas turísticas
- Estar constantemente informado acerca de, promociones, nuevos paquetes, destinos, medios de acceso y alojamientos por medio de boletines electrónicos.

Po otro lado la minería de datos y los sistemas de descubrimiento del conocimiento (KDD) permiten la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados; evaluar e interpretar los patrones para convertirlos en conocimiento; consolidar el conocimiento resolviendo posibles conflictos con conocimiento previamente extraído; y hacer el conocimiento disponible para su uso. Esta definición del proceso clarifica la relación entre el KDD y la minería de datos; el KDD es el proceso global de descubrir conocimiento útil desde las bases de datos mientras que la minería de datos se refiere a la aplicación de los métodos de aprendizaje y estadísticos para la obtención de patrones y modelos [1].

Dentro de los mineros de software, Weka es muy empleado. El sistema está escrito en Java y distribuido bajo los términos de la licencia publica general de GNU [2],[3]. Incluye métodos para todos los problemas de minería de datos estándar: regresión, clasificación, clusters, reglas de asociación y atributos de selección.

En este trabajo se integrará este minero con una herramienta web para generar encuestas dinámicas de acuerdo a las preferencias del usuario de acuerdo al sector turístico para poder analizar las tendencias con la información recopilada de las encuestas. Como experimento se recopilo información proveniente de la Riviera Maya.

2 ALGORITMOS DE WEKA EMPLEADOS

2.1 ALGORITMO COBWEB

Se trata de un algoritmo de clustering jerárquico. COBWEB, se caracteriza porque utiliza aprendizaje incremental, esto es, realiza las agrupaciones instancia a instancia. Durante la ejecución del algoritmo se forma un árbol (árbol de clasificación) donde las hojas representan los segmentos y el nodo raíz engloba por completo el conjunto de datos de entrada. Al principio, el árbol consiste en un único nodo raíz. Las instancias se van añadiendo una a una y el árbol se va actualizando en cada paso.

La actualización consiste en encontrar el mejor sitio donde incluir la nueva instancia, operación que puede necesitar de la restructuración de todo el árbol (incluyendo la generación de un nuevo nodo anfitrión para la instancia y/o la fusión/partición de nodos existentes) o simplemente la inclusión de la instancia en un nodo que ya existía. La clave para saber cómo y dónde se debe actualizar el árbol la proporciona una medida denominada utilidad de categoría, que mide la calidad general de una partición de instancias en un

segmento. La restructuración que mayor utilidad de categoría proporcione es la que se adopta en ese paso. El algoritmo es muy sensible a otros dos parámetros:

Acuity Este parámetro es muy necesario, ya que la utilidad de categoría se basa en una estimación de la media y la desviación estándar del valor de los atributos, pero cuando se estima la desviación estándar del valor de un atributo para un nodo en particular, el resultado es cero si dicho nodo sólo contiene una instancia.

Así pues, el parámetro acuity representa la medida de error de un nodo con una sola instancia, es decir, establece la varianza mínima de un atributo.

Cut-off Este valor se utiliza para evitar el crecimiento desmesurado del número de segmentos. Indica el grado de mejoría que se debe producir en la utilidad de categoría para que la instancia sea tenida en cuenta de manera individual.

Además COBWEB pertenece a los métodos de aprendizaje conceptual o basados en modelos. Esto significa que cada clúster se considera como un modelo que puede describirse intrínsecamente, más que un ente formado por una colección de puntos.

Al algoritmo COBWEB no hay que proporcionarle el número exacto de conglomerados que queremos, sino que en base a los parámetros anteriormente mencionados encuentra el número óptimo.

2.2 Algoritmo EM

EM pertenece a una familia de modelos que se conocen como Finite Mixture Models, los cuales se pueden utilizar para segmentar conjuntos de datos. EM encajaría dentro de los Métodos de Particionado y Recolocación, en concreto Clustering Probabilístico.

El ajuste de los parámetros del modelo requiere alguna medida de su bondad, es decir, cómo de bien encajan los datos sobre la distribución que los representa. Este valor de bondad se conoce como el likelihood de los datos. Se trataría entonces de estimar los parámetros buscados θ, maximizando este likelihood (este criterio se conoce como ML-MaximunLikelihood).

Después de una serie de iteraciones, el algoritmo EM tiende a un máximo local de la función L. Finalmente se obtendrá un conjunto de clusters o conglomerados que agrupan el conjunto de proyectos original. Cada uno de estos cluster estará definido por los parámetros de una distribución normal.

La implementación que lleva a cabo del algoritmo EM lleva asociada la premisa de la independencia de los atributos utilizados, nada más lejos de la realidad existente entre el esfuerzo y los puntos de función.

2.3 Algoritmo de agrupamiento Canopy

Es un método no supervisado de preclustering. Preclustering es un conjunto de tareas de procesamiento previo a un proceso de agrupamiento o clustering, cuyo objetivo es acelerar el proceso de clustering especialmente con grandes conjuntos de datos.

El método de agrupamiento Canopy divide el proceso de segmentación en dos etapas.

En la primera etapa usaremos una métrica sencilla en cálculos, con el objetivo de generar los canopies o subgrupos superpuestos de instancias. Cada instancia pueda pertenecer a más de un canopy y, a su vez, todas las instancias tienen que pertenecer, al menos a un canopy.

En la segunda etapa, utilizaremos el segmento tradicional, como por ejemplo el agrupamiento jerárquico aglomerativo (AHC) o el método k-means, pero lo haremos con la restricción de no calcular la distancia entre los puntos que no pertenecen al mismo canopy.

El algoritmo parte del conjunto de datos y dos valores límite (thresthold): T1 que indica la distancia máxima de la periferia (the loose distance) y T2 que indica la distancia máxima del nucleo (the litght distance), donde T1 > T2. A partir del conjunto inicial de instancias o puntos, se escoge un punto (aleatoriamente) para formar un nuevo canopy. A continuación se calcula la distancia entre este punto y los demás puntos del conjunto de datos, utilizando una métrica más sencilla de calcular que hemos llamado cheapest metric. Se incluyen en el mismo canopy todos los puntos que tengan una distancia inferior al threshold T1. Además, los puntos que tengan una distancia inferior a T2, son eliminados del conjunto de datos. De esta forma se excluye a estos puntos para formar un nuevo canopy y ser el centro de este. Este proceso se repite de forma iterativa hasta que no quedan puntos en el conjunto de datos.

3 MÉTODO

La metodología a emplear para la elaboración del proyecto fue el Proceso Unificado, el paso fundamental fue modelar los requerimientos del sistema, estos se enfocaron a permitir al usuario capturar la encuesta para posteriormente exportar los resultados almacenados en una base de datos al formato ARFF y analizar mediante análisis cluster los resultados obtenidos, a continuación se describen las etapas realizadas.

3.1 Diseño de la Aplicación

La Figura 1 muestra la hoja de estilo con el banner de imágenes de la zona Riviera Maya con el objetivo de familiarizar al turista con el tipo de encuesta a responder. Se incluye un texto corto de Bienvenida en la pestaña de "Inicio".

Fig. 1. Pantalla Inicial de la interfaz para el llenado de la encuesta.

El turista se dirige a la pestaña de la página web "Realizar Encuesta" para seleccionar el País de procedencia en el campo tipo lista de catálogo de países. Para avanzar con la encuesta dar clic en el botón "Continuar", como se muestra en la Figura 2.

Fig. 2. Ejemplo de llenado de los datos de ubicación para la encuesta.

Cuando el usuario ha seleccionado el lugar de procedencia, inmediatamente se mostrará la ventana de la encuesta, conformada por preguntas de selección múltiple. Se observará que todas las preguntas tienen respuestas pre-seleccionadas con el objetivo de que la encuesta sea respondida en el menor tiempo posible. Para terminar la encuesta se tendrá que seleccionar el botón "Finalizar Encuesta", como se muestra en la Figura 3.


1. **¿De qué lugar nos visita? / Where do you come from?**

2. **¿Cuál es su rango de edad? / How old are you?**
☐Menor de 18 años/ Less 18th ☐18 a 30 ☐30 a 40 ☐50 en adelante /more 50

3. **Su sexo es: / What is your gender?**
☐ Femenino /female ☐Masculino/male

4. **Su estado civil es: / Can you tell me your marital status?**
☐ Soltero/Single ☐ Casado/Married ☐Divorciado/Divorced ☐ Viudo/widower

5. **Viaja acompañado de: / Who are your travelling with?**
☐ Familia/Family ☐Pareja/Couple ☐Amigos / Friends ☐Parientes/Relatives

6. **Su nivel de estudios es:/ What is your current level of education?**
☐ Sin estudios/Without higher education ☐Educación Básica (primaria, secundaria) /Elementary-Junior high ☐ Educación Media- Superior (bachillerato o preparatoria)/high school
☐Educación Superior (licenciatura, posgrado)/university

7. **Su ocupación es:/ What do you do for a living?**
☐ Estudiante /student ☐Empresario/ Entrepreneur ☐Empleado en sector privado o público
☐Trabaja por su cuenta/self-employed ☐Labores del Hogar
☐Ninguna/Neither

8. **Para llegar a la zona turística viajó en: /In order to get to Cancun, did you use ?**
☐Avión/airplane ☐Autobús/bus ☐ Vehículo propio/ own car

9. **Partiendo de su lugar de origen, cuantas horas de viaje le tomó llegar a la zona turística?/ How many hours does it take to get to Cancun from where you live?**

Fig. 3. Ejemplo de encuesta.

Respecto al menú del administrador se tienen las siguientes opciones como se muestran en la Figura 5, donde se pueden consultar las encuestas o generar el archivo ARFF correspondiente.

Fig. 4. Graf Menú del administrador

Con el archivo arff generado, se aplicaron distintos algoritmos de clustering para identificar el tipo de turismo que requieren los clientes que contestaron la encuesta, el algoritmo que indicó mejores grupos diferenciados fue el algoritmo FarthestFirst, que tuvo como resultados que en la zona turística de la Riviera Maya prefieren los adultos jóvenes, el turismo de aventura y el ecológico, por lo que esta información es de utilidad para generar campañas y paquetes adecuados al segmento meta, como se muestra en la Figura 5.

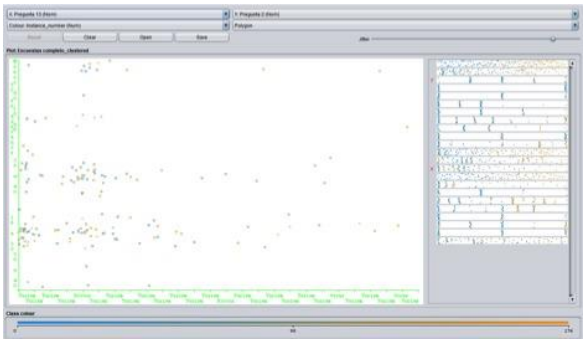

Fig. 5. Ejemplo de uso de los archivos ARFF generados por el sistema.

3.2 PREPROCESAMIENTO DE DATOS

Engloba a todas aquellas técnicas de análisis de datos que permite mejorar la calidad de un conjunto de datos de modo que las técnicas de extracción de conocimiento/minería de datos puedan obtener mayor y mejor información (p.e. mejor porcentaje de clasificación). La preparación de datos puede generar un conjunto de datos más pequeño que el original, lo cual puede mejorar la eficiencia del proceso de Minería de Datos [1].

3.3 ADQUISICIÓN DE LOS DATOS

Para este trabajo se realizó una encuesta la cual fue aplicada directamente a 177 turistas en la Zona de Cancún, Quintana Roo

A los cuáles se les aplicó un pre-procesamiento de limpieza considerando los siguientes aspectos:

- Selección relevante de datos

eliminando registros duplicados

eliminando anomalías

eliminando ruido (valores perdidos)

- Reducción de Datos:

selección de características

selección de instancias manualmente

eliminación de outliers (datos aislados)

Los datos procesados se convirtieron a formatos csv y arff para obtener su comportamiento.

4 RESULTADOS

En la Figura 6 se analizó el impacto de la edad respecto al tipo de turismo que prefieren resultando que los visitantes de edad de 18 a 30 prefieren el ecoturismo más que el turismo de aventura, en el caso de los visitantes de 30 a 40 prefieren ecoturismo.

[fig6.jpg]

Fig. 6. Resultados de aplicar Canopy.

En la Figura 7 nuevamente se analizaron las preguntas 13 y 2 con el algoritmo Cobweb dando como resultado que más turistas de todas las edades prefieren el ecoturismo, también se puede decir que los visitantes son en su mayoría estudiantes de 18 a 30 pero también menores de 18.

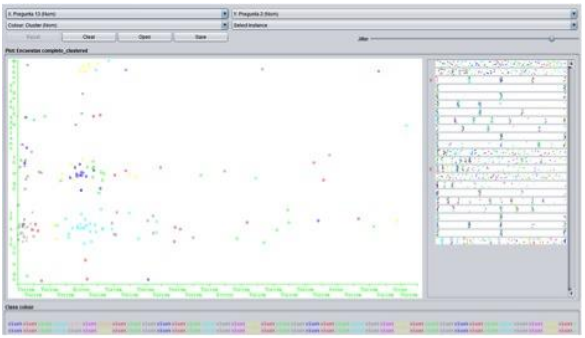

Fig. 7. Resultados de aplicar Cobweb.

Del análisis que se muestra en la Figura 8 se sabe que tanto turistas de 18 a 30 años como turistas de 30 a 40 se hospedaron en hotel, pero también notamos que la misma proporción de este segmento se hospeda con algún familiar o amigo. También se puede observar cómo se forman dos clusters muy concentrados con los siguientes resultados, en el rango de edades de 18 a 30 se observa que no se conectan más de 3 horas al internet por lo que se enfocan a la diversión que le ofrece el centro turístico y en el caso del rango de 30 a 40 años también coincide con utilizar el tiempo de la estancia en el entretenimiento en lugar de utilizar la computadora por motivos de trabajo y la población de 18 a 30 años prefiere utilizar tecnología touch antes que dispositivos tradicionales. la población de 18 a 30 años prefiere conectarse con mucha frecuencia la población de 30 a 40 prefieren sólo emplear la conexión a internet de forma moderada.

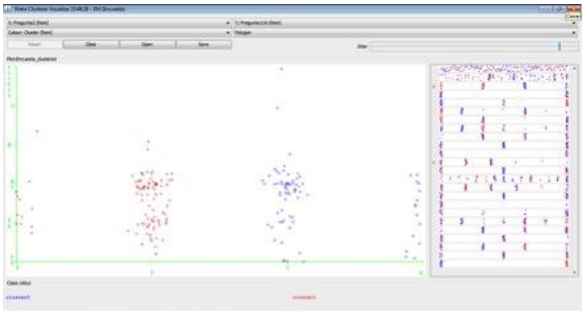

Fig. 8. Resultados de aplicar EM.

5 CONCLUSIONES

Con este sistema se pretende facilitar la vida a los encargados del área de mercadotecnia para recabar la información de sus clientes y si se requiere aplicar mecanismos de inteligencia de negocios para la toma de decisiones a partir de los datos proporcionados.

Como trabajo futuro se plantea realizar una interfaz para hacer consultas de Weka con interpretación personalizada para apoyar al mercadólogo en el aprendizaje de minería de datos.

REFERENCIAS

1. Frank, E., Witten, I.H. *Data Mining Practical Machine Learning Tools And Techniques*. California: Morgan Kaufmann Publishers. 2000
2. Ramírez, C., Hernández, J., Ramírez M.J. Introducción A La Minería De Datos, México: Pearson Prentice Hall. 2004
3. Girones, J., Minguillon, J., Caihuelas, R. *Minería de datos, modelos y algoritmos*, Colombia: editorial UOC. 2017

SEMBLANZA DE LOS AUTORES

M.C. Meliza Contreras González realizó sus estudios de Licenciatura en Ciencias de la Computación y Maestría en Ciencias de la Computación en la Benemérita Universidad Autónoma de Puebla, en el área de Computación Matemática, actualmente realiza estudios en el Doctorado en Ingeniería del Lenguaje y del Conocimiento, sus temas de interés son el procesamiento del lenguaje natural, la economía del comportamiento, las teorías de aprendizaje, los procesos de razonamiento, el web morphing, los chatbots, minería de datos, aplicaciones multimedia, aplicaciones web, web semántica, reading machine, diseño interactivo y experiencia de usuario, plataformas educativas, comercio electrónico, modelado semántico y patrones semánticos, grafos de conocimiento.

M.C. Pedro Bello López realizó sus estudios de Licenciatura en Ciencias de la Computación y Maestría en Ciencias de la Computación en la Benemérita Universidad Autónoma de Puebla, en el área de Programación de Sistemas, actualmente realiza estudios en el Doctorado en Ingeniería del Lenguaje y del Conocimiento, sus temas de interés son la revisión de creencias, la teoría de grafos, los conjuntos independientes, las teorías de aprendizaje, los procesos de razonamiento, cálculo de predicados, cálculo proposicional, minería de datos, aplicaciones multimedia, aplicaciones web, experiencia de usuario, plataformas educativas, grafos de conocimiento, ingeniería web, evaluación en sistemas instruccionales, diseño de reactivos.

Dr. Mario Rossainz López realizó sus estudios de Licenciatura en Ciencias de la Computación en la Benemérita Universidad Autónoma de Puebla. se tituló como licenciado en el año de 1994. Obtuvo el grado de Doctor en Métodos y Técnicas Avanzadas del Desarrollo de Software por la Universidad de Granada, España en el año 2005. Desde el año de 1994 a la fecha trabaja como Profesor Investigador en la Facultad de Ciencias de la Computación de la BUAP y soy miembro del cuerpo académico de Computación Distribuida cuyo estatus es consolidado. Sus áreas de interés son la Programación Concurrente y Paralela, los sistemas distribuidos, la ingeniería de Software, la Programación Orientada a Objetos y el desarrollo de Aplicaciones Web.

Brenda Karen Miranda Hernández actualmente realiza estudios en la Ingeniería en Ciencias de la Computación, se encuentra interesada en proyectos donde los procesos se puedan automatizar permitiendo optimizar tiempo y recursos a las personas y empresas mediante esquemas de inteligencia artificial, sus temas de interés son las aplicaciones móviles, web y programación en redes sociales, diseño inteligente, experiencia de usuario, comercio electrónico y las aplicaciones de inteligencia artificial, el aprendizaje automático, la minería de datos, big data, diseño interactivo, lms, redes de computadoras, chatbots, crawling, la ingeniería de software, la programación orientada a objetos y el desarrollo de aplicaciones web.