

Propuesta de arquitectura para un módulo de inteligencia de negocios basado en minería de datos

Yessica Thalia Apale Lara, Beatriz Alejandra Olivares Zepahua,
Lisbeth Rodríguez Mazahua, Giner Alor Hernandez, Hilarión Muñoz Contreras

Instituto Tecnológico de Orizaba, División de Estudios de Posgrado e Investigación,
Orizaba, Veracruz, México

yapalelara@acm.org, bolivares@ito-depi.edu.mx,
hmunozc189@msn.com, {lrodriguez,galor}@itorizaba.edu.mx

Resumen. Los constantes cambios en las preferencias de los viajeros, nacionales e internacionales influyen en el rendimiento de la industria hotelera creando así la necesidad de que los gerentes hoteleros continúen en la búsqueda de métodos que ayuden a predecir las preferencias y necesidades de los viajeros con el objetivo de obtener mejoras en su planificación y desarrollo de servicios estratégicos. Como solución a esta problemática el presente trabajo propone una arquitectura para el desarrollo e integración de un módulo de Inteligencia de Negocios en el sistema hotelero Softcalli, para que aplicando el proceso KDD se genere un modelo de Minería de Datos que permita identificar grupos de clientes con el fin de realizar la predicción de los servicios más idóneos de acuerdo al perfil de un nuevo cliente.

Palabras clave: Agrupamiento, hotel, inteligencia de negocios, KDD, minería de datos, WEKA.

Proposal of an Architecture for a Business Intelligence Module Based on Data Mining

Abstract. The constant changes in the preferences of national and international travelers influences the performance of the hotel industry thus creating the need for hotel managers to continue in the search of methods that help to predict the preferences and necessities of the travelers with the objective to obtain improvements in their planning and the development of strategic services. As a solution to this problem, the present work proposes an architecture for the development and integration of a Business Intelligence module in Softcalli hotel system, so that by applying the KDD process, a Data Mining model will be generated to identify groups of clients in order to predict the most suitable services according to the profile of a new customer.

Keywords: Clustering, hotel, Business Intelligence, KDD, data mining, Weka.

1. Introducción

La industria hotelera juega un papel muy importante dentro del entorno económico para la mayoría de los países en función de la demanda turística de cada país. La función principal de un Hotel es dar cabida a aquellos que se encuentran fuera de casa. Para muchos países los hoteles son un medio que ayuda a contribuir de manera significativa a la economía local, ya que son una fuente de ingresos en función de la demanda turística de cada país, por lo tanto, es conveniente que logre una ventaja competitiva ofreciendo los servicios de calidad que esperan los turistas [1].

El Banco de México registró que la llegada de turistas nacionales en México, en el mes de noviembre del 2015, fue de 7.6 millones, lo cual representa un incremento de 0.5 millones de viajeros nacionales respecto a noviembre de 2014, que corresponde a un crecimiento del 6.6%. Así mismo, la llegada de turistas internacionales a México, en noviembre del 2015 fue de 2.8 millones, lo que representa un incremento aproximado de 0.2 millones en comparación con noviembre del 2014, es decir 9.0% más.

Estos datos revelan un rápido crecimiento del Turismo en México y, como resultado, una fuerte demanda para el sector hotelero. En el Sistema Nacional de la Información Estadística del Sector Turismo de México (DATATUR) se establece que el porcentaje de ocupación hotelera en 70 centros turísticos, durante el periodo enero-noviembre de 2015, fue superior en 2.50 puntos porcentuales respecto al mismo periodo del año anterior, llegando a 59.6%. En enero-noviembre de 2015, la llegada de turistas nacionales a cuartos de hotel superó los 47.9 millones de turistas (76.2%); mientras que 15.0 millones fueron turistas extranjeros (23.8%) [2].

Uno de los problemas con los que se encuentran los gerentes hoteleros, en función de la alta demanda turística, es que cada vez es más difícil saber cuáles son las preferencias, necesidades y los servicios que esperan los turistas. Actualmente los hoteles en México cuentan con sistemas que permiten llevar la administración del hotel, almacenando información valiosa de las transacciones que se realizan dentro del hotel como son las reservaciones de habitaciones y los servicios extras que solicitan los huéspedes; esta información resultaría útil para solucionar los problemas de negocio si se contase con herramientas que identificaran comportamientos comunes en los huéspedes tal que el personal del hotel contase con una ayuda para ofrecer mejores servicios a sus clientes.

Microcalli del Golfo S.A de C.V, es una empresa que se dedica al desarrollo de sistemas hoteleros, Softcalli es un sistema hotelero integral desarrollado por esta empresa. El sistema Softcalli tiene como función principal administrar el hotel, almacenando altas cantidades de información relevante de los clientes, registra y administra en las bases de datos las reservaciones que realizan los clientes, también realiza el registros de información detallada de los clientes, consumos y servicios entre otros. Un problema detectado en Softcalli es que no cuenta con un módulo de Inteligencia de Negocios que sirva de apoyo para mejorar las estrategias del negocio hotelero.

Como solución a los problemas antes mencionados se propone una arquitectura para el desarrollo e integración al sistema Softcalli de un módulo de Inteligencia de Negocios, empleando técnicas de Minería de Datos como lo son la clasificación y el agrupamiento. Debido a que los hoteles no tienen identificados a sus tipos de clientes,

se aplicarán técnicas de agrupamiento (*clustering*) para formar conjuntos de clientes de acuerdo a una lista de características (variables relacionadas con consumo de servicios y reservaciones), seleccionadas con apoyo de un experto del negocio; una vez que el experto del negocio acepte los grupos identificados, se procederá a etiquetar a los clientes con dichos grupos. Después se aplicarán técnicas de clasificación para obtener un modelo capaz de predecir a qué grupo pertenece un nuevo cliente, de esta forma se conocerá el comportamiento en el consumo de los servicios y las características de la reservación que tendrá dicho cliente.

Este trabajo está estructurado de la siguiente forma: en la sección 2 se mencionan una serie de trabajos relacionados al tema propuesto, en la sección 3 se describen algunos conceptos relevantes a abordar en el desarrollo de este trabajo, en la sección 4 se describe la Arquitectura de la propuesta de solución, en la sección 5 se presentan las conclusiones y los trabajos a futuro.

2. Estado del arte

Con el objetivo de tener una investigación más detallada acerca de necesidades emergentes de los viajeros, en [3] utilizaron una técnica de Minería de Patrones Emergentes, este trabajo se centró en identificar los patrones emergentes de los viajeros internacionales. La Minería de Patrones Emergentes ayudó a descubrir los cambios y tendencias en la atención de los viajeros, se identificaron una serie de características de interés específicas y su usuario objetivo.

En [4], se presentó un estudio comparativo de 4 herramientas de acceso público para Minería de Datos: KNIME (*Konstanz Information Miner*), Orange, RapidMiner y WEKA (*Waikato Environment for Knowledge Analysis*, Entorno para Análisis del Conocimiento de la Universidad de Waikato). Se evaluó el desempeño de dichas herramientas con el fin de proporcionar a los analistas la herramienta y las técnicas a utilizar que arrojaran resultados con mayor precisión y de forma rápida. Los resultados obtenidos indicaron que no existe una sola herramienta o técnica que sea mejor que otra y sea aplicable a cualquier tarea de clasificación. Individualmente, el mejor resultado se logró con Weka y la técnica de árboles de decisión.

En [5] propusieron una novedosa técnica de clasificación *Neuro-fuzzy* para Minería de Datos, con el fin de analizar el método propuesto y comparar su rendimiento con dos potentes algoritmos de clasificación supervisados: Función de Base Radial de Redes Neuronales (RBFNN) y Sistema de Inferencia Adaptativa Neuro-difusa (ANFIS). El método propuesto utilizó una matriz "*fuzzification*". Se evaluó el rendimiento de estos métodos de clasificación en función de las diferentes medidas de desempeño tales como la precisión, el error de raíz cuadrática media, estadística kappa, falsos positivos, falsos negativos y precisión. El NFS (*Neuro-fuzzy system*, Sistema *Neuro-fuzzy*) propuesto fue capaz de mejorar significativamente los métodos de clasificación convencionales.

En [6] propusieron un marco de trabajo que investiga los dos modelos de clasificación más comunes que se utilizan en Minería de Datos para predecir el comportamiento de los clientes. El estudio se realizó en una campaña de mercadotecnia de un banco. Los datos se extrajeron del banco de datos de la Universidad de California en Irvine, los algoritmos evaluados fueron MLPNN (*Multilayer Perception Neural Network*) y NB (*Naïve Bayes*), estos algoritmos se evaluaron utilizando la herramienta

WEKA. Como resultado se encontró que el modelo que obtuvo el mejor rendimiento predictivo fue MLPNN con una tasa de precisión de 88.63%.

En [7] presentaron diferentes técnicas de Minería de Datos para estimar el costo del Software y se evaluaron los resultados de cada técnica. Los modelos propuestos fueron LR (*Linear Regression*), ANN (*Artificial Neural Network*), SRV (*Support Regression Vector*) y KNN (*K-Nearest Neighbors*). Los estudios realizados demostraron que el costo estimado del software dentro de estos modelos tuvo mayor velocidad y la precisión de los modelos algorítmicos fue alta. Las técnicas de Minería de Datos mejoran la precisión de la estimación de los modelos en muchos casos.

En [8], con el fin de transformar adecuadamente los datos en algo útil para el desarrollo de negocio inteligente desde la perspectiva de maximizar los beneficios de la empresa, se diseñó y desarrolló un marco universal para apoyo a las empresas inteligentes. Este marco se forma por un conjunto de técnicas de fusión y un motor de minería de datos basado en el concepto del modelo KID (*Knowledge Information Data*) el cual fue responsable de realizar la recolección universal de datos e información valiosa. El motor desarrollado recogió información significativa extraída e interpretada por algoritmos fusionados, y toda la información se convirtió en conocimiento valioso acerca de la situación de los clientes y las situaciones de los negocios. Este marco se desarrolló con el objetivo de obtener conocimiento, a partir de los datos, para el desarrollo de futuros negocios inteligentes. El desarrollo del motor de minería, basado en técnicas de fusión, es el componente esencial en el marco de negocios inteligentes.

En [9] propuso una extensión de la metodología de Minería de Opiniones Basadas en las Expresiones de Aspectos, termino introducido por Bing Liu, aplicada al ámbito del turismo, con el fin de descubrir las preferencias de los consumidores sobre los productos turísticos, en particular hoteles y restaurantes, obteniendo opiniones y críticas desde la Web. La contribución de la extensión consistió en el desarrollo de nuevas reglas para cubrir la aparición de aquellos aspectos compuestos por más de un término y para abarcar también los aspectos que aparecen más de una vez en una frase. Además se incluyó la formalización de conceptos y la creación de conjuntos de datos especiales, para la evaluación de las propuestas.

Como resultado del análisis realizado en el estado del arte se concluyó que el uso de la Minería de Datos, dentro de la industria hotelera, aplicada en el desarrollo de negocios inteligentes, es de gran ayuda como apoyo para la toma de decisiones, así como para el desarrollo de mejores estrategias de negocio, para así llevar a la empresa a un nivel competitivo más alto.

3. Conceptos

A continuación se describen los conceptos relevantes para el mejor entendimiento del presente documento.

3.1. Minería de Datos

La Minería de Datos (MD) es una herramienta que ofrece un conjunto de técnicas para extraer conocimiento a partir de grandes cantidades de datos. La MD se define como “el proceso de descubrir patrones interesantes y conocimiento desde grandes cantidades

de datos” [10]. Las fuentes de datos que nutren a la MD incluyen bases de datos transaccionales, almacenes de datos, la Web y otros repositorios de información como son archivos de texto, hojas de cálculo, archivos XML (*eXtensible Markup Language*, Lenguaje de Marcas Extensible) o datos que se transmiten al sistema dinámicamente.

3.2. Proceso KDD

El proceso KDD (*Knowledge Discovery from Data*, Descubrimiento de Conocimiento a partir de los Datos), es el proceso de extracción no trivial de información, previamente desconocida y potencialmente útil, a partir de los datos [10]. El proceso KDD implica una secuencia iterativa de una serie de pasos, en la figura 1 se presentan los pasos del proceso KDD.

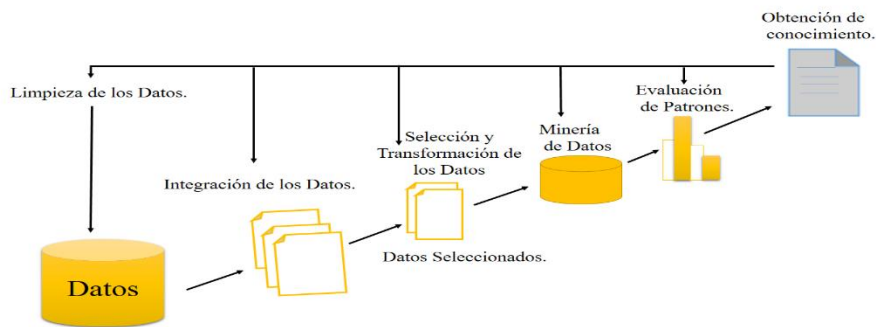


Fig. 1. Pasos del Proceso KDD.

3.3. Inteligencia de negocios

Se conoce como BI (*Business Intelligence*, Inteligencia de Negocios) al conjunto de estrategias, acciones y herramientas enfocada la administración y creación de conocimiento mediante el análisis de datos existentes dentro de una organización o empresa. Las operaciones que se definen como BI son aquellas que se involucran con los sistemas de información y con el desarrollo de métodos que generan un proceso de toma de decisiones inteligente [11].

3.4. Proceso ETL (Extracción, Transformación, y Carga)

ETL (*Extraction, Transformation & Load*, Extracción, Transformación y Carga), es el proceso que permite realizar la integración de la información en un Almacén de Datos (Data Warehouse, DW) o cualquier otro repositorio de datos. La Extracción se refiere a extraer la información de diferentes fuentes de información o Bases de Datos de distintos proveedores. En el proceso de Transformación, los datos extraídos pasan por varios procesos como son la discretización de los datos, limpieza de los datos, reemplazar valores faltantes, eliminación de ruido, entre otros procesos más, de modo que los datos se procesen correctamente y sean datos confiables. Por último, la Carga es el proceso de almacenar los datos ya transformados en el repositorio destino [12].

4. Propuesta de arquitectura para la integración del módulo de inteligencia de negocios al sistema Softcalli

Microcalli del Golfo es una empresa que desarrolla sistemas integrales para la administración de hoteles de cualquier tamaño. Muchos de sus clientes son pequeños hoteles que se encuentran en distintas ciudades del país y que no forman parte de grandes cadenas hoteleras, por lo que el presupuesto que destinan a la tecnología no es muy grande.

Softcalli es una suite de sistemas integrales hoteleros, desarrollada por la empresa Microcalli del Golfo, que cuenta con cuatro módulos con funciones específicas para cada área del hotel. Mk Front es el módulo que permite administrar las reservaciones del hotel, registrar clientes, realizar cambios de habitación, entre otras.

Esta suite se desarrolló con tecnología Microsoft .NET y utiliza una base de datos SQL Server. Con el fin de ofrecer los servicios más idóneos de acuerdo al perfil de cada tipo de turista se integrará a Softcalli un módulo de Inteligencia de Negocios que, aplicando técnicas de Minería de Datos, realice la predicción de los servicios que esperan los huéspedes.

Si bien SQL Server cuenta con herramientas propias para desarrollar Inteligencia de Negocio, específicamente Minería de Datos, dichas herramientas se venden por separado y tienen un costo muy alto, que en ocasiones lo hace prohibitivo para hoteles pequeños o mediados; por lo tanto, se ha considerado como alternativa el uso de la herramienta Weka, que es una aplicación y una biblioteca que contiene un gran número de algoritmos de Minería de Datos para diversas tareas como agrupamiento y clasificación entre otras; Weka trabaja bajo plataforma Java y que, en primer término, es incompatible de forma directa con la tecnología Microsoft.

La solución al manejo de plataformas heterogéneas (Microsoft y Java) será un módulo de Inteligencia de Negocios expuesto como WS (*Web Service*, Servicio Web) basado en SOAP (*Simple Object Access Protocol*, Protocolo Simple de Acceso a Objetos) que hará uso de la biblioteca Weka para aplicar algoritmos de agrupamiento y clasificación y así generar modelos de Minería de Datos que identifiquen los grupos de clientes específicos de un hotel y permitan predecir el comportamiento de nuevos clientes.

El módulo a desarrollar se alojará en un servidor de aplicaciones Glassfish y utilizará el lenguaje de programación Java. Por otra parte, se implementará en Softcalli un cliente del servicio Web como aplicación de escritorio, utilizando el lenguaje de programación C #, que consuma el WS para mostrar los resultados finales de la predicción.

En la figura 2 se muestra de forma general un esquema de solución para el desarrollo del módulo de Inteligencia de Negocios y su integración a Softcalli.

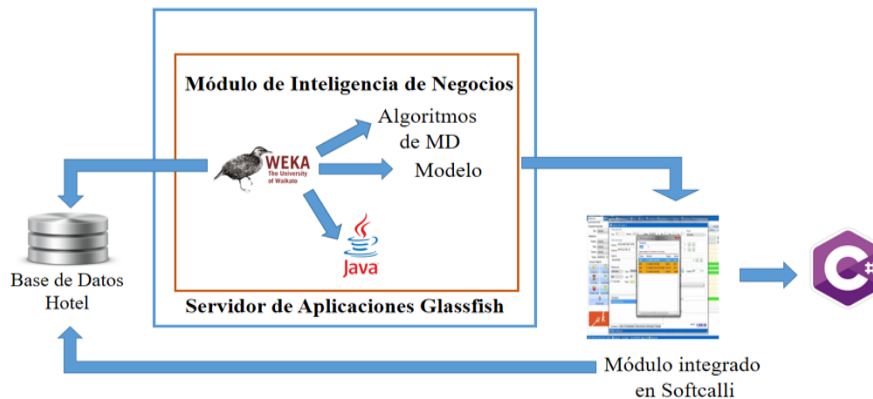


Fig. 2. Esquema de solución.

Es importante señalar que se cuenta con una base de datos (**Base de Datos Hotel** en la figura 2), que tiene almacenados registros históricos de las características y comportamientos de las reservaciones realizadas por los clientes; a partir de esa información, con la ayuda de los expertos de negocio, se identificaron, analizaron y seleccionaron aquellas variables que se consideraron importantes para realizar la identificación de los tipos de clientes y la predicción de los servicios.

En la figura 3 se muestra la estructura de la arquitectura planteada para el desarrollo de la aplicación y los componentes que la integran; esta arquitectura se basa en el Modelo Arquitectónico 4+1 Vistas y la figura corresponde a la vista de Despliegue. A continuación se describen los componentes de la arquitectura.

4.1. Cliente Softcalli.exe

En el sistema Softcalli se implementará el Cliente encargado de consumir los servicios Web que entregará el Proveedor Web Java, será desarrollado bajo la tecnología C# y Microsoft Net Framework 4.5. El Cliente Softcalli mostrará al usuario los algoritmos disponibles para aplicar las técnicas de agrupamiento (*clustering*) y posteriormente las técnicas de clasificación (predicción). Por otro lado esta aplicación también permitirá introducir los valores necesarios para realizar y mostrar una lista con los resultados obtenidos de la predicción, para este caso mostrará una lista con las características de la reservación, consumos y servicios que ofrecerá el hotel a un cliente nuevo.

4.2. Proveedor Web Java

Se trata del componente responsable de la Minería de Datos propiamente dicha, que hará llegar al cliente mediante un WS basado en SOAP. Entre sus elementos, el componente **Acceso a Datos** permite la comunicación con la **Base de Datos del hotel SQL** a través del uso de JDBC (*Java Database Connectivity*), el componente **Modelo** se auxiliará de él para acceder a los datos.

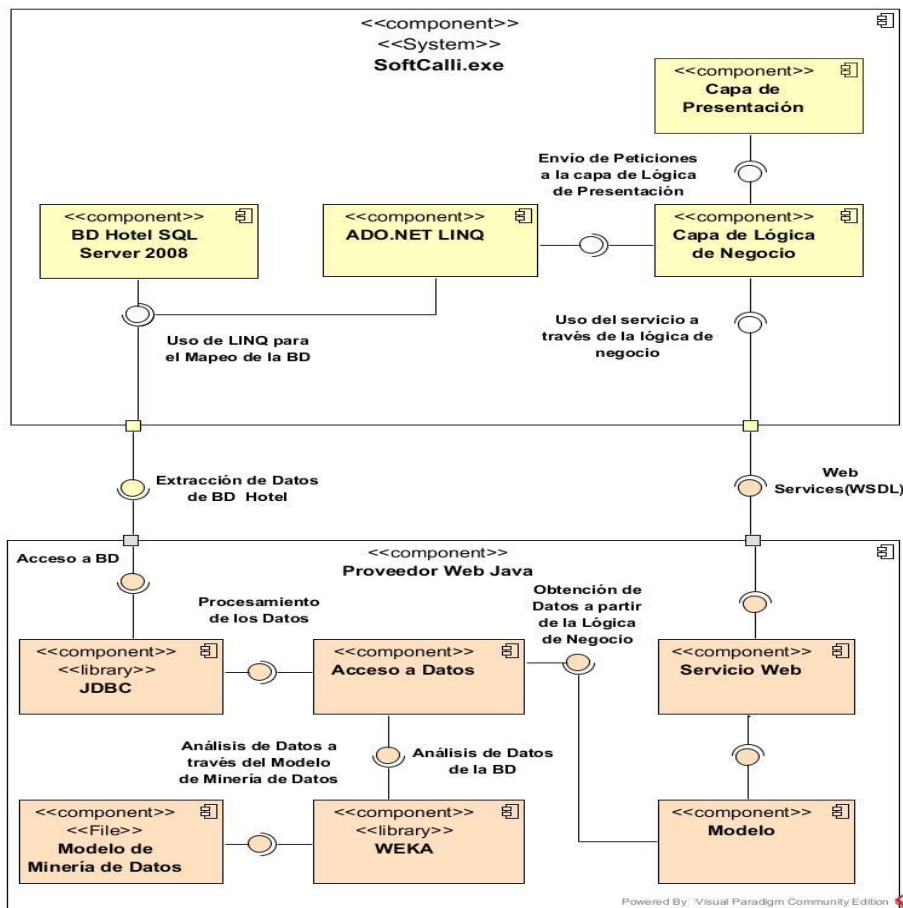


Fig.3. Diagrama de componentes.

En el **Modelo** se encuentran los Bean's que representan la lógica del módulo de Inteligencia de Negocio, donde se encontrarán a) el *ETL Bean*, que contendrá la lógica necesaria para aplicar el proceso ETL descrito en la sección anterior; b) el *Bean para el agrupamiento* y c) el *Bean para la clasificación* que implementarán la lógica necesaria para aplicar las técnicas de Minería de Datos con el objetivo de realizar la predicción de los servicios de acuerdo al perfil del cliente; por último d) el *Bean para aplicar algoritmos* y devolver las métricas de calidad asociadas para permitir al Gerente Hotelero realizar la elección de los algoritmos más óptimos para el proceso de Minería de Datos. Los algoritmos para las técnicas de agrupamiento y clasificación serán proporcionados por **WEKA**. La predicción se realizará en función de las diferentes variables de entrada que se hayan capturado en el módulo de Softcalli.

Para el caso de los algoritmos de agrupamiento se considerarán métricas de evaluación como la homogeneidad, el índice de confianza, la distribución del error y el índice David Bouldin. Para el caso de los algoritmos de clasificación, se considerarán métricas como el índice de certeza, la distribución del error y el error cuadrático. En

todos los casos, se presentarán tanto el valor de las métricas como su interpretación para que el gerente elija el algoritmo que dé los resultados que mejor reflejen la realidad de su negocio.

Para realizar la tarea de predicción, el Modelo de la Lógica de Negocio usará el componente **Modelo de Minería de Datos**, dicho componente contendrá un archivo con extensión .model que se obtiene como resultado de entrenar un conjunto de datos conformado por una lista de variables correspondientes a las características de los clientes cuya etiqueta de clase será el número de grupo, éstas variables son el país, estado y ciudad de procedencia, tarifa elegida, tipo de habitación, consumos, total de noches, total de días, origen de la reservación, si el huésped viene por asuntos de negocio, entre otras. El archivo .model se obtendrá bajo la aplicación del algoritmo que haya elegido el gerente hotelero. El Modelo de Minería de Datos contendrá el conocimiento de los patrones encontrados, y posteriormente se hará uso de él para realizar la predicción de ciertos servicios a ofrecer a un nuevo cliente; los servicios identificados son el tipo de habitación, tipo de tarifa, consumos como servicios de cuarto, restaurante, entre otros.

El Servicio Web proveerá todos los procesos necesarios que permitirán aplicar y regenerar, de ser necesario, los pasos del proceso KDD. Se propone el uso de Servicios Web para el desarrollo del módulo de Inteligencia de Negocios como mejor solución para la fácil integración del módulo con el sistema Softcalli. El servicio web se consumirá por un módulo integrado en Softcalli (Cliente) el cual como se mencionaba anteriormente está desarrollado en el lenguaje C#.

5. Conclusiones y trabajo a futuro

El uso de Minería de Datos apoya a las empresas a plantear mejores soluciones de negocio a los problemas relacionados con ofrecer servicios óptimos a sus clientes; esto es particularmente importante en industrias como la hotelera, donde se busca “enamorar” a los huéspedes al brindarles opciones de acuerdo a sus intereses y presupuesto. Si bien en México existen herramientas computacionales como Softcalli, orientadas a apoyar la operación diaria en los hoteles, en la mayoría de los casos carecen de módulos de Inteligencia de Negocio que exploten todo el potencial de conocimiento que se encuentra almacenado ya en sus bases.

Por otro lado, herramientas específicas de Minería de Datos, como la suite de Inteligencia de Negocios de SQL Server son caras y no todos los hoteles cuentan con el presupuesto para aplicarlas, lo que los coloca en una posición de desventaja frente a grandes cadenas hoteleras; alternativas como Weka proveen una opción viable a este problema, pero presentan la problemática de trabajar en una plataforma distinta a la que utilizan sistemas como Softcalli. En este punto entran tecnologías como los Servicios Web que apoyan en la interoperabilidad de plataformas heterogéneas. El presente trabajo mostró una arquitectura que considera todo lo anterior para proporcionar un módulo de Inteligencia de Negocio que apoye a los hoteleros a identificar mejor las necesidades de sus huéspedes.

Como trabajo a futuro se tiene la implementación de la arquitectura presentada en este documento, desarrollando los componentes indicados para construir el módulo de inteligencia de negocios del Sistema Softcalli, donde se utilizará por una parte C# y

SQL Server y por otra Weka, Java y Glassfish. Además, se refinarán los elementos complementarios del modelo de arquitectura 4+1 vistas (casos de uso, lógica, despliegue y procesos).

Agradecimientos. Al Consejo Nacional de Ciencia y Tecnología, CONACyT, por el apoyo económico otorgado para la realización de esta investigación.

Referencias

1. S. M. a. H. Ingram: *Staying Away from Home, The Business of Hotels*. New York, Elsevier Ltd, pp. 3–12 (2000)
2. Secretaría de Turismo (SECTUR). Available: <http://www.datatur.sectur.gob.mx/SitePages/versionesRAT.aspx> (2015)
3. Gang Li, Rob Law, Huy Quan Vu, Jia Rong, Xinyuan (Roy) Zhao: Identifying Emerging Hotel Preference Using Emerging Pattern Mining Technique. *Tourism Management*, Vol. 46, No. 1, pp. 311–321 (2015)
4. Borges, L. C., Marques, V. M., Bernardino, J.: Comparison of data mining techniques and tools for data classification. In: *Proceedings of the International C* Conference on Computer Science and Software Engineering (C3S2E '13)*, Vol. 34, No. 1, pp. 113–116 (2013)
5. Ghosha, S., Biswas, S., Sarkar, D., Pratim Sarkar, P.: A novel neuro-fuzzy classification technique for data mining. *Egyptian Informatics Journal*, Vol. 15, No. 3, pp. 129–147 (2014)
6. Femina Bahari, T., Sudheep Elayidom, M.: An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour. *Procedia Computer Science*, Vol. 46, No. 1, pp. 725–731 (2015)
7. Abbasi Khalifelum, Z., Soleimani Gharehchopogh, F.: Comparison and evaluation of data mining techniques with algorithmic models in software cost estimation. *Procedia Technology*, Vol. 1, No. 1, pp. 65–71 (2012)
10. Han, J., Kamber, M., Pei, J.: *Data Mining: Concepts and Techniques*. Waltham, USA, Elsevier Inc. (2012)
11. Ahumada Tello, A., Perusquia Velasco, J. M. A.: Business intelligence: Strategy for competitiveness development in technology-based firms. *Contaduría y Administración*, Vol. 61, No. 1, pp. 127–158 (2016)
12. Winda, R. T.: Extraction, Transformation, and Loading (ETL) Module for Hotspot Spatial Data Warehouse Using Geokettle. *Procedia Environmental Sciences*, Vol. 33, No. 1, pp. 626–634 (2016)