Descubrimiento de subgrupos aplicado al portal de comercio de electrónico: OrOliveSur.com

Carmona CJ *, del Jesus MJ, García S
Departamento de Informática

* ccarmona@ujaen.es | 953.21.19.56

Resumen

El descubrimiento de subgrupos es una técnica de minería de datos descriptiva capaz de describir conocimiento con una estadística inusual con respecto a una variable de interés en un conjunto de datos. Algoritmos basados en esta técnica se han aplicado a las visitas registradas por los usuarios del portal de comercio electrónico OrOliveSur.com, que se centra en la venta de aceite de oliva virgen extra de la comarca de Sierra Mágina.

Entre los resultados obtenidos cabe destacar la obtención de unos patrones de comportamiento por parte de los visitantes interesantes de cara al rediseño del portal web y así mejorar las ventas del mismo.

Abstract

Subgroup discovery is a descriptive data mining technique in order to describe knowledge with an unusual statistical with respect to an interest variable of the dataset. An algorithm based on subgroup discovery is applied to the visits registered in the e-commerce website OrOliveSur.com which is focused on the extra virgin olive oil from Sierra Magina.

Results obtained show behaviour patterns of the users very interesting with respect to the design of the website. The improvements indicated in this work could increase the orders of the e-commerce.

1. Introducción

El comercio electrónico es la compra-venta de productos o servicios mediante un medio electrónico, tales como internet o redes de computadores. Originalmente, este término se aplicó mediante la ejecución de transacciones como intercambio de datos electrónicos. Sin embargo, a mediados de los 90 con la aparición de internet se comenzó principalmente a realizar ventas de bienes y servicios en internet, utilizando primordialmente pagos electrónicos. La cantidad de pagos electrónicos ha crecido de forma exponencial en los últimos años. Una amplia variedad de comercios electrónicos han sido publicados en los últimos tiempos [Soares et al. 2008], estimulando la creación y utilización de innovaciones como transferencias electrónicas, marketing en internet, procesamiento de transacciones online, sistemas de recolección automática de datos, etc.
En Andalucía existe una alta concentración de cooperativas olivareras que en los últimos tiempos están proliferando en la exportación de sus productos [Moral-Pajares and Lanzas-Molina, 2009], y el uso de portales de comercio electrónico en las cooperativas y la adopción de Tecnologías de la Información y la Comunicación (TIC) son claves para estas exportaciones.

La utilización de las TICs surge para proponer metodologías de análisis inteligente de los datos para habilitar la extracción de conocimiento útil de los mismos [Fayyad et al, 1996]. Este es el concepto de Descubrimiento de Conocimiento en Grandes Bases de Datos (en inglés, Knowledge Discovery Databases – KDD), que fue definido como el proceso no trivial de identificación de patrones en los datos con las siguientes características: válido, novedoso, útil y comprensible [Han, 2005].

El proceso KDD es un conjunto de pasos interactivos e iterativos, incluyendo entre ellos el pre-procesamiento de los datos para corregir imprecisiones o inconsistencias, reducir el número de registros o encontrar las propiedades más representativas, minería de datos que es la etapa fundamental del proceso donde se extrae el conocimiento, y análisis y visualización de los resultados. KDD combina las técnicas tradicionales de la extracción de conocimiento con numerosos recursos desarrollados en el área de la inteligencia artificial.

En el proyecto abordado se ha descrito una metodología específica para extraer información útil de los datos de registros de usuarios registrados en el portal de comercio electrónico http://www.orolivesur.com. Estos datos de registros de usuarios de OrOliveSur han sido obtenidos mediante la herramienta Google Analytics.

OrOliveSur.com se centra en la venta a nivel nacional e internacional de aceite de oliva virgen extra de la comarca de Sierra Mágina. Las etapas llevadas a cabo en el análisis de este portal son las descritas previamente, es decir, una etapa de preprocesamiento para preparar los datos, extracción de conocimiento y análisis de los resultados obtenidos. A lo largo de este trabajo se presentará un resumen del portal de comercio electrónico OrOliveSur, de las diferentes técnicas y algoritmos de descubrimiento de subgrupos utilizados para obtener conocimiento

relacionado con el comportamiento de los usuarios en el portal, y para finalizar se presentan los resultados obtenidos en este estudio.

2. Materiales y Métodos

En esta sección se presentan las características más destacadas del portal OrOliveSur.com, las principales características de la minería de uso web y los propiedades y algoritmo de descubrimiento de subgrupos aplicados a los datos.

2.1. Portal de comercio electrónico: OrOliveSur.com

OrOliveSur es un proyecto nacido en la provincia de Jaén en Andalucía (España) en el año 2010. El principal propósito es dar a conocer en el mundo el tesoro de nuestra tierra, el aceite de oliva virgen extra. Esta web se centro en el aceite de oliva producido en un particular territorio de Jaén: El parque natural de Sierra Mágina. Esta zona es un área protegida de unas 50.000 hectáreas de parque natural formado por laderas boscosas, valles recónditos y picos montañosos escarpados. El pico más alto, Mágina es el más alto de la provincia de Jaén, llegando a los 2167 metros de altitud sobre el nivel del mar.

Ilustración 1. Página principal del portal web http://www.OrOliveSur.com

El amplio catálogo que presenta OrOliveSur se centra en la variedad de aceite picual. Esta variedad es la más extensa del mundo representando en España el

50% de la producción. La mayoría de esta se encuentra situada en Andalucía, especialmente en la provincia de Jaén. La aceituna es de un tamaño grande y con forma alargada con un pico al final de la misma. Los árboles de esta variedad son de un color plata intenso, abiertos y bien estructurados. Además, la variedad picual tiene unas propiedades excelentes ya que es la variedad con mejor estabilidad y ácido oleico con respecto a otras variedades como arbequina u hojiblanca, entre otras. En la actualidad, este portal de venta de aceite se encuentra traducido íntegramente al inglés, y parcialmente al alemán, francés y danés.

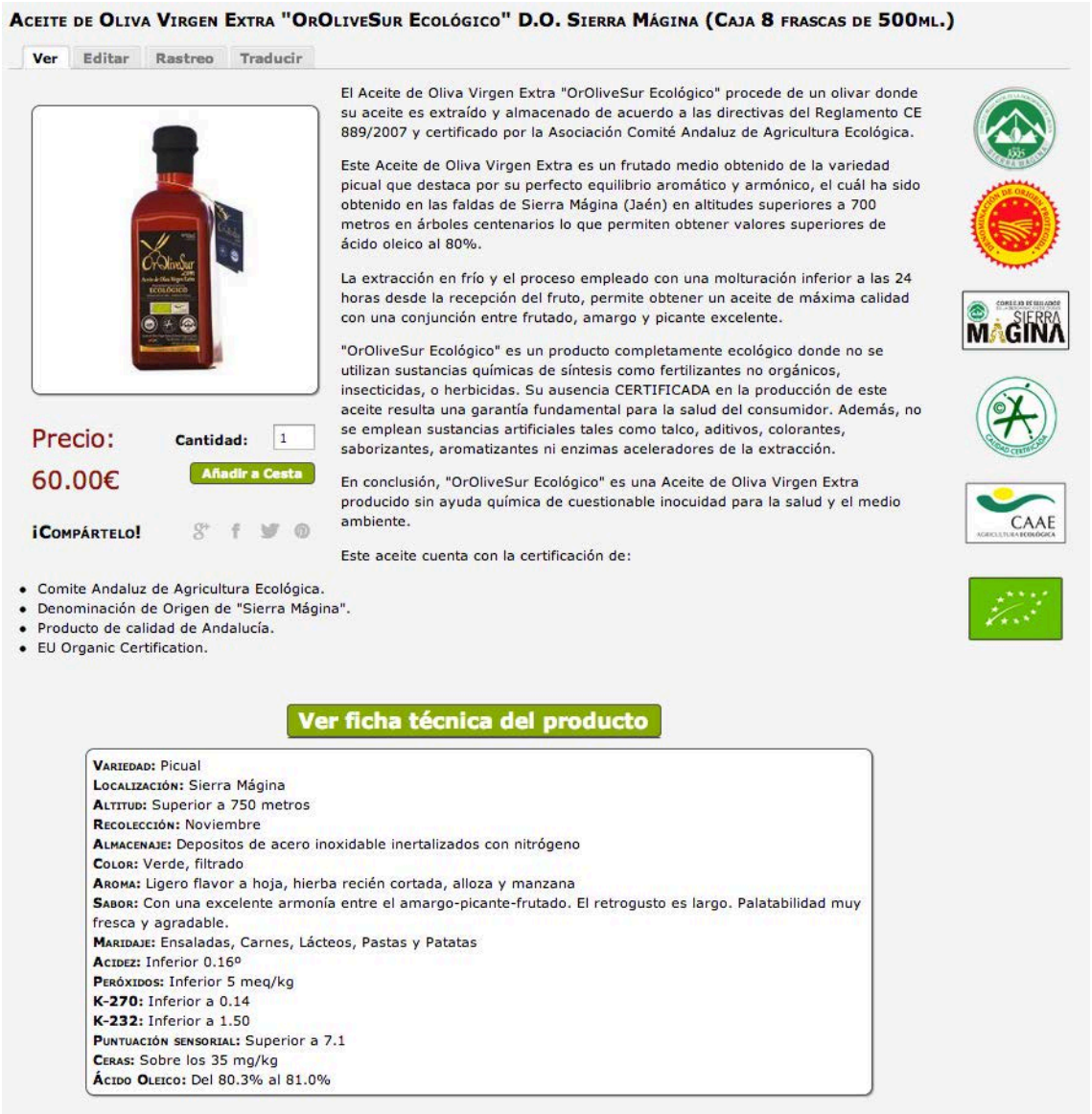

Ilustración 2. Descripción de un producto del portal web http://www.OrOliveSur.com

A lo largo de los últimos años, OrOliveSur ha recibido pedidos tanto nacionales como internacionales desde Dinamarca, Alemania, Reino Unido, Francia, etc., y sus pedidos y visitas incrementan día a día. La característica más destacada de este portal se relaciona con la calidad-precio de sus productos, pues se ofrecen productos de calidad avalados por el Consejo Regulador de la Denominación de Origen “Sierra Mágina” abaratando sus costes en envío y presentando múltiples métodos de pago. Todos los productos llevan una descripción detallada de los

mismos con respecto a propiedades para facilitar a los visitantes la elección de sus aceites. Por ejemplo, en la Ilustración 2 se puede observar la presentación de uno de sus productos.

2.2. Minería de uso web

Etzioni [Etzioni, 1996] definió minería web como el uso de técnicas para descubrir y extraer conocimiento en una web de forma automática, mientras Cooley [Cooley et al, 1999] fue más allá en remarcar la importancia de considerar el comportamiento y preferencias del usuario. En cualquier caso, los autores coinciden en separar la minería web en distintas etapas [Kosala and Bockeel, 2000] [Liu, 2006]:

- Encontrar recursos.
- Seleccionar la información y preprocesar.
- Descubrir el conocimiento.
- Analizar los patrones obtenidos.

La minería web se puede clasificar en tres dominios con respecto a la naturaleza de los datos [Cooley et al, 1997] [Markov and Larose, 2007]: minería web de contenido, minería de estructura de datos y minería de uso web.

En este proyecto nos centramos en la minería de uso web que fue definida por Srivastava [Srivastava et al, 2000] como:

El proceso de aplicar técnicas de minería de datos para el descubrimiento de patrones útiles desde los datos web.

Los patrones se representan como una colección de páginas o ítems visitados por los usuarios. Estos patrones se pueden emplear para comprender las principales características del comportamiento de los usuarios para mejorar la estructura de la web y crear recomendaciones personales y dinámicas sobre el contenido de la web [Mobasher, 2005]. La minería de uso web se puede emplear en diversas propuestas como por ejemplo para analizar secuencias de páginas, calidad de una web o búsquedas globales efectivas. Todas las propuestas han sido clasificadas con respecto a una taxonomía definida en [Facca and Lanzi, 2005]:

- Personalización cuyo objetivos está basado en la recomendación de sistemas.
- Pre-fetching y caching que intenta mejorar el rendimiento de los servidores y aplicaciones en la carga de páginas en caché antes que los usuarios las soliciten.
- Diseño que está relacionado con la usabilidad de una web. Estudios en diseño pueden proporcionar las metas para mejorar el diseño de la web.
- Comercio electrónico donde las técnicas utilizadas dentro de este grupo se relacionan con el *Customer Relationships Management*, que es un modelo de gestión que permite incrementar las ventas de los portales de comercio electrónico.

2.3. Descubrimiento de subgrupos

El concepto de descubrimiento de subgrupos fue introducido inicialmente por Kloesgen [Kloesgen, 1996] y Wrobel [Wrobel, 1997] y definido formalmente como [Wrobel, 2001]:

En descubrimiento de subgrupos, asumimos una población de individuos dada (objetos, clientes, ...) y una propiedad de estos individuos en la que estemos interesados. La tarea del descubrimiento de subgrupos es entonces descubrir los subgrupos de la población que son estadísticamente ``más interesantes'', es decir, individuos que sean tan grandes como sea posible y tenga una distribución estadística los más atípica posible, con respecto a una propiedad de interés.

El descubrimiento de subgrupos intenta buscar relaciones entre diferentes propiedades o variables de un conjunto con respecto a una variable objetivo. Debido a que el descubrimiento de subgrupos está centrado en la extracción de relaciones con características interesantes, no es necesario obtener relaciones completas sino que suele ser suficiente con relaciones parciales. Estas relaciones son descritas en forma de reglas individuales.

Así, una regla *R*, que consiste de una descripción de un subgrupo inducido, puede ser definida formalmente como:

R: Cond -> VarObj

donde VarObj es el valor de la variable de interés o variable objetivo para la tarea de descubrimiento de subgrupos (puede aparecer además en la bibliografía específica como Clase), y Cond es comúnmente una conjunción de funciones (pares atributo-valor) que es capaz de describir una distribución estadística inusual con respecto a la variable objetivo.

En una reciente revisión presentada por Herrera y otros [Herrera et al, 2011] se pueden observar los elementos fundamentales del descubrimiento de subgrupos, medidas de calidad utilizadas, algoritmos y aplicaciones a problemas reales.

A continuación se mencionan los principales elementos del descubrimiento de subgrupos, las medidas de calidad utilizadas en el proceso y el algoritmo empleado en este estudio.

2.3.1. Principales elementos del descubrimiento de subgrupos

Existen diferentes elementos a especificar en el diseño de un algoritmo de descubrimiento de subgrupos. Estos elementos se definen a continuación [Atzmueller et al, 2004]:

- Tipo de la variable objetivo. Se pueden encontrar diferentes tipos de variable objetivo: binaria, nominal o numérica. Para cada una de ellas se pueden aplicar diferentes análisis considerando el tipo de la variable objetivo.

- Lenguaje de descripción. La representación de los subgrupos debe ser adecuada para obtener reglas interesantes. Las reglas deben ser sencillas y por ello se suelen representar mediante pares atributo-valor generalmente en forma normal conjuntiva o disyuntiva. Además, los valores se pueden representar mediante valores positivos y/o negativos, mediante lógica difusa, o mediante el uso de desigualdades o igualdades, entre otros.

- Medidas de calidad. Éstas son un factor clave para la extracción de conocimiento ya que el interés del conocimiento extraído depende directamente de ellas. Además, las medidas de calidad proporcionan al experto la calidad e importancia de los subgrupos obtenidos. Se han presentado diferentes medidas de calidad en la bibliografía especializada [Gamberger and Lavrac, 2003][Kloesgen, 1996][Kloesgen and May, 2002][Lavrac et al, 2004], pero en ningún estudio previo se ha presentado un consenso sobre cuáles son las más adecuadas para usar en descubrimiento de subgrupos. En la siguiente sección se presenta un resumen de las medidas de calidad utilizadas.

- Estrategia de búsqueda. Este elemento es muy importante, ya que la dimensión del espacio de búsqueda tiene una relación exponencial respecto al número de propiedades y valores considerados. Hasta el momento se han utilizado diferentes estrategias, por ejemplo beam search, algoritmos evolutivos, búsqueda en espacios multirelacionales, etc.

2.3.2. Medidas de calidad empleadas en este estudio

Uno de los aspectos más relevantes para resolver un problema de descubrimiento de subgrupos es la elección de las medidas más adecuadas a utilizar para extraer las mejores reglas y evaluarlas. En la actualidad, existe un amplio número de medidas de calidad en la bibliografía. Las medidas más comunes dentro de esta tarea se describen a continuación:

- Confianza difusa: Determina la frecuencia relativa de los ejemplos que satisfacen tanto el antecedente como el consecuente de una regla entre aquellos que satisfacen sólo el antecedente [Del Jesus et al, 2007]. Se calcula como:

$$CnfD(R) = \frac{\sum_{E^k \in K / E^k \in VarObj} APC(E^k, R)}{\sum_{E^k \in K} APC(E^k, R)}$$

  donde APC es el grado de compatibilidad entre un ejemplo (E) y el antecedente de una regla difusa. En el caso de reglas no difusas, los grados de pertenencia son los correspondientes a conjuntos clásicos, es decir 0 ó 1. Esto llevaría a la obtención de los mismos valores, tanto para la confianza difusa, como para la nítida en problemas que contengan únicamente variables discretas.

- Relevancia: La relevancia de una regla se calcula en términos de su razón de verosimilitud, normalizada con la razón de verosimilitud del umbral de relevancia, y se mide como la relación de probabilidad de una regla [Kloesgen, 1996].

$$Rele(R) = 2 \cdot \sum_{k=1}^{n_c} n(VarObj_k \cdot Cond) \cdot log \frac{n(VarObj_k \cdot Cond)}{n(VarObj_k) \cdot p(Cond)}$$

  donde n(VarObj - Cond) es el número de ejemplos que satisfacen la condición y además pertenecen al valor de la variable objetivo en la regla, p(Cond) calculado como n(Cond)/n_s, se utiliza como un factor normalizador, n(Cond) es el número de ejemplos que satisfacen la condición determinada por el antecedente de la regla, n_s es el número de ejemplos, n(VarObj) es el número de ejemplos de la variable objetivo, y n_c es el número de valores de la variable objetivo. Aunque cada regla está definida para un valor específico de la variable objetivo se debe destacar que la medida de relevancia mide la novedad en la distribución imparcialmente, para todos los valores de esta variable.

- Sensibilidad: Esta medida mide la proporción de ejemplos correctamente descritos [Kloesgen, 1996]. Se puede calcular como:

$$Sens(R) = TPr = \frac{TP}{Pos} = \frac{n(VarObj \cdot Cond)}{n(VarObj)}$$

  donde Pos son todos los ejemplos del valor de la variable objetivo que se está analizando n(VarObj). Esta medida de calidad se utiliza para evaluar la calidad de los subgrupos en el espacio ROC (Receiver Operating Characteristic). La medida de sensibilidad combina la precisión y generalidad generada para un valor de la variable objetivo.

- Atipicidad: Esta medida se define como la precisión relativa con pesos [Lavrac et al, 1999]. Se puede calcular como:

$$Atip(R) = \frac{n(Cond)}{n_s}\left(\frac{n(VarObj \cdot Cond)}{n(Cond)} \cdot \frac{n(VarObj)}{n_s}\right)$$

  La atipicidad de una regla se puede describir como el balance entre la cobertura de la regla p($Cond_i$) y su ganancia de precisión p(VarObj - Cond) - p(VarObj).

2.3.4. NMEEF-SD

El algoritmo utilizado en este trabajo se denomina NMEEF-SD, que proviene de las iniciales de Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery [Carmona et al, 2010b]. Este algoritmo es un sistema difuso evolutivo, en [Herrera, 2008] se puede encontrar una amplia descripción de este tipo de sistemas.

El objetivo principal del NMEEF-SD es extraer subgrupos descriptivos difusos y/o nítidos (dependiendo de la naturaleza del problema a resolver) que aporten novedad, precisión e interpretabilidad al problema. El algoritmo utiliza medidas de calidad de reglas para guiar el proceso de aprendizaje, es decir emplea diferentes medidas como objetivos del proceso, y tiene como objetivo obtener reglas que alcancen valores adecuados no solo en estas medidas sino también en otros indicadores de calidad relacionados pero no considerados en este proceso de búsqueda. Además, este modelo permite elegir entre un conjunto de medidas como soporte, cobertura, relevancia, atipicidad y confianza, las medidas de calidad más adecuadas para resolver el problema planteado.

NMEEF-SD está orientado a resolver problemas de descubrimiento de subgrupos y por ello utiliza operadores para extraer subgrupos simples e interpretables, y con una alta calidad en las medidas estudiadas. Como el objetivo general de NMEEF-SD es obtener un conjunto de reglas, que deberían ser generales y precisas, el algoritmo incluye componentes que potencian estas características. Más concretamente, la diversidad se mejora en la población utilizando un operador de re-inicialización basada en cobertura, además de la técnicas de nichos (la distancia de crowding en el operador de selección). Para optimizar la generalidad de los subgrupos, el algoritmo incluye operadores de inicialización sesgada y mutación sesgada. Finalmente, para potenciar la precisión, además de los objetivos empleados por NMEEF-SD para guiar el proceso evolutivo y sobre las reglas, éste solo devuelve como soluciones finales aquellas reglas que alcancen un determinado umbral de confianza.

La estructura de las reglas utilizadas en el algoritmo NMEEF-SD está basada en el uso de la lógica difusa para la representación de las variables continuas. Las variables continuas son consideradas como variables lingüísticas, y los conjuntos difusos correspondientes a las etiquetas lingüísticas se pueden especificar por el usuario o definirse por medio de una partición uniforme si el conocimiento de los expertos no está disponible.

El algoritmo NMEEF-SD permite la obtención tanto de reglas difusas como nítidas, en función de la naturaleza de las variables del problema a estudiar. En caso de trabajar con variables continuas se obtendrán reglas difusas, si se trabaja con variables discretas se obtendrán reglas nítidas, y en caso de trabajar en un problema con ambos tipos de variables se obtendrán reglas que tendrán ambos componentes.

3. Resultados y Discusión

El principal propósito realizado en este trabajo se centra en el estudio del diseño de la web OrOliveSur.com mediante técnicas de minería de uso web. Estas técnicas son aplicadas dentro del proceso KDD que se divide en diferentes fases. En concreto, este estudio se realiza siguiendo las siguientes fases:

3.1. Recopilación y pre-procesamiento de los datos

Los datos son obtenidos mediante la herramienta Google Analytics desde el periodo 1 de enero a 31 de diciembre en el año 2011. Además, se aplican diversos filtros en el conjunto de datos de cara a obtener solo instancias con índices de rebote inferiores al 100%. Este valor es el porcentaje de visitas de una página única o visitas en las que la persona deja el portal en la misma página en la que llega, es decir, solo se consideran visita donde los usuarios han visitado la web durante más de un segundo. En total el conjunto de datos está compuesto por 8832 instancias, junto con distintas propiedades de las visitas que se detallan a continuación:

- Navegador: Esta propiedad contiene el nombre genérico del navegador utilizado por el usuario en su visita. Entre los posibles valores que se pueden encontrar se puede ver: Internet Explorer, Mozilla Firefox, Chrome, Safari, etc.

- Tipo de visitante: Contiene el tipo de visitante. Este valor puede contener el valor de nuevo visitante (N) o recurrente (R).

- Palabra clave: Es la palabra clave de acceso por parte del usuario a la web. Todas las palabras claves han sido clasificadas en seis categorías. Hay que remarcar que las palabras clave se pueden encontrar en distintos idiomas, pero todas ellas han sido clasificados siguiendo la traducción en el inglés:

    - Olive oil: Este valor contiene todas las palabras genéricas relacionadas con aceite de oliva, como por ejemplo: buy olive oil, venta de aceite, aceite ecológico, huile d'olive, etc.

    - Iberian product: En este valor se agrupan todas las palabras genéricas sobre productos ibéricos como jamón ibérico, comprar jamón de bellota, buy ibérico acorn-fed ham, etc.

    - Brand: Esta palabra contiene todas las entradas relacionadas a la marca de los productos del catálogo como La Casona, Verde Salud, Gámez Piñar, OrOlivesur, etc.

    - Gift: Contiene valores relacionados a regalos como boda, cestas de navidad, etc.

    - Other: Este valor agrupa todos los accesos con palabras clave no clasificada previamente.

    - Nothing: Los accesos sin palabras clave son clasificados con esta palabra clave como por ejemplo los accesos directos.

- Recurso: Esta propiedad indica el recurso utilizado por el visitante para acceder a la web:

    - Directo (D): Este valor se utiliza para accesos realizados directos en la web http://www.orolivesur.com

    - Motor de Búsqueda (E): Este valor se utiliza para accesos realizados a través de motores de búsqueda como Google, Yahoo o Bing, por ejemplo.

    - Correo (M): Indica el acceso realizado a través de correos electrónicos con un enlace a la web.

    - Referencia (R): Este valor se encuentra en accesos realizados desde otras webs con un enlace hacia OrOliveSur.

    - Redes Sociales (N): Contiene todos los accesos realizados a través de redes sociales como Facebook, Twitter, Google Plus, etc.

- Nuevas visitas: Indica el número de visitas nuevas realizadas con el mismo navegador, tipo de visitante, palabra clave y recurso.

- Páginas vistas: Indica el número de páginas vistas por el usuario con el mismo navegador, tipo de visitante, palabra clave y recurso.

- Tiempo por visita: Esta propiedad indica el tiempo empleado en la web por los usuarios con el mismo navegador, tipo de visitante, palabra clave y recurso.

- Visitas: Esta propiedad muestra el número de visitas realizadas con el mismo navegador, tipo de visitante, palabra clave y recurso.

- Páginas vistas únicas: Presenta el número de páginas únicas por los usuarios con el mismo navegador, tipo de visitante, palabra clave y recurso.

- Páginas vistas por visita: Muestra el número completo de páginas vistas por cada visita.

- Páginas vistas únicas por visita: Muestra el número completo de páginas únicas vistas por cada visita.

- Tiempo por página: Presenta el tiempo empleado por cada usuario por página vista.

3.2. Minería de datos

Una vez que los datos han sido preparados, ya están listos para pasar a la fase de minería de datos y aplicar el algoritmo NMEEF-SD.

El principal objetivo de la aplicación de NMEEF-SD es proporcionar al equipo de desarrolladores del portal web, información para mejorar el diseño de la web e

incrementar el número de visitas recibidas. En conclusión el objetivo es mejorar la visualización del portal y aumentar las ventas y clientes en el futuro. Esta técnica se ha utilizado en diferentes dominios y se han obtenido muy buenos resultados [Romero et al, 2009][Carmona et al, 2010a] [Carmona et al, 2011a][Carmona et al, 2011b][Carmona et al, 2013].

En la Tabla 1 se describen los parámetros utilizados por NMEEF-SD en el estudio realizado.

Tabla 1. Parámetros utilizados por el algoritmo NMEEF-SD

Tamaño de la población = 50 Número de evaluaciones = 10000 Probabilidad de cruce = 60% Probabilidad de mutación = 10% Confianza mínima = 0.6 Representación de las reglas = Canónicas Etiquetas lingüísticas = 9 {Bastante bajo, Muy bajo, Bajo, Normal, Alto, Muy Alto, Bastante Alto} Objetivo 1 = Sensibilidad Objetivo 2 = Atipicidad

3.3. Análisis y validación de los datos

En esta sección se presentan los resultados obtenidos por el algoritmo NMEEF-SD para los datos obtenidos de la web http://www.OrOliveSur.com.

Como ya hemos mencionado previamente, el objetivo del descubrimiento de subgrupos es obtener relaciones atípicas en los datos con respecto a una variable de interés u objetivo. En concreto para este problema, se analizan propiedades como palabras clave, recursos de tipo de visitante, por ejemplo como variable objetivo.

A continuación, los subgrupos más relevantes que se han obtenido en este estudio para el algoritmo NMEEF-SD con respecto a diferentes variables objetivo y sus medidas de calidad asociadas se muestran en la Tabla 2. En esta tabla se describen las reglas y las medidas de calidad relevancia (RELE), atipicidad (ATIP), sensibilidad (SENS) y confianza difusa (FCNF).

Tabla 2. Reglas y resultados obtenidos por NMEEF-SD

#	Regla	RELE	ATIP	SENS	FCNF
R1	SI recurso = E ENTONCES palabra clave = olive oil	1949.707	0.117	0.999	0.483
R2	SI recurso = E ENTONCES palabra clave = Brand	1949.707	0.073	1.000	0.303
R3	SI tiempo/páginas vistas= Bajo ENTONCES palabra clave = nothing	3.920	0.001	0.999	0.448
R4	SI tiempo = Bajo ENTONCES palabra clave = nothing	11.175	0.005	0,982	0.486
R5	SI palabra clave = nothing Y páginas vistas= Muy bajo Y páginas vistas = Muy bajo ENTONCES recurso = R	2216.810	0.090	0.996	0.373
R6	SI palabra clave = nothing Y únicas páginas vistas= Muy bajo ENTONCES recurso = R	2265.863	0.089	0.999	0.368
R7	SI palabra clave = nothing Y páginas vistas= Muy bajo Y page/visits = Muy bajo ENTONCES recurso = R	2216.810	0.090	0.996	0.372
R8	SI palabra clave = nothing Y únicas páginas vistas= Muy bajo Y únicas page/visits = Muy bajo ENTONCES recurso = R	2265.863	0.089	0.999	0.368
R9	SI tipo visitante = N Y únicas páginas vistas= Bajo ENTONCES recurso = E	90.077	0.038	0.658	0.653
R10	SI navegador = IE Y páginas vistas= Bajo ENTONCES recurso = E	137.419	0.057	0.575	0.709
R11	SI nuevas visitas = 0 ENTONCES tipo visitante = R	2819.825	0.229	1.000	1.000

Como se puede observar en los resultados obtenidos por NMEEF-SD, hay una gran número de reglas con valores aceptables en la mayoría de medidas de calidad. Aunque algunas reglas como R11 es obvia ya que si los visitantes no son nuevos el consecuente es que los usuarios son recurrentes, nos ayudan a mostrar el correcto funcionamiento del algoritmo.

Entre todas las reglas obtenidas por el algoritmo, es interesante remarcar que los usuarios que acceden directamente a la web, es decir sin utilizas palabras clave como indican las reglas R3 y R4, permanecen en la web durante un tiempo aceptable en la web y el tiempo por página es muy interesante. Además, las reglas R5, R6, R7 y R8 muestran que las páginas web que hacen referencia a OrOliveSur, tales como directorios o blogs, son visitas con número muy bajo de páginas vistas y páginas únicas vistas. En este sentido, el equipo de desarrolladores debe mejorar la descripción y la imagen de OrOliveSur en estas páginas porque es probable que los usuarios no encuentren lo que esperaban una vez llegan a la web.

Junto a todo esto, la regla más destacada descubierta por el algoritmo NMEEF-SD es la utilización del navegador Internet Explorer por la mayoría de usuario que visitan OrOliveSur mediante motores de búsqueda como Google o Yahoo, por ejemplo. Estos usuarios visitan un amplio número de páginas dentro del portal. En este sentido, recomendamos al equipo de desarrolladores a analizar el diseño de la web para comprobar que se muestra correctamente en este navegador en cualquier versión.

4. Conclusiones

En este trabajo se ha presentado un estudio basado en técnicas de minería de datos en datos, para analizar el acceso de usuarios a un portal de venta de aceite de oliva online. El propósito era extraer conocimiento sobre la información de acceso de los usuarios al portal de comercio electrónico OrOliveSur.com. Los datos han sido obtenidos mediante herramientas de analítica que facilitan la obtención de los mismos como Google Analytics.

La combinación de minería de datos en datos provenientes de acceso de usuarios en web, se cataloga como minería web. En concreto, en este estudio se ha presentado un estudio de minería de uso web realizado mediante el algoritmo NMEEF-SD para la obtención de subgrupos difusos con respecto a diferentes variables objetivo como recurso de acceso, palabra clave de acceso, etc.

Los resultados obtenidos muestran dos factores clave:

- Primero, el equipo de desarrolladores deben prestar especial atención a los visitantes que llegan desde páginas de referencia porque permanecen muy poco tiempo en el portal.
- Segundo, la mayoría de visitas vienen desde el navegador Internet Explorer. Además estas visitas son usuarios que navegan durante un buen periodo de tiempo a través de la web.

5. Agradecimientos

Este trabajo ha sido soportado por el Ministerio de Economía y Competitividad bajo el proyecto TIN-2012-33856 (Fondos FEDER), por el Plan Andaluz de Investigación bajo el proyecto TIC-3928 (Fondos FEDER), por el Plan de Investigación de la Universidad bajo el proyecto UJA2010/13/07 y patrocinado por la Caja Rural de Jaén.

6. Bibliografía

- [Atzmueller et al, 2004] Atzmueller, M., Puppe, F. & Buscher, H.P. (2004): Towards Knowledge-Intensive Subgroup Discovery. In Proceedings of the Lernen - Wissensentdeckung - Adaptivität - Fachgruppe Maschinelles Lernen, (pp. 111–117).
- [Carmona et al, 2010a] Carmona, C. J., González, P., Del Jesus, M. J., Romero, C., & Ventura, S. (2010). Evolutionary algorithms for subgroup discovery applied to e-learning data. In Proceedings of the IEEE international education engineering (pp. 983–990).
- [Carmona et al, 2010b] Carmona, C. J., González, P., Del Jesus, M. J., & Herrera, F. (2010). NMEEF-SD: Nondominated multi-objective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. IEEE Transactions on Fuzzy Systems, 18, 958–970.
- [Carmona et al, 2011a] Carmona, C. J., González, P., Del Jesus, M. J., Navío, M., & Jiménez, L. (2011). Evolutionary fuzzy rule extraction for subgroup

discovery in a psychiatric emergency department. Soft Computing, 15, 2435–2448.
- [Carmona et al, 2011b] Carmona, C. J., González, P., Del Jesus, M. J., & Ventura, S. (2011). Subgroup discovery in an e-learning usage study based on Moodle, In Proceedings of the international conference of European transnational education (pp. 446–451).
- [Carmona et al, 2013] Carmona CJ, Chrysostomou C, Seker H, del Jesus MJ. (2013). Fuzzy Rules for Describing Subgroups from Influenza A Virus Using a Multi-objective Evolutionary Algorithm. Applied Soft Computing, 13, 3439-3448.
- [Cooley et al, 1997] Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the World Wide Web. On Tools with Artificial Intelligence, 558–567.
- [Cooley et al, 1999] Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining World Wide Web browsing patterns. Knowledge and Information Systems, 1, 5–32.
- [Deb et al, 2002] Deb, K., Pratap, A., Agrawal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions Evolutionary Computation, 6, 182–197.
- [Del Jesus et al, 2007] Del Jesus, M. J., González, P., Herrera, F. & Mesonero, F. (2007) Evolutionary Fuzzy Rule Induction Process for Subgroup Discovery: A case study in marketing. IEEE Transactions on Fuzzy Systems, 15(4), 578–592.
- [Etzioni, 1996] Etzioni, O. (1996). The World Wide Web: Quagmine or gold mine. Communications of the ACM, 39, 65–68.
- [Facca and Lanzi, 2005] Facca, F. M., & Lanzi, P. L. (2005). Mining Interesting Knowledge from Weblogs: A Survey, 53, 225–241.
- [Fayyad et al, 1996] Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In Advances in knowledge discovery and data mining (pp. 1–34). AAAI/MIT Press.
- [Gamberger and Lavrac, 2003] Gamberber, D. & Lavrac, N. (2003) Active subgroup mining: a case study in coronary heart disease risk group detection. Artificial Intelligence in Medicine, 2003, 28(1), 27–57.
- [Han, 2005] Han, J. (2005). Data mining: Concepts and techniques. Morgan Kaufmann Publishers Inc.
- [Herrera, 2008] Herrera F. (2008). Genetic fuzzy systems: taxomony, current research trends and prospects. Evolutionary Intelligence, 1, 27–46.
- [Herrera et al, 2011] Herrera, F., Carmona, C. J., González, P., & Del Jesus, M. J. (2011). An overview on subgroup discovery: Foundations and applications. Knowledge and Information Systems, 29, 495–525.
- [Kloesgen, 1996] Kloesgen, W. (1996). Explora: A multipattern and multistrategy discovery assistant. In Advances in knowledge discovery and data mining (pp. 249–271). American Association for Artificial Intelligence.
- [Kloesgen and May, 2002] Kloesgen, W. & May, M. (2002) Census Data Mining - An application. In Proceedings of the 6th European Conference on principles of data mining and knowledge discovery, pp. 65–79.
- [Kosala and Bockeel, 2000] Kosala, R., & Bockeel, H. (2000). Web mining research: A survey. SIGKDD Explorations, 2, 1–15.

- [Lavrac et al, 1999] Lavrac, N., Flach, P. A. & Zupan, B. (1999) Rule Evaluation Measures: A Unifying View. In Proceedings of the 9th International Workshop on Inductive Logic Programming, vol. 1634 LNCS, pp. 174–185. Springer.
- [Lavrac et al, 2004] Lavrac, N., Cestnik, B., Gamberger, D. & Flach, P.A. (2004) Decision Support Through Subgroup Discovery: Three Case Studies and the Lessons Learned. Machine Learning, 57(1-2), 115–143.
- [Liu, 2006] Liu, B. (2006). Web data mining: Exploring hyperlinks, contents, and usage data (datacentric systems and applications). Springer-Verlag.
- [Markov and Larose, 2007] Markov, Z., & Larose, D. T. (2007). Data mining the web. Uncovering patterns in web content, structure and usage. Wiley-Interscience.
- [Mobasher, 2005] Mobasher, B. (2005). Web usage mining and personalization. CRC Press, LLC.
- [Moral-Pajares and Lanzas-Molina, 2009] Moral-Pajares, E., & Lanzas-Molina, J. R. (2009). La exportacion de aceite de oliva virgen en Andalucia: Dinamica y factores determinantes. Revista de Estudios Regionales, 86.
- [Romero et al, 2009] Romero, C., González, P., Ventura, S., Del Jesus, M. J., & Herrera, F. (2009). Evolutionary algorithm for subgroup discovery in e-learning: A practical application using Moodle data. Expert Systems with Applications, 36, 1632–1644.
- [Soares et al. 2008] Soares, C., Peng, Y., Meng, J., Washio, T., & Zhou, Z. H. (Eds.). (2008). Applications of data mining in e-business and finance. Frontiers in artificial intelligence and applications. IOS Press.
- [Srivastava et al, 2000] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations, 12–23.
- [Wrobel, 1997] Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In Proceedings of the 1st European symposium on principles of data mining and knowledge discovery (pp. 78–87). Springer.
- [Wrobel, 2001] Wrobel, S. (2001). Inductive logic programming for knowledge discovery in databases. Springer [Chapter Relational Data Mining, pp. 74–101].