R E V I S T A

INGENIERIA DE SISTEMAS

Volumen XXI Noviembre 2007

Publicada por el
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

R E V I S T A
INGENIERIA DE SISTEMAS

ISSN 0716 - 1174

Financiado parcialmente por el Proyecto Instituto Milenio "Sistemas Complejos de Ingeniería", como reconocimiento a la difusión de las materias abordadas y de sus participantes.

Instrucciones a los autores:
Los autores deben enviar 2 copias del manuscrito que desean someter a referato a: Comité Editorial, Revista Ingeniería de Sistemas, Av. República 701, Santiago, Chile. Los manuscritos deben estar impresos en hojas tamaño carta, a doble espacio, deben incluir un resumen de no más de 150 palabras, y su extensión no debe exceder las 30 hojas. Detalles en www.dii.uchile.cl/ ris

Correo electrónico: ris@dii.uchile.cl
Web URL: www.dii.uchile.cl/ ris

Representante legal: Rafael Epstein
Dirección: República 701, Santiago, Chile.
Diagramación: Francisco Cisternas
Impresión: Ka2 Diseño e Impresión
Mail: contacto@ka2.cl

Carta Editorial Volumen XXI

Nos es muy grato editar este nuevo número de la Revista de Ingeniería de Sistemas (RIS) dedicado a temas de frontera en Investigación de Operaciones, Gestión y Tecnología. Queremos agradecer al Instituto Milenio "Sistemas Complejos de Ingeniería" y al Departamento de Ingeniería Industrial de la Universidad de Chile por su colaboración para hacer posible esta publicación.

Este número contiene artículos de académicos y estudiantes de nuestro Departamento de Ingeniería Industrial (algunos de ellos incluso son consecuencia de trabajos finales de grado o de tesis de magister), y un artículo escrito por académicas de la Universidad Nacional de Santiago del Estero, en la Argentina.

La revista muestra trabajos sobre la aplicación de técnicas modernas de gestión de operaciones, programación matemática, web-mining, data-mining, redes neuronales y sistemas de información.

Nuestro objetivo a través de esta publicación es contribuir a la generación y difusión de las tecnologías modernas de gestión y administración. La revista pretende destacar la importancia de generar conocimiento en tecnología y administración para nuestras problemáticas, junto con adaptar las tecnologías foráneas a las realidades de nuestro país y de otros similares.

Estamos seguros de que los artículos publicados en esta oportunidad muestran formas de trabajo innovadoras que serán de gran utilidad e inspiración para todos los lectores, ya sean académicos o profesionales, por lo que esperamos que esta iniciativa tenga la recepción que creemos se merece.

Rafael Epstein
Guillermo Durán
Editores

Una metodología para mejorar el contenido de un sitio web a partir de la identificación de sus web site keywords

José I. Fernández*
Juan D. Velásquez*

Resumen

Presentamos una metodología para identificar aproximadamente qué palabras atraen la atención del usuario cuando se encuentra visitando páginas de un sitio web. Estas palabras son llamadas "web site keywords" y pueden ser usadas para la creación de contenidos relacionados a un tópico específico con el que se pretende atraer la atención del usuario.

A través de la utilización de las palabras correctas, se puede mejorar gradualmente el contenido de un sitio web, ayudando de esta forma a los usuarios a encontrar la información que buscan, lo cual se considera clave para el éxito y continuidad del sitio.

Aplicando algoritmos de clustering, y asumiendo que existe una correlacion entre el tiempo invertido en una página y el interés del usuario, se realiza una segmentación de los usuarios según comportamiento de navegación y preferencias de contenidos. A continuación, se identifican las palabras clave del sitio web. Esta metodología fue aplicada en datos originada desde un sitio web real, mostrando su efectividad.

Palabras Clave: Web site keywords, Clustering, Comportamiento del usuario web.

*Departamento de Ingeniería Industrial, Universidad de Chile

1. Introducción

Para muchas compañías y/o instituciones, ya no es suficiente el desarrollo de un sitio web para ofrecer sus productos y servicios en el mercado digital. Lo que a menudo hace la diferencia entre un éxito o fracaso en un e-business es el potencial del sitio web para atraer o retener usuarios. Este potencial depende del contenido del sitio, diseño y aspectos técnicos como tiempo de descarga de páginas del sitio hacia navegador del usuario, entre otros. En términos de contenido, las palabras usadas en el texto libre en las páginas de un sitio web son muy importantes, por cuanto dicen relación con la información que los usuarios buscan. En efecto, la gran mayoría de lo usuarios recurre a motores de búsqueda, tales como Yahoo! y Google, para realizar consultas respecto de un contenido de su interés, a través de consultas basadas en términos en motores de búsqueda para encontrar información en la Web. Estas consultas son realizadas usando palabras clave, es decir, una palabra o grupo de palabras [14] que caracterizan el contenido de un página web dada o un sitio web.

El correcto uso de las palabras con que se crea el contenido textual de una página web, mejora la información presentada a los usuarios, ayuda a la búsqueda efectiva de información, mientras atrae a nuevos usuarios y retiene a los actuales, mediante actualizaciones continuas del contenido textual de la página. El desafío, entonces, es identificar qué palabras son importantes para los usuarios. Lo anterior tiende a relacionarse con cual es "palabra más frecuentemente usada". Algunas herramientas comerciales[1] ayudan a identificar palabras clave objetivo que los consumidores son más propensos a utilizar mientras realizan sus búsquedas en la Web [6].

Mediante la identificación de las palabras más relevantes en las páginas de los sitios, desde el punto de vista del usuario, las mejoras pueden ser realizadas en el sitio web completo. Por ejemplo, el sitio puede ser reestructurado colocando un nuevo hyperlink relacionado con la palabra clave y por supuesto el contenido textual podría ser modificado utilizando las palabras clave relacionadas con un tópico específico para enriquecer el texto libre en una página Web.

En este trabajo se presenta una metodología para analizar el comportamiento de navegación del usuario y sus preferencias de contenido a través de la aplicación de algoritmos de web mining en datos originados en la web, también llamados web data, específicamente registros de un sitio web (web logs) y su contenido textual.

La metodología apunta a identificar aproximadamente cuales palabras atraen

[1] Ver por ejemplo http://www.goodkeywords.com

la atención del usuario cuando esta visitando páginas en un sitio web. Estas palabras son denominadas "palabras clave de un sitio web" [31] y pueden ser utilizadas para la creación de contenidos de texto mejoradas relacionadas con tópicos específicos.

Este paper esta organizado de la siguiente forma: La sección 2 introduce una revisión breve acerca del trabajo relacionado. El proceso de preparación para transformar la web data en vectores de características para ser utilizados como entrada en los algoritmos de web mining es mostrado en la sección 3. En la sección 4, la metodología para identificar las palabras clave de un sitio web es explicada y aplicada en la sección 5. Finalmente, la sección 6 muestra las conclusiones principales de este paper.

2. Trabajos Previos

Cuando un usuario visita un sitio web, datos respecto de qué página visitó son almacenados en archivos de registro llamados web logs. Entonces es directo conocer cuáles páginas son visitadas y cuáles no, e inclusive el tiempo gastado por el usuario en cada una de ellas. Debido a que usualmente las páginas contienen datos acerca de un tópico específico, es posible conocer aproximadamente las preferencias de información de los usuarios. En ese sentido la interacción entre el usuario y el sitio es como una indagación electrónica, entregando los datos necesarios para analizar las preferencias de contenido del usuario en un sitio web particular.

El desafío para analizar las preferencias de texto del usuario en el sitio web es doble. Primero la cantidad de registros en el archivo web log usualmente es enorme, y una parte son datos irrelevantes acerca del comportamiento de navegación del usuario en el sitio. Segundo, el texto libre dentro de las páginas web es comúnmente plano, es decir, sin información adicional que permita conocer directamente cuáles son las palabras que atraen la atención del usuario.

En esta sección se revisan las principales aproximaciones para analizar las web data para extraer patrones significativos relacionados con las preferencias de texto de los usuarios en el sitio web.

2.1. Minando los web data

Las técnicas de web mining emergieron como resultado de la aplicación de teoría de data mining al descubrimiento de patrones desde los web data [8, 16, 25]. El web mining no es una tarea trivial considerando que la Web es una enorme colección de información heterogénea, no clasificada, distribuida, variante en el tiempo, semi estructurada y altamente dimensional. El web mining debe considerar tres importantes pasos: Preprocesamiento, descubrimiento de

patrones y análisis de patrones [27].

Las siguientes terminologías comunes son utilizadas para definir los diferentes tipos de web data.

- Contenido. El contenido de la página web, es decir, imágenes, texto libre, sonidos, etc.
- Estructura. Información que muestra la estructura interna de una página web. En general, tienen etiquetas HTML o XML, alguna de las cuales contienen información acerca de hipervínculos con otras páginas web.
- Uso. Información que describe las preferencias del visitante mientras navega en un sitio web. Es posible encontrar esta información dentro de los archivos web log.
- Perfil del usuario. Colección de información acerca del usuario: Información personal (nombre, edad, etc.), información de uso (por ejemplo, páginas visitadas) e intereses.

Con las definiciones anteriores, y dependiendo de los web data a procesar, las técnicas de web mining pueden ser agrupadas en tres áreas: Minado de contenido web (WCM o Web Content Mining), Minado de la estructura web (WSM o Web Structure Mining), y Minado de la utilización de la web (WUM o Web Usage Mining).

2.1.1. Identificando palabras para la creación de un resumen automático de texto de una página web

La meta es construir automáticamente resúmenes de lenguaje natural de documento [11]. En este caso, una semi estructura relativa es creada por la aplicación de etiquetas HTML desde el contenido textual de una página web, la cual examina temas sin restricción de dominio. En muchos casos, las páginas pueden solamente contener pocas palabras sin elementos textuales (por ejemplo video, imágenes, audio, etc.) [1].

En la investigación de resumen de texto, tres importantes aproximaciones son [18]: basadas en parágrafos, basadas en oraciones y utilización de señales de lenguaje natural en texto.

La primera aproximación consiste en seleccionar un párrafo de un segmento de texto [19] que apunta a un tema en el documento, bajo la suposición que hay varios temas en el texto. La aplicación de esta técnica en una página web no es obvia; los diseñadores web tienen la tendencia de estructurar el texto en párrafos por página. Por lo tanto un documento contiene un solo tema, lo cual hace la aplicación de esta técnica dificultosa.

En la segunda aproximación, las frases más interesantes o frases clave son extraídas y ensambladas en un texto individual [9,37]. Es claro que el texto

resultante puede no ser cohesivo, pero la meta de la técnica es proveer la máxima expresión de información en el documento. Esta técnica es aplicable para páginas web, dado que la entrada puede consistir de pequeñas piezas de texto [6]. La aproximación final es un modelo de discurso basado en la extracción y resumen [14,15] mediante la utilización de señales de lenguaje natural como identificación de nombres propios, sinónimos, frases claves, etc. Este método arma oraciones mediante la creación de una colección de texto con información del documento completo. Esta técnica es más apropiada para documentos dentro de un dominio específico y esto para la implementación en un sitio web es dificultoso.

2.2. Extracción de texto de páginas web y aplicaciones

Las componentes de texto clave son partes de un documento completo, por ejemplo un párrafo, frase y una palabra que contiene información significativa acerca de un tema particular, desde el punto de vista del usuario del sitio web. La identificación de estos componentes puede ser útil para mejorar el contenido textual de un sitio web.

Usualmente, las palabras clave en un sitio web están correlacionadas con las "palabras más frecuentemente utilizadas". En [6], se introduce un método para la extracción de las palabras clave desde un gran conjunto de páginas web. La técnica está basada en la asignación de importancia a las palabras, dependiendo de su frecuencia en todos los documentos. Seguidamente, los párrafos o frases que contienen las palabras clave son extraídos y su importancia es validada a través de pruebas con usuarios reales.

Otro método, en [2], recolecta palabras clave desde un motor de búsqueda. Esto muestra las preferencias globales de palabras de una comunidad web, pero no brinda detalles acerca de un sitio web particular.

Finalmente, en lugar de analizar palabras, en [17] se desarrolla una técnica para extraer conceptos desde el texto de una página web. Los conceptos describen objetos del mundo real, eventos, pensamientos, opiniones e ideas en una estructura simple, como términos descriptivos. Entonces, utilizando el modelo de vector espacial, los conceptos son transformados en vectores de características, permitiendo la aplicación de algoritmos de clustering o clasificación a páginas web.

3. Proceso de preparación de la Web Data

De toda la información web disponible, la más relevante para el análisis del comportamiento y preferencias de navegación del usuario, son los registros (web logs) y las paginas web [33]. Los web logs contienen información acerca

de la secuencia de navegación de páginas y el tiempo gastado en cada página visitada, aplicando el proceso de sesionización. La fuente de la página web es el sitio web en si mismo. Cada página web es definida por su contenido, en particular texto libre. Para estudiar el comportamiento del usuario ambas fuentes - web logs y páginas web - se preparan mediante la utilización de filtros y por la estimación de sesiones reales de usuario. La etapa de preprocesamiento implica, primero, un proceso de limpieza y, segundo, la creación de vectores de características como entrada a los algoritmos de web mining, dentro de la estructura definida por los patrones vistos.

3.1. El proceso de reconstrucción de sesiones

El proceso de segmentación de las actividades de usuarios en sesiones individuales es llamado ***sesionización*** [10] y está basado en los web logs del sitio web. En consideración de los inconvenientes mencionados anteriormente, el proceso no esta libre de errores [26]. La sesionización asume que la sesión tiene un tiempo de duración máximo y que no es posible saber si el visitante ha presionado el botón "volver" (back) en el navegador del sitio web. Si la página esta en el cache del navegador y el visitante vuelve a ella en la misma sesión, podría no quedar registrada en los logs del sitio web. Por esto han sido propuestos el uso de esquemas invasivos como el envío de otra aplicación al browser para capturar el comportamiento de navegación exacto del usuario [3, 10]. Si embargo, este esquema podría ser fácilmente evitado por el visitante.

Muchos autores [3, 10, 20] han propuesto la utilización de heurísticas para la reconstrucción de sesiones por los web logs. En esencia, la idea es crear subgrupos con las visitas de usuarios y aplicando mecanismos sobre los web logs generados para permitir la definición de una sesión como series de eventos entrelazados durante un cierto periodo de tiempo.

La reconstrucción de sesiones apunta a encontrar sesiones de usuarios reales, es decir, cuales páginas fueron visitadas por un ser humano. En ese sentido, cualquiera sea la estrategia utilizada para descubrir las sesiones reales, esta debe satisfacer dos criterios esenciales: las actividades realizadas por una persona real pueden ser agrupadas entre si y el conjunto en actividades que pertenecen a la misma visita (otros objetos requeridos por la página web visitada) también pertenecen al mismo grupo.

Hay varias técnicas para sesionización, las cuales pueden ser agrupadas en dos estrategias mayores: *proactiva y reactiva* [26].

Las **Estrategias Proactivas** intentan identificar el usuario utilizando métodos de identificación como cookies que consisten en una pieza de código asociado al sitio web. Cuando el visitante ingresa al sitio por primera vez, una cookie es enviada al navegador. Luego, cuando la página es revisitada, el navegador muestra el contenido de la cookie al servidor web y automática-

mente la identificación toma lugar. El método tiene problemas desde el punto de vista técnico y también con respecto a la privacidad del usuario. Primero, si el sitio es revisitado después de varias horas, la sesión será considerada muy larga, y será entonces una nueva sesión. En segundo lugar, algunos aspectos de las cookies parecen incompatibles con los principios de protección de datos de algunas comunidades, como la Unión Europea [26]. Finalmente, las cookies pueden ser fácilmente detectadas y desactivadas pro el visitante.

Las Estrategias Reactivas son no invasivas con respecto a la privacidad y hacen uso de la información contenida sólo en los web logs y consiste en el procesamiento de los registros para generar un grupo de sesiones reconstruidas.

En el análisis del sitio web, el escenario general es que los sitios web usualmente no implementan mecanismos de identificación. La utilización de estrategias reactivas puede llegar a ser más útil. Estas pueden ser clasificadas en dos grupos principales [4, 10]:

- Heurísticas orientadas a la navegación: asumen que el visitante llega a páginas a través de hyperlinks desde otras páginas. Si el requerimiento de una página es inalcanzable a través de las páginas previamente visitadas por el usuario, una nueva sesión es iniciada.

- Heurísticas Orientadas al tiempo: se coloca un tiempo máximo de duración, que es usualmente 30 minutos para la sesión completa [7]. Basado en este valor se pueden identificar las transacciones pertenecientes a una sesión específica utilizando filtros programados.

3.1.1. Procesando el contenido textual de una página web

Hay varios métodos para comparar el contenido de dos páginas web, aquí se considera el texto libre dentro de las páginas web. El proceso común es coincidir los términos que componen el texto libre, por ejemplo, mediante la aplicación de un proceso de comparación de palabras. Un análisis más complejo incluye información semántica contenida en el texto libre que involucra también una tarea de aproximación de términos comparados.

La información semántica es fácil de extraer cuando el documento incluye información adicional acerca del contenido del texto, por ejemplo, etiquetas de marcado. Algunas páginas web permiten la comparación de documentos mediante la información estructural contenida en las etiquetas HTML, incluso con restricciones. Este método es utilizado en [28] para comparar páginas escritas en lenguajes diferentes con una estructura HTML similar. La comparación es enriquecida por la aplicación de un proceso de equiparar el contenido textual [29], el cual considera una tarea inicial de traducción a ser completada. El método es altamente efectivo cuando el lenguaje utilizado es el mismo en las páginas que se encuentran en comparación. Una breve encuesta de algorit-

mos para comparar documentos por la utilización de estructuras similares es encontrada en [5].

Las comparaciones son realizadas por una función que retorna un valor numérico mostrando similitudes o diferencias entre dos páginas web. Esta función puede ser utilizada en algoritmos de web mining para procesar un conjunto de páginas web, las cuales pueden pertenecer a una comunidad web o un sitio web aislado. El método de comparación debe considerar un criterio de eficiencia en el procesamiento de contenido de páginas web [13]. Aquí el modelo de vector espacial [24], permite una representación vectorial simple de las páginas web y. mediante el uso de comparación de distancia entre vectores, provee de una medida de las diferencias y similitudes entre páginas.

Las páginas web deben ser limpiadas antes de transformarlas en vectores, tanto para reducir el número de palabras - no todas las palabras tienen el mismo peso - y hacer el proceso más eficiente. Por esto, el proceso debe considerar los siguientes tipos de palabras:

- Etiquetas HTML: En general, estas deben ser limpiadas. Sin embargo, la información contenida en cada etiqueta puede ser utilizada para identificar palabras importantes en el contexto de una página. Por ejemplo, la etiqueta "<titulo>" enmarca el tema central de la página web, es decir, de una noción aproximada del significado semántico de la palabra y, es incluida en la representación vectorial de la página.

- Palabras de detención. (por ejemplo pronombres, preposiciones, conjunciones, etc.).

- Stem de palabras. Después de aplicar el proceso de remoción del sufijo de la palabra (stemización de la palabras [22]), obtenemos la raíz de la palabra o stem.

Para el propósito de representación vectorial, sea R el número total de palabras diferentes y Q el número de páginas en el sitio web. Una representación vectorial del conjunto de páginas es una matriz M de tamaño R x Q.

$$M = (m_{ij}), \text{ con } i = 1, \ldots, R \text{ y } j = 1, \ldots, Q \tag{1}$$

Donde m_{ij} es el peso de la palabra i en la página j.
Basado en *tfxidf-weighting* introducido en [24] los pesos son estimados como:

$$m_{ij} = f_{ij}(1 + sw_i) \log(\frac{Q}{n_i}) \tag{2}$$

Aquí, f_{ij} es el número de ocurrencias de la palabra i en la página j y n_i es el número total de documentos del sitio web que contienen la palabra i.

Adicionalmente, la importancia de las palabras es incrementada por la identificación de palabras especiales, las cuales correspondiente a los términos en la página web que son más importantes que otras, por ejemplo, palabras destacadas (haciendo uso de etiquetas HTML), palabras utilizadas por el usuario en la búsqueda de información y, en general, palabras que implican los deseos y necesidades de los usuarios. La importancia de palabras especiales es almacenada en un arreglo sw de dimensión R, donde sw_i representa un peso adicional para la i-ésima palabra.

El arreglo sw permite al modelo de vector espacial incluir ideas acerca de información semántica contenida en el texto de la página web por la identificación de palabras especiales.

Las fuentes comunes de palabras especiales son:

1. E-Mails: El ofrecimiento de envío de emails por parte del usuario para la plataforma de call center. Este texto enviado es una fuente para identificar las palabras más recurrentes. Sea $ew_i = \frac{w^i_{email}}{TE}$ el arreglo de palabras contenidas en los e-mails, que también están presentes en el sitio web, donde wi email es la frecuencia de la i-ésima palabra y TE es la cantidad total de palabras en el grupo completo del arreglo de palabras de e-mail.

2. Palabras destacadas. En un sitio web, hay palabras con etiquetas especiales, como diferentes fuentes, por ejemplo, itálica, negrita, o palabras pertenecientes al título. Sea $nw_i = \frac{w^i_{marks}}{TM}$ el arreglo de palabras destacadas dentro de las páginas web, donde w^i_{marks} es la frecuencia de la i-ésima palabra y TM es la cantidad de palabras destacadas en el sitio web completo.

3. Palabras de consultas: Un banco, por ejemplo, tiene motores de búsqueda a través de las cuales los usuarios pueden preguntar por asuntos específicos, por la introducción de palabras clave. Sea $aw_i = \frac{w^i_{ask}}{TA}$ el arreglo de palabras usadas por el usuario en el motor de búsqueda y que esta contenida en el sitio web, donde w^i_{ask} es la frecuencia de la i-ésima palabra y TA es la cantidad total de palabras en el grupo completo.

4. Sitios web relacionados. Usualmente un sitio web pertenece a un segmento de mercado, en este caso el mercado de las instituciones bancarias. Luego, es posible recolectar páginas de sitios web que pertenecen a otros sitios en el mismo mercado. Sea $rw_i = \frac{w^i_{rws}}{RWS}$ el arreglo con palabras utilizadas en el mercado de sitios web incluyendo el sitio web bajo estudio,

donde w^i_{rws} es la frecuencia de la i-ésima palabra y RWS es el número total de palabras en todos los sitios web considerados.

La expresión final $sw_i = ew_i + mw_i + aw_i + rw_i$ es la suma simple de los pesos descritos anteriormente.

En la representación vectorial, cada columna de la matriz M es una página web. Por ejemplo, la késima columna mik con i = 1,. . . ,R es la k-ésima página en el grupo completo de páginas.

Definición 1 (Vector de Palabras por página) *es un vector* $WP^k = (wp^k_1, \ldots, wp^k_R) = (m_{1k}, \ldots, m_{rk})$*, con* $k = 1, \ldots, Q$*, es la representación vectorial de la k-ésima página en el grupo de páginas bajo análisis.*

Con las páginas web en representación vectorial, es posible utilizar la medida de distancia para comparar los contenidos de texto. La distancia común es el coseno del ángulo calculado como:

$$dp(WP^i, WP^j) = \frac{\sum_{k=1}^{R} WP^i_k \cdot WP^j_k}{\sqrt{\sum_{k=1}^{R}(WP^i_k)^2}\sqrt{\sum_{k=1}^{R}(WP^j_k)^2}} \quad (3)$$

La ecuación (3) permite comparar el contenido de dos páginas web, retornando un valor numérico entre [0, 1]. Cuando las páginas son totalmente diferentes, $dp = 0$, y cuando son las mismas, $dp = 1$. Otro aspecto importante es que la ecuación 3 cumple con el requerimiento de ser computacionalmente eficiente, lo cual la hace más apropiada para ser utilizada en algoritmos de web mining.

4. Extrayendo las preferencias de contenido del usuario de las páginas web

Diferentes técnicas son aplicadas para analizar el comportamiento del usuario en el sitio web, desde una simple estadística de uso de una página hasta complejos algoritmos de web mining. En el último caso, la investigación se concentra en predicciones acerca de cuales páginas el usuario visitará y la información que esta buscando. Principalmente por la utilización de la combinación de las aproximaciones de WUM y WCM, el propósito es analizar las preferencias de texto del usuario web y por esta vía, identificar cuales palabras atraen la atención del usuario durante su navegación en el sitio. Previamente a la aplicación de una herramienta de web mining, la data relacionada con el comportamiento del usuario ha sido procesada para crear vectores de características, cuyos componentes dependerán de la implementación particular del algoritmo de web mining a utilizar y la preferencia de patrones ha ser extraídos.

4.1. Modelando el comportamiento del usuario web

La mayoría de los modelos de comportamiento de usuario web examinan la secuencia de páginas visitadas para crear vectores de características que representan el perfil de navegación del usuario web [12, 21, 36]. Estos modelos analizan el comportamiento de navegación del usuario en un sitio web mediante la aplicación de algoritmos que extraen los patrones de navegación. El siguiente paso es examinar las preferencias del usuario, definido como el contenido preferido de la página web por el usuario; y este es el contenido de texto que captura la atención especial, dado que es utilizada para encontrar información interesante relacionada a un tema particular por un motor de búsqueda. Por lo tanto, es necesario incluir una nueva variable como parte de la información del vector de comportamiento del usuario web acerca del contenido y tiempo gastado en cada página web visitada.

Definición 2 (Vector de comportamiento del usuario (UBV)) *Es un vector* $\nu = [(p_1, t_1), \ldots, (p_m, t_n)]$ *, donde son los parámetros que representan la i-ésima página del visitante y el tiempo gastado en ella en la sesión, respectivamente. En esta expresión,* p_i *es el identificador de la página.*

En la definición 2, el comportamiento del usuario en un sitio web es caracterizado por:

1. Secuencia de páginas; la secuencia de páginas visitadas y registradas en los archivos log. Si el usuario retorna a una página almacenada en el caché del browser, esta acción puede no ser registrada.

2. Contenido de la página; representa el contenido que puede ser texto libre, imágenes, sonidos, etc. Para propósitos de este paper, el texto libre es el utilizado principalmente para representar una página web.

3. Tiempo gastado, tiempo utilizado por el usuario en cada página. Para la página, el porcentaje de tiempo gastado en cada página durante la sesión del usuario puede ser directamente calculado.

4.2. Analizando las preferencias de texto de los usuarios

El objetivo es determinar las palabras más importantes para un sitio web dado para los usuarios, mediante la comparación de las preferencias de texto libre, a través del análisis de páginas visitadas y de tiempo gastado en cada una de ellas [34]. Sin embargo, difiere de las propuestas mencionadas anteriormente, dado que el ejercicio es encontrar las palabras clave que atraen y retienen a los usuarios en el uso de data disponible en la web. La expectación esta en involucrar usuarios pasados y actuales en un proceso continuo de determinación de palabras clave.

Las preferencias del contenido web del usuario son identificadas por la comparación de contenido de las páginas visitadas, [34, 33, 35] por la aplicación

del modelo de vector espacial a las páginas web, con la variante propuesta en la sección 3.2, ecuación (2). Los temas principales de interés pueden ser encontrados por el uso de la medición de la distancia entre vectores (por ejemplo, distancia euclidiana).

Desde el vector de comportamiento del usuario (UBV), las páginas más importantes son seleccionadas asumiendo que el grado de importancia esta correlacionado al porcentaje de tiempo gastado en cada página. El UBV se ordena de acuerdo al porcentaje de tiempo total gastado en cada página. Las ι página más importantes, es decir, las primeras ι páginas, son seleccionadas.

Definición 3 (Vector de Páginas Importantes (IPV)). *Es un vector* $\vartheta_\iota(\nu) = [(\rho_1, \tau_1), \ldots, (\rho_i, \tau_i)]$ *, donde* (ρ_ι, τ_ι) *es el componente que representa la* ι*-ésima página más importante y el porcentaje de tiempo gastado en ella por la sesión.*

Sean α y β dos UBV. La medida de similitud propuesta entre los dos IPV es introducida en la ecuación 4 como:

$$st(\vartheta_\iota(\alpha), \vartheta_\iota(\beta)) = \frac{1}{\iota} \sum_{k=1}^{\iota} \text{mín}\{\frac{\tau_k^\alpha}{\tau_k^\beta}, \frac{\tau_k^\beta}{\tau_k^\alpha}\} dp(\rho_k^\alpha, \rho_k^\beta) \tag{4}$$

El primer elemento en (4) indica el interés del usuario en las páginas visitadas. Si el porcentaje de tiempo gastado por los usuarios α y β en la k-ésima página visitada es cercano a la otra, el valor de la expresión min.,. será cercano a 1. En el caso opuesto, será cercano a 0. El segundo elemento en (4) es dp, la distancia entre páginas representada en forma vectorial, introducida en (3). En (4) el contenido de las páginas más importantes es multiplicado por el porcentaje de tiempo total gastado en cada página. Esto permite a las páginas con contenidos similares ser distinguidas por intereses diferentes de usuarios.

4.3. Identificando palabras clave del sitio web

Una palabra clave de un sitio web (o web site keyword) es definido como "una palabra o posiblemente un grupo de palabras que hacen de una página web más atractiva para un usuario eventual durante su visita al sitio web" [32]. Es interesante notar que las mismas palabras clave del sitio web pueden ser utilizadas por el usuario en un motor de búsqueda, cuando este está en busca de contenido web.

Para encontrar palabras clave de un sitio web, es necesario seleccionar las páginas web con el contenido textual que es significativo para los usuarios. La suposición es que existe una relación entre el tiempo gastado por el usuario en una página web y su interés en el contenido [31]. La relación es almacenada por el vector de páginas importantes (IPV), dando la información necesaria para extraer las palabras clave de un sitio web a través de la utilización de una herramienta de web mining.

Entre estas técnicas de web mining, se debe colocar especial atención a los algoritmos de clustering. La suposición es, dado un grupo de clusters extraídos de la información generada durante la formación de las sesiones de los usuarios en el sitio en, es posible el extraer las preferencias de los usuarios mediante el análisis de los contenidos del cluster. Los patrones en cada cluster detectado podrían ser suficientes para extrapolar el contenido que él o la usuario esta buscando [20, 23, 30].

En cada IPV, el componente página tiene una representación vectorial presentada por la ecuación (2). En esta ecuación, un paso importante es el cálculo de pesos considerados en el arreglo de palabras especiales swi. Las palabras especiales son diferentes a las palabras normales en el sitio, dado que pertenecen a una fuente alternativa y relacionada o ellas tienen una información adicional mostrando su importancia en el sitio, por ejemplo, una etiqueta HTML que enfatiza una palabra.

El algoritmo de clustering es utilizado para agrupar IPV similares por comparación de la cada componente de tiempo y página del vector, siendo importante el uso de la medida de similitud presentada en la ecuación (4). El resultado debería ser un grupo de clusters cuya calidad debe ser chequeada mediante el criterio de aceptación / rechazo. Un camino simple es aceptar los clusters cuyas páginas comparte un tema principal similar, y en otro caso, rechazar el cluster. En este punto, es necesario conocer que páginas en el sitio son cercanas con los vectores del cluster. Debido a que conoceremos la representación vectorial de las páginas web del sitio y utilizando la ecuación (3) podemos identificar la página más cercana de un cluster dado y de esta forma obtener las páginas adecuadas a un cluster para revisar si las páginas comparten un tema principal en común.

Para cada cluster aceptado y recordando que los centroides contienen páginas donde los usuarios gastan más tiempo durante su sesión respectiva y en la representación vectorial tienen los pesos más altos, el procedimiento de identificación de palabras clave del sitio web es aplicar una medida, descrita en la ecuación (5) (miembro geométrico) para calcular la importancia de cada palabra

$$kw[i] = \sqrt{\prod_{p \epsilon \varsigma} m_{ip}}, \tag{5}$$

donde $i = 1, \ldots, R$ y kw es un arreglo que contiene los pesos para cada palabra relativa a un cluster dado y ς el grupo de páginas representando el cluster. Las palabras clave del sitio web son el resultado del ordenamiento de kw y de la detección de palabras con los pesos más altos, por ejemplo, las 10 palabras con mayor peso.

5. Extrayendo patrones de los datos originados en un sitio web real

Para propósitos experimentales, el sitio web seleccionado debe ser complejo con respecto a varias características: número de visitas, actualización periódica (preferiblemente mensual con el fin de estudiar la reacción de los usuarios a los cambios) y ser rico en contenido textual. La página web de un banco virtual Chileno (sin sucursales físicas, todas las transacciones realizadas electrónicamente) cumple con estos criterios. Cabe destacar que para efectos de privacidad de los datos usados en la investigación, se firmó un acuerdo con el banco, por lo cual so nombre no pude ser mencionado.

Las principales características del sitio web del banco son las siguientes; presentado en Español, con 217 páginas web estáticas y aproximadamente ocho millones de filas en los registros de web log para un periodo de estudio entre Enero y Marzo del 2003.

El comportamiento del usuario en el sitio web del banco es analizado en dos formas. Primero, mediante la utilización de los archivos de registro que contienen información acerca del visitante y del comportamiento de navegación del cliente. Esta información requiere de una reconstrucción previa y limpieza antes de que las herramientas de web mining sean aplicadas. Segundo, la web data en el sitio web en si mismo, específicamente el contenido textual de las páginas web - esto también necesita de un preprocesamiento y limpieza.

5.1. Proceso de reconstrucción de sesiones

La Fig. 5.1 muestra parte de los registros del sitio web bancario e incluye tanto a clientes identificados como visitantes anónimos.

Figura 1: Extracto de un archivo de web log generado en el sitio web de un banco

#	IP	Id	A	Time	Method/URL/Protocol	Statu	Byte	Referer	Agent
1	164.77.129.50	-	-	12/Apr/2003:23:47:44	GET /img/tab.gif HTTP/1.1	200	89	http://www.thebank.cl	MSIE 6.0; Windows 98
2	200.28.206.200	-	20	12/Apr/2003:23:48:31	GET transa/info.htm HTTP/1.1	200	144	/infoeco/info.html	MSIE 4.01; Windows 95
3	200.86.248.170	-	-	12/Apr/2003:23:48:37	GET /img/gen.gif HTTP/1.1	304	0	/ofert/wines/	MSIE 6.0; Windows 98
4	66.249.65.97	-	-	12/Apr/2003:23:48:41	GET /index.htm HTTP/1.1	200	88	-	Googlebot/2.1; google.com/bot.html
5	216.241.8.179	-	31	12/Apr/2003:23:50:03	GET /tx/infoeco/card.htm HTTP/1.1	200	210	/tx/infoeco/prom/	MSIE 6.0; Windows NT 5.1
6	164.77.129.50	-	-	12/Apr/2003:23:48:34	GET /tx/infoeco/ HTTP/1.1	200	186	/tx/infoeco/card.htm	MSIE 6.0; Windows 98
7	200.28.206.200	-	20	12/Apr/2003:23:51:13	GET transa/account.htm HTTP/1.1	200	180	/transa/info.htm	MSIE 4.01; Windows 95
8	216.241.8.179	-	31	12/Apr/2003:23:51:23	GET /tx/infoeco/ind.htm HTTP/1.1	200	300	/tx/infoeco/card.htm	MSIE 6.0; Windows NT 5.1
9	200.86.248.170	-	-	12/Apr/2003:23:51:41	GET /prom/wine.html HTTP/1.1	404	0	/ofert/wines/	MSIE 6.0; Windows 98
10	164.77.129.50	-	44	12/Apr/2003:23:52:04	GET /tx/infoeco/ind.htm HTTP/1.1	200	186	/tx/infoeco/	MSIE 6.0; Windows 98

El acceso de los clientes al sitio es a través de una conexión segura, utilizando un protocolo SSL que permite el almacenamiento de un valor de identificación en el parámetro de autentificación de usuario en el archivo de registros

web. Otro modo de identificación de usuarios es mediante cookies, pero algunas veces estas son desactivadas por los usuarios en sus navegadores. En este caso sería necesario el reconstruir la sesión del visitante.

Durante el proceso de reconstrucción de sesiones, se aplican filtros a los registros del sitio web. En este caso particular, solo se utilizan registros de requerimiento de páginas web para analizar el comportamiento específico del usuario en el sitio. También es importante la limpieza de sesiones anormales, por ejemplo, web crawlers, como es mostrado en la Fig. 1, línea 4, donde un robot perteneciente a Google es detectado.

Las filas de los registros log del sitio web contienen cuatro meses de transacciones, con aproximadamente 8 millones de registros. Sólo se consideran los registros relacionados con páginas web para la reconstrucción de sesiones y análisis del comportamiento del usuario; la información que apunta a otros objetos como imágenes, sonidos, etc, son limpiadas.

5.2. Preprocesamiento del contenido de una página web

Mediante la aplicación de filtros a los textos de las páginas web, se ha encontrado que en el sitio completo contiene R=2034 palabras diferentes para ser utilizadas en el análisis.

Considerando los pesos de las palabras y la especificación de palabras especiales, fue utilizado el procedimiento presentado en la sección 3.2, con el fin de calcular sw_i, en la ecuación 2. Las fuentes de datos fueron:

1. Palabras destacadas. Dentro de las páginas web, se encontraron 743 palabras diferentes después de la aplicación del paso de preprocesamiento y limpieza.

2. Sitios web relacionados: Cuatro sitios web fueron considerados, cada un de ellos con aproximadamente 300 páginas.

El número total de palabras diferentes fue de 9253, con 1842 de ellas contenidas en el contenido del sitio web.

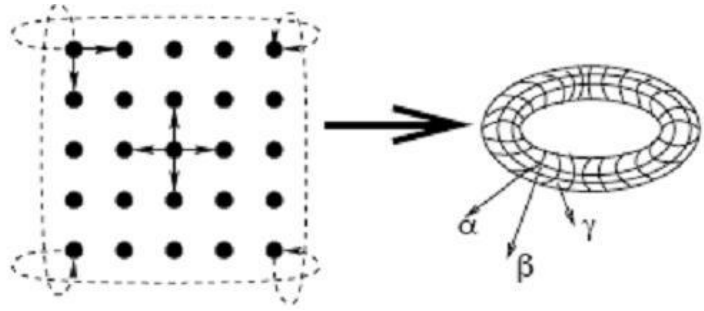
Después de la identificación de palabras especiales y sus respectivos pesos, es posible calcular el peso final para cada palabra en la totalidad del sitio web, por la aplicación de la ecuación (2). Luego, se obtiene la representación vectorial para todas las páginas del sitio

5.3. Analizando las preferencias de texto del usuario

Dos redes neuronales fueron aplicadas al web data para la identificación de clusters. La red neuronal artificial del tipo Kohonen (Self Organizing Feature Map; SOFM) y K-means.

Esquemáticamente, una red SOFM es una red neuronal artificial no supervisada, correspondiente a un arreglo de neuronas de dos dimensiones. Cada neurona esta constituida por un arreglo bidimensional de vectores de n dimensiones cuyos componentes son los pesos sinápticos. Por construcción, todas las neuronas reciben el mismo input en un momento determinado. La noción de vecindad entre neuronas define diversas topologías. Para el caso de este trabajo, se utilizó la topología toroidal [38] que significa que las neuronas localizadas de un borde, son cercanas al borde opuesto. La ventaja de la topología radica en que mantiene la continuidad de los clusters o cuando la data corresponde a secuencias de eventos.

Figura 2: de Vectores de Páginas Importantes en un SOFM con topología toroidal.

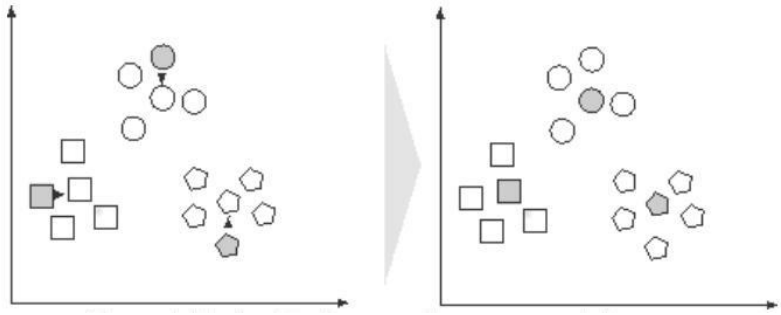


K-means es una red de aprendizaje supervisada y predefiniendo el número de centroides, genera las agrupaciones de vectores llamados miembros en torno a ellos. K-means para detectar las pertenencias a sus centroides tradicionalmente utiliza la distancia euclideana para discernir que centroide es más representativo para un vector. Puesto que la investigación se centra en un vector de comportamiento del usuario con una estructura diferente a la euclideana, se hace modificación de esta red de aprendizaje y se utiliza la medida de similitud presentada en la ecuación (4) para establecer las pertenencias a los centroides correspondientes. Para el caso de esta investigación, el principal input de este algoritmo - los K centroides - será originado por el resultado que entregue la red SOFM que será inicialmente utilizada para el análisis del comportamiento del usuario y que parte de los resultados que retorne serán los clusters detectados. La Fig. 3 muestra el comportamiento de los centroides a medida que se van encontrando mejores representantes.

5.4. Analizando las preferencias del usuario con una red SOFM

Se ha fijado en 3 el número máximo de dimensiones del vector. Luego, un SOFM con 3 neuronas de entrada y 32*32 neuronas de salida fue utilizado

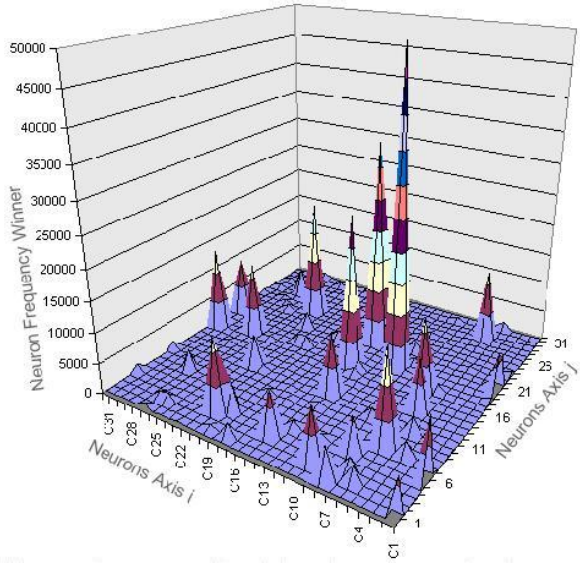
Figura 3: Evolución de centroides en una red K-means



para encontrar los clusters de vectores de páginas importantes.

La Fig. 4 muestra las posiciones de las neuronas en el SOFM en los ejes x e y. El eje z es la frecuencia normalizada de veces que una neurona gana durante el entrenamiento.

Figura 4: Clusters de vectores de páginas importantes desde una red SOFM.



La Fig. 4, muestra 8 cluster principales que contienen información acerca de las páginas más importantes del sitio web. Sin embargo, sólo 5 fueron aceptadas. El criterio de aceptación / rechazo es simple; si las páginas de un centroide de cluster tienen el mismo tema principal, entones el cluster es aceptado, de otra forma se rechaza.

Los centroides de los clusters son mostrados en el Cuadro 1. La segunda columna contiene las neuronas centrales (neuronas ganadoras) para cada cluster y representa las páginas visitadas más importantes.

Cuadro 1: Vectores de páginas importantes obtenidos con SOFM.

Cluster	Páginas Visitadas
1	(171, 130, 159)
2	(76, 58, 130)
3	(175, 78, 10)
4	(78, 32, 130)
5	(130, 171, 159)

5.5. Analizando las preferencias del usuario con una red K-means

Desde el resultado obtenido con la aplicación de SOFM, se inicializa el entrenamiento de la red K-means. La cantidad de centroides es de acuerdo al número final aceptado. Luego, el proceso se inicializó con 5 centroides. La detención de asignación de miembros al clusters se produce cuando los mejores representantes de cada centroide van variando de menor manera llegando en un punto a quedar prácticamente establecido el centroide que será el representante de los miembros.

Para cada centroide se obtiene las páginas visitadas representativas de los grupos. En el Cuadro 2 se muestran las páginas visitadas de los representantes de los clusters y la cantidad de miembros identificados en ellos.

Cuadro 2: Vectores de páginas importantes obtenidos con K-means.

Cluster	Páginas Visitadas
1	(117,192,19)
2	(21,10,179)
3	(205,128,210)
4	(55,18,41)
5	(24,104,95)

5.6. Identificación de web site keywords

Se requiere un paso final para obtener las palabras clave de un sitio web: analizar cuales son las palabras que tienen una mayor importancia relativa con respecto al sitio web completo.

Las palabras clave y su importancia relativa en cada cluster son obtenidas por la aplicación de la ecuación (5). Por ejemplo, si el cluster es ($\varsigma = \{171, 130, 159\}$), entonces $kw[i] = \sqrt[3]{m_{i171}, m_{i130}m_{i159}}$, con $i = 1, \ldots, R$.

Finalmente, ordenando las kw de forma descendente, podemos seleccionar las k palabras más importantes para cada cluster, por ejemplo $k = 5$.

No se nos permite mostrar las palabras clave específicas debido a la cláusula de confidencialidad con el banco, por esta razón las palabras son numeradas. El Cuadro 3 muestra las palabras encontradas con el método propuesto.

El Cuadro 4 muestra un grupo seleccionado de palabras clave de todos los clusters. Las palabras clave en si, sin embargo, no tienen mucho sentido. Estas necesitan un contexto de página web donde ellas podrían ser utilizadas como palabras especiales, por ejemplo, palabras destacadas para enfatizar un concepto o como palabras vinculadas a otras páginas.

Cuadro 3: Las 5 palabras más importantes por cluster

C	Palabras Clave	Peso ordenado
1	$(w_{2032}, w_{1233}, w_{287}, w_{1087}, w_{594})$	(2.35,1.93,1.56,1.32,1.03)
2	$(w_{1003}, w_{449}, w_{895}, w_{867}, w_{1567})$	(2.54,2.14,1.98,1.58,1.38)
3	$(w_{1005}, w_{948}, w_{505}, w_{1675}, w_{1545})$	(2.72,2.12,1.85,1.52,1.31)
4	$(w_{501}, w_{733}, w_{385}, w_{684}, w_{885})$	(2.84,2.32,2.14,1.85,1.58)
5	$(w_{200}, w_{1321}, w_{206}, w_{205}, w_{1757})$	(2.33,2.22,1.12,1.01,0.93)

Cuadro 4: Parte de las palabras descubiertas.

#	Palabras Clave
1	Cuenta
2	Fondo
3	Inversión
4	Tarjeta
5	Hipotecario
6	Seguro
7	Cheques
8	Crédito

La recomendación específica es utilizar las palabras clave como "palabras para escribir" en un sitio web, es decir, los párrafos escritos en la página deberían incluir algunas palabras clave y algunas podrían ser un enlace a otras páginas.

Además es posible sobre la base de este ejercicio el realizar recomendaciones de contenidos de texto. Sin embargo, para reiterar, las palabras clave no funcionan de forma separada sino que requieren de un contexto que las utilice. Revisando el Cuadro 2, para cada cluster, la palabra clave descubierta podría ser utilizada para reescribir un párrafo o una página web completa. Adicionalmente, es importante insertar palabras clave para destacar conceptos específicos.

Las palabras clave también son utilizadas como palabras índice para un motor de búsqueda, es decir, algunas podrían ser utilizadas para personalizar

el crawler que visita sitios web y carga páginas. Luego, cuando un usuario esta buscando por una página en específico en un motor de búsqueda, la probabilidad de obtener el sitio web se incrementa.

5.7. Mejorando el contenido textual el sitio web

Las palabras clave son conceptos para motivar los intereses de los usuarios y hacerlos visitar el sitio web. Están para ser jugadas dentro de su contexto como palabras aisladas que pueden tener un pequeño sentido , dado que los clusters representar contextos diferentes. La recomendación específica es utilizar palabras clave como "palabras para escribir" en el sitio web.

En cuanto cada página contiene un contenido de texto específico, es posible asociar las palabras clave de un sitio web a un contenido de la página; y desde esta sugerir la revisión o reconstrucción de un nuevo contenido en el sitio web. Por ejemplo, si la nueva versión de la página es relacionada con "tarjetas de crédito", entonces las palabras clave del sitio web "crédito, puntos y promociones" deben ser asignadas para la reescritura del contenido textual de la página.

5.8. Testeo de la efectividad de las recomendaciones de texto

La detección y aplicación de web site keywords no garantizan el éxito de aplicación en un contenido textual. Incluso, el riesgo de utilizarlas puede generar disgusto en un usuario habitual y por lo tanto abandonar o dejar de utilizar el sitio web. Como medida precautoria, se realizaron test de efectividad de las web site keywords. Sobre el contenido del sitio web se extrajeron 10 párrafos que contenían para el caso de 5 de ellos web site keywords y otros no las contenían. El resultado se realizó sobre un universo de 10 personas con el fin de conocer la recepción que ellos tenían respecto a párrafos que contenían las palabras detectadas, según el contexto de si entregaban información relevante en un sitio bancario. El resultado de este test es el que se muestra en el Cuadro 6.

Como se puede apreciar, aquellas palabras que contenían web site keywords eran para el usuario mucho más interesantes e importantes en el contexto de navegación en que estaban inmersos, versus aquellos párrafos en que no había presencia de dichos web site keywords. Las web site keywords atraen la atención del usuario y pueden ser una muy buena guía en el diseño de contenidos específicos de un sitio web. Esta combinación de elementos que se alinean a los que el usuario busca puede otorgar un mejor resultado en la satisfacción de los clientes.

Cuadro 5: Párrafos testeados para análisis de keywords.

#	Incluye web site keyword	Párrafo
1	Si	Orientado a empresas que deseen manejar excedentes de caja, así como a Personas que quieran mantener parte de sus recursos **invertidos** en un **fondo mutuo**, cuya cartera esté compuesta exclusivamente por instrumentos de **deuda** nacional, obteniendo rentabilidad y liquidez a corto plazo.
2	Si	Solicitándolos con un día hábil bancario de antelación, se pagarán mediante **cheques** nominativos, **vales** vista o **depósitos** en **cuentas** corrientes, de acuerdo a sus instrucciones.
3	Si	Este plan busca otorgar a tus **Ahorros Previsionales** Voluntarios acumulados a esta fecha y los futuros, una atractiva y segura **rentabilidad** que te permitirá poder mejorar considerablemente tus **ahorros** para una mejor **pensión.**
4	Si	Para obtener información de tu **Cuenta Corriente** y de tu **Línea** de **Sobregiro** debes seguir los siguientes pasos
5	Si	El Servicio de Mensajería es un servicio de entregas y **retiros** de dinero, especies valoradas y documentos que podrás utilizar siendo **cliente**
6	No	Para solicitar tu Plan debes completar la siguiente información y se contactarán contigo.

6. Conclusiones

Cuando un usuario visita un sitio web, hay una correlación entre el máximo de tiempo gastado por sesión en una página y su contenido de texto libre. Esto permite modelar las preferencias del usuario a través del "Vector de Páginas Importantes (IPV)", el cual es la estructura de datos básica de almacenamiento de páginas donde el usuario gasta más tiempo durante e su sesión. Mediante la utilización de IPV como entrada en un SOFM y K-means, se pueden identificar clusters que contienen la navegación del usuario e información de sus preferencias de contenido.

El criterio de aceptación / rechazo de un cluster es simple: si las páginas dentro de cada cluster están relacionadas con un tema principal similar, entonces el cluster es aceptado, en caso contrario, se rechaza. Aplicando este

Cuadro 6: párrafos testeados para análisis de keywords.

		Opinión de aceptabilidad				
#	Incluye web site Keyword	Irrelevante	Moder. irrelevante	Algo relevante	Moder. relevante	relevante
1	Si				8	2
2	Si			4	4	2
3	Si			4	2	4
4	Si				7	3
5	Si			1	2	7
6	No	1	3	5	1	
7	No	3	2	5		
8	No	6	4			
9	No	5	2	1	2	
10	No	7	2	1		

criterio, 5 clusters son aceptados y el patrón contenido en cada una de ellas fue utilizado para extraer las palabras clave del sitio web.

El texto contenido en las páginas web puede ser mejorado utilizando las palabras clave del sitio web, y por esta vía atraer la atención del usuario cuando están visitando un sitio web. Sin embargo, es necesario recordar que estas palabras no pueden ser utilizadas de forma individual, de hecho necesitan de un contexto, el cual es provisto por un ser humano.

Como validación de las palabras detectadas, se realizó un testeo de párrafos que contenían dichos web site keywords versus otros que no contenían. El resultado fue satisfactorio corroborando la importancia de las palabras pues aquellos contenidos con web site keywords parecían más relevantes que otras que no contenían dichas palabras, por lo tanto el interés del usuario en contenidos con los keywords se hace mayor y de ahí la importancia de dar uso a estas palabras en los párrafos del contenido.

Como trabajo futuro, se aplicará la metodología en otros web data, por ejemplo las imágenes y objetos no textuales, con el fin de identificar cuales elementos atraen la atención del usuario en el sitio web.

Agradecimientos: Este trabajo fue parcialmente financiado por el Instituto Milenio Sistemas Complejos de Ingeniería

Referencias

[1] Green, Paul E. and V. Srinivasan (1990), “Conjoint Analysis in Marketing Research: New Developments and Directions”, Journal of Marketing 54, 4, 3-19.

[2] E. Amitay and C. Paris. “Automatically summarizing web sites: Is there

any way around it?" In Procs. of the 9th Int. Conf. on Information and Knowledge Management, pages 173-179, McLean, Virginia, USA, 2000.

[3] R. Baeza-Yates. "Web usage mining in search engines", chapter Web Mining: Applications and Techniques, pages 307-321. Idea Group, 2004.

[4] B. Berendt, A. Hotho, and G. Stumme. "Towards semantic web mining". In Proc. in First Int. Semantic Web Conference, pages 264-278, 2002.

[5] B. Berendt and M. Spiliopoulou. "Analysis of navigation behavior in web sites integrating multiple information systems". The VLDB Journal, 9:56-75, 2001.

[6] D. Buttler. "A short survey of document structure similarity algorithms". In Procs. Int. Conf. on Internet Computing, pages 3-9, 2004.

[7] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. "Focused web searching with pdas". Computer Networks, 33(1- 6):213-230, June 2000.

[8] L. D. Catledge and J. E. Pitkow. "Characterizing browsing behaviors on the world wide web". Computers Networks and ISDN System, 27:1065-1073, 1995.

[9] G. Chang, M. Healey, J. McHugh, and J. Wang. "Mining the World Wide Web". Kluwer Academic Publishers, 2003.

[10] W. Chuang and J. Yang. "Extracting sentence segment for text summarization? a machine learning approach". In Procs. Int. Conf. ACM SIGIR, pages 152-159, Athens, Greece, 2000.

[11] R. Cooley, B. Mobasher, and J. Srivastava. "Data preparation for mining world wide web browsing patterns". Journal of Knowlegde and Information Systems, 1:5-32, 1999.

[12] U. Hahn and I. Mani. "The challenges of automatic summarization". IEEE Computer, 33(11):29-36, 2000.

[13] A. Joshi and R. Krishnapuram. "On mining web access logs". In Proc. of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 63- 69, 2000.

[14] A. P. Jr and N. Ziviani. "Retrieving similar documents from the web". Journal of Web Engineering, 2(4):247-261, 2004.

[15] D. Lawrie, B. W. Croft, and A. Rosenberg. "Finding topic words for hierarchical summarization". In Proc. 24th Int SIGIR Conf. on Research and Development in Information Retrieval, pages 349-357, New Orleans, Louisiana, USA, 2001. ACM Press.

[16] E. Liddy, K. McVearry, W. Paik, E. Yu, and M. McKenna. "Development, implementation and testing of a discourse model for newspaper texts". In Procs. Int. Conf. on ARPA Workshop on Human Language Technology, pages 159-164, Princeton, NJ, USA, 1993.

[17] G. Linoff and M. Berry. "Mining the Web". Jon Wiley & Sons, New York, 2001.

[18] S. Loh, L. Wives, and J. P. M. de Oliveira. "Concept based knowledge discovery in texts extracted from the web". SIGKDD Explorations, 2(1):29-39, 2000.

[19] I. Mani and M. Maybury. "Advances in automatic text summarization". MIT Press, Cambridge, Mass., 1999.

[20] S. Mitra, S. K. Pal, and P. Mitra. "Data mining in soft computing framework: A survey". IEEE Transactions on Neural Networks, 13(1):3-14, 2002.

[21] B.Mobasher, R. Cooley, and J. Srivastava. "Creating adaptive web sites through usage-based clustering of urls". In Procs.Int Conf IEEE Knowledge and Data Engineering Exchange, November 1999.

[22] B. Mobasher, R. Cooley, and J. Srivastava. "Automatic personalization based on web usage mining". Communications of the ACM, 43(8):142-151, 2000.

[23] M. F. Porter. "An algorithm for suffix stripping". Program; automated library and information systems, 14(3):130-137, 1980.

[24] T. A. Runkler and J. Bezdek. "Web mining with relational clustering". International Journal of Approximate Reasoning, 32(2-3):217-236, Feb 2003.

[25] G. Salton, A. Wong, and C. S. Yang. "A vector space model for automatic indexing". Communications of the ACM archive, 18(11):613-620, November 1975.

[26] M. Spiliopoulou. "Data mining for the web". In Principles of Data Mining and Knowledge Discovery, pages 588-589, 1999.

[27] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. "A framework for the evaluation of session reconstruction heuristics in web-usage analysis". INFORMS Journal on Computing, 15:171-190, 2003.

[28] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. "Web usage mining: Discovery and applications of usage patterns from web data". SIGKDD Explorations, 1(2):12-23, 2000.

[29] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. "Recovering traceability links in multilingual web sites". In Procs. Int Conf. Web Site Evolution, pages 14-21. IEEE Press, 2001.

[30] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. "Restructuring multilingual web sites". In Procs. Int. Conf. Software Maintenance, pages 290-299. IEEE Press, 2002.

[31] J. D. Velázquez and V. Palade. "A knowledge base for the maintenance of knowledge extracted from web data". Journal of Knowledge-Based Systems, 20(3):238-248, 2007.

[32] J. D. Velásquez, S. Ríos, A. Bassi, H. Yasuda, and T. Aoki. "Towards the identification of keywords in the web site text content: A methodological approach". International Journal of Web Information Systems, 1(1):11-15, March 2005.

[33] J. D. Velásquez, R. Weber, H. Yasuda, and T. Aoki. "A methodology to find web site keywords". In Procs. IEEE Int. Conf. on e-Technology, e-Commerce and e-Service, pages 285-292, Taipei, Taiwan, March 2004.

[34] J. D. Velásquez, H. Yasuda, and T. Aoki. "Combining the web content and usage mining to understand the visitor behavior in a web site". In Procs. 3th IEEE Int. Conf. on Data Mining, pages 669-672, Melbourne, Florida, USA, November 2003.

[35] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. "Using the kdd process to support the web site reconFig.tion". In Procs. IEEE/WIC Int. Conf. on Web Intelligence, pages 511-515, Halifax, Canada, October 2003.

[36] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. "A new similarity measure to understand visitor behavior in a web site". IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization, E87-D(2):389-396, February 2004.

[37] J. Xiao, Y. Zhang, X. Jia, and T. Li. "Measuring similarity of interests for clustering web-users". In ADC '01: Proceedings of the 12th Australasian conference on Database technologies, pages 107-114, Washington, DC, USA, 2001. IEEE Computer Society.

[38] K. Zechner. "Fast generation of abstracts from general domain text corpora by extracting relevant sentences". In Procs. Int. Conf. on Computational Linguistics, pages 986-989, 1996.

[39] J. D. Velásquez, H. Yasuda, T. Aoki, R. Weber and E. Vera (2003) "Using self organizing feature maps to acquire knowledge about visitor behavior in a web site". Lecture Notes in Artificial Intelligence, 2773(1): 951-958

Programación matemática para seleccionar los aspirantes a un Magíster con criterios de equidad regional, socio-económica y de género

Guillermo Durán*
Rodrigo Wolf Yadlin*

Resumen

El Magíster en Gestión para la Globalización es un nuevo programa de posgrado surgido a través de una alianza entre el Departamento de Ingeniería Industrial de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile y la empresa Minera Escondida. Este Magíster tiene como objetivo contribuir a los desafíos de capital humano y social que enfrenta Chile en esta etapa de su desarrollo, con la formación de excelencia de jóvenes profesionales. En el presente trabajo, mostramos cómo se utilizaron modelos de programación matemática para seleccionar los aspirantes a ingresar a dicho Magíster, utilizando criterios de equidad regional, socio-económicos y de género. La utilización de dichos modelos permite encontrar soluciones robustas en pocos minutos, lo que sería prácticamente imposible de hacer mediante métodos manuales.

Palabras Clave: Equidad, Programación Entera, Selección Robusta.

*Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile

1. Introducción

El Magíster en Gestión para la Globalización es un programa creado en 2007 y tiene como fin la formación de excelencia de jóvenes profesionales para el país. Este programa de posgrado se realiza a través de una alianza entre el departamento de Ingeniería Industrial de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile y la empresa Minera Escondida.

El programa tiene por objetivo contribuir a los desafíos de capital humano y social que enfrenta Chile en esta etapa de su desarrollo, con la formación de jóvenes profesionales provenientes del amplio espectro socioeconómico chileno, en condiciones de desempeñarse eficazmente en la empresa globalizada. Una de las características más importantes del Magíster es que todos los participantes pueden cursar sus estudios en el programa becados por la empresa Minera Escondida. Sin embargo, para poder postular al Magíster hay que cumplir con una serie de requisitos relacionados a la edad del postulante, duración de sus estudios superiores, años de experiencia laboral y su historia académica.

La dirección del Magíster decidió que el número de postulantes a ser aceptados en la primer promoción sería de 53. Asimismo, con el objetivo de aplicar criterios de equidad regional, socio-económica y de género, definió que al menos un 30 % deberían ser mujeres, un 60 % provenir de regiones y un 80 % pertenecer a los cuatro quintiles económicos distintos al superior. Los postulantes que cumplen los requisitos mínimos pasan a una primera etapa del proceso de selección (en este caso fueron más de 600). Además, se le asigna a cada postulante un puntaje, el cual se obtiene en base a sus antecedentes académicos y laborales. De este primer grupo, de acuerdo a su puntaje, 500 personas pasan a la segunda etapa. En dicha etapa deben rendir una serie de pruebas de distintas áreas de conocimiento y someterse a una evaluación psicométrica. El resultado de las pruebas antes mencionadas, más el puntaje que traen de la etapa anterior, se usa para generar un nuevo puntaje. A continuación, 150 personas pasan a la etapa número tres. En ella los postulantes son sometidos a una evaluación psicológica y los que pasan esta prueba, que es eliminatoria, forman el grupo final. De esta forma todo queda listo para elegir a los 53 alumnos del Magíster en Gestión para la Globalización y los 20 postulantes que quedaran en lista de espera en caso de que alguno de los seleccionados desista de participar del programa. Cabe acotar que 87 postulantes de los 150 (un 58 %) pudieron aprobar la evaluación psicológica e integrarse al grupo final de candidatos.

La forma de calcular los puntajes y las condiciones mínimas que deben satisfacer los postulantes para ingresar en la primer etapa del proceso son es-

tablecidas por los organizadores del Magíster y no serán discutidas en este artículo. Lo que nos interesa discutir es cómo, dadas esas condiciones, seleccionamos a los alumnos, teniendo en cuenta que este programa desea tener a los estudiantes con el perfil más adecuado, pero, como ya expresamos, también desea que las ventajas que unas personas puedan tener sobre otras debido al medio en el que nacieron o su género, no sean determinantes en la elección. Los requisitos para definir si un postulante es de los quintiles inferiores o de regiones también son definidos por los organizadores. Cabe destacar que a lo largo de todo el proceso se trabajó con dos definiciones distintas de quintil, decidiendo recién a último momento la Dirección del Magíster con cuál de ellas se quedaban.

El objetivo de este trabajo es mostrar como se utilizaron modelos de optimización lineal entera para seleccionar a los aspirantes que más se adecúan al perfil fijado por el Magíster, bajo la restricción de cumplir con los cupos propuestos. Se pretende conseguir una solución final robusta, en el sentido de que la misma no cambie demasiado ante pequeñas variaciones de los criterios. Cabe destacar que realizar todo este proceso en forma manual en muy corto tiempo hubiera sido prácticamente imposible, y de ahí la importancia del uso de modelos de programación matemática. Algunas de las ideas aquí utilizadas son tomadas de [1]. En la sección 2 se describen en detalle los modelos matemáticos utilizados. En la sección 3 se muestra el funcionamiento del algoritmo que los combina. En la sección 4 se presentan los resultados, mientras que en la última sección se dan algunas conclusiones del trabajo. Queda como anexo una aplicación particular del algoritmo que combina los resultados de los modelos.

2. Los modelos matemáticos

Los modelos desarrollados son tres y cada uno brinda un criterio distinto de selección. A continuación explicamos las restricciones comunes y luego los elementos particulares de cada modelo.

Restricciones Comunes

Supondremos que N es el número de personas que se desea seleccionar, K es un conjunto que incluye a todos los postulantes, M es un conjunto que agrupa a todas las mujeres que postulan, R es un conjunto que posee a todos los candidatos que son de regiones y Q es el conjunto cuyos miembros son los postulantes de los primeros cuatro quintiles que postulan al Magíster. En tanto, p_i es el puntaje del postulante i (sin pérdida de generalidad podemos suponer que los puntajes están ordenados de mayor a menor). Luego todos los

modelos poseen la siguiente variable de decisión y restricciones.

Variables de Decisión

$$X_i = \begin{cases} 1 & \text{si el participante i es seleccionado} \\ 0 & \text{en caso contrario} \end{cases}$$

Restricciones

1. El número de candidatos a ser seleccionados está prefijado por los organizadores del Magíster.

$$\sum_{i \epsilon K} X_i = N \tag{1}$$

2. Al menos un 30 % de los seleccionados deben ser mujeres.

$$\sum_{i \epsilon M} X_i \geq \frac{30}{100} \cdot N \tag{2}$$

3. Al menos un 60 % de los seleccionados debe ser de regiones.

$$\sum_{i \epsilon R} X_i \geq \frac{60}{100} \cdot N \tag{3}$$

4. Al menos un 80 % de los seleccionados debe pertenecer a los cuatro primeros quintiles.

$$\sum_{i \epsilon Q} X_i \geq \frac{80}{100} \cdot N \tag{4}$$

A continuación se describen las funciones objetivo de cada modelo, además de las restricciones adicionales propias de cada uno si es que las posee.

2.1. Modelo 1

El objetivo es maximizar la sumatoria de los puntajes de quienes son seleccionados.

Función Objetivo

$$\text{máx} \sum_{i \epsilon K} X_i \cdot p_i \tag{5}$$

2.2. Modelo 2

El objetivo es minimizar la sumatoria de los rankings de quienes son seleccionados.

Función Objetivo

$$\text{mín} \sum_{i \epsilon K} i \cdot X_i \tag{6}$$

2.3. Modelo 3

El objetivo es minimizar el ranking del último que es seleccionado. A diferencia de los otros modelos aparece una nueva variable de decisión cuyo valor será mayor o igual al ranking de todos los seleccionados (al minimizarla representará el ranking del último candidato que ingresará al Magíster). Para esto es necesario imponer una restricción adicional que exija que esta nueva variable sea mayor o igual que la ubicación en la lista (la cual está ordenada) de todos los seleccionados. Luego, en la función objetivo se minimizará esta variable sumada a la función objetivo del modelo 2 multiplicada por un número muy pequeño, para en caso de empates elegir a los mejores ubicados (está claro que este sumando no afecta el resultado en caso de que no haya empates).

y: posición relativa mayor o igual a todos los candidatos seleccionados.

Restricción

$$i \cdot X_i \leq y \qquad \forall i \tag{7}$$

Función Objetivo

$$\text{mín}(y + 0,0002 \cdot \sum_{i \epsilon K} i \cdot X_i) \tag{8}$$

3. El Algoritmo de Selección

Dado que el objetivo es conseguir una solución final robusta, se tomarán las mejores soluciones otorgadas por cada una de los modelos y se propondrá una forma de combinarlas de modo de obtener una única solución final. Desarrollamos para esto un algoritmo que requiere como punto de partida al menos las

tres mejores soluciones de cada uno de los modelos (consideramos que la idea de considerar tres soluciones de cada modelo ayuda a la hora de buscar la robustez de la solución final, pero este es un parámetro del modelo que puede ser modificado para realizar diferentes tests). Estas corridas brindan el óptimo del problema (corrida 1), la segunda mejor solución (corrida 2) y la tercera mejor solución (corrida 3) de éste. La forma de obtener la segunda mejor solución es agregar una restricción al modelo lineal entero que impida como solución factible a la mejor solución. Análogamente, impidiendo también la segunda mejor solución, se obtiene la tercera mejor solución. Si existe una única mejor solución, una única segunda y una única tercera, se le asigna un coeficiente de 1 a los postulantes que están en la solución de la corrida 1, un coeficiente de 0,6 a los de la corrida 2, y un coeficiente de 0,3 a los de la corrida 3. Luego, se suma para cada postulante sus coeficientes en cada una de las corridas de cada uno de los modelos. Por ejemplo, si un postulante aparece en la corrida 1 de los modelos 1 y 2, en la corrida 2 de los modelos 1 y 2 y en la corrida 3 del modelo 1, esa persona recibe un coeficiente general ponderador de 3,5. Luego este ponderador se multiplica por el puntaje que cada persona posee y así se genera un nuevo puntaje para cada candidato. En caso de empates en alguna de las mejores soluciones de alguno de los modelos, se generaliza este concepto. Por ejemplo, si hubiera 2 mejores soluciones y luego una única tercer mejor solución, las 2 primeras se repartirían el coeficiente 1,6 (1 + 0,6) en dos mitades iguales de 0,8, mientras que la tercera conservaría el coeficiente de 0,3.

El Algoritmo de Selección funciona entonces así:

1. Primera Selección: se chequea cuáles candidatos coinciden en la solución óptima de cada uno de los tres modelos (corridas 1). Las personas que aparezcan en todas ellas son directamente seleccionados para el Magíster. De esta forma si los tres modelos arrojan los mismos resultados ya se tiene a los elegidos y sólo falta confeccionar la lista de espera para lo que se avanza directamente al paso 5, en caso contrario se va al paso 2.

2. Nuevo Puntaje: se calcula el coeficiente ponderador a cada uno de los no seleccionados. Luego este ponderador se multiplica por el puntaje que cada persona posee y así se genera un nuevo puntaje para cada candidato.

3. Segunda Selección: se evalúa la composición de mujeres, postulantes de regiones y de los primeros cuatro quintiles que hay dentro de los ya elegidos con el fin de saber cuántos seleccionados de cada uno de estos grupos faltan para cumplir con el mínimo requerido. Luego, para encontrar los restantes elegidos, se ejecuta el modelo 2 con los puntajes obtenidos en el paso 2, exigiendo en las restricciones que se seleccionen al menos tantas mujeres, postulantes de regiones y no del quinto quintil que permitan

cumplir con la cuota mínima pedida y solicitando que el numero de elegidos por el modelo sea igual al número de personas que falta para tener a la cantidad de estudiantes que cursará el Magíster. Luego se chequea si el óptimo del problema es único o no, si no lo es, vamos al paso 4, si lo es, las personas que conforman la solución pasan a ser parte de la lista de elegidos, quedando así esta completa y se va al paso 5.

4. Tercera Selección: se calcula la sumatoria de los puntajes de cada una de las soluciones encontradas en el paso 3 (o sea, se aplica el modelo 1). El grupo que obtenga el mayor puntaje es el que completa la lista de elegidos. En caso de aún haber igualdad entre dos o más soluciones, se les presentan todas estas alternativas a los organizadores del Magíster para que ellos decidan.

5. Lista de Espera: si el número de postulantes que aparecen en alguna de las nueve corridas, pero que no están dentro de los 53 seleccionados, es superior a 20, se elige a los 20 de mejor puntaje. Por su parte, si la cantidad de postulantes en esa situación es menor que 20, todos ellos pasan a formar parte de la lista de espera, la cual se completa hasta llegar a 20 con los mejores puntajes de todos los postulantes que no están en ninguna de las mejores soluciones de alguno de los modelos.

Para la confección de la lista de espera no se consideran los distintos cupos. Sólo que en caso de ser necesario recurrir a algún postulante de ella para reemplazar a alguno de los elegidos, se tenderá a elegir el de mejor puntaje de modo que el grupo de seleccionados siga cumpliendo con la cuota mínima de mujeres, postulantes de regiones y candidatos pertenecientes a los quintiles inferiores.

La aplicación del Algoritmo de Selección nos da cierta garantía de la robustez de la solución final, dado que los postulantes que terminen siendo seleccionados serán los que figuren en varias de las mejores soluciones de cada modelo. Aquí puede verse la importancia de la aplicación de modelos de programación matemática: sería prácticamente imposible obtener todos estos resultados en pocos minutos de manera manual. Para ilustrar el funcionamiento del algoritmo, en el Anexo 1 se muestra la aplicación del mismo cuando se trabajó con la primera definición de quintil dada por los organizadores.

4. Los Resultados

Los modelos en las primeras dos etapas se usaron para garantizar que hubiesen suficientes mujeres, postulantes de regiones y de los primeros cuatro quintiles en la etapa final, siendo un apoyo para los organizadores a la hora

de decidir quienes avanzaban. En estas etapas nunca se usó el Algoritmo de Selección, es por eso que los resultados que mostraremos y analizaremos en esta sección se centran en la etapa final.

En la tabla siguiente se muestran los resultados de la función objetivo para las 3 mejores soluciones con cada una de las 2 definiciones distintas de quintil que usaron los organizadores.

Cuadro 1: Resultados con ambas definiciones de quintil

Modelo	Mejor Solución	Valor de la F.O. (Def. 1 de Quintil)	Valor de la F.O. (Def. 2 de Quintil)
1	1	3334,0798	3392,6797
1	2	3334,0717	3392,4
1	3	3333,6946	3392,9229
2	1	1792	1470
2	2	1795	1473
2	3	1795	1476
3	1	71,3702	64,294
3	2	71,3714	64,2946
3	3	71,3716	64,2952

Como se puede apreciar los resultados con la definición 2 de quintil tienen mejor valor de la función objetivo. Esto ocurre porque en la segunda definición se agranda el conjunto de personas que pertenece a quintiles distintos del superior. Es interesante notar que el resultado que se obtendría si la única restricción fuera la del número de postulantes seleccionados, para el modelo 2, es 1431, en tanto que para el modelo 3 dicho valor es 53,2862 (el último que entra es el postulante de ranking 53, los decimales se usan para desempatar ante dos soluciones donde el último que entra es el mismo). Así, los resultados, usando la primera definición de quintil, de los modelos 2 y 3 son un 25,22 % y 33,93 % mayor que el mínimo posible, respectivamente. Esto significa que las restricciones tienen un alto impacto en el valor de las funciones objetivo cuando entendemos los quintiles de esta forma. Por su parte, cuando usamos la segunda definición de quintil, la diferencia entre los óptimos de los modelos 2 y 3 con respecto a la cota inferior mostrada en el párrafo anterior es de un 2,72 % y de un 20,65 %, respectivamente. Por lo tanto, el impacto de la discriminación positiva en el modelo 2 es muy bajo, pero en el modelo 3 sigue siendo significativo.

Otro dato interesante es que con la segunda definición de quintil el Algoritmo de Selección salta directamente del paso 1 al paso 5. En tanto con la primera definición es necesario llegar al paso 3 del algoritmo, antes de ir al paso 5, ya que entre las mejores soluciones de los tres modelos hay 48 coincidencias (de un total posible de 53). En el anexo se muestra la aplicación del algoritmo con los resultados para la primera definición de quintil.

En cuanto a la lista de espera, con la primera definición de quintil, hay nueve personas que forman parte de alguna de las corridas, pero que no están

dentro de los elegidos, luego para formar la lista de espera fue necesario incorporar once personas que no aparecieron en ninguna de las mejores soluciones. En tanto con la segunda definición hay sólo tres individuos, que si bien aparecen en alguna mejor solución, igual no son elegidos, por ello se tuvo que elegir a diecisiete personas más para completar la lista. La información del párrafo anterior nos permite señalar que con la primera definición de quintil son 62 las personas que aparecen en alguna de las nueve corridas, mientras que con la segunda definición esa cifra se reduce a sólo 56. Vale decir, con la segunda definición de quintil, no sólo hay alta coincidencia entre las mejores soluciones de cada modelo, sino que también entre las segundas y terceras mejores soluciones de éstos. De hecho las segundas mejores soluciones coinciden, al igual que las terceras soluciones de los modelos 1 y 3, las que sólo discrepan en un candidato con la solución del modelo 2. Finalmente los organizadores optaron por la segunda definición de quintil, con el objetivo de hacer más inclusivo el sector integrado por los quintiles inferiores. Es por ello que de aquí en más cuando hablemos de quintiles o de la solución del problema estaremos haciendo alusión a dicha segunda definición. A continuación se muestra un resumen del resultado final con el puntaje de los postulantes ordenado de mayor a menor.

Cuadro 2: Resultado Final

Puntaje	Género	Es de región?	Es de los quintiles inferiores (Def. 2)?	Elegido
77,3967	Masculino	Si	Si	Si
74,6663	Masculino	No	Si	Si
73,1412	Femenino	No	Si	Si
70,9622	Masculino	No	Si	Si
70,2533	Masculino	Si	Si	Si
70,0854	Masculino	Si	Si	Si
68,9846	Masculino	No	Si	Si
68,3338	Masculino	Si	Si	Si
68,2611	Masculino	No	Si	Si
68,2314	Femenino	Si	Si	Si
67,7061	Masculino	No	No	Si
67,5873	Masculino	Si	Si	Si
67,4197	Masculino	Si	Si	Si
67,3683	Femenino	No	Si	Si
67,336	Masculino	Si	Si	Si
67,148	Masculino	No	No	Si
65,7685	Masculino	No	Si	Si
65,3751	Femenino	Si	Si	Si
65,0443	Masculino	No	Si	Si
64,495	Femenino	No	Si	Si
64,2388	Femenino	Si	Si	Si
63,8693	Masculino	No	Si	Si
63,4154	Masculino	No	Si	Si
63,3793	Masculino	No	Si	Si
63,0156	Masculino	Si	Si	Si
62,8584	Masculino	No	Si	Si
62,7446	Masculino	No	Si	Si
62,6285	Masculino	Si	Si	Si
62,2127	Masculino	Si	Si	Si
62,1483	Masculino	Si	Si	Si
62,1481	Masculino	Si	Si	Si
62,0838	Femenino	Si	Si	Si
62,0832	Masculino	No	Si	Si
62,0342	Masculino	No	Si	Si
61,9296	Femenino	No	Si	Si
61,7264	Masculino	Si	Si	Si
61,4523	Masculino	Si	No	Si
61,1665	Masculino	Si	Si	Si

60,8406	Masculino	No	Si	Si
60,8082	Masculino	Si	Si	Si
60,6791	Masculino	No	Si	Si
60,6263	Masculino	Si	Si	Si
60,4522	Masculino	Si	No	Si
60,3994	Masculino	No	Si	No
60,0838	Masculino	No	Si	No
59,9693	Femenino	Si	Si	Si
59,8533	Masculino	Si	No	Si
59,4615	Femenino	Si	Si	Si
59,4572	Femenino	Si	Si	Si
59,4336	Masculino	No	Si	No
59,2239	Femenino	Si	Si	Si
58,7673	Masculino	No	Si	No
58,659	Femenino	Si	Si	Si
58,6414	Femenino	Si	Si	Si
58,5681	Masculino	Si	Si	Si
57,7766	Masculino	No	Si	No
57,7504	Femenino	Si	Si	Si
57,5946	Masculino	No	Si	No
57,5842	Masculino	No	Si	No
57,5556	Masculino	No	Si	No
57,5294	Masculino	Si	Si	No
57,3431	Masculino	Si	Si	No
57,2793	Masculino	No	No	No
56,9899	Femenino	Si	Si	Si
56,5452	Masculino	Si	Si	No
56,5313	Femenino	No	Si	No
56,4444	Masculino	Si	Si	No
56,421	Masculino	Si	No	No
56,2399	Masculino	No	Si	No
56,1681	Masculino	No	Si	No
55,9509	Femenino	No	Si	No
55,821	Masculino	Si	Si	No
55,6551	Masculino	No	Si	No
55,4727	Femenino	No	Si	No
55,4646	Femenino	No	Si	No
55,4358	Masculino	Si	Si	No
55,2457	Masculino	No	Si	No
55,2024	Masculino	No	Si	No
55,0265	Femenino	No	Si	No
55,0064	Masculino	No	Si	No
55,0007	Masculino	Si	Si	No
54,9741	Femenino	No	Si	No
54,9328	Masculino	Si	Si	No
54,59	Masculino	No	No	No
54,4713	Masculino	Si	Si	No
54,3639	Femenino	Si	No	No
54,1758	Masculino	No	Si	No

5. Análisis y conclusiones

Analizando los resultados del problema encontramos que el número de mujeres y candidatos de regiones seleccionados es exactamente igual al mínimo requerido, lo que no ocurre con los postulantes pertenecientes a los cuatro quintiles inferiores. Por lo tanto, las restricciones alusivas a las mujeres y a las regiones son activas. Luego, eliminar la restricción de los quintiles (habiendo elegido la segunda definición), no altera la solución del problema, pero variar las otras dos restricciones si puede provocar alteraciones en el resultado. Cabe acotar también que si no se exige un mínimo número de mujeres, sale una mujer de la selección que es reemplazada por un hombre, mientras que si no se pide un mínimo número de seleccionados de regiones, habría dos postulantes

menos de regiones dentro de los seleccionados. Por lo tanto, el sacar una de las restricciones manteniendo el resto no provoca grandes cambios en la solución final.

A continuación se puede ver el valor de la función objetivo para cada uno de los modelos en sus distintas corridas sin exigir un mínimo número de mujeres y luego sin demandar una cupo de postulantes de regiones. Por lo ya expuesto, los resultados sin pedir un mínimo número de seleccionados pertenecientes a los cuatro quintiles inferiores son los mismos que aparecen en la sección anterior.

Cuadro 3: Resultados sin requisito sobre mínimo número de mujeres

Modelo	Mejor Solución	Valor de la F.O.
1	1	3393,2192
1	2	3393,0329
1	3	3392,9395
2	1	1467
2	2	1468
2	3	1470
3	1	61,2394
3	2	61,294
3	3	61,2946

Cuadro 4: Resultados sin requisito sobre mínimo número de seleccionados de regiones

Modelo	Mejor Solución	Valor de la F.O.
1	1	3393,7415
1	2	3393,511
1	3	3393,2829
2	1	1457
2	2	1459
2	3	1459
3	1	64,2914
3	2	64,2918
3	3	64,2926

Si comparamos estos resultados con los obtenidos considerando todas las restricciones se puede apreciar que las variaciones son pequeñas, lo que es esperable, ya que los cambios entre los seleccionados son mínimos.

Por su parte, un punto de gran interés para los organizadores del programa es el impacto del test psicológico en la selección de candidatos. De hecho, dicha evaluación eliminó a un 42 % de los participantes que habían llegado a la tercera etapa de la selección. De aquí se concluye que la evaluación psicológica tiene un alto impacto para ver quienes serán los elegidos para el Magíster. Es más, si no se hubiese llevado a cabo el test psicológico, 13 personas que están en la lista de seleccionados hubiesen sido reemplazadas por postulantes que quedaron eliminados por la aplicación de este test.

Es interesante ver que cuando buscamos los elegidos sin aplicar el test es necesario llegar al paso 3 del Algoritmo de Selección, lo que no ocurrió con la aplicación del mismo.

El impacto también se puede analizar en el valor de las funciones objetivo. La siguiente tabla ilustra dichos resultados.

Cuadro 5: Resultados sin la aplicación del test psicológico

Modelo	Mejor Solución	Valor de la F.O.
1	1	3470,1457
1	2	3470,1241
1	3	3470,0253
2	1	1479
2	2	1480
2	3	1480
3	1	68,2982
3	2	68,2984
3	3	68,2992

La mejor solución del modelo 1 sin la aplicación del test es un 2,28 % mejor que la que se obtiene al aplicarlo. En cambio en los modelos 2 y 3 el valor de la función objetivo de la mejor solución para cada uno empeora en un 0,61 % y en un 6,22 % respectivamente al no llevar a cabo el test. Así, es posible concluir que en términos de las funciones objetivo no hay una gran diferencia en los resultados, cosa que si ocurre con quienes son seleccionados.

No deja de ser interesante el hecho de que el resultado del modelo 1 mejore sin la aplicación del test, en cambio los resultados de los otros modelos empeoren. Lo que ocurre es que la aplicación de la evaluación psicológica disminuye la región factible. Por lo tanto, si reducimos el número de candidatos y aplicamos el modelo 1 es imposible obtener una mejor solución que la que se obtiene sin aplicar el test. En cambio, en los modelos 2 y 3, esta reducción en el número de los participantes no trae consigo un peor valor para las funciones objetivo, todo va a depender de quienes fueron dados de baja (la reducción de la región factible no afecta a priori sobre la suma de los rankings de los que ingresan o sobre el ranking del último que es seleccionado).

Finalmente, y como conclusión general del trabajo, nos parece importante destacar el aporte social de la investigación de operaciones y la programación matemática con el fin de obtener la lista final de postulantes seleccionados que más se adecúan al perfil fijado por el Magíster, bajo criterios de equidad regional, socio-económicos y de género. Encontrar soluciones robustas a este problema en pocos minutos hubiera sido prácticamente imposible mediante métodos manuales. La herramienta utilizada también agrega transparencia al proceso de selección.

Agradecimientos: A Lysette Henríquez, Patricio Meller y todo el grupo de organizadores del MGG, con quienes fue un gran placer desarrollar este proyecto. A los revisores, por las múltiples sugerencias que contribuyeron sensiblemente a mejorar el trabajo. El primer autor está parcialmente financiado por el proyecto Fondecyt 1050747 y por el Instituto de Ciencias Milenio "Sistemas Complejos de Ingeniería".

Referencias

[1] Epstein, R., L. Henríquez, J. Catalán, G. Weintraub, C. Martínez and F. Espejo, "A Combinatorial Auction Improves School Meals in Chile: A Case of OR in Developing Countries´´. International Transactions in Operational Research 11, 593-612, 2004.

6. Anexo: Aplicación del Algoritmo de Selección a los resultados con la primera definición de quintil

Para ilustrar el funcionamiento del Algoritmo de Selección se muestra la aplicación del mismo cuando se trabajó con la primera definición de quintil. A continuación aparecen dos tablas. La primera contiene a las personas que forman parte de al menos alguna de las 3 mejores soluciones en alguno de los modelos. Los números de la primer columna corresponden a la identificación (ID) de los postulantes. Durante todas las etapas de esta selección se trabajo con el ID, para darle más transparencia al proceso. Notar que el modelo 2 tiene dos mejores segundas soluciones, que las llamamos 2a) y 2b).

La segunda tabla muestra un resumen con los seleccionados directos (Paso 1), los candidatos a ser elegidos junto con su ponderador, quienes son elegidos dentro de este grupo de postulantes (Paso 3) y, por último, los miembros de la lista de espera (Paso 5). En ambas tablas si el candidato i posee el atributo j, la componente i,j de la tabla es marcada con una X.

Cuadro 6: Resultados

ID Persona	MS 1 Mod 1	MS 2 Mod 1	MS 3 Mod 1	MS 1 Mod 2	MS 2A Mod 2	MS 2B Mod 2	MS 1 Mod 3	MS 2 Mod 3	MS 3 Mod 3
036	X	X	X	X	X	X	X	X	X
039	X	X	X	X	X	X	X	X	X
060	X	X	X	X	X	X	X	X	X
077	X	X	X	X	X	X	X	X	X
083	X	X	X	X	X	X			
093	X	X	X	X	X	X			
108	X	X	X	X	X	X	X	X	X
130	X	X	X	X	X	X	X	X	X
133	X	X	X	X	X	X	X	X	X
144	X	X	X	X	X	X	X	X	X
166		X							
186					X		X	X	X
191	X	X	X	X	X	X	X	X	X
192	X	X	X	X	X	X	X	X	X
198	X	X	X	X	X	X	X	X	X
202	X	X	X	X	X	X	X	X	X
209	X	X	X	X	X	X	X	X	X
232	X	X	X	X	X	X	X	X	X
243	X	X	X	X	X	X	X	X	X
249	X	X	X	X	X	X	X	X	X
253	X	X	X	X	X	X	X	X	X
254				X	X	X		X	
255	X	X	X	X	X		X	X	X
257							X	X	X
269	X	X	X	X	X	X	X	X	X
273									X
274	X	X	X	X	X	X	X	X	X
280	X	X	X	X	X	X	X	X	X
293	X	X	X	X	X	X	X	X	X
312	X	X	X	X	X	X	X	X	X
323	X	X	X	X	X	X	X	X	X
324	X		X			X			
334	X	X	X	X	X	X	X	X	X
354	X	X	X	X	X	X	X	X	X
358	X	X	X	X	X	X	X	X	X
364	X	X	X	X	X	X	X	X	X
369							X	X	X
370	X	X	X	X	X	X	X	X	X
374	X	X	X	X	X	X	X	X	X
384	X	X	X	X	X	X	X	X	X
407	X	X	X	X	X	X	X	X	X
422	X	X	X	X	X	X	X	X	X
469	X	X	X	X	X	X	X	X	X
471	X	X	X	X	X	X	X	X	X
494	X	X	X	X	X	X	X	X	X
496	X	X	X	X	X	X	X	X	X
507	X	X	X						
511	X	X	X	X	X	X	X	X	X
516			X						
531	X	X		X	X	X			
533	X	X	X	X	X	X	X	X	X
551	X	X	X	X	X	X	X	X	X
587				X		X	X	X	X
589	X	X	X	X	X	X	X		X
592	X	X	X	X	X	X	X	X	X
595	X	X	X	X	X	X	X	X	X
604	X	X	X	X	X	X	X	X	X
605	X	X	X	X	X	X	X	X	X
608	X	X	X	X	X	X	X	X	X
613	X	X	X	X	X	X	X	X	X
617	X	X	X	X	X	X	X	X	X
621							X	X	

Cuadro 7: Resumen

ID Persona	Elegidos (Paso 1)	Candidatos a ser elegidos (ponderador)	Elegidos (Paso 3)	Lista de espera
36	X			
39	X			
60	X			
77	X			
83		X(3,8)	X	
93		X(3,8)	X	
108	X			
130	X			
133	X			
140				X
144	X			
150				X
166		X(0,6)		X
186		X(2,35)		X
191	X			
192	X			
195				X
198	X			
201				X
202	X			
209	X			
232	X			
243	X			
249	X			
253	X			
254		X(2,5)	X	
255	X			
257		X(1,9)		X
258				X
269	X			
273		X(0,3)		X
274	X			
280	X			
293	X			
312	X			
319				X
323	X			
324	X	(1,75)		X
334	X			
354	X			
358	X			
364	X			
369		X(1,9)		X
370	X			
374	X			
384	X			
407	X			
422	X			
454				X
469	X			
471	X			
494	X			
496	X			
507		X(1,9)		X
511	X			
516		X(0,3)		X
531		X(3,5)	X	
533	X			
551	X			
581				X
582				X
587		X(3,35)	X	
589	X			
592	X			
595	X			
604	X			
605	X			
608	X			
613	X			
617	X			
618				X
621		X(1,6)		X
626				X

Un Enfoque Memético de los Sistemas de Información

Elena Durán*
Silvina Unzaga*

Resumen

Diferentes investigaciones destacan la importancia que se debe dar a los factores culturales y a las visiones de las organizaciones para proponer procesos de cambio viables. Ante esto, es preciso replantear el enfoque utilizado para desarrollar SI, con el fin de contemplar los aspectos culturales de quienes los operan. En este articulo se presenta un Modelo Memetico para la construcción de SI. Este modelo permitirá desarrollar sistemas sobre la base de un análisis de los aspectos culturales de la organización y, posee características evolutivas que posibilitará que el SI se adapte a los cambios culturales del entorno.

Palabras Clave: : Sistemas de Información, Memética, Cultura Organizacional, Variables y Dimensiones Culturales, Modelo Memético.

*Departamento de Informática - Facultad de Ciencias Exactas y Tecnologías -Universidad Nacional de Santiago del Estero

1. Introducción

La cultura es aquello que distingue y da identidad a un grupo humano; es la forma como interactúan los integrantes del grupo entre si y con los de afuera, y el modo como acostumbran realizar lo que hacen (Carrada Bravo, 2002).

En las organizaciones laborales conviven personas cuyos comportamientos y relaciones responden a una cultura en la que prevalecen ritos, pautas, códigos que se manifiestan en actitudes que reflejen valores sustentados en creencias arraigadas tanto en el espíritu individual como colectivo, que muchas veces no coincide con las nuevas misiones, con los nuevos paradigmas y con nuevas maneras de abordar las diversas y complejas situaciones y relaciones sociales que origina la modernidad (Davis y Newstrom, 1985).

Entre estas nuevas maneras de abordar las situaciones, están los SI, los que muchas veces son rechazados por los trabajadores de la organización ya que alteran su cultura de trabajo.

En este sentido Rodríguez Ulloa (Rodríguez Ulloa, 1994), sistemista peruano, realiza un análisis que se centra en la relación que existe entre los grupos culturales, el manejo estratégico de las instituciones y el cambio organizacional. En este estudio, se indica la importancia que hay que darle a los factores culturales y a las visiones de las organizaciones para entender y proponer procesos de cambio estratégicos y viables. Según esto, la cultura de la organización aparece como sumamente significativa para los éxitos o fracasos de sus SI.

Sin embargo, las actuales metodologías de desarrollo de SI, en su mayoría, no consideran factores culturales. Esto trae aparejado la dificultad en la implantación de SI, y SI con estructuras rígidas que les impiden evolucionar e ir adaptándose a los cambios de la cultura organizacional.

Ante esto, nuestra investigación se centra en desarrollar un enfoque para el diseño de SI, que considere los aspectos culturales de la organización, como un modo de asegurar el éxito frente a los cambios organizacionales. Este nuevo enfoque busca integrar, como parte de sus herramientas de análisis, aspectos de la cultura organizacional en la que se encuentra inserto el SI; permitiendo la producción de sistemas más flexibles y mejor adaptados a quienes los usan.

Para concretar el diseño del enfoque se realizó un estudio inicial, tendiente a identificar las principales variables de análisis de la cultura organizacional, sus dimensiones e indicadores. Estos últimos se aplicaron, en un estudio de campo, a los usuarios de SI de una organización pública del medio, con el fin de identificar los memes compartidos por la comunidad de usuarios.

En el presente artículo se describe el estudio inicial realizado y el modelo memético de los SI, obtenido como producto de la investigación. Este modelo permite desarrollar sistemas sobre la base de un análisis de los aspectos culturales de la organización, y posee características evolutivas que permiten al SI ir adaptándose a los cambios culturales del entorno.

En la siguiente sección se presentan las bases conceptuales en las que se sustenta la propuesta, en la sección 3 se mencionan algunos antecedentes relevantes sobre el tema, en la sección 4 se describe el estudio inicial realizado para identificar los memes compartidos por la comunidad de usuarios, en la sección 5 se presenta el Modelo Memético propuesto, y por último, en la sección 6 se resumen algunas conclusiones obtenidas durante el desarrollo de la investigación.

2. Bases Conceptuales

El presente trabajo posee como bases conceptuales: la *Memética*, el concepto de *Cultura Organizacional*, y los *Sistemas de Información* con un abordaje desde un enfoque sociotécnico (Laudon y Laudon, 1996). Las mismas se describen brevemente a continuación.

2.1. Memética

Richard Dawkins, zoólogo especializado en Teoría de la Evolución, define a la Memética como un nuevo campo científico que analiza la transmisión de la cultura a través de memes (Dawkins, 1985). Dawkins sostiene que la evolución cultural es análoga a la evolución biológica y, en general, a todo proceso evolutivo. Él manifiesta que, así como la vida evoluciona por la supervivencia diferencial de los genes, sometidos a la selección natural, la cultura evoluciona mediante la supervivencia diferencial de los replicadores culturales, que también se someten a un proceso de selección. Estas unidades mínimas de información y replicación cultural fueron bautizadas por Dawkins como memes. Así, podemos citar como ejemplos de memes: slogans, frases hechas, melodías, íconos, inventos, modas, ideas, etc.

Un meme posee las características propias de todo proceso evolutivo: *fecundidad*, *longevidad* y *fidelidad* en la replicación. A su vez, estas se dan en un amplio campo de variación, se replican a sí mismas por mecanismos de imitación y transmisión, de cerebro a cerebro, y engendran un amplio abanico de copias que subsisten en diversos medios. Con ello tenemos el marco general de un proceso evolutivo que, Dawkins compara con la evolución biológica, e incluso llega a aceptar que los memes deben ser considerados como estructuras vivientes no sólo metafóricamente, sino técnicamente.

Al concepto de meme, por analogía con la evolución biológica, se le aplica la teoría de la *Selección natural*, la cual se basa en dos pilares fundamentales: *la supervivencia del más apto y la reproducción diferenciada.* La supervivencia del más apto declara que los organismos que mejor se adaptan a su ambiente son los que poseen más probabilidades de sobrevivir. La reproducción diferenciada indica que los organismos mejor adaptados dejan en promedio a más descendientes. En definitiva la selección natural ocurre siempre que existan las siguientes condiciones (Dennet, 1990):

- *Variación:* una continua abundancia de elementos diferentes.
- *Herencia o replicación:* los elementos tienen la capacidad de crear copias o replicarse entre si.
- *Diferenciar "la aptitud":* el número de copias de un elemento que se crea en un tiempo dado varía, dependiendo de la interacción entre los rasgos del elemento (esto es lo que hace diferente de otros elementos) y rasgos del ambiente en el que persiste.

Sobre esta base teórica se construye el modelo de interacción descripto en el apartado 5.

2.2. Cultura organizacional

Otro concepto sobre el que se ha trabajado es el de *cultura organizacional.* Dado que existen numerosas definiciones de este concepto, consideramos importante presentar la definición al que adherimos en nuestro trabajo. En este caso hemos tomado la definición dada por Schein, (Schein,1992) para quien cultura organizacional es .[el] patrón de premisas básicas que un determinado grupo inventó, descubrió o desarrolló en el proceso de aprender a resolver sus problemas de adaptación externa y de integración interna y, que funcionaron suficientemente bien a punto de ser consideradas validas y, por ende, de ser enseñadas a nuevos miembros del grupo como la manera correcta de percibir,

pensar y sentir en relación a estos problemas". La cultura es el "pegamento"social que mantiene unida a una organización. Expresa los valores o ideales y creencias que los miembros de la organización llegan a compartir, manifestados en elementos simbólicos, como mitos, rituales, historias, leyendas y un lenguaje especializado. Podemos develar la cultura de una organización a partir de la observación de tres niveles.

El **primer nivel** es el de los *artefactos visibles* que comprende el ambiente físico de la organización, su arquitectura, los muebles, los equipos, el vestuario de sus integrantes, el patrón de comportamiento visible, documentos, cartas, etc.

El **segundo nivel**, es el de los *valores*, que dirigen el comportamiento de los miembros de la empresa. Dentro de los valores podemos distinguir:

- *Valores grupales:* constituyen aspiraciones o propósitos que benefician a un cierto grupo dentro de la organización.
- *Valores organizacionales:* son las cuestiones importantes que comparte la organización en su conjunto.

El **tercer nivel** es el de los *supuestos inconscientes*, que revelan la forma como un grupo percibe, piensa, siente y actúa. Estos supuestos son construidos a medida que se soluciona un problema eficazmente. En un primer momento, fueron valores conscientes que guiaron las acciones de los miembros de la organización en la solución de problemas, tanto internos como externos. Con el pasar del tiempo dejaron de ser cuestionados, constituyéndose en "verdades", volviéndose inconscientes. Este último nivel está compuesto por las dimensiones siguientes:

Dimensión 1: Relación de la organización con el ambiente externo. Refleja los supuestos que la organización tiene sobre su misión principal en la sociedad, su razón de ser", el tipo de producto o servicio que ofrece, su mercado, su clientela, etc.

Dimensión 2: Naturaleza de la verdad y de la realidad. Son los supuestos básicos, las reglas verbales y comportamentales sobre la realidad, la verdad, el tiempo, el espacio y la propiedad, que sirven de base para la toma de decisiones.

Dimensión 3: Naturaleza de la naturaleza humana. Refleja la visión de hombre que posee la organización y su aplicación a los diferentes niveles de funcionarios y empleados.

Dimensión 4: Naturaleza de la actividad humana. Refleja la concepción de trabajo y de descanso que se tiene en la organización.

Dimensión 5: Naturaleza de las relaciones humanas. Se refiere a la manera considerada correcta para que las personas se relacionen unas con otras. Verifica además, en qué patrones está fundamentada la relación de la organización con los funcionarios.

2.3. Enfoque Sociotécnico de los SI

Se fundamenta en el análisis realizado por Rodríguez Ulloa (Rodriguez Ulloa, 1994) de la relación que puede existir entre los grupos culturales, el manejo estratégico de las instituciones y el cambio organizacional. Según esto, la cultura de la organización aparece como sumamente significativa para los éxitos o fracasos de sus SI. Este nuevo enfoque de los SI lleva a Laudon y Laudon (Laudon y Laudon, 1996) a considerar a los SI como sistemas *sociotécnicos*, compuestos por máquinas, dispositivos y tecnología dura (hardware) pero que requieren de una investigación organizacional y social para que el trabajo sea adecuado. Por lo tanto, para entenderlos es necesario conocer en amplitud las *tecnologías*, la *administración* y la *organización*. Los *administradores* son los que perciben los retos de negocio que impone el entorno, deciden la estrategia que debe adoptar la institución para responder a esos retos, y asignan los recursos humanos y financieros para cumplir esa estrategia y coordinar el trabajo. La *tecnología* de los SI es una de las muchas herramientas de las que los administradores pueden disponer para enfrentar el cambio. Los SI son, además, parte de la organización. La figura 1 muestra esta visión de los SI.

Figura 1. Visión de los SI

3. Trabajos Relacionados Relevantes

Existen varios antecedentes respecto a la consideración de la cultura organizacional en el desarrollo de SI. Algunos de ellos se remontan a la década del 90. Por ejemplo, en (Shirani et al, 1994), se presenta un modelo de satisfacción de usuarios de información, que explica la satisfacción como una consecuencia de la combinación de características: del usuario, organizacionales y del sistema. Entre las características de la organización consideran: la estructura, la cultura y las políticas. Estas características, junto con las del usuario, son consideradas claves para comprender las expectativas del usuario.

En (Butler y Fitzgerald, 1997) se presenta un estudio de la participación de los usuarios en el proceso de desarrollo de SI. Se adopta un enfoque cualitativo, basado en casos, para investigar y proporcionar una descripción en profundidad de la naturaleza social compleja de los fenómenos que se manifiestan en una organización. La conclusión central de ese estudio es que la no satisfacción de los usuarios respecto a los SI, se debe al pobre manejo de los cambios organizacionales, por lo que ponen de manifiesto la necesidad de un alto nivel de participación de todos los actores en las prácticas de desarrollo de SI, ya que esto tiene una influencia positiva en la cultura y el clima del ambiente de desarrollo.

En (Kelly F., 1994) se propone un sistema de información ejecutivo, para el que se reconoce una influencia fundamental de la cultura organizacional en la adopción y uso del sistema. Cita además, una serie de características de la cultura que contribuyen directamente al éxito o fracaso del proyecto, tales como: el aprendizaje organizacional, la mejora continua, el manejo de la crisis, el trabajo en grupo, las decisiones basadas en datos, los objetivos específicos y la información compartida.

Entre los antecedentes más recientes podemos citar a (Mitev N., 2000), donde se presenta un estudio de la implementación de un sistema computarizado de reserva de pasajes en una organización de transporte. El principal objetivo del estudio es trasladar las explicaciones comunes de los fracasos y éxitos de los SI y, encontrar formas mas complejas y ricas de comprensión del uso de SI en las organizaciones, a través de la inclusión de factores sociales, económicos, políticos, culturales e históricos.

En (Umarji y Seaman, 2005) se plantea la posibilidad de estudiar la mejora

de los procesos software desde una perspectiva de aceptación de la tecnología. Para ello, es preciso modificar los Modelos de Adaptación de la Tecnología (TAM) (Davies F, 1989) y los Modelos de Teoría del comportamiento Planificado (TPB) (Hardgrave B. y Johnson R., 2003), agregándoles extensiones para tomar en consideración la cultura organizacional, el impacto de los cambios causados por el SI y las características propias de los desarrolladores de software. Así, se obtiene un Modelo de Aceptación de las Mejoras de los Procesos de Software.

En cuanto a trabajos de desarrollo de software que apliquen el enfoque memetico, se han encontrado muy pocos antecedentes. Los más significativos son los desarrollados en el Laboratorio Meme Media de la Universidad de Hokkaido, Sapporo, Japón, liderado por Yuzuru Tanaka. Allí, desde 1992, están trabajando en la implementación de la tecnología de meme media en los SI (Tanaka Y, 2003). Los meme media brindan un recurso de conocimiento para la sociedad de consumo y la cultura del consumo, que requiere nuevos servicios de distribución, administración y recuperación de la información. En uno de sus últimos trabajos (Tanaka Y., 2005) se aplica esta tecnología de los meme media a los sistemas de e-learning permitiendo obtener sistemas con las siguientes características: fácil extracción y reuso de contenidos existentes en la web, herramientas y servicios, fácil personalización y combinación de sistemas de e-learning publicados en la web, y aceleración de la evolución de la cultura memética de los contenidos de e-learning. Para conseguir estos resultados el grupo ha dirigido investigaciones y desarrollos sobre arquitecturas meméticas, proponiendo las arquitecturas de "Pad Inteligentesz Cajas Inteligentes"(Tanaka Y, 2003).

En base a lo expuesto, es posible concluir que a excepción de los trabajos desarrollados en el Laboratorio de Meme-Media, el resto de los trabajos presentados, si bien varios de estos trabajos consideran las características relevantes de la organización en el desarrollo de SI, no lo hacen desde una perspectiva de evolución (memetica).

Por otra parte, los desarrollos del Laboratorio de Meme- Media son de aplicación en un ámbito específico (e-learning), mientras que el enfoque que se presenta en este trabajo es aplicable a la creación de distintos tipos de SI.

4. Identificación de memes en la comunidad de usuarios

En esta sección, se describe brevemente el estudio inicial, realizado para identificar los memes compartidos por la comunidad de usuarios de SI, de una organización del medio (Universidad Nacional de Santiago del Estero - UNSE, Argentina). Un análisis detallado del estudio se encuentra en (Maldonado et al., 2005).

En este estudio, en primer término; se seleccionaron indicadores para identificar los rasgos culturales presentes en usuarios de SI. Para ello se consideraron las variables y dimensiones de análisis de la cultura organizacional definidas en el apartado 2.2. Los indicadores seleccionados se resumen en la Tabla 1.

A continuación, tomando como referencia los indicadores mencionados, se elaboró una encuesta destinada a recolectar información sobre la opinión de los usuarios de SI en la UNSE. Para cada uno de los indicadores se formuló una pregunta (abierta o cerrada), y luego éstas fueron agrupadas en cuatro categorías que se detallan a continuación:

- *Perfil del encuestado.* Permite recabar datos personales, profesionales y laborales del encuestado.
- *Opiniones del encuestado sobre los SI en la organización.* Esta categoría permite obtener información, sobre: políticas o normas existentes en la organización que condicionan el trabajo con los SI, el clima organizacional y su relación con la operación de SI, y el grado en el que los SI contribuyen al logro de los objetivos organizacionales, entre otros aspectos.
- *Opiniones del encuestado sobre los SI que opera.* Permite obtener información sobre la operatividad de los usuarios con los SI, teniendo en cuenta la calidad de los sistemas, la calidad de los equipos de computación y, el nivel de comunicación que existe entre los compañeros.
- *Opiniones sobre sus preferencias como usuario.* En esta categoría se obtiene información que indica las predilecciones que tiene el usuario, con respeto a las características de los SI que él opera.
- *Opiniones sobre los usuarios en general.* Se obtiene información para caracterizar los rasgos culturales de los usuarios de SI.

Variables	Dimensión	Indicadores
Artefactos Visibles	Infraestructura física y material	▪ Equipos de computación - Tecnología
	Estructural	▪ Documentos, normas y procedimientos
		▪ Servicios otorgados
		▪ Políticas informáticas
		▪ Clima organizacional
		▪ Motivación
Valores	Organizacionales	▪ Áreas que ofrecen mayor recompensa
		▪ Imagen
	Grupales	▪ Tabúes
		▪ Comunicación
		▪ Productividad
		▪ Creatividad
		▪ Calidad
Supuestos Inconscientes	Relación del sistema con el contexto	▪ Objetivos
	Naturaleza de la verdad y de la realidad	▪ Reglas y acuerdos no escritos (tipo, espacio, propiedad, información, planificación)
	Naturaleza humana	▪ Concepción de valores
	Naturaleza de la actividad humana	▪ Horarios de trabajo ▪ Capacitación
	Naturaleza de las relaciones humanas	▪ Relaciones entre los miembros

Tabla 1- Indicadores para cada variable de análisis de la cultura organizacional

A los efectos de obtener la muestra para aplicar la encuesta, se llevó a cabo un relevamiento de todas las áreas de la UNSE, con el fin de establecer en cuáles se operaban SI y quiénes eran los encargados de esa tarea. De este relevamiento, resultaron seis áreas Biblioteca Central, Area Contable, Mesa Gral. de Entrada y los Departamentos Alumnos de las Facultades de Cs. Exactas, Humanidades y Agronomía, con un total de 27 usuarios. Dado que la población resultante era pequeña, se decidió encuestar al total de la población.

Luego de realizar el encuestamiento y, con el fin de determinar los rasgos meméticos presentes en la comunidad de usuarios de SI de la UNSE, se efectuó el procesamiento de la encuesta. Los memes resultantes fueron categorizados teniendo en cuenta su grado de incidencia, es decir, según la cantidad de usuarios que comparten el mismo meme. En esta categorización, se consideró que los memes compartidos son aquellos que infectaron a más de un 50 % de la población. El resultado de la categorización se muestra en la Tabla 2.

MEMES IDENTIFICADOS	NIVEL DE INCIDENCIA
Se califica a los SI como útiles.	81%
Los SI inciden aumentando la calidad del servicio que se ofrece en la organización.	81%
Los SI a veces satisfacen los requerimientos del usuario.	74%
Los SI inciden aumentando la productividad del trabajo.	74%
La principal característica que debe poseer un usuario de un SI es responsabilidad.	74%
Los SI no inciden en la carga horaria del operador.	63%
Los usuarios de SI no reciben un trato diferenciado.	59%
La característica que motivan a un usuario a utilizar un SI es la facilidad de ayuda, que el sistema proporcione consejos y ayudas cuando el usuario lo necesite.	52%
Los SI contribuyen parcialmente al logro de los objetivos de la organización.	52%
El usuario prefiere realizar el trabajo con un SI en forma conjunta con sus compañeros.	52%
No existen prejuicios en los usuarios de SI que les impida utilizarlos libre y cómodamente.	48%
Para hacer un uso adecuado de un SI es necesario contar con una formación previa y capacitación continua.	48%
A veces se utilizan los SI como medio de comunicación entre los operadores, niveles superiores y subordinados de la organización.	44%
Cuando un usuario trabaja con un SI prefiere que le permita optar por distintas formas para realizar una tarea.	44%
La documentación que oficialmente establece los procedimientos para el uso de los SI existen y se respetan.	41%
Que las políticas de la organización condicionan el trabajo con los SI.	41%
Los equipos en los que se ejecutan los SI son obsoletos.	37%
Los acuerdos verbales para el uso del SI no existen.	37%
La categoría organizacional seleccionada condiciona parcialmente el trabajo con el SI.	33%
La organización tiene un clima paternalista.	30%

Tabla 2. Nivel de incidencia de los memes en la comunidad de usuarios

5. Modelo Memético de los SI

Para construir el modelo se tomó como base los resultados del estudio descripto en la sección anterior y el Enfoque Sociotécnico de los SI, descripto en la sección 2.3. Este último propone que los SI se diseñen como facilitadores del cambio organizacional, incorporando elementos cuyo eje central es el ser humano y su cultura. Sobre esta base se diseño un modelo de interrelación entre los SI y la cultura organizacional, que se muestra en la Figura 2 (Duran et al., 2004). Esta relación se da en dos sentidos: la cultura condiciona a los SI y los SI se ven condicionados por la cultura. Si consideramos el primer sentido, entonces los SI deberían diseñarse y mantenerse teniendo en cuenta los rasgos que presenta la cultura de la organización. Para ello, es necesario que los modelos de proceso de construcción de software contemplen actividades tendientes a identificar los principales rasgos culturales de las organizaciones, a través del análisis de manifestaciones estructurales, simbólicas-conceptuales, materiales y conductuales de la cultura. El reconocimiento de estas manifestaciones, que se materializan como indicadores, permite analizar las variables de interés que definen la cultura de las organizaciones. En el segundo de los sentidos, la cultura de una organización va cambiando y evolucionando en la medida que la organización y los individuos cambian, a través del entrenamiento y el aprendizaje necesario para que los SI puedan operar.

A partir de los antecedentes nombrados, se diseña el Modelo Memético de los SI. Modelo que debería seguirse en cualquier proceso de construcción de software, para obtener aplicaciones que se adapten a la cultura organizacional. En la Figura 3 se presenta el modelo resultante.

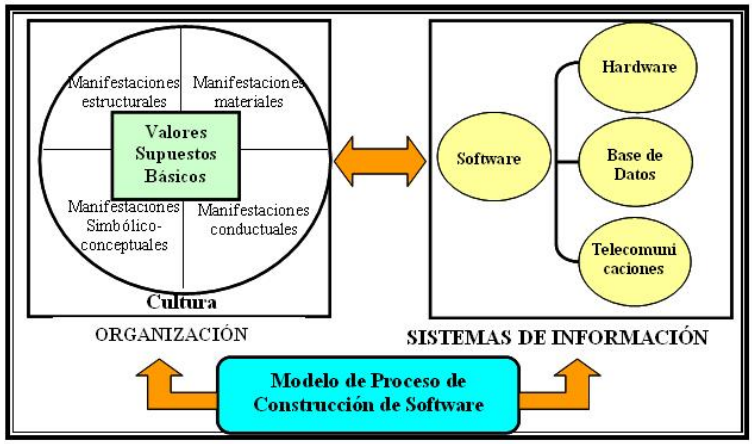


Figura 2. Relación de los SI y la cultura organizacional (Fuente [Duran et al, 2004])

En este modelo, se propone, en primer lugar, identificar los principales rasgos culturales de la organización, a través del análisis de las manifestaciones estructurales, simbólicas-conceptuales, materiales y conductuales de la cultura. El reconocimiento de estas manifestaciones, permite captar el núcleo de los rasgos culturales de la organización, que en el modelo se identifican como *valores y supuestos básicos.*

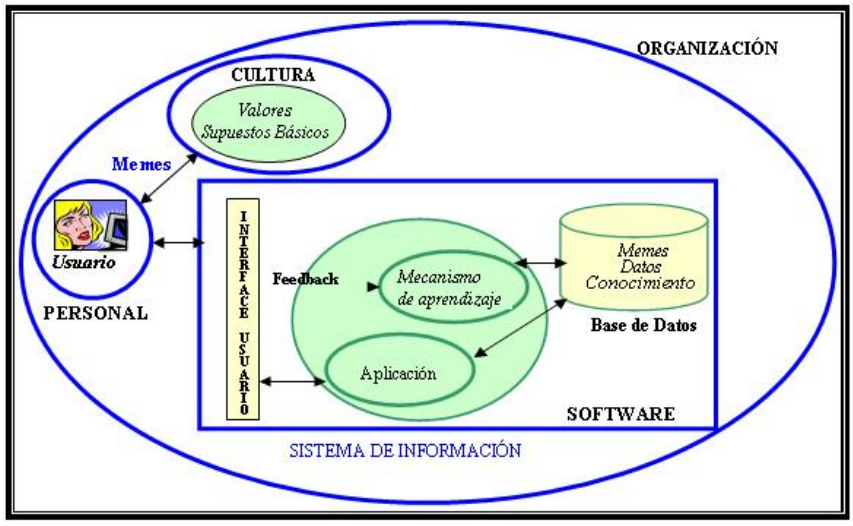
Sobre la base de que la evolución cultural responde a los mismos principios de evolución biológica, se hace necesario identificar las unidades básicas de transmisión cultural o "memes". Estos memes, en el modelo representan justamente el núcleo de los rasgos culturales y, son los que sesgan el comportamiento del personal dentro de esa organización.

El *personal* es otro de los componentes de una organización que mantiene un estrecho vínculo con los SI. Son quienes hacen uso de los mismos, de ahí el nombre de *usuarios*, con el que los hemos identificado en el modelo. Son ellos quienes proporcionan los datos de entrada al SI y quienes reciben los resultados generados por el sistema. Son ellos también, quienes necesitan sistemas acordes a su realidad organizacional, sistemas que se adapten a los cambios que la organización sufre. Particularmente, necesitan sistemas que se adapten a la cultura de la organización, es decir, sistemas que se adapten a los memes que identifican esa cultura. Tomando la afirmación hecha por Rosnay (Ros-

nay, 1996) de que la información se integra por capas concéntricas en niveles superiores: los signos en los saberes, los saberes en los conocimientos y los conocimientos en las culturas; y dado que el avance tecnológico ha permitido que los SI puedan también hoy gestionar conocimiento además de datos; en nuestro modelo, proponemos subir un peldaño más y que el SI de información pueda gestionar memes. Para ello, el SI debe almacenar estos memes como unidad de información cultural. Se propone entonces, que la *base de datos* del sistema, contenga tanto datos y conocimiento, como memes. El sistema debe poder gestionar esos memes, aplicando las operaciones básicas de la Teoría Memética: selección natural y replicación.

En el modelo, el software que conforma el SI debe estar compuesto por la aplicación, con la que interactúa el usuario a través de una *interfaz de usuario*, y por un *mecanismo de aprendizaje*, que capta el feedback explícito y no explícito del usuario y, actualiza la base de memes. El mecanismo de aprendizaje, se basa en técnicas de Inteligencia Artificial, que deberán seleccionarse adecuadamente.

De esta manera, la *aplicación* se presentará ante el usuario respetando los rasgos culturales de la organización, los que serán aprendidos por el sistema, almacenados para su posterior mantenimiento y consultados para adaptar la interfaz de usuario a los memes predominantes.



6. Conclusiones

A partir de la problemática planteada y, siendo conscientes desde un principio que un problema difícilmente pueda ser resuelto si no se logra una comprensión del mismo; buscamos indagar sobre los rasgos culturales que condicionan el trabajo de los usuarios de sistemas. Se ha descripto en este artículo el estudio inicial realizado. Con la concreción de este estudio se logró mejorar la comprensión de la incidencia de los factores culturales en los SI de las organizaciones.

Comprender, la cuestiones relacionadas con la inserción de la cultura organizacional en los SI, nos ha permitido además modelarlas, de forma tal que ese modelo sirva de base para el diseño de SI que se adapten a los memes compartidos por la comunidad de usuarios.

Se ha plantado en este articulo una versión preliminar del Modelo Memetico de los SI, el que será refinado en posteriores procesos de validación del mismo.

Actualmente estamos trabajando en la implementación de este Modelo en un nuevo enfoque memético para la construcción de software de aplicación.

Como línea de acción futura, se aplicará el enfoque memetico en la construcción de SI adaptativos en distintas áreas de aplicación. Esto permite además poder validar el modelo memetico propuesto.

Con esta investigación, ha sido posible establecer las bases para replantear el enfoque utilizado en el desarrollo de SI y contribuir a mejorar el nivel de aceptación de los SI, proporcionando a los usuarios ayuda adaptada a sus rasgos culturales.

Referencias

[1] Green, Paul E. and V. Srinivasan (1990), Çonjoint Analysis in Marketing Research: New Developments and Directions,"Journal of Marketing 54, 4, 3-19.

[2] Butler y Fitzgerald, 1997, .ª Case of User Participation in the Information Systems Development Process", en Proc. ACM International Conference on Information Systems, Atlanta, Georgia, United States, pp. 411-426, ISBN: ICIS1997

[3] Carrada Bravo, 2002 "La cultura organizacional en los sistemas de salud ¿Por qué estudiar la cultura?", Revista Medica del Instituto Mexicano del Seguro Social, Vol 40, Nº3, Pag 203-211.

[4] Davies F, 1989, "Perceived usefulness, perceived ease of use, and end user aceptance of information technology.en MIS Quarterly, Vol.13. pp. 318-339

[5] Davis K y Newstrom J.W, 1985, Çomportamiento humano en el trabajo: comportamiento organizacional", Mc Graw-Hill, México.

[6] Dawkins R., 1985, El Gen Egoísta, 1ª edición, Ed. Salvat, España.

[7] De Souza A., 1998, Çultura Organizacional". http://www.rrhh.net. (julio 2003)

[8] Dennet, D., 1990, "Memes and the Exploitation of Imagination", Journal of aesthetics and art criticism, 48, pp.127-35.

[9] Duran E. y Unzaga S., 2004 Ün Enfoque Memetico de la Cultura Organizacional y su Relación con los Sistemas de Información", Segundas Jornadas de Ingeniería (JUI-2004), Catamarca, Argentina

[10] Hardgrave B. y Johnson R., 2003, "Toward and Information Systems Development Aceptance Model: The Case of Object-oriented System Development", en Proc. IEEE Trans. On Eng. Management. Vol. 50 (3), pp. 322-336.

[11] Kelly Floyd, "Implementing an Executive Information System", www.itmweb.com/essay519, 2005.

[12] Laudon K. y Laudon J., 1996, Administración de los Sistemas de Información. Organización y Tecnología, 3ª edición, Prentice Hall, México.

[13] Maldonado M. Unzaga S, Duran E., y Costaguta R., 2005, .Estudio de los memes compartidos por los usuarios de sistemas información", en Proc. 1º Simposio Internacional de Investigación", Universidad Católica de Santiago del Estero, Dpto. San Salvador de Jujuy, Argentina.

[14] Mitev Natalie, 2000, "Toward Social Constructivist Understandings of is Success and Failure: Introducing a New Computarized Reservation System", en Proc. ACM International Conference on Information Systems,PP. 84-93.

[15] Rodríguez Ulloa R., 1994, La Sistémica, los Sistemas Blandos y los Sistemas de Información, Universidad del Pacífico, Lima, Perú.

[16] Rosnay De J., 1996, El Hombre Simbiótico, Ediciones Cátedra S.A., Madrid, España.

[17] Schein E., 1992, Organizational Culture and Leadership, Jossey-Bass, San Francisco.

[18] Shirani et al, 1994, .A Model of User Information Satisfaction", en Proc. ACM SIGMIS Database. Vol. 25(4), pp.17-23.

[19] Tanaka Y., 2005, "Memetic Approach to the Dissemination of e-Learning Objects", en Proc. ACM International Symposium on Information and Communication Technologies. Vol. 92 pp. 32-37.

[20] Tanaka Y., 2003, Meme Media and Meme Market Architectures. IEEE Press and John Wiley)

[21] Umarji y Seaman, 2005, "Predicting Acceptance of Software Process Improvement", en Proc ACM International Conference on Software Engineering, Workshop on Human and Social Factors of Software Engineering, pp. 1-6.

PRONÓSTICO DEL PRECIO DEL COBRE MEDIANTE REDES NEURONALES

Cristian Foix[*]
Richard Weber[**]

Resumen

El alto nivel que ha alcanzado el precio del cobre en los últimos años, sumado a su gran impacto en la actividad minera nacional motivan a la aplicación de nuevas técnicas para su modelamiento y pronóstico. Un pronóstico ajustado del precio tiene un gran valor no sólo para las empresas de la industria, quienes se podrían beneficiar gracias a la correcta evaluación de proyectos y negocios mineros, sino que también para la autoridad económica en la definición del presupuesto fiscal. En el presente trabajo se entregan evidencias respecto a la potencia de las redes neuronales como herramienta para el pronóstico del precio anual del cobre. Los resultados conseguidos se contrastaron con los generados mediante la aplicación de los más tradicionales y exitosos modelos de series de tiempo. Adicionalmente, se construyeron modelos híbridos combinando modelos ARIMA y redes neuronales. Los resultados revelaron un mejor desempeño de los modelos de pronóstico basados en redes neuronales en el periodo de evaluación considerado (1977-2006), especialmente en pronósticos a más de dos años. Se concluye que las redes neuronales, aunque requieren un mayor esfuerzo en su diseño, pueden ser una herramienta valiosa para el pronóstico del precio del cobre[1].

Palabras Clave: Pronóstico, redes neuronales, series de tiempo.

[*]Dirección de Estudios, Codelco
[**]Departamento de Ingeniería Industrial, Universidad de Chile
[1]El contenido de este trabajo no compromete de manera alguna a Codelco.

1. Introducción

A partir del año 2003, el precio del cobre ha mostrado un extraordinario repunte, llegando a niveles no vistos en más de 30 años. Esto se ha reflejado en nuestra economía donde el cobre ha llegado a representar sobre el 55 % del valor de nuestras exportaciones y sólo los aportes de CODELCO y ENAMI a los ingresos fiscales ya superan el 15 %.

El desconocimiento de los futuros valores de esta importante variable motiva la búsqueda de metodologías de pronóstico eficientes que permitan contar con estimaciones de la mayor precisión posible para su uso tanto en la industria como en la definición del presupuesto de la nación.

En este trabajo se evaluó el desempeño predictivo de diferentes redes neuronales de tipo multilayer perceptron, construidas sobre la base del precio rezagado del cobre y variables derivadas del mismo, tales como la última variación del precio y la desviación estándar de los últimos periodos. Los resultados conseguidos se contrastaron con los generados mediante la aplicación de los más tradicionales y exitosos modelos de series de tiempo (ARIMA, caminata aleatoria y promedio móvil), estableciéndose así el tipo de modelamiento que es capaz de aprovechar mejor la información contenida en los precios históricos considerados en el estudio.

El capítulo 2 muestra antecedentes del presente estudio. En el capítulo 3 se encuentra el modelamiento empleado con redes neuronales y modelos híbridos. El capítulo 4 exhibe los resultados obtenidos. Las conclusiones del presente estudio se encuentran en el capítulo 5 mientras el capítulo 6 describe posibles trabajos futuros.

2. Antecedentes del estudio

2.1. El impacto del precio del cobre en la actividad minera y en la economía chilena

La minería del cobre se ve directamente afectada por los vaivenes del precio. En los periodos de altos precios, aumentan las utilidades de las faenas en operación y se reactivan las exploraciones, los nuevos proyectos, las expansiones y las reaperturas. Por el contrario, en los periodos de bajos precios, disminuyen las utilidades de las operaciones existentes, disminuyen las inversiones en exploraciones, se postergan los proyectos y se cierran las faenas de

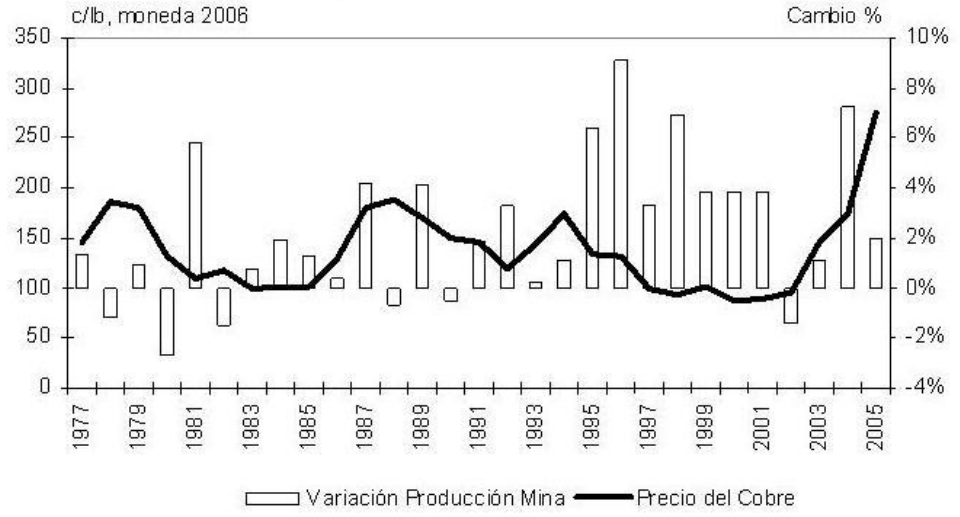
altos costos.

De este modo las fluctuaciones del precio del cobre se traducirán en variaciones de la producción minera, aunque con un desfase dado por la capacidad de reacción de la industria (Gráfico Nº1).

En Chile, estas variaciones de la actividad minera repercuten en las inversiones, en el tipo de cambio vía exportaciones, y en el presupuesto fiscal a través de los impuestos, transmitiéndose al resto de la economía gracias al efecto de estas variables sobre la demanda agregada, la inflación, los precios relativos y el tipo de cambio real [19][24]. Así, en los últimos 20 años, mientras el precio registró un mínimo anual de 87,4 c/lb y un máximo anual de 188,5 c/lb (valores en moneda 2006), la minería del cobre representó entre el 34 % y el 50 % de las exportaciones y los aportes de la Corporación del Cobre (Codelco) llegaron a significar entre el 2 % y el 25 % de los ingresos fiscales [6].

Por lo anterior, contar con proyecciones precisas del precio anual del cobre tiene atractivo no sólo para los productores del metal, quienes resultarían favorecidos por el menor error en un parámetro clave del proceso de planificación minera y evaluación de proyectos, sino que también para el gobierno de la nación. Por ejemplo, el ajuste del gasto fiscal o el endeudamiento en un escenario de bajos precios del cobre dependerá de la duración esperada de la fase depresiva del precio [10].

Gráfico Nº1: Precio del Cobre y Variación Porcentual de la Producción Minera Mundial



Nota: Precio del Cobre entre los años 1977 y 2006, c/lb moneda 2006. 2006, promedio a junio.
Fuente: Cochilco [7] y World Bureau of Metal Statistics.

2.2. ¿Es pronosticable el precio del cobre?

Una variable será pronosticable en la media que su comportamiento histórico nos revele patrones o relaciones que nos permitan estimar su evolución futura. De este modo, el éxito de los pronósticos basados en modelos depende de [5]:

- La existencia de regularidades a ser capturadas.
- El que las regularidades sean informativas acerca del futuro.
- La adecuada captura de estas regularidades por parte del modelo construido.

Con esta definición como referente podemos examinar el caso del cobre. La evolución del precio del cobre desde 1913 en adelante está marcada por su alta volatilidad (Gráfico N°2). Las causas que provocan estas variaciones son diversas y de distinta naturaleza [18]:

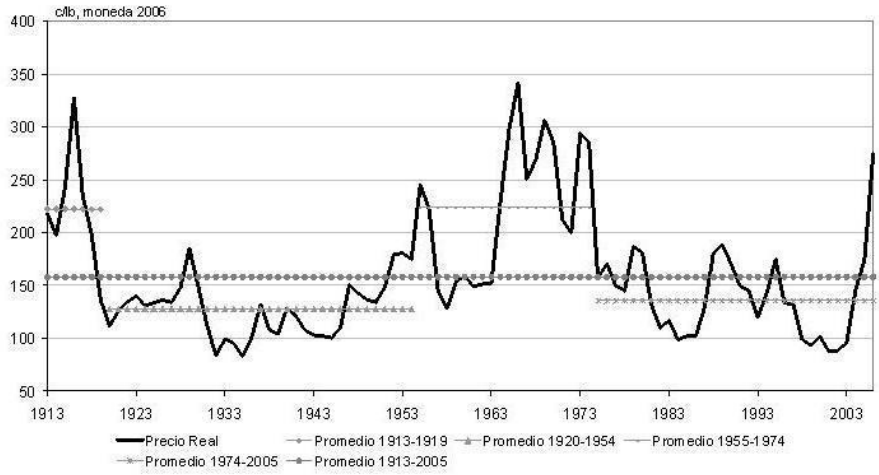
- Fuertes cambios en la industria derivados ya sea de shocks en el precio de los factores productivos, grandes innovaciones tecnológicas o del desarrollo de nuevos mercados, cuyos efectos pueden extenderse por decenas de años. A estos trastornos se les denomina shocks con efecto permanente.
- Desajustes transitorios entre la oferta y la demanda de cobre provocados por los ciclos de la actividad económica mundial y la falta de flexibilidad de los productores para alterar sus tasas de operación, cerrar/reabrir faenas o adelantar el desarrollo de proyectos.
- Fallas de equipos, accidentes, huelgas, terremotos, conflictos políticos y, desde los años noventa, cada vez con mayor fuerza, operaciones financieras, tanto de cobertura como de inversión o especulación, que podrían afectar el precio por días, semanas o meses.

La suma de todos estos eventos determinará el precio del cobre en cada momento. La serie histórica de precios refleja esta descripción: el precio del cobre se mueve por decenas de años, sujeto a shocks temporales, en torno a un nivel definido por distintos shocks permanentes (Gráfico N°2).

La dificultad en el pronóstico del precio tiene que ver con lo complejo que resulta adelantarse a los shocks mencionados. Por otra parte, el potencial de pronóstico del precio del cobre se basa en el decaimiento esperado de los shocks temporales recientemente producidos [21]. La predictibilidad del decaimiento de los shocks temporales obedece al comportamiento de la producción de cobre que, siguiendo a los precios, siempre terminará ajustándose a la demanda en

el mediano y largo plazo. A modo de ejemplo, desde fines del año 2003, la demanda mundial de cobre ha superado con creces a la oferta, lo que ha elevado considerablemente los precios. La oferta ha reaccionado lentamente a esta mayor demanda, esperándose que recién supere al consumo el año 2007 o 2008.

Gráfico Nº2: Precio Promedio del Cobre a través del Tiempo, c/lb moneda 2006.



Fuente: Precio del Cobre: CRU y Cochilco. Deflactor del Precio: U.S. Department of Labor, Bureau of Labor Statistics.

Distinta podría ser la situación de más corto plazo, donde la mayor participación de los agentes financieros en las bolsas de metales, realizando operaciones que involucran cobre en función de la rentabilidad de otros activos, dificulta la proyección de la evolución del precio en los próximos minutos, semanas o meses [18].

En síntesis, estando el precio del cobre en el mediano y largo plazo ligado fuertemente al mercado físico, y evidenciando éste un patrón de comportamiento, las regularidades resultantes podrían ser captadas por modelos de pronóstico.

3. Construcción de modelos para pronosticar el precio del cobre

3.1. ¿Por qué pronosticar el precio del cobre utilizando redes neuronales?

El precio del cobre comúnmente se pronostica recurriendo a la econometría [4][10][14][21][29][30][31]. Para ello se utilizan ya sea grandes modelos estructurales con numerosas ecuaciones y variables, o los más sencillos modelos de series de tiempo construidos únicamente sobre la base del precio rezagado. Existen pruebas de un mejor desempeño de los modelos lineales de series de tiempo más simples (caminata aleatoria, AR1) [10], hecho que les ha ganado la categoría de benchmark en el pronóstico del precio.

Sin embargo, las evidencias de no linealidad en el comportamiento del precio [10], sumada a la falta de estudios sobre el pronóstico del precio del cobre mediante redes neuronales motivan la realización de este trabajo.

3.2. Diseño general del ejercicio de pronóstico

Dado que la evaluación del error de pronóstico será más representativa en la medida que se tengan más mediciones, sobre la base de distintos orígenes (años de altos precios, años de bajos precios) y de distintos alcances (pronósticos a 1 año, a 2 años, etc.), se optó por la generación de pronósticos dinámicos (pronósticos sobre pronósticos), a partir de modelos recalibrados (reestimados) periódicamente, utilizando una muestra de tamaño creciente y con un origen móvil [27]. Se optó por el uso de una muestra de tamaño creciente y no una de tamaño constante debido a que, a la luz de la coyuntura actual de precios altos, se consideró que las observaciones históricas podrían ser de utilidad [21].

En cuanto al alcance de los pronósticos, en el presente trabajo, el foco estará puesto en las proyecciones de corto y mediano plazo, en un horizonte de 1 a 6 años por sus potenciales aplicaciones en la industria del cobre (evaluación y control de los resultados de la gestión y de negocios, planificación de la producción y evaluación de proyectos) y en las finanzas públicas (presupuesto fiscal y Fondo de Compensación del Cobre o similar). No se evaluarán pronósticos de largo plazo, más de 6 años, dado que los modelos ARIMA, utilizados como referente, no tienen un buen desempeño en proyecciones de largo alcance [10].

3.3. Datos utilizados

Las redes neuronales utilizadas en este estudio fueron construidas y entrenadas sobre la base del precio histórico del cobre y variables derivadas del mismo, no se consideró el efecto de variables externas como por ejemplo, el crecimiento de la economía. Los pronósticos fuera de muestra generados con dichas redes se contrastaron con los obtenidos con la aplicación de los más exitosos modelos de series de tiempo. Dado que las redes neuronales pueden alimentarse con variables distintas al precio, la comparación puede parecer injusta, al, eventualmente, subestimarse la capacidad de pronóstico de las redes. Sin embargo, si las redes neuronales demuestran tener un mejor desempeño que los modelos de series de tiempo, se habrá establecido que las redes neuronales pueden sacar mayor provecho de los precios históricos y se tendrá una evaluación conservadora de su eficiencia predictiva que podrá alentar a la realización de investigaciones adicionales.

Por otra parte, la generación de proyecciones del precio empleando otras variables explicativas tiene el potencial problema de requerir el pronóstico de dichas variables, lo que complica su utilización práctica.

Los datos de precios fueron sometidos a dos transformaciones para la construcción de modelos y la realización de pronósticos:

- Escalamiento de valores para eliminar el efecto de la inflación, expresando el precio en moneda constante.
- Aplicación de logaritmo natural a los datos en moneda constante para reducir su dispersión y atacar problemas de heterocedasticidad [16][17].

Los datos utilizados en este estudio corresponden al precio anual del cobre refinado de la Bolsa de Metales de Londres durante el periodo comprendido entre los años 1913 y 2006 (2006, promedio a junio). La elección de este periodo se fundamenta en tres razones:

- Si bien existen datos para el precio del cobre en los Estados Unidos durante el siglo XIX, dichos valores corresponden a un precio de productores, no a un precio de mercado y obedecen a un tipo de minería muy distinta a la gran minería que comenzó a desarrollarse a partir de las primeras décadas del siglo XX (flotación, minería a rajo abierto, mecanización). Por este motivo se excluyeron los datos previos a 1910.
- El deflactor empleado por la Comisión Chilena del Cobre (COCHILCO) para expresar el precio del metal rojo en moneda constante es el Índice de Precios de Productores de los Estados Unidos (PPI All Commodities, Not Seasonally Adjusted) que registra valores desde 1913 en adelante [28].

- En el apartado "¿Es pronosticable el precio del cobre?"se describió al precio como una variable que presenta reversión a la media, en particular, a distintas medias a través del tiempo. Este comportamiento de mediano/largo plazo se observa en la ventana seleccionada, donde el logaritmo natural del precio no presenta tendencia y los test practicados rechazan la presencia de raíz unitaria.

Para medir el error de los pronósticos, se optó por un periodo de evaluación de 30 años, similar al empleado en otros trabajos [10][21] que permitió contar con un número de pronósticos de entre 30 (proyecciones a un año) y 25 (proyecciones a 6 años), abarcando tanto periodos con precios ascendentes como descendentes.

3.4. Modelamiento

Se mostrará el modelamiento con redes neuronales, modelos de series de tiempo (en particular modelos de la familia ARIMA) y modelos híbridos.

3.4.1. Redes Neuronales

Una red neuronal puede ser descrita como un modelo de regresión no lineal cuya estructura se inspira en el funcionamiento del sistema nervioso. En términos generales, una red consiste en un gran número de unidades simples de proceso, denominas neuronas, que actúan en paralelo y están conectadas mediante vínculos ponderados [13][23].

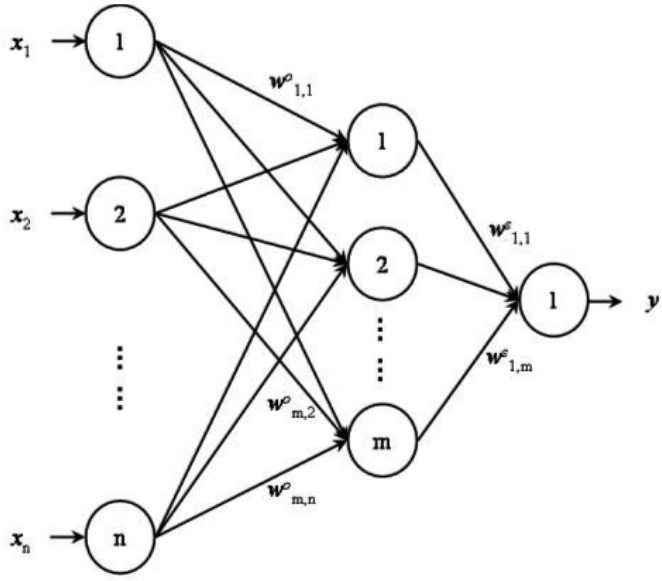
Cada neurona recibe entradas desde otras neuronas y genera un resultado que depende sólo de la información localmente disponible, ya sea almacenada internamente o plasmada en los ponderadores de las conexiones. El resultado generado por la neurona servirá de entrada para otras neuronas.

Mediante la adecuada modificación de los ponderadores de la red, en un proceso denominado aprendizaje, la red mejorará su desempeño en el desarrollo de la tarea para la cual fue construida.

El diseño de una red neuronal es una tarea compleja por la gran cantidad de decisiones que involucra tanto a nivel de su arquitectura como de su mecanismo de aprendizaje. Por los buenos resultados conseguidos en numerosas aplicaciones [15], en este estudio se emplearon redes MLP con una sola capa oculta, una sola neurona en la capa de salida y conexiones hacia delante (Figura Nº1).

En cada una de las neuronas utilizadas, los datos de entrada se combinaron a través de una suma ponderada. Dicho resultado se transformó en las neuronas de la capa oculta, mediante la aplicación de la función tangente hiperbólica, antes de ser transferido a la capa de salida.

Figura Nº1: Red Multilayer Perceptron.



Fuente: Hilera, J. y Martínez, V. Redes Neuronales Artificiales. Fundamentos, Modelos y Aplicaciones. [13]

La red así construida, al recibir el vector de entrada (x1, x2,. . . ,xn), genera el siguiente output:

$$y = w^{sc} + \sum_{j=1}^{m} w^{s}_{1,j} \cdot tanh(w^{oc}_{j} + \sum_{j=1}^{n} w^{o}_{j,i} \cdot x_i) \quad (1)$$

Donde:
w^{oc}_{j}: Ponderador de la conexión con entrada de valor unitario y la neurona j de la capa oculta.
w^{sc}: Ponderador de la conexión con entrada de valor unitario y la neurona de la capa de salida.
$w^{o}_{j,i}$: Ponderador de la conexión entre la neurona i de la capa de entrada y la neurona j de la capa oculta.
$w^{s}_{1,j}$: Ponderador de la conexión entre la neurona j de la capa oculta y la neurona 1 de la capa de salida.

En el mecanismo de aprendizaje aplicado en este estudio, se separan las fases de entrenamiento y aplicación de red, y por corrección de error, donde el ajuste de los ponderadores obedece al error respecto de la respuesta deseada. El método de aprendizaje elegido fue el de Levenberg-Marquardt por la mayor velocidad de convergencia en comparación con el método de descenso del

gradiente. Adicionalmente, evaluaciones preliminares mostraron la obtención de un menor error de pronóstico.

Para el entrenamiento de la red los datos muestrales se dividieron en tres conjuntos disjuntos:

- Entrenamiento: conjunto que contiene los ejemplos que servirán para la modificación de los conectores neuronales.
- Testeo: fracción de los datos muestrales que no participa directamente en el entrenamiento de la red. Durante el proceso de aprendizaje, a intervalos regulares, se evaluó el desempeño de la red en este conjunto para verificar el cumplimiento de la meta en algún indicador crítico, como por ejemplo máximo error cuadrático medio deseado. Una vez alcanzado el valor meta se detuvo el entrenamiento (early-stopping).
- Evaluación: conjunto de datos que no participan ni directa ni indirectamente del proceso de aprendizaje. Los datos del conjunto de evaluación son los que se utilizan para la evaluación de las proyecciones fuera de muestra.

Para el resto de los parámetros de diseño se siguió una estrategia de evaluación de distintos de valores, mediante el uso de grillas [8].

La aplicación de early-stopping requiere definir el tamaño del conjunto de testeo. Se seleccionó una fracción variable de los últimos ejemplos disponibles dentro de la muestra: 10 %, 20 % o 30 % [16], que no participó, directamente, en el entrenamiento de las redes neuronales.

Respecto al número de neuronas en la capa de entrada, se construyeron y entrenaron redes con 3, 6 o 9 neuronas en la capa de entrada, alimentadas con los correspondientes rezagos del precio, mientras que para la capa oculta se utilizaron entre 1 y 6 neuronas. Las combinaciones seleccionadas fueron determinadas por el número de conexiones asociadas y el número de ejemplos del conjunto de entrenamiento. Si bien sólo existen teoremas que relacionan el número de conexiones de la red con el número de ejemplos de entrenamiento para ciertos tipos de redes (multicapas, con entradas y salidas binarias y función de activación escalón) [25], al menos debe respetarse que el número de ejemplos supere, con alguna holgura, el número de coeficientes a estimar (ponderadores) para así evitar el overfitting. Las combinaciones presentadas en la Tabla N°1 cumplen con este requisito, ya que, en promedio, el número de ejemplos de entrenamiento quintuplica el número de coeficientes [16] (el número de ejemplos de entrenamiento depende del tamaño de la muestra, de la fracción de datos muestrales destinada al conjunto de testeo y del número de neuronas de la capa de entrada, moviéndose en un rango de entre 55 y 90 ejemplos).

Tabla Nº1: Neuronas en la Capa de Entrada, en la Capa Oculta y Número de Conexiones.

Neuronas en la capa de entrada	Neuronas en la capa oculta	Número de conexiones
3	1	6
3	2	11
3	3	16
3	4	21
3	5	26
3	6	31
6	1	9
6	2	17
6	3	25
6	4	33
9	1	12
9	2	23
9	3	34

Nota: El número de conexiones considera las conexiones entre capas (incluyendo la capa de salida) y las conexiones correspondientes a las entradas de valor unitario. (Una para cada neurona de la capa oculta y una para la neurona de la capa de salida)

Dado que los ponderadores iniciales de las conexiones determinan el punto de origen del proceso de aprendizaje (minimización de error del modelo), se recomienda repetir el entrenamiento de cada arquitectura unas 15 veces, redefiniendo aleatoriamente dichos pesos [8][26], obteniéndose así 15 redes distintas para una misma arquitectura. Si a esto agregamos que, como ya fue mencionado, cada arquitectura se entrenó utilizando conjuntos de testeo definidos de 3 maneras distintas, el total de redes construidas para una misma combinación de neuronas en la capa de entrada y neuronas en la capa oculta llega a 45 y el total de redes a entrenadas y evaluadas llega a 585.

Las redes neuronales se recalibraron para incorporar la información adicional disponible a medida que se realizaban las evaluaciones. En evaluaciones preliminares, se observó que un mayor número de recalibraciones no siempre se traducía en un mejor pronóstico. Por este motivo, cada red se recalibró 3, 6 y 10 veces.

Para entrenar y evaluar las distintas redes, se empleó el lenguaje de programación MATLAB y su toolbox de redes neuronales (MATLAB 6.1, The MathWorks Inc.). Si bien el entrenamiento y evaluación de cada red tomó pocos segundos, el gran número de redes estimadas requirió de tiempos de cálculo de hasta 6 horas (Pentium IV, 3.0 GHz, 1 GB RAM).

3.4.2. Series de tiempo

En este trabajo se emplearon modelos estocásticos y deterministas de series de tiempo. Los modelos estocásticos utilizados forman parte de la familia ARIMA [22]:

$$\Delta^d Y_t = \delta + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \epsilon_t + \Theta_1 \epsilon_{t-1} + ... + \Theta_q \epsilon_{t-q} \qquad (2)$$

Donde:
$\Delta Y_t = Y_t - Y_{t-1}$
$\Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1}$
ϕ_i, θ_j: coeficientes de las porciones autorregresivas (Y_{t-i}) de media móvil (ϵ_{t-j}) del proceso.

El examen de la función de autocorrelación parcial de la serie en niveles, evaluaciones preliminares realizadas y los resultados conseguidos en otros estudios [10][21], llevaron a la construcción de un set de modelos tentativos cuyos coeficientes son significativos para los datos en niveles en el periodo de interés[2]:

Tabla Nº2: Modelos ARIMA Utilizados en el Ejercicio de Proyección.

AR1	MA3	MA1MA2
AR2	MA4	AR3MA1MA2
AR3	AR1MA1	AR1AR2AR3
AR4	AR1MA2	AR1AR2AR3MA1MA2MA3
MA1	AR2MA1	MA1MA2MA3MA4
MA2	AR3MA1	MA1MA2MA3MA4MA5

Se realizaron pruebas con modelos de varianza condicional constante y modelos GARCH (Heterocedasticidad Condicional Autorregresiva Generalizado) [9]:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + ... + \alpha_p \epsilon_{t-p}^2 + \lambda_1 \sigma_{t-1}^2 + ... + \lambda_q \sigma_{t-q}^2 \tag{3}$$

Adicionalmente, y sólo con fines comparativos, se modeló el precio del cobre como un camino aleatorio con y sin drift.

$$y_t = \delta + y_{t-1} + \epsilon_t \tag{4}$$

Por último, en la línea de los modelos deterministas, se modeló el precio del cobre como un promedio móvil de 3, 6 y 9 años. Es interesante incluir el promedio móvil, dado que, hasta hace algunos años, fue el método utilizado para determinar el precio de referencia del Fondo de Compensación del Cobre (promedio móvil de 6 periodos) [3][10].

$$y_{t+1} = \frac{1}{6} \cdot (y_t + y_{t-1} + .. + y_{t-5}) \tag{5}$$

Para la construcción de modelos estocásticos se utilizó el software EViews 4.1 (Quantitative Micro Software). Las rutinas programadas en EViews para la

[2]Para los modelos de datos en primeras diferencias sólo se obtienen coeficientes significativos con los modelos AR2, MA1 y MA2.

generación y evaluación de proyecciones fuera de muestra sólo necesitaron un par de segundos para ejecutar los cálculos correspondientes a cada modelo, y la evaluación del set completo de modelos tentativos fue cosa de pocos minutos. Para el resto de los modelos se empleó la planilla de cálculo Microsoft Excel 2002 (Microsoft Corporation).

Las evaluaciones realizadas mostraron que los mejores resultados se conseguían recalibrando el modelo cada vez que un nuevo dato histórico se incorporaba al conjunto de entrenamiento. Este efecto ha sido observado en otros trabajos de pronósticos [11].

3.4.3. Modelos Híbridos

Adicionalmente se combinaron los modelos de series de tiempo y las redes neuronales de la siguiente manera [1], [2]:

- Con modelos de series de tiempo:
  - Se realizaron proyecciones fuera de muestra.
  - Se calcularon los errores de proyección fuera de muestra.
- Con redes neuronales:
  - Se interpretó los errores calculados con series de tiempo como una nueva serie de tiempo
  - Se entrenó una red neuronal para pronosticar dicha serie de tiempo.
  - Se pronosticaron los errores de proyección fuera de muestra con la red neuronal.
- Se corrigieron las proyecciones fuera de muestra vía modelos de series de tiempo con los errores pronosticados mediante redes neuronales.

3.5. Evaluación del desempeño predictivo: medición del error

Para la evaluación del desempeño predictivo se emplean diferentes indicadores que cuantifican qué tan cerca está la variable pronosticada de su serie de datos correspondiente [22]. Una de las medidas más utilizadas es el Promedio del Error Porcentual Absoluto (MAPE) [12][22]:

$$MAPE = \frac{1}{T}(\sum_{t=1}^{T} APE_t) = \frac{1}{T}(\sum_{t=1}^{T} \frac{|Y_t^s - Y_t^a|}{Y_t^a}) \cdot 100 \qquad (6)$$

Donde:
APE : error porcentual absoluto
Y_t^s: valor pronosticado de Y_t

Y_t^a: valor real de Y_t
T: número de periodos

El MAPE mide el valor medio del error absoluto en términos porcentuales al valor real de la variable.

Para evaluar la dispersión de los errores se puede calcular el Desvío Estándar del error porcentual absoluto (APE).

$$\text{Desvío Estándar APE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(APE_t - MAPE)^2} \tag{7}$$

Otra medida del error de pronóstico comúnmente empleada es la Raíz Cuadrática Media del Error (RMSE):

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t^s - Y_t^a)^2} \tag{8}$$

Donde:
Y_t^s: Valor pronosticado de Y_t
Y_t^a: Valor real de Y_t
T: número de periodos

El RMSE mide la dispersión de la variable simulada en el curso del tiempo, penalizando fuertemente los errores grandes al elevarlos al cuadrado. Esta característica hace que el RMSE se recomiende cuando el costo de cometer un error es aproximadamente proporcional al cuadrado de dicho error [20].

No siempre el modelo que genere pronósticos con un menor MAPE generará los pronósticos con el menor RMSE y viceversa, por lo que en la selección de los mejores modelos de pronóstico se hace necesario establecer la medida de error a utilizar para la elaboración del ranking de desempeño.

Dado que una mala estimación del precio fututo del cobre se traduce en una pérdida de ingresos proporcional al tamaño del error, el MAPE, y no el RMSE, parece ser la medida de desempeño más adecuada. A esto se suma la ventaja práctica del MAPE de no requerir ser acompañado por la media para dimensionar la magnitud del error. Luego, la medida de error que se empleará para identificar los modelos de mejor desempeño será el MAPE.

Sin perjuicio de lo anterior, para los mejores modelos identificados se presentarán tanto el MAPE, el desvío estándar del APE y el RMSE de sus pronósticos.

4. Resultados obtenidos

4.1. Resultados y comparación de modelos

En esta sección se comparan los errores de pronóstico, a distintos plazos, arrojados por los mejores modelos de series de tiempo, las mejores redes neuronales y la caminata aleatoria. En todos los alcances, las redes neuronales entregaron mejores pronósticos (menor MAPE) que los modelos de series de tiempo, y estos últimos, a su vez, superaron a la caminata aleatoria. La mayor exactitud que registran las redes neuronales respecto a los modelos de series de tiempo es marginal en pronósticos a 1 año, con una reducción del MAPE de 4 %, pero se hace más significativa en pronósticos a 2 y más años, con una reducción de 30 %. En cuanto a la dispersión del error, las redes que generan pronósticos de menor MAPE muestran una dispersión en sus errores similar a la de los mejores modelos de series de tiempo.

En la medición del error aplicando RMSE, nuevamente los mejores modelos de redes neuronales tuvieron un mejor desempeño que los modelos ARIMA y la caminata aleatoria, con reducciones del RMSE de entre 1 % y 23 % y de entre 4 % y 37 % respectivamente.

La notación empleada para identificar las redes es Red(a1; a2; a3; a4; a5), donde:

- a1: Número de neuronas de la capa de entrada.
- a2: Número deneuronas de la capa oculta.
- a3: Razón entre el número de ejemplos efectivamente utilizados en el entrenamiento y el número de ejemplos empleados en el testeo (early-stopping).
- a4: Número correlativo de la inicialización de la red.
- a5: Número de recalibraciones efectuadas.
- uvp: Última variación del precio. Red incluye la diferencia entre los dos últimos rezagos del precio en logaritmo natural.
- dep: Desviación estándar del precio en los últimos periodos. Red incluye la desviación estándar de los precios rezagados que alimentan la capa de entrada.

En las Tablas Nº4 y Nº5, se observa que la ventaja conseguida por las mejores redes neuronales por sobre los modelos ARIMA al evaluar mediante

Tabla N°3: Comparación de Resultados. MAPE (%) en Pronósticos a Distintos Alcances.

	a 1 año
Red(3; 5; 80/20; 14; 3)	13,48%
AR2MA1	14,08%
Caminata Aleatoria	15,26%
% Reducción Error	
Redes v/s ARIMA	-4,20%
Redes v/s Caminata Aleatoria	-11,70%

	a 2 años
Red(6; 3; 70/30; 8; 10)	16,13%
AR1.GARCH(0,2)	22,98%
Caminata Aleatoria	23,82%
% Reducción Error	
Redes v/s ARIMA	-29,80%
Redes v/s Caminata Aleatoria	-32,30%

	a 3 años
Red(3; 2; 80/20; 4; 3)(dep)	18,66%
AR2MA1.GARCH(0,2)	27,02%
Caminata Aleatoria	29,20%
% Reducción Error	
Redes v/s ARIMA	-30,90%
Redes v/s Caminata Aleatoria	-36,10%

	a 4 años
Red(3; 6; 80/20; 11; 3)(dep)	20,63%
AR2MA1.GARCH(0,2)	28,87%
Caminata Aleatoria	33,30%
% Reducción Error	
Redes v/s ARIMA	-28,50%
Redes v/s Caminata Aleatoria	-38,00%

	a 5 años
Red(3; 1; 80/20; 13; 3)(dep)	21,54%
AR2MA1.GARCH(0,2)	31,08%
Caminata Aleatoria	37,37%
% Reducción Error	
Redes v/s ARIMA	-30,70%
Redes v/s Caminata Aleatoria	-42,40%

	a 6 años
Red(6; 3; 70/30; 9; 6)(uvp)	20,33%
AR2MA1.GARCH(0,2)	31,07%
Caminata Aleatoria	40,14%
% Reducción Error	
Redes v/s ARIMA	-34,60%
Redes v/s Caminata Aleatoria	-49,30%

MAPE, se hace menos significativa al emplear una mediada como el RMSE que penaliza los errores de mayor tamaño. Sin embargo, se debe recordar que los mejores modelos de redes neuronales fueron seleccionados por su menor MAPE y no por su menor RMSE.

En cuanto a la desviación estándar del error porcentual absoluto, mediante el uso de redes neuronales se observa una reducción de la dispersión en la mayor parte de los alcances.

Utilizando un modelo híbrido se generaron proyecciones a un año, buscando determinar la mayor exactitud que se podría conseguir con el uso combinado de series de tiempo y redes neuronales. Para la aplicación del modelo híbrido se trabajó con el modelo AR2MA1, por ser éste el de mejor desempeño en pronósticos a un año, mientras que, a nivel de las redes se aplicó la misma estrategia de búsqueda en torno a una grilla descrita en la sección de redes neuronales. Sin embargo, dado que se debe trabajar con residuos de proyecciones fuera de muestra y que para construir el primer modelo AR2MA1 se utilizan los datos históricos del periodo 1913 a 1942, sólo se cuenta con 64 datos

Tabla Nº4: Comparación de Resultados. Disminución del MAPE Conseguido con las Mejores Redes Neuronales Seleccionadas.

Reducción del MAPE respecto a:	a 1 año	a 2 años	a 3 años	a 4 años	a 5 años	a 6 años
ARIMA	-4,2%	-29,8%	-30,9%	-28,5%	-30,7%	-34,6%
Caminata Aleatoria	-11,7%	-32,3%	-36,1%	-38,0%	-42,4%	-49,3%

Tabla Nº5: Comparación de Resultados. Disminución del RMSE Conseguido con las Mejores Redes Neuronales Seleccionadas.

Reducción del RMSE respecto a:	a 1 año	a 2 años	a 3 años	a 4 años	a 5 años	a 6 años
ARIMA	-1%	-23%	-6%	-4%	-13%	-18%
Caminata Aleatoria	-4%	-30%	-20%	-22%	-31%	-37%

Tabla Nº6: Comparación de Resultados. Disminución Desvío Estándar del Error Porcentual Absoluto Conseguido con las Mejores Redes Neuronales Seleccionadas.

Reducción del MAPE respecto a:	a 1 año	a 2 años	a 3 años	a 4 años	a 5 años	a 6 años
ARIMA	0,4%	-8,0%	13,5%	-0,3%	-1,5%	-25,4%
Caminata Aleatoria	-2,4%	-23,3%	-18,5%	-23,5%	-22,7%	-39,6%

de residuos para entrenar y evaluar redes. Debido a esto se trabajó con una grilla más pequeña cuidando mantener una adecuada razón entre el número de datos disponibles y el número de conexiones.

Tabla Nº7: Neuronas en la Capa de Entrada, en la Capa Oculta y Número de Conexiones.

Neuronas en la capa de entrada	Neuronas en la capa oculta	Número de conexiones
3	1	6
3	2	11
3	3	16
3	4	21
6	1	9
6	2	17
6	3	25

Nota: El número de conexiones considera las conexiones entre capas (incluyendo la capa de salida) y las conexiones correspondientes a las entradas de valor unitario.

A diferencia de lo ocurrido con las redes empleadas para proyectar el precio, en la proyección de los residuos, los mejores resultados no se obtuvieron actualizando los ponderadores con el método de Levenberg-Marquardt, sino que con el método de descenso del gradiente. Esto podría indicar que la función de error para este problema es menos suave, por lo que el método de Levenberg-Marquardt no converge al mínimo global. La mejor red resultante fue Red(3; 2; 80/20; 9; 3).

La siguiente tabla compara los errores conseguidos con los mejores modelos identificados.

En el conjunto de 30 pronósticos evaluados, el modelo híbrido entregó mejores

Tabla Nº8: Comparación de Resultados. MAPE (%) en Pronósticos a 1 Año.

	a 1 año
Método Híbrido	13,284%
Red(3; 5; 80/20; 14; 3)	13,481%
AR2MA1	14,078%
Caminata Aleatoria	15,264%
% **Reducción Error**	
Método Híbrido v/s Red	-1,5%
Método Híbrido v/s ARIMA	-5,6%
Método Híbrido v/s Caminata Aleatoria	-13,0%

resultados que el más exitoso modelo ARIMA en 16 oportunidades, logrando una disminución del MAPE de 5,6 %.

Otro dato interesante al momento de elegir una herramienta de pronóstico es la capacidad de respuesta frente a grandes cambios de precios. Entre 1977 y 2006, en 14 oportunidades el cambio del precio respecto al año anterior superó el 15 %, entregando el modelo híbrido los mejores pronósticos en 8 ocasiones.

Si la ventana de pronósticos se divide en 3 subperiodos de igual duración, se aprecia que el MAPE del modelo híbrido es menor al MAPE del modelo ARIMA durante los primeros 20 años. En el tercer subperiodo, la comparación favorece al modelo ARIMA, gracias a los menores errores conseguidos en los últimos 3 años.

A pesar de estos buenos resultados, la comparación del modelo híbrido con la mejor red neuronal muestra sólo una mejora marginal del MAPE y un deterioro, también marginal, del RMSE.

Tabla Nº9: MAPE (%) por subperiodo.

	Pronóstico AR2MA1	Método Híbrido
1977-1986	13,030%	10,453%
1987-1996	13,290%	12,985%
1997-2006	15,914%	16,414%

4.2. Comparación con proyecciones de expertos

Es interesante comparar los resultados conseguidos con las proyecciones publicadas en el pasado por analistas del mercado. No fue posible encontrar pronósticos de expertos para el periodo completo entre los años 1977 y 2006,

por lo que la comparación se restringió a los últimos 20 años. Este es un elemento que debe ser considerado en la comparación, dado que los mejores modelos de pronóstico identificados se seleccionaron por la calidad de su desempeño en el periodo de 30 años mencionado y no en la ventana de tiempo 1987 - 2006, por lo que el contraste puede subestimar la precisión de los modelos híbridos y de las redes neuronales.

La Tabla N°10 muestra los errores de pronóstico conseguidos con los mejores modelos identificados y los errores de proyección del consultor Brook Hunt & Associates, una de las empresas más reputadas de la industria.

Tabla N°10: Precio, Pronósticos y Error Porcentual Absoluto (APE).

	Precio Histórico	Pronósticos a 1 año		Pronósticos a 2 años	
		APE Método Híbrido	Proyección Analista	APE Red	Proyección Analista
1987	129,2	14,120%	21,441%		
1988	180,8	24,734%	33,015%	37,414%	39,714%
1989	188,5	5,718%	14,628%	11,768%	36,490%
1990	170,3	5,923%	22,744%	3,698%	34,645%
1991	149,2	2,925%	27,533%	9,405%	38,021%
1992	144,8	11,035%	35,748%	7,800%	43,536%
1993	119,5	6,846%	9,712%	18,976%	20,265%
1994	142,8	30,341%	34,611%	4,327%	38,038%
1995	175	12,549%	18,167%	20,231%	42,942%
1996	133,3	15,657%	20,315%	10,137%	12,411%
1997	132,5	1,923%	7,967%	20,919%	1,721%
1998	98,7	36,032%	6,704%	11,381%	13,373%
1999	93,2	6,515%	4,735%	45,606%	6,472%
2000	101,6	4,369%	1,572%	15,780%	14,939%
2001	87,4	19,739%	39,731%	0,340%	24,361%
2002	88,3	5,832%	4,746%	17,570%	55,704%
2003	95,8	1,522%	6,483%	0,673%	15,193%
2004	145,3	31,039%	21,208%	21,782%	35,871%
2005	173,9	14,547%	23,920%	29,818%	31,506%
2006	275,3	42,624%	43,568%	36,949%	63,820%
MAPE 1987-2006		13,284%	19,928%	17,083%	29,910%
Desvío Estándar APE		10,657%	12,463%	12,527%	15,991%

Fuente: Brook Hunt & Associates. Datos corresponden a proyecciones realizadas en Diciembre, salvo para los años 1990, 1991 y 1994, donde las proyecciones fueron realizadas en Septiembre.

En el conjunto de 20 pronósticos a 1 año, el modelo híbrido entregó mejores resultados que el analista en 15 oportunidades, consiguiendo disminuir el MAPE en 33 % y disminuir el desvío estándar del APE en 14 %.

En proyecciones a 2 años (19 evaluaciones), el modelo híbrido entregó mejores resultados que el analista en 16 oportunidades, consiguiendo disminuir el MAPE en 43 % y el desvío estándar del APE en 22 %.

En cuanto al pronóstico en situaciones con grandes cambios de precios,

entre 1987 y 2006, en 10 oportunidades el cambio del precio respecto al año anterior superó el 15 %, entregando el modelo híbrido los mejores pronósticos, en términos de MAPE, en 8 ocasiones.

Finalmente, en términos de RMSE, los modelos de Inteligencia Computacional entregaron menores errores con reducciones de 16 % en proyecciones a 1 año y de 40 % en proyecciones a 2 años.

Así los mejores modelos de pronóstico identificados generaron proyecciones que, en promedio y en desviación estándar del error, son más precisas que las proyecciones publicadas por Brook Hunt and Associates.

5. Conclusiones

En las evaluaciones realizadas, los modelos no lineales de pronóstico, basados en redes neuronales multilayer perceptron, superaron la exactitud de los modelos lineales de series de tiempo más comúnmente usados. De esta manera, las redes neuronales demostraron ser capaces de aprovechar mejor la información contenida en los precios históricos considerados en el estudio.

El modelo con mejores pronósticos a un año, para el periodo 1977 - 2006, fue un modelo híbrido que, utilizando conjuntamente modelos AR2MA1 y redes neuronales entrenadas con residuos, promedió un error porcentual absoluto (MAPE) de 13,284 %. Esta cifra significa una reducción del error, en términos relativos, de 5,6 % respecto al mejor modelo ARIMA, y de 13 % respecto a la caminata aleatoria.

Dicho modelo híbrido también mostró un mejor desempeño que ARIMA en situaciones con grandes cambios de precios (variaciones de precios superiores a 15 % de un año a otro), generando proyecciones con menor MAPE en 8 de 14 eventos ocurridos entre 1977 y 2006.

La ventaja de los modelos basados en redes neuronales por sobre los modelos ARIMA se acrecienta en pronósticos a 2 y más periodos, llegando a una reducción del MAPE del orden de 30 % en términos relativos.

Aún cuando los mejores modelos basados en redes neuronales fueron seleccionados por su menor MAPE y no por su menor RMSE, estos también consiguieron una reducción del RMSE respecto a los mejores modelos de series de tiempo, aunque de menor tamaño (entre 1 % y 23 %).

Así como la estimación con modelos GARCH permitió mejorar los resultados en los modelos ARIMA, la inclusión de variables derivadas del precio, en especial su desviación estándar, también mejoró las proyecciones de las RN.

En cuanto a la arquitectura de las redes neuronales más exitosas, éstas presentaron 3 o 6 neuronas alimentadas por rezagos del precio en la capa de entrada, y, en proyecciones a 3 y más años, una neurona adicional alimentada

con la desviación estándar del precio o su última variación. En la capa oculta, las redes neuronales más exitosas mostraron un diverso número de neuronas (1, 2, 3, 5, 6).

Se observa que, con un modelamiento no lineal, los últimos 6 rezagos del precio aportan información útil para la generación de pronósticos, hecho que no se aprecia en los modelos ARIMA, donde modelos de ese orden fueron descartados por tener coeficientes no significativos.

La comparación de los pronósticos publicados por expertos en el mercado del cobre versus los resultados conseguidos con modelos de Inteligencia Computacional reveló la mayor precisión de estos últimos con reducciones del MAPE de 33 % y 43 % en proyecciones a 1 y 2 años respectivamente.

El mejor modelo híbrido también mostró un mejor desempeño que las proyecciones de expertos en situaciones con grandes cambios de precios (variaciones de precios superiores a 15 % de un año a otro), generando proyecciones con menor MAPE en 8 de 10 eventos ocurridos entre 1987 y 2006.

La mayor exactitud de los pronósticos conseguidos utilizando redes neuronales va acompañada por una mayor complejidad en el diseño. En este estudio se optó por sensibilizar sólo algunos parámetros del diseño de una red y, aún así, el número de cálculos y el tiempo requerido fue muy superior al de los modelos de series de tiempo (horas versus minutos).

Si bien las redes neuronales superaron a los modelos de series de tiempo, los resultados conseguidos con la aplicación de modelos híbridos muestran que, mediante el trabajo conjunto con modelos de series de tiempo y redes neuronales, es posible conseguir mejores pronósticos para el precio del cobre.

Las evidencias entregadas avalan a las redes neuronales como una herramienta atractiva para el pronóstico del precio del cobre, animando el desarrollo de futuros estudios que aporten más antecedentes sobre su potencial y motivando la investigación de metodologías de diseño que posibiliten el aprovechamiento de dicha capacidad de generalización.

6. Trabajos Futuros

Continuando con la evaluación del potencial de las redes neuronales como herramienta de pronóstico para el precio del cobre, podría evaluarse el uso de otras estrategias de aprendizaje: otros algoritmos, aplicación de momentum y pruning. Así mismo, también resulta interesante probar el efecto de una definición aleatoria de los conjuntos de entrenamiento y testeo en la capacidad de generalización.

A nivel de la arquitectura de las redes, se podría explorar el uso de n neuronas en la capa de salida, siendo n el número de periodos hacia adelante

que se desea pronosticar.

En el área de los modelos híbridos, sería un aporte contar con una evaluación de su desempeño en pronósticos a 2 y más periodos.

Otro trabajo interesante para la entrega de evidencias sobre el atractivo de las redes neuronales consistiría en el entrenamiento de redes incluyendo variables adicionales al precio y sus derivaciones, como por ejemplo: stocks en semanas de consumo, indicadores de actividad económica y expectativas.

En la línea del diseño de redes neuronales, futuros estudios deberían explorar algunas de las heurísticas propuestas en la literatura [8], evaluándolas y proponiendo variantes que aseguren la construcción de redes con elevada capacidad de generalización.

Una mayor comprensión del funcionamiento de las redes también puede facilitar su diseño. La extracción de reglas a partir de su funcionamiento, la sensibilización de los parámetros de la red y de las variables de entrada, junto con el análisis detallado de los residuos son áreas de interés para futuros trabajos.

Finalmente, los resultados conseguidos en este trabajo pueden ser un referente para la evaluación del potencial de otras herramientas de pronóstico, como las máquinas de soporte vectorial (support vector regression); ver por ejemplo [11].

Agradecimientos: Los autores les agradecen al Instituto Científico Milenio "Sistemas Complejos de Ingeniería" P04-066-F (www.sistemasdeingenieria.cl), por el apoyo brindado en la elaboración y financiamiento de este estudio.

Referencias

[1] Aburto, L., Weber, R. (2007a): Improved Supply Chain Management based on Hybrid Demand Forecasts. Applied Soft Computing 7, No. 1, 136-144

[2] Aburto, L., Weber, R. (2007b): A Sequential Hybrid Forecasting System for Demand Prediction. In: Petra Perner (Ed.): Machine Learning and Data Mining in Pattern Recognition. LNAI 4571, Springer Verlag, Berlin, Heidelberg, 518-532

[3] Ceballos, J. y Tilton, J. (2005). Análisis del Fondo de Compensación del Cobre de Chile. En: LAGOS, G (Edit.). Minería y Desarrollo. Foro en Economía de Minerales, vol. III.

[4] Ciudad, J. (2005). Determinantes del Precio Spot del Cobre en las Bolsas de Metales. En: COCHILCO (Edits). Comisión Chilena del Cobre. 145-187.

[5] Clements, M. Y Hendry D. (2001). Forecasting Non-Stationary Economic Time Series. The MIT Press.

[6] Comision Chilena Del Cobre. (2006). [en línea]Anuario: Estadísticas del Cobre y Otros Minerales 1986-2005. http://www.cochilco.cl/anm/articlefiles/456-ANUARIO2005-PDF %2811-AGO-06 %29.pdf .

[7] Comision Chilena Del Cobre. [en línea] Precio del Cobre. http://www.cochilco.cl .

[8] Crone, S. (2005). Stepwise Selection of Artificial Neural Network Models for Time Series Prediction. Journal of Intelligent Systems 14 (2-3). 99-122.

[9] Enders, W. (2004). Applied Econometric Time Series. 2ªed., John Wiley & Sons, Inc.

[10] Engel, E. y Valdes, R. (2002). Prediciendo el Precio del Cobre: ¿Más Allá del Camino Aleatorio?. En: MELLER, P. (Edit.). Dilemas y Debates en Torno al Cobre. Santiago, Dolmen Ediciones. 269-290.

[11] Guajardo, J., Weber, R., Miranda, J. (2006): A Forecasting Methodology Using Support Vector Regression and Dynamic Feature Selection. Journal of Information & Knowledge Management 5, No. 4, 329-335

[12] Hanke, J. y Reitsch, A. (1995). Estadística Para Negocios. Irwin Profesional Publishing.

[13] Hilera, J. y Martinez, V. (1995). Redes neuronales Artificiales. Fundamentos, Modelos y Aplicaciones. Madrid, RA-MA, Addison-Wesley Iberoamericana, S.A.

[14] International Monetary Fund. 2006. World Economic Outlook September (2006), Financial Systems and Economic Cycles. Washington, D.C., USA..

[15] Isasi, P. y Galvan, I. (2004). Redes de Neuronas Artificiales. Un Enfoque Práctico. Madrid, Pearson Educación, S.A.

[16] Kaastra, I. y Boyd, M. (1996). Designing a Neural Network for Forecasting Financial and Economic Time Series. Neurocomputing 10(3): 215-236.

[17] Maddala, G. S. (1996). Introducción a la Econometría. 2ªed., Prentice-Hall Hispanoamericana, S. A.

[18] Marshall, I. y Silva, E. (1998). Fluctuaciones del Precio del Cobre. Informe Macroeconómico para la Empresa. Nº35: 38-60. Instituto de Economía, Pontificia Universidad Católica de Chile.

[19] Meller, P. (2002). El Cobre Chileno y la Política Minera. En: MELLER, P. (Edit.). Dilemas y Debates en Torno al Cobre. Santiago, Dolmen Ediciones. 11-77.

[20] Nau, R. (2006). What's the Bottom Line? How to Compare Models. [en línea]. http://www.duke.edu/ rnau/compare.htm .

[21] Phillips, S. y Swiston, A. (2002). Forecasting Copper Prices in the Chilean context. Chile: Selected Issues. IMF Country Report No.02/163. International Monetary Fund.

[22] Pindyck, R. S. y Rubinfeld D. L. (2001). Econometría: Modelos y Pronósticos. 4ªed. México, D.F., McGraw-Hill/Interamericana Editores, S. A.

[23] Reed, R. D. y Marks, R. J. II. (1999). Neural Smithing. The MIT Press.

[24] Romaguera, P. (1991). Las Fluctuaciones del Precio del Cobre y su Impacto en la Economía Chilena. CIEPLAN. Notas técnicas Nº143.

[25] Silipo, R. (2003). Neural Netwroks. En: BERTHOLD, M. y HAND, D. (Edits.). Intelligent Data Analysis. Nueva York, Springer-Verlag New York, Inc. pp: 269-320.

[26] Tang, Z. y Fishwick, A. (1993). Feed-Forward Neural Nets as Models for Time Series Forecasting. ORSA Journal of Computing, 5(4): 374-386.

[27] Tashman, L. (2000). Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review. International Journal of Forecasting (16): 437-450.

[28] U.S. Department Of Labor, Bureau of Labor Statistics. Producer Price Index. Commodity Data. All Commodities Not Seasonally Adjusted. http://data.bls.gov/cgi-bin/surveymost . [Consulta: 5 de agosto de 2006].

[29] Ulloa, A. (2002). Tendencia y Volatilidad del Precio del Cobre. En: MELLER, P. (Edit.). Dilemas y Debates en Torno al Cobre. Santiago, Dolmen Ediciones. 291-337.

[30] Vial, J. (1988). An Econometric Study of the World Copper Market, Ph.D. Dissertation. University of Pennsylvania. CIEPLAN. Notas Técnicas Nº112.

[31] Vial, J. (2004). Modeling Commodity Markets in the Global Economy: Familiar Finding and New Strategies. The Earth Institute at Columbia University. Center on Globalization and Sustainable Development. Working Paper Nº18.

Segmentación de los Contribuyentes que Declaran IVA Aplicando Herramientas de Clustering

Sandra Lückeheide C.*
Juan D. Velásquez*
Lorena Cerda**

Resumen

En este trabajo se llevó a cabo una caracterización de contribuyentes que declaran IVA a través de la aplicación de algoritmos de clustering, con el fin de aportar nueva información de apoyo a la labor fiscalizadora del SII. La segmentación se realizó a partir de la información tributaria consignada por los contribuyentes, en sus declaraciones de IVA (Formulario F29) y en su Inicio de Actividades.

En primer lugar, aplicando un proceso de limpieza y reducción de los datos, se confeccionó un vector de características compuesto por la información del formulario F29, del documento de Inicio de Actividades, tales como el conjunto de Actividades Económicas (actecos) y la comuna. Sobre este vector de características, se aplicaron los algoritmos de clustering Self Organizing Feature Map (SOFM) y K-means. Comparando los resultados de distintas aplicaciones de estos dos métodos, se obtuvo el vector de características final y la segmentación de los contribuyentes. En ambos casos, SOFM y K-means, los resultados de la segmentación son comparables, lo cual valida el modelo de comportamiento del contribuyente desarrollado.

A partir de la segmentación propuesta, el organismo fiscalizador mejora su labor, haciendo más certero el proceso de validación de la información que declara el contribuyente.

Palabras Clave: Data Mining, Segmentación, Clustering, Clasificación.

*Departamento de Ingeniería Industrial, Universidad de Chile
**Servicio de Impuestos Internos de Chile

1. Introducción

El Servicio de Impuestos Internos (SII) es una institución del Estado, cuya labor es la administración tributaria, siendo el principal ente fiscalizador tributario. El SII es responsable de administrar el sistema de tributos internos, facilitar y fiscalizar el cumplimiento tributario, propiciar la reducción de costos de cumplimiento y potenciar la modernización del Estado; lo anterior en pos de fortalecer el nivel de cumplimiento tributario y del desarrollo económico de Chile y de su gente.

De acuerdo a la Ley, las funciones del SII son la "aplicación y fiscalización de todos los impuestos internos actualmente establecidos o que se establecieren, fiscales o de otro carácter en que tenga interés el Fisco y cuyo control no esté especialmente encomendado por la ley a una autoridad diferente".

El SII usa actualmente diversos métodos de fiscalización. Algunos de ellos se basan en lo que el contribuyente declara en algún determinado código de un formulario, lo que conlleva a que algunos contribuyentes puedan ser fiscalizados más de una vez, al ser seleccionados en distintos tipos de fiscalizaciones. Esto resulta ineficiente para el SII y molesto para el contribuyente. Otro método se basa en lo que se esperaría que el contribuyente declare, según sus actividades económicas declaradas en su Inicio de Actividades. Este último, se determina en base a la experiencia de los fiscalizadores, es decir, de forma cualitativa y subjetiva.

Una forma alternativa factible de enfrentar este problema, es la realización de una caracterización de los contribuyentes que declaran IVA, a partir de un agrupamiento basado en su información tributaria, contenida en los formularios de Declaración Mensual y Pago Simultáneo de Impuestos (IVA, Formulario 29) y en su Inicio de Actividades. De esta forma, se buscan los mejores grupos de equivalencia de comportamiento económico, determinado por la información tributaria declarada por los contribuyentes en el formulario F29, e información cualitativa declarada en el Inicio de Actividades, tales como el conjunto de Actividades Económicas (actecos).

La segmentación de contribuyentes, constituye un apoyo a la labor fiscalizadora del SII, a través del cual se podrá analizar el comportamiento de contribuyentes con características similares, y no a la gran masa, compuesta por personas con comportamientos y actividades muy diversas. De esta forma, se puede identificar las características principales que definen a cada grupo, para luego jerarquizarlos y priorizarlos, para una fiscalización más eficiente.

2. Trabajos Previos

El rápido avance de la tecnología, la gran cantidad de datos actualmente disponible y el bajo costo relativo su almacenamiento, han incentivado la creación y el uso de distintas técnicas y algoritmos que permiten procesar los datos, extrayendo conocimientos y patrones ocultos, que de otra forma no se podrían obtener. Entre estas técnicas se encuentran las herramientas de clustering, que permiten agrupar objetos similares (y separar los objetos disímiles). Estas herramientas son actualmente muy útiles en la investigación del comportamiento humano, para el apoyo de áreas como el marketing, o en la toma de decisiones importantes.

2.1. Clustering

Las herramientas de clustering son muy populares en la extracción de patrones de conjuntos de datos, particularmente en el análisis de comportamiento humano. Esto, debido a que la formación de grupos de personas con características comunes es una tendencia natural: comunidades sociales (por ejemplo civilizaciones, países, cuyas características comunes son el idioma, raza, aspectos culturales), y dentro de estas, se forman subgrupos, por ejemplo basados en antecedentes socio-económicos. Específicamente, el análisis de cluster es muy útil en marketing, dado que las compañías buscan crear el producto preciso para un grupo específico de consumidores [17].

El análisis de clusters o clustering, también llamado segmentación de data, tiene una variedad de objetivos, todos ellos relacionados con agrupar o segmentar una colección de objetos en subconjuntos o "clusters", tal que aquellos objetos dentro de cada cluster están más cercanamente relacionados que los asignados a clusters diferentes [7]. Un objeto puede ser descrito por un conjunto de medidas, o por su relación con otros objetos. Adicionalmente, el objetivo puede ser ordenar los clusters en una jerarquía natural. Esto involucra agrupar los clusters sucesivamente, de modo que en cada nivel de la jerarquía, los clusters en un mismo grupo son más similares entre ellos, que aquellos en diferentes grupos.

Centro de todos los objetivos del clustering, es la noción de grado de similitud (o diferencia) entre los objetos individuales a ser clusterizados, y por ello es fundamental para todas las técnicas de clustering, la elección de la medida de distancia o similitud entre dos objetos. Un método de clustering intenta agrupar los objetos basados en la definición de similitud que se le provee. La situación es algo parecida a la especificación de una función de pérdida o costo, en problemas de predicción (aprendizaje supervisado). El costo asociado con

una predicción inexacta depende de consideraciones externas a la data.

Sea Ω un conjunto de m vectores $\omega_i \in \Re^n$, con $i = 1, \ldots, m$. El objetivo es particionar Ω en K grupos, donde C_j es el j-ésimo cluster. Luego, $\omega_i \in C_j$ significa que ω_i es más parecido a los elementos dentro del cluster C_j, con $j = 1, \ldots, K$ que a los elementos pertenecientes a cualquier otro cluster [17].

El clustering requiere una medida de similitud, $\zeta(w_p, w_q)$ para comparar dos vectores de Ω. La forma de determinar el número de clusters, depende del método usado.

2.1.1. Vector de Características

El vector de características (feature vector) es el conjunto de atributos (o variables) seleccionados para representar a cada objeto del conjunto de datos, luego de haberlo preprocesado, limpiado y transformado, y sobre el cual se aplican los algoritmos de Data Mining.

La elección de este vector para la aplicación de técnicas de clustering, influye directamente en los resultados del análisis, por ello es un aspecto muy importante en Data Mining, pues los resultados dependen en gran medida, de las variables consideradas en el estudio [22].

Para la aplicación de toda herramienta de data mining, se requiere generar un vector de características, compuesto por un conjunto de variables que representan las características intrínsecas del fenómeno en estudio y que luego es usado como entrada para del algoritmo de clustering. Para ello, en una primera etapa, se debe realizar una selección, limpieza y preprocesamiento del conjunto de datos y variables sobre las cuales se lleva a cabo el estudio. Es decir, dentro del conjunto de datos, se deben tratar los fuera de rango y/o inconsistentes, se transforman, normalizando las variables y los vectores, y cuando se requiere, se realiza una reducción de las variables transformadas. La reducción de dimensiones, consiste en la selección de atributos (o feature selection), es decir se selecciona el conjunto mínimo de atributos, tal que la distribución de probabilidad de las diferentes clases, dados los valores de esos atributos, sea lo más parecida posible a la distribución original, considerando los valores de todos los atributos. Existen diversos métodos de reducción de dimensiones, entre estos el Análisis de Componentes Principales [14]. Luego de esta etapa, se ha obtenido el vector de características final, al que se le aplicará los algoritmos de Data Mining, donde se comprueba si el vector seleccionado es el indicado. Por ello, esta parte puede ser iterativa, pues se debe experimentar hasta dar con el mejor vector de características.

Se debe tener en cuenta que, al usar distintas técnicas en cada uno de los pasos mencionados, se puede llegar a distintos resultados, por lo que se debe hacer una cuidadosa elección de cada técnica a aplicar.

2.2. Self Organizing Feature Maps

Self-Organizing Maps (SOFM) es uno de los modelos más populares de Redes Neuronales. Elabora una cuantización del espacio formado por los datos de entrenamiento, y simultáneamente lleva a cabo una proyección con preservación topológica en una grilla regular de baja dimensión [18].

Una red de Kohonen, o SOFM (Self-Organizing Map) es una RNA no supervisada, competitiva, distribuida de forma regular en una grilla de, normalmente, dos dimensiones, cuyo fin es descubrir la estructura subyacente de los datos introducidos en ella. A lo largo del entrenamiento de la red, los vectores de características son introducidos en cada neurona y se comparan con el vector de peso característico de cada neurona. La neurona que presenta menor diferencia entre su vector de peso y el vector de datos es la neurona ganadora (o BMU) y ella y sus vecinas verán modificados sus vectores de pesos.

En las SOFM de dos dimensiones, se pueden distinguir dos tipos de rejillas:

- Rejilla hexagonal: en ella cada neurona tiene seis vecinos (excepto los extremos).
- Rejilla rectangular: cada neurona tiene cuatro vecinos.

Cada neurona de la red tiene asociado un vector de pesos (o prototipo) de la misma dimensión que los datos de entrada. Éste sería el espacio de entrada de la red, mientras que el espacio de salida sería la posición en el mapa de cada neurona.

Las neuronas mantienen con sus vecinas relaciones de vecindad, las cuales son claves para conformar el mapa durante la etapa de entrenamiento.

En cada paso se introduce un vector de datos en cada neurona y se calcula la "similitud" entre éste y el vector de peso de cada neurona.

$$\|X_j - m_{BMU}\| = \min_j \{\|X_i - m_j\|\} \tag{1}$$

$$m_k \leftarrow m_k + \alpha(X_i - M_k) \tag{2}$$

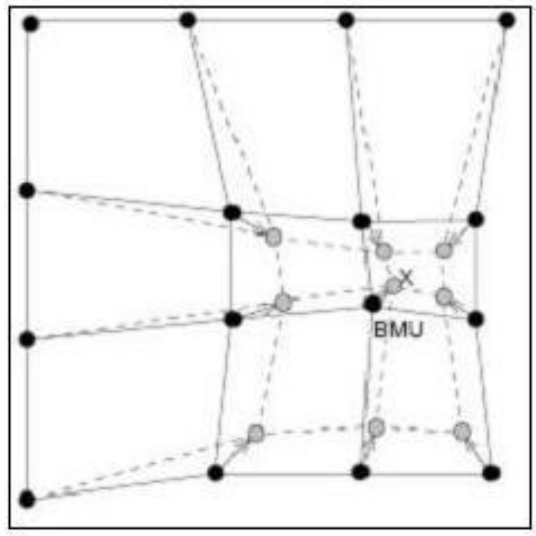
La neurona más parecida al vector de entrada, es la ganadora (o BMU, Best-Matching Unit, Unidad con mejor ajuste). Para medir la similaridad se utiliza usualmente la distancia euclideana. Tras ello, los vectores de pesos de la BMU y de sus vecinos son actualizados, de tal forma que se acercan al vector de entrada.

SOFM tiene propiedades tanto de algoritmos de cuantización de vectores, como de proyección de vectores, lo que permite hacer una reducción del conjunto de datos original, manteniendo la representatividad, además de hacer análisis posteriores como clustering o visualización.

Además de los beneficios computacionales ofrecidos por la cuantización de vectores, las principales ventajas del SOFM son [18]: a) su robustez, dado que todos los prototipos son afectados por todos los datos, b) su sintonización local, pues se trabaja en la vecindad de cada unidad del mapa, se sintoniza localmente con la densidad de los datos y c) su facilidad de visualización.

Entre sus desventajas, se destacan: los efectos de borde, dado que las definición de las vecindades no es simétrica en los bordes del mapa, y la contracción del rango de valores de las variables, en que se dejan algunos valores afuera que bajo algún punto de vista podrían ser interesantes.

Figura 1: Actualización del BMU y sus vecinos, hacia X



Fuente: [18]

2.2.1. K-Means

El algoritmo K-means, es uno de los métodos de clustering iterativos más usados. Es destinado a situaciones en las cuales todas las variables son de tipo cuantitativo, y la distancia euclideana es generalmente escogida como medida de disimilitud.

La dispersión intra-puntos puede escribirse como:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{c(i)=K}\sum_{C(i')=K}\|X_i - X_{i'}\|^2 = \sum_{K=1}\sum_{C(i)=K}\|X_i - \bar{X}_k\|^2 \quad (3)$$

Donde $\bar{X}_k = (\bar{X}_{1k}, ..., \bar{X}_{pk})$ es el vector promedio sociado al cluster K.

Luego, este criterio es minimizado asignando las N observaciones de los K clusters de tal forma que dentro de cada cluster, la disimilitud promedio de

las observaciones desde la media de los clusters definida por los puntos en ese cluster, es minimizada.

Algoritmo K-means:

1. Para una asignación de clusters dada C, la varianza de cluster total

$$\bar{X}_s = arg \min_m \sum_{i \epsilon S} \|X_i - m\| \tag{4}$$

es minimizada respecto a $m_1, ..., m_k$ dando las medias de los clusters asignados actualmente

$$C^* = \min_c \sum_{K=1}^{K} \sum_{C(i)=K} \|X_i - \bar{X}_k\|^2 \tag{5}$$

2. Dado un conjunto actual de medias m1,...,mk, (4) es minimizado asignando cada observación al cluster cuya media es la más cercana. Esto es,

$$C(i) = arg \min_{1 \leq k \leq K} \|X_i - m_k\|^2 \tag{6}$$

3. Los pasos 1 y 2 son iterados hasta que las asignaciones no cambian

Las principales ventajas de los algoritmos K-means son su simplicidad, sencillez, no es sensible al orden de los datos, y es basado en el sólido fundamento del análisis de varianzas [7]. Entre las desventajas, se cuentan la fuerte dependencia del resultado de la asignación inicial de los centroides, el óptimo encontrado es local y puede estar bastante lejos del global, la dificultad de una buena elección en el número de clusters a encontrar, es un proceso sensible a los datos fuera de rango, el algoritmo carece de escalabilidad, se limita a abarcar sólo atributos numéricos y los clusters resultantes pueden ser desequilibrados.

3. Segmentación de Contribuyentes

A continuación se describe la elaboración del vector de características, la aplicación de dos herramientas de clustering, Self Organizing Feature Maps y K-means, y la comparación de los resultados de ambos métodos.

La herramienta utilizada para usar los algoritmos es R[1] , un paquete Open Source estadístico y de Data Mining.

[1] www.r-project.org

3.1. Construyendo el vector de características

Inicialmente, los datos usados para la realización de este estudio, correspondieron a la información presentada en el año 2005, por los contribuyentes que declaran IVA (Impuesto al Valor Agregado), en el formulario F29 (Declaración Mensual y Pago Simultáneo de Impuestos), y en el formulario de Inicio de Actividad Económica. El número de contribuyentes considerados en un principio es de 597.082, y se tomaron en cuenta gran parte de códigos del formulario F29.

La declaración de IVA se hace mensualmente. Por lo tanto, para transformar los datos de mensual a anual, y comenzar a reducir de esta forma la dimensionalidad del vector de características, se decidió considerar estadísticos por cada código, como el número de no nulos en el año, la suma de los montos mensuales, el promedio y la desviación estándar (considerando sólo los meses declarados), para cada una de las variables seleccionadas.

Luego de consolidar la información, se hizo una selección y preprocesamiento de los datos. En esta etapa, se realizó la limpieza, eliminando los datos fuera de rango (outliers) y aquellos considerados inconsistentes, excluyendo de esta forma aproximadamente el 6 % de los registros iniciales. Después de la limpieza, se llevó a cabo la reducción de los datos y selección de las variables.

Dada la gran dimensionalidad del problema, tanto en número de registros (contribuyentes) como en cantidad de dimensiones, se hizo indispensable la reducción de éstas, para posibilitar y hacer más eficiente el análisis. Debido a que muchos de los códigos del fromulario de declaración de IVA, en un gran porcentaje de registros, se encuentran en blanco, proporcionando así muy poca información relevante en la discriminación de grupos, se optó por considerar sólo los códigos en los que al menos un 10 % de los contribuyentes tienen valores no nulos. Por lo tanto, teniendo la base de datos limpia, a estos códigos (o más bien a los estadísticos de estos códigos) se les aplicó un Análisis de Componentes Principales (ACP), y se procedió a seleccionar las variables (códigos del formulario) que mejor representen la variabilidad de los datos.

Considerando que las variables pueden tener distintas escalas, que puede conllevar a que aquellas con un mayor rango de valores le quiten importancia a otras con un menor rango, todas las variables consideradas fueron escaladas según la normalización "Min-Max", en el rango [0,1], según la fórmula y'= ((y-min) / (max - min)) (max'- min') + min'. Además, se tomó en cuenta que al normalizar las columnas, se pierde la relación original entre los componentes de cada vector o fila. Por ello, se llevó a cabo la normalización de los vectores (norma 1), es decir, se calculó el módulo de cada vector, y cada una de sus componentes se dividió por este valor. Para ello fue necesario extraer aquellos contribuyentes que solo tenían valores nulos en todas las variables seleccionadas (es decir, vectores de norma 0), que representan alrededor del

12 %.

La medida de distancia seleccionada para la aplicación de los algoritmos de clustering en el vector de características generado, fue la Euclidiana, que por ser la más comúnmente utilizada, viene por defecto en la mayoría de los algoritmos en R (como en gran parte de las herramientas de data mining).

Para la incorporación de la actividad económica de los contribuyentes y su comuna, se decidió cuantizar esta información, en base a lo que cada actividad y cada comuna, en promedio, genera en impuestos.

1. Primer Experimento:
   Se usaron las 10 primeras componentes principales.

   Se extrajo una muestra aleatoria de 200 mil contribuyentes, y sus respectivos valores en cada una de las variables del vector de características (las 10 componentes principales que explican el mayor porcentaje de la varianza). Se quitó de esta muestra aquellos contribuyentes cuyo vector tuviese norma 0 (es decir todas las variables con valor nulo), que corresponden a 24.599 datos, por lo tanto la muestra empleada en el análisis tiene un tamaño de 175.401 contribuyentes, que tienen al menos una variable con valor positivo.

   Luego de normalizar cada variable, según la normalización Min-Max y la normalización de los vectores, se aplicó el algoritmo K-means, con 8 clusters como condición inicial, y 10 semillas iniciales.

   Como resultado, se obtuvo 8 clusters, con los siguientes tamaños:

Cuadro 1: 8 clusters y sus tamaños

Cluster	Tamaño		Cluster	Tamaño
1	61 (0,03 %)		5	133 (0,076 %)
2	170.848 (97,4 %)		6	48(0,027 %)
3	2.667 (1,52 %)		7	111 (0,063 %)
4	563(0,32 %)		8	970 (0,553 %)

Fuente: Elaboración propia

   Se puede observar que mediante este método, se creó un gran cluster, que abarca más del 97 % de la muestra, y los clusters restantes contienen el otro 3 % (con diferencias considerables de tamaño entre algunos de ellos también).

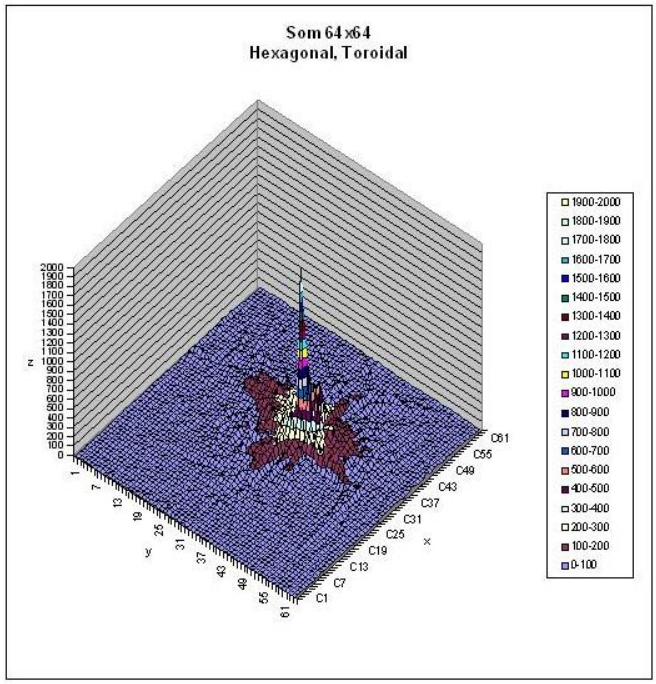
   Se probó con distintos números de clusters, entre 3 y 20 clusters, obteniéndose los mismos resultados anteriores: un solo grupo contiene más del 90 % de los datos.

Para verificar si el problema era el método empleado, se probó aplicando el algoritmo SOFM, al mismo conjunto de datos.

Utilizando el paquete "som" de R , se aplica el algoritmo SOFM a la misma muestra anterior, de 175.401 contribuyentes, y las primeras 10 Componentes Principales. Como se observa en la Figura 2, mediante este análisis se obtuvo, al igual que con el Kmeans, un gran cluster que concentra la mayoría de los datos.

Considerando la posibilidad de que se estén incluyendo observaciones que produzcan ruido (que no aportan información, sino sólo distorsionan los resultados) en el estudio, se tomaron sólo los datos pertenecientes al "gran" cluster, y sobre esta nueva muestra (de un tamaño de 148.681 observaciones), se aplicaron nuevamente el método Kmeans, para evaluar si existe algún cambio en los resultados.

Figura 2: Mapa SOFM 64x64, topología hexagonal y cerrada (toroidal), muestra de 175.401 contribuyentes, 10 componentes principales (eje z y colores corresponden a número de observaciones por celda).



Fuente: Elaboración propia

Se aplicó el algoritmo Kmeans a esta nueva muestra de 148.681 observaciones, con 8 clusters como condición inicial, y 15 semillas iniciales.

Como resultado, se obtuvo 8 clusters, con los tamaños que se muestran en el Cuadro 2.

Cuadro 2: 8 clusters y sus tamaños (Análisis incluyendo sólo observaciones pertenecientes al cluster gigante)

Cluster	Tamaño
1	258
2	1.783
3	305
4	1.028
5	135.315
6	189
7	6.567
8	491

Fuente: Elaboración propia

En el Cuadro 2 se puede ver que el resultado fue similar a los anteriores: el cluster 5 contiene 135.315 observaciones, que corresponde a más del 90 % de la muestra. Se probó con distintos números de clusters, obteniendo resultados similares.

En el mundo real, los contribuyentes no se comportan igual en absoluto. A pesar de ello, los resultados de los experimentos anteriores, al entregar un gran cluster que abarca la mayoría de la muestra, muestran lo contrario. Esto podía deberse a una mala elección del vector de características, que no definía comportamientos distintos entre contribuyentes. Por ello, se experimentó luego con un nuevo vector de características. En esta ocasión, en vez de usar las componentes principales como variables, se usaron las variables originales que tengan mayor peso en las componentes principales que explicaban mayor varianza.

2. Segundo Experimento:

Se utilizaron como vector de características, los mismos 14 códigos del formulario 29, usados para el análisis de componentes principales (aquellos que tienen un porcentaje de contribuyentes con valor positivo mayor a 10 %), además del número de declaraciones del formulario en el año 2005, pero en este casom las variables originales que tenían mayor importancia en las CP. Para cada código y contribuyente, se toma en cuenta la suma total del año 2005 (en pesos) y el número de meses en que el monto es mayor que cero.

Se realizó un análisis de componentes principales, con un total de 29 variables (la cantidad de declaraciones en el año y, para los 14 códigos, suma y número de no nulos).

La importancia por código resultó bastante similar al primer análisis realizado. Pero al seleccionar las variables puras en vez de las componentes principales, se puso atención a las correlaciones entre las variables, para no trabajar con variables muy correlacionadas que distorsionen el análisis.

Luego de analizar las correlaciones, las 16 las variables seleccionadas para el análisis fueron: c142s, c142nn, c111s, c538s, c538nn, c511s, c511nn, c525s, c525nn, c504s, c504nn, c48s, c48nn, c151s, c151nn y la cantidad de declaraciones.

Del conjunto de datos inicial, se consideraron aquellos datos en que al menos una variable era no nula, que correspondía a una muestra de 173.935 contribuyentes. Se normalizó las variables mediante la normalización "Min-max", y luego se normalizó los vectores, a módulo unitario.

Se aplicó el algoritmo Kmeans, imponiendo 8 clusters como condición inicial. Los tamaños de los grupos generados por este método, se observan a continuación:

Cuadro 3: Tamaño 8 clusters (16 Variables)

Cluster	Tamaño
1	21.905
2	16.689
3	13.745
4	11.429
5	9.667
6	50.972
7	30.360
8	19.168

Fuente: Elaboración propia

Tan sólo observando el tamaño de los clusters, se nota un cambio drástico respecto a los experimentos anteriores, en los que se consideraba como variables las componentes principales. En esta prueba, el cluster de menor tamaño contiene casi el 5.5 % de la muestra, y el de mayor tamaño el 29 %.

Al observar el Cuadro 4 correspondiente a los vectores de los centros de los clusters, llama la atención la gran importancia de las variables correspondientes al número de no nulos de cada código, no así del monto total declarado en los mismos.

Cuadro 4: Centros de los 8 clusters (16 variables)

Cluster	c142nn	c142s	c111s	c538nn	c538s	c511nn	c511s	c525nn
1	0.007	0.000	0.000	0.075	0.000	0.021	0.000	0.003
2	0.012	0.000	0.000	0.488	0.001	0.059	0.000	0.011
3	0.013	0.000	0.000	0.026	0.000	0.005	0.000	0.001
4	0.069	0.001	0.004	0.345	0.013	0.221	0.001	0.075
5	0.546	0.003	0.001	0.118	0.000	0.011	0.000	0.003
6	0.009	0.000	0.001	0.546	0.001	0.024	0.000	0.007
7	0.007	0.000	0.000	0.671	0.001	0.031	0.000	0.006
8	0.070	0.000	0.002	0.512	0.002	0.421	0.000	0.026
	c525s	c504nn	c504s	c151nn	c151s	c48nn	c48s	cant
1	0.000	0.661	0.002	0.023	0.000	0.004	0.000	0.673
2	0.000	0.477	0.002	0.099	0.001	0.005	0.000	0.632
3	0.000	0.077	0.000	0.560	0.006	0.008	0.000	0.668
4	0.002	0.104	0.005	0.388	0.024	0.466	0.011	0.482
5	0.000	0.032	0.000	0.330	0.009	0.074	0.001	0.583
6	0.000	0.065	0.000	0.546	0.001	0.003	0.000	0.562
7	0.000	0.042	0.000	0.023	0.000	0.002	0.000	0.642
8	0.000	0.073	0.000	0.399	0.002	0.008	0.000	0.518

Fuente: Elaboración propia

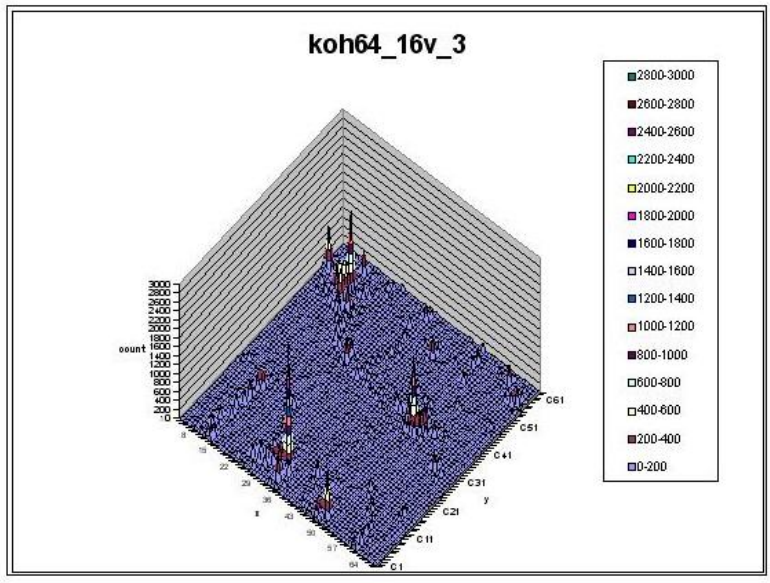
Para complementar el análisis hecho con el método Kmeans, se realizó un análisis de SOFM a la misma muestra de datos, y con las mismas variables. Utilizando el paquete "kohonen" de R, se aplica sobre esta muestra el método SOFM, con una grilla de 64x64 celdas, topología hexagonal y cerrada (toroidal), obteniéndose un mapa como el de la Figura 3.

El mapa generado por el SOFM, resultó bastante diferente a aquellos de los experimentos anteriores. En este caso, se distinguen ciertos agrupamientos visiblemente diferenciados y algunos bastante más concentrados que otras. Analizando las características de cada agrupamiento, se observó que estos se diferencian entre sí principalmente por las variables relacionadas con el número de "no nulos" (es decir cuantas veces declaró un valor positivo en el código, en el año).

De acuerdo a lo anterior, se pudo concluir que con el vector de características empleado en este experimento, ambos métodos (Kmeans y SOFM) agruparon contribuyentes basados principalmente en el número de valores no nulos para cada código, sin considerar el monto en pesos. Esto se debe a que gran parte de las observaciones tiene valores 0 o 12 en las variables "no nulos", y aquellos que tienen valores entre 1 y 11 se distribuyen uniformemente , en cambio para las variables relativas a sumas (montos), la mayoría se concentra en valores relativamente muy bajos y unos pocos en valores muy altos, por lo tanto en el análisis, luego

de la normalización de las variables y de los vectores, son los "no nulos" los que más pesan y le restan importancia a las sumas.

Figura 3: SOFM de 64x64, topología hexagonal y toroidal, muestra de datos con al menos 1 código no nulo, 16 variables (eje z y colores corresponden a número de observaciones por celda).



Fuente: Elaboración propia

Sin embargo, se consideró el hecho de que el monto declarado por el contribuyente en un determinado código, es más importante que el número de veces que declare un monto no nulo en el año . Para entender mejor el problema que puede generar la presencia de las variables "nn" en el análisis, se ejemplifica con el siguiente caso:

Suponiendo que se tiene un apicultor que produce miel (puede aplicarse a productores de cualquier otro producto) y que vende su producción anual en lotes, a grandes supermercados, los cuales realizan 4 pedidos en el año. A este contribuyente le corresponde declarar sus ventas en el código c502 (Facturas Emitidas), por lo que en la variable c502s tendrá un monto (en pesos) relativamente grande, y en la variable c502nn tendrá un valor de 4 (4 meses en que emitió facturas). Sin embargo este apicultor, todos los meses vende en su domicilio una reducida cantidad de miel a clientes de paso, a los que les entrega boleta por estas ventas. Estas ventas significan un porcentaje despreciable en comparación a lo que vende a supermercados, aún así declarará en el código c111 (Boletas), por lo que la variable c111s tendrá un valor pequeño en comparación al valor en la variable c502s y la variable c111nn tendrá un valor de 12

(dado que entregó boletas los 12 meses del año). Luego de normalizar las variables y vectores, la diferencia que existía entre el c502s y el c111s se torna despreciable y aquella entre la variable c502nn y c111nn se vuelve a su vez muy importante, y determinará la diferencia de comportamiento entre este contribuyente y los demás. Finalmente, se concluirá que este contribuyente se dedica principalmente a la venta directa, cuando en realidad ocurre todo lo contrario.

Por esta razón, se decidió sacar del análisis las variables relativas a la cantidad de "no nulos", dado el ruido que provocaban, así como también la cantidad de declaraciones en el año, dejando sólo las sumas (en pesos) para cada código, es decir 8 variables.

De esta forma, se selecciónó la suma del año 2005 de los siguientes códigos: Del recuadro Débitos y Ventas:

- c142 (Ventas y/o Servicios prestados Internos Exentos o No Gravados).
- c111 (Boletas)
- c538 (Total Débitos).

Del recuadro Créditos y Compras:

- c511 (IVA por documentos electrónicos recibidos)
- c525 (Facturas activo fijo)
- c504 (Remanente Crédito Fiscal mes anterior)

Y finalmente del recuadro Impuesto a la Renta:

- c48 (Retención Impuesto único a los Trabajadores)
- c151 (Retención de Impuesto con tasa de 10

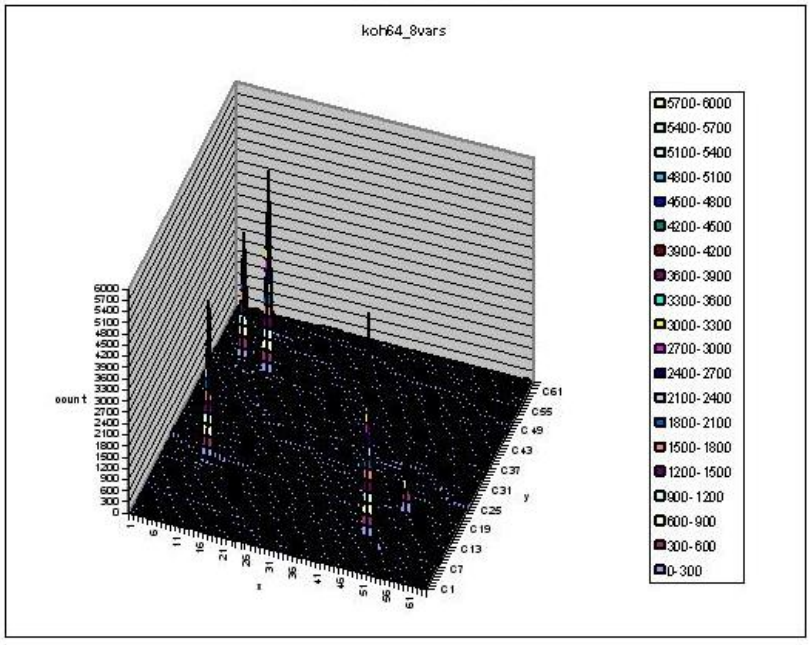
Se incluyó también la actividad económica y la comuna, pero nuevamente esto llevaba a la agrupación de la mayoría de los datos en un solo gran cluster, por lo que se decide no incorporarlas pues introdujeron más ruido que información relevante para la discriminación de grupos.

3.2. Aplicando el SOFM

Utilizando el paquete "kohonen" de R, se aplica el método SOFM, con una grilla de 64x64, de topología hexagonal y cerrada (toroidal), sobre una muestra (normalizada) de 100 mil contribuyentes.

En este caso, como se observa en el mapa generado por el SOFM, con el vector de características de 8 variables (Figura 4), se distinguen claramente

Figura 4: SOFM de 64x64, topología hexagonal y toroidal, 8 variables (eje z y colores corresponden a número de observaciones por celda).



Fuente: Elaboración propia

5 "peaks", correspondientes a celdas con una gran concentración de observaciones. En primera instancia, se consideraron estas celdas como centroides de los posibles clusters.

Para cada una de estas concentraciones, se analizaron las características tributarias de sus contribuyentes (sus declaraciones en el formulario de declaración de IVA), para determinar las similitudes existentes dentro de cada una, obteniéndose lo siguiente:

- Cluster 1: tiene una media alta en los códigos c504 (Remanente Crédito Fiscal mes anterior) y c537 (Total Créditos), y valores nulos en todos los otros códigos. En el código c91, que corresponde al total de impuesto a pagar, posee media = 0. Se puede decir que este grupo corresponde a contribuyentes que obtuvieron pérdidas en el año 2005 y se etiquetó como "Remanentes".

- Cluster 2: se caracteriza por tener montos positivos en los códigos c111 (Boletas), c538 (Total Débitos), c520 (Facturas recibidas del giro y Facturas de compra emitidas), c537 (Total Créditos), c62 (PPM 1ª Categoría), y consecuentemente en el c91 (Total a pagar). Este centroide puede estar constituido por contribuyentes que realizan actividades de

venta directa, dado que declaran en el código "Boletas" (consecuentemente en el código "Total Débitos"), y en el de "Facturas Recibidas" (consecuentemente en el código "Total Créditos"), y el código que corresponde al Pago Provisional Mensual (PPM) de 1ª Categoría. Por lo tanto, este grupo se etiquetó como "Ventas Directas".

- Cluster 3: todos los contribuyentes tienen un valor positivo en el código c142 (Ventas y/o Servicios prestados Internos Exentos, o No Gravados), y una gran parte tiene valores positivos en el c62. Luego, este grupo se llamó "Exentos".

- Cluster 4: en esta celda, los contribuyentes declaran principalmente en el código c151 (Retención de Impuesto con tasa de 10 % sobre las rentas), y en el resto de las variables poseen, en su mayoría, valores nulos. Por lo tanto, este grupo se denominó "Retenedores".

- Cluster 5: los contribuyentes de este centroide, se caracterizan por tener valores positivos en los códigos c502 (Facturas emitidas), c538 (Total Débitos), c520 (Facturas recibidas del giro y Facturas de compra emitidas), c537 (Total Créditos), c62 (PPM Neto Determinado) y c91 (Total a Pagar). Por lo tanto, está constituido por contribuyentes que realizan actividades de venta indirecta, dado que declaran en el código "Facturas Emitidas", por lo que se etiquetó como "Ventas Indirectas" o "Mayoristas".

Una vez definidos los clusters, se debió probar el clasificador. Dado que se había usado una muestra de 100 mil contribuyentes, de un total de 173.935 cuyas variables fueron inicialmente normalizadas, se seleccionaron aleatoriamente 30 mil contribuyentes del grupo que no fue considerado en el proceso de clustering, y mediante la función "map" del paquete "kohonen" de R, estas 30 mil observaciones fueron dispuestas en el mapa entrenado por las 100 mil originales.

Al colocar una nueva muestra de datos sobre el mapa entrenado inicialmente por la muestra de 100 mil datos, se generó un mapa muy similar. Las mismas celdas seleccionadas como centroides en el mapa original, en este caso también formaron "peaks" en el mapa, por su gran concentración de observaciones y se observó que los contribuyentes de cada una de estas celdas tenían características similares a los de la muestra original.

Luego de caracterizar o etiquetar cada centroide, y por ende cada cluster, se procedió a asignar todos los contribuyentes al cluster que más se le asimile, según la información contenida en él. La medida de distancia utilizada para encontrar el cluster más cercano a cada vector, fue la distancia Euclideana.

3.3. Aplicando el K-Means

Utilizando el paquete "Kmeans" de R, se aplicó el algoritmo Kmeans a la misma muestra de 100.000 contribuyentes, tomada de la muestra inicial de tamaño 173.935, con 5 clusters como condición inicial, y 20 semillas iniciales (es decir 20 pruebas con distintos centros de clusters iniciales, de las que se escoge la que entrega el mejor resultado), obteniéndose lo siguiente:

Analizando los vectores correspondientes a los centros de los clusters, se observa:

- Cluster 1: se encuentran valores altos en la variable c504s, y un poco más bajos en la variable c538s.
- Cluster 2: se encuentran valores altos en la variable c111s, y valores menores en las variable c538s y c151s.
- Cluster 3: predominan valores altos en la variable c142s y en menor medida en la variable c151s,
- Cluster 4: predomina la variable c151s y en menor medida la variable c538s.
- Cluster 5: predomina la variable c538s, y valores menores en la variable c151s.

Cuadro 5: Tamaño 5 clusters (8 Variables)

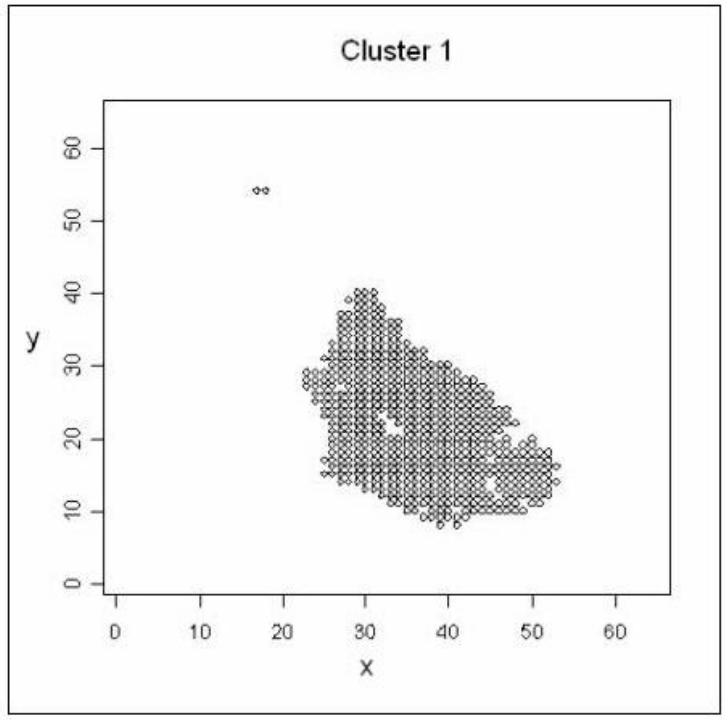
Cluster	Tamaño
1	15.583
2	32.797
3	3.405
4	27.662
5	20.533

Fuente: Elaboración propia

3.4. Comparación de Resultados

Los vectores de los centros de los clusters generados por el Kmeans, resultaron con características muy similares a aquellos de los centros del SOFM: valores relativamente altos en la variable c504s en el primero, valores altos de la variable c111s en el segundo (y consecuentemente también cierta presencia del código c538), presencia importante de la variable c142s en el tercer cluster, de la variable c151s en el cuarto y finalmente valores altos de la variable c538s en el quinto.

Luego, se confirmó gráficamente si los contribuyentes unidos por el método Kmeans, se encuentran unidos en el mapa generado por el SOFM. Para ello, se elaboraron gráficos, en los que se observa la ubicación en el mapa generado por el SOFM, de los contribuyentes de cada cluster formado por el Kmeans. A modo de ejemplo, en la Figura 5 se muestra el Cluster 1 ("Remanentes") generado por el método Kmeans, y su ubicación en el mapa del SOFM. Se puede ver que, a excepción de un par de celdas, todas las celdas que contienen contribuyentes del Cluster 1 del Kmeans, se encuentran juntas en el SOFM. Algo similar se puede concluir respecto a los demás clusters. Por lo tanto, se puede concluir que el agrupamiento estuvo bien hecho, al llegar a resultados similares, por caminos diferentes.



Fuente: Elaboración propia

4. Aplicación del Clasificador

Luego de comprobar que efectivamente existían diferencias en el total de tributación (de IVA) por grupo y de evaluar la calidad de la segmentación, se procedió a generar indicadores que permitan, dentro de cada grupo, encontrar aquellos contribuyentes cuyo comportamiento se aleje significativamente del

resto del grupo.

En general, se esperaría que para cualquier contribuyente, la suma de los montos en el código c538 (Total Débitos) sea superior a la suma de los montos en el código c537 (Total Créditos), es decir que los ingresos deben ser superiores a los egresos . En primer lugar, se creó el indicador "Débitos / Créditos". Dado que algunos tienen valor cero en Créditos, se debió transformar a "Débitos / (Créditos + 1)".

La razón "Débito/Crédito" (o Débito - Crédito <= 0) es usualmente utilizada como uno de los métodos de fiscalización. Sin embargo, este cálculo no tiene mucho sentido para ciertos tipos de contribuyentes, como los del cluster 3 ("Exentos"), que en su mayoría corresponden contribuyentes cuyas actividades no generan débito fiscal, o el cluster 1 ("Remanentes"), donde la media de la suma de Débitos es menor a la de Créditos. Pero este indicador sí tiene sentido principalmente para aquellos del cluster 2 ("Ventas Directas") y del 5 ("Ventas Indirectas"), en los que en el primer cuartil, este indicador ya tiene un valor mayor a 1 (es decir que en estos casos, la mayoría cumple que Débitos < Créditos). Por lo tanto, para los contribuyentes pertenecientes a estos clusters, se debe poner principal atención en aquellos que tienen un valor inferior a 1 (o un valor levemente mayor a 1) en el indicador.

Luego, a partir del indicador creado, se generó una medida del comportamiento de pago de cada uno de estos 2 grupos, que son aquellos que tienen más incentivo (dado que todas las transacciones que realizan se encuentran gravadas) y oportunidades de evadir. El cluster 3 ("Exentos") no es tan interesante de analizar en este sentido, pues dado que las actividades que realizan la mayor parte se sus contribuyentes se encuentran exentas, estos no tienen mayor incentivo a distorsionar los montos que declaran. El cluster 1 ("Remanentes"), puede ser sujeto a mayor estudio en trabajos posteriores, pues corresponde a contribuyentes en que los valores mas importantes corresponden al código c504 (Remanente Crédito Fiscal mes anterior), es decir en su mayoría son contribuyentes que declaran pérdidas, y por ello la mayor parte tiene valor nulo en el código c91 (Total a Pagar).

Por lo tanto, considerando sólo los clusters 2 ("Ventas Directas") y 5 ("Ventas Indirectas"), para cada uno de ellos se extrae el valor del primer cuartil en el conjunto de contribuyentes cuyo indicador es superior a uno. Estos valores se consideran como el valor mínimo esperado para la razón Débito/Crédito, que deberían tener todos los contribuyentes según el cluster al que pertenezcan.

Luego, para ambos clusters, se calculó el promedio de tributación (usando el código 91, Total a Pagar) de aquellos cuyo indicador se encontrara muy cercano a los valores calculados en la etapa anterior.

A continuación, para crear una medición del comportamiento de pago en cada cluster, para cada contribuyente cuya razón "Débito/Crédito" es menor al umbral calculado para el cluster (es decir primer cuartil), se calculó la

diferencia entre el valor declarado por éste en el código c91 y el promedio en el código c91 calculado en la etapa anterior, para el cluster correspondiente, que es considerado como el valor mínimo esperado a pagar. Finalmente, la suma de las diferencias entre el valor real y el esperado, calculadas para cada contribuyente de un determinado cluster, corresponde al indicador de comportamiento de pago del mismo. Mientras más alto es el valor de éste, más diferencia (negativa) hay entre lo que pagan los contribuyentes "bajos" (cuyo indicador Débito/Crédito es bajo).

De esta forma, además de generar una "alarma" en aquellos contribuyentes bajo el umbral determinado, se obtuvo la diferencia total para cada cluster considerado, entre lo que estos pagan y lo que se esperaría que paguen.

5. Conclusiones y Trabajos Futuros

A través de este trabajo, se realizó una caracterización de los contribuyentes que declaran IVA usando su información tributaria del año 2005. De esta forma, obtuvo información novedosa y potencialmente útil para el SII, en particular en el proceso de selección de contribuyentes a fiscalizar.

La elección del vector de características es fundamental, en este y en la mayoría de los trabajos de Data Mining, lo que quedó claramente demostrado, al obtener resultados absolutamente diferentes entre una y otra elección de vector de características, a veces incluso bajo pequeños cambios. Por lo tanto, la elección del vector determina en gran medida el resultado final del análisis. Luego de varios experimentos, se concluyó que el vector de características que mejor discrimina entre los contribuyentes, dada la calidad del clustering resultante, es aquel compuesto por los siguientes códigos, declarados en el Formulario de Declaración Mensual y Pago Simultáneo de Impuestos: c142 (Ventas y/o Servicios prestados Internos Exentos o No Gravados), c111 (Boletas), c538 (Total Débitos), c525 (Facturas Activo Fijo), c511 (IVA por documentos electrónicos recibidos), c504 (Remanente Crédito Fiscal mes anterior), c48 (Retención Impuesto único a los Trabajadores) y c151 (Retención de Impuesto con tasa del 10 %).

Usando el vector de características seleccionado, se agruparon los contribuyentes, utilizando los algoritmos K-means y SOFM. Se seleccionó este último, obteniéndose 5 grupos claramente diferenciados respecto a los montos declarados en diferentes códigos del formulario F29. Mediante un análisis estadístico, se verificó que estos grupos son significativamente diferentes respecto al Impuesto Total a pagar.

Para caracterizar el comportamiento de un contribuyente dentro de su grupo, se creó un indicador (razón entre Débitos y Créditos), que resulta útil

principalmente en 2 de los clusters encontrados (Cluster de "Ventas Directas" y el de "Ventas Indirectas" por tener estos grupos mayor incentivo y oportunidades de evadir impuestos. A partir de esto, se generó otro indicador, caracterizando el comportamiento de pago de los contribuyentes de cada grupo.

Quedó además demostrado que existen otras formas de agrupar a los contribuyentes, además de las que actualmente se conocen como el tamaño de la empresa o el rubro o sector al cual pertenezca. En este caso, la agrupación se hizo en base a los códigos que declaran (independiente de los montos declarados en ellos), obteniéndose un grupo caracterizado por generar pérdidas ("Retenedores"), otro grupo consistente en los que venden directamente al consumidor final ("Ventas Directas"), otro en que los contribuyentes realizan actividades exentas ("Exentos", entre los que se incluyen por ejemplo los centros médicos), otro compuesto por los contribuyentes intermediarios ("Ventas Indirectas") y el grupo donde se reúnen los contribuyentes que emplean y retienen ("Retenedores").

La metodología generada en este trabajo para agrupar contribuyentes de comportamiento similar, resulta bastante confiable y perpetuable en el tiempo. Esto, debido a que ella no es tan sensible a los montos declarados en cada código, sino más bien al código en sí mismo, es decir, si este es usado o no por el contribuyente.

A partir de estas conclusiones, se proponen las siguientes recomendaciones:

Resulta interesante realizar un análisis más profundo del cluster de "Remanentes" (cluster 1), correspondiente a aquellos contribuyentes que declaran pérdidas, dado que se comprobó que es un grupo bastante importante en cuanto a tamaño (15,7 % de los contribuyentes considerados) y que en este estudio no fue posible de caracterizar y estudiar con profundidad. Sobre todo, se debería estudiar el comportamiento de los contribuyentes pertenecientes a este grupo, en los años anteriores y posteriores, tomando como hipótesis el hecho de que hay incentivo a permanecer en una actividad, solo mientras las utilidades son positivas.

En trabajos futuros, se puede elaborar otros indicadores, que permitan evaluar el comportamiento en los otros clusters, como el 3 ("Actividades Exentas") y el 4 ("Retenedores").

Se recomienda realizar este estudio, usando otros métodos para el preprocesamiento de las variables y el agrupamiento de los datos, que eventualmente podrían llevar a resultados diferentes. Para ello se puede por ejemplo: cambiar la topología del SOFM en su forma y tamaño de la grilla, re-muestrear, usar otros tipos de escalamiento de variables o indagar en métodos de clustering que trabajen con otras medidas de disimilaridad, con el fin de incorporar variables cualitativas. Este último aspecto puede ser muy relevante, al permitir la incorporación de la Actividad Económica.

Agradecimientos: Este trabajo fue parcialmente financiado por el Insti-

tuto Milenio Sistemas Complejos de Ingeniería.

Referencias

[1] P. Berkhin, "Survey of clustering data mining techniques", Technical Report, Accrue Software, San Jose CA, 2002.

[2] U. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", Article, American Association for Artificial Intelligence, 1996, Vol.17, Nº 3, pp. 37-54.

[3] G. Fung, "A Comprehensive Overview of Basic Clustering Algorithms", June 2001. http://www.cs.wisc.edu/ gfung/clustering.pdf

[4] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On clustering validation techniques", Journal Article, Journal of Intelligent Information Systems, Dec. 2001, Vol. 17, pp. 107-145.

[5] J. Handl, J. Knowles, D. B. Kell, "Computational Cluster Validation in Post-genomic Data Anlysis" School of Chemistry, University of Manchester UK, Bioinformatics Review, Vol. 21, Mayo 2005, pp. 3201-3212.

[6] J. Hartigan And M. Wong, "Algorithm AS136: A K-means clustering algorithm", Applied Statistics, Vol. 28, pp. 100-108, 1979.

[7] T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning: Datamining, Inference and Prediction", Springer, New York, 2001, Cap. 14, pp. 437-508.

[8] E. Paradis, "R para Principiantes", Institut des Sciences de lÉvolution, Universit Montpellier II, France, http://cran.r-project.org/doc/contrib/rdebuts-es.pdf.

[9] J. W. Sammon, JR, "A Nonlinear Mapping for Data Structure Analysis", Transactions on Computers, Mayo 1969, Vol. C-18, Issue 5, pp. 401-409.

[10] Servicio De Impuestos Internos, Formulario Inscripción al Rol único Tributario y/o Declaración de Inicio de Actividades, http://www.sii.cl/formularios/imagen/4415.PDF

[11] Servicio De Impuestos Internos, Formulario Declaración Mensual y Pago Simultáneo de Impuestos (F29), http://www.sii.cl/formularios/anverso-f29.pdf

[12] Servicio De Impuestos Internos, Suplemento Formulario 29, www.sii.cl

[13] B. Silverman, "Density Estimation for Statistics and Data Analysis", Monographs on Statistics and Applied Probability, 1986, Chapman and Hall, London

[14] L.I. Smith, "A Tutorial on Principal Components Analysis", Febrero 2002, http://www.cs.otago.ac.nz/cosc453/student tutorials/principal-components.pdf

[15] R. Tibshirani, G. Walther, T. Hastie, "Estimating the Number of Clusters in a Dataset via the Gap Statistic", Journal of the Royal Statistical Society: Series B (Statist. Methodol.), Vol. 63, Marzo 2000, pp. 411-423.

[16] L. Torgo, "Data Mining with R: learning by case studies", LIACC-FEP, University of Porto, 22 Mayo 2003, http://www.liac.up.pt/ ltorgo

[17] J.D. Velásquez, V. Palade., "Adaptive web site: A knowledge extraction from web data approach", IOS Press, chapter 3: "Knowledge discovery from web data".

[18] J. Vesanto, "Using SOM in Data Mining", Licentiate´s Thesis, Finland, Abril 2000.

[19] A. Weingesel, E. Dimitriadou, S. Dolnicar, "An Examination of Indexes For Determining The Number Of Clusters In Binary Data Sets", Psychometrika, 2002, Vol. 67, Nº 1, pp. 137-160.

[20] I. Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Junio 2005

[21] <http://csnet.otago.ac.nz/cosc453/student-tutorials/principal-components.pdf> [Consulta: Noviembre 2006]

[22] <http://datamining.anu.edu.au/student/math3346-2006/3.4up.pdf> [Consulta: Noviembre 2006).

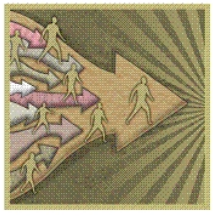
[23] R Project, <http://www.r-project.org> [Consulta: Junio 2006 a Marzo 2007]

[24] Servicio De Impuestos Internos, <http://www.sii.cl> [Consulta: Junio 2006 a Marzo 2007]

[25] http://www.ir.iit.edu/ñazli/cs422/CS422-Slides/DM-Preprocessing.pdf [Consulta: Noviembre 2006]

Programas de Postgrado Impartidos por el DII

INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE CHILE
M A G I S T E R



MGO

El Magíster en Gestión de Operaciones busca formar profesionales de excelencia en esquemas de gestión, uso de modelos y tecnologías de información, con capacidad de resolución de problemas complejos en el ámbito de la gestión de operaciones.

Programa impartido por Ingeniería Industrial, reacreditado por 5 años.

Postulación: Octubre a 15 de Diciembre para ingreso en Marzo de cada año. Abril a 15 de Junio para ingreso en Julio de cada año.

Informaciones:
Teléfonos: 9784018-9784073
email: julie@dii.uchile.cl
www.dii.uchile.cl/mgo

Gestión de Operaciones

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

MAGCEA
MAGÍSTER EN ECONOMÍA APLICADA

El Magíster en Economía Aplicada (MAGCEA) del Departamento de Ingeniería Industrial de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, busca formar profesionales de gran competencia analítica y una sólida base en economía.

El MAGCEA es impartido por el Centro de Economía Aplicada (CEA) del Departamento de Ingeniería Industrial, líder en investigación en economía en Chile y en el desarrollo de propuestas para las políticas públicas. Adicionalmente, destacados académicos y profesionales son invitados a dictar clases en el programa. El programa está reacreditado por CONAP.

Requisitos de Admisión

Título Profesional, nacional o extranjero, que exija al menos 5 años de estudio o el grado de licenciado en campos disciplinarios afines a la especialidad.

Calendario Académico

Semestre Otoño: Período de Postulaciones: Octubre a 15 de Diciembre. Inicio de clases Marzo de cada año.

Semestre Primavera: Período de Postulaciones: Abril a 15 de Junio. Inicio de clases Julio de cada año.

Postulación en línea: www.magcea-uchile.cl

MAYOR INFORMACIÓN:

Domeyko 2313, Piso 1, Santiago de Chile

Teléfonos: (562) 9784084- (562) 9784073

Email: magcea@dii.uchile.cl

Página web: www.magcea-uchile.cl

INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE CHILE

MAGÍSTER EN GESTIÓN Y POLÍTICAS PÚBLICAS-**MGPP**

LÍDERES DE EXCELENCIA PARA AMÉRICA LATINA

Fecha de inicio

Junio de 2008

Duración

Un año y medio

Postulaciones

- Hasta el **15 de Noviembre de 2007** para personas que postulan a becas de instituciones.
- Hasta el **15 de abril de 2008** para personas que cuentan con fondos propios.

Nueva Versión en Horario Ejecutivo

Para información sobre plazos de postulación, fecha de inicio y duración, consultar: **www.mgpp.cl**

Antecedentes y postulaciones

Comité de Admisiones
Av. República 701, Santiago. Chile.
Teléfono: (56 2) 978 4067. Fax: (56 2) 689 4987.
mgpp@dii.uchile.cl

www.mgpp.cl

fcfm FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

Doctorado en Economía

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

FACULTAD ECONOMÍA Y NEGOCIOS
UNIVERSIDAD DE CHILE

El Doctorado en Economía busca ser un referente regional en la formación de economistas de alto nivel y de esta forma atraer a estudiantes de excelencia de Chile y América Latina, que tengan claros intereses académicos.

El Doctorado en Economía es el primer programa de Doctorado en Economía en Chile y es impartido por la Facultad de Ciencias Físicas y Matemáticas y la Facultad de Economía y Negocios de la Universidad de Chile, a través de sus respectivos Departamentos de Ingeniería Industrial y Economía.

Calendario Académico

Primer proceso: Finaliza el segundo viernes de octubre de 2007.

Segundo proceso: Finaliza el primer viernes de diciembre de 2007.

Inicio de clases: Marzo de 2008. El Programa es de dedicación exclusiva.

MAYOR INFORMACIÓN:

Domeyko 2313, Piso 1, Santiago de Chile

Teléfonos: (562) 9784073

Email: fmelis@dii.uchile.cl

Doctorado en Sistemas de Ingeniería (DSI)

Es impartido por investigadores de excelencia en Gestión de Operaciones, Optimización, Energía y Transporte, de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile.

El programa integra investigación metodológica con aplicaciones sofisticadas, en la frontera del conocimiento.

Los egresados serán capaces de desenvolverse en ambientes académicos de alto nivel, áreas de innovación en la empresa, instituciones públicas y consultoras. Ellos podrán tomar decisiones en sistemas complejos en los que interactúan infraestructura y comportamiento humano; que combinan gran tamaño, aleatoriedad, aspectos dinámicos y/o externalidades

Mayor información:
julie@dii.uchile.cl
www.sistemasdeingenieria.cl/doctorado
Fonos: (56-2) 978 4017- 978 4073

Único en su tipo en Latinoamérica

Calendario de postulaciones:
Inicio de Clases: marzo y julio de cada año.
Recepción final de postulaciones: 1 de diciembre del año anterior y el 1 de junio, respectivamente.