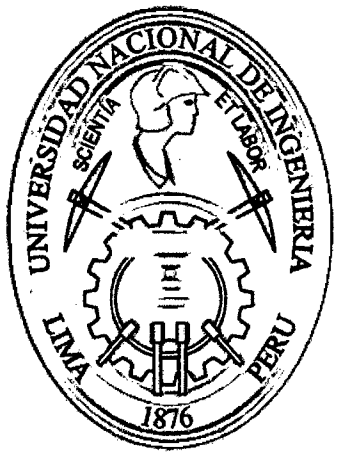
UNIVERSIDAD NACIONAL DE INGENIERÍA
FACULTAD DE INGENIERÍA INDUSTRIAL Y DE SISTEMAS



RECONOCIMIENTO DE PATRONES DE COMPORTAMIENTO DE USUARIOS EN EL PORTAL WEB USANDO WEB MINING

TESIS

Para optar el Título Profesional de:

INGENIERO DE SISTEMAS

MARTÍNEZ ROMERO, Miguel Iván

MUÑOZ DOMÍNGUEZ, Amancio Edwin

Lima - Perú

2011

RECONOCIMIENTO DE PATRONES DE COMPORTAMIENTO DE USUARIOS EN EL PORTAL WEB USANDO WEB MINING

MARTÍNEZ ROMERO, Miguel Iván

MUÑOZ DOMÍNGUEZ, Amancio Edwin

05 de Setiembre del 2011

DEDICATORIA

De Muñoz Domínguez Amancio Edwin:

A Dios por darme la vida y porque creo en su grandeza, a mis Padres Julián y Dionisia quienes creyeron en mi educación y formación profesional como mejor legado, a mis hermanos Mary, JJ, Irma, Ivan y Elvis en especial a ti Mary que con tu alegría y comprensión unes la familia y a todos que estuvieron, están y estarán en mi vida. Les dedico y doy gracias a todos Ustedes.

De Martínez Romero Miguel Iván:

A Dios por iluminarme y darme sabiduría, a mis padres Sabina Romero Blas y Zenobio Urbano Martinez Ramírez por acompañarme en todos mis logros y metas de mi vida y a mis hermanos Nilton, Jacky, Ericka, Ezio y Crhistian por darme su apoyo y comprensión.

ÍNDICE GENERAL

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

DESCRIPTORES TEMÁTICOS

1. Inteligencia de Negocio.
2. Data Mining.
3. Web Mining.
4. Técnicas de Agrupamiento.
5. Cluster.
6. Clustering.
7. K-Means.
8. Portal Web.
9. Patrón de Comportamiento.
10. Sesión.

RESUMEN

La siguiente investigación tiene la finalidad de encontrar los patrones de comportamiento de los usuarios de un Portal Web, para ésto se utiliza la metodología de Web Mining, la técnica del *Clustering* y el algoritmo de K-Means: esta metodología permite formar grupos con características iguales o similares (preferencias y cantidad de visitas a uno o más servicios). Una vez que se tenga el conocimiento de los grupos se personaliza el Portal Web para cada uno éstos grupos sobre la base de sus preferencias. Con esto se logra un menor tiempo de acceso al servicio deseado y un mayor grado de satisfacción de los usuarios. Los datos a analizar se extrae de los registros de acceso de usuarios hacia el Portal Web (archivos *Logs*) en los cuales se registran los datos de navegación de los usuarios, y éstos se almacenan en el servidor web.

ABSTRACT

The following research aims to find the patterns behavior of users of a Web Portal for this use the methodology of Web Mining, Clustering technique and K-means algorithm, this methodology will allow us to form groups with same or similar characteristics (preferences and number of visits to one or more services). Once we have knowledge of the groups, customize the Web Portal for each group based on their preferences. This will achieve a shorter access to the service desired and greater user satisfaction. The data analyzed is drawn from the access logs (log files) in the which records the navigation data users, and these stored in the web server.

INTRODUCCIÓN

Los entornos Web ofrecen nuevos escenarios para la interacción del cliente y la empresa, de allí la importancia de conocer el comportamiento y las preferencias de los usuarios Web, esta información permite tomar decisiones en el rediseño y la mejora del Portal Web sobre la base de los patrones de comportamiento identificado con el análisis de los archivos *Logs*.

El presente estudio aborda especialmente el contexto de los Portales Webs que brindan servicios: aplicables también a páginas comerciales, redes sociales, etc. Se Analiza el Portal Web de la SUNAT (Superintendencia Nacional de Administración Tributária) cuyos usuarios son los contribuyentes que realizan trámites, consultas y/u otras acciones en dicho Portal Web.

El estudio se justifica debido a la poca importancia que le dan las empresas peruanas al estudio de Data Mining y Web Mining como mecanismo para conocer el comportamiento de los usuarios, pues

presentan sus servicios y/o productos en el Portal Web sin ningún estudio del registro de los usuarios, teniendo como resultado: la poca aceptación del Portal Web.

El objetivo del estudio es llegar a determinar, a través de la metodología Web Mining y la técnica de *Clustering*, patrones de comportamiento de los usuarios web y la aplicación de aquellos patrones en la personalización de los portales de acuerdo a las necesidades de los usuarios.

La hipótesis planteada es:

Si se mejora un Portal Web a través de la personalización de dicho Portal Web para cada Cluster, esto gracias a los patrones de comportamiento de los usuarios, entonces habrá una disminución en el tiempo de acceso hacia los servicios del Portal Web así también un aumento en el nivel de satisfacción de los usuarios.

Se dará validez a dicha hipótesis a través de encuestas a los usuarios tanto antes como después de la mencionada mejora.

> *"La mayor parte de las personas son mucho más predecibles de lo que creen"*
>
> Andreas Weigend[1]

[1] Andreas Weigend, Científico, ex jefe de amazon.com y ha escrito más de 100 artículos científicos sobre redes sociales, finanzas y negocios. Actualmente es conferencista en las universidades de California, Berkeley, Stanford y de la universidad de Tsinghua.

CAPÍTULO I

DIAGNÓSTICO ACTUAL

I.1 FORMULACIÓN DEL PROBLEMA

I.1.1 DESCRIPCIÓN DE LA SITUACIÓN PROBLEMÁTICA

La necesidad de saber cómo se comportan los usuarios, es debido a que no siempre los Portales Web son desarrollados y diseñados sobre la base de las preferencias y/o el contenido no es lo que el usuario espera encontrar y debido a esto el usuario no concretiza su objetivo o simplemente abandona el Portal Web, felizmente, existe abundante información registrada sobre las navegaciones de usuarios, por ejemplo en los servidores Web encontramos a los archivos *Log* en el cual se encuentra registrado su IP, las *URLs* de los módulos visitados, hora y fecha, etc. de los usuarios que han realizado su visita al Portal Web, sin

embargo en la mayoría de los casos a esta información no se le da la importancia debida o simplemente se la ignora.

El nuevo escenario que vienen experimentando las empresas, que consiste en mostrar y ofrecer productos y/o servicios a través de la Internet, obliga a dichas empresas a conocer: quiénes interactúan con su Portal Web, cuál es el comportamiento y cuáles son sus expectativas: con el objetivo de fidelizar y captar nuevos usuarios.

Los datos de los usuarios que navegan en un Portal Web están registrados en los archivos *Log,* mencionados anteriormente, entonces mediante el análisis de estos archivos se encuentra los patrones de comportamiento y/o preferencias: ello permitirá mejorar la presentación y el contenido del Portal Web, respondiendo a las necesidades y preferencias de los usuarios. Además nos ayudará a entender la evolución del comportamiento de los usuarios del Portal Web.

I.1.2 DESCRIPCIÓN DEL PROBLEMA

¿Por qué los usuarios de los Portales Web que requieren consumir un servicio, tienen la intención de hacerlo pero no concretizan su objetivo o no satisfacen su necesidad? ¿Se puede predecir mediante patrones de comportamiento el servicio o módulo que el usuario desea visitar?

Mediante la personalización del Portal Web, para cada *Cluster* y sobre la base de patrones de comportamiento hallados, se puede:

¿Incrementar la aceptación de los usuarios con respecto a los servicios que brinda el Portal Web? ¿Disminuir el tiempo de acceso hacia los servicios del Portal Web? ¿Incrementar el nivel de satisfacción de los usuarios del Portal Web?

I.2 JUSTIFICACIÓN Y VIABILIDAD

La tesis se argumenta en descubrir información relacionada a los patrones de comportamiento de los usuarios de los Portales Web. Por lo tanto la técnica de *Clustering* utilizando el algoritmo *K-Medias* se justifica por:

- Los escasos estudios de las empresas sobre el reconocimiento de los patrones de comportamiento de los usuarios de Portal Web.
- Los métodos clásicos estadísticos de *Clustering* soportan una cantidad determinada de variables, pero a medida que estas aumentan los resultados no son precisos. La tesis propone una técnica de *Clustering* cuyos resultados no son alterados a medida que las variables de análisis van aumentando, permitiendo con esto la escalabilidad de estudio.

I.3 HIPÓTESIS

Se plantea la hipótesis global y las hipótesis específicas.

I.3.1 HIPÓTESIS GLOBAL

Si se mejora un Portal Web a través de la personalización de dicho Portal Web para cada *Cluster*, esto gracias a los patrones de comportamiento de los usuarios, entonces habrá una disminución en el tiempo de acceso hacia los servicios del Portal Web así también un aumento en el nivel de satisfacción de los usuarios.

I.3.2 HIPÓTESIS ESPECÍFICAS

Tenemos las siguientes hipótesis específicas:

1. La aplicación de los patrones de comportamiento de los usuarios de un Portal Web, en la mejora del mismo, permite disminuir el tiempo de acceso hacía los servicios encontrados para cada *Cluster*: *Cluster* 1, *Cluster* 2, *Cluster* 3, . . . *Cluster* k.

   Donde k es el número de *Clúster* óptimos encontrados.

2. El nivel de satisfacción de los usuarios del Portal Web se incrementa con el nuevo Portal Web, personalizado para cada *Cluster*.

I.3.3 IDENTIFICACIÓN Y OPERACIONALIZACIÓN DE LAS VARIABLES

Del problema de investigación:

¿Cómo influye la aplicación de los patrones de comportamiento de los usuarios de un Portal Web en la disminución del tiempo de acceso hacia los servicios de dicho Portal Web?

Variable Independiente

Patrones de comportamiento de usuarios Web.

Variable Dependiente

Tiempo de acceso hacia los servicios del Portal Web.

I.3.4 OPERACIONALIZACIÓN

VARIABLE INDEPENDIENTE	DIMENSIONES	INDICADORES
Patrones de comportamiento de usuarios Web.	Social y Gubernamental.	Disminución del tiempo para acceder hacia los servicios administrativos del Portal Web de la SUNAT por parte de los contribuyentes. Aumento de los contribuyentes conectados al Portal Web de la SUNAT. Aumento de consultas hacia los servicios del Portal Web.
VARIABLE DEPENDIENTE	**DIMENSIONES**	**INDICADORES**
Optimización del acceso a los servicios del Portal Web.	Procedimentales.	Tiempo y Beneficios.

Cuadro I.1: Operacionalidad de las variables.

I.4 OBJETIVO DE LA INVESTIGACIÓN

I.4.1 OBJETIVO GENERAL

Determinar la incidencia de la aplicación de los patrones de comportamiento de los usuarios del Portal Web en la disminución del tiempo de acceso hacia los servicios del Portal Web y en el aumento del nivel de satisfacción de los usuarios.

I.4.2 OBJETIVOS ESPECÍFICOS

1. Determinar un conjunto de servicios para cada *Cluster* encontrado y relacionar al IP del usuario que entra a un *Cluster*.
2. Entender la metodología Web Mining para determinar el patrón de comportamiento de un usuario de un Portal Web.
3. Describir conceptos relacionados con las técnicas de reconocimiento de patrones de comportamiento.
4. Desarrollar un modelo solución en la cual se describan los procedimientos para determinar un patrón de comportamiento.
5. Desarrollar un prototipo sobre la base del modelo solución y que tenga como finalidad encontrar un número de *Cluster* óptimos.

I.5 ALCANCES Y LIMITACIONES

I.5.1 ALCANCES

En esta investigación se pretende identificar los patrones de comportamiento de los usuarios que acceden hacia los servicios de un Portal Web, esta propuesta tendrá como ámbito de estudio a los portales de empresas que ofrecen servicios por Internet.

Para esta propuesta de investigación, se desarrollará un prototipo que identifique los patrones de comportamiento de los usuarios del Portal Web de la SUNAT que acceden sin necesidad de identificarse, para utilizar cualquier servicio libre que ofrece dicho Portal Web y el registro de acceso que se utilizará para el experimento es del año 2010.

I.5.2 LIMITACIONES

No se hace un estudio general de todas las páginas Web, solamente se enfoca en los Portales Web que ofrecen servicios y/o productos por dicho portal. El estudio no pretende conocer al usuario como individuo con sus características, sino que sobre la base de sus acciones en el Portal Web lo vamos acercar a un patrón a la que llamaremos "Patrón de Comportamiento".

CAPÍTULO II

MARCO TEÓRICO

II.1 MARCO TEÓRICO CONCEPTUAL

A continuación se elabora las definiciones de los términos básicos y conceptos con la finalidad de dar la base teórica para el desarrollo del tema: Reconocimiento de Patrones de Comportamiento de los Usuarios de un Portal Web Usando Web Mining.

II.1.1 PATRÓN DE COMPORTAMIENTO

Modelo que sirve de muestra para representar las singularidades de una realidad en un determinado contexto, el cual es repetitivo cada cierto periodo, es reusable, lo que significa que es aplicable a diferentes problemas se comprueba su efectividad resolviendo problemas similares de anteriores ocasiones.

II.1.2 PÁGINA WEB

Archivo - generalmente HTML - que constituye una unidad de información accesible a través de un programa navegador de Internet (Explorer, Mozilla, Chrome, Safari, etc).

Puede ser un texto corto o un gran conjunto de textos, fotografías, gráficos estáticos o animados, sonido, etc. La página web no es el contenido global de un sitio Web sino que es una parte de dicho sitio.

II.1.3 PORTAL WEB

Espacio Web que sirve de punto de partida para navegar por Internet y que, normalmente, ofrece una gran diversidad de servicios tales como listado de sitios Web, buscador, noticias, e-mail, información, *chat*, grupos de discusión, comercio electrónico, etc.

II.1.4 SESIÓN

Es el periodo que va desde que el usuario accede al Portal Web hasta cuando lo abandona, teniendo en cuenta los diferentes accesos hacia los servicios Web que el usuario pueda realizar.

II.1.5 URL (*UNIVERSAL RESOURCE LOCATOR*)

En castellano significaría Identificador Universal de Recursos. Se trata del Sistema unificado de identificación de recursos en la red. Ejemplo: *http://www.elpais.es*, la sintaxis de una dirección http, es:

"Http://" host.domain [':' puerto] "?" [abs_ruta [consulta]]

Dónde:

- *Host.domain[:puerto];* es el nombre del servidor del sitio. El puerto TCP /IP[2] es opcional (el puerto por defecto es 80).
- *Ruta absoluta*; es la ruta del recurso solicitado en el servidor.
- Consulta; es una colección de parámetros opcionales, que se pasa como entrada a un recurso que en realidad es un programa ejecutable, por ejemplo, un *Script CGI.* (Goldmann, [10]).

II.1.6 USUARIO WEB

Se define como aquel individuo que interactúa con la Web. En particular, si nuestro interés está centrado en el conocimiento de un determinado Portal Web, un usuario es cada persona que interactúa con el portal, es decir, que accede al mismo.

---

[2] TCP/IP es un modelo de descripción de protocolos de red creado en la década de 1970 por DARPA, una agencia del Departamento de Defensa de los Estados Unidos.Este modelo, describe un conjunto de guías generales de diseño e implementación de protocolos de red específicos para permitir que una computadora pueda comunicarse en una red.

II.1.7 SERVIDOR WEB

Es una máquina conectada a la red en la que se guardan físicamente las páginas web que componen un sitio Web. También se conoce con este nombre al programa que sirve dichas páginas.

II.1.8 DIRECCIÓN IP

Es una etiqueta numérica que identifica, de manera lógica y jerárquica, una interfaz (elemento de comunicación/conexión) de un dispositivo (habitualmente una computadora) dentro de una red que utilice el protocolo IP (Internet Protocol), que corresponde al nivel de red del protocolo TCP/IP. Dicho número no se ha de confundir con la dirección MAC que es un identificador de 48 bits para identificar de forma única a la tarjeta de red y no depende del protocolo de conexión utilizado ni de la red. La dirección IP puede cambiar muy a menudo por cambios en la red o porque el dispositivo encargado internamente de asignar las direcciones IP, decida asignar otra IP (por ejemplo, con el protocolo DHCP), a esta forma de asignación de dirección IP se denomina dirección IP dinámica (normalmente abreviado como IP dinámica).

Los sitios de Internet que por su naturaleza necesitan estar permanentemente conectados, generalmente tienen una dirección IP fija o IP estática, la cual no cambia con el tiempo. Los servidores de correo, DNS, FTP públicos y servidores de páginas web necesariamente deben

contar con una dirección IP fija o estática: de esta forma se permite su localización en la red.

II.1.9 ARCHIVO *LOG*

Para administrar de manera efectiva los Servidores Web, es necesario tener registro de la actividad y el rendimiento del servidor así como de cualquier problema que pudo ocurrir durante su operación. Es por ello que los servidores, independientemente del producto, ofrecen registros de estos datos, el cual podemos configurar para que registre los datos que nosotros consideremos necesarios. Entre estos registros se encuentra la información acerca de: como el registro de errores, registros de accesos, advertencia de seguridad, etc.

Para nuestro estudio el registro de accesos será el archivo al que hay que analizar, y para ello es necesario conocer este registro.

El registro de accesos, que en adelante lo llamaremos como archivo *log*, guarda información sobre todas las peticiones que procesa, por cada acción que el usuario realiza se guarda una fila, en el archivo *log*, separado por espacios en blanco la información que guarda, por lo general tiene la siguiente estructura, aunque puede variar de acuerdo a la configuración realizada.

[IP Protocolo ID_Usuario Hora "Metodo y Recurso" Cod_Estado Tamaño_Recurso "URL_del_Recurso" "Idenf_Navegador"]

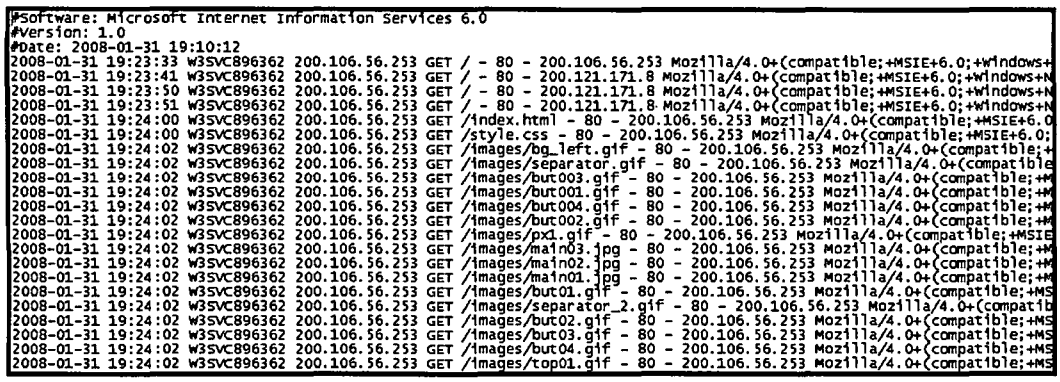
- **Protocolo:** Protocolo RFC 1413[3], que por lo general se guarda un "-", porque la información es poco confiable.

- **ID_Usuario:** Identificador del usuario de la persona que solicita el documento o recurso determinado por la autenticación HTTP[4], pero si el documento no está protegido por contraseña se mostrara un "-", lo cual es nuestro caso ya que el archivo log es de acceso sin contraseña (opciones libres).
- **Hora:** Es la hora que se recibió la petición y el formato es [dia/mes/año:hora:minuto:segundo: zona_horaria].
- **"Método y Recurso":** Es la petición del cliente, se muestra en dobles comillas. La primera, es el método usado por el cliente puede ser el GET o POST, para nuestro estudio tomaremos la peticiones que hayan utilizado el método GET. Segundo, es la petición al recurso que hecho el usuario.

[3] Protocolo RFC 1413, Protocolo de identificación (también conocido como IDENT) proporciona un medio para determinar la identidad del usuario de una conexión TCP.
[4] HTTP (*Hypertext Transfer Protocol*): Es un protocolo de red de distribución, sistemas de colaboración y de la información hipermedia. Es el fundamento de la comunicación de datos de la *World Wide Web*.

Tercero, es el protocolo que el cliente utilizo como es el HTTP.

- **Cod_Estado:** Código de estado que envía el servidor al usuario, revela si la petición fue respondida con éxito por el servidor o hubo un error en el servidor.
- **Tamaño_Recurso:** Es el tamaño del recurso u objeto retornado al usuario.
- **"URL_del_Recurso":** Es la dirección de la página que contiene un enlace o contiene al recurso solicitado por el usuario.
- **"Idenf_Navegador":** Es la información que el navegador del usuario incluye sobre sí mismo.



```
#Software: Microsoft Internet Information Services 6.0
#Version: 1.0
#Date: 2008-01-31 19:10:12
2008-01-31 19:23:33 W3SVC896362 200.106.56.253 GET / - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+
2008-01-31 19:23:41 W3SVC896362 200.106.56.253 GET / - 80 - 200.121.171.8 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+N
2008-01-31 19:23:50 W3SVC896362 200.106.56.253 GET / - 80 - 200.121.171.8 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+N
2008-01-31 19:23:51 W3SVC896362 200.106.56.253 GET / - 80 - 200.121.171.8 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+N
2008-01-31 19:24:00 W3SVC896362 200.106.56.253 GET /index.html - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+MSIE+6.0
2008-01-31 19:24:00 W3SVC896362 200.106.56.253 GET /style.css - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+MSIE+6.0;
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/bg_left.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/separator.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatible
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/but003.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+M
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/but001.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+M
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/but004.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+M
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/but002.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+M
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/px1.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+MSIE
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/main03.jpg - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+M
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/main02.jpg - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+M
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/main01.jpg - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+M
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/but01.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+MS
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/separator_2.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatib
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/but02.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+MS
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/but03.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+MS
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/but04.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+MS
2008-01-31 19:24:02 W3SVC896362 200.106.56.253 GET /images/top01.gif - 80 - 200.106.56.253 Mozilla/4.0+(compatible;+MS
```

Figura II.1: Muestra de un Archivo *Log*.

II.1.10 USABILIDAD

La Organización Internacional para la Estandarización[5] definió la usabilidad como "*el grado en el que un producto puede ser utilizado por usuarios específicos para conseguir objetivos específicos con efectividad, eficiencia y satisfacción en un determinado contexto de uso*".

II.1.10.1 LA USABILIDAD EN EL ENTORNO WEB

La usabilidad para la Web surge a partir del uso masivo de la Internet como sistema de comunicación. El desarrollo de la tecnología permite la aparición de sitios más complejos y de diseños más sofisticados con interfaz más difíciles de usar por los usuarios.

La usabilidad aporta un enfoque para la confección de entornos Web, de tal manera que los diseños elaborados sean fáciles de usar y de aprender, efectivos, eficaces y que cubran las expectativas tanto de los diseñadores como de los futuros usuarios.

La usabilidad engloba una amplia gama de aspectos. Éstos abarcan desde los aspectos puramente de diseño gráfico, como las tipografías, colores e imágenes, hasta los aspectos más específicos, como el estilo narrativo, la estructura de la información, los elementos que integran los

[5] **ISO**, Organismo Internacional de Normalización, es el organismo encargado de promover el desarrollo de normas internacionales de fabricación, comercio y comunicación.

sistemas de navegación, etc. Principalmente podemos agrupar estos aspectos en cuatro grandes bloques:

- **La Información:** Podemos englobar en este apartado todos los aspectos referentes a los sistemas de organización de la información, que vendrán determinados por las características de la misma y que serán decisivos para los sistemas de navegación del Portal Web.
- **La Navegabilidad:** Este apartado engloba todos los aspectos que permitirán a los usuarios moverse a través de las diferentes páginas del sitio. Este apartado está estrechamente relacionado con la arquitectura de la información del Portal Web.
- **La Interfaz:** Es uno de los elementos fundamentales, ya que es la herramienta a través de la cual el usuario se comunica con el sistema. Debe ser fácil de usar y de aprender, eficaz, cómodo y agradable. En este apartado se englobaría los aspectos puramente de diseño, como los colores, tipografía, imágenes, etc.

II.1.10.2 REESTRUCTURACIÓN DE UN CICLO DE VIDA EN INGENIERÍA DE USABILIDAD

Habitualmente se suele otorgar poca importancia a los usuarios en los modelos de desarrollo *Software*, ya que típicamente la figura del usuario exclusivamente aparece al principio del desarrollo (Ingeniería de Requerimientos), al final del mismo o al final de cada etapa, pero no durante el proceso de desarrollo.

Para contribuir a cambiar esta situación, se han agrupado todas las actividades en usabilidad que describe Nielsen en (Nielsen, Jakob [2]) de tal manera que resulten en un ciclo de vida de fácil inclusión en diferentes modelos de desarrollo *Software* (hemos experimentado en los de desarrollo evolutivo e incremental). Cada una de las actividades en usabilidad se ha enmarcado dentro de etapas genéricas (como diseño, implementación, etc), que son unidades conceptuales que engloban actividades similares en un mismo momento del desarrollo. El ciclo de vida se recoge en el siguiente cuadro:

ETAPA	ACTIVIDAD QUE TIENE LUGAR
Recolección de información	Perfiles de usuario, análisis de tareas, presupuesto, análisis de funciones, establecer metas y análisis competitivo.
Diseño	Diseño paralelo, diseño participativo, diseño conceptual, consistencia en la

	interfaz y una interfaz de usuario propuesta.
Implementación	Fijar directrices del proyecto, prototipos horizontales, prototipos verticales, elaboración del prototipo final e interfaz final a evaluar.
Evaluación	Evaluación heurística, evaluación con usuarios reales, métodos de prototipaje y problemas de usabilidad.
Diseño iterativo	Volver a realizar todas las fases, dependiendo de los problemas hallados.
Seguimiento	Recoger información del sistema ya instalado para mejorar la usabilidad en aplicaciones futuras.

Cuadro II.1: Etapas y actividades del ciclo de vida en ingeniería de usabilidad.

Dentro del ciclo de vida de la ingeniería de usabilidad de Nielsen, la etapa de evaluación no es considerada con la importancia debida es decir los Portales Web no son evaluados teniendo en cuenta las necesidades de los usuarios por parte de la empresas que tienen un Portal Web, por ello se hace difícil a los usuarios navegar y es justamente la propuesta de esta tesis la que abarca un estudio sobre los patrones de comportamiento de los usuarios Web y su aplicación en la mejora de los Portales Web

para satisfacer la necesidades de usabilidad, navegabilidad y accesibilidad de los usuarios.

II.1.11 ACCESIBILIDAD

La accesibilidad en un sitio Web consiste en garantizar el acceso a la información y a los servicios de sus páginas sin limitación ni restricción alguna por razón de discapacidad de cualquier carácter o condicionantes técnicos, debiendo tener en cuenta que muchas personas que acceden a la información incluida en páginas Web lo hacen desde diferentes dispositivos y contextos.

II.2 MARCO TEÓRICO INSTRUMENTAL

En este segmento se describe las técnicas y metodologías aplicadas para el Reconocimiento de los Patrones de los usuarios de un Portal Web.

II.2.1 INTELIGENCIA DE NEGOCIO

Una interesante definición para inteligencia de negocios o BI, por sus siglas en inglés, según el Data Warehouse Institute, lo define como. "*La combinación de tecnologías, herramientas y procesos que nos permiten transformar los datos almacenados en información, esta información en conocimiento y este conocimiento dirigido a un plan o una estrategia comercial*". La integración de los negocios debe ser parte de la estrategia empresarial, esa le permite optimizar la utilización de recursos, monitorear el cumplimiento de los objetivos de la empresa y la capacidad de tomar buenas decisiones para así obtener mejores resultados.

Es importante visualizar de alguna forma en qué consiste una arquitectura de inteligencia de negocios. La siguiente *Figura II.2* representa esta arquitectura. Se Analiza este diagrama de izquierda a derecha. Los primeros dibujos representan las distintas fuentes de datos (cubos, bases de datos, archivos planos, archivos xml, hojas de cálculo, etc) que pudieran utilizarse para extraer los datos de múltiples fuentes simultáneamente. El segundo dibujo representa el proceso de extracción, transformación y carga (ETL). En este proceso se define qué campos se

van a utilizar, si se necesita algún tipo de modificación y/o transformación y en donde quiero ubicar estos datos: a este procesos se le conoce como *Maping*. El tercer dibujo representa el repositorio de datos, en este repositorio se encuentran los datos transformados y presentados visualmente en modelos multidimensionales con tablas de datos.

Existe un proceso entre el repositorio de datos y la interfaz de acceso al usuario, éste es el motor de Inteligencia de Negocios, que permite habilitar componentes, administrar consultas, monitorear procesos, cálculos, métricas. La interfaz de acceso a usuarios permite interaccionar con los datos y presentar de forma gráfica.

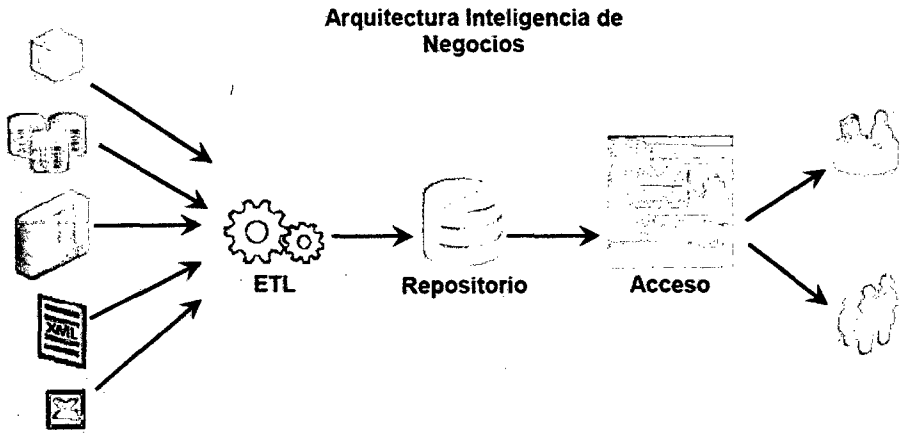


Figura II.2: Arquitectura Inteligencia de Negocios.

Fuente: Oracle BI [3].

II.2.2 INTRODUCCIÓN A LA MINERÍA DE DATOS WEB

Las técnicas de Data Mining son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos de negocios fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. Data Mining toma este proceso de evolución más allá del acceso y la navegación retrospectiva de los datos, hacia la entrega de una información prospectiva y proactiva. Data Mining está lista para su aplicación en la comunidad de negocios porque está soportado por tres tecnologías que ya están suficientemente maduras:

- Recolección masiva de datos.
- Potentes computadoras con multiprocesadores.
- Algoritmos de Data Mining.

II.2.3 MINERÍA DE DATOS

Es la extracción no trivial de la información oculta y predecible de grandes bases de datos, para luego ser transformados y representados en modelos que permitan hacer predicciones o tomar decisiones, ver *Figura II.3.* Las herramientas de Data Mining predicen futuras tendencias y comportamientos, estas herramientas explotan las bases de datos en busca de patrones ocultos, hallando información predecible que un experto no podría llegar a encontrar porque está fuera de sus expectativas.

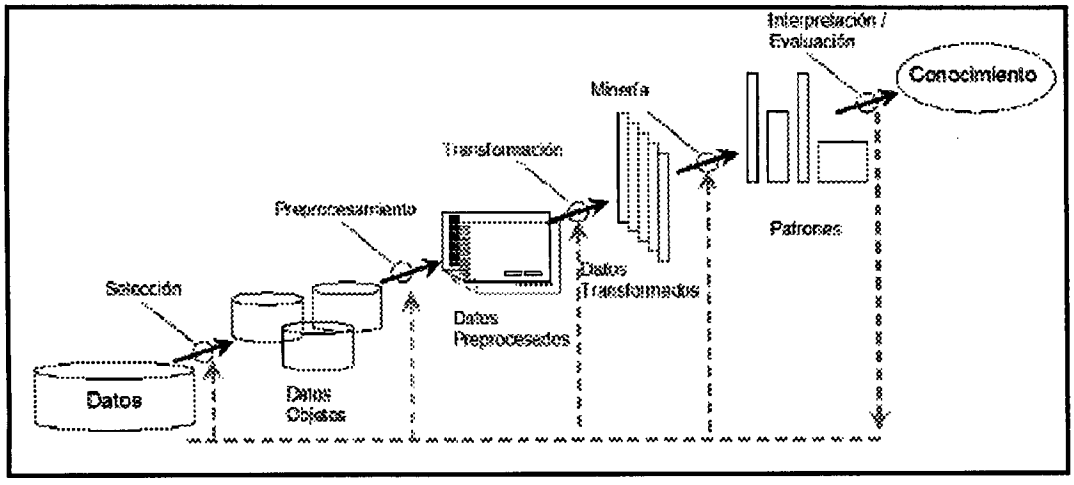


Figura II.3: Metodología de Data Mining.

II.2.4 AGRUPAMIENTO O *CLUSTERING*

Análisis de conglomerados, también llamado segmentación de datos, tiene una variedad de objetivos. Todos se refieren a la agrupación o división de una colección de objetos (también llamado observaciones, individuos, casos, o líneas de datos) en subconjuntos o *Clusters*, de tal forma que los objetos de cada grupo están estrechamente relacionados entre sí que los objetos asignados a los distintos grupos.

II.2.5 MINERÍA WEB

Algunos autores definen a la Web Mining como el uso de técnicas para descubrir y extraer de forma automática información de los documentos y servicios de la Web. Según M. Scotto, "*La Web Mining es el proceso de descubrir y analizar información útil de los documentos de la Web*" (M. Scotto [4]). Sin embargo y tomando en cuenta lo expuesto en la introducción la minería web se puede definir como el descubrimiento y el análisis de información relevante que involucra el uso de técnicas y acercamientos basados en la minería de datos, orientados al descubrimiento y extracción automática de información de documentos y servicios de la Web, teniendo en consideración el comportamiento y preferencias del usuario.

En la Web Mining, los datos pueden ser coleccionados en diferente niveles; en el área del servidor, en el lado del cliente (Cookies), en los servidores proxy (*log* files), etc.

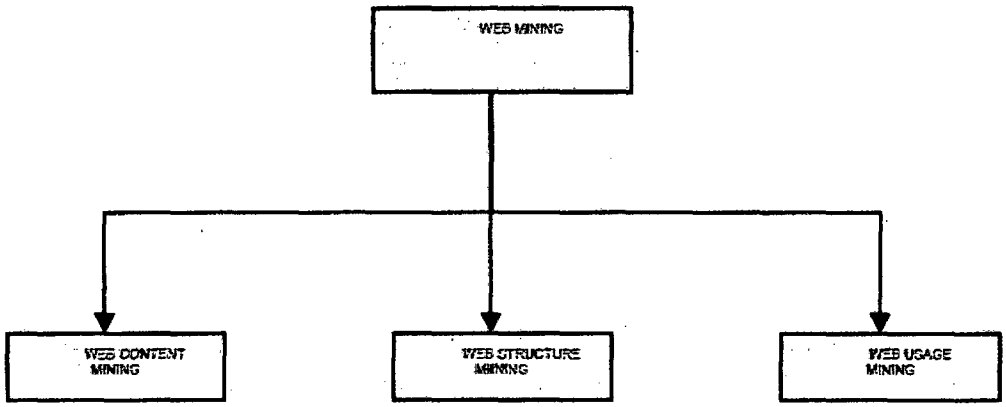


Figura II.4: Tipos de metodologías Fuente: Francisco de Gyves, [14].

II.2.5.1 TIPOS DE MINERÍA WEB

Tenemos los siguientes tipos:

a) MINERÍA DE CONTENIDO WEB

Su objetivo es la recolección de datos e identificación de patrones relativos a los contenidos de la Web y a las búsquedas que se realizan sobre los mismos. Es decir, son los datos reales que se entregan a los usuarios, los datos que almacenan de los sitios Web (Baeza-Yates, R. Pobrete [5]).

La minería de contenidos consiste de datos desestructurados tales como tablas o páginas generadas con datos de bases de datos y semi-estructurados tales como documentos HTML, textos libres, etc. Existen dos grupos de estrategias sobre minería de contenidos: aquellas que minan directamente el contenido de los documentos y aquellas que mejoran en la búsqueda de contenidos.

b) MINERÍA DE ESTRUCTURA DE LA WEB

La minería de estructura intenta descubrir el modelo subyacente de las estructuras de los enlaces de la Web. El modelo se basa en la topología de los hiperenlaces con la descripción de los enlaces, o sin ella.

Este modelo puede ser usado para categorizar las páginas Web y es útil para generar información tal como la similitud y relación entre diferentes páginas Web. Es decir revela la estructura real de un sitio Web a través de la recolección de datos referentes a su estructura y principalmente a su conectividad. Típicamente consta de dos tipos de enlaces: estáticos y dinámicos.

c) MINERÍA DE USO DE LA WEB

La minería de uso intenta dar sentido a los datos y comportamientos generados en las sesiones de navegación de la Web. Es decir, son aquellos datos que describen el uso al cual se ve sometido un sitio, registrados en los archivos *Logs* de acceso hacia los servidores Web. A partir de esta información se podría concluir, por ejemplo, qué documento visitado no tiene razón de ser, o si una página no se encuentra en los primeros niveles de jerarquía de un sitio Web (Baeza-Yates, R. Pobrete [5]). Analizar los *Logs* de diferentes servidores Web, puede ayudar a entender el compartimiento del usuario, como la estructura de la web, permitiendo de este modo mejorar el diseño de esta colección de recursos (Galeas, [7]).

II.2.6 TÉCNICAS DE AGRUPAMIENTO

Tenemos las siguientes técnicas:

II.2.6.1 REGLAS DE ASOCIACIÓN

Por lo general esta técnica es utilizada para descubrir la correlación entre los accesos de los clientes a varios archivos disponibles en el servidor. Cada transacción está conformada por un conjunto de URL accedidas, por el usuario en una visita al Portal Web.

II.2.6.2 *PATH* ANÁLISIS

Este análisis es una extensión del modelo de regresión, usada para probar las correlaciones entre dos a más modelos causales que están siendo comparados. La regresión está hecha para cada variable, como un dependiente de los otros donde el modelo indica causas. Los pesos de regresión predichos por el modelo son comparados en una matriz de correlación para las variables, y así se calcula el índice de bondad de ajuste. El mejor ajuste de dos o más modelos es seleccionado por el investigador como el mejor modelo.

II.2.6.3 SECUENCIAS DE PATRONES

Esta técnica se basa en descubrir patrones de un conjunto de ítems en orden temporal. Analizando estos datos se puede determinar el comportamiento de los usuarios con respecto al tiempo.

II.2.6.4 *CLUSTERING*

Análisis de conglomerados, también llamado segmentación de datos, tiene una variedad de objetivos. Todos se refieren a la agrupación o la división de una colección de objetos (también llamado observaciones, individuos, casos, o líneas de datos) en subconjuntos o *Clusters*, de tal forma que los objetos de cada grupo están estrechamente relacionados entre sí.

II.2.7 ALGORITMO DE AGRUPAMIENTO

Tenemos los siguientes tipos de algoritmos:

II.2.7.1 ALGORITMO JERÁRQUICO

Los llamados métodos jerárquicos tienen por objetivo agrupar *Clusteres* para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud.

Los métodos jerárquicos se subdividen en aglomerativos y disociativos. Cada una de estas categorías presenta una gran diversidad de variantes:

1. Los métodos aglomerativos, también conocidos como ascendentes, comienzan el análisis con tantos grupos como individuos haya. A partir de estas unidades iniciales se van formando grupos, de forma ascendente, hasta que al final del proceso todos los casos tratados están englobados en un mismo conglomerado.

2. Los métodos disociativos, también llamados descendentes, constituyen el proceso inverso al anterior.

Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van

formando grupos cada vez más pequeños. Al final del proceso se tienen tantas agrupaciones como casos han sido tratados.

II.2.7.2 ALGORITMO *FUZZY K-MEANS*

El *Fuzzy K-Means* es una generalización del algoritmo *K-Means* en el ámbito de la lógica difusa. Si como información de origen se dispone únicamente de proximidades entre objetos, puede obtenerse una partición *Fuzzy K-means* a partir de expresiones basadas en las mismas proximidades sin necesidad de disponer de coordenadas de representación de los objetos. Por otra parte, debido a que la participación de un individuo en un *Cluster* influye en su posterior reasignación, se propone una modificación del algoritmo basada en una validación cruzada (*cross-validación*) que elimina dicho efecto.

II.2.7.3 ALGORITMO *K-MEANS*

Es uno de los más simples y conocidos algoritmos de agrupamiento, sigue una forma fácil y simple para dividir una base de datos dada en k grupos (fijados a priori). La idea principal es definir k centroides (uno para cada grupo) y luego tomar cada punto de la base de datos y situarlo en la clase de su centroide más cercano. El próximo paso es recalcular el centroide de cada grupo y volver a distribuir todos los

objetos según el centroide más cercano. El proceso se repite hasta que ya no haya cambio entre los grupos de un paso al siguiente.

El problema de uso de estos esquemas es que fallan cuando los puntos de un grupo están muy cerca del centroide de otro grupo, también cuando los grupos tienen diferentes tamaños y formas.

II.2.7.4 ALGORITMO *GRASP K-MEANS*

Un procedimiento de búsqueda voraz, aleatoria y adaptativa (*GRASP*) es una meta heurística propuesta por Feo y Resende para encontrar soluciones aproximadas de problemas de optimización combinatoria, mediante un proceso iterativo. En cada iteración se realizan dos fases de operaciones: construcción y búsqueda local. En la fase de construcción se genera un conjunto solución S de una instancia E de un problema combinatorio, y en la fase de búsqueda local se determina una posible mejor solución a S: finalmente, se elige la solución mejor entre la solución de la iteración anterior y la actual. La mejor solución será indicada por una función objetivo f. Cada iteración es realizada un número máximo de veces (MAX_ITER). A continuación se presenta en notación de pseudocódigo o algoritmo *GRASP* básico tal como fue descrito:

Algoritmo Grasp (E, MAX_ITER, α)

1. Inicializar solución S: = ∅ y f* := ∞

2. Repetir MAX_ITER veces

2.1. Obtener una solución S* de Construcción_Grasp (E, α)

2.2. Obtener una solución S* de Búsqueda_Local_Grasp (S*)

2.3. Si f(S*) < f*, entonces

2.3.1. Actualizar S:= S* y f* := f(S*)

2.4. Fin Si

3. Fin Repetir

4. Solución S

El hecho de que la fase de búsqueda local toma como entrada la solución obtenida en la fase de construcción proporciona una diferencia notable frente a los algoritmos de búsqueda local tradicionales.

II.2.7.5 CUADRO COMPARATIVO DE ALGORITMOS

CRITERIO DE COMPARACION	JERÁRQUICO	FUZZY KMEANS	KMEANS	GRASP KMEANS
CRITERIO DE AGRUPACIÓN	El criterio de agrupación es la distancia. Los objetos que estén cerca uno del otro pertenecerían al mismo conglomerado o *Cluster*, y los objetos que estén lejos uno del otro pertenecerían a *Clusters* diferentes.	Si como información de origen se dispone únicamente de proximidades entre objetos, puede obtenerse una partición *Fuzzy K-means*	El criterio de agrupación es la distancia agrupando a objetos con distancias más cortas en centroides.	Agrupa los *Clusters* mediante la proximidad de los objetos situándolos en el espacio haciendo uso de vectores.
CANTIDAD DE OBJETOS EN LA MUESTRA	Es el más útil cuando se desea agrupar un número pequeño (menos que algunos cientos) de objetos.	Es más útil cuando se desea agrupar una cantidad pequeña de objetos (de 50 a 100)	Es más útil cuando se desea agrupar una cantidad regular de objetos (de 100 a 500)	Si puede ser útil para grandes colecciones de objetos más de 100.

Cuadro II.2: Tipos de Algoritmos de Agrupación.

El algoritmo elegido para el presente trabajo es el *Grasp K-Means*, debido a que es un algoritmo que tiene una secuencia repetitiva que hace reducir el error en el agrupamiento de elementos es eficaz para grandes agrupaciones, de más de cien elementos, además el agrupamiento es mediante la proximidad de los elementos a través de la distancia entre ellos. A continuación veremos con detalle este algoritmo.

II.2.7.6 DETALLE DEL ALGORITMO *GRASP K-MEANS*

Adaptamos la meta heurística Grasp para resolver de manera eficiente el problema del *Clustering* minimizando la función objetivo:

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} d(x_i, \bar{x}_j)^2$$

En ese sentido, enfocamos el algoritmo *K-Means* considerando las dos fases de la arquitectura de *Grasp* y una fase adicional previa a éstas, llamada inicialización KM. El algoritmo formulado debe realizar repetidas veces (MAX_ITER veces) la secuencia de las tres fases mencionadas.

Así, en la fase de inicialización (Inicialización KM) se obtienen los K centros iniciales y se definen los *Clusters* Cj en torno de sus centros iniciales. Estos *Clusters* sirven de base para la fase de construcción (ConstrucciónKM)

donde se refina la solución inicial, evitando caer en óptimos locales. La siguiente fase, búsqueda de la mejor solución (MemoriaKM), se basa en la exploración de nuevas soluciones alterando heurísticamente la estructura de los *Clusters* obtenidos en la fase de construcción para mejorar la solución.

Por último, retiene la mejor solución entre la anterior y la actual. Presentamos la estructura del algoritmo *GraspKM*, para después presentar en detalle cada una de las fases mencionadas. Se consideran como datos de entrada el conjunto de objetos X, un número K de *Clusters* a generar, la relajación α, y el máximo número de iteraciones MAX_ITER. Debemos esperar como resultado el conjunto de *Clusters* C.

a) ALGORITMO GRASPKM (X, K, MAX_ITER)

1. f* := ∞, C := {}

2. Repetir MAX_ITER veces

2.1. C' := InicializacionKM(X, K)

2.2. C' := ConstruccionKM(X, K, C', α_)

2.3. C' := MemoriaKM(X, K, C')

2.4. Si f(C') < f*, entonces

2.4.1. C:= C'

2.4.2. f*:= f(C')

2.5. Fin Si

3. Fin Repetir

4. Solución C

b) CONFIGURACIÓN INICIAL

De manera similar al algoritmo K-Means, en esta fase se seleccionan K centros aleatoriamente, luego se forman los *Clusters* iniciales asociando el objeto x ∈ X al *Cluster* Cj si el centro x_j es el menos distante al objeto. Finalmente, se calcula el nuevo centro o la media del *Cluster* Cj, (j = 1,..., K) haciendo uso de la siguiente expresión:

$$\underline{x_j} = \frac{1}{|C_j|} \sum x_i \in C_j,$$

Seguidamente, presentamos el algoritmo Inicialización KM:

1. Seleccionar K centros iniciales $\{\overline{x}_i = \text{Random}(X)\}_{i=1,\ldots,K}$
2. Para cada $x \in X$,
2.1. Asignar x a C_j, cuando $j = \text{ArgMin}\{d(x, \overline{x}_i)\}_{i=1,\ldots,K}$
3. Fin Para
4. Calcular los centros $\{\overline{x}_j := \text{Media}(C_j)\}_{j=1,\ldots,K}$
5. Resultado $C = \{C_j\}_{j=1,\ldots,K}$

c) CONSTRUCCIÓN DE SOLUCIONES

Esta fase es una adaptación del algoritmo K-Means con el objetivo de evitar la convergencia a óptimos locales. La adaptación se realiza principalmente sobre la función golosa que asigna los objetos del K-Means, la cual establece que un objeto $x \in X$ se asigna al *Cluster* Cj , si x^j es el centro menos distante al objeto x. Al aplicar el parámetro de relajación sobre la función golosa, se crea un conjunto RCL de posibles *Clusters* a los cuales puede ser reasignado un objeto.

El RCL estará conformado por una vecindad alrededor del *Clúster* más próximo al objeto evaluado. Esta fase se implementa mediante el algoritmo Construcción KM que se presenta a seguir, considerando como datos de entrada el número de *Clústers* K, el conjunto de *Clústers* C = { Cj } donde j=1,2,3,..., K con sus respectivos centros { x_i } donde i=1,2,3,.., K generados en la fase anterior, y el parámetro de relajación α.

Construcción KM (X,K,C, α)

1. Repetir

1.1. Para cada x ∈ X tal que x ∈ Cj para algún j = 1, ..., K

1.1.1. β := Max{d(x, x l): d(x, x l) ≤ d(x, x j)}l =1, ..., K

1.1.2. β := Min{d(x, x l)}l=1, ..., K

1.1.3. RCL := {Ct : d(x, x t) ≤β+α(β+β)}t =1, ..., K

1.1.4. Ct := Random(RCL)

1.1.5. Si t ≠ j

Ct := Ct ∈{x}

Cj:=Cj- {x}

1.1.6. Fin de Si

1.2. Fin de Para

1.3. Recalcular centros { x j: = Media (Cj)} j = 1, ..., K

Media(Cj)}j=1,..., K

2. Hasta que no haya más reasignaciones

3. Resultado C = {Cj}j=1,..., K

En un proceso K-Means un objeto x ε Cj será asignado a otro *Cluster* Cp si la distancia d(x, x_P) es la mínima entre las distancias hacia los *Clusters* y j ≠ p.

En el proceso Construcción KM, los posibles *Clusters* que contendrían al objeto x en análisis son agrupados en un conjunto RCL que contiene un número menor de *Clusters* cuyas distancias de sus centros al objeto x están en un intervalo definido por β, que es el valor máximo de las distancias menores que la distancia a su centro de origen, β que es el mínimo de las distancias menores que la distancia a su centro de origen, regulada linealmente por el parámetro de relajación. Del conjunto RCL será elegido aleatoriamente un *Cluster* al cual será reasignado el objeto x, desde luego retirándolo del *Cluster* al cual correspondía antes de ese proceso.

En los pasos que corresponden al segmento 1.1.4. del algoritmo ConstruccionKM son realizados las operaciones descritas. Esa operación de reasignación será realizada iterativamente para cada objeto x ϵ X. Después de la reasignación de todos los objetos de X en los diferentes *Clusters*, es lógico que el centro haya variado; lo que justifica que, nuevamente, se deban recalcular los centros de cada *Cluster* a través de la media aritmética de los objetos de los respectivos *Clusters*, { x_j = Media(Cj)}j=1,...,K, donde la función Media es definida como :

$$\underline{x}_j = \frac{1}{|C_j|} \sum x_i \in C_j ,$$

Propuesto de esta manera, la fase de construcción evita el determinismo del algoritmo KMeans, teniendo una gran probabilidad de encontrar una mejor solución, aunque también puede encontrar peores. Es por ese motivo que en el procedimiento Grasp, la fase de construcción debe procesarse repetidas veces para obtener una gran variedad de soluciones que pueden ser mejoradas en la fase de búsqueda local.

d) BÚSQUEDA DE LA MEJOR SOLUCIÓN.

La fase de construcción de soluciones genera soluciones que no son necesariamente óptimas, debido a que la búsqueda se realiza de manera aleatoria en un espacio de soluciones restringido por el RCL.

En la fase de búsqueda de la mejor solución, denominada Mejoría KM, se debe alterar la estructura de la solución generada en la fase de construcción con el fin de obtener una mejor solución.

Fränti y Kivijärvi (Fränti y Kivijärvi, [8]) proponen un método de búsqueda local aleatorio para obtener una solución al problema del *Clustering*. El método está basado en un proceso de búsqueda local que altera la estructura de la solución generando un centro aleatorio, seleccionado del conjunto de objetos X, y elimina el *Cluster* con menor error cuadrático; luego, se reasignan los objetos a los *Clusters* que tengan el centro más cercano; y, finalmente, se realiza un proceso de refinamiento de la solución mediante el algoritmo K-Means. El proceso es repetido un número determinado de veces, y el mejor resultado es devuelto como la solución del problema.

El método favorece la eliminación de *Clusters* con menor error cuadrático que probablemente sea producto de alcanzar un óptimo local y evita este inconveniente a través de la generación de un nuevo centro.

Basados en la propuesta de Fränti y Kivijärvi diseñamos la fase MejoriaKM. La idea básica es, dada una solución, ignorar y regenerar *Clusters* según ciertos criterios establecidos. A diferencia de Fränti y Kivijärvi, se propone eliminar el *Cluster* que contiene menos objetos y genera un nuevo centro aleatoriamente dentro del *Cluster* más disperso. Luego, todos los objetos de X son reasignados a los *Clusters* más cercanos; de esta manera, los objetos del *Cluster* eliminado son repartidos entre el resto de *Clusters* y, a la vez, un *Cluster* se forma alrededor del nuevo centro. La configuración del *Cluster* obtenida hasta este punto no es la óptima, por lo cual la solución entra en un proceso de refinamiento, que para el caso sería el mismo de la fase de construcción Grasp. Todo el proceso es repetido iterativamente hasta que no se pueda encontrar otra mejor solución.

La forma heurística con que se modifica la estructura de la solución, obedece a la posibilidad de encontrar una mejoría, asumiendo que los *Clusters* con menor cantidad de elementos y mayor dispersión ocurren porque el algoritmo alcanzó un óptimo local y que se puede encontrar una mejor solución.

En la mejoría descrita, podemos notar dos procesos que se pueden realizar de manera independiente. El primero de ellos es la alteración de la solución; es decir, la eliminación de un *Cluster*, la generación de un nuevo *Cluster* a partir de un centro elegido de manera aleatoria y la agrupación de los objetos

alrededor del nuevo centro y de los existentes. En adelante a este proceso se le denominará de reagrupación.

El segundo es el proceso de refinamiento que corresponde al mismo de ConstruccionKM, lo que da un valor agregado a esta fase en cuanto a evitar alcanzar óptimos locales.

La estructura general del algoritmo MejoriaKM, se presenta a continuación. Considera como datos de entrada conjunto de datos X, el número de *Clusters* K y los *Clusters* generados por la fase construcción C = { Cj } donde j=1,2,3..., K, con respectivos centros x_j = { x_i } donde i=1,2,3,..., K. El proceso entra en una iteración hasta que se alcance una solución estable, realizando las dos etapas descritas anteriormente: la de reagrupación de los elementos de Clusters críticos denominada ReagrupacionKM y el proceso ConstruccionKM usado para el refinamiento de la solución.

MejoriaKM (X,K,C')
1. Repetir
1.1. C := C'
1.2. C' := ReagrupacionKM(X,K,C)
1.3. C' := ConstruccionKM(X,K,C', α)
2. Mientras (f (C') < f(C))
3. Resultado C = $\{C_j\}_{j=1,...,K}$

e) REAGRUPACIÓN

El proceso de Reagrupación KM requiere la identificación de los *Clusters* C^l de menor número de elementos y C^h de mayor dispersión, y un nuevo centro x_r, elegido de manera aleatoria de los elementos pertenecientes a C^h. Si C^l es el *Cluster* con menor número de elementos, entonces *l* está dado por:

$$l = \text{ArgMin}\{|C_j|\}_{j=1,\ldots,k}$$

Para la identificación del *Cluster* de mayor dispersión, se necesita primero el cálculo del error promedio del *Cluster*, el cual está dado por:

$$Ej = \frac{\sum x \in C_j d(x, \bar{x}_j)}{|C_j|}$$

Antes de identificar el *Cluster* más disperso, se describirán dos casos extremos: el primero de un *Cluster* compacto y el segundo de un *Cluster* disperso. El primer caso es un *Cluster* con error promedio bajo y que tiene

una buena cantidad de objetos; esta situación nos da la idea de que el *Cluster* está bastante compacto.

Por el contrario, si tenemos un *Cluster* con error promedio alto y con gran cantidad de elementos, entonces diremos que el *Cluster* está disperso. Es decir, cuanto mayor sea la relación Ej / |Cj|, mayor será la dispersión del *Cluster*, y a menor valor de la relación, menor será la dispersión y por consiguiente mayor será su compactación.

La *Figura II.5* muestra la idea de un *Cluster* compacto y otro disperso. Los *Clusters* dispersos son los que nos interesa reagrupar, por ello se generará un nuevo centro dentro del *Cluster* más disperso con la finalidad de que contribuya a una mejor distribución de los objetos entre los *Clusters*.

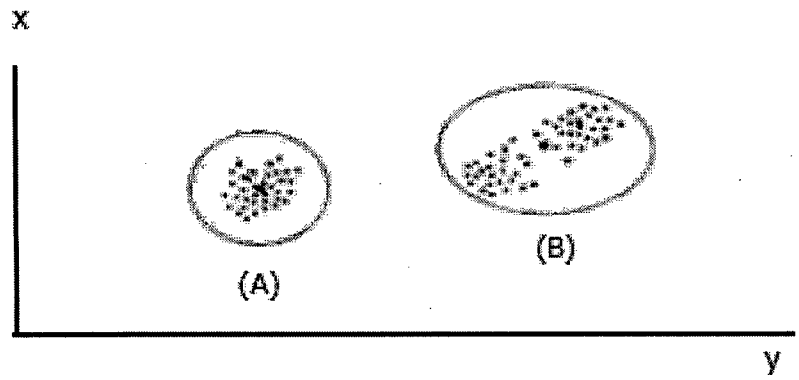


Figura II.5: (A) *Cluster* Compacto (B) *Cluster* Difuso [13].

Entonces, si C^h es el *Cluster* con mayor dispersión, *h* está determinado por:

$$h = \text{ArgMax}\left\{\frac{Ej}{|C_j|}\right\}_{j=1,\ldots,K}$$

El valor del nuevo centro x^r es tomado desde C^h cuya selección se realiza de manera aleatoria y está dado por $x^r = \text{Random}(C^h)$.

Los cálculos descritos anteriormente se llevan a cabo dentro del algoritmo ReagrupacionKM, cuyo pseudocódigo se presenta a continuación, donde los datos de entrada son los mismos de MejoriaKM:

ReagrupacionKM (X, K, C)

1. l := ArgMin$\{|C_j|\}_{j=1,\ldots,K}$
2. h := ArgMax Ej $\left\{ |Cj| \right\}_{j=1,\ldots,K, j \neq l}$
3. $\overline{x}_r$:= Random(C_h)
4. Reemplazar $\overline{x}_l$ por $\overline{x}_r$
5. Para cada x ∈ X // Generando clusters

5.1. Asignar x a C_j, donde j = ArgMin$\{d(x,x_i)_{j=1,\ldots,K}$

6. Fin Para
7. Calcular los centros $\{ \overline{x}_j$:= Media(C_j)$\}_{j=1,\ldots,K}$
8. Resultado C = $\{C_j\}j_{=1,\ldots,K}$

CAPÍTULO III

INVESTIGACIÓN

III.1 METODOLOGÍA DE LA INVESTIGACIÓN

En este capítulo se detalla la metodología de la investigación, su fundamento y tipo de metodología.

III.1.1 TIPO DE INVESTIGACIÓN

El tipo de investigación es sustantiva, explicativa, aplicada, tecnológica, experimental y cuantitativa.

Tipo	Descripción
Sustantiva	Está orientada a describir cómo explicar, la realidad.
Explicativa	Tiene relación causal intenta encontrar las causas del problema.
Aplicada	Busca obtener conocimientos e información sobre hechos o fenómenos para aplicarlos en el enriquecimiento de la ciencia y la solución de los problemas.
Tecnológica	Demuestra la validez de ciertas técnicas en la modificación de los factores que intervienen en una situación problemática.
Experimental	Se encuentra dirigida a la validación de hipótesis que explique las causas de un fenómeno especifico, con el objetivo de sentar las bases para posteriores predicciones.
Correlacional	De acuerdo a los hipótesis de la tesis en base a los datos del patrón de comportamiento (variable independiente) se harán mejoras en el portal de servicios sobre la base de encuestas antes y después de la mejora: se obtendrán datos de tiempo de acceso y satisfacción de los usuarios (variables dependiente) los cuales se cruzan para dar

	validez o no a la hipótesis y determinar la correlación entre estas variables.
Cuantitativa	Se medirá las variables dependiente e independiente las cuales serán relacionadas, para dar validez a la hipótesis planteada.

Cuadro III.1: Tipos de Investigación

III.1.2 FUNDAMENTO METODOLÓGICO DE LA INVESTIGACIÓN

La metodología a seguir es el método inductivo, ya que el reconocimiento del comportamiento de usuarios de un Portal Web está regido por el agrupamiento y el reconocimiento de características y comportamientos comunes del grupo en un periodo en particular y que puede ser aplicado de manera general para predecir cómo va a ser el comportamiento de usuarios en otros periodos, a continuación se describe las dos perspectivas: la tecnológica y la social.

Perspectiva	Descripción
Tecnológica	Se utilizará el Data *Minig* como una herramienta tecnológica para el reconocimiento de patrones de usuarios en un Portal de Servicio con ello se busca mejorar los accesos hacia los Portales de Servicios.
Sociedad	Se hará un análisis del comportamiento de los usuarios del Portal Web de la SUNAT, éstos son contribuyentes que están vinculados a la SUNAT así mismo utilizan el Portal Web para acceder hacia los servicios que esta brinda.

Cuadro III.2: Tipos de Perspectivas.

III.1.3 ESTRUCTURA METODOLÓGICA

En la tesis no se va detallar y redundar técnicas sobre la elaboración de una metodología que envuelva la elaboración de la investigación, sin embargo en la *Figura III.1* se presenta un esquema de la metodología de investigación.

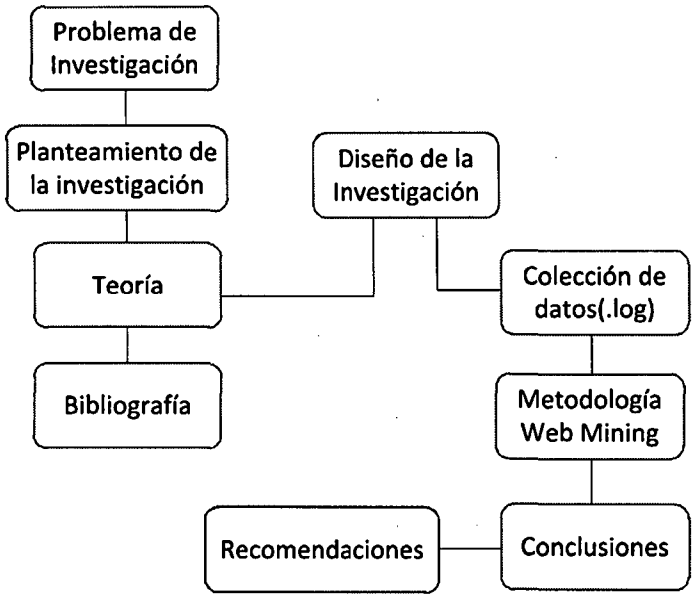


Figura III.1: Metodología de la Investigación.

III.2 DISEÑO DE LA INVESTIGACIÓN

El diseño es experimental.

Es experimental, porque la variable independiente "*Patrón de Comportamiento*" va a ser medida, utilizando la metodología Web Mining, utilizando la técnica de *Clustering* se determinarán grupos con características comunes, estas características están determinadas por los accesos de los usuarios a los diferentes módulos del Portal Web, para analizar el comportamiento se define como variables dependientes:

- Menor tiempo en el acceso al módulo o servicio deseado.
- Grado de satisfacción de los usuarios, con los servicios ofrecidos en el Portal Web.

Para ello se utilizará la técnica de la encuesta antes y después de la mejora del Portal Web, de esta manera poder explicar la correlación que hay entre la variable *"Patrón de Comportamiento"* y con las variables *"Tiempos de Acceso"* al Portal Web y el *"Grado de satisfacción de los Usuarios"*; para llegar a ello se realizará una secuencia de controles que utiliza la correlación.

1º Se observa que el tiempo de acceso de los usuarios hacia el servicio deseado dentro del Portal Web de la SUNAT mejora por la facilidad que se le brinda (accesos directos en una ventana emergente, sobre la base al patrón al que pertenece).

Sean las variables:

Tpi = Tiempo promedio de acceso en el Portal Web sin mejora.

Tpf = Tiempo promedio de acceso en el Portal Web con mejora.

2º Se observa que el grado de satisfacción de los usuarios que acceden hacia los servicios del Portal Web de la SUNAT mejora, esto se logra ofreciendo los servicios deseados, sobre la base a patrones encontrados.

Spi = Grado de Satisfacción en el Portal Web sin mejora.

Spf = Grado de Satisfacción en el Portal Web con mejora.

3º Se compara:

Tpi : Tpf.

Spi : Spf.

4º A través de encuestas se determina la relación(R) entre la variable independiente: X (Patrón de Comportamiento) y las variables dependientes Y (Tiempo de Acceso hacia los servicios del Portal Web y el Grado de Satisfacción): es positiva si la probabilidad de aprobación sobre la optimización del Portal Web (Pi) es mayor a cuando no se hizo la optimización (Pj) en caso contrario será negativa.

Según: $R = Pi - Pj$

Si $R > 0$; la hipótesis es verdadera

Si $R \leq 0$; la hipótesis es falsa

III.2.1 OBJETO DE LA INVESTIGACIÓN

El objeto de la investigación o unidad de análisis será la sesión de un usuario de un Portal Web, para ello se analiza: los tiempos de acceso, la cantidad de visitas, los tiempos de sesión, los tipos de usuarios así como la realización de las transacciones en el Portal Web.

Determinar el reconocimiento de patrones de comportamiento de usuarios en un Portal Web de servicios permitiendo con ésto la disminución del tiempo de acceso hacia los servicios del Portal Web

III.2.2 POBLACIÓN

Son las sesiones de los usuarios de un Portal Web registradas en el archivo log para un periodo determinado, este periodo se define sobre la base del objeto de estudio y el comportamiento cíclico de la empresa, puede ser mensual, trimestral (productos estacionales), semestral, anual, etc. Para el estudio presente estudio se toma el periodo anual ya que se quiere saber cuál es el comportamiento de los contribuyentes para un ciclo o renta anual de la SUNAT.

III.2.3 TAMAÑO DE LA POBLACIÓN

El tamaño de la población está determinado por las sesiones registradas en los archivos log para el periodo anual se toma el año 2010 como referencia.

Mes de Registro Log	Cantidad de Sesiones
2010 – 01	129526
2010 – 02	131834
2010 – 03	157561
2010 – 04	166022
2010 – 05	139324

Mes de Registro Log	Cantidad de Sesiones
2010 – 06	128463
2010 – 07	126446
2010 – 08	135576
2010 – 09	147562
2010 – 10	154927
2010 – 11	140553
2010 – 12	133802
Total	**1691596**

Cuadro III.3: Cantidad de sesiones por mes, del año 2010.

III.2.4 TAMAÑO DE LA MUESTRA

En el estudio se determina la muestra representativa de la población debido a que el costo de procesamiento de toda la población hace que el algoritmo sea lento, para tener una mayor performance de la misma se calcula una muestra representativa y sin perder la confiabilidad del mismo. Se utiliza el teorema de Chernoff Bound *(Balaji, [9])* para determinar la muestra, ver *Fig. III.2.*

$$
\begin{aligned}
Pr\left[Y \geq K(1+\epsilon)\right] + \quad & Pr \quad \left[Y \leq K(1-\epsilon)\right] \\
& \leq \quad 0 + e^{K\epsilon^2/2} \\
& \leq \quad e^{K\epsilon^2/2} \leq \delta \\
\Rightarrow K \quad & \geq \quad \frac{2}{\epsilon^2} \log\left(\frac{1}{\delta}\right)
\end{aligned}
$$

Figura III.2: Tamaño de la muestra según el Modelo Chernoff Bounds *(Balaji, [9])*.

Dónde:

k : Tamaño de la muestra.

1-ε : Nivel de confianza.

1-δ : Probabilidad.

ς : Grado de error.

Para diferentes valores de ε y de δ, podemos encontrar valor para el tamaño de la muestra: ver *Cuadro III.4*.

ϵ	δ	Size of sample
0.01	0.01	105967
0.01	0.02	92104
0.01	0.05	73778
0.01	0.1	59915
0.02	0.01	26492
0.02	0.02	23026
0.02	0.05	18445
0.02	0.1	14979
0.05	0.01	4239
0.05	0.02	3685
0.05	0.05	2952
0.05	0.1	2397
0.1	0.01	1060
0.1	0.02	922
0.1	0.05	738
0.1	0.1	600

Cuadro III.4: Tamaños de la muestra, para diferentes valores de ε y δ *(Balaji, [9])*.

Por ejemplo: si queremos que el 90% de los cálculos del algoritmos se encuentre dentro del grado de error declarado (probabilidad) y con un nivel de confianza del 90% tenemos que hacer el experimento, al menos con 600 muestras, ver el *Cuadro III.4* para ε = 0.1 y δ = 0.1.

Para el presente estudio se toma: una probabilidad del 95% y con un 95% de grados de confianza, entonces ε = 0.05 y δ = 0.05 y para este valor se observa en el *Cuadro III.4* que el tamaño de la muestra es mayor o igual a 2952:

$k \geq 2952$

Nuestra unidad muestral son las sesiones: cada sesión es distinta y todas tienen la misma probabilidad de ser elegidas, es por eso que el método apropiado para la selección de la muestra es el método aleatorio, por lo que se utiliza esta técnica para seleccionar la muestra, en el *Capítulo VI Procesos de Sesión*, *Sección VI.4.2.3 Selección de la Muestra Representativa*, se detalla la selección de la muestra.

III.2.5 VARIABLES DE LA INVESTIGACIÓN

A continuación se detalla las variables dependientes e independientes sujeta al modelo.

III.2.5.1 VARIABLES INDEPENDIENTES DEL MODELO

La investigación está sujeta sobre la base de las siguientes variables independientes:

X_1: IP del usuario.
X_2: La fecha de visita al Portal Web.
X_3: La hora de visita al Portal Web.
X_4: La cantidad de visitas a un servicio Web.
X_5: Servicio elegido.

III.2.5.2 VARIABLE DEPENDIENTE DEL MODELO

Para definir la variable dependiente la de investigación se centra en el porcentaje de visitas de un usuario a los diferentes módulos del Portal Web en un periodo determinado, la variable dependiente es:

Y1: Porcentaje de visitas al Portal Web.

III.2.5.3 PARÁMETROS DEL MODELO

Entre los parámetros del modelo tenemos:

P_1: Tiempo de sesión de usuario.
P_2: Cantidad de *Cluster*.

Se coloca claramente cuáles son las variables independientes y dependientes.

El diseño de la investigación comprende tanto la definición del objeto de investigación, las sesiones de los usuarios como la población a estudiar conjuntamente con la selección de la muestra. Ésto servirá para realizar los experimentos necesarios para recolectar datos y realizar las operaciones con las variables del modelo (X_1, X_2, X_3, X_4, X_5 e Y_1) que permiten realizar juicios sobre la validez de la hipótesis.

III.3 RECOLECCIÓN Y ELABORACIÓN DE DATOS

III.3.1 INSTRUMENTOS DE RECOLECCIÓN

El instrumento que se utilizó en la investigación fue el cuestionario, el cual fue aplicado a los contribuyentes que hacen uso de Portal de Servicios de la SUNAT.

La intención de hacer una encuesta es medir los tiempos de acceso al Portal Web así como los niveles de satisfacción de los usuarios antes y después de haber efectuado la mejora en el Portal Web, vale resaltar que esta mejora se concretiza con la personalización del portal para cada *Cluster* encontrado sobre la base a los patrones de comportamiento de los usuarios; obtenidos en el proceso experimental.

III.3.2 TÉCNICAS DE LA INVESTIGACIÓN

Para la recolección de los datos necesarios para poder contrastar nuestra hipótesis, es decir, para analizar, comparar y correlacionar las variables de estudio, se utilizarán las siguientes técnicas de estudio.

TECNICA	INSTRUMENTO	MOMENTO	CONTENIDO	SUJETO
Encuesta.	Cuestionario.	Único.	Los servicios administrativos del Portal Web de la SUNAT.	Ciudadanos que son contribuyentes.

Cuadro III.5 Técnica de Investigación.

CAPÍTULO IV

MODELO DE SOLUCIÓN

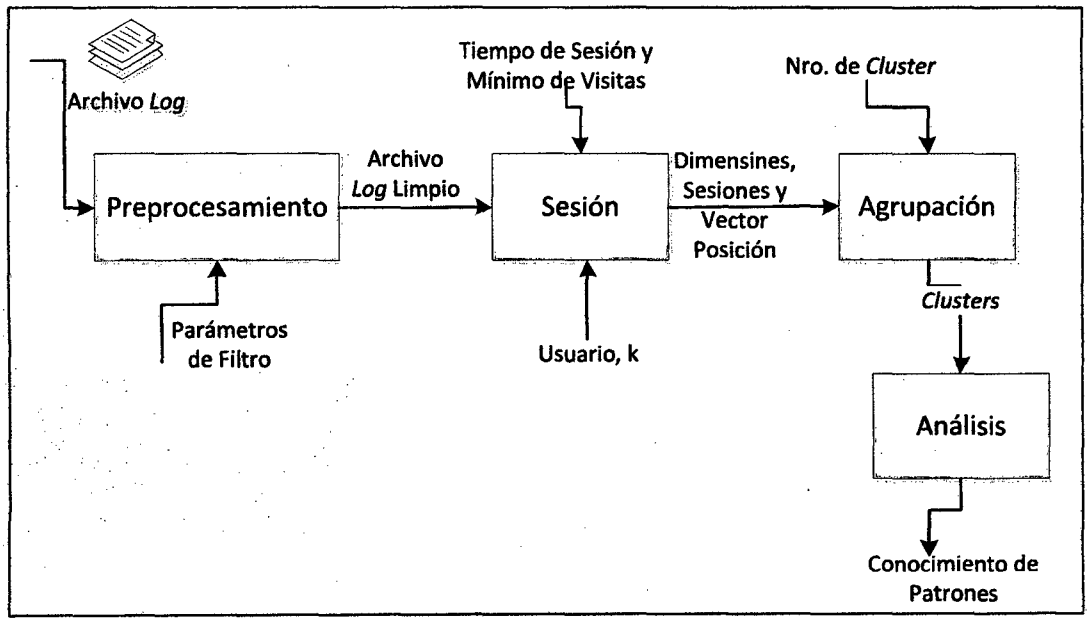


Figura IV.1: Modelo Solución.

IV.1 DESCRIPCIÓN DEL MODELO DE SOLUCIÓN

Este modelo está basado en el modelo *Frequent Pattern Mining in Web Log DataHungary* según *Renata Ivancsy, Istvan Vajk* (Renata & Istvan [10]).

El modelo solución consiste en cuatro procesos: Pre-procesamiento, Sesión, Agrupamiento y Análisis. La entrada al modelo solución es un archivo *log* donde se registran los sucesos realizadas por las visitas de los usuarios al Portal Web, estos sucesos a la vez están conformados por un conjunto de atributos como son: la hora, IP del usuario, URL, etc; la salida del modelo solución es un gráfico estadístico el cual se determina a través del análisis de los grupos de sesiones de usuarios, el proceso general del modelo solución tiene como finalidad extraer, generar y representar información que se encuentra oculta en los archivos *log* de los servidores Web, siguiendo la metodología de Web Mining.

IV.2 PROCESOS DE MODELO SOLUCIÓN

Se tiene cuatro grandes procesos: Pre-procesamiento, Sesión, Agrupamiento y Análisis, cada uno estos procesos serán tratados en capítulos separados para mayor detalle.

IV.2.1 PROCESO DE PRE-PROCESAMIENTO

Este proceso tiene como objetivo la limpieza del archivo log.

Tiene como entrada al archivo *log*, en este archivo se registran los sucesos que vienen hacer las visitas realizadas por los usuarios, estos sucesos están conformados por un conjunto de atributos como son: la hora, IP del usuario, URL del servicio, etc.

Cuando un usuario solicita una página, ese pedido se graba en el archivo de *log*, pero además, si la página posee imágenes, se guardará una línea por cada imagen solicitada y así sucede con otros recursos: scripts, estilos, etc. En la mayoría de los casos, estos registros adicionales almacenados en los archivos *log* no son necesarios para la tarea de identificación de hábitos de navegación de los usuarios. Se podrían filtrar todos los registros cuyos archivos solicitados tengan las extensiones *jpg, jpeg, gif, js, css, swf, avi, mov*, etc

IV.2.2 PROCESO DE SESIÓN

El objetivo de este proceso es identificar las sesiones de los usuarios, para ello se necesita dividir las distintas peticiones realizadas por los usuarios en una o más sesiones.

Tiene como entrada al archivo *log* limpio (*archlim*) y además de un parámetro de umbral: tiempo que puede durar una sesión, para lo cual se han realizado investigaciones que buscan encontrar el valor que mejor divida las sesiones de los usuarios estableciendo un valor óptimo en forma empírica, sobre esto tenemos la investigación realizado por Catleedge y Pitkow, quienes determinaron de manera empírica el valor de 25.5 minutos como tiempo de sesión máxima (Catleedge y Pitkow [11]). Sin embargo, generalmente, es utilizado el valor de 30 minutos como valor máximo entre dos peticiones de una misma sesión. Por lo que en esta investigación el parámetro tendrá el valor de 30 minutos, este viene hacer el umbral para poder generar las sesiones es decir si el tiempo entre el primer y último suceso es menor a este umbral, entonces se considera que todos los sucesos dentro del tiempo, establecido, pertenecen a la misma sesión.

Para la formación de sesiones se utiliza generalmente un tiempo máximo (umbral) entre peticiones de un mismo usuario, de modo que, si dos peticiones de un usuario se realizan con un intervalo de tiempo menor al máximo, las dos peticiones son consideradas como parte de la misma

sesión, así también si dos peticiones de un usuario se realizan con un intervalo de tiempo mayor al umbral, las dos peticiones corresponden a sesiones distintas; la primera es la última petición de la sesión y la otra es la primera de una nueva sesión.

Este proceso tiene como salida al conjunto de Dimensiones, Sesiones y Vector Posición.

IV.2.3 PROCESO DE AGRUPAMIENTO

El objetivo de este proceso es descubrir los patrones, a través de la técnica del agrupamiento o *Clustering*.

Este proceso tiene como entrada a las sesiones y como salida a los grupos o *Cluster* de sesiones, cada grupo está constituido por sesiones que tiene una característica en común así como también pertenecen a un mismo centroide.

La técnica de *Cluster* es utilizada para juntar sesiones que poseen características similares, en este caso las sesiones estarán agrupadas teniendo en cuenta la frecuencia con la cual han sido visitados los diferentes servicios de un Portal Web: el agrupamiento de sesiones de los usuarios con características comunes ayuda a interpretar el comportamiento de los

usuarios que navegan en un Portal Web, y es justamente esta característica la que se quiere analizar.

IV.2.4 PROCESO DE ANÁLISIS

El objetivo de este proceso es analizar a los patrones encontrados, es decir, de los grupos encontrados en la fase de agrupamiento, para ésto se contabiliza la cantidad de peticiones en una sesión de un usuario, luego se calcula los porcentajes de cada servicio visitado con respecto al total de peticiones de una sesión de un usuario, con esto se busca conocer cómo ha sido el comportamiento de los usuarios.

Este proceso tiene como entrada a los grupos de sesiones (sesiones) estos grupos de sesiones tienen características comunes en este caso se han agrupado las sesiones teniendo en cuenta su frecuencia de visitas a los diferentes servicios del Portal Web como se mencionó en el proceso de agrupamiento, como salida de este proceso tenemos la información y el conocimiento del comportamiento de los usuarios que están representado en gráficos estadísticos, estos gráficos estadísticos plasman los porcentaje de visitas a los diferentes servicios de un Portal Web por usuario en cada sesión.

CAPÍTULO V

PRE–PROCESAMIENTO

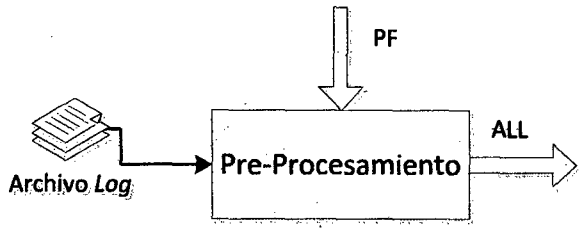


Figura V.1: Proceso de Pre-Procesamiento.

El objetivo de este proceso es convertir el archivo *log*[6] en un archivo limpio que solo contenga registros que servirán para hacer análisis de *Clustering*: los demás registros se descartan.

Cuando un usuario solicita una página, este pedido se graba en el archivo *log*, pero además, si la página posee imágenes, se guarda una línea

[6] Del archivo *log* se trató en el capítulo II, Sección 1 y punto 9 (Marco Teórico conceptual: Archivo *log*)

adicional por cada imagen solicitada. Este comportamiento ocurre con cualquier recurso que esté referenciado desde la página solicitada originalmente, como pueden ser archivos con *scripts* JavaScript, hojas de estilo, animaciones Flash, videos, etc. Para la identificación de patrones de comportamiento no interesan estos registros adicionales. Por ello es necesario filtrar todos los registros del *log* donde los recursos solicitados pertenezcan a estos tipos, para este filtro se consideran parámetros que básicamente dependen del objeto de estudio.

V.1 PARÁMETROS DE FILTRO (PF)

Para obtener el archivo *log* limpio, necesitamos definir con qué parámetros filtrar y para esto tenemos 3 parámetros:

1. **URL–Alias:** Que la URL contenga "*Alias*" con lo cual se identifica a los servicios del Portal Web, ya que todos los servicios del portal libre de la SUNAT, están gestionados por un Servlet y éste tiene un "*Alias*".

   Ejemplo:

   - "http://www.sunat.gob.pe/ol-ti-itinsrucsol/**iruc001Alias**".
   - "http://www.sunat.gob.pe/cl-ti-itmrconsruc/**jcrS00Alias**".

2. **Método GET:** En el registro *log*, el parámetro método contenga *"GET"*, ya que es el método utilizado para las peticiones de los servicios que brinda el Portal Web.

   Ejemplo:

   - "**GET** /cl-ti-itmrconsruc/frameCriterioBusqueda.jsp HTTP/1.1".
   - "**GET** /cl-ti-itmrconsruc/jcrS00Alias HTTP/1.1".

3. **Recurso .jsp** Que el recurso contenga "*.jsp*" por lo que los servicios del Portal están representados por Java Server Page (JSP).

   Ejemplo:

   - "GET **/cl-at-itageban/buscarbcosuc.jsp** HTTP/1.1".
   - "GET **/cl-ti-itmrconsruc/frameCriterioBusqueda.jsp** HTTP/1.1".

Por tanto considerando estos parámetros se halla el algoritmo de limpieza y del registro del nuevo archivo *log* limpio.

V.2 PROCESO DE PRE-PROCESAMIENTO

Este proceso consiste en de los siguientes pasos:

1. La variable de entrada es el archivo *log* (*archlog*) está representa en la dirección donde se encuentra físicamente el *archlog*.
2. Se lee línea por línea el *archlog* desde la primera línea hasta el final.
3. Para cada línea se verifica que la petición **no sea nula**, que el campo de la dirección (URL) contenga la palabra reservada "**Alias**", que el campo de recurso tenga la extensión "**.jsp**" y que el campo de la petición contenga "**GET**".
4. A todos los registros que cumplan con el filtro anterior se capturan IP, Fecha, Hora, Método, Recurso y URL y se registra en la tabla *t1archlogfiltrado*, con campos adicionales Periodo, Sub-Periodo, Fecha del sistema y Usuario que actualiza.
5. Hacer ésto hasta terminar con los registros del *archlog*.

V.3 DIAGRAMA DE FLUJO

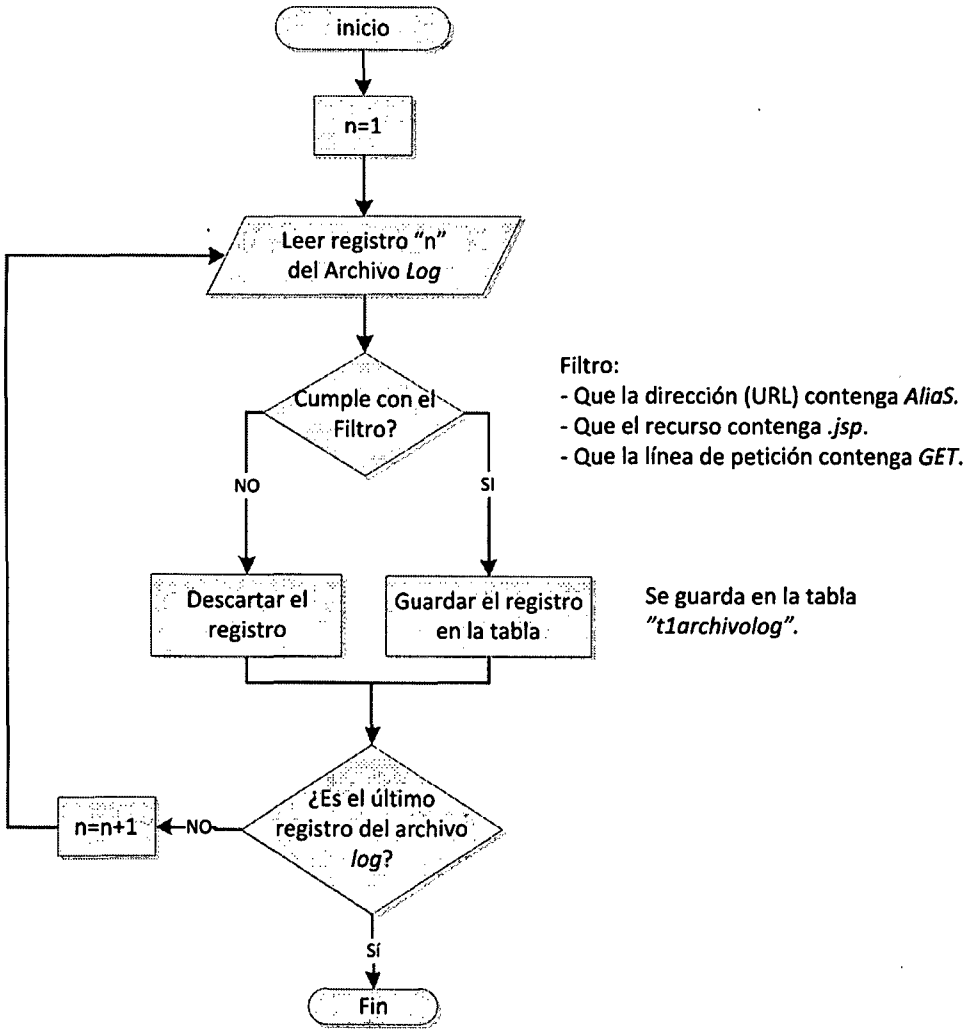


Figura V.2: Diagrama de flujo para la limpieza del archivo *Log*.

V.4 PSEUDOCÓDIGO

1. leer archlog
2. Instancias las conexiones a mydb.t1archlogfiltrado
3. n=1
4. linea=archlog.radline(n)
5. Mientras linea != null entonces
6. st=linea.pasear (" ")
7. ipuser =st.NexTokenizer()
8. idmaquina =st.NexTokenizer()
9. iduser =st.NexTokenizer()
10. fechora1 = st.NexTokenizer()
11. fechora2 = st.NexTokenizer()
12. peticion1 = st.NexTokenizer()
13. peticion2 = st.NexTokenizer()
14. peticion3 = st.NexTokenizer()
15. estado = st.NexTokenizer()
16. tamano = st.NexTokenizer()
17. urlpage = st.NexTokenizer()
18. infouser1 = st.NexTokenizer()
19. Si (urlpage contiene "Alias" y parametro2 contiene ".jsp" y peticion1 contiene "GET") entonces

20. insert into t1archlogfiltrado values(periodo, subperiodo, ipusuario, fecha, hora, método, recurso, url)

21. fin si

22. n=n+1

23. si existe archlog.radline(n) entoces

24. ir al punto 3

25. else

26. fin

27. sin si

V.5 ARCHIVO *LOG* LIMPIO (ALL)

Este archivo *log* limpio contiene únicamente los datos que sirven para hacer el análisis y determinar los patrones. Se registrarán en la tabla *t1archloglimpio*, lo que servirá como entrada para el siguiente proceso de Sesión.

CAPÍTULO VI

PROCESO DE SESIÓN

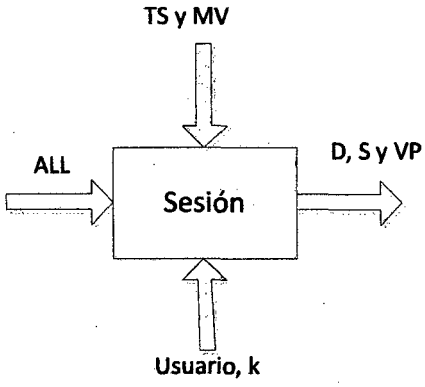


Figura VI.1: Proceso de Sesión.

El objetivo de este proceso es encontrar y registrar las sesiones de los usuarios, tomando en cuenta los parámetros y los umbrales definidos.

VI.1 IDENTIFICACIÓN DE LOS USUARIOS (U)

Luego de la limpieza de los *log*, se debe identificar a los distintos usuarios. Existen distintos métodos para identificar a los usuarios, cada uno de los cuales poseen sus ventajas y desventajas. Un método de identificar a los usuarios es mediante la utilización de los *Cookies*. La W3C (www, [1]) define a una *Cookie* como:

"*Datos enviados por el servidor web al cliente, el cual los almacena localmente y los envía nuevamente al servidor en los sucesivos pedidos*".

En otras palabras, una *Cookie* es simplemente una cabecera *http* que consiste de una cadena de texto. Uno de los problemas del uso de las *Cookies* para la identificación de los usuarios, es que los usuarios puedan deshabilitar el soporte para *Cookies* en sus navegadores, con lo cual ya no se almacenan las *Cookies* en el cliente y no se tendría la posibilidad de identificarlo en las sucesivas visitas. El otro problema es que las *Cookies* son almacenadas en la computadora del usuario. Otra forma de identificar a los usuarios es mediante la utilización de Identidad. Identidad es un protocolo de identificación especificado en el RFC 1413 que permite identificar a los usuarios de una conexión TCP particular. Dado un par de números de puertos TCP, devuelve una cadena de texto que identifica al dueño de esa conexión en el sistema del servidor. El problema del uso de "identidad" para

la identificación de los usuarios reside en que el cliente debe estar configurando para el soporte de identidad.

El otro método y la que se emplea en nuestro modelo de solución, es la identificación de usuarios mediante su dirección IP, en cada línea del archivo de log se almacena la dirección IP del cliente que realizó el pedido. Cada IP diferente viene a ser un usuario diferente.

Usuario = Dirección IP

VI.2 TIEMPO DE SESIÓN (TS)

Luego de la identificación de los usuarios, se deben identificar sus sesiones. Para ello se necesita dividir las distintas peticiones realizadas por un mismo usuario en una o más sesiones. Debido a que las peticiones a los recursos de otros servidores Web no están disponibles, es difícil saber cuándo un usuario abandona el sitio Web. Para la formación de sesiones se utilizan generalmente un tiempo máximo entre sucesivas peticiones, de modo que, si dos peticiones consecutivas de un usuario se realizan con un intervalo de tiempo menor al umbral, las dos peticiones son consideradas como parte de la misma sesión. Si dos peticiones consecutivas se realizan con un intervalo de tiempo mayor al umbral, las dos peticiones corresponde a

sesiones distintas; la primera es la última petición de la sesión y la otra es la primera de una nueva sesión. Se debe seleccionar un tiempo máximo entre peticiones lo que estableceremos como nuestro umbral, para lo cual se han realizado investigaciones que buscan encontrar el valor que mejor divida las sesiones de los usuarios estableciendo un valor óptimo en forma empírica, sobre esto tenemos la investigación realizado por Catleedge y Pitkow, quienes determinaron de manera empírica el valor de 25.5 minutos como tiempo de sesión máxima (Catleedge y Pitkow [11]). Sin embargo, generalmente, es utilizado el valor de 30 minutos como valor máximo entre dos peticiones de una misma sesión. Por lo que en esta investigación el parámetro tendrá el valor de 30 minutos, este viene hacer el umbral para poder generar las sesiones es decir si el tiempo entre el primer y último suceso es menor a este umbral, entonces se considera que todos los sucesos dentro del tiempo, establecido, pertenecen a la misma sesión.

TS = 30'

Para la formación de sesiones se utiliza generalmente un tiempo máximo (umbral) entre peticiones de un mismo usuario, de modo que, si dos peticiones de un usuario se realizan con un intervalo de tiempo menor al máximo, las dos peticiones son consideradas como parte de la misma sesión, así también si dos peticiones de un usuario se realizan con un intervalo de tiempo mayor al umbral, las dos peticiones corresponden a

sesiones distintas; la primera es la última petición de la sesión y la otra es la primera de una nueva sesión.

VI.3 UMBRAL MÍNIMO DE VISITAS (MV)

Otro parámetro a considerar es el número mínimo de visitas, para nuestro estudio consideramos como sesión válida aquella que registre como mínimo la visita de tres servicios pudiendo ser el mismo servicio.

Nro. MV >= 3

A continuación se presenta el algoritmo que consigna estos parámetros como filtro para registrar las sesiones, cantidad de visitas y vector posición, pero antes se deben registrar las dimensiones.

VI.4 PROCESO DE SESIÓN

El objetivo de este proceso es primero determinar las dimensiones luego se determinarán las sesiones y el vector posición.

VI.4.1 PROCESO PARA DETERMINAR LAS DIMENSIONES (D)

Las dimensiones son todos los servicios libres que ofrece la SUNAT a través de su Portal Web, estas dimensiones pueden aumentar con el tiempo y también pueden darse de baja, por lo que se tomará en cuenta este dinamismo.

Este proceso para el registro de las dimensiones o módulos que tiene el Portal Web consta de los siguientes pasos.

1. Se selecciona todas las URL distintas de la tabla *t1archloglimpio*.
2. Todas las URL distintas se guarda en un cursor.
3. Se lee uno por uno las URL del cursor y se compara si ya existe dicha URL en la tabla *t2dimesion*.
    a. Si Existe la URL se descarta.
    b. Si No existe la URL se guarda en la tabla *t2dimension*.

4. Se verifica si es el último URL del cursor.
    a. Si es el último terminar, ir al paso 5.
    b. Si no es el último pasar al siguiente valor del cursor, ir al paso 3.
5. Fin.

VI.4.1.1 DIAGRAMA DE FLUJO PARA DETERMINAR LAS DIMENSIONES

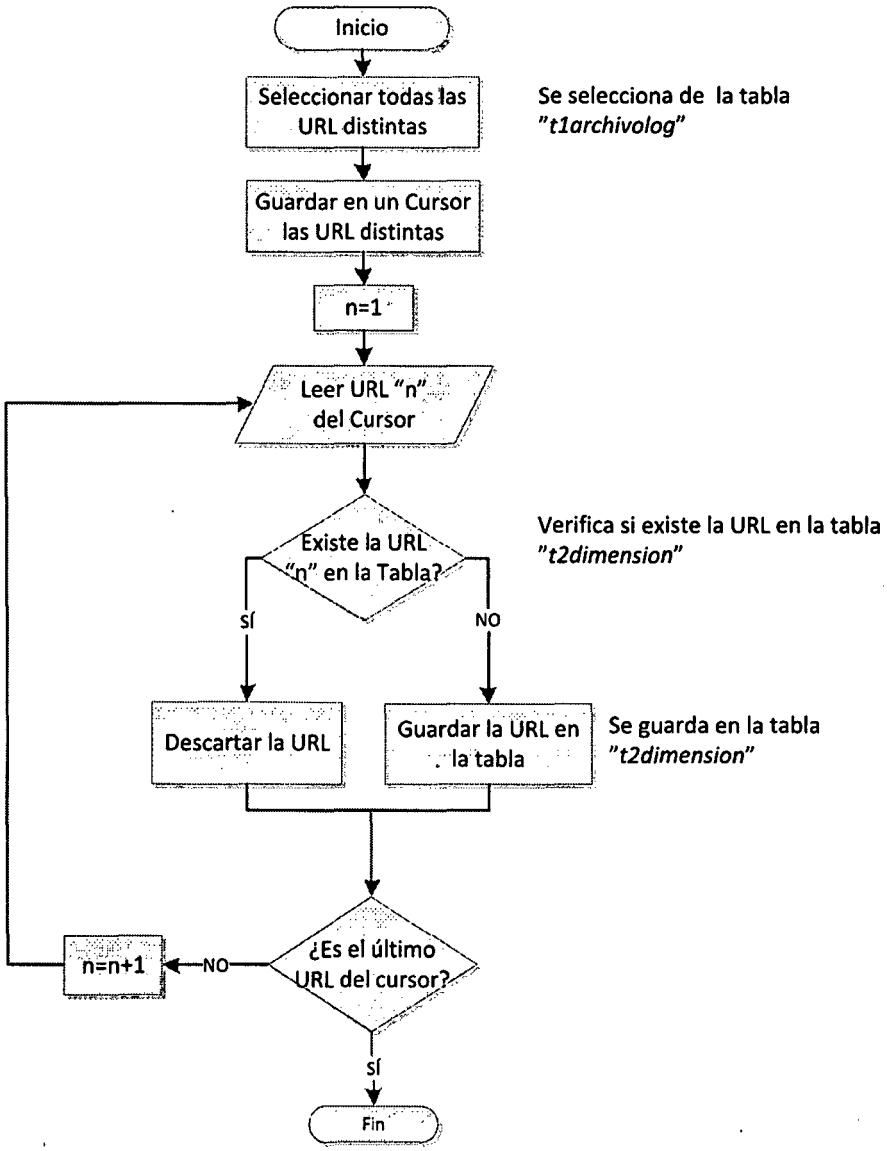


Figura VI.2: Diagrama de Flujo del sub-proceso para determinar las dimensiones.

VI.4.1.2 PSEUDOCÓDIGO PARA DETERMINAR LAS DIMENSIONES

1. Instancias las conexiones a mydb.t1archlogfiltrado; mydb.t2dimension.
2. Asignar cursor **dimen** for select distinct t1_urlpage from mydb.t1archlogfiltrado.

   2.1 abrir el cursor **dimen.**

   2.2 mientras exista elemento del cursor hacer.
3. asignar al urlpage ← **dimen.**
4. si noexiste(SELECT para urlpage) entonces.
5. insert into t2dimension values(periodo, valor dimensión).
6. fin si.
7. Fin mientras.

VI.4.2 PROCESO PARA DETERMINAR LAS SESIONES (S) Y EL VECTOR POSICIÓN (VP)

A continuación la serie de pasos para el proceso de determinación de sesiones y vector posición.

1. Se selecciona el IP, Fecha, Hora y la URL de la tabla *t1archlogfiltro* ordenando por IP, Fecha y Hora.
2. Guardar la selección en un cursor.
3. Inicializan los parámetros; IP_ini, umb y n.

a. IP_ini : 0.0.0.0 ; Un IP como auxiliar, que no exista en el archivo *log*.

b. umb : 0 ; Cantidad de visitas por sesión de un usuario.

c. n : 0 ; Contador

4. Se lee el elemento "n" del cursor, n es el valor de la posición de los registros del cursor.
5. Comparar la IP del registro "n" con el IP auxiliar (IP del registro "n-1"). Si son iguales pasar a punto 6 de lo contrario al punto 7.
6. Se captura la fecha y hora de este registro y se calcula la diferencia con la fecha y hora del registro n-1, y luego se evalúa que la diferencia no sea mayor a 30 minutos; si son iguales pasar al punto 7 de lo contrario al punto 8.
7. Compara el valor del umbral (umb) que sea mayor o igual a 3, si es verdadero pasar al punto 9 de lo contrario al punto 10.
8. Se aumenta la cantidad de visitas en una unidad, ir al punto 4.

$$\mathbf{umb = umb + 1}$$

9. Se guardar la sesión con los valores de IP, Periodo, Fecha y Hora de Inicio, Fecha y Hora Fin y Total de Visitas. Y también se registra la cantidad de visitas con los valores del Periodo, Número de Vector, Total de Visitas, Número de Dimensión, Cantidad de Visitas a la Dimensión y Porcentaje de Visita.

10.Se inicializan los parámetros con los nuevos valores.

```
IP_ini=IP_n
fec_hora_ini=fec_hora_cur
fec_hora_fin=fec_hora_cur
umb=1
```

11.Verificar si es el último registro, si es así terminar (12), de lo contrario ir al punto 4.

12.Fin

VI.4.2.1 DIAGRAMA DE FLUJO PARA DETERMINAR LAS SESIONES Y EL VECTOR POSICIÓN

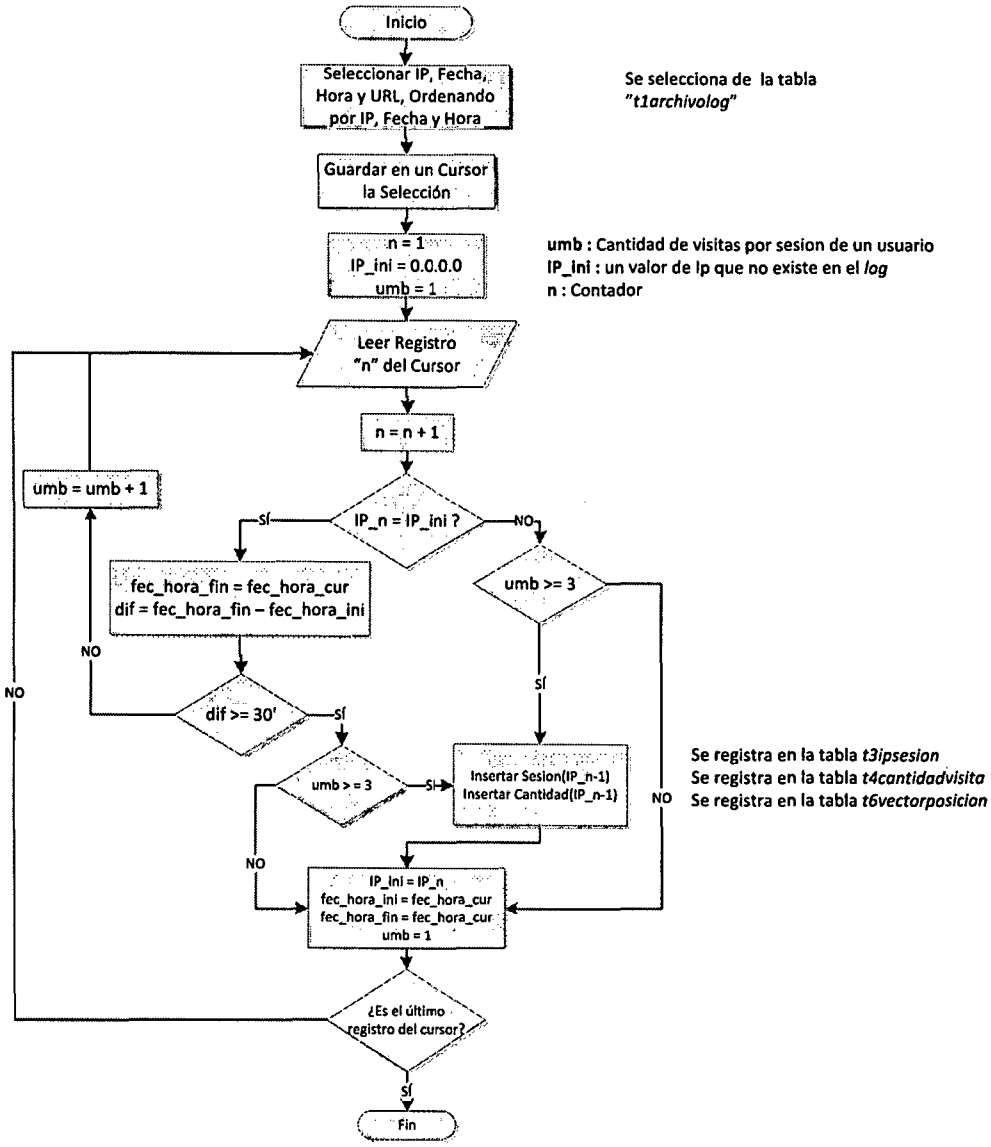


Figura VI.3: Diagrama de Flujo del Sub-proceso para determinar las Sesiones.

VI.4.2.2 PSEUDOCÓDIGO PARA DETERMINAR LAS SESIONES Y EL VECTOR POSICIÓN

1. Instancias las conexiones a mydb.t1archlogfiltrado; mydb.t2dimension, mydb.t3ipsesion, mydb.t6vectorposicion.
2. Asignar cursor **sesión_vector** for select distinct t1_urlpage from mydb.t1archlogfiltrado.
3. ip_diferente = 0.0.0.0 ; n=1 ; umb = 1.
    a. abrir el cursor **sesión_vector.**
    b. mientras exista elemento del cursor **sesión_vector** hacer.
4. Asignar al (ipusuario, fecha, hora, turlpage) ← **sesión_vector.**
5. Si ipusuario = ip_diferente hacer.
6. fecha_hora_fin = fecha_hora_cur;
   dif = fecha_hora_fin – fecha_hora_ini.
7. Si **dif** >=30'.

7.1 Si **umb** >= 3.

7.1.1 Registrar sesión de IP_n-1 en la tabla t3ipsesion.

7.1.2 Registrar la cantidad de visitas de la IP_n-1 en la tabla t4cantidadvisita.

7.1.3 Asignar IP_ini=IP_n.

7.1.4 Asignar fec_hora_ini=fec_hora_cur.

7.1.5 Asignar fec_hora_fin=fec_hora_cur.

7.1.6 Asignar umb=1.

7.2Fin sí.

7.3Si umb < 3.

7.3.1 Asignar IP_ini=IP_n.

7.3.2 Asignar fec_hora_ini=fec_hora_cur.

7.3.3 Asignar fec_hora_fin=fec_hora_cur.

7.3.4 Asignar umb=1.

7.4Fin sí.

7.5Si es último registro del cursor.

7.5.1 Fin.

7.6Fin sí.

7.7Si no es el último registro del cursor.

7.7.1 Ir al punto 4.

7.8Fin sí.

8. Fin sí.

VI.4.3 SELECCIÓN DE LA MUESTRA REPRESENTATIVA

El objetivo de este sub-proceso es determinar la muestra representativa a partir del valor de "*k*", donde "*k*" es el tamaño de la muestra calculada en el *Capítulo III* Investigación y en la *Sección III.2.3* Tamaño de la Población, seleccionaremos de manera aleatoria "*k*" elementos del grupo de sesiones y estos elementos se registran en una tabla para hacer el experimento.

VI.4.3.1 DIAGRAMA DE FLUJO

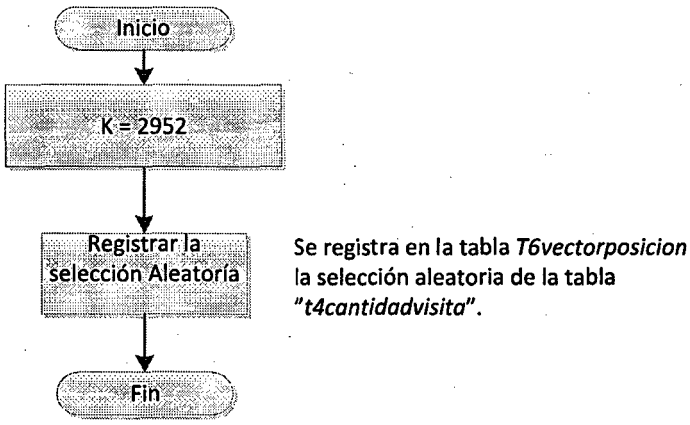


Figura VI.4: Diagrama de Flujo de Selección de la muestra representativa.

VI.4.3.2 PSEUDOCÓDIGO

1. Instancias las conexiones a mydb.t4cantidadvisita, mydb.t6vectorposicion.
2. Insertar un Select Aleatorio de *2952* elementos INSERT mydb.t5muestrasesion() SELECT() FROM mydb.t4cantidadvisita WHERE() LIMIT 2952.
3. Fin.

CAPÍTULO VII

PROCESO DE AGRUPAMIENTO

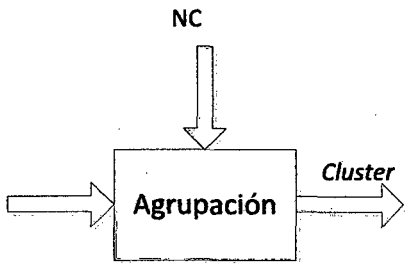


Figura VII.1: Proceso de Agrupamiento.

El objetivo de este proceso es aplicar el algoritmo *k-means*, que mediante un programa y con los parámetros iniciales permite encontrar los *Clusters*.

VII.1 NÚMERO DE *CLUSTER* (NC)

Es la variable para determinar el número de *Clusters* (NC), si le asignamos NC = 2, el proceso de agrupación determina dos *Clusters*, si le asignamos NC = 3, el proceso de agrupación determina tres *Clusters* y así sucesivamente hasta que en el proceso de análisis se determine el *k-óptimo* (*Clusters* óptimo). No se considera NC=1, ya que estaríamos pretendiendo que todos los elementos formen un grupo.

VII.2 PROCESO DE AGRUPACIÓN

En este proceso se ejecuta el algoritmo *k-means* para lo cual se ingresa el número de *Cluster* (NC), Dimensiones (D), Sesiones(S) y Vector Posición (VP). La generación de *Cluster* consiste en ir agrupando las sesiones, cada sesión está determinado por un vector, este vector es un conjunto de porcentajes de visitas a las diferentes Dimensiones(D), la agrupación de las sesiones se determina a través de las distancias euclidianas de las sesiones hacia los centroides de los *Clusters,* según la cercanía se van agrupando.

1. Se instancia el objeto *Clustering* que está compuesto por los métodos de inicialización, construcción y memoria.
2. La inicialización consiste en generar los *Cluster* iniciales, las variables iniciales son arreglo de sesiones (sesiones), el tamaño de las

dimensiones (dim), número de *Clusteres* (numclus) y la cantidad de sesiones (cantses).

a. Se instancia el vector posición de centroides.
b. Se instancia el arreglo de centroides.
c. Se instancia el arreglo de arreglo de *Clusters*.
d. Se escoge aleatoriamente sesiones como centroides según la cantidad de *Clusters*.
e. Se guarda la posición del centroide elegido.
f. Se guardan los centroides elegidos en el arreglo de centroides.
g. Se generan grupos que adelante serán llamados *Clusters* con respecto a los centroides elegidos.
h. Se recalcula para los *Clusters* generados nuevos centroides.
i. Se actualizan los nuevos centroides calculados en el arreglo de centroides.
j. Se registran los *Clusters* en el arreglo de *Clusters*.

3. Se recorre una cantidad **n** de veces los métodos de construcción y memoria para hallar los *Clusters* óptimos.
    a. La Construcción, esta función evita la convergencia a óptimos locales, las variables de entrada son arreglo de sesiones y la dimensión del vector de sesión.
        i. Se instancia el arreglo de centroides
        ii. Se instancia el arreglo de *Clusters*

iii. Se recorre el arreglo de *Clusters*

1. Se extrae un *Cluster* del arreglo de *Clusters*
2. Se recorre el *Cluster*

a. Se extrae vector sesión del *Cluster*.

b. Se recorre el arreglo de centroides

i. Se extrae un centroide y se halla la distancia con el vector sesión.

ii. Si la distancia hallada es menor a la distancia del vector sesión a su centroide entonces el *Cluster* es guardado en un arreglo RCL de óptimos locales.

iii. Se vuelve al punto *3.a.iii.2.b* hasta recorrer todos los centroides.

c. Escoges al azar un *Cluster* del arreglo RCL.

d. Agregas a dicho *Cluster* el vector sesión.

e. Se genera nuevos centroides tanto para el *Cluster* que se agregó al vector posición en correspondencia al vector, que se quitó.

f. Se actualiza los *Clusters* modificados en el arreglo de *Clusters* así también los centroides generados en el arreglo de

centroides.

3. Se vuelve al punto *3.a.iii.2* hasta terminar de recorrer del *Cluster*.

iv. Se vuelve al punto *3.a.iii* hasta terminar de recorrer el arreglo de *Clusters*.

b. La Memoria, esta función te genera las mejores soluciones, las variables de entrada son arreglo de *Clusters* y la dimensión del vector de sesión.

i. Ingresa al método Reagrupación, este método elimina el *Cluster* de menos elementos y genera un nuevo centroide al *Cluster* más disperso generando una reagrupación de los elementos del *Cluster* eliminado.

1. Se instancia un arreglo de *Clusters*.
2. Se instancia un arreglo de centroides.
3. Se recorre el arreglo de *Clusters*.
    a. Se escoge un *Cluster*.
    b. Se escoge su centroide.
    c. Se halla el error promedio que viene hacer la división entre la suma de las distancias de los vectores sesiones a sus centroide entre el módulo de los mismo: error = sumadistancia/modulo.

d. Se halla el tamaño del *Cluster.*

e. Se halla el parámetro de dispersión:

rel = error/tamaño.

f. Se hallan las posiciones del *Cluster* de menos cantidad y el que tiene mayor dispersión.

4. Se vuelve al punto *3.b.i.3* hasta terminar de recorrer todos los *Clusters.*
5. Se extrae del arreglo de *Clusters*, a los *Clusters* de mayor dispersión y a los *Cluster* de menor cantidad de sesiones.
6. Se elimina del arreglo de *Clusters*, a los *Cluster* de menos cantidad de sesiones.
7. Se divide el *Cluster* de mayor dispersión en dos *Clusters.*
8. Se calcula el centroides de los dos nuevos *Clusters.*
9. Se agregan los nuevos *Clusters* al arreglo de *Clusters.*
10. Se agrega los dos nuevos centroides al arreglo de centroides.
11. Recorre el *Cluster* de menos elementos.

a. Se extrae el vector sesión del *Cluster*.

b. Se recorre el arreglo de centroides.

i. Se extrae un centroide.

ii. Se halla la distancia del vector sesión al centroide.

iii. Se halla la posición del *Cluster* de menos distancia.

c. Se vuelve al punto *3.b.i.11.b* hasta terminar de recorrer todo el arreglo de centroides.

12. Se extrae el *Cluster* de menor distancia.

13. Se agrega el vector sesión al *Cluster* de menor distancia.

14. Se halla el centroide al *Cluster* de menor distancia.

15. Se agrega al arreglo de *Clusters* el *Cluster* modificado.

16. Se agrega al arreglo de centroides el centroide hallado.

17. Se vuelve al punto *3.b.i.11* hasta recorrer todos los vectores sesión del *Cluster*.

ii. Se ingresa nuevamente a la función construcción, para evitar la convergencia a óptimos locales, las variables de

entrada son arreglo de sesiones y la dimensión del vector de sesión.

iii. Se obtiene el arreglo de *Clusters* de la función construcción.

iv. Se ingresa el arreglo de *Clusters* al método función objetivo.

1. Se recorre el arreglo de *Clusters*.
    a. Se escoge un *Cluster*
    b. Se escoge su centroide del arreglo de centroides.
    c. Se recorre el *Cluster*
        i. Se escoge un vector sesión.
        ii. Se halla la distancia del vector sesión al centroide.
        iii. En un parámetro suma se van sumando las distancias halladas.
    d. Se vuelve al punto *3.b.iv.1.a* hasta recorrer todos los elementos del *Cluster*.
2. Se vuelve al punto *3.b.iv.1* hasta recorrer todo el arreglo de *Clusters*.
3. Se retornó el parámetro suma "a". Se halla la función objetivo (b) para arreglo de *clusters*

hallados en el método memoria.

v. Si la condición es a<b se volverá al punto *3.b.i* y se hará nuevamente la iteración hasta que la condición sea a>b.

vi. Se retornara el arreglo de *Clusters* que cumplan con la condición.

c. Se halla la función objetivo para el arreglo de *Clusters* hallado en el método de memoria.

d. Si en caso la función objetivo es menor a la función objetivo anterior entonces se escoge al arreglo de *Clusters* de la menor función objetivo.

e. Se vuelve al punto *3* hasta terminar de recorrer las n veces.

4. Se guardan los *Clusters* en la tabla *t7Cluster* así también se actualizan los identificadores de *Cluster* en la tabla *t6vectorposicion* los vectores sesiones según corresponda.

VII.2.1 PSEUDOCÓDIGO

Los llamados métodos jerárquicos

Proceso: [grupos]= Agrupamiento (sesiones, dim, numclus, cantses)

1. *Clustering Clustering*= new Clustering ()
2. Inicializacion=*Clustering*.Inicializacion(arregloVectoresUsuarios, dim, numClus, cantses)

2.1 ArrayList arreglocentroides=new ArrayList()

2.2 ArrayList arregloClustering=new ArrayList()

2.3 Mientras i < numClus

2.3.1 int pos= Random (0, cantses-1)

2.3.2 Float [] centroides= new float [dim vector]

2.3.3 ArrayList Clustering=new ArrayList ()

2.3.4 centroides= (float []) arregloVectoresUsuarios.get (pos)

2.3.5 arreglocentroides.add(centroides)

2.3.6 arregloClustering.add(Clustering)

2.4 Fin Mientras

2.5 Mientras i < cantses

2.5.1 float [] vectorusuario= new float [dimvector]

2.5.2 vectorusuario= (float []) arregloVectoresUsuarios.get (i)

2.5.3 Mientras j < arreglocentroides.size ()

2.5.3.1 centroide= (float []) arreglocentroides.get(j)

2.5.3.2 distancia=distancia (vectorusuario, centroide)

2.5.3.3 Si distancia<mindis

2.5.3.3.1 mindis=distancia

2.5.3.3.2 posclus=j

2.5.3.5 Fin Si

2.5.4 Fin Mientras

2.5.5 Clustering = arregloClustering (posclus)

2.5.6 Clustering.add(vectorusuario)

2.5.7 arregloClustering.set(posclus,Clusteringaux)

2.6 Fin Mientras

2.7 Mientras i < arregloClustering.size()

2.7.1 *Clustering*= (ArrayList) arregloClustering.get(i)

2.7.2 centroide=Centroide (Clustering,dimvector)

2.7.3 arreglocentroides.set(i,centroide)

2.8 Fin Mientras

2.9 ArregloClustering.add(arregloClustering)

2.10 ArregloClustering.add(arreglocentroides)

2.12 inicializacion = ArregloClustering

2.11 return inicializacion

2.12 Fin Inicializacion

3. Mientras i < Max_Iteracion

3.1 construccion=Clustering.Construccion(inicializacion,dim)

3.1.1 arregloClustering = inicializacion.get(0)

3.1.2 arreglocentroides = inicializacion.get(1)

3.1.3 Mientras j < arregloClustering.size()

3.1.3.1 Clustering = arregloClustering.get(j)

3.1.3.2 Mientras k< Clustering.size()

3.1.3.2.1 vectorusuario = (float []) Clustering.get (k)

3.1.3.2.2 ArrayList RCL = new ArrayList()

3.1.3.2.3 ArrayList RCLpos = new ArrayList()

3.1.3.2.4 Mientras r < arreglocentroides.size()

1. centroide =(float[]) arreglocentroides.get(r)

2. distancia=distancia (vectorusuario,centroide)

3. Si distancia<distini

3.1 posclus=r

3.2 ClusteringRCL= arregloClustering.get(r)

3.3 RCLpos.add(posclus)

3.4 RCLpos.add(ClusteringRCL)

3.5 RCL.add(RCLpos)

4. Fin Si

3.1.3.2.3 Fin Mientras

3.1.3.2.4 pos=RCLpos.get(0)

3.1.3.2.5 remoto=(ArrayList)

arregloClustering.get(pos)

3.1.3.2.6 remoto.add(vectorusuario)

3.1.3.2.7 Index=Clustering.indexOf(vectorusuario)

3.1.3.2.8 Clustering.remove(index)

3.1.3.2.9 cenremoto=Centroide

(remoto,dimvector)

3.1.3.2.10 clusmodi=Centroide

(Clustering,dimvector)

3.1.3.2.11 arregloClustering.set(j,Clustering)

3.1.3.2.12 arregloClustering.set(pos,remoto)

3.1.3.2.13 arreglocentroides.set (j, clusmodi)

3.1.3.2.14 arreglocentroides.set (pos, cenremoto)

3.1.3.3 Fin Mientras

3.1.4 Fin Mientras

3.1.5 ArregloCluster.add(arregloClustering)

3.1.6 ArregloCluster.add(arreglocentroides)

3.1.7 construccion = ArregloCluster

3.1.8 return construccion

3.2 Fin construccion

3.3 memoria=Clustering.Memoria(construccion, dim)

3.3.1 int a=0

3.3.2 int b =0

3.3.3 Mientra (a<b)

3.3.3.1 reagrupacion = Reagrupacion (construccion, dimvector)

3.3.3.1.1 arregloClustering = construccion.get (0)

3.3.3.1.2 arreglocentroides = construccion.get (1)

3.3.3.1.3 Mientras i<arregloClustering.size()

1. Clustering = arregloClustering.get(i)
2. centroide = arreglocentroides.get(i)
3. error =ErrorPromedio (Clustering,centroide,dimvector)
4. tamaño=Clustering.size()
5. rel=error/tamaño
6. Si tamaño<tamañomin
   6.1 tamañomin=tamaño
   6.2 post=i
7. Fin Si
8. Si rel<relmayor
   8.1 relmayor=rel
   8.2 post=i
9. Fin Si

3.3.3.1.4 Fin Mientra

3.3.3.1.5 Clusteringmin=
arregloClustering.get(post)

3.3.3.1.6 Clusteringdis=
arregloClustering.get(posdis)

3.3.3.1.7 arregloClustering.remove(post)

3.3.3.1.8 arregloClustering.remove(posdis)

3.3.3.1.9 arreglocentroides.remove(post)

3.3.3.1.10 arreglocentroides.remove(posdis)

3.3.3.1.11ArregloCluster=DividirClustering(Clust
eringdis,dimvector)

3.3.3.1.12 arregloCluster =
ArregloCluster.get(0)

3.3.3.1.13 arregloCentroide =
ArregloCluster.get(1)

3.3.3.1.14 Cluster1 = arregloCluster.get(0)

3.3.3.1.15 Cluster2 = arregloCluster.get(1)

3.3.3.1.16 centroide1 = arregloCentroide.get(0)

3.3.3.1.17 centroide2 = arregloCentroide.get(1)

3.3.3.1.18 arreglocentroides.add(centroide1)

3.3.3.1.19 arreglocentroides.add(centroide2)

3.3.3.1.20 arregloClustering.add(Cluster1)

3.3.3.1.21 arregloClustering.add(Cluster2)

3.3.3.1.22 Mientras i < Clusteringmin.size()

1. vectorposicion = Clusteringmin.get(i)
2. Mientras j < arreglocentroides.size()

   2.1 Centroide = arreglocentroides.get(j)

   2.2 d = distancia (vectorposicion,centroide)

   2.3 Si (d<distmin)

      2.3.1 distmin=d

      2.3.2 pos = j

   2.4 Fin Si

3. Fin Mientras
4. Clustering = arregloClustering.get (pos)
5. Centroide=Centroide (Clustering,dimvector)
6. Arreglocentroides.set (pos, centroide)
7. ArregloClustering.set(pos,

Clustering)

3.3.3.1.23Fin Mientras

3.3.3.1.24 ArregloCluster.add(arregloClustering)

3.3.3.1.25 ArregloCluster.add(arreglocentroides)

3.3.3.1.26 reagrupacion = ArregloCluster

3.3.3.1.27 Return reagrupacion

3.3.3.2 Fin Reagrupación.

3.3.3.3 construaux= Construccion (reagrupacion, dimvector)

3.3.3.4 a = FuncionObjetivo(construaux, dimvector)

3.3.3.5 b = FuncionObjetivo(construccion, dimvector)

3.3.3.6 Fin mientras

3.3.3 Memoria = Construccion

3.3.5 Return memoria

3.4 Fin memoria

3.5 faux = Clustering.FuncionObjetivo(memoria,dim)

3.6 Si (faux<f) entonces

3.6.1 ClusteringFinal = memoria

3.6.2 f = faux

3.7 Fin Si

4. Fin Mientras

5. ArregloClustering = ClusteringFinal.get (0)

6. Arreglocentroides = ClusteringFinal.get (1)

7. Mientras i < arregloClustering.size ()

    7.1 pos= i +1

    7.2 Map params = new HashMap()

    7.3 t7Clusterdao.insert(params)

    7.4 Clustering =arregloClustering.get(i)

    7.5 Centroide = arreglocentroides.get(i)

    7.6 t6vecposdao.insertVector(centroide , i, num, periodo)

    7.7 Mientras s < Clustering.size()

        7.7.1 Vectorposicion = Clustering.get(s)

        7.7.2 posi = Vectorposicion[vectorposicion.length]

        7.7.3 t6vecposdao.update(pos, posi)

    7.8 Fin Mientras

8. Fin Mientras

VII.2.2 DIAGRAMA DE FLUJO

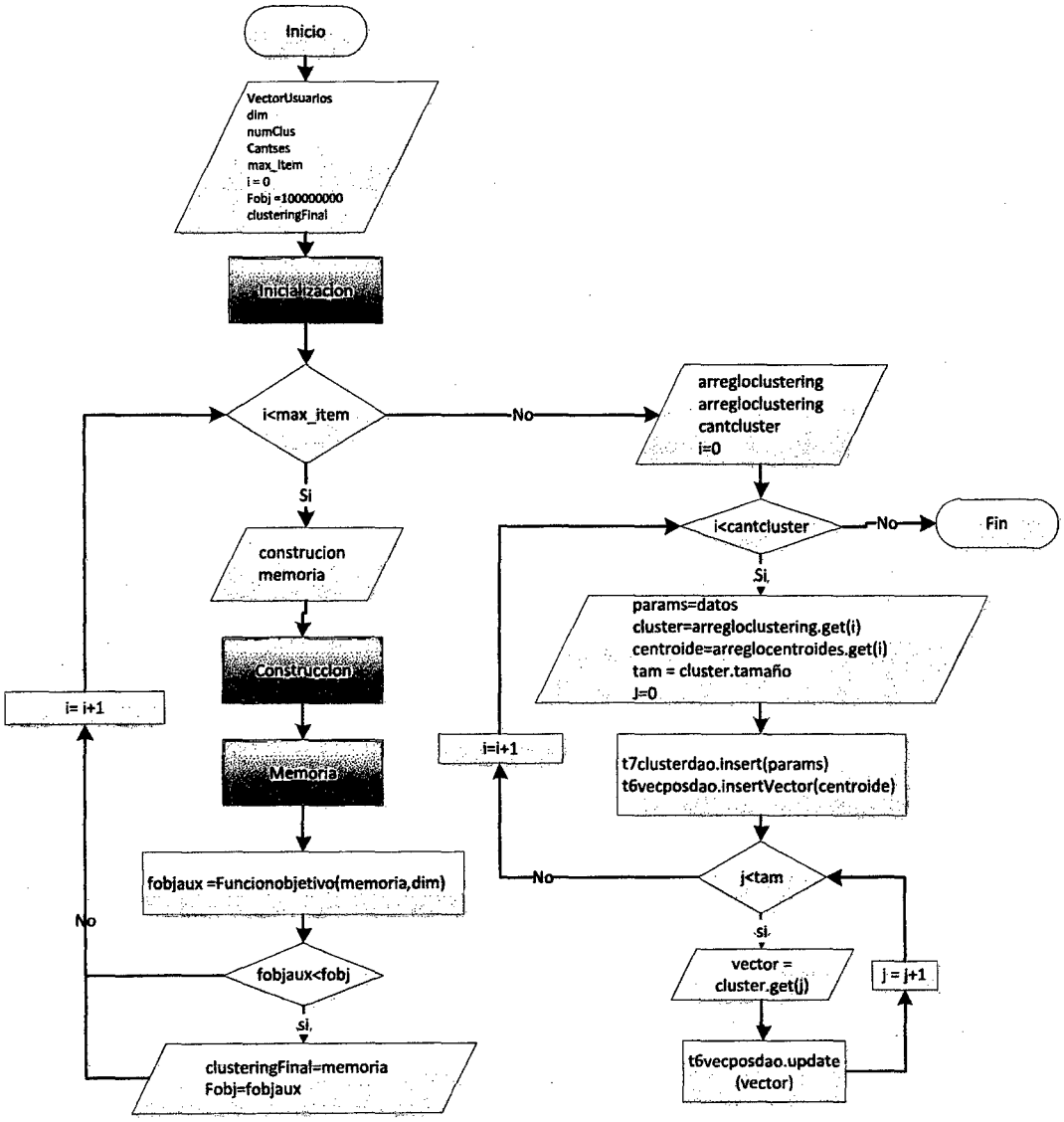


Figura VII.2: Diagrama de Flujo del Algoritmo de Agrupamiento.

VII.3 *CLUSTER*

Conjunto de sesiones que tienen características comunes, las cuales se representan por los porcentajes de visitas a las diferentes dimensiones, estas dimensiones representan a los servicios del portal, los *Clusters* que se obtienen están determinados por el número de *Cluster* (NC) que es la entrada al Proceso de Agrupación.

CAPÍTULO VIII

PROCESO DE ANÁLISIS

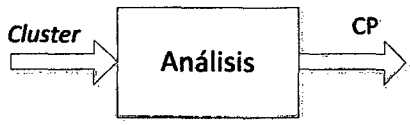


Figura VIII.1: Proceso de Análisis.

El objetivo de este proceso es encontrar el número de *Cluster* óptimo, hacer el análisis a dicho *Cluster* y determinar el patrón para cada *Cluster.*

VIII.1 OBTENCIÓN DE NÚMERO DE CLUSTER ÓPTIMO

Para determinar el número óptimo de *Cluster* se establece un parámetro de distancia mínima, que determina cuándo dos *Clusters* son similares y cuándo no lo son: en tal sentido el parámetro definido es la distancia entre los centroides de los *Clusters*, cuando estos están muy cercanos entonces se considera que los *Clusters* son similares en caso contrario se consideran como distintos (Jain, Murty [12]).

Se considera que la distancia es pequeña si está en el intervalo de:

0.001 < DISmin < **0.005** aproximadamente.

a) CORRIDA DE ALGORITMO PARA K=2

Como se observa en el *Cuadro VIII.1* la distancia entre los dos centroides no es muy pequeña, entonces se considerar que pueden haber más de dos *Clusters*.

K=2	C1	C2
C1		0.05059924
C2		

Cuadro VIII.1: Distancia entre centroides para K=2.

b) CORRIDA DE ALGORITMO PARA K=3

Como se observa en el *Cuadro VIII.2* la distancia entre los tres centroides C1:C2, C1:C3 y C2:C3 no es muy pequeña entonces se considera que pueden haber más de tres *Clusters*.

K=3	C1	C2	C3
C1		0.1198244	0.05012438
C2			0.0711653
C3			

Cuadro VIII.2: Distancia entre centroides para K=3.

c) CORRIDA DE ALGORITMO PARA K=4

Como se observa en el *Cuadro VIII.3* la celda C2:C4 representa la mínima distancia, entonces el centroide 2 con el centroide 4 están muy cercanos por lo tanto los *Cluster* 2 y 4 se consideran como si fueran uno solo, en tal sentido el número *Cluster* óptimo es 3.

K=4	C1	C2	C3	C4
C1		0.01763838	0.5682182	0.018047523
C2			0.58518068	0.000413689
C3				0.585574567
C4				

Cuadro VIII.3: Distancia entre centroides para K=4.

d) CORRIDA DE ALGORITMO PARA K=5

Como se observa en el *Cuadro VIII.4* las celdas C2:C4, C2:C5 y C4:C5 representan la mínima distancia, entonces el centroide 2 con el centroide 4 y centroide 5 están muy cercanos por lo tanto los *Cluster* 2, 4 y 5 se consideran como si fueran uno solo, en tal sentido el número *Cluster* óptimo sería de 3.

K=5	C1	C2	C3	C4	C5
C1		0.76491972	0.7608053	0.760359373	0.76384668
C2			0.00625482	0.004882167	0.00168456
C3				0.005136369	0.00542122
C4					0.00414239
C5					

Cuadro VIII.4: Distancia entre centroides para K=5.

e) CORRIDA DE ALGORITMO PARA K=6

Como se observa en el *Cuadro VIII.5* las celdas C2:C4, C2:C5, C2:C6, C4:C5, C4:C6 y C5:C6 representan la mínima distancia, entonces el centroide 2 con el centroide 4, centroide 5 y centroide 6 están muy cercanos por lo tanto los *Cluster* 2, 4, 5 y 6 se consideran como si fueran uno solo, en tal sentido el número *Clusters* óptimo es 3.

=6	C1	C2	C3	C4	C5	C6
C1		0.75589939	0.75242693	0.756838937	0.75519646	0.75511809
C2			0.00413013	0.00212992	0.00138652	0.0015131
C3				0.005606991	0.00395212	0.00384834
C4					0.00222269	0.00220386
C5						0.00120944
C6						

Cuadro VIII.5: Distancia entre centroides para K=6.

Como podemos observar la tendencia de la cantidad óptima de *Cluster* es 3 en todas las corridas.

VIII.2 PROCESO DE ANÁLISIS DE *CLUSTER*

En este proceso se a analiza los *Clusters* obtenidos para *k-óptimo*, como se vio en la *Sección VIII.1*; el k-óptimo es k=3.

a) *CLUSTER* 1

En el cuadro *Cuadro VIII.6*, se observa que los servicios más ingresados para el *Cluster* 1 son:

- Consulta RUC.
- Suspensión de 4ta Categoría-Formulario 1609.
- Cronograma de obligaciones mensuales - ejercicio 2011.
- Inscripción de RUC.
- Consulta de solicitudes de suspensiones de 4ta-categoría Formulario 1609.
- Atención de Quejas y Sugerencias.
- Estado de la Queja Presentada.

En la *Figura VIII.2* podemos observar mejor, mediante la gráfica de la torta.

Nro de Dimen.	URL de la Página Web	Porcentaje
1	/cl-ti-itmrconsruc/jcrS00Alias	89.59%
2	/cl-at-itageban/bcosucS01Alias	0.00%
3	/ol-ti-itinsrucsol/iruc001Alias	1.81%
4	/ol-ti-itsusprenta/srS01Alias	4.70%
5	/cl-ti-itcronobligme/fvS01Alias	2.68%
6	/cl-at-itconcompag/ccS02Alias	0.08%
7	/cl-ti-itpresqueja/sqsS21Alias	0.33%
8	/ol-ti-itdenuncia/denS01Alias	0.00%
9	/cl-ti-itpresqueja/sqsS31Alias	0.16%
10	/cl-ti-itconsrenta/srS01Alias	0.65%
11	/cl-at-itrecone/roS01Alias	0.00%
12	/ol-ti-itfichaseleccion/fichaS01Alias	0.00%
13	/cl-at-itentdeu/iisS01Alias	0.00%
14	/cl-ti-itconspredios/sfpS02Alias	0.00%
15	//cl-ti-itmrconsruc/jcrS00Alias	0.00%
16	/cl-ti-itmrconsruc//jcrS00Alias	0.00%
17	/ol-ti-itpresf5030/rsdS01Alias	0.00%
18	/cl-ti-itconsdenuncia/denS02Alias	0.00%
19	/ol-ti-itpdtpred/sfp01Alias	0.00%
20	/cl-ad-ittabladescripcionconsulta/TablaDescripcionS01Alias	0.00%
21	/wtc/wapTcS01Alias	0.00%
22	/ol-ti-itreciboelectronico/cpelec003Alias	0.00%
23	/ol-ti-itreciboelectronicovalarch/ceS01Alias	0.00%
24	/cl-ti-itmrconsruc/jcrS03Alias	0.00%
25	/cl-ad-ittipocambioconsulta/TipoCambioS01Alias	0.00%

Cuadro VIII.6: Porcentaje de visitas promedio a las diferentes dimensiones o servicios, del *Cluster* 1.

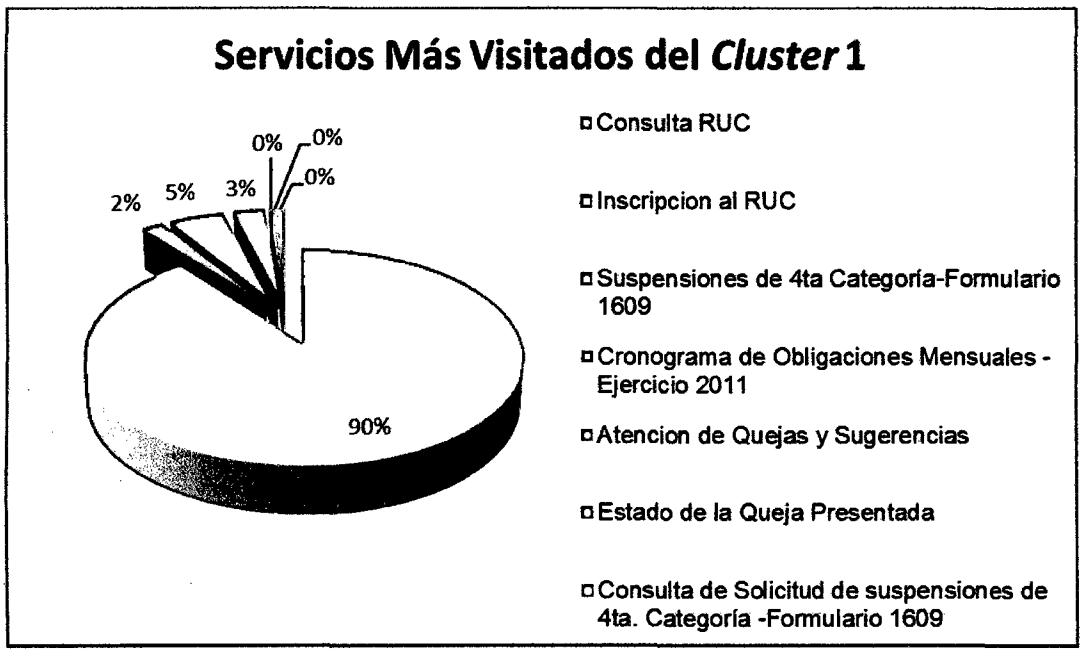


Figura VIII.2: Servicios más visitados en el *Cluster* 1.

b) *CLUSTER 2*

En el *Cuadro VIII.7*, se observa que los servicios más ingresados para el *Cluster* 2 son:

- Consulta RUC.
- Consulta de Agencias Bancarias a nivel Nacional.
- Inscripción del RUC.
- Suspensión de 4ta Categoría-Formulario 1609.
- Cronograma de Obligaciones Mensuales.
- Consulta de Compensación de Pagos.
- Consulta de solicitudes de suspensiones de 4ta-categoría Formulario 1609.

En la *Figura VIII.3* podemos observar mejor, mediante la gráfica de la torta.

Nro de Dimen	URL de la Página Web	Porcentaje
1	/cl-ti-itmrconsruc/jcrS00Alias	99.01%
2	/cl-at-itageban/bcosucS01Alias	0.02%
3	/ol-ti-itinsrucsol/iruc001Alias	0.13%
4	/ol-ti-itsusprenta/srS01Alias	0.34%
5	/cl-ti-itcronobligme/fvS01Alias	0.35%
6	/cl-at-itconcompaq/ccS02Alias	0.04%
7	/cl-ti-itpresqueja/sqsS21Alias	0.00%
8	/ol-ti-itdenuncia/denS01Alias	0.00%
9	/cl-ti-itpresqueja/sqsS31Alias	0.00%
10	/cl-ti-itconsrenta/srS01Alias	0.06%
11	/cl-at-itrecone/roS01Alias	0.02%
12	/ol-ti-itfichaseleccion/fichaS01Alias	0.01%
13	/cl-at-itentdeu/iisS01Alias	0.02%
14	/cl-ti-itconspredios/sfpS02Alias	0.00%
15	//cl-ti-itmrconsruc/jcrS00Alias	0.00%
16	/cl-ti-itmrconsruc//jcrS00Alias	0.00%
17	/ol-ti-itpresf5030/rsdS01Alias	0.00%
18	/cl-ti-itconsdenuncia/denS02Alias	0.00%
19	/ol-ti-itpdtpred/sfp01Alias	0.00%
20	/cl-ad-ittabladescripcionconsulta/TablaDescripcionS01Alias	0.00%
21	/wtc/wapTcS01Alias	0.00%
22	/ol-ti-itreciboelectronico/cpelec003Alias	0.00%
23	/ol-ti-itreciboelectronicovalarch/ceS01Alias	0.00%
24	/cl-ti-itmrconsruc/jcrS03Alias	0.00%
25	/cl-ad-ittipocambioconsulta/TipoCambioS01Alias	0.00%

Cuadro VIII.7: Porcentaje de visitas promedio a las diferentes dimensiones o servicios, del *Cluster* 2.

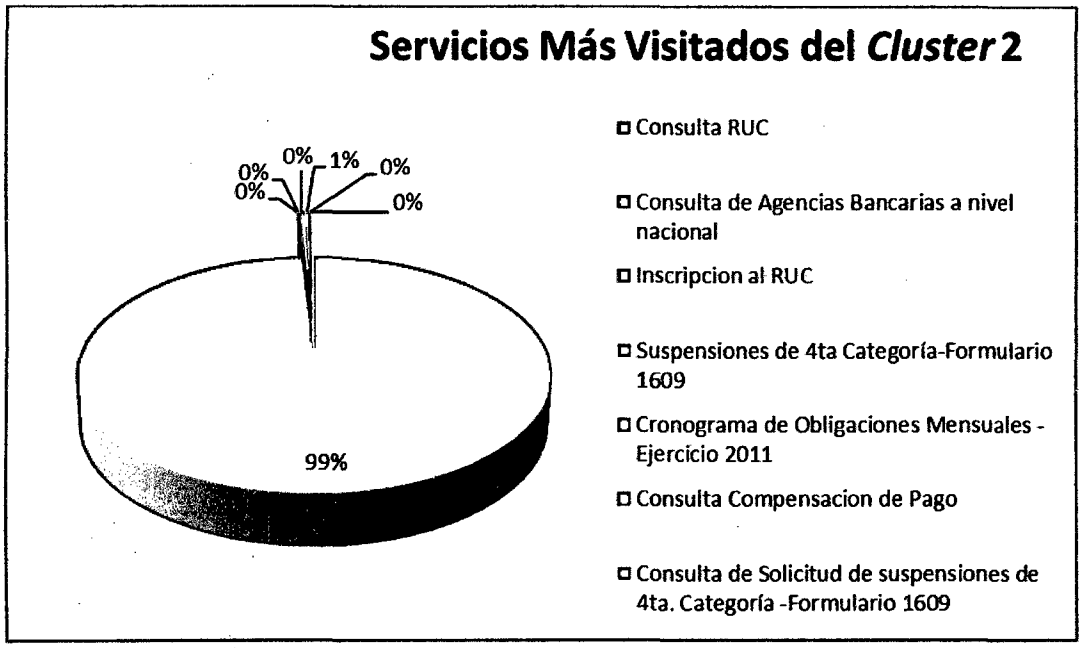


Figura VIII.3: Servicios más visitados en el *Cluster* 2.

c) *CLUSTER* 3

En el cuadro *Cuadro VIII.8*, Vemos que los servicios más ingresados para el *Cluster* 3 son:

- Consulta RUC.
- Inscripción del RUC.
- Suspensión de 4ta Categoría-Formulario 1609.
- Cronograma de Obligaciones Mensuales.
- Consulta de Compensación de Pagos.
- Consulta de solicitud de suspensiones de 4ta. Categoría-Formulario 1609.
- Declaraciónón de Predios Formulario Virtual 1630.

En la *Figura VIII.4* podemos observar mejor, mediante la gráfica de torta.

Nro de Dimen.	URL de la Página Web	Porcentaje
1	/cl-ti-itmrconsruc/jcrS00Alias	93.89%
2	/cl-at-itageban/bcosucS01Alias	0.33%
3	/ol-ti-itinsrucsol/iruc001Alias	0.84%
4	/ol-ti-itsusprenta/srS01Alias	2.08%
5	/cl-ti-itcronobligme/fvS01Alias	1.34%
6	/cl-at-itconcompag/ccS02Alias	0.40%
7	/cl-ti-itpresqueja/sqsS21Alias	0.00%
8	/ol-ti-itdenuncia/denS01Alias	0.00%
9	/cl-ti-itpresqueja/sqsS31Alias	0.00%
10	/cl-ti-itconsrenta/srS01Alias	0.66%
11	/cl-at-itrecone/roS01Alias	0.00%
12	/ol-ti-itfichaseleccion/fichaS01Alias	0.00%
13	/cl-at-itentdeu/iisS01Alias	0.00%
14	/cl-ti-itconspredios/sfpS02Alias	0.06%
15	//cl-ti-itmrconsruc/jcrS00Alias	0.00%
16	/cl-ti-itmrconsruc//jcrS00Alias	0.00%
17	/ol-ti-itpresf5030/rsdS01Alias	0.00%
18	/cl-ti-itconsdenuncia/denS02Alias	0.00%
19	/ol-ti-itpdtpred/sfp01Alias	0.40%
20	/cl-ad-ittabladescripcionconsulta/TablaDescripcionS01Alias	0.00%
21	/wtc/wapTcS01Alias	0.00%
22	/ol-ti-itreciboelectronico/cpelec003Alias	0.00%
23	/ol-ti-itreciboelectronicovalarch/ceS01Alias	0.00%
24	/cl-ti-itmrconsruc/jcrS03Alias	0.00%
25	/cl-ad-ittipocambioconsulta/TipoCambioS01Alias	0.00%

Cuadro VIII.8: Porcentaje de visitas promedio a las diferentes dimensiones o servicios del *Cluster* 3.

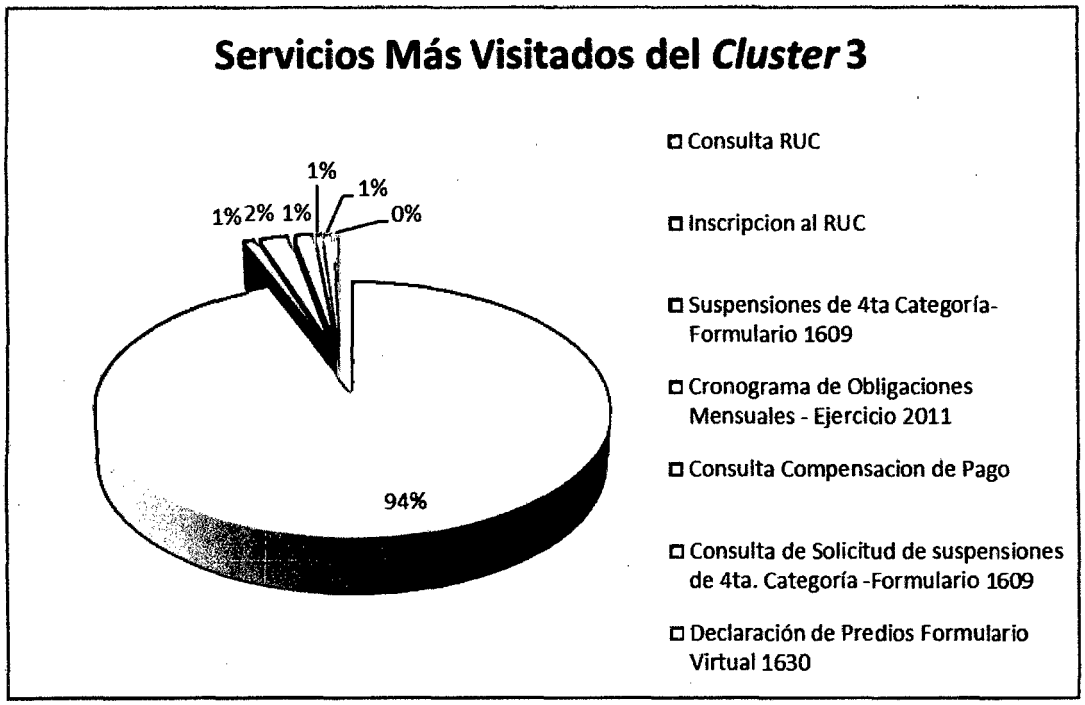


Figura VIII.4: Servicios más visitados en el *Cluster* 3.

VIII.3 CONOCIMIENTO DE PATRONES (CP)

En el análisis de los *Clusters* de usuarios descubiertos es posible acceder a la información sobre la identificación de hábitos de uso, con la cantidad de sesiones de usuarios en cada grupo o *Cluster* descubierto, sobre el detalle de las sesiones y el porcentaje de acceso a cada servicio por parte de los usuarios de cada grupo.

A continuación se muestra la cantidad de sesiones en cada *Cluster* [*Cuadro VIII.9*] y el porcentaje correspondiente a cada uno [*Figura VIIII.4*].

Cluster	Cantidad de Sesiones	Porcentaje (%)
C1	203	6.88
C2	2014	68.27
C3	733	24.85
Total	2950	100

Cuadro VIII.9: Cantidad de Sesiones por *Cluster.*

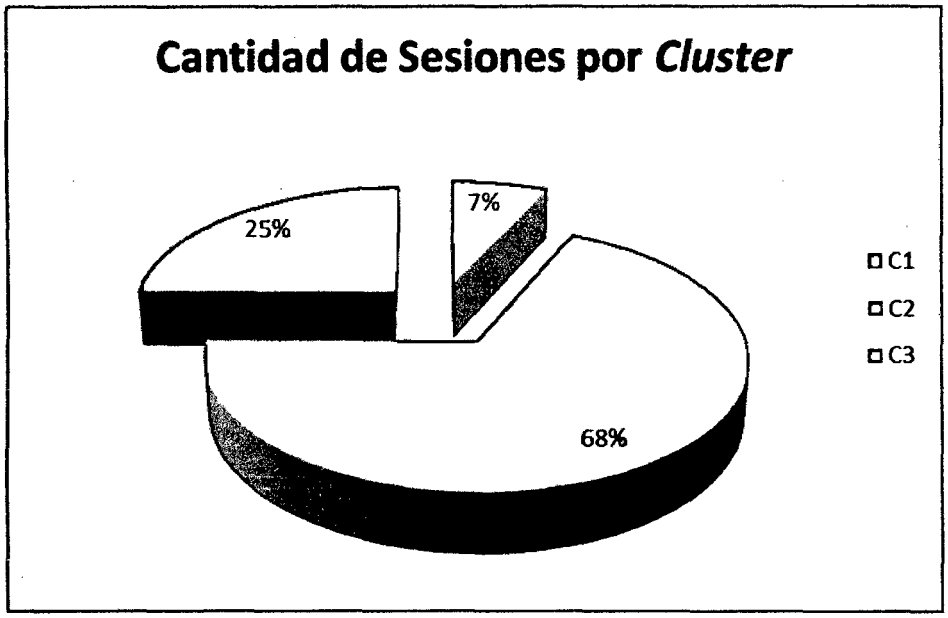


Figura VIII.5: Cantidad de sesiones por *Cluster*.

a) DEL *CLUSTER* 1 - PATRÓN 1

Viendo el porcentaje de visitas (ver *Cuadro VIII.6*): se observa que los servicios más visitados son la Consulta Ruc y otros siete servicios más, que se lista a continuación:

- Consulta RUC.
- Suspensión de 4ta Categoría-Formulario 1609.
- Cronograma de obligaciones mensuales - ejercicio 2011.
- Inscripción de RUC.
- Consulta de solicitudes de suspensiones de 4ta-categoría Formulario 1609.
- Atención de Quejas y Sugerencias.
- Estado de la Queja Presentada.

La probabilidad de que los usuarios del *Cluster* 1 puedan ingresar a estos servicios es alta y es por ello que estos servicios deben ser puestos como acceso directo para que cuando entre los usuarios del *Cluster* 1 les aparezca un *Pop Up* con estos servicios.

b) DEL *CLUSTER* 2 - PATRÓN 2

Viendo el porcentaje de visitas (ver *Cuadro VIII.7*): se observa que los servicios más visitados son la Consulta Ruc y otros siete servicios más, que lista a continuación:

- Consulta RUC.
- Consulta de Agencias Bancarias a nivel Nacional.
- Inscripción del RUC.
- Suspensión de 4ta Categoría-Formulario 1609.
- Cronograma de Obligaciones Mensuales.
- Consulta de Compensación de Pagos.
- Consulta de solicitudes de suspensiones de 4ta-categoría Formulario 1609.

La probabilidad de que los usuarios del *Cluster* 2 puedan ingresar a estos servicios es alta y es por ello que estos servicios deben ser puestos como acceso directo para que cuando entre los usuarios del *Cluster* 2 les aparezca un *Pop Up* con estos servicios.

c) DEL *CLUSTER* 3 - PATRÓN 3

Viendo el porcentaje de visitas (ver *Cuadro VIII.8*): se observa que los servicios más visitados son la Consulta Ruc y otros siete servicios más, que lista a continuación:

- Consulta RUC.
- Inscripción del RUC.
- Suspensión de 4ta Categoría-Formulario 1609.
- Cronograma de Obligaciones Mensuales.
- Consulta de Compensación de Pagos.
- Consulta de solicitud de suspensiones de 4ta. Categoría-Formulario 1609.
- Declaraciónón de Predios Formulario Virtual 1630.

La probabilidad de que los usuarios del *Cluster* 3 puedan ingresar a estos servicios es alta y es por ello que estos servicios deben ser puestos como acceso directo para que cuando entre los usuarios del *Cluster* 3 les aparezca un *Pop Up* con estos servicios.

VIII.4 ANÁLISIS DESCRIPTIVO

Una vez implementada la mejora en el Portal Web de la SUNAT, teniendo en cuenta los resultados del análisis experimental en el cual se obtuvo los servicios más visitados y así como los perfiles de los usuarios: se procede a dar validez a la hipótesis planteada:

"Si se mejora un Portal Web a través de la personalización de dicho Portal Web para cada Cluster, esto gracias a los patrones de comportamiento de los usuarios, entonces habrá una disminución en el tiempo de acceso hacia los servicios del Portal Web así también un aumento en el nivel de satisfacción de los usuarios".

Para lo cual se procede a realizar una encuesta (ver *Anexo1.2*) con el objetivo de medir los indicadores de tiempo de acceso y el nivel de satisfacción antes y después de la implementación de la mejora en el Portal Web y luego hacer una comparación y verificar si los resultados son óptimos y dan validez a la hipótesis.

VIII.4.1 RESULTADO DE LA APLICACIÓN DE ENCUESTA

Se toma una muestra de 20 personas entre los 25 a 40 años, se verificó que sean contribuyentes de la SUNAT y que antes hayan utilizado el portal de la SUNAT para hacer algún trámite o que al menos conozcan de la

existencia del Portal Web. El experimento consistió en presentarles dos opciones uno con el Portal sin la implementación de la mejora y otra con la implementación de la mejora, luego se hizo pruebas de navegabilidad en ambos contextos midiendo los tiempos de acceso y los niveles de satisfacción una vez terminada la prueba.

Cuadres Generales: Datos estadísticos no influenciado por la mejora en el Portal Web.

a) SOBRE LA PREGUNTA: ¿HA UTILIZADO EL PORTAL WEB DE LA SUNAT PARA HACER TRÁMITES O CONSULTAS?

Se Tiene las respuestas en el *Cuadro VIII.10* y en la *Figura VIII.6*

¿Usó el portal de la SUNAT?	Cantidad	%
Sí	15	75.00
No	5	25.00
Totales	20	100.00

Cuadro VIII.10: Uso del Portal Web.

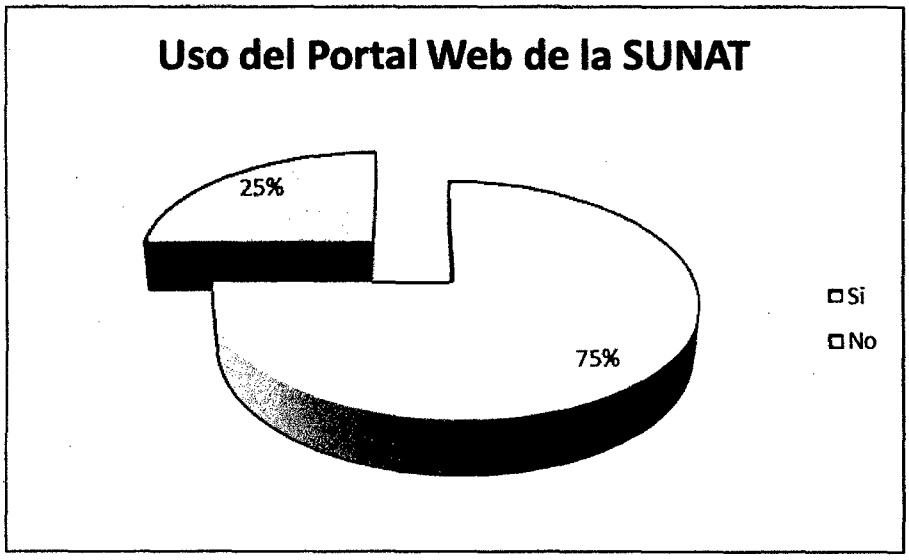


Figura VIII.6: Uso del Portal Web.

DESCRIPCIÓN–INTERPRETACIÓN

Del total de encuestados el 75% había realizado antes un trámite o consulta por el Portal Web de la SUNAT, mientras que el 25% sabía de la existencia del Portal pero preferían hacer los trámites por agencias de la SUNAT, ver *Cuadro VIII.11* y *Figura VIII.6*

b) SOBRE LA PREGUNTA: ¿CON QUÉ FRECUENCIA UTILIZA EL PORTAL WEB?

Frecuencia	Cantidad	%
Una vez	7	35.00
Dos veces	6	30.00
Tres veces	4	20.00
Más de tres veces	3	15.00
Totales		100.00

Cuadro VIII.11: Frecuencia de uso del Portal Web.

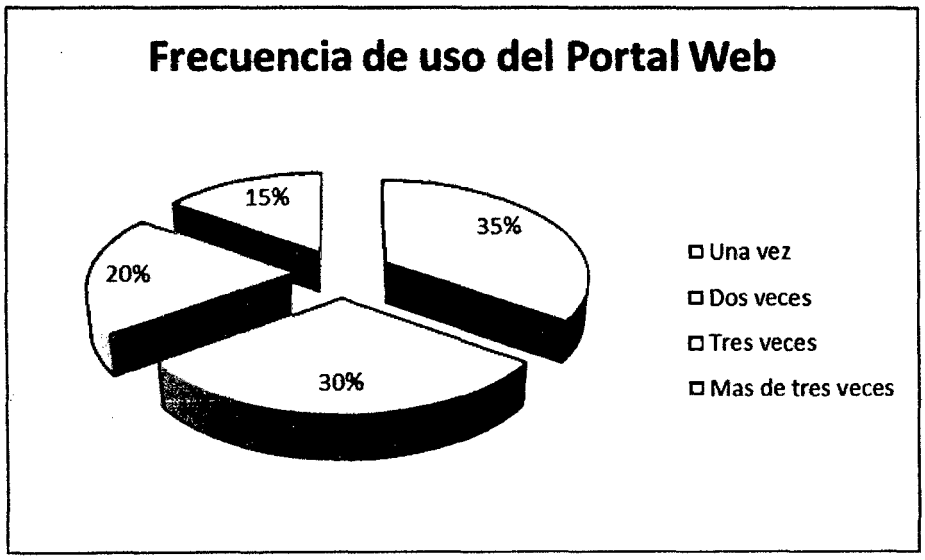


Figura VIII.7: Frecuencia de uso del Portal Web.

DESCRIPCIÓN–INTERPRETACIÓN

Según el cuadro estadístico (*Cuadro VIII.11*) la frecuencia semanal del uso del portal de la SUNAT tiene un 35% para una vez, 30% para dos veces, 20% para tres veces y un 15% para una vez. Por lo que podemos afirmar que hay un 70%, aproximadamente, que no usan frecuentemente el Portal de la SUNAT lo cual es un signo de que el Portal no satisface las necesidades o no es de fácil de uso.

CUADROS ESPECÍFICOS

Se presentan cuadros estadísticos para dos contextos antes y después de la mejora y según el diseño de la investigación se determinará la relación (R) entre las variables independiente: X (Patrones de comportamiento) y las variables dependiente Y (Tiempo de Acceso hacia los Servicios del Portal Web y Nivel de Satisfacción) : es positiva si la probabilidad de aprobación sobre la optimización del Portal Web (Pi) es mayor que cuando no se hizo la optimización del portal (Pj), en caso contrario es negativa.

Según: $R = Pi - Pj < 0$, este valor será calculado cuando se tenga las probabilidades de aprobación antes y después de la mejora en el Portal Web.

c) **SOBRE LA PREGUNTA: DIFICULTAD AL INGRESAR EN EL PORTAL WEB.**

ANTES DE LA MEJORA DEL PORTAL

Rango de Problemas	Cantidad	%
Demora mucho tiempo.	9	45.00
Difícil de Navegar.	5	25.00
Diseño o contenido difuso.	4	20.00
No tiene dificultades.	2	10.00
Totales	20	100.00

Cuadro VIII.12: Dificultades para ingresar al Portal, antes de la mejora.

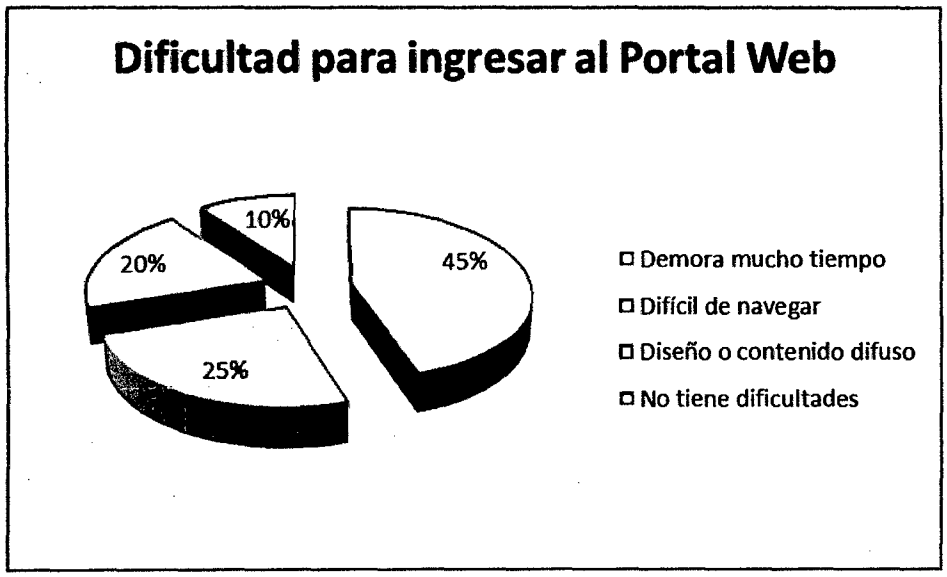


Figura VIII.8: Dificultades para ingresar al Portal Web, antes de la mejora.

DESCRIPCIÓN–INTERPRETACIÓN

Según el cuadro estadístico (*Cuadro VIII.12*) la dificultad más remarcada para el ingreso hacia los servicios del Portal es la *"demora en el acceso"* con un 45%, y es justamente el tiempo de acceso lo que se va optimizar. Con un 25 % encontramos a la *"dificultad de navegar"* por el Portal Web lo cual hace engorroso la búsqueda, otra dificultad es el *"mal diseño"* del Portal con un 20 % es decir un diseño sin considerar necesidades y preferencias de los usuarios, en el estudio se listan los servicios más utilizados por cada *Cluster* con lo que tendríamos mapeado los servicios para una mejora o rediseño del Portal Web y finalmente un 10 % de los encuestados no tuvieron dificultad para encontrar a los servicios del Portal.

DESPUÉS DE LA MEJORA DEL PORTAL

Rango de Problemas	Cantidad	%
Demora mucho tiempo.	3	15.00
Difícil de Navegar.	3	15.00
Diseño o contenido difuso.	4	20.00
No tiene dificultades.	10	50.00
Totales	20	100.00

Cuadro VIII.13: Dificultades para ingresar al Portal Web, después de la mejora.

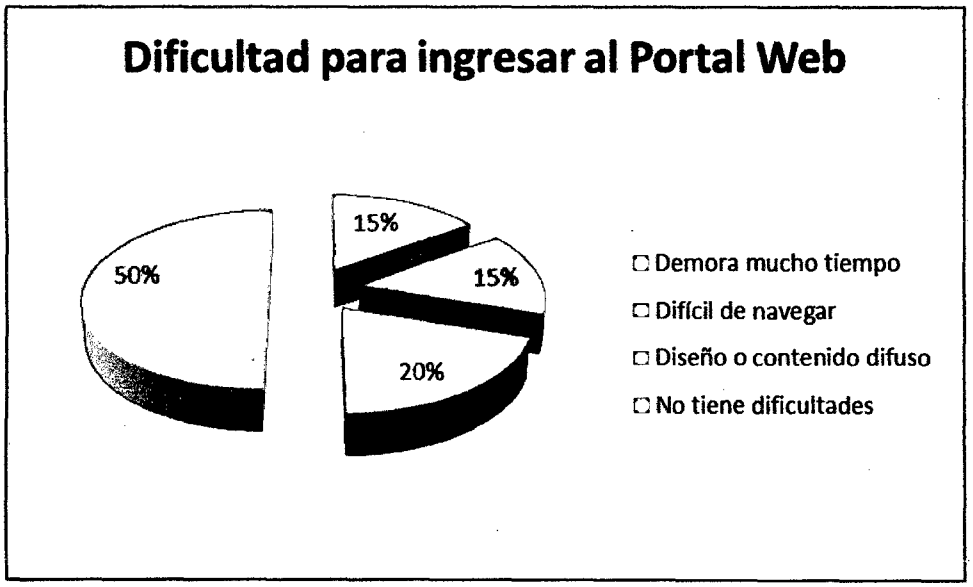


Figura VIII.9: Dificultades para ingresar al Portal Web, después de la mejora

DESCRIPCIÓN–INTERPRETACIÓN

Según el cuadro estadístico (*Cuadro VIII.13*) hay un 50% de usuarios que mencionan no tener dificultades para ingresar hacia los servicios del Portal Web ésto es por la mejora en el Portal Web que se realizaron, este porcentaje aumentó en un 40% con respecto a 10% que se tenía cuando el Portal no fue modificado (*Cuadro VIII.12*), mientras que la dificultad de "*Demora mucho tiempo*" tiene un 15% que es 30% menos respecto a un 45% que cuando no se hicieron las modificaciones el Portal Web, sobre la *"Dificultad de navegar"* tiene un 15% que 10% menos respecto a un 25% cuando no se hicieron las modificaciones del Portal Web y finalmente la

dificultad "*Diseño y contenido difuso*" no tuvo variación y se conserva en 20%.

d) SOBRE LA PREGUNTA: TIEMPO PARA ENCONTRAR EL SERVICIO DESEADO EN EL PORTAL WEB.

ANTES DE LA MEJORA DEL PORTAL

Rango de Problemas	Cantidad	%
Menos de un minuto	5	25.00
Más de un minuto	8	40.00
Más de dos minutos	7	35.00
Total	20	100.00

Cuadro VIII.14: Tiempo para encontrar el servicio deseado en el Portal Web, antes de la mejora.

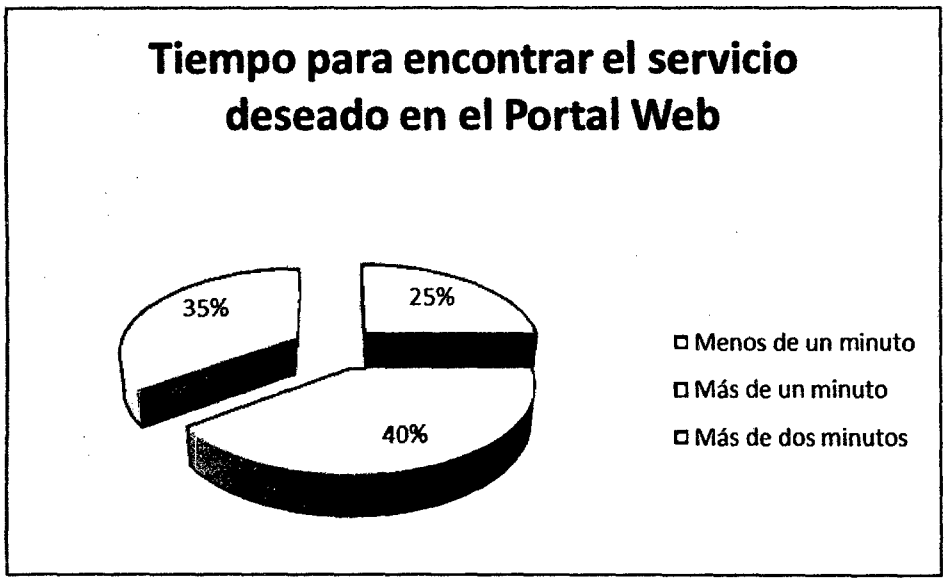


Figura VIII.10: Tiempo para encontrar el servicio deseado en el Portal Web, antes de la mejora.

DESCRIPCIÓN–INTERPRETACIÓN

Según el cuadro estadístico (*Cuadro VIII.14*) hay un porcentaje de 40% de usuarios que se demoran entre 1 a 2 minutos en encontrar el servicio que quieren utilizar en el Portal Web, un 35 % se demora más de 2 minutos y finalmente hay un 25% de usuarios que se demoran menos de un minuto.

DESPUÉS DE LA MEJORA DEL PORTAL

Rango de Problemas	Cantidad	%
Menos de un minutos	14	70.00
Más de un minuto	4	20.00
Más de dos minutos	2	10.00
Total	20	100.00

Cuadro VIII.15: Tiempo para encontrar el servicio deseado en el Portal Web, después de la mejora.

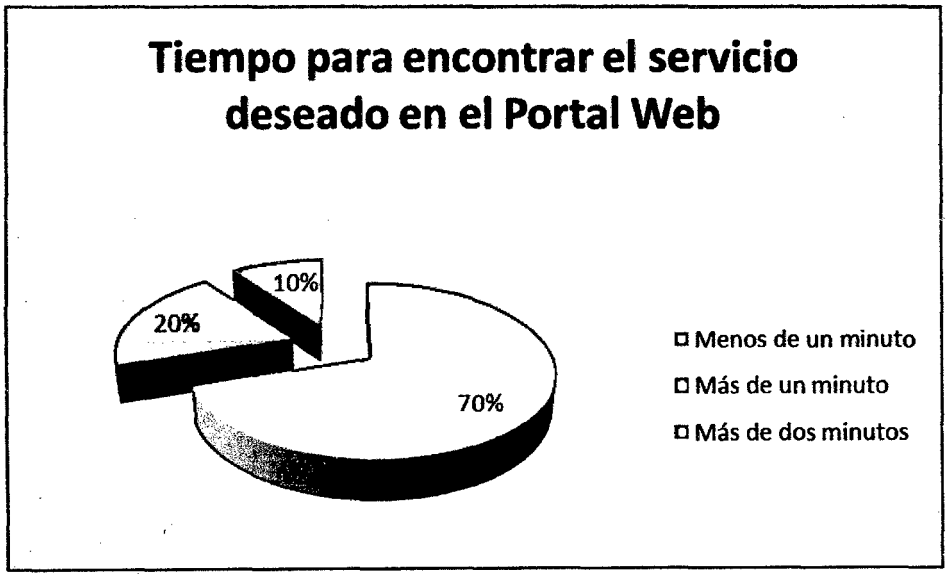


Figura VIII.11: Tiempo para encontrar el servicio deseado en el Portal Web, después de la mejora.

DESCRIPCIÓN–INTERPRETACIÓN

Según el cuadro estadístico(*Cuadro VIII.15*) hay un 70% de usuarios que se demoraron *"menos de un minuto"* para encontrar el servicio deseado en el Portal Web, este porcentaje aumentó en un 50% con respecto al 25% que se tenía cuando el portal no fue modificado (ver *Cuadro VIII.14*) con esto se da validez a la hipótesis:

"Si se mejora un Portal Web a través de la personalización de dicho Portal Web para cada Cluster, esto gracias a los patrones de comportamiento de los usuarios, ***entonces habrá una disminución en el tiempo de acceso hacia los servicios del Portal Web*** *así también un aumento en el nivel de satisfacción de los usuarios".*

Mientras que *"más de un minuto"* tiene un 20%, que es 20% menos respecto a un 40% que se registró antes de los cambios y finalmente el 10% se *"demora más de dos minutos"*, que es 25% menos respecto a un 35% que se registró antes de los cambios.

e) SOBRE LA PREGUNTA: NIVEL DE SATISFACCIÓN AL USAR EL PORTAL WEB DE LA SUNAT.

ANTES DE LA MEJORA DEL PORTAL

Rango de Problemas	Cantidad	%
Malo	5	25.00
Regular	8	40.00
Bueno	4	20.00
Muy bueno	3	15.00
Total	20	100.00

Cuadro VIII.16: Nivel de Satisfacción al usar el Portal Web, antes de la mejora.

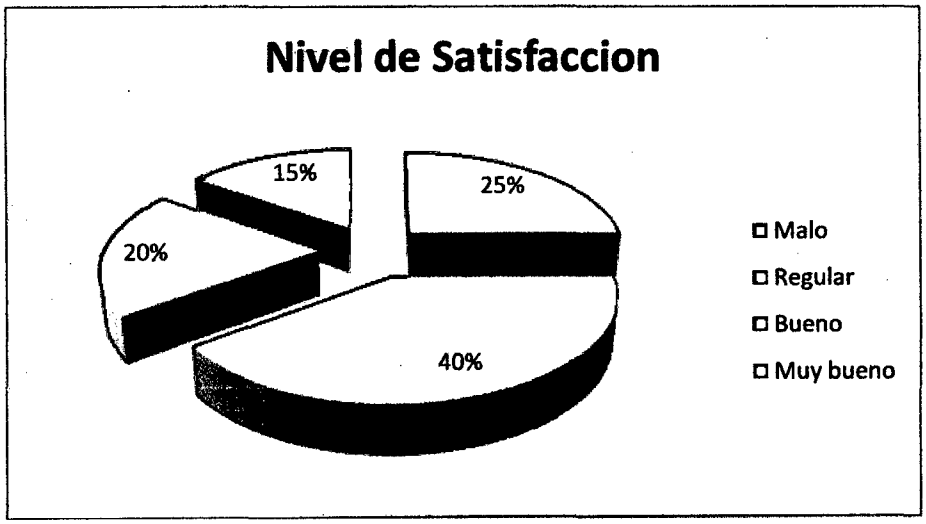


Figura VIII.12: Nivel de Satisfacción al usar el Portal Web, antes de la mejora.

DESCRIPCIÓN–INTERPRETACIÓN

Según el cuadro estadístico (*Cuadro VIII.16*) un 40% de los usuarios no se sienten *"muy satisfecho"* con el uso del Portal Web, un 25 % consideran *"malo"* o no están satisfechos, un 20% consideran que es *"bueno"* o que están satisfechos y apenas 15% consideran *"muy bueno"* o que están muy satisfechos.

DESPUÉS DE LA MEJORA DEL PORTAL

Rango de Problemas	Cantidad	%
Malo	2	10.00
Regular	4	20.00
Bueno	8	40.00
Muy bueno	6	30.00
Total	20	100.00

Cuadro VIII.17: Nivel de Satisfacción al usar el Portal Web, después de la mejora.

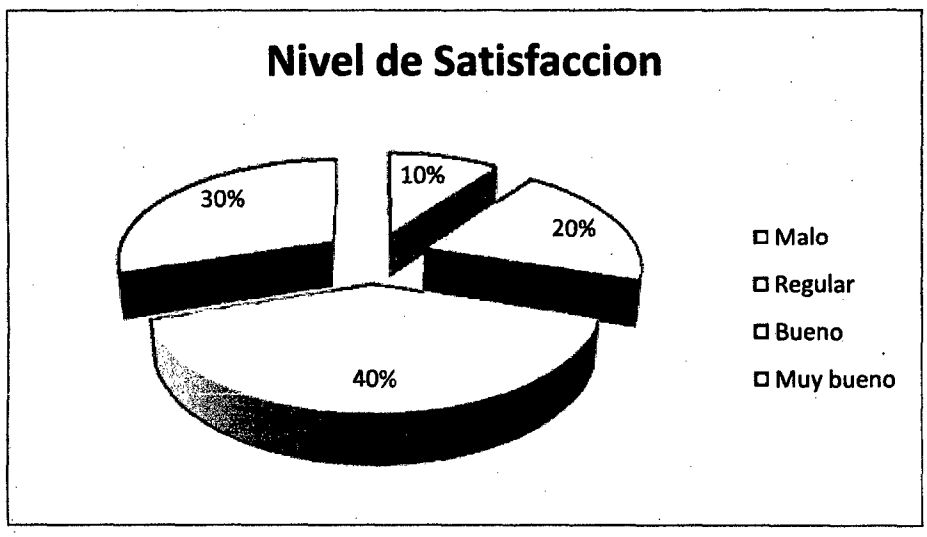


Figura VIII.13: Nivel de Satisfacción al usar el Portal Web, después de la mejora.

DESCRIPCIÓN – INTERPRETACIÓN

Según el cuadro estadístico (*Cuadro VIII.17*) hay un 40% que consideran *"bueno"* o que están satisfechos (este porcentaje aumento en un 20% con respecto al 20% que se tenía cuando el Portal Web no fue mejorada), un 30% considera *"muy bueno"* o que están muy satisfechos, un 20% consideran que es *"regular"* y sólo el 10% consideran malo. Con esto se da validez a la hipótesis:

"Si se mejora un Portal Web a través de la personalización de dicho Portal Web para cada Cluster, esto gracias a los patrones de comportamiento de los usuarios, entonces habrá una disminución en el tiempo de acceso

hacia los servicios del Portal Web así también ***un aumento en el nivel de satisfacción de los usuarios"****.*

CAPÍTULO IX

EXPERIMENTACIÓN

El experimento consistió en personalizar el Portal Web teniendo en cuenta los patrones identificados para cada *Clúster* en el presente estudio, los patrones están definidos por los diferentes servicios que acceden los usuarios en los tres grupos o *Clusters* identificados.

La personalización del Portal Web estará representada por la evocación de un *Pop Up* (ventana emergente) una vez que se ingrese al Portal Web, en este *Pop Up* estarán los servicios más visitados por cada grupo o *Cluster* al cual corresponda el usuario.

El usuario está identificado por una IP y esta IP está mapeada a un *Cluster* (representado por un grupo de servicios del Portal Web), cuando un usuario

ingrese al Portal Web se leerá primero la IP del usuario que realiza la consulta, luego con esta IP se buscará a qué *Cluster* pertenece y una vez que se identifique el *Cluster* se mostrará el correspondiente *Pop Up* al que está relacionado, con los accesos directos hacia el grupo de servicios del *Cluster*. Si es *Cluster* 1 se muestra el *Pop Up* 1, Si es *Cluster* 2 se muestra el *Pop Up* 2 y si es *Cluster* 3 se muestra el *Pop Up* 3. Si la IP no se encuentra en la lista no se mostrará ningún *Pop Up* y simplemente se cargará el Portal Web principal.

Nº de *Cluster*	**Portal**
1	Portal–*Pop Up*1
2	Portal–*Pop Up*2
3	Portal–*Pop Up*3

Cuadro IX.1 Relación *Cluster* y Portal Web con el *Pop Up* correspondiente.

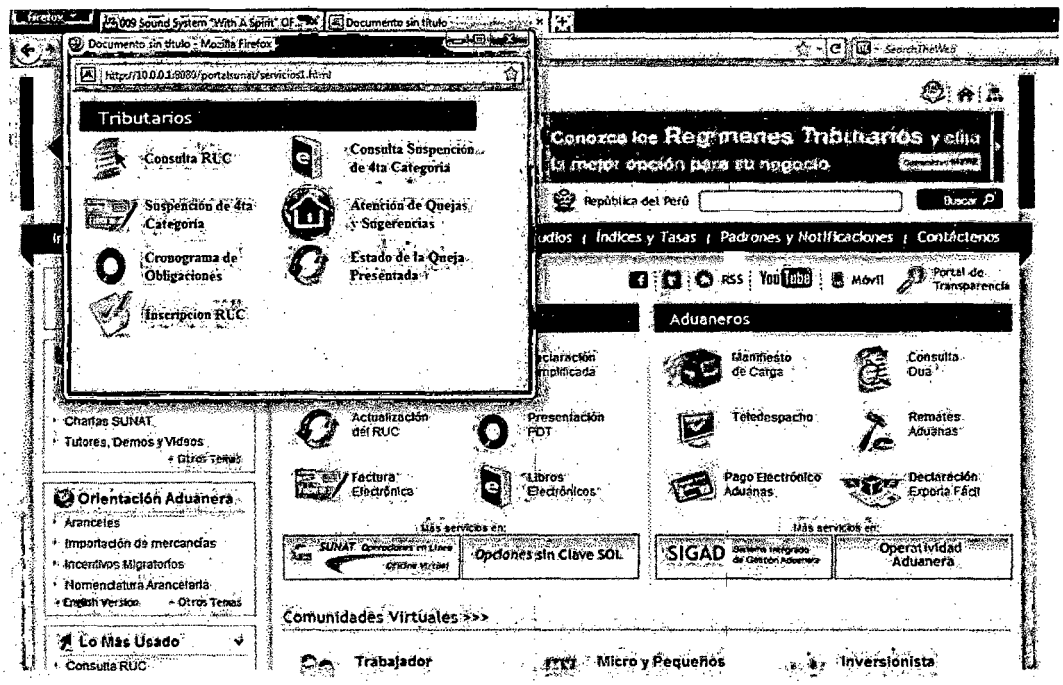


Figura IX.1 Invocación del *Pop Up* 1 en el Portal Web.

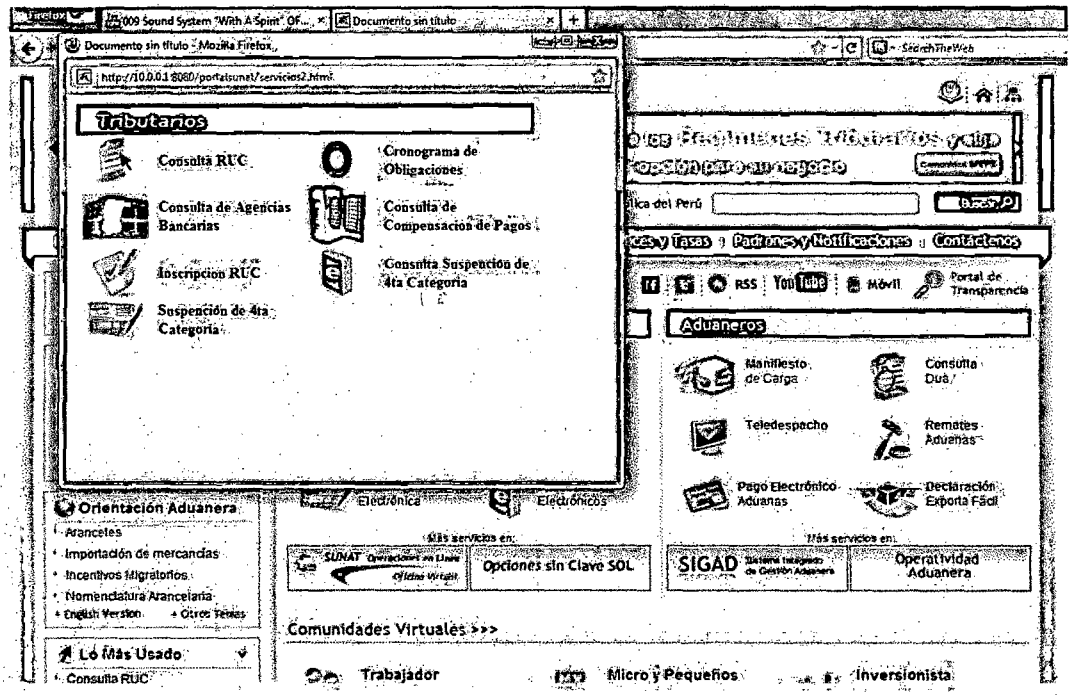


Figura IX.2 Invocación del *Pop Up* 2 en el Portal Web.

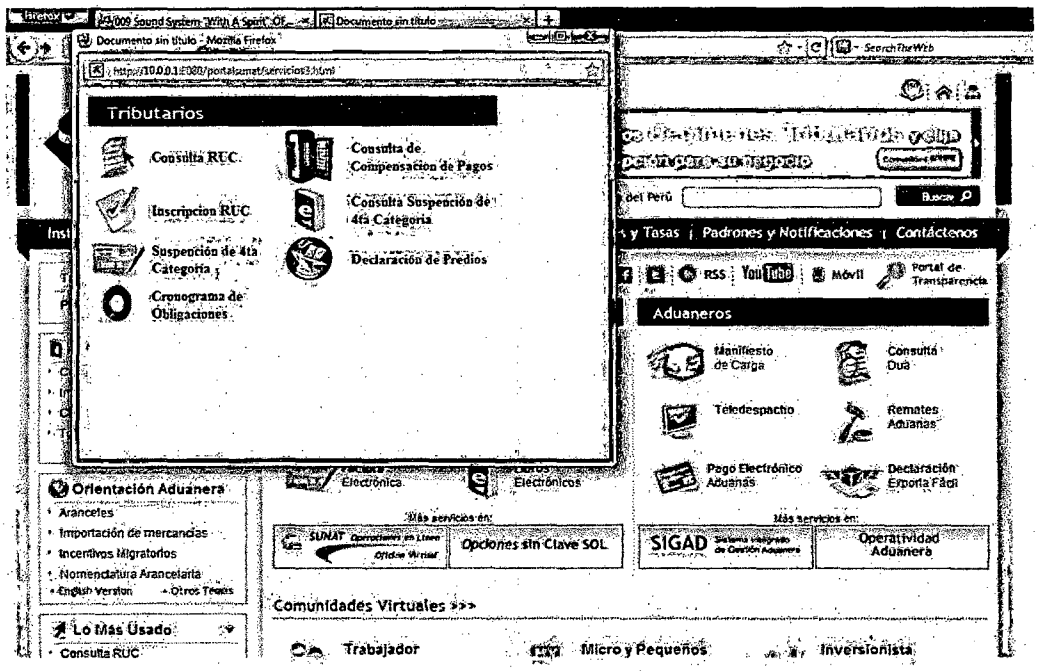


Figura IX.3 Invocación del Pop Up 3 en el Portal Web.

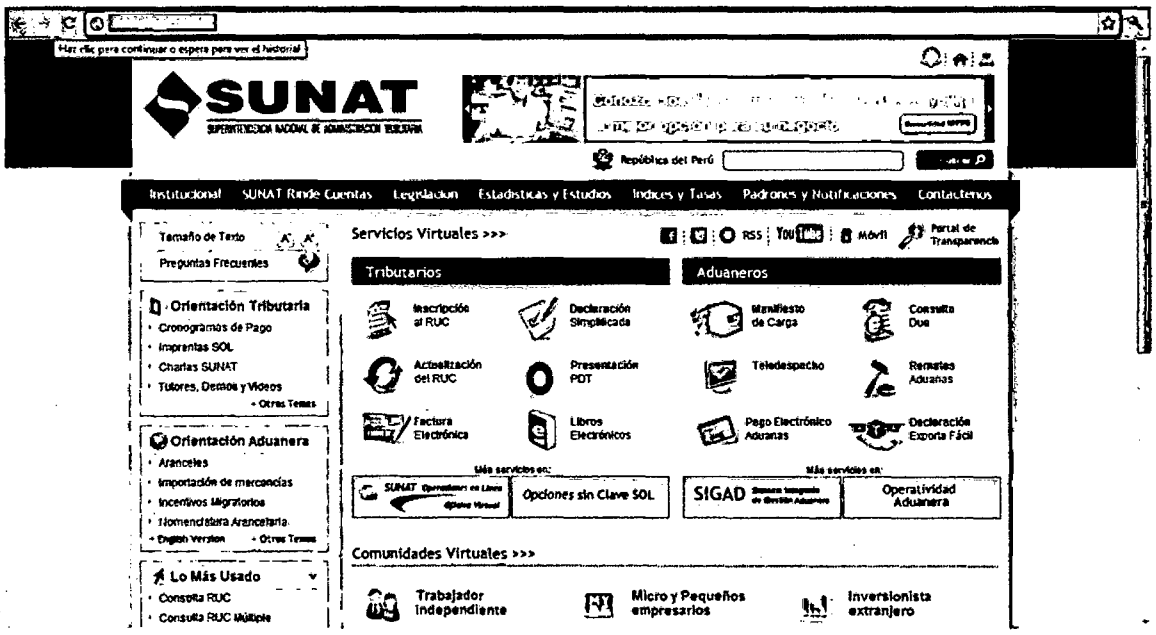


Figura IX.4 Portal Web Principal si ningún *Pop Up*.

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

1. Los patrones de comportamiento de los usuarios en el Portal Web de la SUNAT, están determinados por los grupos o *Clusters* identificados en el estudio.
2. La cantidad óptima de grupos o *Clusters* que se halló con el experimento es de 3 *Clusters*.
3. Se determinó los siguientes patrones.

   Patrón A.- Usuarios que mayormente usan los siguientes servicios.

   - Consulta RUC.
   - Suspensión de 4ta Categoría-Formulario 1609.
   - Cronograma de obligaciones mensuales - ejercicio 2011.

- Inscripción de RUC.
- Consulta de solicitudes de suspensiones de 4ta-categoría Formulario 1609.
- Atención de Quejas y Sugerencias.
- Estado de la Queja Presentada.

Patrón B.- Usuarios que mayormente usan los siguientes servicios.

- Consulta RUC.
- Consulta de Agencias Bancarias a nivel Nacional.
- Inscripción del RUC.
- Suspensión de 4ta Categoría-Formulario 1609.
- Cronograma de Obligaciones Mensuales.
- Consulta de Compensación de Pagos.
- Consulta de solicitudes de suspensiones de 4ta-categoría Formulario 1609.

Patrón C.- Usuarios que mayormente usan los siguientes servicios.

- Consulta RUC.
- Inscripción del RUC.
- Suspensión de 4ta Categoría-Formulario 1609.
- Cronograma de Obligaciones Mensuales.
- Consulta de Compensación de Pagos.
- Consulta de solicitud de suspensiones de 4ta. Categoría-Formulario 1609.
- Declaraciónón de Predios Formulario Virtual 1630.

4. Hay un 70% de usuarios que se demoraron menos de un minuto para encontrar los servicios del Portal Web de la SUNAT.

Rango de Problemas	***Cantidad***	***%***
Menos de un minutos	***14***	***70.00***
Más de un minuto	*4*	*20.00*
Más de dos minutos	*2*	*10.00*
Total	*20*	*100.00*

Cuadro VIII.15: Tiempo para encontrar el servicio deseado en el Portal Web, después de la mejora.

Este porcentaje aumento en un 50% con respecto al 25% que se tenía cuando el portal no fue modificado.

Rango de Problemas	***Cantidad***	***%***
Menos de un minuto	***5***	***25.00***
Más de un minuto	*8*	*40.00*
Más de dos minutos	*7*	*35.00*
Total	*20*	*100.00*

Cuadro VIII.14: Tiempo para encontrar el servicio deseado en el Portal Web, antes de la mejora.

Con ésto se da validez a la hipótesis.

"Si se mejora un Portal Web a través de la personalización de dicho Portal Web para cada Cluster, esto gracias a los patrones de comportamiento de los usuarios, ***entonces habrá una disminución en el tiempo de acceso hacia los servicios del Portal Web*** *así también un aumento en el nivel de satisfacción de los usuarios."*

5. Hay un 40% de usuarios que mencionan que el Portal Web de la SUNAT brinda un servicio bueno.

Rango de Problemas	***Cantidad***	***%***
Malo	*2*	*10.00*
Regular	*4*	*20.00*
Bueno	***8***	***40.00***
Muy bueno	*6*	*30.00*
Total	*20*	*100.00*

Cuadro VIII.17: Nivel de Satisfacción al usar el Portal Web, después de la mejora.

Este porcentaje aumento en un 20% con respecto al 20% que se tenía cuando el Portal Web no fue modificado

Rango de Problemas	***Cantidad***	***%***
Malo	*5*	*25.00*
Regular	*8*	*40.00*
Bueno	*4*	*20.00*
Muy bueno	*3*	*15.00*
Total	*20*	*100.00*

Cuadro VIII.16: Nivel de Satisfacción al usar el Portal Web, antes de la mejora.

Con esto se da validez a la hipótesis.

"Si se mejora un Portal Web a través de la personalización de dicho Portal Web para cada Cluster, esto gracias a los patrones de comportamiento de los usuarios, entonces habrá una disminución en el tiempo de acceso hacia los servicios del Portal Web así también un ***aumento en el nivel de satisfacción de los usuarios****."*

RECOMENDACIONES

1. Recomendamos utilizar Data Mining en el estudio estadístico de data no estructurada, ya que ésta permite la identificación de patrones que no son fáciles de identificar por los paquetes estadísticos tradicionales: ya que existen muchas variables y la escalabilidad de la misma no es soportada por aquellos paquetes estadísticos.
2. A las empresas peruanas que tienen un Portal Web o Página Web les recomendamos hacer un estudio de los archivos *log* que tienen en sus servidores Web, a través de la metodología Web Mining: porque dichos datos registran las preferencias que tienen los usuarios por algún(os) módulo(s) en particular del Portal Web. Si se procesan los mencionados archivos *log* se obtendría información sobre el patrón de comportamiento de los usuarios, lo cual es muy importante para la toma de decisiones a la hora planificar o realizar alguna modificación de su Portal Web, buscando con ello la personalización del Portal Web sobre la base de necesidades y preferencias de los usuarios.
3. Por último, el conocimiento de los grupos o *Cluster*, podrían ser utilizados para la creación de nuevos productos y/o servicios o para ofrecer publicidad personalizada sobre la base de las preferencias y/o necesidades de un grupo, logrando así una mayor aceptación.

BIBLIOGRAFÍA

[1] World Wide Web Consortium [http://www.w3.org/].

[2] Nielsen, Jakob, Usability Engineering.: Morgan Kaufmann, San Francisco (1993).

[3] Oracle Business Intelligence [www.oracle.com/bi].

[4] Scotto, M. Sillitti, A. Succi, G. Vernazza, T. "Managing Web-Based Information", International Conference on Enterprise Information Systems (ICEIS 2004), Porto, Portugal, April 2004. Page 1-3.

[5] Baeza-Yates, R. Pobrete, B. "Una herramienta de minería de consultas para el diseño del contenido y la estructura de un sitio Web "Actas del III Taller Nacional de Minería de Datos y Aprendizaje TAMIDA2005", pp.39-48, 2005.

[6] Kosala, R. and Blockeel, H. Web Mining Research: A Survey. ACM SIGKDD Explorations, Newsletter of the Special Interest Group on

Knowledge Discovery and Data Mining. Page 1-9, 2000.

[7] Galeas, P. Web Mining by Patricio Galeas. http://www.galeas.de/webmining.html.

[8] Fränti y Kivijärvi, Randomised Local Search Algorithm for the Clustering Problem, Pag. 360.

[9] Balaji Vasan Srinivasan, Department of Computer Science, University of Maryland, College Park, USA Random Sampling For Estimating The Performance Of Fast Summations(59).

[10] Frequent Pattern Mining in Web Log Data : Budapest University of Technology and Economics; Goldmann Gy. tér 3, H-1111 Budapest, Hungary Autor: Renata Ivancsy, Istvan Vajk Pag. 81.

[11] Catledge y Pitkow, Characterizing Browsing Strategies in the World-Wide Web, Computer Networks and ISDN Systems, 27(6): 1065–1073, Abril. Pág.4.

[12] A.K. Jain, M.N. Murty and P.J. Flynn: Data *Clustering*: A Review Pag. 273.

[13] Erick Vicente, Luis Rivera y David Mauricio; Grasp en la Resolución del problema de *Clustering*, Universidad Nacional de San Marcos – Lima Perú, 2005, Pág. 22.

[14] Francisco Manuel de Gyves, Web Mining: Fundamentos Básicos. Universidad de Salamanca, Pág. 5.

ANEXOS

ANEXO 1: INSTRUMENTO DE EVALUACIÓN

INSTRUCCIONES

Estimado usuario, conteste con sinceridad el presente cuestionario porque los resultados servirán para un estudio para el mejoramiento del Portal de la SUNAT, para una mejor prestación de los servicios.

ENCUESTA PARA IDENTIFICAR EL TIEMPO DE USO DEL PORTAL WEB DE LA SUNAT EN UNA SESIÓN DE UN USUARIO

1. ¿Conoce el portal de la SUNAT?	**2. ¿Ha realizado alguna transacción o consulta por el portal de la SUNAT?**
Sí 1 (pas a 2) No 2 (termina)	Sí 1 (pas a 3) No 2 (termina)

3 ¿Cuánto tiempo en promedio se queda usando un servicio en el portal de la SUNAT?	
Menos de diez minutos	1
Más de diez minutos	2
Menos de veinte minutos	3
Más de veinte minutos	4
Menos de treinta minutos	5
Más de treinta minutos	6

ENCUESTA PARA IDENTIFICAR EL TIEMPO DE ACCESO Y EL NIVEL DE SATISFACCIÓN EN EL USO DE LOS SERVICIOS DEL PORTAL DE LA SUNAT

1. ¿Conoce el portal de la SUNAT?	**2. ¿Ha realizado alguna transacción o consulta por el portal de la SUNAT?**
Sí 1 (Pasa 2) No 2 (termina)	Sí 1 (Pasa 3) No 2 (termina)
3. ¿Con que frecuencia a la semana ha realizado alguna transacción o consulta por el portal de la SUNAT?	**4. ¿Qué dificultades tuvo al ingresar al portal de la SUNAT?**
Una vez: 1 Más de tres veces: 4 Dos veces: 2 Tres veces: 3	Demora mucho tiempo: 1 Solo tiene información: 2 Diseño o contenido difuso: 3 No tiene dificultades: 4
4. ¿Cuánto tiempo ha demorado en encontrar el servicio en el portal de la SUNAT?	**5. ¿Cuánto tiempo en promedio se queda usando un servicio en el portal de la SUNAT?**
Menos de un minuto: 1 Más de un minuto y menos de dos minutos: 2 Más de dos minutos: 3	Menos de diez minutos: 1 Más de diez minutos: 2 Menos de veinte minutos: 3 Más de veinte minutos: 4 Menos de treinta minutos: 5 Más de treinta minutos: 6

6. ¿Qué servicio de la lista es la que más ha usado?	**7. ¿Entrando al portal de la SUNAT satisfizo sus necesidades? ¿Es decir encontró lo que buscaba? ¿Cómo lo definiría?**
Consulta RUC: 1 Formularios y Solicitudes varias: 2 Suspensión de retenciones: 3 Otros: 4	Malo: 1 Regular: 2 Bueno: 3 Muy bueno: 4

ANEXO 2: DESCRIPCIÓN DEL SISTEMA, BAJO LA METODÓLOGIA RUP

1. MODELO DEL NEGOCIO

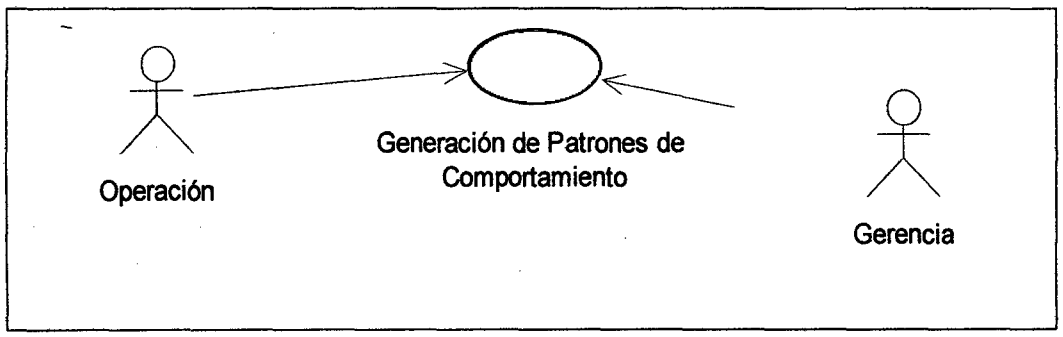
1.1 MODELAMIENTO DE PROCESOS DE NEGOCIO

1.1.1 IDENTIFICACIÓN DE LOS PROCESOS DEL NEGOCIO

- Casos de Uso de Negocio.

Número	Proceso de Negocio
CUN-001	Reconocimiento de Patrones.

- Diagrama de Casos de Uso del Negocio.



1.1.2 IDENTIFICACIÓN DE LOS ACTORES DEL ENTORNO DEL NEGOCIO

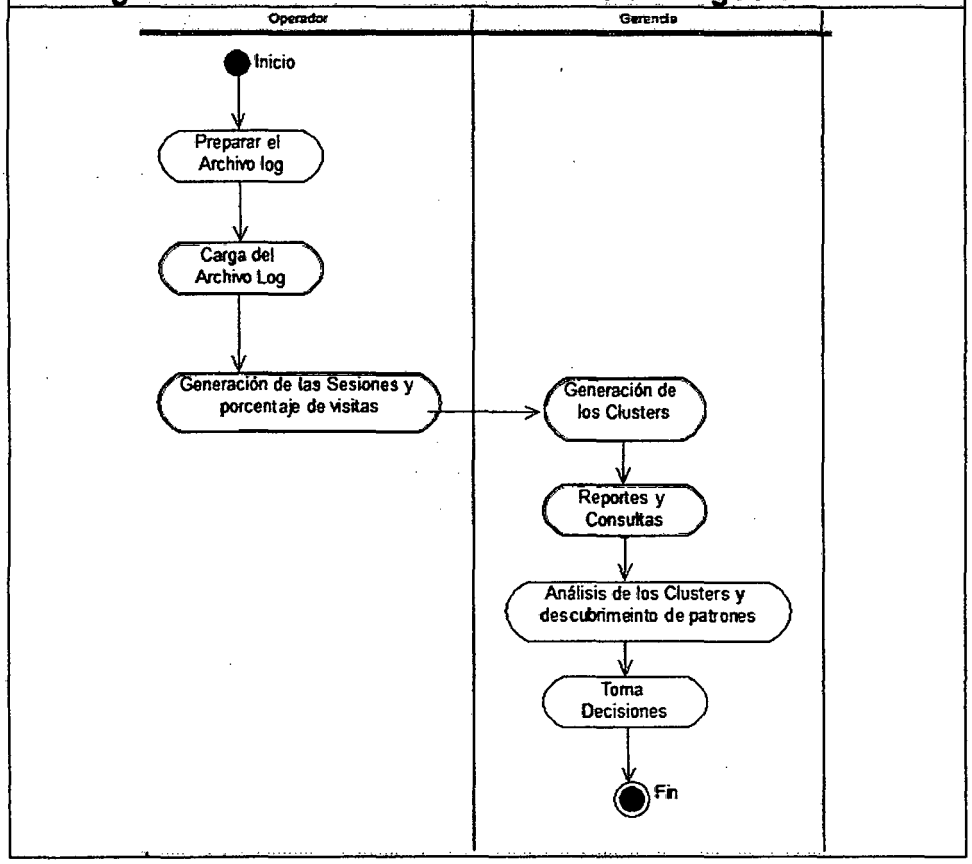
Identificación de los Actores

Número	Actor	Roles/Responsabilidades
ACT-001	Gerencia.	Es la persona o trabajador de la organización que requiere el conocimiento de patrones para tomar decisión.
ACT-002	Operación.	Es la persona o trabajador de la organización que se encargar de cargar y generar los reportes.

1.1.3 DESCRIPCIÓN DE LOS CASOS DE USO DEL NEGOCIO

1.- Proceso de Negocio	Generación de Patrones de Comportamiento
2.- Objetivo	Generar los patrones de comportamiento de usuarios web, para así tener conocimiento y hacer tomada de decisiones en la mejora de la aceptación de los servicios.
3.- Actores	Gerencia y Operador
4.- Precondiciones	
5.- Flujo de Eventos	
1. El Operador prepara el archivo *log.* 2. El Operador Cargar el archivo *log*, se carga en una tabla haciendo la limpieza. 3. El Operado genera las sesiones y el porcentaje de visitas a las dimensiones. 4. La Gerencia genera los *Clusters* con los parámetros considerados. 5. La Gerencia visualiza los reportes. 6. La Gerencia analiza los *Clusters* y descubre patrones. 7. Toma decisión sobre la base del conocimiento de los patrones. 8. Se finaliza.	

6.- Pos condiciones	
7.- Excepciones	
8.- Versión	1.0
9.- Tiempo de Ejecución	
10. Diagrama de Actividades del Proceso de Negocio	



1.1.4 ESPECIFICACIÓN DE REGLAS DE NEGOCIO

GLOSARIO DE TÉRMINOS

Término	Descripción
Gerencia.	Área de a organización que toma decisión sobre la base del análisis y conocimiento de la empresa de caras a la mejora en el servicio al cliente.
Operación.	Área de la organización que se encargar de actualizar la bases de datos y los reportes.

CATÁLOGO DE REGLAS DEL NEGOCIO

Regla del Negocio	Descripción
RN-001	El archivo log a cargar debe tener el formato estándar de accesos a página web.
RN-002	El número mino de visitas debe ser mayor o igual a 3.
RN-003	El tiempo máximo de sesión será menor o igual a 30 minutos.

2. ANÁLISIS DE SISTEMAS

2.1 REQUERIMIENTOS DEL SISTEMA DE INFORMACION

2.1.1 OBTENCIÓN DE REQUERIMIENTOS

Requerimientos Funcionales

Número	Requerimiento	Descripción	Prioridad
RF-001	Cargar el archivo *Log*	El sistema permitirá limpiara y cargar el archivo *log* en una tabla.	Alta
RF-002	Definir las Dimensiones	De la carga del archivo *log* el sistema permitirá encontrar las dimensiones y registrarla en una tabla.	Alta
RF-003	Generar Sesiones	El sistema permitirá generar las sesiones con los parámetros de tiempo máximo sesión y la cantidad mínima de visitas y registrar en una tabla de sesiones.	Alta
RF-004	Determinar el Porcentaje de visitas.	El sistema permitirá calcular el porcentaje de visitas por dimensión de cada una de las sesiones y registrarla en una tabla de cantidad de visitas.	Alta
RF-005	Seleccionar la muestra representativa.	El sistema debe permitir seleccionar una muestra representativa y registrar en una tabla.	Alta
RF-006	Generar *Clusters*.	El sistema debe permitir generar *Clusters* con el parámetro de número de *Cluster* y actualizar la tabla de sesiones, con el número de *Cluster* a que pertenece.	Alta

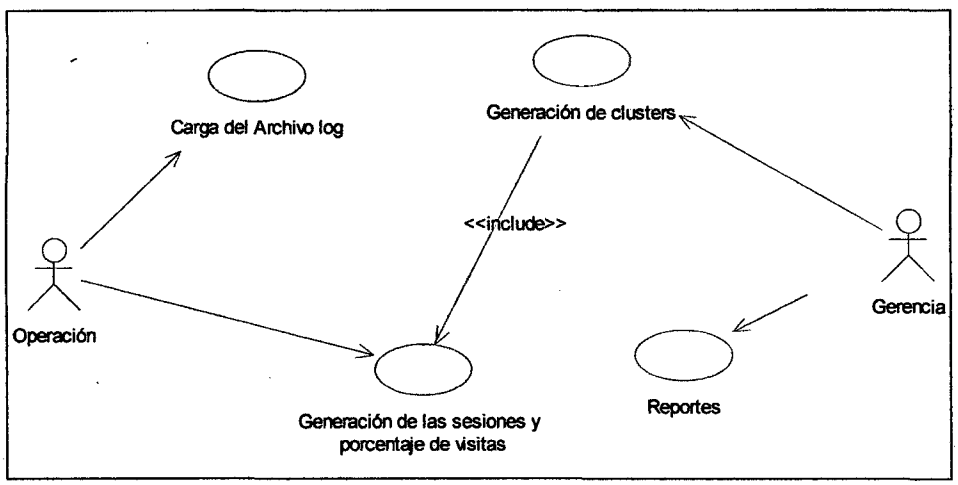
RF-007	Reportes.	El sistema debe permitir genera los siguientes reportes: • Dimensiones. • Sesiones. • Vector Posición. • *Clusters.* • Diferencia de Centroides. • IPsx *Cluster.* • *Cluster*_1 (Centroide). • *Cluster*_2 (Centroide). • *Cluster*_3 (Centroide).	

Requerimientos No Funcionales (RNF)

Número	Requerimiento	Descripción	Prioridad
RNF-001	Seguridad.	El sistema debe restringir las operaciones a realizar por el usuario, de acuerdo a su nivel de acceso.	Alta
RNF-002	Desempeño.	El sistema contará con un manual de usuario para un mejor manejo de las opciones y una mayor eficiencia.	Alta
RNF-003	Configuración.	El sistema debe ser construido e implantado de tal manera que un cambio en los parámetros del negocio no obligue a la generación de una nueva versión.	Alta
RNF-004	Escalabilidad.	El sistema debe estar en capacidad de permitir en el futuro el desarrollo de nuevas funcionalidades, modificar o eliminar funcionalidades después de su construcción y puesta en marcha inicial.	Alta
RNF-005	Mantenibilidad.	El sistema debe contar con una interfaz de administración que incluya: Administración de usuarios, Administración de módulos y Administración de parámetros. En cada una de éstas secciones deberá ofrecer todas las opciones de administración disponibles para cada uno.	Media

2.1.3 OBTENCIÓN DEL MODELO DE CASOS DE USO DEL SISTEMA

Diagrama de Casos de Uso del Sistema



Descripción de Casos de Uso del Sistema

1.- Caso de Uso del Sistema	CARGA DEL ARCHIVO *LOG*
2.- Descripción del caso de uso	
El sistema verifica si tiene la estructura establecida de carga del archivo que se desea cargar, luego filtra sólo aquellos que cumplan con las reglas establecidas y finalmente parsear y registrar en una tabla.	
3.- Actor(es)	
Operación.	
4.- Precondiciones	
Tener el archivo *log*.	
5.- Postcondiciones	
• Registro de Archivo *log* Limpio.	

6.- Pasos (Flujo de Eventos)

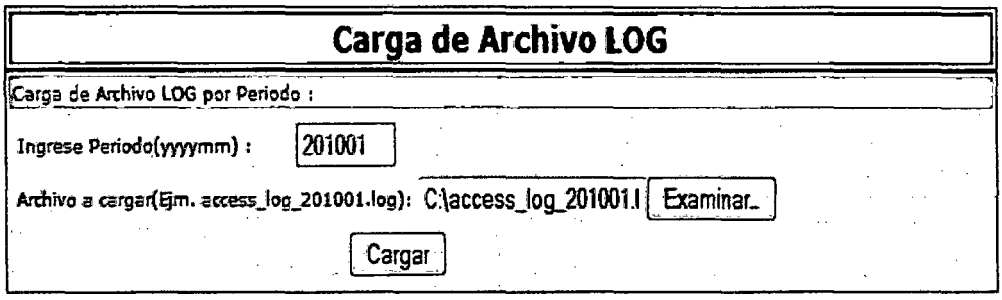
Nº	Acción del Actor	Respuesta del Sistema
1	Recibe, verifica el archivo *log* e ingresa al modulo de carga.	Muestra la pantalla de Carga del archivo *Log*.
2	Ingresa los parámetros, selecciona el archivo a cargar y clic en cargar.	El sistema lee los parámetros y el archivo log, procesa e inserta en la tabla t1archlogfiltrado, y al final le muestra el mensaje del estado de la ejecución. • Mensaje de Error. • Mensaje de Carga Exitosa.
3	Confirma el Mensaje, si: • La carga fue exitosa: Cierra • La carga fue con error: reporta el error y Cierra.	Se cierra la ventana.

7.- Requerimiento asociado

RF-001

8.- Prototipo de interfaz de usuario

IU-003



Inicio

9.- Casos de Uso Alternos

CARGA T2DIMENSION

```
DELIMITER $$
                DROP PROCEDURE IF EXISTS
                `mydb`.`proc_dimensiones` $$
                CREATE DEFINER=`root`@`localhost`
                PROCEDURE `proc_dimensiones`(periodo CHAR(6),
                usuario VARCHAR(20))
                BEGIN
                DECLARE urlpage VARCHAR(400);
                DECLARE no_more_dimension INT;
                DECLARE cant INT;
                DECLARE cur_dimension CURSOR FOR
                SELECT DISTINCT t1_urlpage FROM
                mydb.t1archlogfiltrado WHERE t1_periodo=periodo;
                DECLARE CONTINUE HANDLER FOR NOT FOUND
                SET no_more_dimension=1;
                SET no_more_dimension = 0;
                   IF NOT EXISTS (SELECT * FROM t2dimension)
                THEN
                   SET cant = 0;
                   ELSE
                   SET cant=(SELECT
                MAX(t2dimension.t2_nrodimension) FROM t2dimension);
                   END IF;
                OPEN cur_dimension;
                dimension_loop:WHILE(no_more_dimension=0) DO
                 FETCH cur_dimension INTO urlpage;
                  SET cant=cant+1;
                 IF no_more_dimension=1 THEN
                   LEAVE dimension_loop;
                 END IF;
                 IF NOT EXISTS(SELECT *
                FROM t2dimension WHERE
                t2dimension.t2_valordimension=urlpage) THEN
                      INSERT INTO t2dimension
                VALUES(null,cant,periodo, urlpage,null,NOW(),usuario);
                 END IF;
                END WHILE dimension_loop;
                CLOSE cur_dimension;
                SET no_more_dimension = 0;
                END $$

                DELIMITER ;
```

1.- Caso de Uso del Sistema	GENERACIÓN DE LAS SESIONES Y LOS PORCENTAJE DE LAS VISITAS
2.- Descripción del caso de uso	
El sistema permitirá generar las sesiones y los vectores posición. Para ésto primero actualizará la tabla de dimensiones si en la última carga se agregaron nuevas dimensiones entonces actualizará la tabla de dimensiones, luego determinará las sesiones, cantidad de visitas y el vector posición según el tamaño de la muestra, para ésto es necesario ingresar los siguientes parámetros: Mínima cantidad de visitas. Tiempo máximo de sesiones. Tamaño de la muestra. Y el intervalo de periodo y para nuestro estudio se determinó que será para un periodo de un año	
3.- Actor(es)	
4.- Precondiciones	
Que se haya cargado el archivo *log* en la t1archloglimpio para el periodo a procesar.	
5.- Postcondiciones	
Tabla de las Dimensiones actualizada(t2dimension) Tabla de Sesiones actualizad(t3ipsesion) Tabla de cantidad de visitas actualizada(t4cantidadvisit) Tabla de Muestra de Sesiones(t5muestrasesion) Tabla de Vector posición actualizada(t6vectorposicion)	

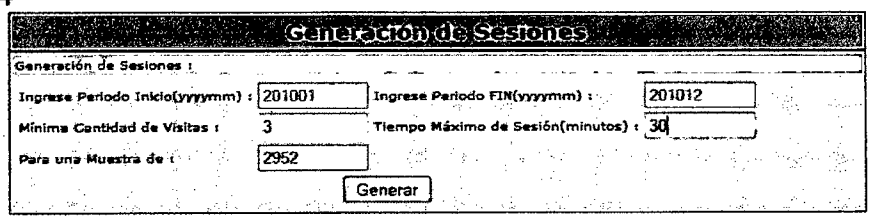
6.- Pasos (Flujo de Eventos)		
Nº	**Acción del Actor**	**Respuesta del Sistema**
1	Ingresa al módulo de generación de sesión.	Muestra la pantalla de Generación de Sesiones.
2	Ingresa los parámetros y clic en generar.	El sistema procesa y genera las sesiones, insertando en las tablas t3ipsesion, t4cantidadvisita, t5muestrasesion, t6vectorposicion y al final le muestra el mensaje del estado de la ejecución. • Mensaje de Error. • Mensaje de Carga Exitosa.
3	Confirma el Mensaje, si: • La carga fue exitosa Cierra. • La carga fue con error, reporta el error y Cierra.	Se cierra la ventana.

7.- Requerimiento asociado

RF-002, RF-003, RF-004, RF-005

8.- Prototipo de interfaz de usuario

IU-004



Inicio

9.- Casos de Uso Alternos

CARGA T2DIMENSION

```
DELIMITER $$
DROP PROCEDURE IF EXISTS `mydb`.`proc_dimensiones`
$$
CREATE DEFINER=`root`@`localhost`
PROCEDURE `proc_dimensiones`(periodo CHAR(6), usuario
VARCHAR(20))
BEGIN
DECLARE urlpage VARCHAR(400);
DECLARE no_more_dimension INT;
DECLARE cant INT;
DECLARE cur_dimension CURSOR FOR
SELECT DISTINCT t1_urlpage FROM mydb.t1archlogfiltrado
WHERE t1_periodo=periodo;
DECLARE CONTINUE HANDLER FOR NOT FOUND SET
no_more_dimension=1;
SET no_more_dimension = 0;
   IF NOT EXISTS (SELECT * FROM t2dimension) THEN
   SET cant = 0;
   ELSE
   SET cant=(SELECT MAX(t2dimension.t2_nrodimension)
FROM t2dimension);
   END IF;
OPEN cur_dimension;
dimension_loop:WHILE(no_more_dimension=0) DO
 FETCH cur_dimension INTO urlpage;
  SET cant=cant+1;
 IF no_more_dimension=1 THEN
   LEAVE dimension_loop;
 END IF;
 IF NOT EXISTS(SELECT *
FROM t2dimension WHERE
t2dimension.t2_valordimension=urlpage) THEN
  INSERT INTO t2dimension VALUES(null,cant,periodo,
urlpage,null,NOW(),usuario);
 END IF;
END WHILE dimension_loop;
CLOSE cur_dimension;
SET no_more_dimension = 0;
END $$
DELIMITER ;
```

1.- Caso de Uso del Sistema	GENERACIÓN DE LAS SESIONES Y LOS PORCENTAJE DE LAS VISITAS

2.- Descripción del caso de uso

El sistema permitirá generar las sesiones y los vectores posición. Para ésto primero actualizará la tabla de dimensiones si en la última carga se agregaron nuevas dimensiones entonces actualizará la tabla de dimensiones, luego determinará las sesiones, cantidad de visitas y el vector posición según el tamaño de la muestra, para ésto es necesario ingresar los siguientes parámetros:

Mínima cantidad de visitas.
Tiempo máximo de sesiones.
Tamaño de la muestra.

Y el intervalo de periodo y para nuestro estudio se determinó que será para un periodo de un año.

3.- Actor(es)

4.- Precondiciones

Que se haya cargado el archivo *log* en la t1archloglimpio para el periodo a procesar.

5.- Postcondiciones

Tabla de las Dimensiones actualizada(t2dimension).
Tabla de Sesiones actualizad(t3ipsesion).
Tabla de cantidad de visitas actualizada(t4cantidadvisit).
Tabla de Muestra de Sesiones(t5muestrasesion).
Tabla de Vector posición actualizada(t6vectorposicion).

6.- Pasos (Flujo de Eventos)

Nº	Acción del Actor	Respuesta del Sistema
1	Ingresa al módulo de generación de sesión	Muestra la pantalla de Generación de Sesiones
2	Ingresa los parámetros y clic en generar	El sistema procesa y genera las sesiones, insertando en las tablas t3ipsesion, t4cantidadvisita, t5muestrasesion, t6vectorposicion y al final le muestra el mensaje del estado de la ejecución. • Mensaje de Error. • Mensaje de Carga Exitosa.

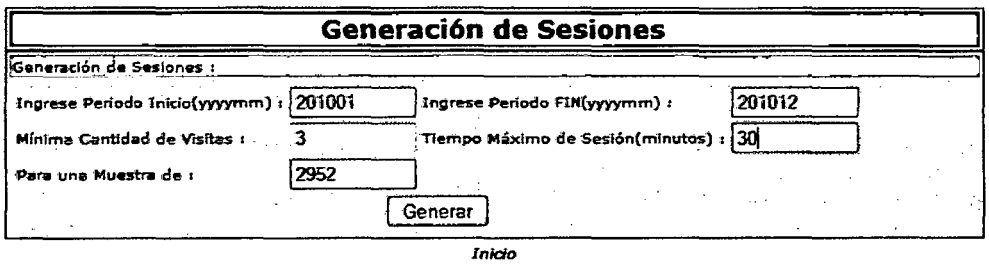
2	Ingresa los parámetros y clic en generar.	El sistema procesa y genera las sesiones, insertando en las tablas t3ipsesion, t4cantidadvisita, t5muestrasesion, t6vectorposicion y al final le muestra el mensaje del estado de la ejecución. • Mensaje de Error. • Mensaje de Carga Exitosa.
3	Confirma el Mensaje, si: • La carga fue exitosa: Cierra. • La carga fue con error: reporta el error y Cierra.	Se cierra la ventana.

7.- Requerimiento asociado

RF-002, RF-003, RF-004, RF-005

8.- Prototipo de interfaz de usuario

IU-004



Inicio

9.- Casos de Uso Alternos

CARGA: T4CANTIDADVISITA Y T3IPSESION

```
DELIMITER $$
DROP PROCEDURE IF EXISTS `mydb`.`proc_sesionvalor`
$$
CREATE DEFINER=`root`@`localhost` PROCEDURE
`proc_sesionvalor`(periodo CHAR(06), subperiodo CHAR(02),
usuario VARCHAR(20))
BEGIN
DECLARE t_ipusuario VARCHAR(15);
DECLARE t_fecha CHAR(11);
DECLARE t_hora TIME;
DECLARE t_urlpage VARCHAR(400);
DECLARE ipdiferente VARCHAR(15);
DECLARE t_fechaini CHAR(11);
DECLARE t_horaini TIME;
DECLARE t_fechafin CHAR(11);
DECLARE t_horafin TIME;
DECLARE no_more_sesionvalor INT;
DECLARE ipaux VARCHAR(15);
DECLARE cant INT;
DECLARE dimen INT;
DECLARE sesion INT;
DECLARE umbral_cant INT;
DECLARE fechora DATETIME;
DECLARE cur_sesionvalor CURSOR FOR
SELECT t1_ipusuario, t1_fecha, t1_hora, t1_urlpage
FROM mydb.t1archlogfiltrado
WHERE t1_periodo=periodo AND t1_subperiodo=subperiodo
ORDER BY 1;
DECLARE CONTINUE HANDLER FOR NOT FOUND SET
no_more_sesionvalor=1;
SET fechora = NOW();
SET no_more_sesionvalor=0;
SET ipdiferente="0.0.0.0";
SET umbral_cant=1;
SET dimen=(SELECT MAX(t2dimension.t2_nrodimension)
FROM t2dimension);
#    IF NOT EXISTS (SELECT * FROM t3ipsesion LIMIT 1)
THEN
 #   SET cant = 0;
  #  ELSE
    SET cant=(SELECT MAX(t3ipsesion.t3_nrosesion) FROM
t3ipsesion);
  # END IF;
```

```
#       IF NOT EXISTS (SELECT * FROM t4cantidadvisita
LIMIT 1) THEN
 #         SET sesion = 0;
 #       ELSE
          SET sesion=(SELECT
MAX(t4cantidadvisita.t4_nrovector) FROM t4cantidadvisita);
  #      END IF;
OPEN cur_sesionvalor;
sesionvalor_loop:WHILE(no_more_sesionvalor=0) DO
 FETCH cur_sesionvalor INTO t_ipusuario, t_fecha, t_hora,
t_urlpage;
 SET ipaux=t_ipusuario;
 IF no_more_sesionvalor=1 THEN
   LEAVE sesionvalor_loop;
 END IF;
 IF ipaux!=ipdiferente THEN
   SET cant=cant+1;
   IF umbral_cant>=3 THEN
        SET sesion = sesion+1;
   INSERT INTO t3ipsesion
   VALUES(null,cant,periodo,ipdiferente,t_fechaini,t_horaini,
   t_fechafin,t_horafin,umbral_cant,fechora,usuario);
   SET @cadena1=CONCAT('INSERT
   t4cantidadvisita(t4_periodo, t4_t3_nrosesion, t4_nrovector,
   t4_totalvisita, t4_t2_nrodimension, t4_cantidadvisitaxdim,
   t4_porcentaje, t4_fecactualiza, t4_usuario)
   SELECT ',periodo,', ',cant,', ',sesion,', ',umbral_cant,',
   t2_nrodimension, count(*) AS cantidad,
count(*)/',umbral_cant,' ,'',fechora,'', '',usuario,''
   FROM t1archlogfiltrado, t2dimension
   WHERE t1_periodo='',periodo,''
   AND t1_subperiodo='',subperiodo,'' AND
t1_ipusuario='',ipdiferente,''
   AND t1_urlpage=t2_valordimension GROUP BY
t1_urlpage');
   PREPARE stmt1 FROM @cadena1;
   EXECUTE stmt1;
     WHILE dimen > 0 DO
       IF NOT EXISTS (SELECT *
       FROM t4cantidadvisita
       WHERE t4_t3_nrosesion = cant
       AND t4_t2_nrodimension = dimen) THEN
       INSERT INTO t4cantidadvisita(t4_periodo,
```

```
        t4_t3_nrosesion, t4_nrovector, t4_totalvisita,
t4_t2_nrodimension,
        t4_cantidadvisitaxdim, t4_porcentaje, t4_fecactualiza,
t4_usuario)
        VALUES (periodo, cant, sesion, umbral_cant, dimen,
0, 0.0 ,fechora, usuario);
        END IF;
        SET dimen = dimen-1;
      END WHILE;
   SET ipdiferente=ipaux;
   SET t_fechaini=t_fecha;
   SET t_horaini=t_hora;
   SET t_fechafin=t_fecha;
   SET t_horafin=t_hora;
   SET umbral_cant=1;
   ELSE
   SET ipdiferente=ipaux;
   SET t_fechaini=t_fecha;
   SET t_horaini=t_hora;
   SET t_fechafin=t_fecha;
   SET t_horafin=t_hora;
   SET umbral_cant=1;
   END IF;
 ELSE
   SET t_fechafin=t_fecha;
   SET t_horafin=t_hora;
   SET umbral_cant=umbral_cant+1;
 END IF;
END WHILE sesionvalor_loop;
COMMIT;
CLOSE cur_sesionvalor;
SET no_more_sesionvalor=0;
END $$
DELIMITER ;
```

1.- Caso de Uso del Sistema	GENERACIÓN DE LOS *CLUSTER*

2.- Descripción del caso de uso

El sistema permitirá generar las Cluster en función a los siguientes parámetros.

Para el periodo que se calculó la muestra.
Para la cantidad de *Cluster* que se desea obtener.

La generación de *Cluster* está sujeto al análisis del interesado, ya que puede seguir generando hasta obtener un número óptimo de *Cluster* y una vez generada la cantidad de *Cluster* óptimo el interesado analizará los patrones que se pueden encontrar, recomendaciones y toma de decisión en la mejora del portal.

3.- Actor(es)

Gerencia y/o responsables de los proyectos en la mejora del Portal Web.

4.- Precondiciones

Para el periodo a analizar, deben estar actualizados:

Tabla de las Dimensiones actualizada(t2dimension).
Tabla de Sesiones actualizad(t3ipsesion).
Tabla de cantidad de visitas actualizada(t4cantidadvisita).
Tabla de Muestra de Sesiones(t5muestrasesion).
Tabla de Vector posición actualizada(t6vectorposicion).

5.- Postcondiciones

Actualizar la tabla de Vector Posición t6vectorposicion (el campo de Cluster al que pertenece el vector e insertar los centroides de cada *Cluster*) e inserta a la tabla t7Cluster

Generación de los archivos:

Clustering.txt.
DatosCentroide.txt.
DatosClustering.txt.

6.- Pasos (Flujo de Eventos)

N°	Acción del Actor	Respuesta del Sistema
1	Ingresa al módulo de generación de *Clusters.*	Muestra la pantalla de Generación de *Clusters.*

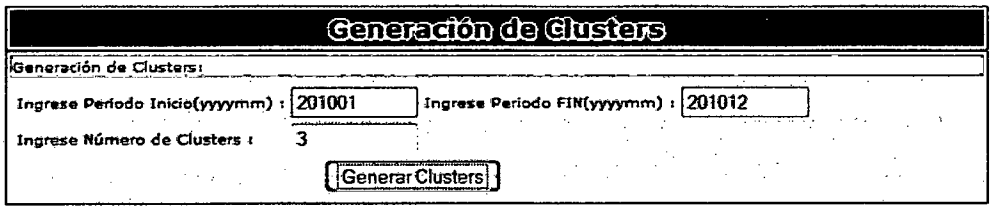
2	Ingresa los parámetros y clic en generar *Clusters*	El sistema procesa y genera los *Clusters*, actualizando la tabla t6vectorposicion e insertando la tabla *t7Cluster* y al final le muestra el mensaje del estado de la ejecución • Mensaje de Error • Mensaje de Carga Exitosa
3	Confirma el Mensaje, si: • La carga fue exitosa Cierra • La carga fue con error, reporta el error y Cierra.	Se cierra la ventana

7.- Requerimiento asociado

RF-006

9.- Prototipo de interfaz de usuario

IU-005



Inicio

8.- Casos de Uso Alternos

Carga: t6vectorposicion y t5muestrasesion

```
DELIMITER $$
                DROP PROCEDURE IF EXISTS
                `mydb`.`proc_vectorposicion` $$
                CREATE DEFINER=`root`@`localhost` PROCEDURE
                `proc_vectorposicion`(periodo CHAR(06), usuario
                VARCHAR(20))
                BEGIN
                DECLARE cant INT;
                DECLARE no_more_muestra INT;
                DECLARE t_periodo CHAR(06);
                DECLARE t_nrovector INT;
```

```
#SET fechora = NOW();
DECLARE cur_muestra CURSOR FOR
SELECT t5_periodo, t5_t4_nrovector FROM
mydb.t5muestrasesion ORDER BY 2;
DECLARE CONTINUE HANDLER FOR NOT FOUND SET
no_more_muestra=1;
SET no_more_muestra=0;
   IF NOT EXISTS (SELECT * FROM t6vectorposicion
LIMIT 1) THEN
   SET cant = 0;
   ELSE
   SET cant=(SELECT MAX(t6vectorposicion.t6_nrovector)
FROM t6vectorposicion);
   END IF;
   SET @cadena1=CONCAT
   ('INSERT t5muestrasesion(t5_t4_nrovector, t5_periodo,
t5_fecactualiza,
   t5_usuario) SELECT DISTINCT(b.t4_nrovector),
"201013",  NOW(), '',usuario,''
   FROM t4cantidadvisita b ORDER BY RAND() LIMIT
2952');
   PREPARE stmt1 FROM @cadena1;
   EXECUTE stmt1;
SET cant = 0;
OPEN cur_muestra;
muestravalor_loop:WHILE(no_more_muestra=0) DO
FETCH cur_muestra INTO t_periodo, t_nrovector;
SET cant=cant+1;
   SET @cadena1=CONCAT('INSERT
t6vectorposicion(t6_periodo, t6_nrovector, t6_tipo,
    t6_t7_nroCluster, t6_nrodimension, t6_porcentaje,
t6_fecactualiza, t6_usuario)
    SELECT t4_periodo, ',cant,', 0 , 0 , t4_t2_nrodimension,
t4_porcentaje, NOW(), '',usuario,''
    FROM t4cantidadvisita WHERE t4_nrovector =
',t_nrovector,' ORDER BY 2, 5');
   PREPARE stmt1 FROM @cadena1;
   EXECUTE stmt1;
END WHILE muestravalor_loop;
COMMIT;
CLOSE cur_muestra;
SET no_more_muestra=0;
END $$
DELIMITER ;
```

CARGA: T9IPCLUSTER

INSERT INTO t9ipCluster(t9_idipCluster, t9_nroip, t9_Cluster, t9_periodo, t9_fecactualiza, t9_usuario)
SELECT NULL, t3_ipusuario, t6_t7_nroCluster, "201013", NOW(), "AEMD"

FROM t4cantidadvisita, t3ipsesion, t6vectorposicion WHERE t3_nrosesion =

t4_t3_nrosesion AND t4_nrovector = t6_nrovector GROUP BY 2

1.- Caso de Uso del Sistema	REPORTES Y CONSULTAS
2.- Descripción del caso de uso	
El sistema muestra el menú de reportes: • Dimensiones • Sesiones • Vector Posición. • Clusters • Diferencia de Centroides • IPsx *Cluster* • *Cluster*_1 (Centroide) • *Cluster*_2 (Centroide) • *Cluster*_3 (Centroide) En cada uno de ellos visualizamos en paginado.	
3.- Actor(es)	
Gerencia	
4.- Precondiciones	
Haber generado las sesiones, porcentaje de visitas y los *Clusters*	
5.- Postcondiciones	
• Analizar los reportes para la toma de decisiones.	
6.- Pasos (Flujo de Eventos)	

N°	**Acción del Actor**	**Respuesta del Sistema**
	Ingresa al modulo de reporte y consulta.	Muestra la pantalla de reporte y consultas
1	Clic en el reporte que desea visualizar.	El sistema consulta la BD y le muestra el reporte o de lo contrario muestra el mensaje de error.

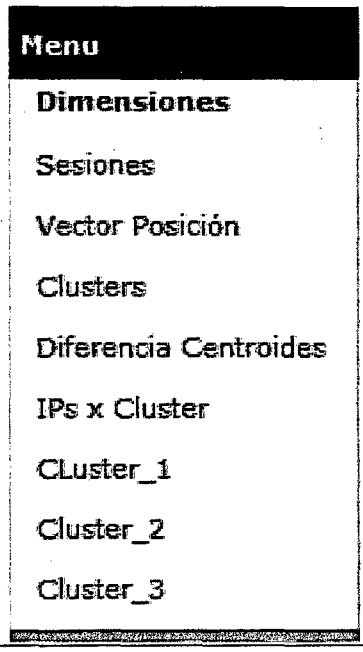
2	Si el reporte es exitoso • Visualiza el reporte, analiza y cierra. Si muestra el mensaje de error. • Reporta el error y Cierra.	Se cierra la ventana.

7.- Requerimiento asociado

RF-00

8.- Prototipo de interfaz de usuario

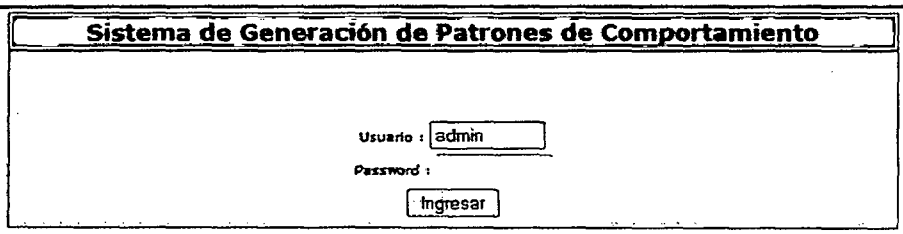
IU-005



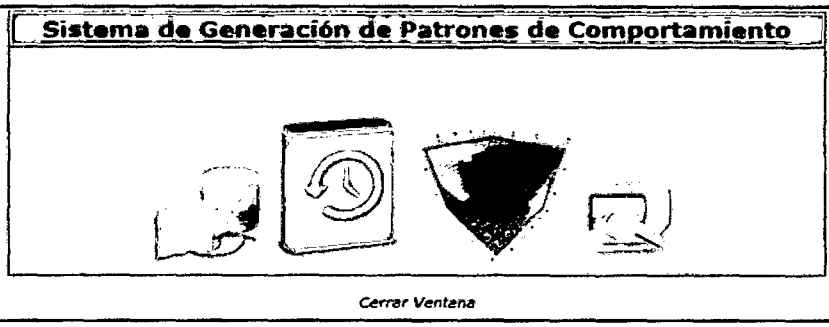
9.- Casos de Uso Alternos

2.1.5 ESPECIFICACIÓN DE LA INTERFAZ DE USUARIO

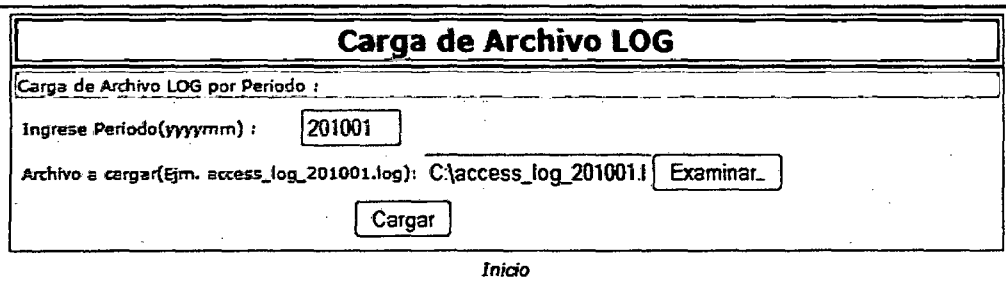
Número	IU-001
Propósito de la interfaz	Autentificación
Gráfica de la interfaz	



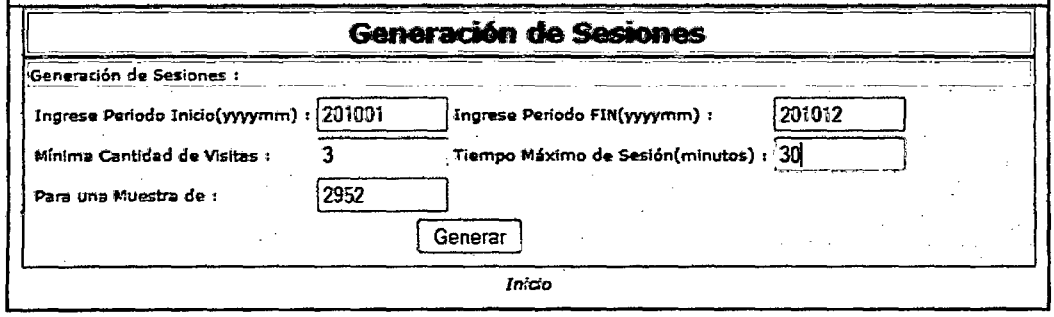
Número	IU-001
Propósito de la interfaz	Menu (Carga de Archivo *LOG*, Generación de Sesiones, Generación de *Cluster* y Reportes)
Gráfica de la interfaz	



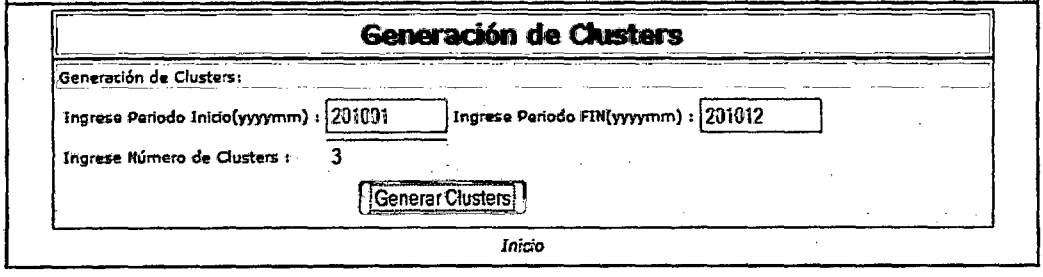
Número	IU-002
Propósito de la interfaz	Carga de Archivo *LOG*
Gráfica de la interfaz	



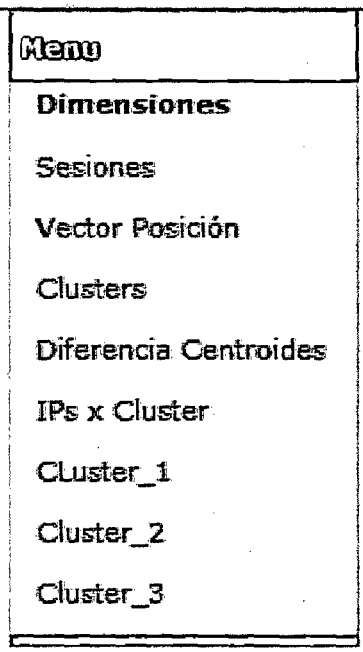
Número	IU-003
Propósito de la interfaz	Generación de Sesiones
Gráfica de la interfaz	



Número	IU-004
Propósito de la interfaz	Generación de *Clusters*
Gráfica de la interfaz	



Número	IU-005
Propósito de la interfaz	Menú de Reportes
Gráfica de la interfaz	



Dimensiones

Dimensiones

Nro. Dimension	Periodo	URL Dimension	Descripcion
1	201006	"http://www.sunat.gob.pe/cl-ti-itmrconsruc/jcrS00Alias"	
2	201006	"http://www.sunat.gob.pe/cl-at-itageban/bcosucS01Alias"	
3	201006	"http://www.sunat.gob.pe/ol-ti-itinsrucsol/iruc001Alias"	
4	201006	"http://www.sunat.gob.pe/ol-ti-itsusprenta/srS01Alias"	
5	201006	"http://www.sunat.gob.pe/cl-ti-itcronobligme/fvS01Alias"	
6	201006	"http://www.sunat.gob.pe/cl-at-itconcompag/ccS02Alias"	
7	201006	"http://www.sunat.gob.pe/cl-ti-itpresqueja/sqsS21Alias"	
8	201006	"http://www.sunat.gob.pe/ol-ti-itdenuncia/denS01Alias"	
9	201006	"http://www.sunat.gob.pe/cl-ti-itpresqueja/sqsS31Alias"	
10	201006	"http://www.sunat.gob.pe/cl-ti-itconsrenta/srS01Alias"	
11	201006	"http://www.sunat.gob.pe/cl-at-itrecone/roS01Alias"	
12	201006	"http://www.sunat.gob.pe/ol-ti-itfichaseleccion /fichaS01Alias"	
13	201006	"http://www.sunat.gob.pe/cl-at-itentdeu/iisS01Alias"	
14	201006	"http://www.sunat.gob.pe/cl-ti-itconspredios/sfpS02Alias"	
15	201006	"http://www.sunat.gob.pe//cl-ti-itmrconsruc/jcrS00Alias"	
16	201006	"http://www.sunat.gob.pe/cl-ti-itmrconsruc//jcrS00Alias"	
17	201006	"http://www.sunat.gob.pe/ol-ti-itpresf5030/rsdS01Alias"	
18	201006	"http://www.sunat.gob.pe/cl-ti-itconsdenuncia/denS02Alias"	
19	201006	"http://www.sunat.gob.pe/ol-ti-itpdtpred/sfp01Alias"	
20	201007	"http://www.sunat.gob.pe/cl-ad-ittabladescripcionconsulta /TablaDescripcionS01Alias"	

Records 1 - 20 of 25. First | Previous | 1 | 2 | Next | Last

Sesiones

Sesiones

Nro. Sesion	Nro. Vector	Periodo
1	293,052	201013
2	302,861	201013
3	54,702	201013
4	607,762	201013
5	457,476	201013
6	633	201013
7	31,517	201013
8	712,775	201013
9	810,573	201013
10	230,967	201013
11	368,980	201013
12	733,747	201013
13	769,620	201013
14	437,245	201013
15	682,312	201013
16	359,553	201013
17	295,898	201013
18	724,156	201013
19	548,199	201013
20	859,395	201013

Records 1 - 20 of 2952 First | Previous | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Next | Last

Vector Posición

Vectorposicion

Nro. Vectorposicion	Periodo	Nrovector	Tipo	Nro Cluster	Nro. Dimension	Porcentaje
1	201006	1	0	3	1	1.00
2	201006	1	0	3	2	0.00
3	201006	1	0	3	3	0.00
4	201006	1	0	3	4	0.00
5	201006	1	0	3	5	0.00
6	201006	1	0	3	6	0.00
7	201006	1	0	3	7	0.00
8	201006	1	0	3	8	0.00
9	201006	1	0	3	9	0.00
10	201006	1	0	3	10	0.00
11	201006	1	0	3	11	0.00
12	201006	1	0	3	12	0.00
13	201006	1	0	3	13	0.00
14	201006	1	0	3	14	0.00
15	201006	1	0	3	15	0.00
16	201006	1	0	3	16	0.00
17	201006	1	0	3	17	0.00
18	201006	1	0	3	18	0.00
19	201006	1	0	3	19	0.00
20	201006	1	0	3	20	0.00

Records 1 - 20 of 73900 First | Previous | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Next | Last

Clusters

Clusters

Nro. Cluster	Nro. Corrida	Nombre	Descripcion	Cantidad x Cluster	Porcentaje
1	0			203	0.0688
2	0			2,014	0.6827
3	0			733	0.2485

Records 1 - 3 of 3 First | Previous | 1 | Next | Last

IPs x *Cluster*

IP Cluster

Nro.	Nro. IP	Cluster	Periodo
1	12.170.7.2	3	201013
2	12.45.100.26	3	201013
3	12.48.16.2	2	201013
4	13.8.137.11	2	201013
5	136.166.1.3	2	201013
6	138.103.17.18	1	201013
7	140.212.213.40	2	201013
8	144.191.148.3	2	201013
9	148.168.127.10	3	201013
10	148.175.49.1	2	201013
11	148.243.149.12	2	201013
12	148.243.149.14	2	201013
13	149.128.8.245	3	201013
14	15.227.249.73	2	201013
15	15.227.249.74	2	201013
16	15.227.249.75	2	201013
17	153.2.247.34	2	201013
18	155.70.222.29	1	201013
19	161.132.144.102	2	201013
20	161.132.168.200	3	201013

Records 1 - 20 of 2956 First | Previous | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Next | Last

Cluster_1

Cluster 1

Nro Dimension	URL de la Página	Porcentaje
1	"http://www.sunat.gob.pe/cl-ti-itmrconsruc/jcrS00Alias"	0.8959
4	"http://www.sunat.gob.pe/ol-ti-itsusprenta/srS01Alias"	0.0470
5	"http://www.sunat.gob.pe/cl-ti-itcronobligme/fvS01Alias"	0.0268
3	"http://www.sunat.gob.pe/ol-ti-itinsrucsol/iruc001Alias"	0.0181
10	"http://www.sunat.gob.pe/cl-ti-itconsrenta/srS01Alias"	0.0065
7	"http://www.sunat.gob.pe/cl-ti-itpresqueja/sqsS21Alias"	0.0033
9	"http://www.sunat.gob.pe/cl-ti-itpresqueja/sqsS31Alias"	0.0016
6	"http://www.sunat.gob.pe/cl-at-itconcompag/ccS02Alias"	0.0008
23	"http://www.sunat.gob.pe/ol-ti-itreciboelectronicovalarch/ceS01Alias"	0.0000
24	"http://www.sunat.gob.pe/cl-ti-itmrconsruc/jcrS03Alias"	0.0000
22	"http://www.sunat.gob.pe/ol-ti-itreciboelectronico/cpelec003Alias"	0.0000
21	"http://www.sunat.gob.pe/wtc/wrapTcS01Alias"	0.0000
20	"http://www.sunat.gob.pe/cl-ad-ittabladescripcionconsulta /TablaDescripcionS01Alias"	0.0000
19	"http://www.sunat.gob.pe/ol-ti-itpdtpred/sfp01Alias"	0.0000
18	"http://www.sunat.gob.pe/cl-ti-itconsdenuncia/denS02Alias"	0.0000
17	"http://www.sunat.gob.pe/ol-ti-itpresf5030/rsdS01Alias"	0.0000
16	"http://www.sunat.gob.pe/cl-ti-itmrconsruc//jcrS00Alias"	0.0000
15	"http://www.sunat.gob.pe//cl-ti-itmrconsruc/jcrS00Alias"	0.0000
14	"http://www.sunat.gob.pe/cl-ti-itconspredios/sfpS02Alias"	0.0000
13	"http://www.sunat.gob.pe/cl-at-itentdeu/iisS01Alias"	0.0000

Records 1 - 20 of 25 First | Previous | 1 | 2 | Next | Last

Cluster_2

Cluster2

Nro Dimension	URL de la Página	Porcentaje
1	"http://www.sunat.gob.pe/cl-ti-itmrconsruc/jcrS00Alias"	0.9901
5	"http://www.sunat.gob.pe/cl-ti-itcronobligme/fvS01Alias"	0.0035
4	"http://www.sunat.gob.pe/ol-ti-itsusprenta/srS01Alias"	0.0034
3	"http://www.sunat.gob.pe/ol-ti-itinsrucsol/iruc001Alias"	0.0013
10	"http://www.sunat.gob.pe/cl-ti-itconsrenta/srS01Alias"	0.0006
6	"http://www.sunat.gob.pe/cl-at-itconcompag/ccS02Alias"	0.0004
2	"http://www.sunat.gob.pe/cl-at-itageban/bcosucS01Alias"	0.0002
12	"http://www.sunat.gob.pe/ol-ti-itfichaseleccion/fichaS01Alias"	0.0001
13	"http://www.sunat.gob.pe/cl-at-itentdeu/iisS01Alias"	0.0001
11	"http://www.sunat.gob.pe/cl-at-itrecone/roS01Alias"	0.0001
7	"http://www.sunat.gob.pe/cl-ti-itpresqueja/sqsS21Alias"	0.0000
14	"http://www.sunat.gob.pe/cl-ti-itconspredios/sfpS02Alias"	0.0000
9	"http://www.sunat.gob.pe/cl-ti-itpresqueja/sqsS31Alias"	0.0000
24	"http://www.sunat.gob.pe/cl-ti-itmrconsruc/jcrS03Alias"	0.0000
23	"http://www.sunat.gob.pe/ol-ti-itreciboelectronicovalarch/ceS01Alias"	0.0000
22	"http://www.sunat.gob.pe/ol-ti-itreciboelectronico/cpelec003Alias"	0.0000
21	"http://www.sunat.gob.pe/wtc/wapTcS01Alias"	0.0000
20	"http://www.sunat.gob.pe/cl-ad-ittabladescripcionconsulta /TablaDescripcionS01Alias"	0.0000
19	"http://www.sunat.gob.pe/ol-ti-itpdtpred/sfp01Alias"	0.0000
18	"http://www.sunat.gob.pe/cl-ti-itconsdenuncia/denS02Alias"	0.0000

Records 1 - 20 of 25 First | Previous | 1 | 2 | Next | Last

Cluster_3

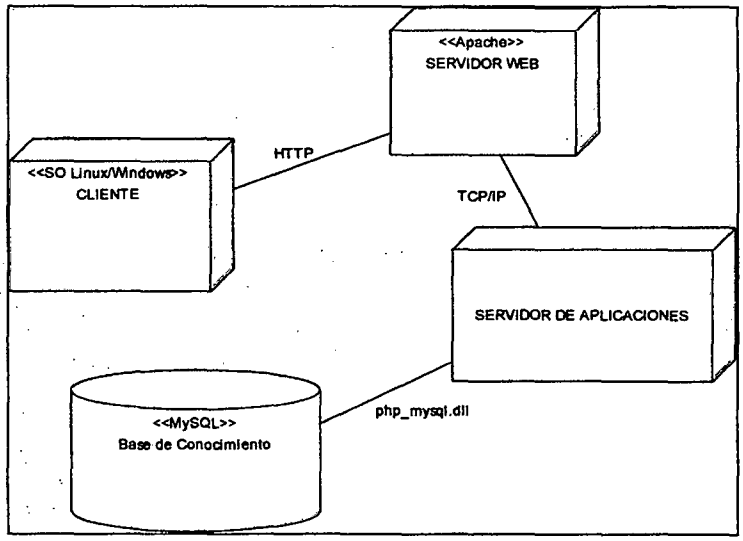
Cluster3		
Nro Dimension	**URL de la Página**	**Porcentaje**
1	"http://www.sunat.gob.pe/cl-ti-itmrconsruc/jcrS00Alias"	0.9389
4	"http://www.sunat.gob.pe/ol-ti-itsusprenta/srS01Alias"	0.0208
5	"http://www.sunat.gob.pe/cl-ti-itcronobligme/fvS01Alias"	0.0134
3	"http://www.sunat.gob.pe/ol-ti-itinsrucsol/iruc001Alias"	0.0084
10	"http://www.sunat.gob.pe/cl-ti-itconsrenta/srS01Alias"	0.0066
19	"http://www.sunat.gob.pe/ol-ti-itpdtpred/sfp01Alias"	0.0040
6	"http://www.sunat.gob.pe/cl-at-itconcompag/ccS02Alias"	0.0040
2	"http://www.sunat.gob.pe/cl-at-itageban/bcosucS01Alias"	0.0033
14	"http://www.sunat.gob.pe/cl-ti-itconspredios/sfpS02Alias"	0.0006
18	"http://www.sunat.gob.pe/cl-ti-itconsdenuncia/denS02Alias"	0.0000
20	"http://www.sunat.gob.pe/cl-ad-ittabladescripcionconsulta /TablaDescripcionS01Alias"	0.0000
21	"http://www.sunat.gob.pe/wtc/wapTcS01Alias"	0.0000
22	"http://www.sunat.gob.pe/ol-ti-itreciboelectronico/cpelec003Alias"	0.0000
23	"http://www.sunat.gob.pe/ol-ti-itreciboelectronicovalarch/vceS01Alias"	0.0000
24	"http://www.sunat.gob.pe/cl-ti-itmrconsruc/jcrS03Alias"	0.0000
17	"http://www.sunat.gob.pe/ol-ti-itpresf5030/rsdS01Alias"	0.0000
16	"http://www.sunat.gob.pe/cl-ti-itmrconsruc//jcrS00Alias"	0.0000
15	"http://www.sunat.gob.pe//cl-ti-itmrconsruc/jcrS00Alias"	0.0000
13	"http://www.sunat.gob.pe/cl-at-itentdeu/iisS01Alias"	0.0000
12	"http://www.sunat.gob.pe/ol-ti-itfichaseleccion/fichaS01Alias"	0.0000

Records 1 - 20 of 25 First | Previous | 1 | 2 | Next | Last

3. DISEÑO DE SISTEMAS

3.1 DEFINICIÓN DE ARQUITECTURA DEL SISTEMA

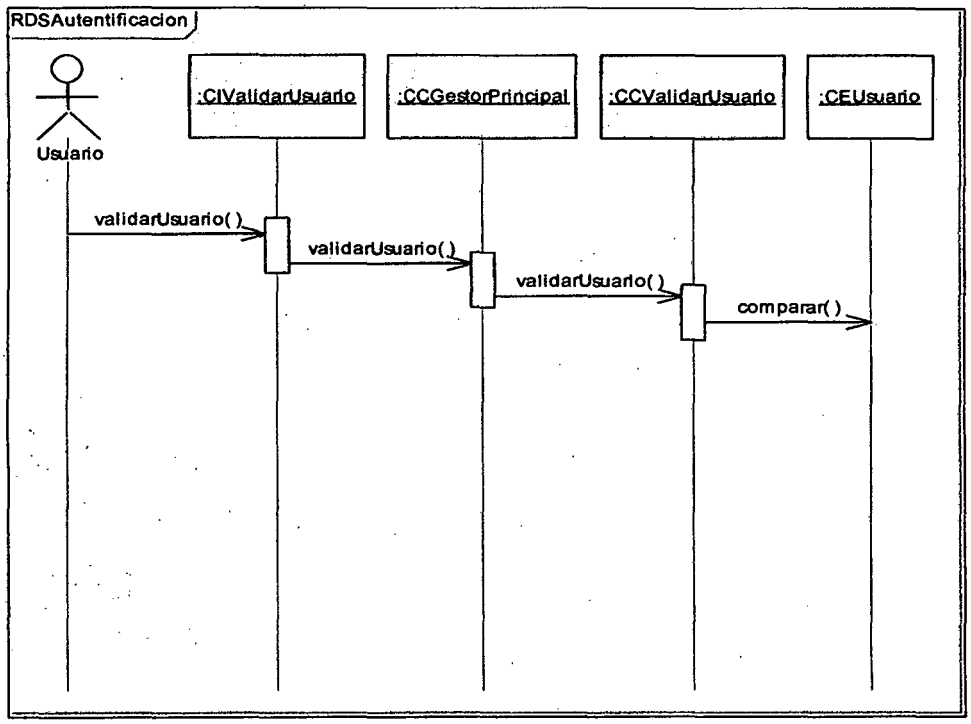
3.1.1 DEFINICIÓN DE NIVELES DE ARQUITECTURA



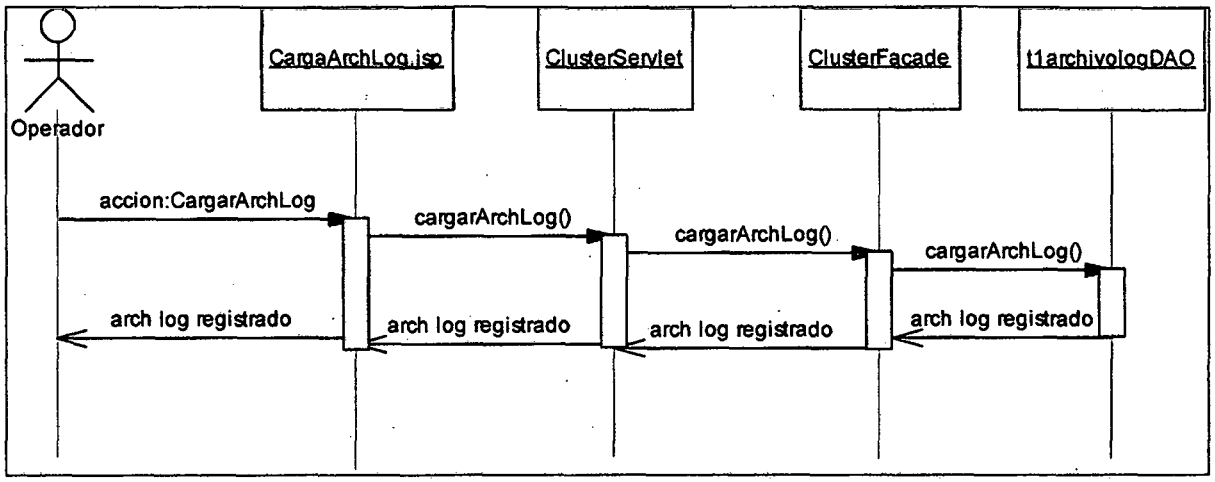
1.1.2 ENTORNO TECNOLÓGICO DEL SISTEMA (DSI 1.1)

Hardware	2 PCs Core 2 Duo u otro superior para servidor web y Servidor de Aplicaciones. 1 PC Dual Core u otro superior para la Administración del Sistema.
Software	**Sistemas operativos:** Linux / Windows. **Gestor de base de datos:** MySQL **Servicios:** Servidor Apache, PHP
Comunicaciones	**Servicios:** Conexión a internet. **Protocolo:** • HTTP (Cliente – Servidor Web) • TCP/IP (Servidor Web – Servidor de Aplicaciones) • (Servidor de Aplicaciones – Gestor de Base de Datos)

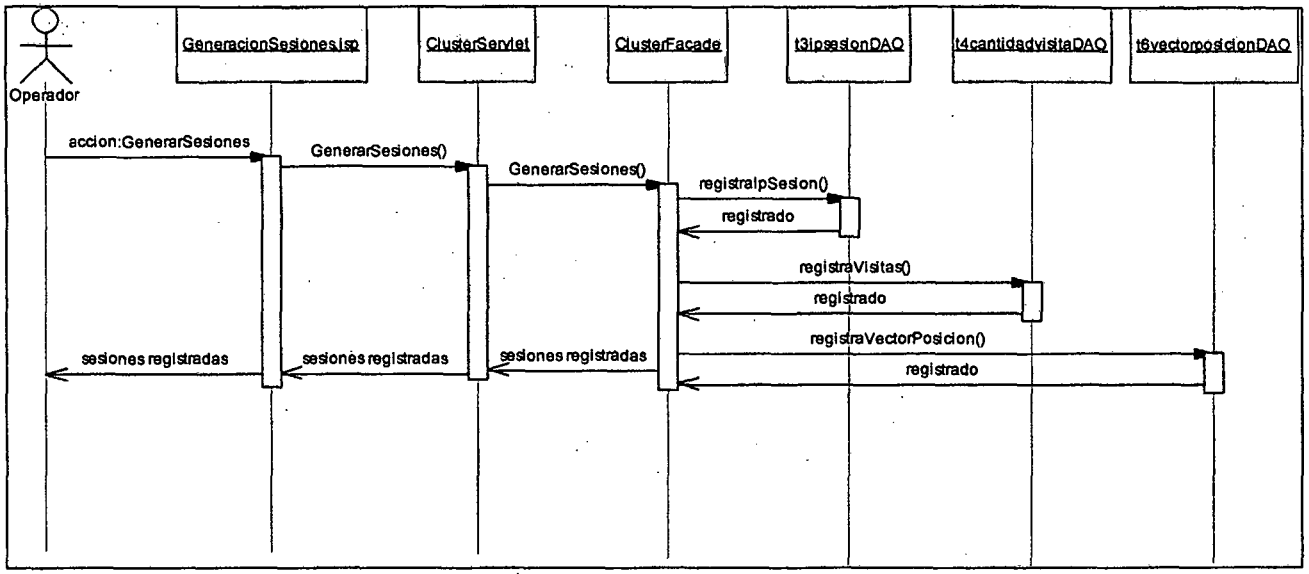
3.2 DISEÑO DE CASOS DE USO (DSI 2)



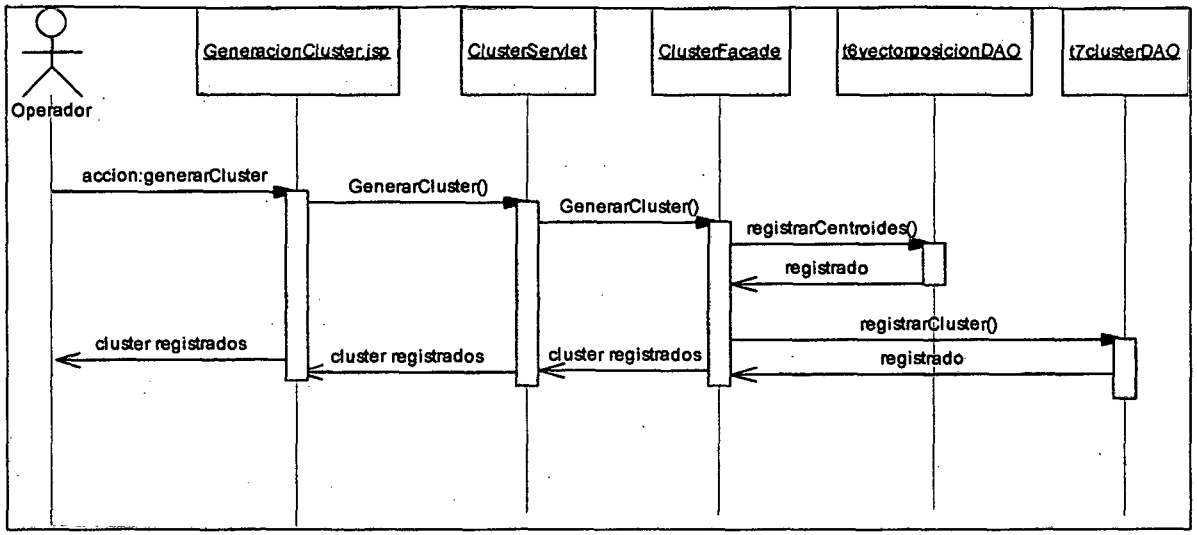
DS Autentificación



DS Carga Archivo *Log*

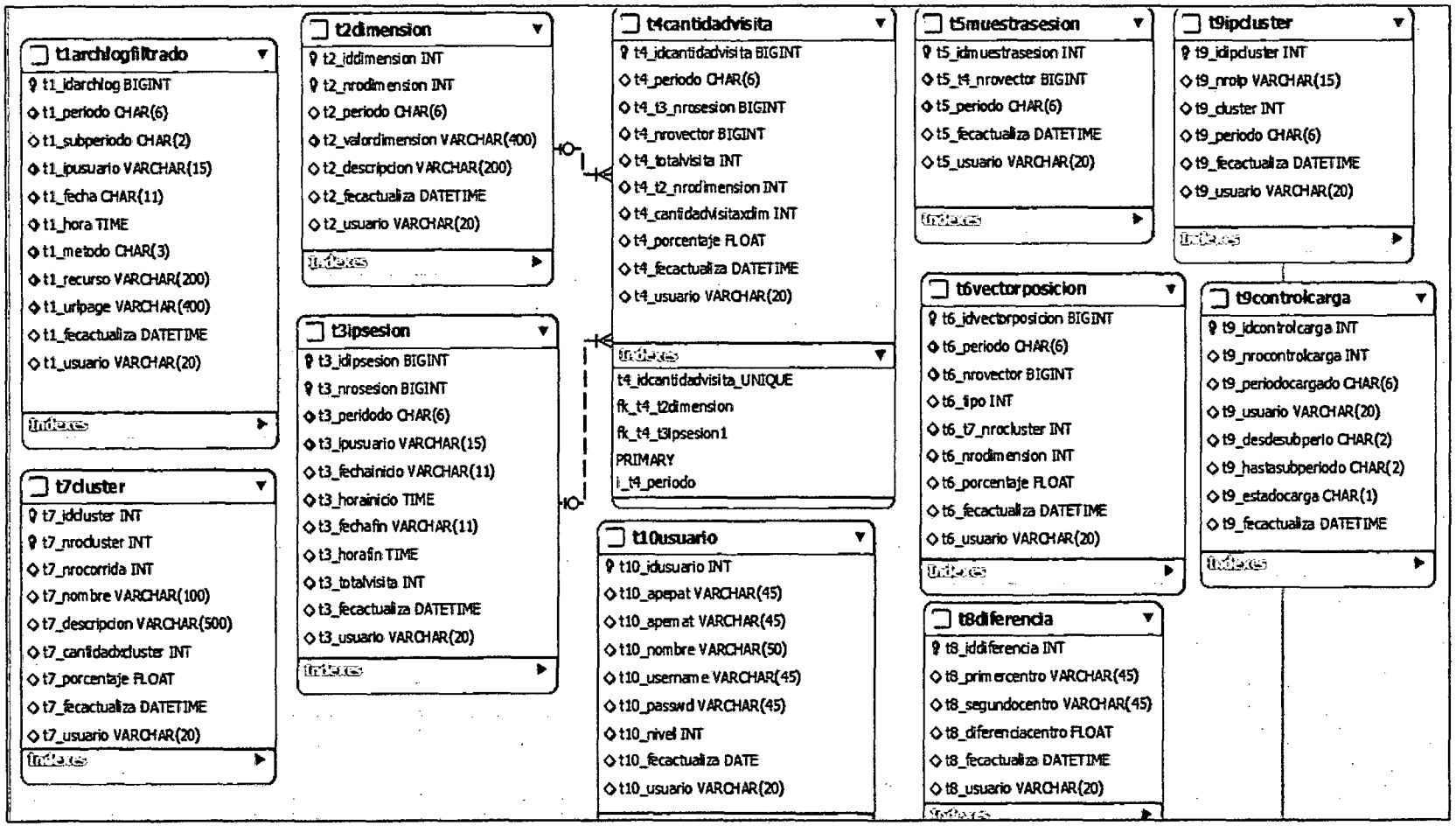


DS Generación de Sesiones

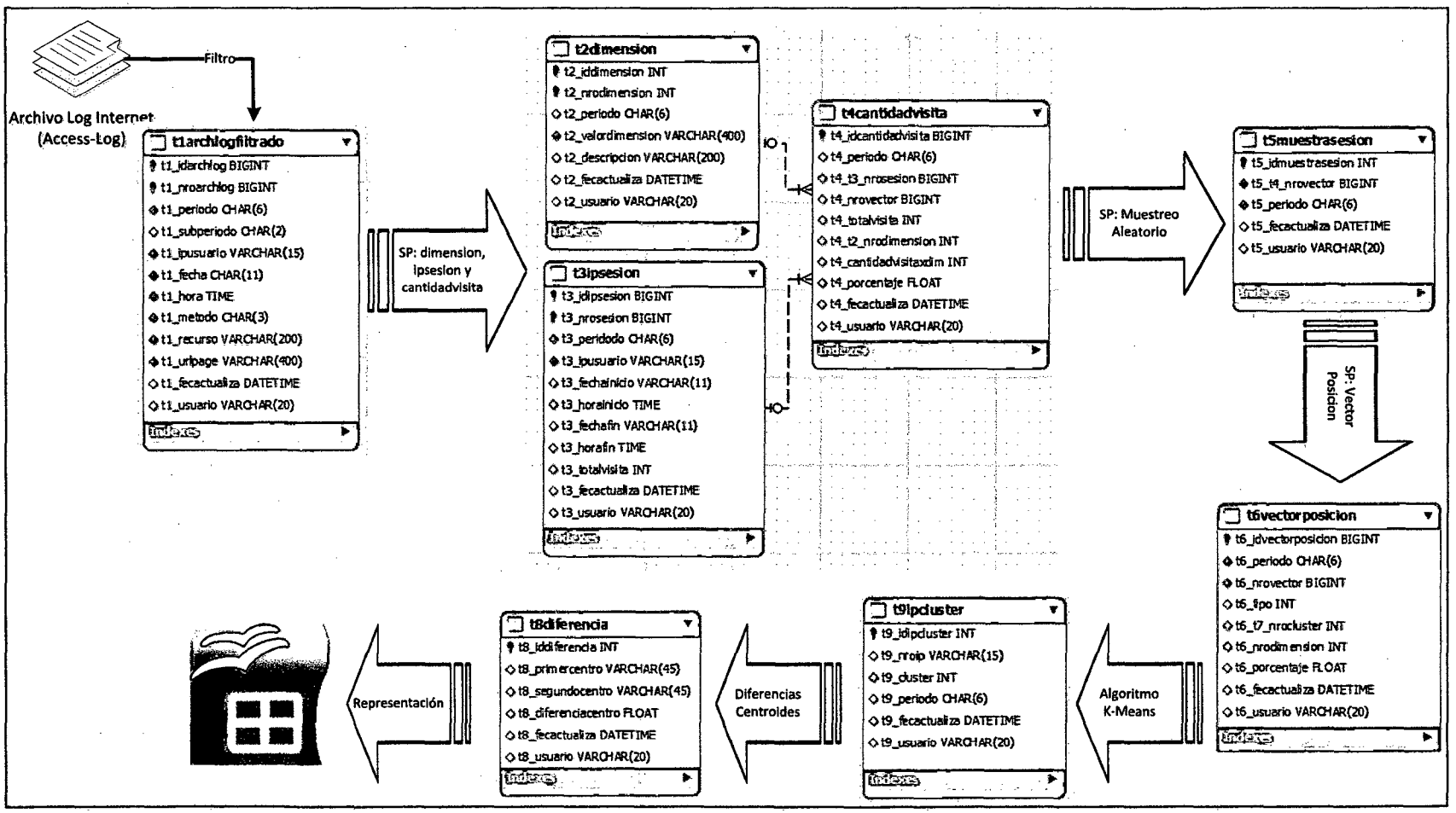


DS Generación de *Cluster*

4. DISEÑO FÍSICO DE DATOS

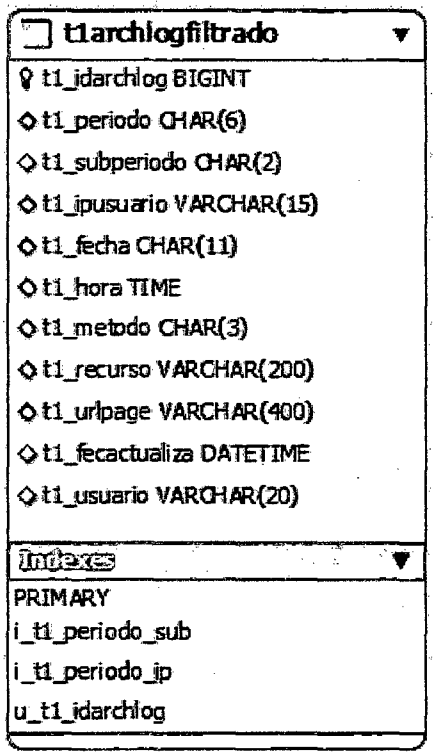


4. FLUJO DE LOS STORE PROCEDURES



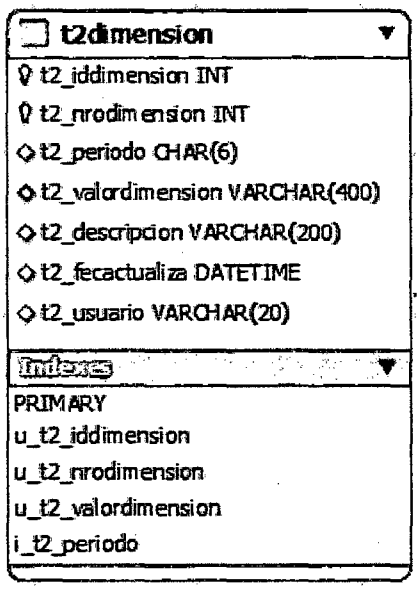
1.- ENTIDAD: Archlogfiltrado

- **Propósito:** Almacenamiento, de los registros logs filtrados y parseados.
- **Atributos :**



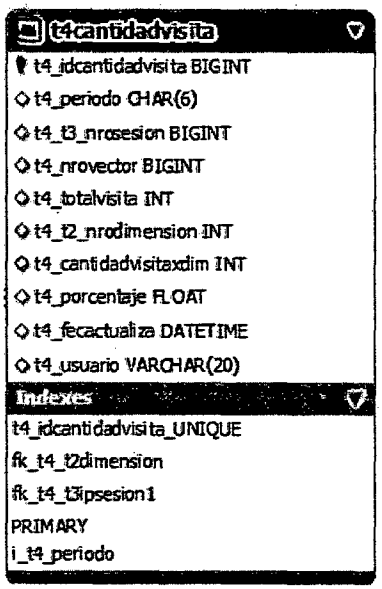
2.- ENTIDAD: Dimensión

- **Propósito:** Almacenamiento, de las URL o módulos del Portal Web
- **Atributos :**



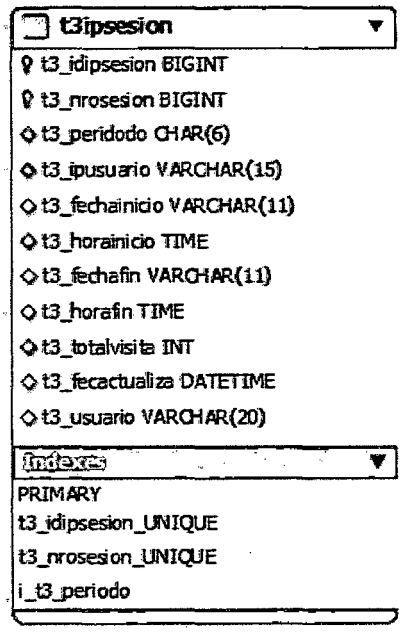
3.- ENTIDAD: Cantidadvisita

- **Propósito:** Almacenamiento, guarda la cantidad de URL visitadas dentro de cada sesión.
- **Atributos :**



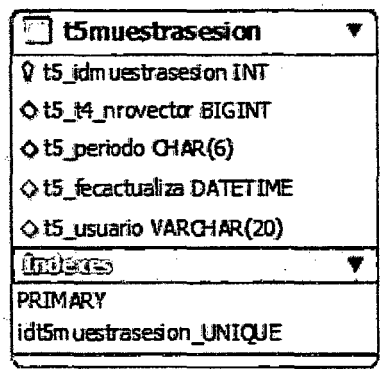
4.- ENTIDAD: IPSesion

- **Propósito:** Almacenamiento, guarda las sesiones de los usuarios, con los umbrales considerados.
- **Atributos :**



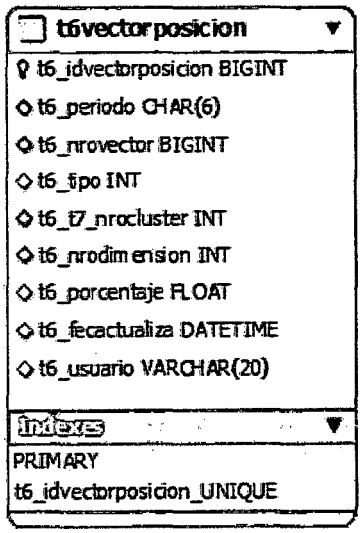
5.- ENTIDAD: Muestrasesion

- **Propósito:** Almacenamiento, solamente guarda una cantidad representativa de las sesiones (muestra).
- **Atributos :**



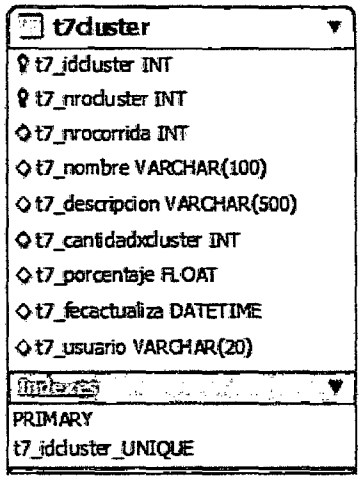
6.- ENTIDAD: Vectorposicion

- **Propósito:** Almacenamiento, guarda los vectores posición que serán utilizadas por el algoritmo k-means y luego actualizados con el nro. de Cluster al que pertenece.
- **Atributos :**



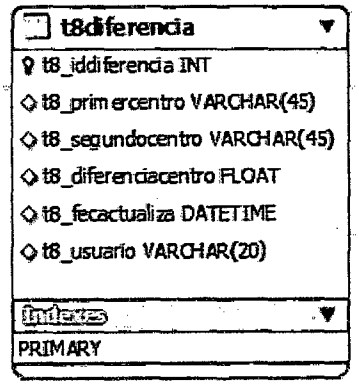
7.- ENTIDAD: *Cluster*

- **Propósito:** Almacenamiento, guarda los *Clusters* y su descripción.
- **Atributos :**



8.- ENTIDAD: Diferencia

- **Propósito:** Almacenamiento, guarda las diferencias entre centroides de los *Clusters*.
- **Atributos :**



9.- ENTIDAD: *IPCluster*

- **Propósito:** Almacenamiento, guarda las IPs con el *Cluster* al que pertenece cada una de las IP.
- **Atributos :**