Análisis Longitudinal de Accesos a un Servidor WWW

por

Ing. David De La Rosa Hilario (IPN)

Sometida a revisión al Departamento de Ciencias de la Computación
en cumplimiento de los requisitos para el grado de

Maestro en Ciencias de la Computación y Matemáticas Industriales

en el

CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS A.C.

Enero 2006

Autor .....
Departamento de Ciencias de la Computación
27 Enero de 2006

Aceptada por .....
Johan Jozef Lode Van Horebeek
Asesor de Tesis
Investigador del CIMAT

Aceptada por .....
Rogelio Hasimoto Beltrán
Revisor de Tesis
Investigador del CIMAT

Aceptada por .....
Ramón Reyes Carrión
Revisor de Tesis
Investigador del CIMAT

Aprovada por .....
Mariano J. J. Rivera Meraz
Coordinador de la Maestria en el Departamento de Ciencias de la Computación

Análisis Longitudinal de Accesos a un Servidor WWW

por

Ing. David De La Rosa Hilario (IPN)

Sometida a revisión al Departamento de Ciencias de la Computación
el 27 Enero de 2006, en cumplimiento de los
requisitos para el grado de
Maestro en Ciencias de la Computación y Matemáticas Industriales

Resumen

El presente trabajo es un estudio sobre como los usuarios accesan el sitio WWW del CIMAT. Enfatizando el comportamiento de los datos en función del tiempo además de modelar los datos con diversas técnicas.

Describimos de una manera general las características más importantes de los datos a través de técnicas de proyecciones como PCA e ICA y relaciones entre atributos de las sesiones.

Al ser nuestro énfasis en la parte temporal del comportamiento de los navegantes, buscamos formas de modelar las cadenas mismas. Implementamos modelos Markovianos los cuales son considerados el estado del arte.

Encontramos que los modelos Markovianos presentan la desventaja no ser intuitivos, de dificl estimación y carecen de una interpretación evidente de sus resultados.

A través del último capítulo ofrecimos una alternativa a éstas desventajas proponiendo medidas en las cuales no es necesaria la asumpción de Markovianidad.

Con éstos modelos queremos proporcionar un mapa más intuitivo del comportamiento de los navegantes, en el cual relacionamos a los grupos temáticos de acuerdo al comportamiento que tienen los usuarios sobre ellos.

Supervisor de Tesis: Johan Jozef Lode Van Horebeek
Title: Asesor de Tesis

Supervisor de Tesis: Rogelio Hasimoto Beltrán
Title: Revisor de Tesis

Supervisor de Tesis: Ramón Reyes Carrión
Title: Revisor de Tesis

No os conforméis a este siglo, sino transformaos
por medio de la renovación de vuestro entendimiento,
para que comprobéis cuál sea la buena voluntad de Dios,
agradable y perfecta.

–Romanos 12:2

Mirad que nadie os engañe por medio de filosofas
y huecas sutilezas, según las tradiciones de los hombres,
conforme a los rudimentos del mundo, y no según Cristo.

–Colosenses 2:8

Agradecimientos

Son pocas las palabras que al momento de escribir éstas lineas bienen a mi mente para agradecer a todos aquellos involucrados directamente en la realización de ésta tesis, así como a quienes en un acto de verdadero heroísmo pensaron que pudiera ser terminada a tiempo.

La verdad es que meses después de la presión y titánica labor (para mi) de terminar éste trabajo veo con ojos diferentes todo su proceso. Como cada pieza fue encajando en su lugar para poder formar esa maquinaria de ideas y trabajo para cristalizar y plasmar en letras lo que en algún momento paso por la mente de mi asesor Johan y mia.

Cada pieza encajó a la perfección como si fuera diseñada mucho tiempo antes para que en el preciso momento se dieran las cosas. Primeramente por la tolerancia y paciencia que el Dr. Horebeek tuvo a lo largo de nuestras seciones, las cuales agradezco de infinita manera, porque no solamente me dirigió en este trabajo sino también me enseño que como todo trabajo humanamente perfectible a nunca estar conforme con lo obtenido sino que siempre se puede mejorar continuamente.

A mis amigos Ivvan y Luis por sus interminables discuciones, desatadas bajo cualquier pretexto, para poder exponer tan diferentes puntos de vista como nosotros mismos. Gracias por su amistad.

A la beca alimenticia Hernández-Zavala la cual me mantuvo a flote varios meses con la cantidad de risas y nutrientes que uno requiere para levantar el espiritu en cualquier momento.

A mis papas por creer en mi y siempre estar ahí para mi cuando lo necesitase, dandome consejo y una cálida bienvenida cada vez que visitaba mi casa.

A la Srta. Ma. Isabel Hernández por no solamente darme su amistad sino por permitirme ser su compañero de travesuras y demás circo maroma y teatro que hicimos juntos, gracias por todo.

Y no por último menos importante al CIMAT y a su comunidad que me permitieron a mi, hasta ahora un turista de la ciencia, incursionar en éste ambiente de la ciencias rigurosamete exactas y apreciarlas con nuevos ojos, con un nuevo vocabulario más rico para expresar ideas y finalmente con mayor conocimiento que antes.

Ésta tesis fué patrocinada por Conacyt en su programa de becas de posgrado y también gracias al proyecto SEP 2004-C01-47972.

Domo Domo Arigato

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Análisis de los Datos

El presente trabajo es un caso de estudio en el cual se quieren responder preguntas respecto al comportamiento de los usuarios que transitan por el website del CIMAT.

A diferencia de la literatura consultada [9], [13], [14], [15], [16], [17], [18], [19], [20] queremos enfatizar la manera en que los usuarios navegan en el tiempo y su interacción con el contenido que se les presenta; dando a conocer a través de éste trabajo ¿Cómo es que los usuarios navegan? así como ¿De qué manera se lleva a cabo en el tiempo , el acto de visitar las páginas?. Es por eso que en el análisis de los datos la característica temporal tendrá un especial énfasis.

Con éste estudio queremos poder distinguir grupos de usuarios (en caso que existan) y modelarlos de alguna manera, de acuerdo a sus preferencias de navegación de las páginas.

El modelo propuesto no solo debe de clasificar usuarios de acuerdo a su patrón de navegación, también debe ser posible comparar y visualizar dichos modelos de usuarios de una manera intuitiva y fácilmente comprensible.

1.1. Descripción de los datos

De los datos con los que vamos a trabajar, podemos distinguir dos características fundamentales en ellos, la primera: una estructura estática relacionada en la disposición e interacción de la información contenida en el website y la Segunda una estructura dinámica relacionada con la manera en que esta información es consultada.

1.1.1. Estructura Estática

Ésta característica de los datos corresponde a la manera en que el website está conformado, la información contenida y la manera en que ésta es presentada al navegante para su uso.
El website está constituido por páginas cuya información tiene relación con otras páginas, además de contar con referencias (*hipervínculos*) las cuales impactan directamente en la manera en que se consulta (navega) la información.

Los hipervínculos contenidos en las páginas y el orden en que éstos referencian a otras páginas son los ejes fundamentales para poder obtener un diagrama que nos represente el website. Una representación muy adecuada es un grafo, donde cada página es un nodo y cada hipervínculo es una arista entre dos nodos. Es necesario tener en cuenta que las aristas entre nodos tienen dirección y es posible crear con ellas rutas entre nodos que formen ciclos[1] obteniendo así un Grafo Dirigido Cíclico (DGC *Directed Cyclic Graph*).

El grafo implícito en la estructura estática del website fue obtenido a través de un script en Perl que busca de manera exhaustiva las páginas contenidas en el website y los hipervínculos en ellas. Para lograrlo se tuvieron que hacer varios filtros de páginas y ligas, debido al obstáculo técnico que es un sitio dinámico (programado con PHP) por lo que una estructura de archivos en el disco duro no tiene sentido ni corresponde estructuralmente al contenido y presentación de la información. Además se tuvo que filtrar y mapear la información contenida en ambas versiones de la página del CIMAT, ya que cuenta con versión de texto y versión gráfica, las cuales contienen la misma información pero es presentada de una forma visualmente más atractiva. Gracias a este script es posible manejar las páginas de acuerdo a la notación de hemos definido, garantizando que la información en cada página listada sea única.

También es a través de éste script que removemos los datos que no son generados por usuarios humanos (robots), como lo son los buscadores *google* o *yahoo*, los cuales no son considerados en el presente estudio.

Un problema relacionado con la obtención de este grafo es la manera de visualizarlo; el mostrar este tipo de estructuras gráficas presenta problemas de una alta complejidad por la necesidad de representar los datos de una forma informativa, ordenada y limpia a los ojos del usuario.

Por ejemplo podemos tomar un tipo grafo muy sencillo como es el árbol binario (figura 1-1), el cual solamente tiene dos nodos hijo por nodo, pero a pesar de su sencillez, el número de nodos hijos en el nivel más bajo crece a un ritmo casi exponencial (x^2)

De querer distribuir uniformemente los nodos para cada nivel del árbol, es necesario un espacio de trabajo muy grande para pocos niveles. Establaciendo que es un problema de espacio podemos recurrir al uso de un

[1] Un ciclo en un grafo se define como una secuencia de conexiones en un grafo que parten de un nodo A y termina en A

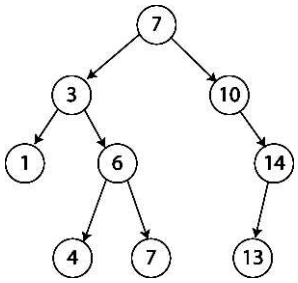


Figura 1-1: En un caso tan sencillo de un grafo como el árbol binario, el espacio entre cada nodo decrece casi exponencialmente

tipo de geometría diferente a la euclidiana, que nos proporcione una forma de medir el espacio entre nodos de manera que nos convenga.

Tal es el caso de la geometría en el espacio hiperbólico la cual cumple todos los axiomas de la geometría euclidiana, con excepción de su último postulado en el que solamente existe una línea l paralela a k que contenga un punto P que no este en k. Éste espacio esta denotado por $H^n = \{x \in \mathbb{R}^{(n,1)} \mid \langle x, x\rangle = -x_0^2 + x_1^2 + x_2^2 \ldots + x_n^2 y x_0 > 0\}$ [8]

Es a través del espacio hiperbólico y su métrica [2] que podemos asignar una porción de espacio entre nodos y distribuirlos de mejor manera a pesar de los niveles que pueda tener el árbol.

Para visualizar éste y otros tipo de grafos mucho más complicados, es posible utilizar un modelo que nos proyecte todo el espacio hiperbólico en una porción finita del espacio euclidiano.

Para ejemplificar esto tomemos un caso sencillo como un eje hiperbólico H^1 ; que vive en un espacio $\mathbb{R}^2$. Es por eso que a pesar de considerar el caso más sencillo tenemos un espacio que contiene dos componentes, en la figura 1-2 podemos observar un eje hiperbólico el cuál solamente puede ser graficado utilizando $\mathbb{R}^2$ como la mitad de una hipérbola.

Para construir una proyección de todo este espacio H^1 sobre un segmento de $\mathbb{R}^1$ (resaltado en rojo en la figura 1-2) alineamos el centro del segmento con el centro del eje hiperbólico y tomamos un punto llamado 'punto de vista' proyectando cada punto en el eje sobre el segmento como muestra la figura.

Es de notar que los puntos en los extremos del segmento representan puntos en el infinito del eje hiperbólico, lo cual cumple lo que buscábamos que es confinar todo éste espacio en un segmento finito, obviamente a costa de la deformación del espacio original.

Existen dos formas de hacer esta proyección y es a través de dos modelos: el de Klein y el de Poincaré,

[2] La métrica utilizada es la de Minkowski

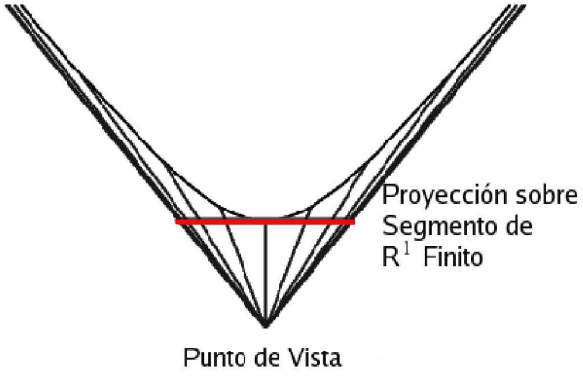


Figura 1-2: Proyección de un eje hiperbólico en una proporción de espacio euclidiano finito

conocidos también como modelo proyectivo y conformal respectivamente, los cuales tienen ciertas diferencias en cuanto a la distorsión que hacen en el espacio original. El modelo proyectivo mantiene las líneas rectas, pero distorsiona los ángulos entre ellas a diferencia del conformal que preserva los ángulos distorsionando las lineas dándonos una mejor proyección en términos de contexto de nodos en el grafo de acuerdo a [6].

Gracias a esta proyección en el espacio euclidiano obtendremos una buena vista foco+contexto (Focus+context) donde apreciamos la conectividad del nodo de interés (focus) sin perder su localización general en el grafo (context) ya que al deformarse el espacio los vecinos del nodo de interés son reacomodados dejando ver la conectividad entre estos nodos [4], [5].

Un ejemplo de este tipo de proyección para el caso tridimensional se encuentra en la figura 1-3 en la cual podemos ver la esfera sobre la cuál se está proyectando un grafo de ejemplo, permitiendo así la visualización de los datos consultados. Es de notar que todos los nodos están colocados en la superficie de la esfera con excepción del nodo de interés el cuál se encuentra en el centro de la misma, esto se debe que éste nodo se encuentra alineado con el llamado punto de vista y gracias a esto nos permite ver las conexiones existentes entre el nodo de interés con sus vecinos.

Usando éste tipo de proyección del espacio hiperbólico para el caso cuando $n = 3$ podemos visualizar el grafo a través del software Walrus el cual forma parte del proyecto CAIDA (*Cooperative Association for Internet Data Analysis*).

El método de visualización del grafo (GDC) del website por medio de éste programa está basado en proveer un GDA (*Grafo Acíclico Dirigido*) auxiliar que será el esqueleto de la visualización. Éste GDA auxiliar es obtenido de un GDC el cuál es un grafo más completo, con las restricciones que debe contener

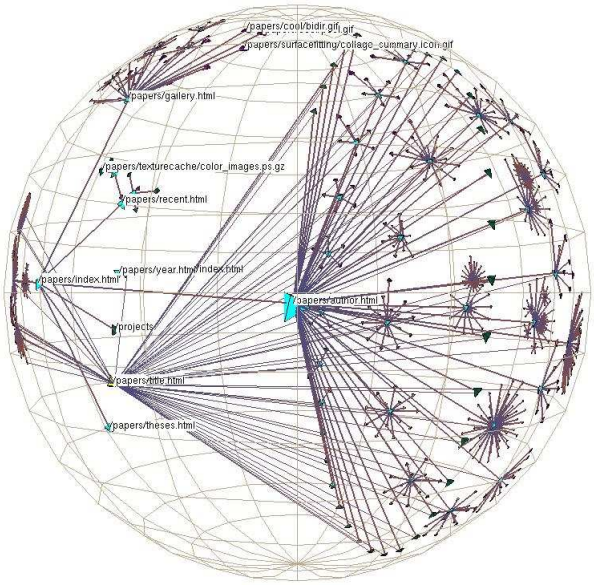


Figura 1-3: Grafo Acíclico Dirigido (DAG) construido usando geometría hiperbólica en $\mathbb{R}^{3+1}$ proyectado sobre la superficie de una esfera $\mathbb{R}^3$

todos sus nodos y cada nodo debe de tener una conexión a el.

En la figura 1-4 podemos ver la estructura del GDC que conforma el website y apreciar el GDA el cual está definido por las conexiones de color sólido y cada nodo representa una página del sitio.

Los colores de los nodos en la figura 1-4 únicamente reflejan el orden en que los nodos fueron buscados, pero nos da una idea de como es la topología del sitio. Por ejemplo el grupo de nodos de color naranja de las figuras pertenecen al grupo de páginas del NotiCimat, el verde a Eventos y los azules en su mayoría al contenido de la página principal. También podemos observar una estructura circular dominante que se aprecia en la parte superior de la figura 1-4 y que pertenece a los nodos que se encuentran en la página principal, de estas se desprenden varias páginas de cada uno de los temas principales del CIMAT.

Hemos podido obtener una representación fiel del grafo implícito en las páginas del website y los hipervínculos entre ellas, dándonos una mejor idea de cómo está conformado el sitio. Es de aclarar que es posible que haya páginas albergadas en el servidor del CIMAT que no hayan sido representadas en este grafo, pero esto se debe a que no hay un hipervínculo a ellas dentro de las páginas visitadas, tal es el caso de algunas páginas personales y de eventos especiales cuya sede es el CIMAT.

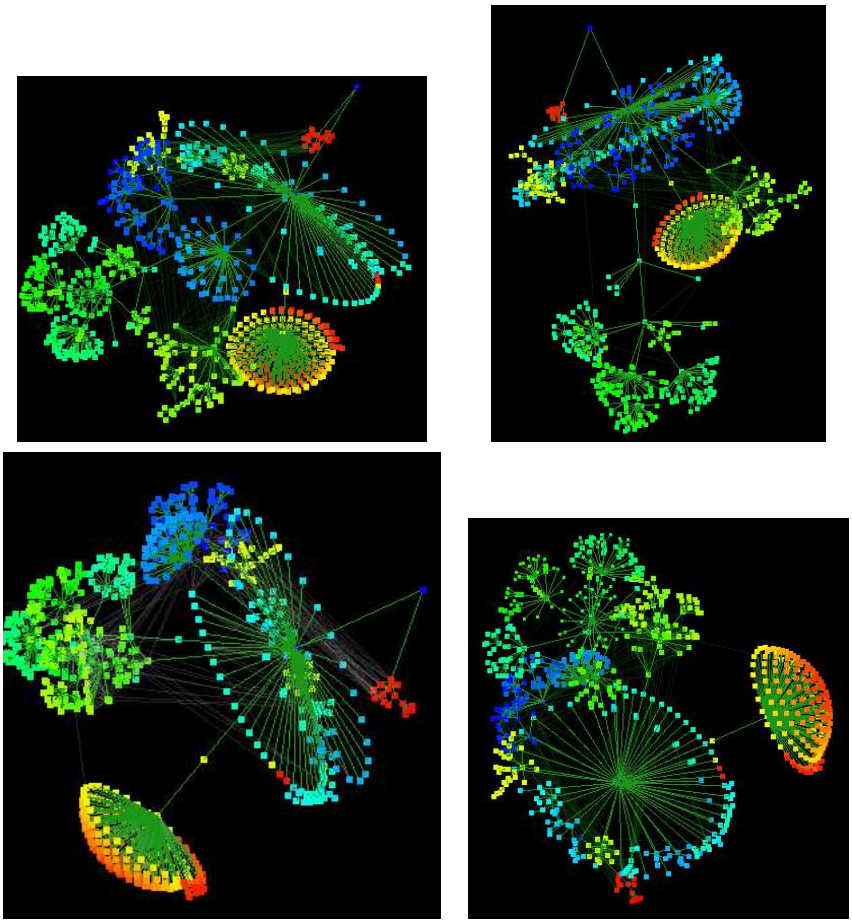


Figura 1-4: Grafo de las páginas del CIMAT, obtenido de la búsqueda exhaustiva de las ligas en sus páginas

1.1.2. Estructura Dinámica

Además de la estructura estática se encuentra también la estructura dinámica de los datos, éstos se generan cuando los navegantes consultan las páginas del website. Los datos con los que se trabajarán son obtenidos del servidor de páginas del CIMAT y solamente se trabajará con los datos generados por personas que navegan fuera de la red interna. También se removerán automáticamente las consultas hechas por programas buscadores llamados 'robots' que consultan el website en busca de información.

Éstos datos son almacenados en un archivo con un formato dado por la W3C (*Wold Wide Web Consortium*) generando enormes cantidades de información. El archivo es una bitácora (logfile) de información que es descargada del servidor de páginas web a un lugar remoto. Para su análisis debe ser preprocesado eliminando la información redundante así como la no necesaria, por ejemplo archivos de imágenes y peticiones al servidor fallidas. [1].

El formato del archivo log que utilizaremos es el siguiente :

IP - Fecha - Mét odo Archivo - Código - Locación Origen - Información Extra

A continuación tenemos un breve ejemplo del archivo donde podemos ver la información extra que se nos proporciona como el tipo de navegador y sistema operativo.

```
211.22.230.68 -- [10/Jan/2005:20:16:32 -0600] "GET /index/html HTTP/1.1" 200 3277 "http://www.cimat.mx/emo2005/" "Mozilla/4.0 (compatible; MSIE 5.01; Windows NT 5.0)"
211.22.230.90 -- [10/Jan/2005:20:16:32 -0600] "GET /botones/10_b.gif HTTP/1.1" 200 3428 "http://www.cimat.mx/emo2005/" "Mozilla/4.0 (compatible; MSIE 5.01; Windows NT 5.0)"
211.22.230.90 -- [10/Jan/2005:20:16:32 -0600] "GET /botones/10_b.gif HTTP/1.1" 200 3428 "http://www.cimat.mx/emo2005/" "Mozilla/4.0 (compatible; MSIE 5.01; Windows NT 5.0)"
211.22.230.68 -- [10/Jan/2005:20:16:32 -0600] "GET /botones/12_b.gif HTTP/1.1" 200 4891 "http://www.cimat.mx/emo2005/" "Mozilla/4.0 (compatible; MSIE 5.01; Windows NT 5.0)"
211.22.230.68 -- [10/Jan/2005:20:16:32 -0600] "GET /botones/12_b.gif HTTP/1.1" 200 4891 "http://www.cimat.mx/emo2005/" "Mozilla/4.0 (compatible; MSIE 5.01; Windows NT 5.0)"
```

Como parte del preproceso del log file, es necesario considerar las características de los datos enumerados a continuación :

- Datos incompletos: No es posible reconstruir la sesión real de un usuario debido a que solamente contamos con las requisiciones de archivos que el usuario haga al servidor. Ésto se debe a que los programas de navegación en Internet (navegadores *Browsers*) al hacer uso de memoria temporal (*cache*) por lo que existen páginas que fueron visitadas, sin que el servidor se percate de ello, todo esto se debe al hecho que en la computadora del usuario en *cache* fueron almacenadas y consultadas sin hacer peticiones de envio de información al servidor.
- Duplicidad de IP : Debido a que en Internet existen una infinidad de redes y subredes y es necesario aplicar un enmascaramiento. Una misma IP se utiliza para un grupo de máquinas resultando en

IP's con cualquier cantidad de usuarios consultando el website usando la misma dirección. Al no ser especifica una IP por usuario no es posible separar exactamente las sesiones ni saber cuándo ocurre el enmascaramiento.

- Análisis de datos de longitud variable: Suponiendo que la extracción de las sesiones de los navegantes fuera acertada, no existe forma sencilla de comparar 2 cadenas de longitud diferente. Los métodos tradicionales de clasificación y análisis trabajan sobre espacios fijos de caracterísitcas por observaciónes complicando la estandarización de cadenas bajo cierto criterio.

Para resolver el primer problema de los datos no hay una técnica estándar, aunque el uso de 'galletas' (*cookies*) en muchos casos puede ser la solución; pero no es el caso en la página del CIMAT, así que será un problema con el que lidiaremos a lo largo del análisis que haremos.

La duplicidad de IP's ha sido abordada en varios trabajos [1], [2], [3] validando dos supuestos para la obtención de las sesiones, el primero de ellos es que a cada dirección IP en el archivo se considera como una persona que está navegando dentro del website y el segundo es que después de un numero determinado de minutos de inactividad la sesión se reinicia suponiendo que consultas posteriores pertenecen a otro navegante.

La manera en que se extrajeron estas características del archivo fue a través de scripts en Perl que obtienen las páginas visitadas por los usuarios dentro de un rango de tiempo además el orden y fecha en que se visitaron las páginas.

La longitud de las sesiones es en extremo variable, aún aplicando el criterio de reiniciar una sesión después de cierto intervalo de inactividad, se llegan a producir sesiones de cientos de brincos los cuales suponemos no son naturales, ni prácticos para su análisis [1]. A esto debemos añadir que no existe una métrica sencilla para medir la distancia entre 2 sesiones, debido a que hay que tomar en cuenta su estructura temporal, así como sus diferencias de longitud.

Una vez aplicado el script a los datos de entrada obtenemos una lista de visitas agrupadas por IP en el formato de duplas, donde el primer elemento es el número de página que visitó y el segundo una marca de tiempo indicando el momento en que lo hizo. Un ejemplo de aplicara el script a los datos de entrada es el siguiente :

```
1 63676949 12 63676949
9 63861957 9 63861957 9 63861957 9 63861957 9 63861958 9 63861958 9 63861958 9 63861958 9
14 7 63560138 7 63560138 7 63560148 7 63560148 7 63560153 7 63560153 7 63560162 7 63560162 7 63560167 7 63560167 7 63560172 7 63560172 7 63560182 7 63560182
2 63968034 4 63968079 13 63968104 10 63968110
1 63628755 11 63628755
1 64292856 11 64292856
9 63951650 9 63951652 9 63951652 9 63951652 9 63951654 9 63951654 9 63951672 9 63951719 9 63951808 12 63951808
```

```
1 63764852 6 63764852
1 64232787 11 64232787
12 63773643 12 63773648 11 63773648
11 64188759
1 63627876 11 63627876
1 63581858 9 63581858
1 63717199 11 63717199
12 64218645 12 64218667 12 64218670 12 64218670 12 64218724 12 64218744
 7 63868620 7 63868620 7 63868621 7 63868621 7 63868704 7 63868704 7 63868903 7 63868903 7 63868905 7 63868905 7 63868997 7 63868997
```

Éste proceso nos permitirá analizar los datos junto con su estructura dinámica y es precisamente del análisis de ésta parte de los datos lo que dará luz sobre la manera en que los usuarios consultan la información y sobre este tipo de datos será en los que nos enfocaremos. Nuestro énfasis consiste en responder ¿Cómo es la dinámica con la que los usuarios navegan y buscan en el sitio ? ¿Cómo es que lo hacen ?. Para ello es necesario un análisis no a nivel de páginas sino de temas o tópicos por ser ahí donde el interés de los usuarios se encuentra.

Para poder analizar los datos bajo este esquema es necesario asociar a cada página un grupo temático. Técnicamente también existe el problema de que hay varias direcciones URL (*Uniform Resource Locator*) las cuales conducen a la misma información o página, consecuencia de tener dos versiones del website, por lo que se construyó un manejador de páginas en C++, el cual administra y asocia cada página en el website a un grupo de interés que se defina.

El contenido del website en su totalidad puede ser encapsulado en los temas mostrados en el cuadro 1.1. Éstos temas son los más relevantes [3] y es a través de la interacción entre ellos que se hará el presente estudio.

Cuadro 1.1: Grupos de páginas

1 Consulta a la Biblioteca	2 Ligas externa (al CIMAT)
3 NotiCimat	4 Páginas Personales
5 Vinculación	6 Unidad Aguascalientes
7 Ingeniería de Software (Ingsoft)	8 Eventos
9 Aniversario	10 Información General
11 Publicaciones	12 Docencia
13 Investigación	14 Administración

De esta manera logramos una reducción bastante considerable en cuanto al espacio de trabajo de los datos ya que el número de páginas es del orden de los miles mientras que el número de grupos es de decenas.

[3] Relevantes hasta abril de 2005 como el 25 aniversario de la institución

Esta simplificación es bastante útil, pero lleva a la pregunta si deberían tomarse igualmente importantes todos los grupos de páginas, debido a la enorme diferencia en la cantidad de miembros de cada grupo como el porcentaje del archivo del servidor.

Además del preproceso los datos han sido filtrados para obtener un comportamiento más homogéneo entre los navegantes; el filtro aplicado toma en cuenta únicamente los accesos que hayan empezado su sesión en la página inicial del CIMAT. A menos que se indique lo contrario todos los datos deben de ser vistos bajo esta perspectiva.

En el cuadro 1.2 se muestran algunas estadísticas del archivo analizado: el porcentaje de páginas por grupo temático, el pocentaje de visitas por grupo temático, comparando el acceso a website con y sin el filtro de entrada antes mencionado.

Cuadro 1.2: Estadísticas de la estructura dinámica

Porcentaje de	páginas por grupo
00.153 %	Consulta a la Biblioteca
00.076 %	Ligas externa (al CIMAT)
15.680 %	NotiCimat
00.153 %	Páginas Personales
13.220 %	Vinculación
01.383 %	Unidad Aguascalientes
06.840 %	Ingsoft
33.051 %	Eventos
01.229 %	Aniversario
02.459 %	Información General
02.075 %	Publicaciones
14.527 %	Docencia
04.073 %	Investigación
05.073 %	Administración

Entrada General (%)	Grupo	temático	Entrada filtrada (%)
20.3	Docencia	Docencia	19.5
17.1	Información General	Información General	19.3
12.0	NotiCimat	NotiCimat	13.4
11.8	Eventos	Administración	12.1
10.6	Administración	Eventos	09.8
08.6	Investigación	Investigación	08.9
05.6	Vinculación	Vinculación	05.9
03.5	Ligas externa	Ingsoft	02.9
03.1	Ingsoft	Aguascalientes	02.0
01.8	Aguascalientes	Aniversario	01.9
01.8	Publicaciones	Publicaciones	01.9
01.7	Aniversario	Biblioteca	01.4
01.3	Biblioteca	Ligas externa	00.3
00.0	Páginas Personales	Páginas Personales	00.0

Éstas estadísticas son un poco más específicas para el caso que estamos analizando, pero comparándolas con los resultados obtenidos con otro software (libre y comercial) cuyo propósito es dar al administrador una

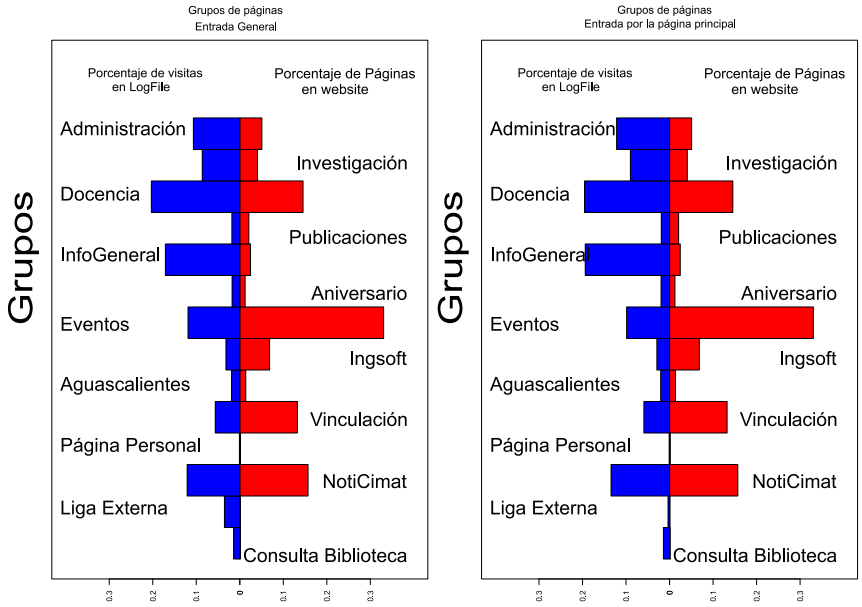


Figura 1-5: Visitas generales (izquierda), Visitas por la entrada principal (derecha)

idea de como se comportan sus usuarios, es mucho mejor ya que no toman en cuenta las complicaciones que padecen los datos que estamos manejando debido a la programación en PHP. En algunos casos solamente se consiguió ubicar dos páginas en todo el sitio lo cual es un grave error para nuestros propósitos, en otros casos las estadísticas son tan burdas como simplemente frecuencia de visita de cada elemento del sitio, resultando la página más buscada el archivo de: '`Error 404 Archivo no encontrado`'. Una ventaja que presentan estos estadísticos es generar una jerarquización de páginas en grupos temáticos hecha por el manejador de páginas.

La figura 1-5 contiene dos barplot comparando el porcentaje de accesos al sitio con el número de páginas en cada grupo temático, en la parte izquierda se encuentra el comparativo de accesos en general del sitio y a la derecha aplicando el flitro de entrar por la página principal; en azul las consultas hechas en línea (*online*) que se le hacen a un grupo de páginas tomadas del logfile [4] y en rojo el porcentaje de páginas por grupo en el website.

Como contextualización de los datos si cada grupo de páginas fuera visitado proporcionalmente al número de páginas contenidas el barplot sería simétrico, lo cual no sucede sugiriéndonos preferencias de páginas o grupos temáticos.

Con la reducción del espacio de trabajo simplificamos las sesiones obtenidas como una sucesión de saltos entre grupos temáticos.

Una manera de analizar estos datos es obtener de ellos características las cuales son mapeadas a un vector

[4] El archivo tomado para esta gráfica pertenece a las fechas de 3 de junio al 12 de junio de 2005

de longitud fija (*feature space*), esperando que los atributos extraidos de cada sesión reflejen su estructura y así poder trabajar con ellos a través de métodos más tradicionales.

Las características usadas para representar las sesiones son las siguientes :

Cuadro 1.3: Características extraidas de las sesiones

1	Longitud de la cadena.
2	¿Cuántas veces cambia de grupo temático ?
3	Máximo número de visitas consecutivas a un mismo grupo.
4	¿Cuántos grupos temáticos fueron visitados?
5	¿Qué grupo temático fue el más visitado?
6	¿Cuántas veces fue visitado el grupo más frecuente?.
7	Tiempo promedio de latencia entre visitas.
8	Máximo tiempo de estadía entre visitas.
9	¿En cuál grupo temático permaneció más tiempo?.
10	Duración de la sesión (tiempo).

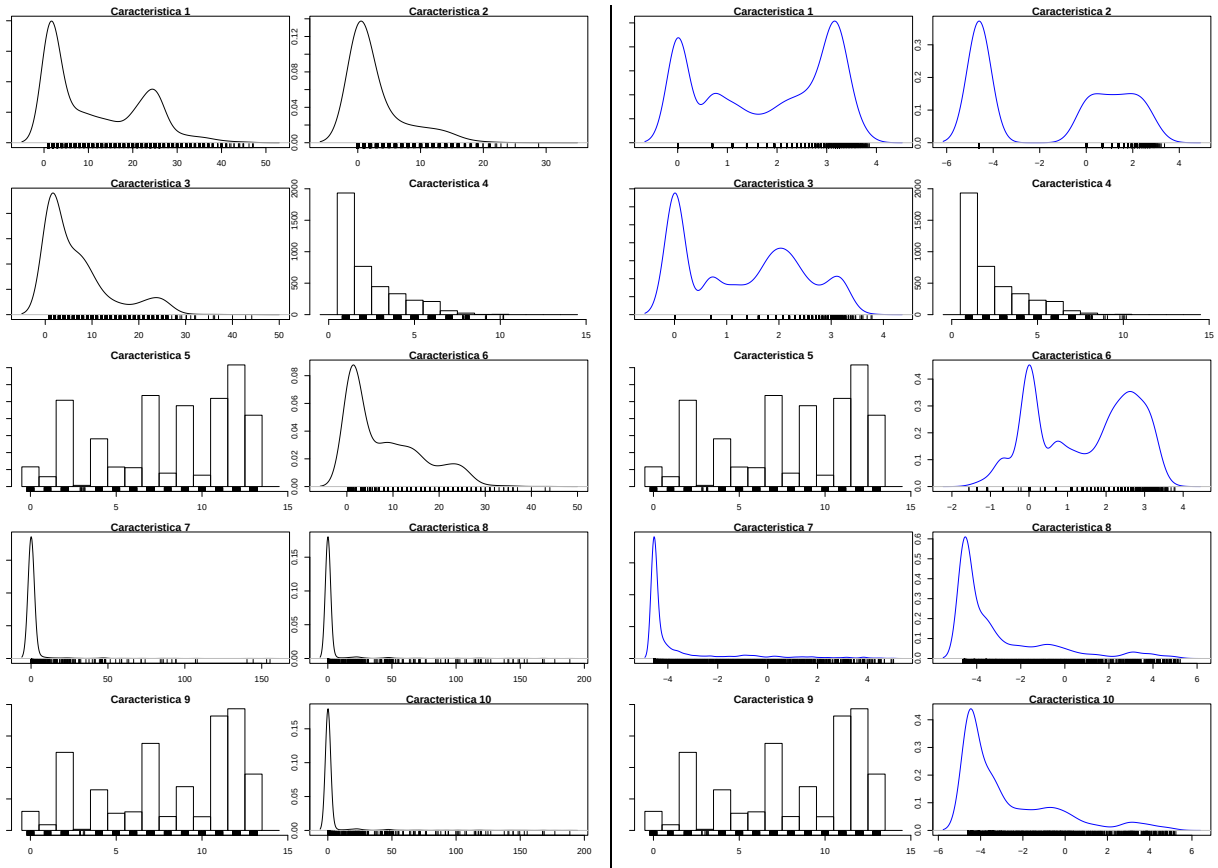
A cada característica del *feature space* se le ha estimado su densidad encontrando cierta estructura en ellas, aunque es necesario tener en cuenta que las variables (5) y (9) son variables categóricas.

Existen atributos que presentan densidades multimodales, hecho que nos hace pensar que cada moda pudiera ser causada por diferentes grupos de usuarios o clusters de usuarios con preferencias en común, tal es el caso de las características (1), (3), (4) y (6).

También hay un efecto visual en cuanto a la presentación de los datos, causado por el rango de valores que toman las caracterísitcas (7) y (8) donde hay observaciones mucho muy alejadas de la media. Éste efecto se ha visto corregido aplicándo una transformación logarítmica a los predictores lo cual mejora visualmente la apreciación de dichas densidades. En la figura 1-6 se ha colocado un cuadro comparativo entre las densidades de las características, a la izquierda tenermos las densidades calculadas sin transformación alguna y en la parte derecha de color azul transformados a escala logarítmica de base 20 para apreciar mejor el comportamiento de los usuarios. Es de notar que las características que aparecen como histogramas es debido a su carácter entero.

A continuación enumeramos algunas de las características de los datos que hemos encontrado las cuales consideramos más importantes para su comprensión:

1. Existe un claro distintivo entre usuarios con sesiones largas y sesiones cortas debido a la bimodalidad de longitud de las sesiones. La densidad de la caracterísitica de longitud de sesión contiene dos modas,



Escala Natural *Escala Logarítmica*

Figura 1-6: Izquierda características en escala natural, derecha características en escala logaritmica

una en sesiones cortas 63.48 % de longitud menor o igual a 15 y otra en sesiones largas 36.51 % de longitud entre mayores a 15.

2. El comportamiento de las sesiones de acuerdo a longitud y duración se resume de la siguiente forma:

Cuadro 1.4: Comporamiento por duración y longitud

	Sesiones largas (>15)	Sesiones cortas ($\leq$ 15)	Todas las sesiones
Latencia promedio entre visitas	6 seg aprox. (0.10 min.)	2.10 min	1.37 min.
Duración de sesión	2.51 min.	3.51	3.10 min.

3. El porcentaje de sesiones que revisita al menos un grupo de páginas es 23.96 %. Hemos definido la revisitación a un grupo de páginas como la aparición de un grupo previamente visitado una vez que se ha visitado un grupo diferente.

4. El 48.21 % de las sesiones visitan un solo grupo temático. En la figura 1-7 tenemos el histograma de los temas más recurrentes por la sesiones que solo visitan un grupo temático. En el cuadro 1.5 el porcentaje de sesiones de éste tipo por grupo temático. Ésto nos suguiere que hay un especial interés por algunos grupos destacándose: Docencia, NotiCimat, Investigación y Eventos.

Cuadro 1.5: Grupos de páginas únicamente visitados por sesión

Consulta a la Biblioteca 02.16 %	Ligas externa (al CIMAT) 00.00 %
NotiCimat 19.50 %	Páginas Personales 00.16 %
Vinculación 06.50 %	Unidad Aguascalientes 03.50 %
Ingeniería de Software (Ingsoft) 03.16 %	Eventos 17.16 %
Aniversario 01.16 %	Información General 04.33 %
Publicaciones 01.16 %	Docencia 19.50 %
Investigación 18.33 %	Administración 03.33 %

La figura 1-8 muestra para cada sesión la longitud (eje X) y el porcentaje que conforman las visitas al grupo más popular para cada sesión (eje Y). En verde la media de Y para cada X y en marron la media omitiendo el grupo de observaciones que solamente visitan un solo grupo ($y = 1{,}0$). Coloreamos de azul las cadena cortas y de rojo las cadenas largas.

Observamos que el porcentaje de páginas que uno dedica al grupo más popular no cambia significativamente con la longitu de la sesión.

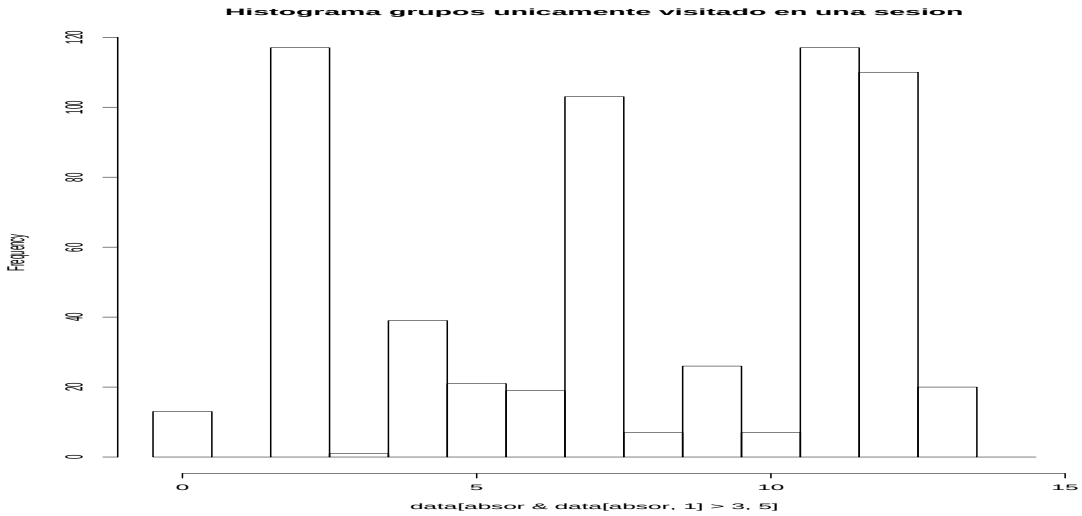


Figura 1-7: Histograma de grupos visitados únicamente

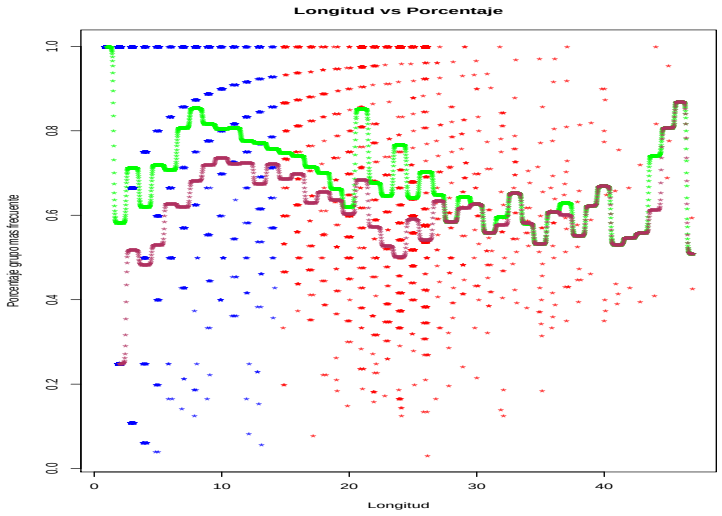


Figura 1-8: Razón entre grupo más visitado y longitud de sesión en función del largo de la cadena

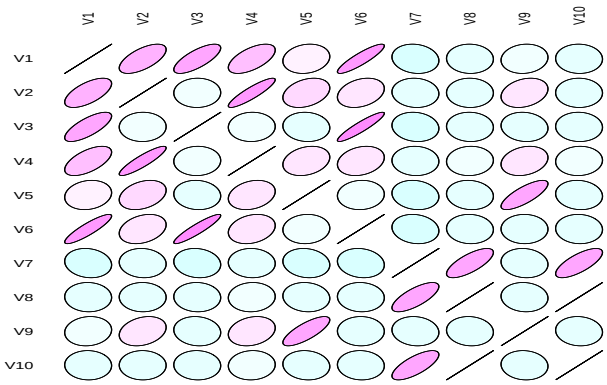


Figura 1-9: Correlación entre predictores

Además éste porcentaje es bastante alto ($> 50\,\%$).

5. Hay pocos atributos altamente correlacionados. La figura 1-9 es una representación de la matriz de correlación entre los predictores donde la elipticidad de cada elemento refleja el valor de correlación y el color junto con la orientación su signo (azul e inclinación a $-45°$ grados indica negativo).

   El décimo predictor (duración de la sesión) es áltamente correlacionado con el máximo tiempo de estadía en un grupo (predictor 8).

6. El atributo (2) (cambio de grupos temáticos con media alrededor de 3 (2.79)) junto con la visitación a pocos grupos diferentes (2.30 en promedio) nos sugieren que la consulta de información se hace visitando varias veces el mismo grupo temático ántes de cambiar de tema, evitando la revisitación de grupos.

7. En la figura 1-10 se muestra para cada sesión la logitud (eje X) y el porcentaje que el tiempo máximo de estadia entre páginas representa de la duración total de cada sesión (eje Y). En verde la media de Y para cada valor de X y coloreados de acuerdo a su longitud cadenas largas de rojo y cortas de azul.

   A mayor la duración de la cadena mayor resulta ser el tiempo máximo entre páginas.

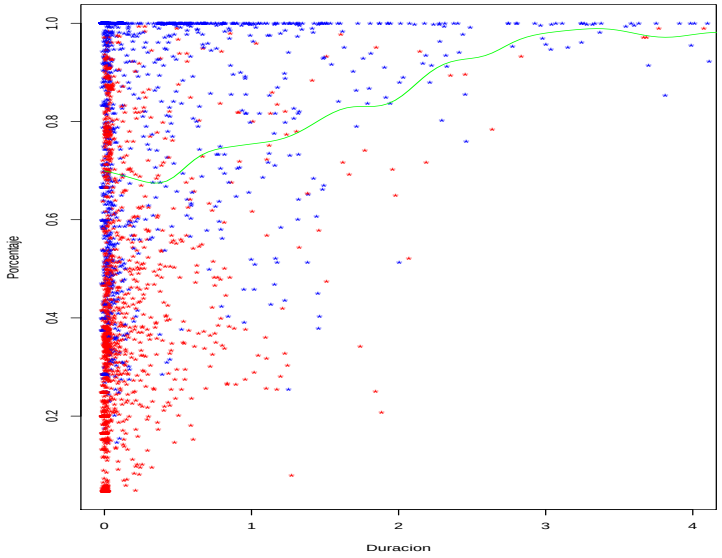


Figura 1-10: Razón entre tiempo máximo de estadia y duración de sesión en función de la duración de la sesión

1.2. Análisis de Proyecciones

Una manera diferente de ver como se distribuyen datos es a través de transformaciones que permitan apreciar mejor sus características, que debido a su dimensionalidad no es posible visualzar de una manera sencilla.

Dos técnicas usadas para éste propósito son: ***PCA*** (*Principal Component Analysis*) e ***ICA*** (*Independent Component Analysis*) [23].

Éstas técnicas buscan proyecciones que ayudan a 'ver' mejor los datos. En el caso de PCA se buscan proyecciones ortonormales entre si cuya varianza sea la mayor posible. Técnicamente la manera de buscar estas componentes es a través de la obtención de los eigenvectores de la matriz empirica de covarianza de los datos, los cuales corresponden a las direcciones de máxima varianza.

Para ICA se asume un modelo generativo de los datos, el cual supone que éstos son una mezcla lineal de variables latentes, mutuamente independientes. La proyecciones 'interesantes' las definimos para éste análisis como menos gausianas.

Para estos dos tipos de análisis queremos características muy descriptivas de las sesiones por lo que hemos modificando algunos atributos de la tabla 1.3 convirtiéndolos en atributos más cualitativos que cuantitativos. Una forma de hacerlo es modificarlos por medidas relativas, tal es el caso de los atributos (2), (3), (6), (7) y (8) del espacio de características original a los cuales hemos aplicado este cambio.

El espacio de características con el que trabajaremos será el siguiente:

Cuadro 1.6: Características extraidas de las sesiones

1	Longitud de la cadena.
2	Cambios de grupo temático por longitud.†
3	Porcentaje de visitas consecutivas a un mismo grupo por longitud.†
4	¿Cuántos grupos temáticos fueron visitados?.
5	¿Qué grupo temático fue el más visitado?.
6	Porcentaje de grupo más popular por longitud.†
7	Porcentaje del Tiempo promedio de latencia entre visitas.‡
8	Porcentaje que representa el máximo tiempo de estadía en un grupo de páginas.‡
9	¿En cuál grupo temático permaneció más tiempo?.
10	Duración de la sesión (tiempo).

† Características relativas a longitud de sesión.

‡ Características relativas a duración de sesión.

En la figura 1-11 mostramos las primeras 3 componentes de las 2 técnicas antes mencionadas, la razón de hacer esto es que a través de éste número de componentes principales (PCA) podemos captar el 87.61 % de la varianza de los datos. Arriba de la diagonal se encuentran las proyecciones del análisis con PCA y debajo con ICA, en la diagonal el boxplot de cada componente. Para éste análisis hemos descartado las variables categóricas (5), (9) por la naturaleza de los datos.

La figura 1-11 es una comparación de dos características de los datos: número de grupos vistados durante la sesión y lóngitud de la sesión. A la izquierada indicamos el número de visitas con una escala de colores donde rojo significa alta intensidad (un solo grupo visitado) y amarillo baja intensidad (10 grupos visitados). A la derecha de color rojo las cadenas largas (≥ 15) y azul las cadenas cortas (<15).

Ésta comparación crea un mapa de los datos usándo la primeras dos componentes de PCA que captura en 75.90 % de la varianza. Recorriendo de la esquina superior derecha a la esquina inferior izquierda se encuentran ordenados por su longuitud de menor a mayor; de la esquina superior izquierda a la esquina inferior derecha en orden ascendente por el número de grupos visitados. La proyección de estos dos componetes captura el 75.90 % de la varianza de los datos.

Los primeros 3 componente del análisis de PCA se encuentran en el cuadro 1.7

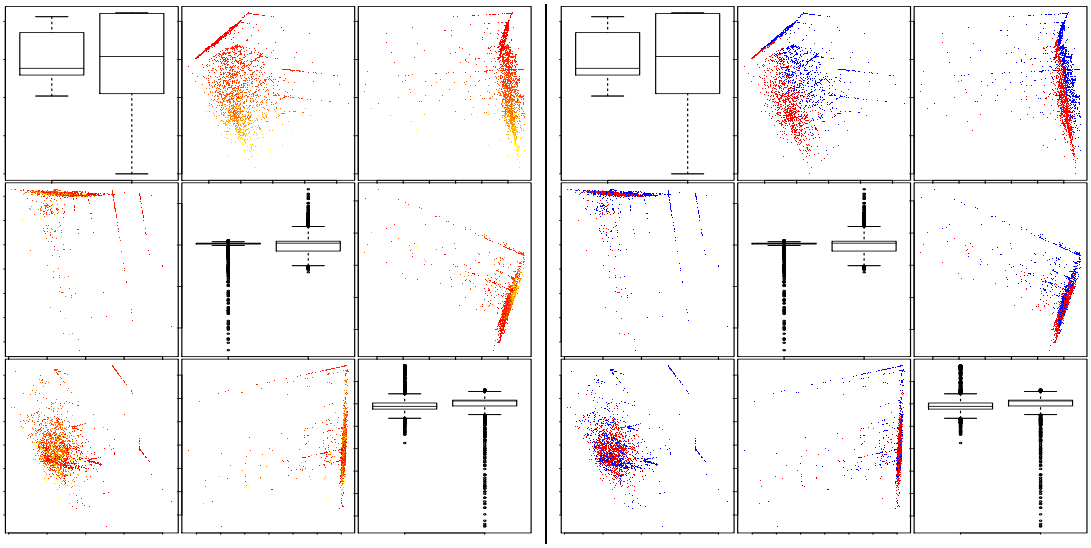


Figura 1-11: Proyecciones encontradas por PCA (arriba de la diagonal) proyecciones de ICA (abajo de la diagonal) coloreadas de acuerdo al número de grupos visitados (izquierda) y longitud (derecha)

Cuadro 1.7: Primeros 3 componentes del PCA

Primer Componente		Segundo Componente		Tercer Componente	
Atributo	*PCA*	*Atributo*	*PCA*	*Atributo*	*PCA*
1	−0,334911	1	−0,430076	1	−0,222839
2	−0,396088	2	0,353726	2	0,163211
3	0,440214	3	−0,214212	3	−0,090967
4	−0,426614	4	0,087008	4	0,061957
6	0,382843	6	−0,366770	6	−0,125470
7	0,353074	7	0,393324	7	0,209903
8	0,287834	8	0,486311	8	0,000189
10	−0,032896	10	0,332765	10	−0,922924

Desaforfunadamente no es fácil hacer una interpretación de los componentes que obtenemos de éste análisis, aunque del primer componente pudieramos decir que nos da una idea de contraste entre diversidad de grupos distintos que visita y una permanencia en cada grupo.

Con éste análisis podemos obtener un diagnóstico de como se comportan los datos pero no podemos establecer una relación clara de la importancia que juegan las características que hemos medido.

1.2.1. Componentes Robustos

Existen versiones robustas de PCA que atacan el problema de la sensibilidad de los componentes al ser estimados a través de la matriz de covarianza. Los métodos robustos buscan minimizar el efecto de

observaciones atípicas que modifican la dirección de las componentes y no capturan la variación de los datos regulares.

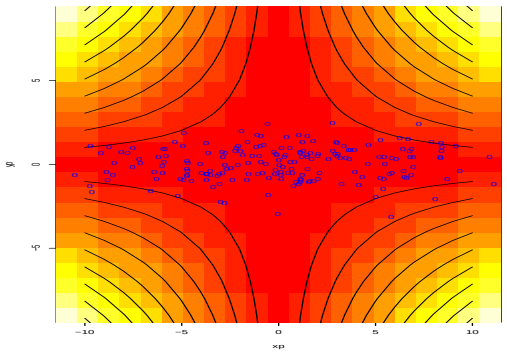


Figura 1-12: Observaciones en azul, de rojo a amarillo el valor de los pesos asignados de mayor a menor

Rob-PCA (*Robust Principal Component Analysis*) es una técnica robusta que pesa los datos de acuerdo a una medida que toma en cuenta la distancia a los componentes. La figura 1-12 es un ejemplo donde se muestran las observaciones y los pesos asociados a la medida. Para éste caso los componentes coinciden con los ejes X y Y, en azul las observaciones, de rojo a amarillo los pesos de la medida donde rojo tienen mucho peso y amarillo poco peso, en negro las curvas de nivel de los pesos.

Gracias a ésta medida podemos catalogar las observaciones en regulares y atípicas (outliers). Robusteciendo el método inensibilizando a datos atípicos.[24]

Aplicando ésta técnica obtenemos que un componente es suficientes para explicar el 98.14 % de su varianza. La figura 1-13 muestra el primer componente del análisis donde en el cuadro de la izquierda se muestra la densidad de los datos separándolos en sesiones cortas (azul) y largas (rojas) en negro la densidad total y el la derecha la proyección de cada sesión sobre el componente coloreandolos por el número de grupos diferentes visitados.

En el cuadro 1.8 resume los datos mostrados en la figura.

Cuadro 1.8: Resumen de atributos de ROB-PCA

	Cadenas Largas	Cadenas Cortas	Visita un solo grupo	Visita varios grupos	Total
Outliers	15.61 %	21.60 %	5.71 %	31.50 %	37.21 %
Regulares	22.39 %	40.38 %	42.50 %	20.77 %	62.78 %
Total	38.01 %	61.98 %	48.21 %	51.78 %	

El componente de ROB-PCA es:

Cuadro 1.9: Componentes de ROB-PCA

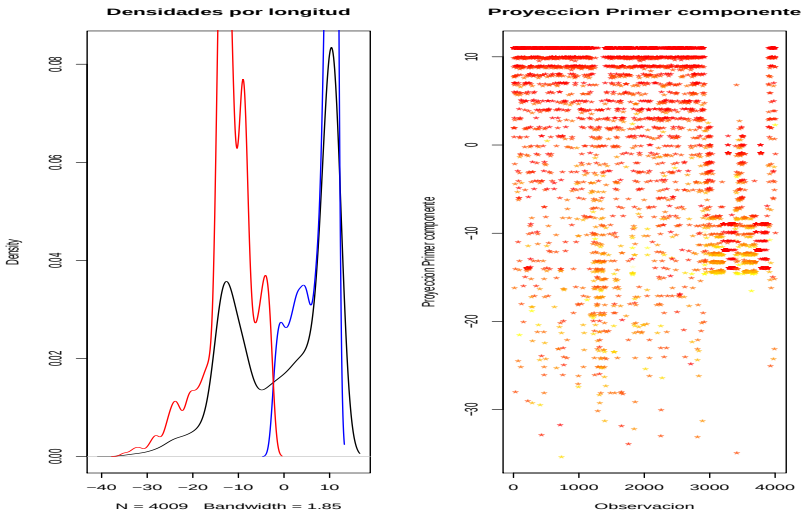


Figura 1-13: Únicamente es necesario un componente para explicar el 98.14 % de la varianza

Atributo	*Componente*
1	−0,994143130
2	−0,008466196
3	0,016287056
4	−0,098718996
6	0,009452763
7	0,032009825
8	0,021855871
10	−0,002320462

- Primer Componente ROB-PCA: Dominantemente referente a la longitud de la sesión.

Haciendo uso de la técnica ROB-PCA reducimos el número de componentes que necesitamos para explicar la varianza. Corroborando a la caracterísica de longitud como un atributo muy importante para nuestro análisis.

1.2.2. Inclusión de variables categóricas

Hemos descrito brevemente el comportamiento de las sesiones extraidas del logfile a través del espacio de características que hemos definido, en ésta sección consideramos utilizar con un espacio de características mucho más complicado como lo sería al añadir las características nominales (5) y (9).

La inclusión de variables categóricas no puede hacerse sin afectar predictores previamente definidos ya que optamos por sustituir a cada variable nominal por un vector de longitud igual a el número de grupos de páginas (n) el cuál contiene valores que miden el atributo para cada grupo de páginas. Es necesario remover 2 características ya que serían redundantes al agregar los vectores.

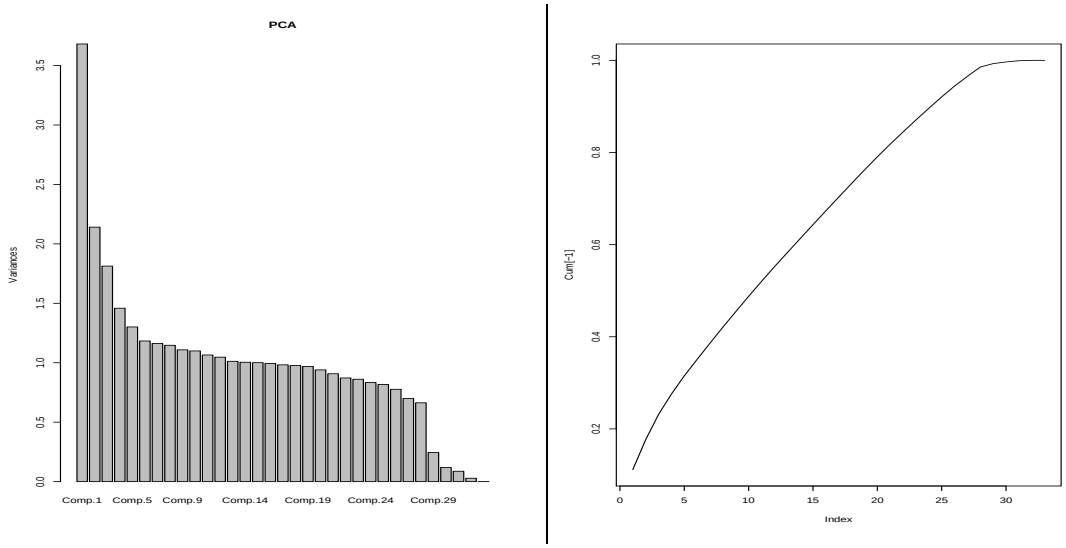


Figura 1-14: Proporción de la varianza por componente (izquierda). Varianza acumulada (derecha)

El espacio de características extendido es de la siguiente forma:

Cuadro 1.10: Características extraidas de las sesiones con variables nominales

1	Longitud de la cadena.
2	Porcentaje que representan los cambios de grupo temático por longitud.†
3	Porcentaje de visitas consecutivas a un mismo grupo por longitud.†
4	¿Cuántos grupos diferentes de páginas fueron visitados?
5-18	$F_1 = \{f_1, f_2, \dots f_n\}$ porcentaje de la sesión en que se visitó el grupo f_i.†
19	Porcentaje que representa el tiempo promedio de latencia entre visitas.‡
20-33	$G_2 = \{g_1, g_2, \dots g_n\}$ Porcentaje que representa el tiempo de estadia en el grupo g_i. ‡
34	Duración de la sesión.

† Características relativas a longitud de sesión.

‡ Características relativas a duración de sesión.

Aplicando la técnica de PCA obtenemos que son necesarios 16 componentes para explicar apenas el 91.26 % de la varianza de estos datos. En la figura 1-14 vemos a la izquierda la varianza de cada una de las componentes obtenidas por el método, a la derecha la varianza acumulada para cada una.

La existencia de datos atípicos en las proyecciones nos sugiere el uso de un método robusto para éstas características.

Componentes Robustos con variables categóricas

Haciendo uso nuevamente de la técnica de ***ROB-PCA*** obteniendo que con 2 componentes puede explicar el 99.71 % de la varianza de los datos contenidos en 34 dimensiones lo cual es muy bueno para nuestros propósitos. Los resultados están resumidos en la figura 1-15 donde abajo de la diagonal se encuentran las observaciones regulares de los datos coloreadas de acuerdo a su longitud cortas (azul) y largas (rojo), arriba de la diagonal las mismas proyecciones con los datos atípicos, en la diagonal la densidad de las observaciones regulares segmentadas en cadenas cortas y largas, en color negro las densidades sin hacer dicha segmentación.

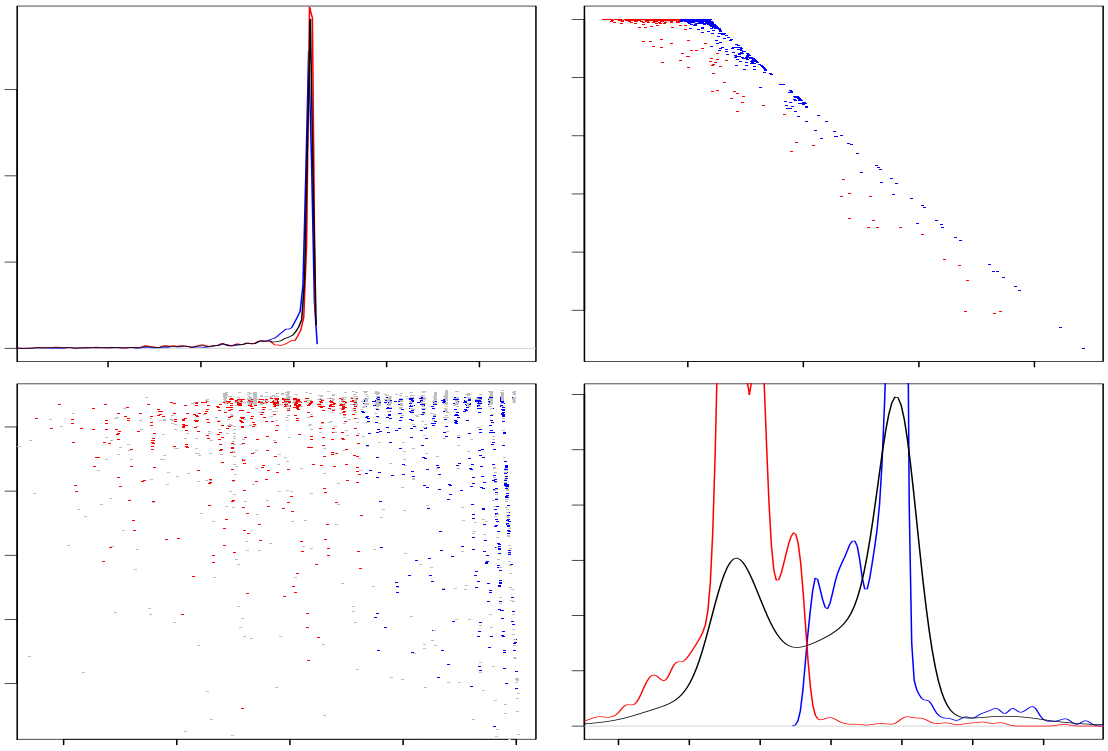


Figura 1-15: ROB-PCA puede resumir la varianza en 2 componentes de un espacio de 34 dimensiones

El cuadro 1.11 resume el comportamiento en general de este análisis

Cuadro 1.11: Distribución cadenas largas y cortas sobre cadenas típicas y regulares

	Cadenas regualres	Cadenas atípicas	Total
Cadenas Cortas	41.90 %	20.07 %	61.98 %
Cadenas Largas	24.54 %	13.54 %	38.01 %
Total	66.45 %	33.54 %	100.0 %

En el cuadro 1.12 vemos mostramos los dos componentes

Cuadro 1.12: Componentes de ROB-PCA

Atributo	*Componente*1	*Componente*2
1	$-1{,}353253e-02$	$-9{,}947372e-01$
2	$-2{,}241197e-03$	$-7{,}252782e-03$
3	$2{,}496991e-03$	$1{,}600136e-02$
4	$-9{,}449645e-03$	$-9{,}988754e-02$
5	$9{,}114274e-07$	$1{,}749531e-05$
6	$-3{,}618703e-06$	$7{,}334132e-06$
7	$-3{,}365945e-06$	$-4{,}250359e-05$
8	$-3{,}189950e-08$	$1{,}488248e-06$
9	$-1{,}489881e-06$	$1{,}515122e-05$
10	$2{,}408465e-06$	$1{,}094820e-05$
11	$-3{,}139627e-06$	$1{,}200425e-06$
12	$1{,}012117e-05$	$8{,}502618e-05$
13	$-1{,}577314e-06$	$-9{,}482800e-07$
14	$-1{,}021892e-05$	$-8{,}090436e-05$
15	$2{,}084914e-07$	$6{,}047387e-06$
16	$5{,}040333e-06$	$-5{,}254684e-05$
17	$4{,}998062e-06$	$1{,}058012e-04$
18	$-2{,}456621e-07$	$-7{,}358924e-05$
19	$5{,}260423e-05$	$5{,}364782e-04$
20	$9{,}828693e-05$	$4{,}361504e-04$
21	$1{,}017324e-04$	$4{,}430624e-04$
22	$8{,}604414e-05$	$3{,}375679e-04$
23	$1{,}023501e-04$	$4{,}410430e-04$
24	$9{,}013456e-05$	$4{,}215973e-04$
25	$9{,}993188e-05$	$4{,}298762e-04$
26	$9{,}768364e-05$	$4{,}276008e-04$
27	$1{,}035585e-04$	$4{,}310149e-04$
28	$9{,}978116e-05$	$4{,}278744e-04$
29	$1{,}057049e-04$	$3{,}874423e-04$
30	$9{,}956867e-05$	$4{,}292353e-04$
31	$1{,}063411e-04$	$2{,}852636e-04$
32	$8{,}386206e-05$	$4{,}130392e-04$
33	$9{,}433642e-05$	$2{,}937623e-04$
34	$-9{,}998581e-01$	$1{,}446405e-02$

Las interpretación que les podemos dar a éstos componentes es:

- Primer Componente: Se debe a que las características relativas (5) a (18) y (20) a (33). Éstas características forman hiperplanos al ser atributos que suman uno, es por eso que éste componente busca éste plano.
- Segundo Componente: Éste componente separa las cadenas en largas y cortas.

Capítulo 2

Modelos Markovianos

En el capítulo anterior analizamos las sesiones usando un espacio de características derivadas de las sesiones. En éste capítulo trabajaremos directamente con las sesiones a través de modelos Markovianos, con ello buscamos modelar el comportamiento de los usuarios al nivel de visitas de grupos temáticos.

Presentamos un análisis basado en éstos modelos y proponemos algunas modificaciones.

2.1. Cadenas de Markov

Las cadenas de Markov (MC *Markov Chains*) asumen la posibilidad de resumir el comportamiento futuro del sistema observando su estado actual (ver figura 2-1). Dado el estado presente, el futuro es condicionalmente independiente del pasado (ecuación 2.1).

$$P_{E_t}(j \mid E_{t-1} = e_{t-1},\ E_{t-2} = e_{t-2},\ E_{t-3} = e_{t-3},\ \ldots E_0 = e_0) = P_{E_t}(j \mid E_{t-1} = e_{t-1}) \tag{2.1}$$

Supongamos que conocemos todos los posibles estados de un sistema S en función del tiempo, de dicho sistema desconocemos su modo de operación pero a través de mediciones sabemos en que estado se encuentra el sistema.

Una manera de conceptualizar éste sistema es como se muestra en la figura 2-2 en la cual los n posibles

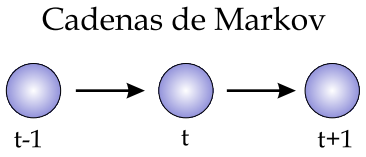


Figura 2-1: El estado futuro es condicionalmente independiente del pasado

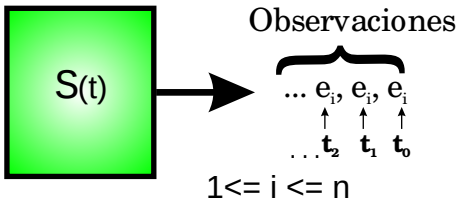


Figura 2-2: Sistema a modelar

Modelo de Cadena de Markov

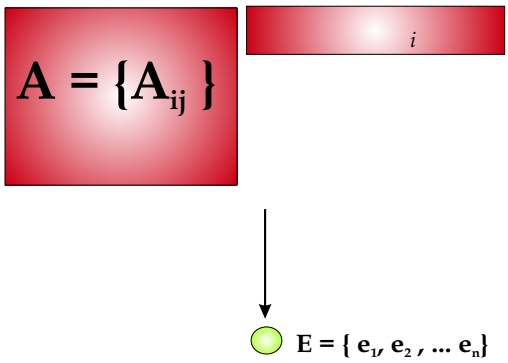


Figura 2-3: Modelo generativo de Cadenas de Markov

estados del sistema son obtenidos por medio de mediciones generando nuestras observaciones.

La dinámica del sistema $S(t)$ es parametrizada a través los elementos $A_{i,j}$ de la matriz de transición A de dimensión $n \times n$. Cada uno de éstos elementos es la probabilidad de transición del estado i al estado j, $A_{i,j} = P_{E_t}(e_j \mid E_{t-1} = e_i)$. Al conocer los n posibles estados sabemos que $\sum_{j=1}^{n} A_{i,j} = 1$.

Éste modelo explica la dinámica del sistema para el tiempo $t > 0$ pero es necesario poder tener un estado inicial en el tiempo $t = 0$ y es por eso que utilizamos un vector inicial Π donde cada uno de sus elementos Π_i representa la probabilidad de que en el tiempo t_0 el sistema se encuentre en el estado i, $\Pi_i = P_{E_t}(e_i)$ donde $t = 0$, éste vector tiene también la propiedad de $\sum_{i=1}^{n} P_{E_0}(e_i) = 1$.

Aplicando éste modelo a nuestro análisis suponemos que cada visita a un grupo temático es un estado y nuestras observaciones las transiciones hechas por los navegantes entre estados. A través de este modelo queremos explicar la dinámica de los usuarios a través del tiempo obteniendo información sobre las preferencias de los usuarios en cuanto a navegación. Éste es un modelo generativo de los datos del cual estimaremos sus parámetros $(\Pi_i, A_{i,j})$ a través de las observaciones que tenemos $(E = e_1, e_2, e_3 \ldots e_n)$ ver figura 2-3.

2.1.1. Estimación del Modelo

La estimación de éste modelo se hizo a través de las frecuencias relativas de las transiciones de un estado a otro a lo largo de todo el archivo de sesiones.

$\Pi_i \setminus A_{i,j}$	*Bibl*	*Extern*	*Noti*	*Perso*	*Vin*	*Agu*	*IngS*	*Eve*	*Aniv*	*Info*	*Publ*	*Doc*	*Inv*	*Adm*
0,0159	0,7118	0,0034	0,0051	0,0000	0,0102	0,0017	0,0034	0,0222	0,0120	0,0583	0,0051	0,0205	0,1252	0,0205
0,2797	0,0248	0,0000	0,1062	0,0015	0,0919	0,0058	0,0554	0,0570	0,0015	0,0137	0,0348	0,4688	0,1231	0,0147
0,0811	0,0008	0,0034	0,9612	0,0000	0,0022	0,0001	0,0022	0,0062	0,0022	0,0069	0,0035	0,0025	0,0062	0,0020
0,0016	0,0000	0,0000	0,0000	0,0810	0,0540	0,0000	0,0270	0,5405	0,0000	0,0000	0,0270	0,0000	0,2702	0,0000
0,0485	0,0042	0,0042	0,0061	0,0003	0,6387	0,0219	0,0057	0,0401	0,0019	0,0482	0,0131	0,1115	0,0744	0,0289
0,0178	0,0011	0,0023	0,0011	0,0000	0,2595	0,6089	0,0034	0,0149	0,0046	0,0299	0,0057	0,0161	0,0219	0,0299
0,0193	0,0006	0,0040	0,0040	0,0000	0,0127	0,0013	0,7671	0,0771	0,0060	0,0140	0,0046	0,0127	0,0577	0,0375
0,1240	0,0014	0,0018	0,0047	0,0016	0,0211	0,0020	0,0275	0,8259	0,0173	0,0143	0,0178	0,0205	0,0167	0,0267
0,0155	0,0000	0,0000	0,0058	0,0023	0,0140	0,0046	0,0046	0,1219	0,7209	0,0457	0,0023	0,0164	0,0457	0,0152
0,0763	0,0021	0,0010	0,0076	0,0000	0,0201	0,0020	0,0018	0,0135	0,0037	0,7768	0,0050	0,0231	0,0651	0,0777
0,0142	0,0060	0,0108	0,0072	0,0000	0,0472	0,0036	0,0108	0,1477	0,0072	0,0423	0,2723	0,2615	0,0581	0,1246
0,1039	0,0005	0,0078	0,0017	0,0000	0,0109	0,0017	0,0052	0,0162	0,0011	0,0345	0,0027	0,7569	0,0323	0,1277
0,1401	0,0251	0,0063	0,0031	0,0023	0,0326	0,0023	0,0145	0,0155	0,0052	0,0711	0,0127	0,0711	0,7205	0,0169
0,0613	0,0005	0,0005	0,0034	0,0013	0,0110	0,0437	0,0019	0,0277	0,0062	0,0670	0,0572	0,1957	0,0169	0,5664

Figura 2-4: Matriz de transición

El resultado de este método se encuentra el la figura 2-4 donde en la columna de la izquierda se encuentran los valores de Π_i y en la derecha los elementos $A_{i,j}$ de la matriz de transición. Arriba de cada columna se encuentran los nombres de cada grupo temático los cuales pueden ser consultados en la tabla 1.1.

Ésta estimación nos da una perspectiva general del comportamiento de los usuarios a través del sitio y sus preferencias sobre grupos temáticos al hacerlo. El análisis de los resultados obtenidos tiene dificultades para interpretarse a simple vista ya que no es presentado de una manera informativa, cuestión que resolveremos en la siguiente sección.

Podemos observar que la matriz A es diagonal dominante si excluimos la columna 2 correspondiente al grupo de páginas externas al CIMAT, las cuales no son de nuestro interés en éste estudio. La dominancia sobre los demás estados por parte de la diagonal nos indica que existe una fuerte tendencia por parte de los navegantes de hacer varias consultas a un mismo grupo de páginas antes de cambiar de grupo temático.

2.1.2. Visualización de Cadenas de Markov

Retomando el resultado del modelo de cadenas de Markov (ver figura 2-4) queremos una manera visual de representar éste tipo de resultados. Una forma natural de hacerlo es utilizando un grafo donde cada nodo representa un grupo temático y a través de aristas la probabilidad de pasar de un estado a otro como en la fig 2-5

Las aristas han sido coloreadas de acuerdo a la paleta mostrada en la figura 2-6 donde rojo representa alta probabilidad y blanco representa baja probabilidad. A lo largo de este trabajo se utilizará la misma representación excepto cuando se indique lo contrario.

Éste tipo de visualización es muy poco informativa por el excesivo número de conexiones que presenta al usuario y le es necesario incorporar la información contenida en el vector de inicio Π. Un modificador a esta propuesta se presenta en la figura 2-7 donde únicamente se muestran los valores de la matriz mayores a 0.2 y la intensidad de color en el nodo indica la probabilidad dada por Π_i

La restricción de presentar los valores mayores a 0.2 genera representaciones de la matriz visualmente más

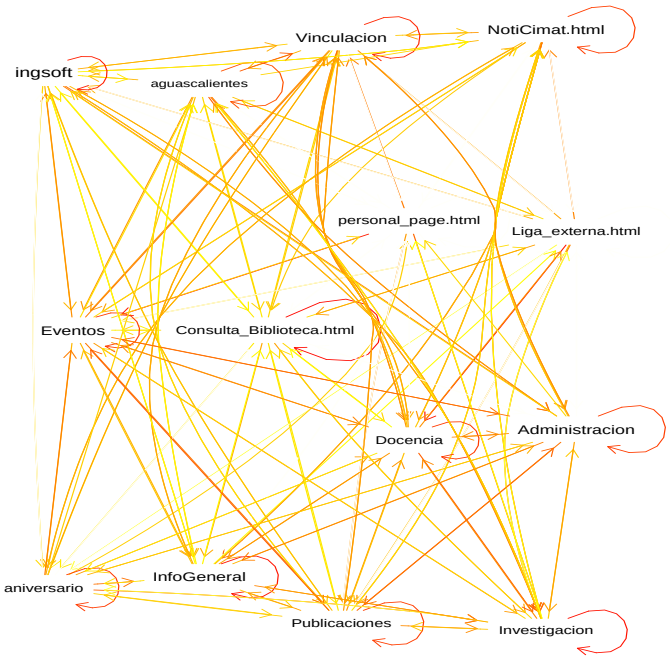


Figura 2-5: Grafo de grupos de páginas proveniente de la matriz de transición

atractivas, pero debido a la dominancia de la diagonal el escoger una restricción cuantitativa resulta poco práctica. Para remediar el fenómeno condicionaremos las probabilidades a salir del grupo, es decir mostrando las probabilidades $A_{i,j}$ condicionando a $i \neq j$.

El resultado de aplicar éste condicionamiento se puede ver en la figura 2-8 el cual es mucho más informativo. El grafo solo muestra las dos probabilidades de transición más altas por cada estado.

Hacemos ésta propuesta de visualización para obtener una mejor percepción de como los navegantes consultan los grupos. Empíricamente hemos observado que mostrar dos estados por grupo de páginas son suficientes para tener una visión clara del comportamiento de los navegantes.

Ésta propuesta permite enfocarnos en las interaciones que tienen los navegantes con los grupos y tener una imagen mental del orden y preferencias de los usuarios al consultar la información.

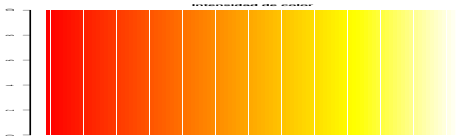


Figura 2-6: Derecha (blanco) baja intensidad, izquierda (rojo) alta intensidad

2.2. Cadenas Ocultas de Markov

Con las cadenas de Markov existe solamente un solo modelo generador de los datos el cual captura el comportamiento general de los navegantes. Con cadenas ocultas de Markov se extiende ésta idea agrupando a los usuarios de acuerdo a sus preferencias de navegación.

Éste método utiliza las mismas asunciones en cuanto a markovianidad (ver ecuación 2.1) generalizando el modelo usando varios subsistemas de cadenas de Markov. Cada subsistema $S_i(t)$ es específico del i-ésimo tipo de usuario y explica el comportamiento de un grupo de usuarios con hábitos de navegación similares donde cada navegante pertenece a un grupo de usuarios solamente.

La extensión la hacemos agregando una capa oculta donde suponemos k diferentes tipos de usuarios y la pertenencia del i-ésimo navegante a un grupo está dado por una variable indicadora. Es a través del uso de ésta variable que extendemos el modelo a más tipos de usuarios con la desventaja que no podemos hacer mediciones sobre ella.

En la figura 2-9 se muestra de forma esquemática el modelo generativo de los datos usando cadenas ocultas de Markov.

2.2.1. Estimación del modelo

La estimación del modelo de cadenas ocultas busca maximizar la probabilidad de observar los datos D dado en modelo M; es decir

$$arg\ max\ _M\ P(D \mid M) \tag{2.2}$$

Un problema donde el conjunto de datos D contiene dos tipos de elementos: los datos observados $E = \{e_i^t\}$ y los datos ocultos $U = \left\{U_i^k\right\}$ donde e_i^t es el estado observado en el tiempo t producido por la sesión i y $U_i^k = \{0, 1\}$ una variable indicadora que toma valor uno cuando la i-ésima sesión pertenece al k-ésimo submodelo y cero de cualquier otra forma.

La ecuación 2.2 no se puede maximizar directamente por lo que recurrimos al uso del algoritmo **E-M** (*Expectation-Maximization*) [10], [11], [12].

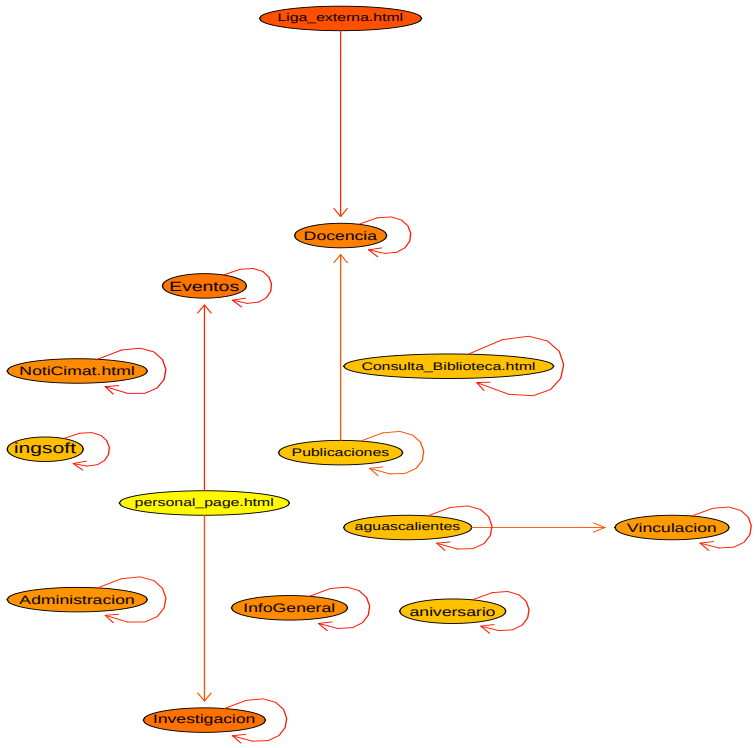


Figura 2-7: Matriz de transición A y vector Π en su representación por un grafo

Algoritmo E-M

Éste algoritmo es una herramienta popular para resolver problemas de máxima verosimilitud complicados como el planteado por la ecuación 2.2. Supongamos que la función de verosimilitud a maximizar es la siguiente:

$$arg\ max_{M,U} \quad \ell(M, E, U) = \sum_{j=1}^{k} \sum_{i=1}^{N} U_i^k \log \left[\phi_j(\{e_i^t\})\right] \tag{2.3}$$

donde $\phi_j(\{e_i^t\})$ es la probabilidad de generar toda la sesión i por el j-ésimo subsistema $\phi_j(.)$.

Al no contar con los valores de U_i^k haremos un asignamiento suave de cada observación a cada submodelo usando su valor esperado (ecuación 2.4).

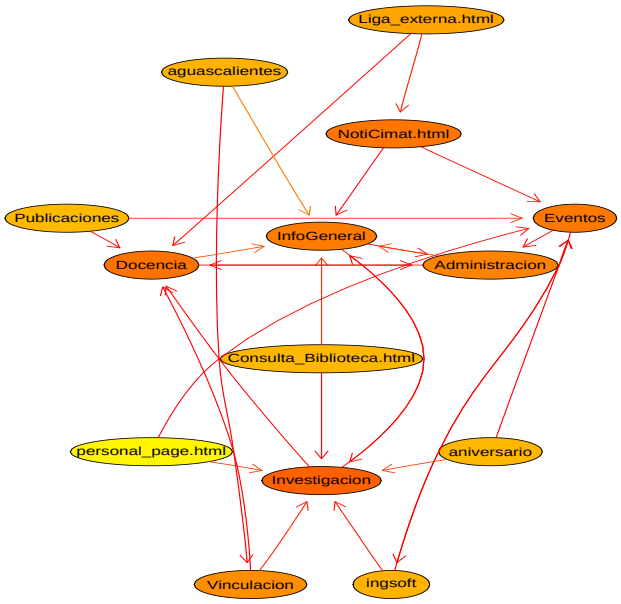


Figura 2-8: Matriz de transición A y vector Π en su representación por un grafo al condicionar a los navegantes a salir del grupo

$$\gamma_i^k = E\left[U_i^k \mid M^k, \{e_i^t\}\right] = P(U_i^k = 1 \mid M^k, \{e_i^t\}) \tag{2.4}$$

donde M^k son los parámetros del k-ésimo subsistema y $\{e_i^t\}$ toda la sesión i.

El algoritmo E-M es un procedimiento en dos partes:

- Expectación: Calculamos la probabilidad de cada sesión de pertenecer al k-ésimo submodelo M^k (γ_i^k) condicionando en los parámetros de cada uno de los k submodelos.
- Maximización: Estimamos los parámetros de M^k de cada submodelo condicionando en el valor esperado de U_i^k.

Éste esquema puede interpretarse también como un esquema de maximización conjunta en dos pasos donde en el paso de Expectación maximizamos el valor esperado de U_i^k fijando los submodelos y en el paso de Maximización maximizamos la verosimilitud de los parámetros de M^k dado U. El proceso de maximización conjunta es'ta representado en la figura 2-10 donde el eje horizontal es el espacio de los parámetros latentes del sistema (Expectación) y el eje horizontal los parámetros del modelo (Maximización), las curvas en la figura muestran curvas de contorno de la ecuación 2.2 y el avance hacia el óptimo en la función con cada paso del algoritmo.

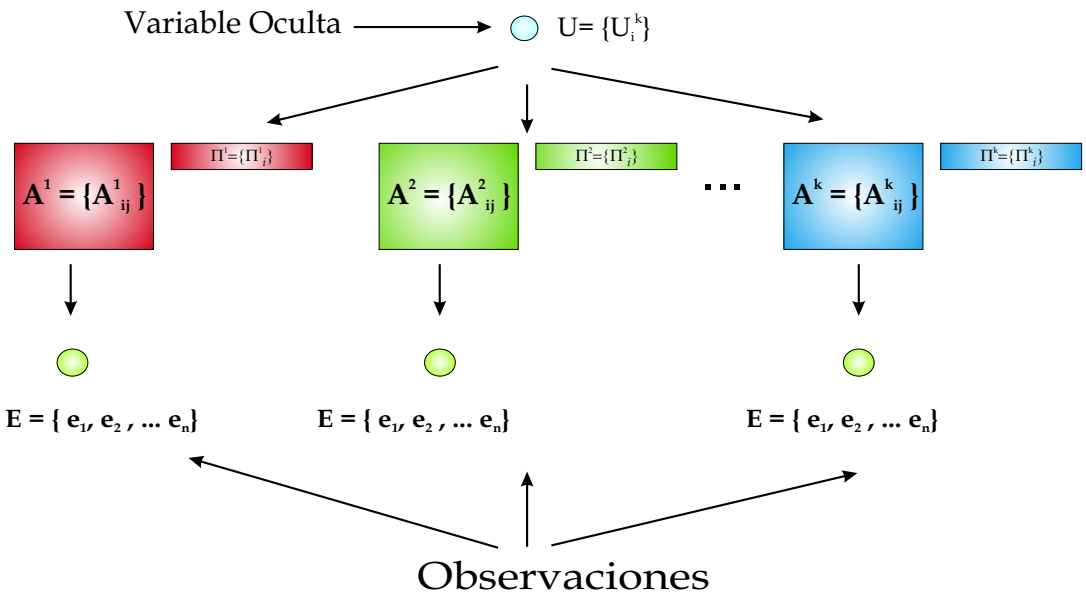


Figura 2-9: Modelo de Generativo de Cadenas Oculta de Markov

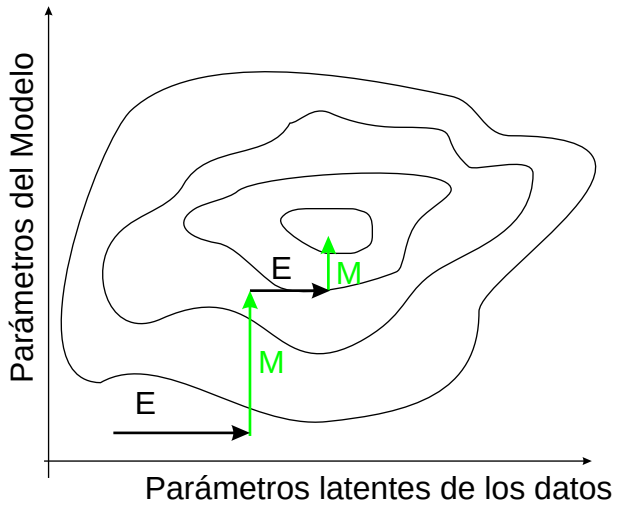


Figura 2-10: Pasos del Algoritmo E-M

Es necesario aclarar que éste algoritmo es muy suceptible a estancarse en óptimos locales debido a las multiples de soluciones que tiene el sistema.

La manera en que implementamos éste modelos utiliza la siguiente notación:

- **Notacion**

n: número de grupos de páginas.

N: número de sesiones.

T_i: longitud de la i-ésima sesion ; $i = \{1, \dots N\}$.

E: $= \{e_i^t\}]$ estados (grupos de páginas) que la sesión i visita en el tiempo t, $t = \{0, \dots T_i - 1\}$ y $1 \leq e_i^t \leq n$.

k: número de submodelos o tipos de usuario .

γ_i^k : $\in [0, 1]$probabilidad de la i-ésima sesión de ser generada por el k-ésimo submodelo.

M^k: $= \left\{A^k, \Pi^k\right\}$ los parámetros del k-ésimo submodelo. Donde $A^k = \left\{A_{i,j}^k\right\}$ y $A_{i,j}^k = P_{E_t}(e_j \mid E_{t-1} = e_i)$ es la probabilidad de estar en el estado j dado que en el tiempo anterior se visitó el estado i; $\Pi^k = \left\{\Pi_i^k\right\}$ y $\Pi_i^k = P_{E_{t=0}}(e_i)$ es la probabilidad de estar en el estado i en el tiemp $t = 0$.

Es de observar las siguientes propiedades :

1. $\sum_{j=1}^{n} A_{i,j}^q = 1 \quad q = 1, \dots, k.$
2. $\sum_{i=1}^{n} \Pi_i^q = 1 \quad q = 1, \dots, k.$
3. $\sum_{q=1}^{k} \gamma_i^q = 1 \quad i = 1, \dots, N \quad q = 1, \dots, k.$

A través del algoritmo E-M es posible estimar los k submodelos de usuario. A diferencia de [13] el agrupamiento de páginas en grupos temáticos evita ejercer un *pre-clustering* sobre las páginas al no hacer uso de una capa oculta extra porque desconocemos la asignación de cada página a un grupo temático. El inconveniente de tener un modelo con más capas ocultas recide en que el pre-clustering de páginas puede resultar incorrecto ya que estima la pertenencia de las páginas a un grupo a través de cómo son visitadas y no por su contenido.

Usando el conocimiento *a priori* de la asignación de páginas a cada grupo, el cálculo del modelo se ve tremendamente simplificado en comparación a [13].

La estimación del modelo se hace a través del algoritmo E-M con las siguientes ecuaciones:

Paso Expectación:

$$\gamma_i^k = P(U_i^k = 1 \mid M^k, \left\{e_i^t\right\}) = \frac{\alpha_k P(e_i^t \mid M^k)}{\sum_x \alpha_x P(\{e_i^t\} \mid M^x)} \tag{2.5}$$

donde :

$\alpha_k = \frac{1}{N}\sum_{i=1}^{N}\gamma_i^k$, es la probabilidad de pertenecer al k-ésimo tipo de usuario y $\{e_i^t\}$ es la i-ésima sesión

Paso de Maximización para cada uno de los k submodelos:

$$A_{i,j}^k = \frac{\sum_{m=1}^{N}\gamma_m^k \ \sum_{t=1}^{T_l-1}\xi_{m,k}^t(i,j)}{\sum_{m=1}^{N}\gamma_m^k \ \ \sum_{t=1}^{T_l-1}\nu_{m,k}^t(i)} \tag{2.6}$$

$$\Pi_i^k = \frac{\sum_{m=1}^{N}\gamma_m^k \ \nu_{m,k}^{t=0}(i)}{\sum_{m=1}^{N}\gamma_m^k \sum_{z=1}^{n}\nu_{m,k}^{t=0}(z)} \tag{2.7}$$

Donde:

$$\nu_{m,k}^t(i) = \begin{cases} 1 & \text{si } e_i^t = i \\ 0 & \text{en cualquier otro caso} \end{cases}$$

$$\xi_{m,k}^t(i,j) = \begin{cases} 1 & \text{si } e_m^t = i \ y \ e_m^{t-1} = j \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Podemos ver que cuando el número de tipos de usuarios $k = 1$, el sistema se reduce al caso de la cadenas de Markov donde la estimación se hizo a través del conteo de frecuencias relativas.

Cabe mencionar que para éste modelo es necesario proponer el número de tipos de usuarios que queremos obtener, llevándonos al problema de determinar el número de los k tipos de usuarios óptimo.

Una manera muy sencilla de buscar el número 'real' de los k tipos de usuarios es usar el método de *crossvalidation*: resguardamos una porción de los datos y calculamos el modelo con los datos restantes hasta convergencia; una vez hecho esto medimos que tan bien se ajustan los datos resguardados al modelo previamente calculado y ésto da una idea de que bien fue la elección del valor de k.

Un ejemplo sintético es el que se presenta en la figura 2-11. Para cada valor de k se hicieron 30 corridas y se obtuvo el promedio de la logverosimilitud [22] para los dos tercios con los que se entrenó el algoritmo mostrado en negro, para el faltante en rojo.[1].

Éste ejemplo sintético fue generado de un modelo bastante sencillo con 5 tipos de usuario. En la figura 2-11 vemos que en el punto donde $k = 5$ la línea roja es máxima lo cual es el resultado que esperábamos de

[1]A la figura 2-11 se le agregó un offset a la línea roja puesto que esta en otra escala, con fines de una mejor comparación

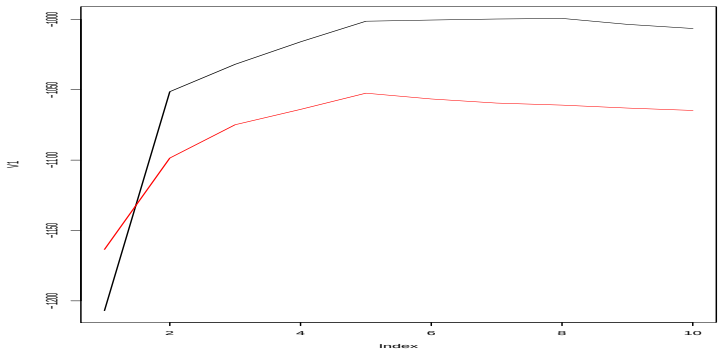


Figura 2-11: Caso sintético de k usuarios donde $1 \leq k \leq 10$. Número óptimo de usuarios es $k = 5$ en negro Logverosimilitud de los datos de entrenamiento en rojo logverosimilitud de datos restantes

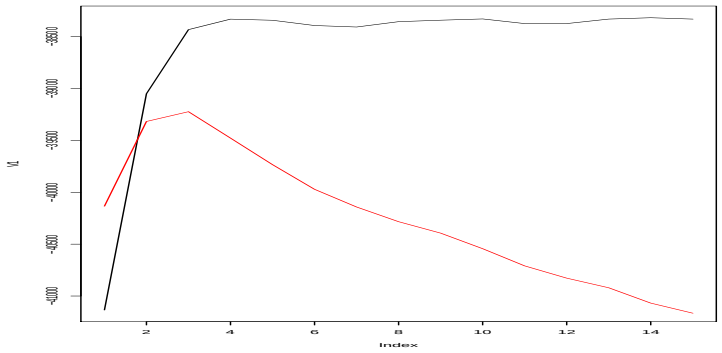


Figura 2-12: Caso real de k usuarios donde $1 \leq k \leq 10$ y número óptimo de usuarios es $k = 3$ en negro Logverosimilitud de los datos de entrenamiento en rojo logverosimilitud de datos restantes

experimento, validando nuestra propuesta.

En la figura 2-12 los resultados [2] hechos para los datos reales, los cuales nos dejan ver el número de usuarios k más probable de acuerdo a la logverosimilitud es 3.

Dado que el número de tipos de usuario encontrados hasta ahora es relativamente pequeño, podemos seguir usando para nuestras comparaciones este número de usuarios.

La generalización de las cadenas de Markov permite segmentar en tipos de usuario el conjunto de navegantes además de modelar a cada uno por sus hábitos de navegación, pero complica significativamente el método de estimación del modelo y aumentando el número de parámetros a estimar.

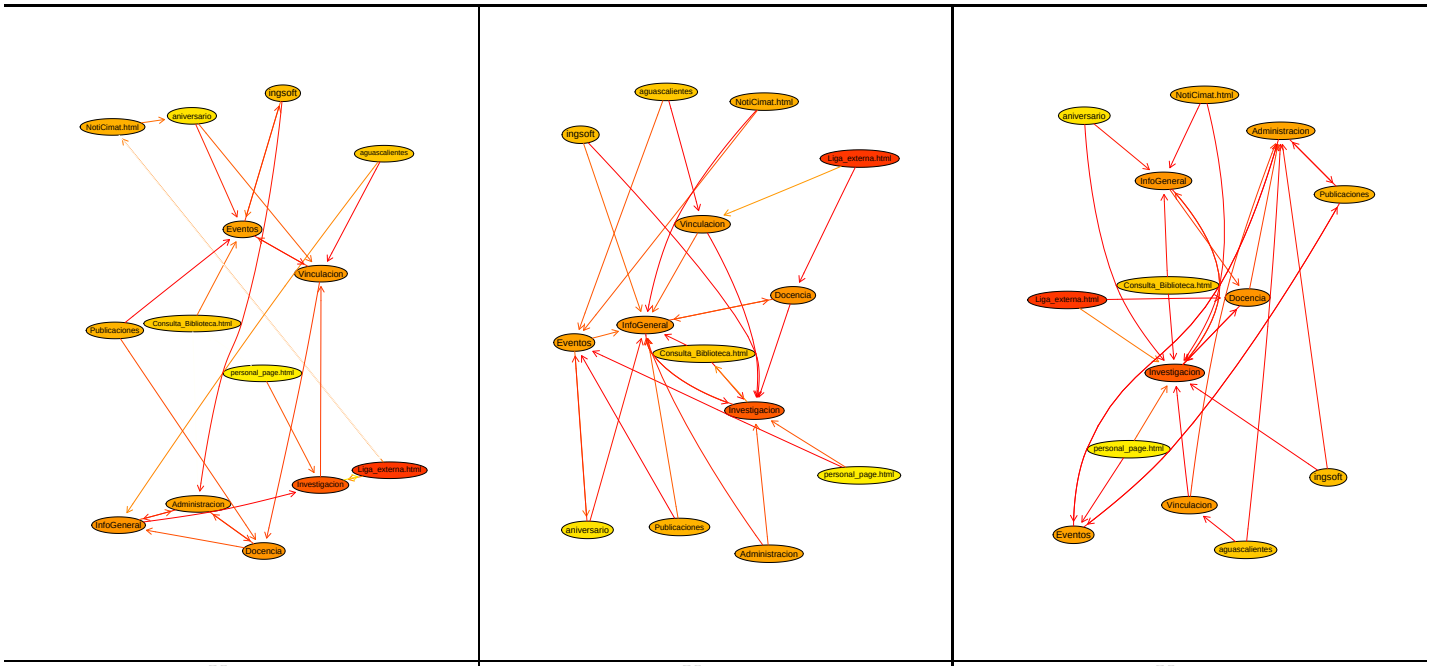
Los resultados obtenidos por éste algoritmo pueden variar al quedar estancado en óptimo local; sin embargo dentro de los experimentos efectuados los resultados son muy similares, inclusive al usar archivos de otras fechas, en todas las corridas hechas al fijar el parámetro k.

[2] A la figura 2-12 se le agregó un offset a la línea roja puesto que esta en otra escala, con fines de una mejor comparación

Una característica que obtenemos en nuestros experimentos es la aparición de un tipo de usuario con la una proporción de navegantes considerable atribuidos a éste modelo de navegación.

2.2.2. Visualización de Modelos de Cadenas Ocultas

Hemos obtenido una propuesta de visualización con las matrices de transición de las cadenas de Markov, la cual extrapolaremos al caso de las cadenas ocultas como se muestra en la figura 2-13. En la figura mostramos las representaciones de las matrices encontradas por el algoritmo E-M para el caso cuando el número de tipos de usuarios $k = 3$.



$U = 1$ | $U = 2$ | $U = 3$

Figura 2-13: Representación de matrices de transición de grupos de usuarios con el parámetro $k = 3$

Con ésta representación podemos hacer comparaciones más intuitivas entre tipos de usuarios, entre las cuales se destacan los grupos Investigación, Eventos, Información General como grupos de alta probabilidad de vistarlos debido al número de aristas de llegada.

También existe una similitud en cuanto a los primeros grupos en vistar como Docencia e Investigación.

2.3. Modelo de contraste

De los tipos de usuarios obtenidos en la sección anterior queremos obtener un contraste mayor entre ellos y así apreciar mejor las diferencias en su comportamiento. Para ello introducimos en ésta sección un usuario de contraste.

Éste tipo de usuario de contraste nos permitirá acentuar mejor diferencias entre los tipos de usuarios al agrupar sesiones cuyas preferencias de navegación conozcamos.

Los contrastes que proponemos son:

- Contraste A: El usuario no presenta preferencias por algún grupo temático; su interés es uniforme sobre cada grupo de páginas: $\Pi_i = \frac{1}{Número\ de\ grupos}$; $A_{i,j} = \frac{1}{Número\ de\ grupos}$
- Contraste B: El usuario presenta interés uniforme sobre todo el contenido del sitio; su interés es proporcional al tamaño de cada grupo temático. $\Pi_i = \frac{Número\ de\ Páginas\ en\ Grupo\ i}{Páginas\ totales}$; $A_{i,j} = \frac{Número\ de\ Páginas\ en\ Grupo\ j}{Páginas\ totales}$

La introducción del usuario de contraste nos dará información acerca de los usuarios que no se ajustan al modelo de caminatas aleatorias contrastándolos con los navegantes que visitan el sitio con preferencias distintas.

Los resultados obtenidos con estos usuarios de contraste con $k = 3$ donde el primer usuario es de contraste son :

Cuadro 2.1: Estadísticas con Grupo de contraste con $k = 3$

	Contraste A	Contraste B
Porcentaje de navegantes segmentados como de contraste U=1	0.97 %	2.74 %
Porcentaje de navegantes segmentados como Tipo de usuario regular U=2	25.01 %	24.76 %
Porcentaje de navegantes segmentados como Tipo de usuario regular U=3	74.00 %	72.48 %

La proporción en que aparecen los tipos de contraste es baja indicándonos la poca probabilidad de modelar usuarios como caminatas aleatoreas sobre el sitio. Y no hay una diferencia significativa entre los resultados obtenidos con los diferentes contrastes encontrando una semejanza en la asignación de usuarios a cada navegante de 96,90 %

La figura 2-14 nos muestra como se distribuyen las caracterísitcas de la sesiones (excepto (5) y (9)) por cada tipo de usuaro a través de boxplots. Cada color indica el grupo al que pertencen y a la izquierda los usuarios con tipo de contraste A y a la derecha con tipo de contraste B usando el color negro las características de los usuarios de contraste.

Con ésta figura podemos apreciar un comportamiento diferente por parte de los usuarios de contraste destácandose las variables (2), (3) y (6).

Siendo la característica más importante de los usuarios de contraste es que visitan en promedio más grupos temáticos pero con menos visitas por grupo aunque duran más tiempo.

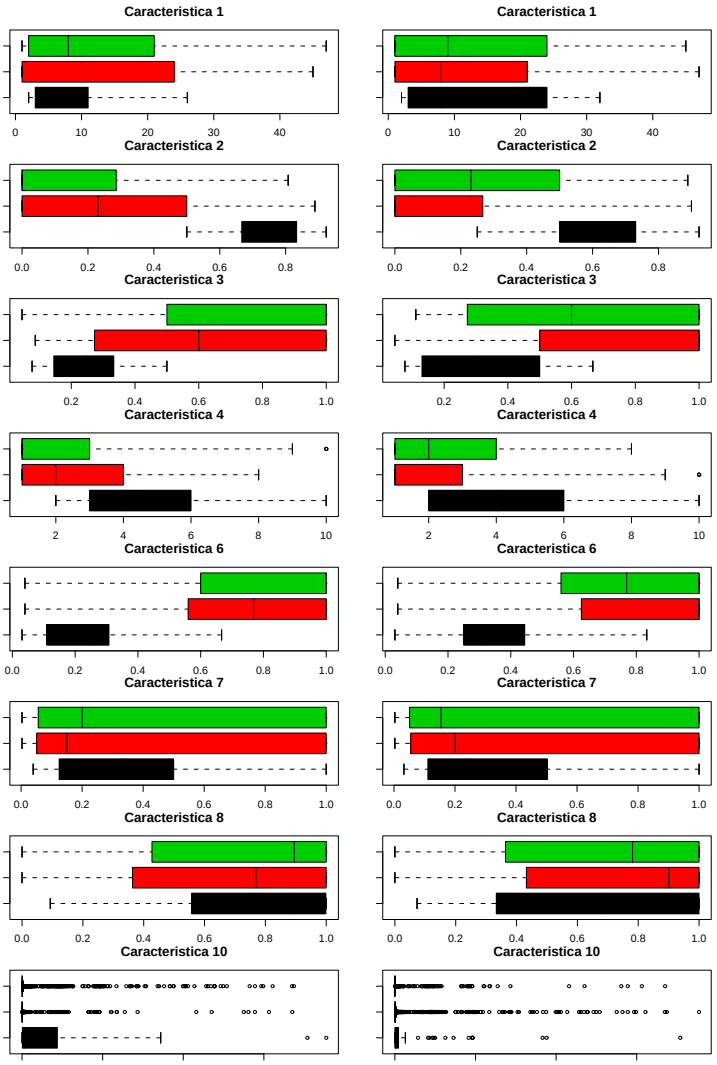


Figura 2-14: Variables por tipos de usuario de contraste

A continuación presentamos una comparación visual de las relaciones encontradas en los 3 escenarios propuestos con los HMC, usando las estimaciones de los k tipos de usuario e incluyendo los modelos de usuario de contraste tipo A o B.

En la figura 2-15 se muestran los 3 escenarios de los HMC con $k = 3$ aplicando la misma restricción de mostrar los 2 grupos más probables condicionando a salir del presente.

El primer renglón es la misma representación de los HMC mostrada en la figura 2-5, el segundo renglón es utilizando el contraste A y el tercer renglón usando el contraste B.

En el caso de las corridas con contraste el grafo en la columna de la izquierda corresponde al usuario de contraste.

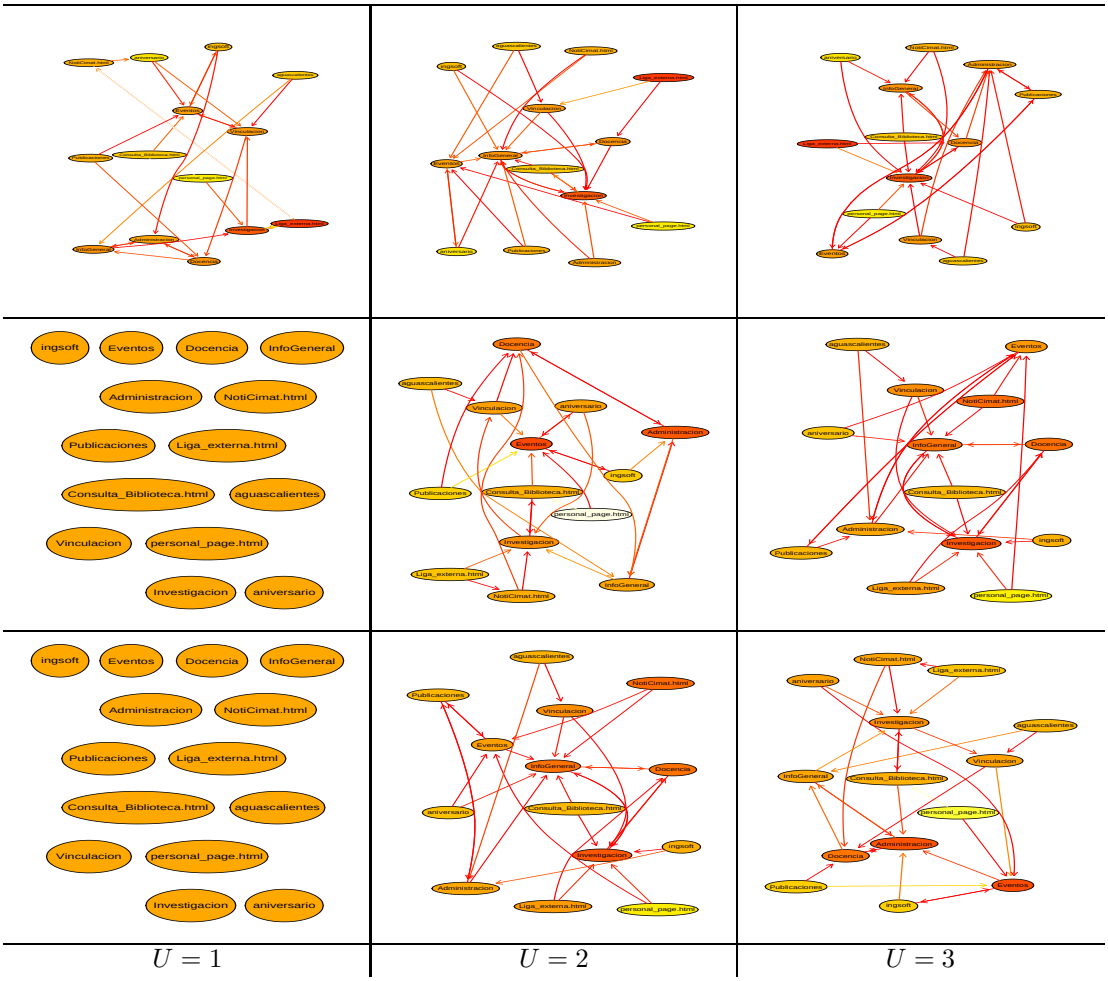


Figura 2-15: HMC normal (primer renglón), Usuario de Contraste A (segundo renglón), Usuario de contraste B (tercer renglón)

La inclusión de un modelo de contraste no evita obtener relaciones entre grupos semejantes a las obtenidas con el modelo simple. Existen relaciones que a pesar del modelo de contraste permanecen en nuestros resultados como el número de aristas de llegada a grupos como Información General e Investigación.

Éstos resultados nos dan mucho más confianza en las relaciones que hemos encontrado entre los usuarios regulares, además de darnos información acerca de los usuarios que divagan más en el sitio como los modelados con los contrastes propuestos.

Comparativo con Análisis de Componentes

Por medio de los HMC hemos obtenido una parametrización de dónde navegan los usuarios que visitan el website, qué grupos de páginas están más fuertemente relacionados y una segmentación de navegantes en clusters de usuario.

Una posible manera de confirmar estas tendencias y segmentación es a través del uso de componentes principales (PCA *Principal Component Analysis*) e independientes (ICA *Independent Component Analysis*) suponiendo que el *feature space* pueda reflejar el comportamiento de usuarios y capturar los rasgos que el HMC ha determinado.

Retomando el espacio de de características de la tabla 1.6 queremos ver si existe alguna relación o semejanza entre el espacio generado por estas características y los resultados obtenidos por los *HMC*. Para el análisis presentado se excluyeron las variables catergóricas (5) y (9).

Usando ambos tipos de transformaciones (*PCA* e *ICA*) sobre el espacio de características queremos comparar la segmentación de navegantes hecha por el HMC sin usar contraste con las proyecciones que estas técnicas nos ofrecen. En el análisis multidimencional que habiamos hecho encontramos una característica importante para distinguir las cadenas y esto es su longitud; encontramos cadenas largas (≥ 15) y cortas (< 15).

En la figura 2-16 mostramos solamente 3 componentes las cuales son suficientes para explicar el 87.51 % de la varianza de los datos en téminos de PCA; vemos la comparación de como éste atributo se comporta sobre los datos transformados.

Arriba de la diagonal tenemos las proyecciones sobre los componentes principales y debajo los independientes, sobre la diagonal un boxplot de cada una de las respectivas proyecciones. Las observaciones están coloreadas (negro,rojo,verde) de acuerdo al tipo de usuario que los HMC obtuvieron y en la parte inferior de la figura mostramos las densidades de cada componente por tipo de usuario, en el primer renglón los componentes de PCA y en el segundo de ICA.

En la tabla 2.2 se encuentra el porcentaje de navegantes de acuerdo a su tipo de usuario:

Cuadro 2.2: Distribución de navegantes en tipo de usuario para cadenas largas y cortas

	$U = 1$	$U = 2$	$U = 3$	Total
Navegantes en el Sitio	39.18 %	21.55 %	39.26 %	100.0 %
Navegantes con Cadenas Cortas	55.43 %	55.08 %	72.28 %	61.98 %
Navegantes con Cadenas Largas	44.56 %	44.87 %	27.68 %	38.01 %

Algunas conclusiones que podemos obtener del análisis son que el espacio de características utilizadas no captura la interacción entre grupos de páginas como lo hacen los HCM dado que en los tipo de usuario existen navegantes con cadenas largas como con cadenas cortas. Era de esperarse que una segmentación basada en una sola característica por muy relevante que fuera no reflejaría el tipo de interacciones que los HMC analizan.

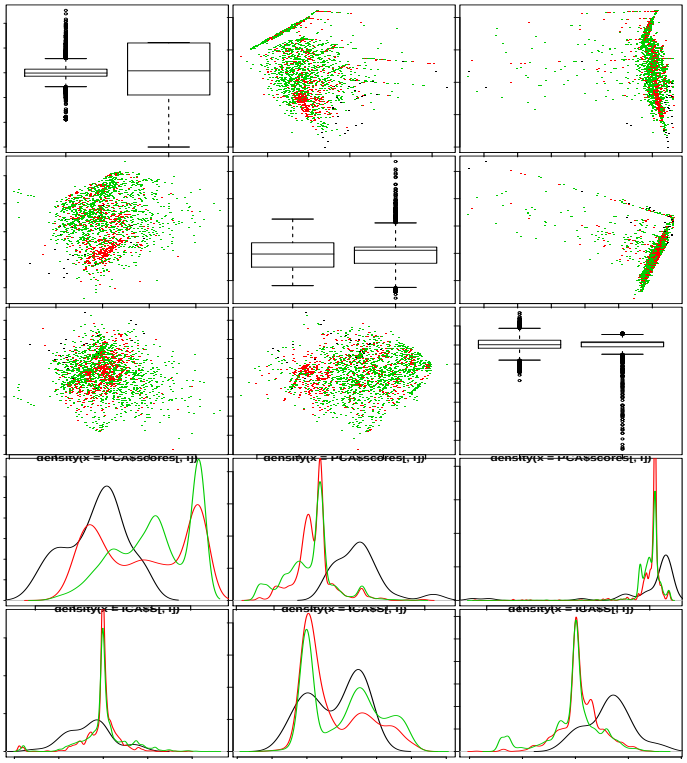


Figura 2-16: Comparativo de HMC sobre PCA e ICA, en colores los tipos de usuario y en los renglones las densidades por cada tipo de usuario.(PCA arriba, ICA abajo)

Un remedio para ésta situación es el utilizar el espacio de características extendidas (ver capítulo 1) en el cuál de una manera más detallada analizamos las caracterísitcas de las sesiones. Este espacio tiene la característica que es demasiado complicado para poder análizarlo con técnias como ICA o PCA, es por eso que hacemos uso de ROB-PCA para ello.

Los resultados obtenidos se encuentran el la figura 2-17 donde 2 componentes son suficientes para explicar el 98.14 % de la varianza de los datos. Los datos se han coloreado de acuerdo a el tipo de usuario que los HMC obtuvo. Hemos dividido los datos por la clasificación que ROB-PCA ha encontrado en datos regulares (abajo de la diagonal) y atípicos(arriba de la diagonal) mostramos las densidad de la componente segmentada sobre cada tipo de usuario indicado por los colores (negro, rojo y verde). Sin importar el uso de técnicas robustas los resultados obtenidos a través de HMC y componentes principales difiere sustancialmente aún usando un espacio de atributos más complicado.

Ésto sucede debido a que los dos enfoques empleados analizan distintas aspectos de las cadenas y el empalme entre ambas perspectivas es poco.

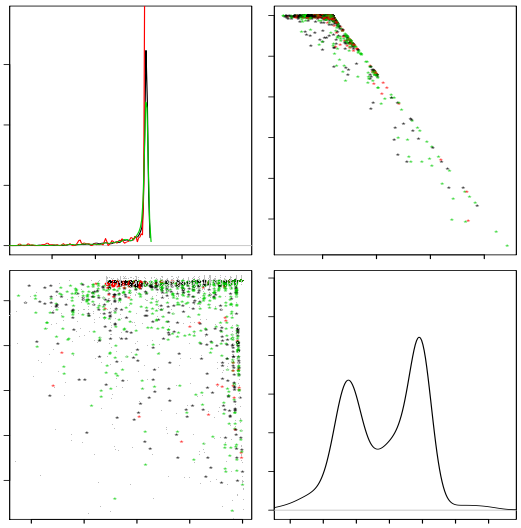


Figura 2-17: Comparativo de HMC sobre Rob-PCA, en colores los tipos de usuario y en los renglones las densidades por cada tipo de usuario.(PCA arriba, ICA abajo)

2.4. Estados Absorbentes

Por definición las matrices de transición encontradas por los HMC poseen lo que se llaman estados absorbentes [21], estados con una probabilidad de transición igual a uno; es decir el estado i es absorbente si, y solo si, $A_{i,i} = 1$. El término absorción se debe a que una vez que la cadena transita por éste estado no puede salir de él.

Las matrices de transición obtenidas en el secciones anteriores no cumplen estríctamente con la definición de estado absorbentes, pero en nuestro estudio hemos flexibilizado la definición considerando como absorbentes aquellos estados que sean mayores a un umbral Θ_{th} (típicamente 0.9) ya que el 10 % restante de la masa de probabilidad tendrá que estar distribuida entre los 13 grupos restantes y es por eso que lo consideramos como absorbente.

De la definición de estado absorbente se desprenden propiedades interesantes de las matrices de trancisión contestan las siguientes preguntas:

1. ¿Cuáles estados son absorbente y cuáles son transitivos?
2. ¿Cuál es el tiempo promedio en que una cadena es absorbida?
3. ¿Cuál es el tiempo promedio en que una cadena se encuentra en un estado transitivo antes de ser absorbida?

Donde las propiedades relacionadas con el tiempo dependen del estado inicial de la cadena (Estado en t_0), una forma de mostrar éstas propiedades se encuentra en la figura 2-18 donde a cada estado i de una

matríz de transición se le asocia la siguiente representación:

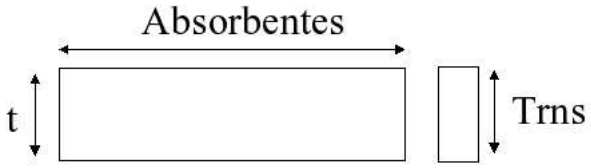


Figura 2-18: Visualzación de estados absorbentes

Cada estado absorbente m divide a el rectángulo principal horizontalmente en partes proporcionales indicando la probabilidad de ser absorbido por el estado m dado que la cadena inició en el estado i. La altura del rectángulo principal indica el tiempo esperado en que la cadena deambulará por los estados transitorios antes de presentarse la absorción.

El cuadro a la derecha representa el tiempo promedio antes de la absorción de la cadena y esta dividido proporcionalmente en los estados transitorios, mostrando el tiempo esperado que la cadena visite dichos estados antes de ser absorbida.

Un ejemplo se muestra en la figura 2-19 en la cual se tomaron las matrices de transición de tipos de usuario de los HMC sin considerar usuarios de contraste y con el parámetro $k = 3$.

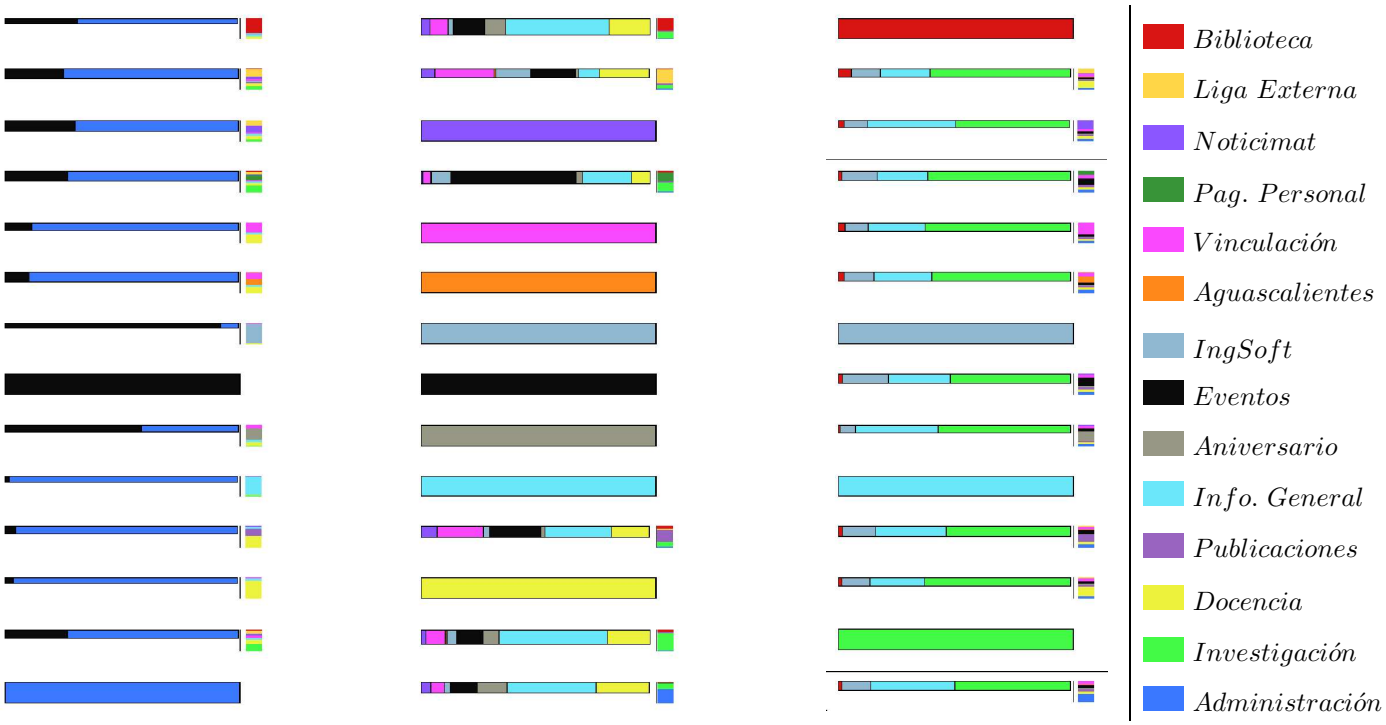


Figura 2-19: Estados absorbentes para 3 tipos de clusters de usuarios (izquierda) tabla de nombres de los grupos de páginas (derecha)

Dentro de las conclusiones que obtuvimos de éste análisis una reincidencia de grupos de páginas detectados como absorbentes por parte de varios tipos de usuario.

Al hacer cadenas más homogeneas a través del filtro de entrada vemos que las sesiones al principio de la cadena tienen mucho impacto en la información que los navegantes consultan. A través del enfoque de los estados absorbentes observamos como los intereses de los usuarios cambian dependiendo del grupo temático en el que se encuentran.

Los navegantes tienen una influencia en las visitas futuras que realizarán dependiendo del grupo temático en que se encuentren.

2.5. Experimentos Extra

Como un último experimento queremos obtener información acerca de como los usuarios interactuan con la información presentada y ? cuál es es contenido que es más interesante para ellos?. A diferencia de las secciones anteriores el contenido de site no será considerado por secciones sino por páginas.

En éste experimento pesaremos a cada entrade en el log file de acuerdo a dónde está ubicada en la sesión siendo las páginas con mayor peso las que se encuentran al final de la misma y promediando el peso obtenido para cada página.

A través de ésto queremos encontrar cuál es el contenido que los usuarios consultan al final. Éste resultado puede interpretarse como información en que los usuarios están interesados pero tardan toda la sesión para encontrarla.

Las páginas que obtuvieron un promedio mayor de 0.8 en una escala de 0.0 a 1.0 fueron:

- `famat/cursos/computacion/tsc_arquitectura_de_computadoras.html`
- `pag_externa.php?ext=tra_in03_1103.html`
- `pag_interna.php?id=122`
- `Eventos/oebn-conference/top.html`
- `pag_externa.php?ext=tra_p_apoy.html`
- `pag_externa.php?ext=tra_in03_0804.html`
- `pag_externa.php?ext=tra_08_0202.html`
- `famat/cursos/ecuaciones_diferenciales/cursos_ecuaciones_diferenciales_ordinarias.html`
- `Eventos/oebn-conference`

- `pag_externa.php?ext=99_grobner.html`
- `famat/cursos/computacion/cursos_programacion_no_lineal.htm`
- `pag_interna.php?id=118`

Éstas páginas coresponden a páginas que pertenecen a Inscripciones de Eventos , Proyectos de Vinculación y principalmente a información de cursos sobre la Facultad de Matemáticas (FAMAT). Sobre éstos últimos creemos que debería de hacerse una reubicación puesto que es contenido "escondido.º de dificil acceso y los navegantes parecen tener un interés particular en ellos.

Capítulo 3

Modelos de Interés

Los modelos generativos utilizados en el capítulo anterior analizan el comportamiento de los navegantes a un nivel de secuencias, el cuál es muchas veces demasiado fino para llegar a resultados fácilmente interpretables y/o hacen supuestos que dificilmente se cumplen.

En éste capítulo queremos proponer algunas alternativas que hacen un comporomiso entre supuestos y detalle.

Uno de los supuestos que hacemos a lo largo de éste capítulo es suponer que la forma en que está acomodada la información impacta directamente en cómo se consulta. Al contener la pa'gina principal todos los temas relevantes del sitio provee al navegante de la información necesaria para saber que temas son de su interés.

El limitarnos a sesiones que comenzaron en ésta página fortalece el supuesto de considerar como más útil y/o interesante las información consultada en las primeras entradas de la sesión.

Presentamos 4 propuestas las cuales miden diferentes características de interés para poder analizar el comportamiento de los navegantes. Con éstas propuestas queremos obtener una representación menos fina de los datos conservando la utilidad que nos presenta la informacón comparada con los modelos anteriores.

3.1. Medidas de interés por grupo

De la misma manera como mapeamos cada sesión a un espacio de características (ver capítulo 1) buscamos un espacio que capture de cada sesión el interés del navegante por cada uno de los grupos temáticos del sitio.

Una propuesta para hacer esto es a través de 2 medidas que obtendremos de las sesiones: **rapidez** con la que se visita un grupo una vez se ingresa al sitio y **revisitación** al grupo.

- Rapidez: Lo medimos a través de la ubicación de la primera vista a un grupo en la sesión, siendo una combinación lineal de cuántos grupos se visitaron ántes y después de dicha visita, denotamos con g_i^k

al valor de interés por el k-ésimo grupo en la i-ésima sesión generado por la primera visita al grupo k.

$$g_i^k = \begin{cases} d_{ik} - a_{ik} = L_i - 2 * p_k^i & \\ 0 & \text{Si no hay visita al grupo } k \text{ en la sesión } i \end{cases} \tag{3.1}$$

donde d_{ik} es el número de grupos visitados después de la primera aparición del k-ésimo grupo (incluido el presente), a_{ik} es el número de grupos visitados antes de la primera visita al k-ésimo grupo, L_i la longitud de la i-ésima sesión y p_i^k es la posición de la primera aparición del k-ésimo grupo en la i-ésima cadena.

- Revisitación: Lo medimos a través de cuántas veces el grupo k es revisitado. Denotamos al valor de h_i^k como la medida de revisitación al k-ésimo grupo por la i-ésima sesión.

$$h_i^k = \begin{cases} \text{número de revisitas al grupo } k \text{ en la sesión } i \end{cases} \tag{3.2}$$

Ambas medidas son normalizadas por sesión en el intervalo $[0, 1]$.

Las ecuaciones 3.1 y 3.2 miden de una forma 'pura' las características de rapidez y revisitación respectivamente. Supongamos que el interés por el k-ésimo grupo T_k es un reflejo de éstas dos medidas. Por eso definimos el siguiente funcional que minimiza la distancia entre los valores de g_i^k y h_i^k para T_k.

$$arg\ min\ _{T_q}\ F(T_q) = \sum_{i=1}^{\#sesiones} \sum_{q=1}^{\#grupos} \frac{(T_q - g_i^q)^2}{\#sesiones} + \lambda \sum_{i=1}^{\#sesiones} \sum_{q=1}^{\#grupos} \frac{(T_q - h_i^q)^2}{\#sesiones} \quad \forall\ q = 1, \ldots, k \tag{3.3}$$

donde el parámetro λ pesa el término de revisitación de grupos, el valor de T_q indica el interés en el k-ésimo grupo y se encuentra en el intervalo $[0, 1]$ siendo 1 el de mayor interés.

En la figura 3-1 ilustramos el problema a resolver donde los valores de g_i^k y h_i^k están representados por los puntos verdes y rojos respectivamente y T_k por los puntos azules.

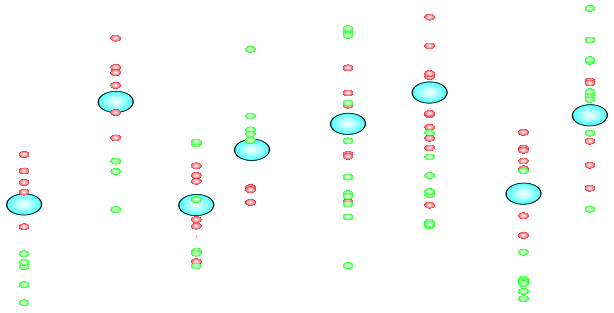


Figura 3-1: Buscamos la menor distancia de T_k (azules) a las mediciones g_i^k 'y h_i^k (verde y roja) para cada grupo

Una manera de percibir la sensibilidad de los resultados al parámetro λ es la siguiente tabla en la cual se comparan dos corridas con $\lambda = 0$ y $\lambda = 1000$.

Cuadro 3.1: Comparación de corridas modelo

$\lambda = 0$	$\lambda = 1000$
T[k]	T[k]
0.304 Investigación	0.151 Docencia
0.269 Docencia	0.147 Investigación
0.235 Eventos	0.139 InfoGeneral
0.222 NotiCimat	0.116 Administración
0.217 InfoGeneral	0.097 Eventos
0.207 Administración	0.086 NotiCimat
0.177 Vinculación	0.058 Vinculación
0.132 Liga Externa	0.028 Ingsoft
0.130 Ingsoft	0.022 Aniversario
0.129 Aguascalientes	0.018 Consulta Biblioteca
0.124 Consulta Biblioteca	0.015 Aguascalientes
0.123 Publicaciones	0.015 Publicaciones
0.122 Aniversario	0.001 Página personal
0.101 Página personal	0.000 Liga externa

Los resultados de la ecuación 3.3 cambian con respecto al parámetro λ que pesa la importancia de cada una de las medidas. El uso de éste parámetro nos proporciona todo el espectro de posibles soluciones compromiso entre la rapidez y revisitación.

En la figura 3-2 se muestran los valores de T_k en rojo para el caso de $\lambda = 1000$ y en verde $\lambda = 0$ los cuales podemos tomar como valores extremos es el espectro de soluciones en azul una posible solución con un valor intermedio de λ.

3.2. Medidas de interés conjunto

En ésta sección elaboramos una función que además de seguir midiendo la rapidez y revisitación al grupo considera que el interés del navegante en la información no cambia drásticamente a lo largo de la sesión. Asignando un valor de interés similar a grupos que se consultan cercanos en el tiempo.

La función de interés conjunto se divide en dos partes: Interés individual y relaciones entre grupos.

- Interés individual: Lo definimos como una medida que fusiona la rapidez y la revisitación en un solo valor $P = \{P_k\}$ donde P_k es el valor de interés individual del k-ésimo grupo y es proporcional al número

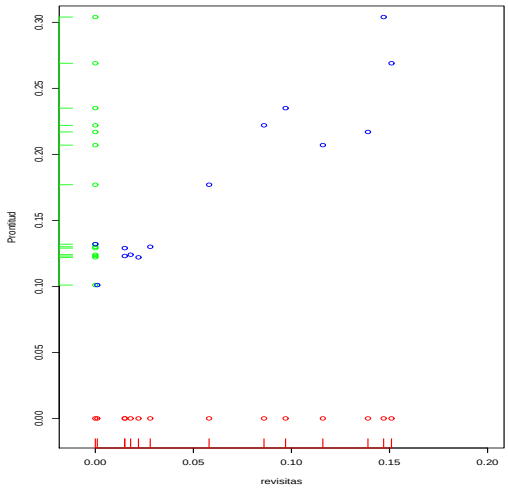


Figura 3-2: Soluciones compromiso entre dos medidas de interés. Casos cuando $\lambda = 1000$ y $\lambda = 0$ (rojo y verde respectivamente). Valores de T_k (azul) dando una posible solución intermedia.

de veces que un usuario entra y sale del grupo. En ésta medida consideramos como valores de mayor peso a las visitas al inicio de la sesión.

$$P_k = \frac{1}{\#sesiones} \quad 2 * q_k - r_k \tag{3.4}$$

donde

$$q_k = \sum_{i=1}^{\#sesiones} \sum_{t=0}^{L_i - 1} S_{it}^k \, \frac{L_i - t - 1}{L_i - 1}$$

$$r_k = \sum_{i=1}^{\#sesiones} \sum_{t=0}^{L_i - 1} E_{it}^k \, \frac{L_i - t - 1}{L_i - 1}$$

con $L_i =$ la longitud de la i-ésima cadena junto con las funciones indicadoras S_{it}^k y E_{it}^k donde

$$S_{it}^k = \begin{cases} 1 & \text{cuando en la sesión i al tiempo t se sale del grupo k} \\ 0 & \text{de otra forma} \end{cases}$$

$$E_{it}^k = \begin{cases} 1 & \text{Cuando en la sesión i al tiempo t se llega al grupo k} \\ 0 & \text{De otra forma} \end{cases}$$

- Relaciones entre grupos: Proporcional al número de trancisiones del grupo i al grupo j teniendo más

peso las visitas al principio de la sesión. Denotamos a K_{ij} como la relación entre dos grupos por lo que $K_{ij} = K_{ji}$.

$$K_{ij} = \sum_{l=1}^{\#sesiones} \sum_{t(i,j,l)} 0{,}99^t \tag{3.5}$$

donde L_i es la longitud de la cadena i y $t(i, j, l)$ en tiempo donde se producen las transiciones del grupo i al grupo j en la sesión l.

Con las medidas dadas por las ecuaciones 3.4 y 3.5 construiremos una función que tome en cuenta ambos aspectos, obteniendo un valor de interés conjunto $\Theta = \{\Theta_i\}$ que refleje estos elementos.

La figura 3-3 representa un sistema equivalente a la ecuación 3.6 donde la altura de cada cubo está dada por P_k reflejando la intensidad con la que es visitado el k-ésimo grupo, el interés de cada grupo dado por la altura de cada una de las esferas Θ_i conectadas entre si con un resorte de rigidez K_{ij}^N donde entre más rigido sea el resorte más dificil es deformarlo. Ésta rigidez tiene el efecto de posicionar grupos altamente relacionados con intereses similares.

Cada esfera de interés Θ_i además de estar conectada a todas las demás está sujeta con otro resorte con valor de rigidez K a la intensida de visitación dada por P_k. Colocando cada esfera de interés en un valor cercano a P_k.

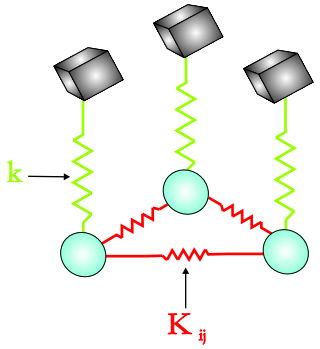


Figura 3-3: Modelo de resortes propuesto

$$arg\ min_{\ \Theta}\ F(\Theta) = K\sum_{i}^{n}(P_i - \Theta_i)^2 + \sum_{i=1}^{n}\ \sum_{\substack{j=1 \\ i \neq j}}^{n} K_{ij}^N(\Theta_i - \Theta_j)^2 \tag{3.6}$$

donde P_i es el valor de interés individual asociado a cada grupo de páginas de la ecuación 3.4 y ha sido normalizado de $[0, 100]$, n el número de grupos de páginas, Θ_i los valores de interés conjunto que buscamos para cada grupo de páginas, K la constante de rigidez de los resortes entre P_k y las esferas de interés, K_{ij}^N son las constantes de rigidez entre esferas de interés.

Para este caso K tiene un valor de 100 y K_{ij}^N es la normalización por por renglones de K_{ij} en el rango $[0, 100]$. El propósito de la normalización es para poder comparar los dos términos de la ecuación 3.6 sin problemas de escalas.

Los resultados son bastante consistentes para los experimentos que realizamos modificando los datos de entrada para diferentes fechas. Al obtener un interés conjunto el grupo correspondiente a ligas externas al CIMAT modifica los valores obtenidos, en el cuadro 3.2 mostramos los resultados al descartar la interacción de dicho grupo con los demás.

A través de éstos modelos hemos podido obtener un ranqueo de los grupos temáticos de acuerdo a su interés ya sea individual o conjunto.

Cuadro 3.2: Grupos de páginas

Con referencia externa

Θ_i Interés Conjunto	Grupo	P_i Interés Individual
74.707	Docencia	100.000
67.741	InfoGeneral	83.727
64.766	Administracion	61.351
62.797	Investigacion	65.254
61.347	NotiCimat	68.357
56.885	Eventos	47.326
55.877	Vinculación	29.542
50.025	Publicaciones	8.815
48.839	Liga Externa	1.992
47.560	Ingsoft	13.648
45.961	Aniversario	9.605
45.693	Aguascalientes	9.064
41.653	Consulta Biblioteca	5.883
40.103	Página personal	0.015

Parámetros K_{ij}^{N} con referencia externa

0.0	5.9	3.9	0.0	8.3	1.0	1.5	10.2	3.1	23.1	4.1	7.3	100.0	7.2	Consulta Biblioteca
24.1	0.0	63.1	2.2	37.3	8.8	37.2	15.4	0.0	6.2	17.0	100.0	65.4	8.8	Liga externa
8.7	34.0	0.0	0.0	31.1	2.3	18.8	67.0	19.1	100.0	29.5	34.2	55.1	32.7	NotiCimat
0.0	4.0	0.0	0.0	7.6	0.0	4.0	100.0	7.6	0.0	3.9	0.0	62.9	25.8	Páina personal
4.5	4.9	7.6	0.6	0.0	73.4	9.0	56.6	4.5	64.7	16.6	100.0	93.6	34.6	Vinculación
0.8	1.6	0.8	0.0	100.0	0.0	1.9	9.1	2.8	14.2	2.7	9.9	9.9	91.3	Aguascalientes
1.2	7.2	6.8	0.4	13.2	2.0	0.0	100.0	5.2	12.4	5.1	24.8	59.3	25.0	Ingsoft
7.2	2.6	21.4	9.6	73.9	8.7	89.0	0.0	60.4	56.4	74.0	88.0	59.8	100.0	Eventos
3.6	0.0	10.1	1.2	9.6	4.4	7.6	100.0	0.0	34.6	4.3	13.5	39.2	25.7	Aniversario
5.5	0.4	10.8	0.0	28.4	4.6	3.7	18.9	7.0	0.0	7.3	54.3	93.2	100.0	InfoGeneral
2.2	2.2	7.1	0.3	16.3	1.9	3.4	55.7	1.9	16.3	0.0	59.3	30.1	100.0	Publicaciones
0.7	2.3	1.4	0.0	17.2	1.3	2.9	11.6	1.1	21.2	10.4	0.0	29.9	100.0	Docencia
25.5	4.1	6.4	2.2	44.1	3.4	19.0	21.6	8.5	100.0	14.4	82.2	0.0	22.4	Investigación
0.7	0.2	1.4	0.3	5.9	11.5	2.9	13.1	2.0	39.1	17.5	100.0	8.2	0.0	Administración

Sin referencia externa

Θ_i Interés Conjunto	Grupo	P_i Interés Individual
75.077	Docencia	100.000
67.998	InfoGeneral	83.727
64.987	Administracion	61.351
63.142	Investigacion	65.254
62.418	NotiCimat	68.357
57.162	Eventos	47.326
56.181	Vinculacion	29.542
50.256	Publicaciones	8.815
47.760	Ingsoft	13.648
46.182	Aniversario	9.605
45.873	Aguascalientes	9.064
41.705	Consulta Biblioteca	5.883
40.187	Página personal	0.015
0.000	Liga externa	0.000

Parámetros K_{ij} sin referencia externa

0.0	0.0	3.9	0.0	8.3	1.0	1.5	10.2	3.1	23.1	4.1	7.3	100.0	7.2	Consulta Biblioteca
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Liga externa
8.7	0.0	0.0	0.0	31.1	2.3	18.8	67.0	19.1	100.0	29.5	34.2	55.1	32.7	NotiCimat
0.0	0.0	0.0	0.0	7.6	0.0	4.0	100.0	7.6	0.0	3.9	0.0	62.9	25.8	Página personal
4.5	0.0	7.6	0.6	0.0	73.4	9.0	56.6	4.5	64.7	16.6	100.0	93.6	34.6	Vinculación
0.8	0.0	0.8	0.0	100.0	0.0	1.9	9.1	2.8	14.2	2.7	9.9	9.9	91.3	Aguascalientes
1.2	0.0	6.8	0.4	13.2	2.0	0.0	100.0	5.2	12.4	5.1	24.8	59.3	25.0	Ingsoft
7.2	0.0	21.4	9.6	73.9	8.7	89.0	0.0	60.4	56.4	74.0	88.0	59.8	100.0	Eventos
3.6	0.0	10.1	1.2	9.6	4.4	7.6	100.0	0.0	34.6	4.3	13.5	39.2	25.7	Aniversario
5.5	0.0	10.8	0.0	28.4	4.6	3.7	18.9	7.0	0.0	7.3	54.3	93.2	100.0	InfoGeneral
2.2	0.0	7.1	0.3	16.3	1.9	3.4	55.7	1.9	16.3	0.0	59.3	30.1	100.0	Publicaciones
0.7	0.0	1.4	0.0	17.2	1.3	2.9	11.6	1.1	21.2	10.4	0.0	29.9	100.0	Docencia
25.5	0.0	6.4	2.2	44.1	3.4	19.0	21.6	8.5	100.0	14.4	82.2	0.0	22.4	Investigación
0.7	0.0	1.4	0.3	5.9	11.5	2.9	13.1	2.0	39.1	17.5	100.0	8.2	0.0	Administración

3.3. Medida de Interés Dinámico

Queremos en ésta sección introducir una formulación la cual nos permita obtener medidas de interés que cambie a lo largo del tiempo.

Proponemos una función que nos permita apreciar la dinámica del interés en cada cada grupo con respecto al tiempo, brindándonos más información para entender los datos.

El planteamiento general considera a T distribuciones discretas $\psi^t = \{\psi^t(k)\}$ donde $\psi^t(k)$ es la probabilidad de visistar el k-ésimo grupo de páginas en el tiempo t y $t \in \{1, \ldots, T\}$ donde T es tiempo máximo de las sesiones.

Definimos a $\phi(k,t)$ como el valor de interés por el k-ésimo grupo de páginas al tiempo t el cual es proporcional a $\psi^t(k)$.

En nuestro experimento podemos estimar las distribuciones $\psi^t(k)$ a través de los accesos en el logfile obteniendo g_{kt}.

$$g_{kt} = \text{ Propoción de visitas al grupo } k \text{ en el tiempo } t \text{ en todo el logfile} \tag{3.7}$$

Los valores de g_{kt} son contaminados con ruido ξ_{kt} el cual suponemos log normal independiente (ver ecuación 3.8).

$$g_{kt} = \phi(k,t)\xi_{kt} \qquad \xi_{kt} \sim LogN(0, \sigma^2) \tag{3.8}$$

Para hacer la notación más compacta definimos las siguientes variables:

$$\begin{aligned} G_{kt} &= \log(g_{kt}) \\ \Phi(k,t) &= \log(\phi(k,t)) \end{aligned}$$

Reescribiendo la ecuación 3.8 tenemos :

$$G_{kt} = \Phi(k,t) + \eta_{kt} \qquad \eta \sim N(0, \sigma^2) \tag{3.9}$$

Para calcular la probabilidad de observar G_{kt} dado los valores de $\Phi(k,t)$ observamos que es equivalente obtener $P_{\eta_{kt}}(G_{kt} - \Phi(k,t))$.

$$P(G_{kt} \mid \Phi(k,t)) = P_{\eta_{kt}}(G_{kt} - \Phi(k,t)) = \frac{1}{Z_\eta} e^{\frac{-(G_{kt} - \Phi(k,t))^2}{2\sigma^2}} \tag{3.10}$$

donde σ^2 es la varianza del ruido y $Z_\eta = \sqrt{2\pi\sigma^2}$ una constante de normalización para la distribución gaussiana.

Si suponemos que Φ proviene de una distribución de Gibbs sabemos que existe una $U(.)$ tal que

$$P(\Phi) = \frac{e^{-\lambda U(\Phi)}}{\sum_{\Phi' \in \mathbb{F}} e^{-\lambda U(\Phi')}} \tag{3.11}$$

donde $\mathbb{F}$ son todos los posibles valores que Φ' puede tomar, $U(\Phi)$ es un potencial definido y λ una constante de temperatura.

Aplicando la regla de Bayes tenemos :

$$P(\Phi \mid G_{kt}) = \frac{P(G_{kt} \mid \Phi)P(\Phi)}{P(G_{kt})} = \frac{e^{\frac{-(G_{kt}-\Phi(k,t))^2}{2\sigma^2}} e^{-\lambda U(\Phi)}}{\sum_{\Phi' \in \mathbb{F}} e^{-\lambda U(\Phi')} P(G_{kt}) Z_\eta} \tag{3.12}$$

y ahora para todos los valores G_{kt}

$$P(\Phi \mid G_{11} \dots G_{kt}) = \prod_k^K \prod_t^T \frac{P(G_{kt} \mid \Phi)P(\Phi)}{P(G_{kt})} = \prod_k^K \prod_t^T \frac{e^{\frac{-(G_{kt}-\Phi(k,t))^2}{2\sigma^2}} e^{-\lambda U(\Phi)}}{P(G_{kt}) Z_\eta \sum_{\Phi' \in \mathbb{F}} e^{-\lambda U(\Phi')}} \tag{3.13}$$

donde K es el número máximo de grupos en el website y T la longitud máxima de las sesiones que utilizamos para calcular G_{kt}

Queremos maximizar la ecuación 3.13 ($arg\ max_{\ \Phi}\ P(\Phi \mid G_{11} \dots G_{kt})$) donde el denominador es una constante para cualquier valor de Φ y no influye en el proceso de optimización.

Aplicando logaritmo a la ecuación 3.13 y multiplicando por (-1) el problema a resolver es el siguiente:

$$arg\ max_{\ \Phi(k,t)} \left[\frac{1}{2\sigma^2} \sum_k^K \sum_t^T \frac{(G_{kt} - \Phi(k,t))^2}{2\sigma^2} + \lambda U(\Phi) \right] \tag{3.14}$$

donde $\frac{1}{2\sigma^2}$ puede ser absorbido por la constante λ sin afectar nuestro proceso de minimización.

Parametrizaremos $\phi(k,t)$ a través de n duplas $\{\theta_k, \alpha_k\}$ que sustituiremos en la ecuación 3.14 de la forma $\theta_k e^{\alpha_k t}$. De ésta forma modelaremos la dinámica del interés por el k-ésimo grupo θ_k en el tiempo, donde α_k es un factor de atenuación/amplificación en el tiempo.

Queremos que el potencial $U(\theta_k, \alpha_k)$ condicione la estimación de $\phi(k,t)$ a ser suave en a lo largo del tiempo. Una forma de hacerlo será a través del potencial propuesto en la ecuación 3.15 evitando cambios bruscos en el interés por cada grupo favoreciendo valores de α_k cercanos a cero.

$$U(\theta_k, \alpha_k) = \sum_{i=1}^K (\alpha_i)^2 \tag{3.15}$$

Obtenemos finalmente la función :

$$arg\ max\ _{\theta_k,\alpha_k} E(\theta_k,\alpha_k) = \sum_k^K \sum_t^T (G_{kt} - \log\left[\theta_k e^{\alpha_k t}\right])^2 + \lambda \sum_{i=1}^K (\alpha_i)^2 \tag{3.16}$$

Si definimos como $a_k = \log(\theta_k)$ y $b_k = \alpha_k$ tenemos :

$$E = \sum_k^K \sum_t^T (\log(g_{kt}) - a_k - b_k t)^2 + \lambda \sum_{i=1}^K (b_k)^2. \tag{3.17}$$

queremos encontrar $arg\ min_{a_k,b_k} E(g_{kt}, a_k, b_k)$ por lo que derivamos con respecto a cada dupla de variables $\{a_i, b_i\}$

$$\frac{\delta E}{\delta a_i} = -2 \sum_t^T \log(g_{it}) + 2a_i \sum_t^T + 2b_i \sum_t^T t = 0 \tag{3.18}$$

$$\frac{\delta E}{\delta b_i} = -2 \sum_t^T t \log(g_{it}) + 2a_i \sum_t^T t + 2b_i \sum_t^T t^2 + 2\lambda b_i = 0 \tag{3.19}$$

que en su forma matricial es:

$$\begin{bmatrix} \sum_t^T & \sum_t^T t \\ \sum_t^T t & \lambda + \sum_t^T t^2 \end{bmatrix} \begin{bmatrix} a_i \\ b_i \end{bmatrix} = \begin{bmatrix} \sum_t^T \log(g_{it}) \\ \sum_t^T t \log(g_{it}) \end{bmatrix}$$

al resolver esta matrix de 2 x 2 de la forma

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

Obtenemos:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{x_{22}y_1 - x_{12}y_2}{D} \\ \frac{-x_{21}y_1 - x_{11}y_2}{D} \end{bmatrix}$$

Donde $D = x_{11}x_{22} - x_{12}x_{21}$

Si resolvemos para cada variable k obtenemos los siguiente:

Cuadro 3.3: Grupos de páginas

λ=0		Grupos	$\lambda = 1000$	
θ_k	α_k	$NombredeGrupo$	θ_k	α_k
0,005293	0,052462	$ConsultaBiblioteca$	0,013473	0,001962
0,000000	−0,160453	$Ligaexterna$	0,000000	−0,006000
0,159283	−0,034644	$NotiCimat$	0,085947	−0,001296
0,000000	−1,016740	$PáginaPersonal$	0,000000	−0,038023
0,041046	0,013799	$Vinculación$	0,052480	0,000516
0,006587	0,073377	$Aguascalientes$	0,024332	0,002744
0,011304	0,062322	$Ingsoft$	0,034295	0,002331
0,066538	0,019992	$Eventos$	0,094992	0,000748
0,006586	0,065679	$Aniversario$	0,021213	0,002456
0,099782	0,027720	$InfoGeneral$	0,163468	0,001037
0,008534	0,044939	$Publicaciones$	0,018998	0,001681
0,143299	0,016874	$Docencia$	0,193527	0,000631
0,090723	0,015378	$Investigación$	0,119303	0,000575
0,123303	−0,019830	$Administración$	0,086619	−0,000742

Éstos resultados pueden verse también en la figura 3-4 donde por cada grupo temático se coloca la estimación g_{kt} en color negro, en color rojo en intervalo de confianza de la estimación dada por $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$, y en azul y verde los resultados de la estimación con los parámetros $\lambda = 0, 1000$ respectivamente.

En los experimentos hemos usado valores de g_{kt} donde tengamos más de 60 observaciones porque los comprometen demasiado la estimación de los valores de interés ya que habrá lugares donde solamente existe una observación

De la figura apreciamos el impacto que tiene λ en la obtención de un modelo que se ajuste a nuestros datos,sin embargo con esta modelación logramos una reducción enorme computacionalmente por el hecho de trabajar con resumenes de datos como lo son las observaciones g_{kt} y obtener una forma cerrada de calcular el modelo.

3.4. Medida de Interés longitudinal No Paramétrico

Al introducir la idea de interés longitudinal propusimos ajustar una medida de interés a un modelo de comportamiento exponencial de navegación en función del tiempo. Lo cual es de fácil interpretación, pero carece no relaciona grupos e información tomando sus semejanzas y diferencias.

En éste capitulo buscamos una forma de crear un mapa que involucre a ambas características, proponiendo un método que no paramétrico para hacerlo y de interpretación intuitiva. Con la ventaja adicional de medir

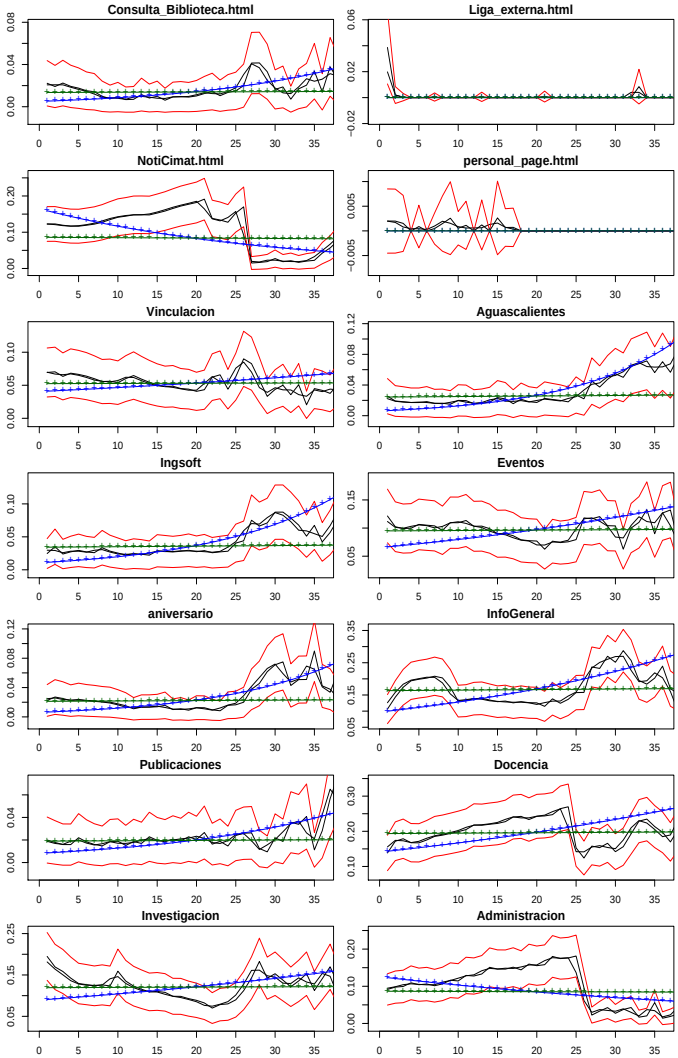


Figura 3-4: Estmación de los modelos de interés dinámico

el interés por cada grupo de manera relativa y agrupándo aquellos con intereses similares por parte de los usuarios.

Usando la misma idea de una distribución subyacente en los datos (ver sección anterior), podemos obtener una función para cada grupo temático que indique una característica particulamente interesante de los datos.

La característica que buscamos en éste caso es la primer visita al q-ésmio grupo temático. En la figura 3-5 mostramos la estimación de ésta característica denotada por $P = \{P_t^q\}$ para cada uno de los grupos dada por:

$$P_t^q = \frac{1}{d_t} \sum_{i=1}^{n} p_{i,t}^q \qquad \forall t = 0, \ldots, T; \quad \forall\ q = 1, \ldots, k \tag{3.20}$$

donde d_t es el número de sesiones de longitud $\leq t$, T el tiempo máximo, k el número de grupos, n es el número de sesiónes y $p^q_{i,t}$ la función indicadora

$$p^q_{i,t} = \begin{cases} 1 & \text{si el } q\text{-ésimo grupo es vistado por primera vez en el tiempo } t \text{ en la sesión } i \\ 0 & \text{de otra manera} \end{cases}$$

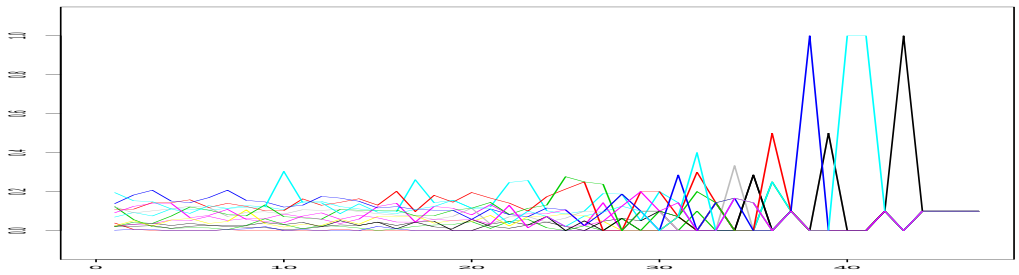


Figura 3-5: Estimación de la función de la primera visita a cada grupo de páginas

Obtenemos k funciones en el tiempo con zonas de alta variabilidad sobre su valor debido a que el número de observaciones disponibles para cada tiempo varia.

Una manera de evitar la alta incertidumbre de los datos es restringir el tamaño de la sesión para poder tener datos mejor estimados, el número mínimo de sesiones que utilizamos para estimar será 60 por lo que la longitud de la sesión se vera reducida de 47 a 24.

En la figura 3-6 mostramos los datos con los que trabajaremos, donde de color negro se muestra la estimación y de color rojo la incertidumbre calculada por $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$

A diferencia métodos anteriormente empleados no vemos a los datos no como una serie de observaciones sino como funciones en el tiempo.

3.4.1. Análisis de Datos Funcionales

Los datos originales (ver figura 3-5) han sido suavisados usando un kernel no paramétrico como lo es el método de Nadaraya-Watson [27] a fin de obtener una estimación de las funciones más suave. Los resultados son mostrados en la figura 3-7

Con estos datos aplicamos PCA con el enfoque de datos funcionales [27] obteniendo los siguientes componentes mostrados en la figura 3-8.

Solamente 4 componentes son necesarios para explicar el 93.13 % del comportamiendo de las funciones. Una manera en que se pueden presentar los resultados es a través de un biplot de las primeras dos componentes, las cuales capturan el 83.79 % del comportamiento, mostrado en la figura 3-9. Donde cada número indica la proyección del grupo temático correspondiente.

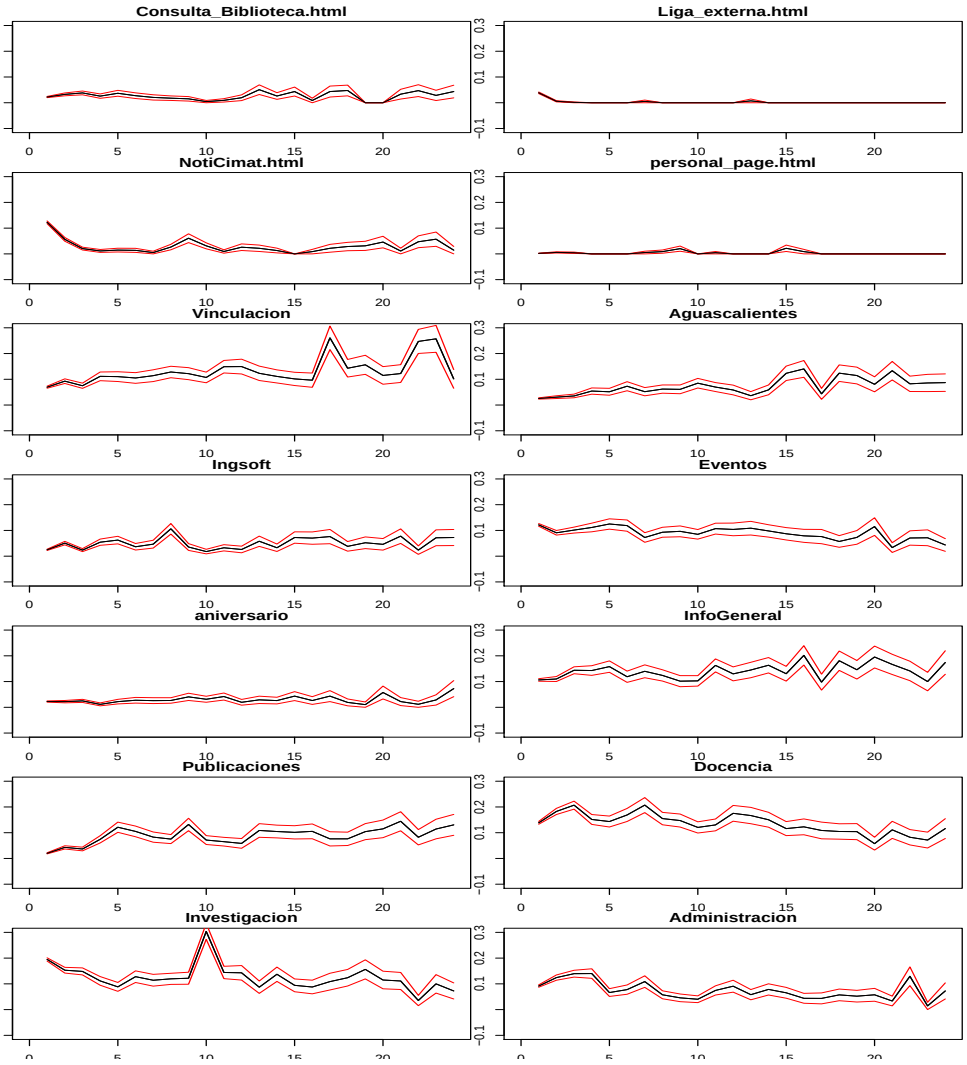


Figura 3-6: Estimación de la función primera visita a cada grupo de páginas con incertidumbre

En la proyección del biplot observamos 3 áreas en las cuales podemos encapsular el comportamiento de los navegantes de acuerdo a su interés en el grupo.

Las áreas incluyen los siguientes grupos:

- Area 1: Docencia, Investigación, Eventos y Administración.
- Area 2: NotiCimat,Liga Externa, Página Personal, Consulta Biblioteca,Aniversario e Insoft.
- Area 3: Aguascalientes, Publicaciones, Vinculación e Información General.

Una técnica para interpretar las proyecciones es a cada función añadirle un componente principal multiplicado por una constante ϵ donde ϵ es un valor pequeño positivo o negativo. En la figura 3-10 se muestran

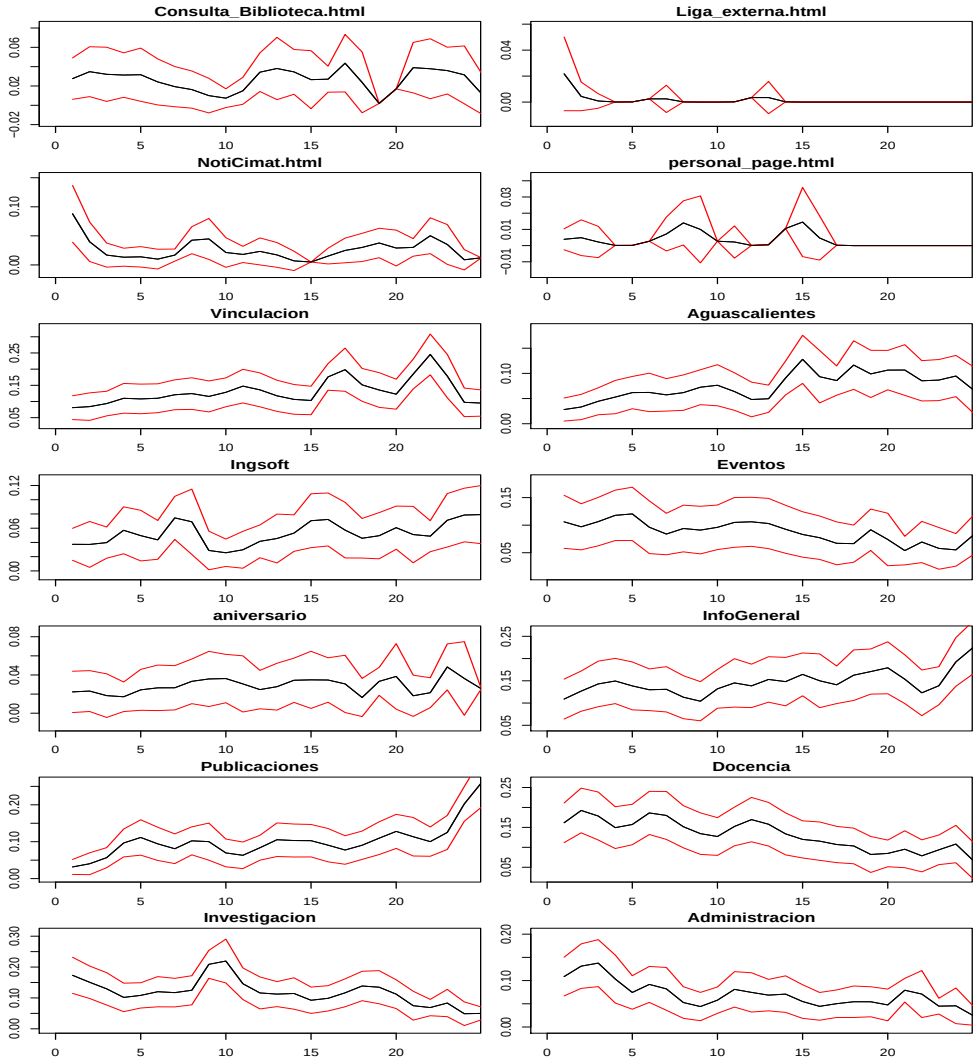


Figura 3-7: Datos suavizados con kernel deNadaraya-Watson

las primeras dos componentes sumadas a cada medición del grupo de interés, para valores negativos de ϵ se han usado colores rojos y para valores positivos valores amarillos, en negro se muestra la variable original.

La interpretación que podemos dar a estos componentes es la siguiente:

- Componente 1: Indica un promedio de visitación.
- Componente 2: Contraste entre visitar el grupo temprana o tardiamente.

Con esta interpretación podemos ver la figura 3-9 con de la siguiente forma:

- Eje Horizontal: De izquierda a derecha los grupos ordenados de forma descendente en la intensidad de visitas (Primer Componente). Siendo de gran interés con respecto a visitas los grupos del área 1 y 3.

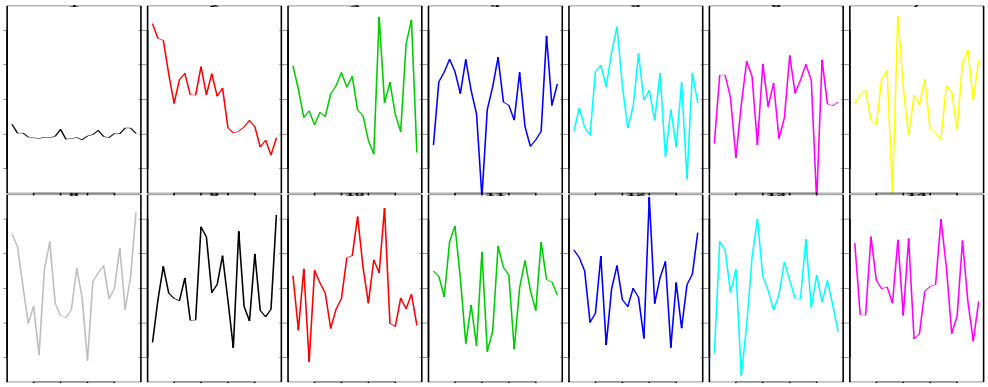


Figura 3-8: Componentes PCA datos funcionales

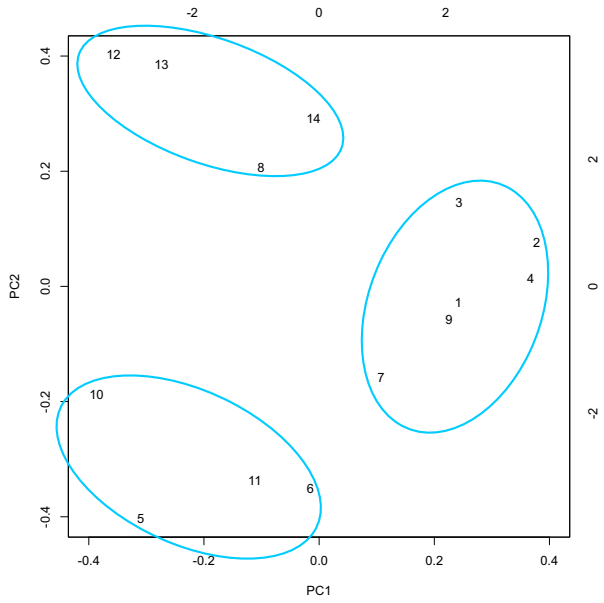


Figura 3-9: Biplot datos funcionales, lsa primeras dos componentes capturan el 83.79 % de la varianza

- Eje Vertical: De abajo hacia arriba los grupos ordenados en forma de visitas tempranas y tardías respectivamente (Segundo Componente). Llamándonos la atención el comportamiento de los grupos del área 2 los cuales no generan contraste, éstos grupos son visitados en la misma intensidad a independiente de dónde se encuentren en la sesión.

Al obtener una función del tiempo del comportamiento de los usuarios hemos ampliado las interpretaciones que les damos a los datos y con una representación tan amplia podemos relacionar a los grupos de páginas por su ubicación en el biplot de componentes y analizar su comportamiento individualmente.

Esto nos ofrece ventajas con respecto a otras técnicas antes utilizadas además de reducir la complejidad computacional del cálculo del modelo al trabajar con el resúmen de los datos.

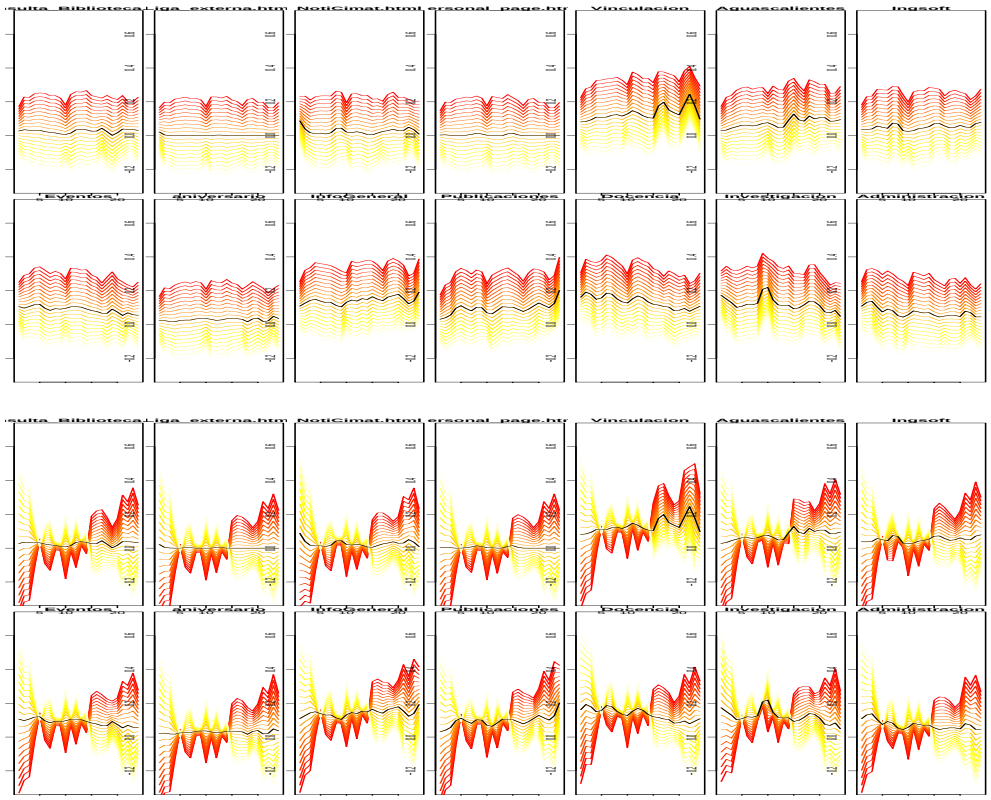


Figura 3-10: Datos más ϵ veces el componente

Capítulo 4

Aportaciones y Conlcusiones

A través del análisis en éste trabajo hemos obtenido resultados sobre como se accesan las páginas WWW del CIMAT además de construir herramientas éspecíficas para éste propósito.

Usando scripts en Perl y un manejador de páginas obtuvimos la topología del sitio además de analizar los archivos logfile y separar las páginas en grupos temáticos.

Éstos grupos de páginas fueron eje fundamental en el estudio de los datos simplificando cálculos y análisis al reducir las visitas entre páginas a visitas entre grupos.

Describimos de una manera general las características más importantes de los datos a través de técnicas de proyecciones como PCA e ICA y relaciones entre atributos de las sesiones obteniendo resultados como la importancia de la longitud de la sesión y la poca revisitación entre grupos temáticos.

Una de las caracterísitcas que encontramos entre los navegantes es que el interés por un grupo temático no cambia a pesar de la longitud de la sesión.

El uso de técnicas robustas como Rob-PCA y la adición de un espacio de características extendido mejoró la percepción que teníamos de los datos, corroborando la importancia de la longitud de la sección que habiamos obtenido anteriormente.

Al ser nuestro énfasis en la parte temporal del comportamiento de los navegantes, buscamos formas de modelar las cadenas mismas. Implementamos modelos Markovianos los cuales son considerados el estado del arte.

Trabajamos directamente con las sesiones a través de modelos que asumen Markovianidad y buscamos métodos graficos de como representarlos.

Al extender los modelos Markovianos con cadenas ocultas obtuvimos que el comportamiento de los navegantes puede resumirse en tres tipos de usuario y a través del enfoque de los estados absorbentes de las cadenas ocultas podemos visualizar como los intereses de los usuarios cambian dependiendo del grupo temático en el que se encuentran.

Encontramos que los modelos Markovianos presentan la desventaja no ser intuitivos, de dificl estimación y carecen de una interpretación evidente de sus resultados.

A través del último capítulo ofrecimos una alternativa a éstas desventajas proponiendo medidas en las cuales no es necesaria la asumpción de Markovianidad.

Propucimos que el análisis de las sesiones a un nivel menos fino nos lleva a la obtención de las características más relevantes sin perder demasiado detalle, ofreciéndonos una clara interpretación de resultados.

Presentamos 4 medidas con los que podemos modelar los datos elaborándolos a través de la idea que los usuarios consultan la información porque tienen interés en ella.

Cada función hace un énfasis en una característica de interés de los navegantes, destacándose los últimos dos modelos en los cuales permitimos que el interés de los usuarios cambie longitudinalmente en la sesión a través del tiempo.

Con éstos modelos queremos proporcionar un mapa más intuitivo del comportamiento de los navegantes, en el cual relacionamos a los grupos temáticos de acuerdo al comportamiento que tienen los usuarios sobre ellos.

Bibliografía

[1] Walker Cooley R. "Web Usage Mining: Discovery and Applications of Internet Patterns from Web Data " , University of Minesota PhD Thesis.

[2] R. Cooley, B. Mobasher, and Jaiideep, "Data preparation for mining wild wide web browsin patterns" Knoledge and Information systems 1999.

[3] Berendt B. Mobasher B. Nakagawa M. Spiliopoulou M. "The impact of Site Structure and User Enviroment on Session Reconstruction in Web Usage Analysis ", Humblod University Berlin

[4] Andrews, Keith "Information Visualization Tutorial Notes ", IICM Graz University of Technology Inffledgasse

[5] I. Herman, G. Melancon and M. S. Marshall "Graph Visualization and Navigation in Information Visualization: a Survey ", Centre for Mathematics and Computer Sciences Amnsterdam, The Netherlands

[6] T. Munzner,"Interactive Visualization of Large Graphs and Networks ",Ph.D. dissertation, Stanford University, June 2000.

[7] T. Munzner P. Burchard "Visualizing the Structure of the World Wide Web in 3D Hyperbolic Space ", Proceedings of VRML'95 Symposium ACM SIGGRAPH ACM Press 1995

[8] Marvin Jay Greenberg "Euclidean and Non-Euclidean Geometries, Development and history " , W.H. Frreman and Company

[9] De la Rosa D. .ªnálisis de Secuencias: Revisión de técnicas usadas en los últimos años " B de éste documento.

[10] Lawrence R .Rabiner , .ª Tutorial on Hidden Markov Models and Selected Aplications in Speach Recognition",IEEE.

[11] Jeff A. Bilmes , .ª Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models",International Computer Science Institute Berkeley California.

[12] Alan B. Poritz "Hidden Markov Models : a Guided Tour ", Institute of Defence Analysis, Communicatons Research Division Princeton NJ

[13] Ypma, A, Heskes T. Çategorization of Web pages and user clustering with mixtures of hidden Markov models ", SSN Nijimegen, Geert Grooteplein The Netherlands

[14] Cooley R. Mobasher B. Srivastava J. ,"Data preparation for Mining WWW Broswsing patterns ", Department of Computer Science and Engineering University of Minnesota.

[15] Batista P. Silva M. ,"Mining web access logs of an Online newspaper " ,Departamento de Informatica Universidad de Lisboa

[16] Smith K. Ng A. ,"Web page clustering using a self-organizing map of user navigation patterns ", School of bussines SystemsMonash Univ. Australia

[17] Ajith A. ,"i-Miner: A Web usage Mining Framework Using Hierarchical Intelligent Systems ", Deparment of Computer Science Oklahoma University.

[18] Nakagawa M. , Bamshad Mobasher, "Impact of Site Characteristics on Recomendation Model Based on Association Rules and Sequencial Patterns " ,School of Computer Science, Telecommunication and Information Systems DePaul University Chicago Ill.

[19] Glover E. J., Tsioutsiouliklis K. Lawrence S. Pennocj D, Flake G. ,Üsing web structure for classifying and describing web pages " ,Computer Scence Department Princeton University.

[20] Buchner A.G Baumgarten M. Anand S.S. ,"Navigation Pattern Discovery from Internet Data " ,Northen Irelan Knoledge Engineering Laboratory University of Ulser

[21] Charles M. Grinsead, J. Laurie Snell "Introduction to Probability " , Dartmouth College

[22] J. Van Horebeek "Introducción a los Métodos Probabilisticos en Ciencias de la Computación ", Centro de Investigación en Matemáticas CIMAT Gto. México

[23] T. Hastie, R. Tibshirari, J. Friedman, "Elements of Statistical Learning" , Springer

[24] M. Hubert, P. J. Rousseeuw, K. Vanden Branden .ª New Approach to Robust Principal Component Analysis: Rob-PCA"

[25] J. L. Marroquín , E. Arce , S. Botello, "Hidden Markov Measure Fields Models for Image Segmentation", Comunicación Técnica CIMAT No.I-02-05/12-04-2002

[26] J. L. Marroquín , E. Arce , S. Botello, "Hidden Markov Measure Fields Models for Image Segmentation", IEE Transactions on patter Analysis and Machine Inteligence Vol25 No. 11

[27] J.O. Ramnsay, B. W. Silverman, "Functional Data Analysis", Springer, Series in Statistics.

Apéndice A

Documentación

A.1. Documentación

A lo largo de éste trabajo se desarrollaron varias herramientas las cuales permitieron hacer el análisis de los datos en la figura A-1 se presenta el esquema de como son usadas éstas herramientas

A.2. Perl

A.2.1. StripRealSession

Dado un archivo log de entrada junto con sus parámetros crea una lista de páginas visitadas en el archivo, genera cadenas de visitas de páginas en orden cronológico por IP junto con un resumen del orden en que se visitan las páginas. Imprime en orden alfabetico las páginas encontradas en el archivo log.

Ejemplo :

```
StripRealSession.pl  <Log-file>  <parametros>
```

Entradas	
<Log-file>	Archivo log del website
<parametros >	parametros para poder procesar el archivo log
	Ejemplo
	diaInic=01 & mesInic=01 & yearInic=2005 & horaInic=00 & minInic=0 &
	segInic=0 & diaFin=3 & mesFin=1 & yearFin=2006 & horaFin=22
	& minFin=00 & segFin=9 & gifs=off & jpgs=off & bmps=off &
	javas=off & Tiempo=300
Salidas	
IPs.dat	Cadenas de visitas a página agrupadas por IP
Pages.dat	Nombre de páginas enumeradas
FinalLog.dat	brincos entre páginas a lo largo del logfile

A.2.2. StripIPlogfile

Parte las sesiones encontradas por IP en sesiones más pequeñas ya sea por venir de una referencia externa por ser demasiado larga o por haber demasiado tiempo entre visitas.

Ejemplo :

```
StripIPlogfile.pl  <IPs.dat>  <UMBRAL_DE_TIEMPO> <MAX_LARGO> <Pages.dat> > <RawLogFile.dat>
```

Entradas	
<IPs.dat>	Archivo de cadenas agrupadas por IP
< UMBRAL_DE_TIEMPO >	Tiempo máximo entre visitas en el log file
<MAX_LARGO>	Largo máximo de sesión
<Pages.dat>	Nombre de las páginas
Salidas	
<RawLogFile.dat>	Archivo

A.2.3. ClusterSplit

Busca patrones en las URL de las páginas del log file para agruparlos en clusters y facilitar la asignación de grupos a cada página.

Ejemplo :

```
ClusterSplit.pl <Pages.dat> >  ClusterMapping.dat
```

Entradas

<Pages.dat>	Nombres de las páginas encontradas en el log file

Salidas

<ClusterMapping >	Lista de las páginas asignadas a un grupo encontrado por la busqueda de patrones
<AviableClusters.dat>	Lista de patrones y clusters encontrados de forma automática.

A.2.4. MergeSession

Fusiona dos archivos log previamente procesados y unifica el nombre de las páginas y sus asignaciones de grupos.

Ejemplo :

```
MergeSessions.pl Pages2.dat RawLogFile2.dat
```

Entradas

<Pages.dat>	Lista de paginas originales
<ClusterMapping.dat>	Asignacion de cada página a un grupo de paginas (cluster)
<RawLogFile.dat>	Archivo log pre procesado al que se le agregarán nuevas entradas
<Pages2.dat>	Lista de páginas las cuales se agregarán.
<Rawlogfile2>	Archivo log preprocesado al cual se agregará.

Salidas

<RawLogFile.dat >	Se modifica el archivo original y se añade el segundo log file unificando las páginas y clusters

A.2.5. MapsAndSupress

Tomando un archivo log preprocesado se eliminan páginas y se mapea el ID de la página a el ID del cluster al que pertenece

Ejemplo :

```
MapsAndSupress2.pl Pages.dat SupresedPagesMOD.dat ClusterMappingMOD2.dat  RawLogFile.dat > FinalSessionData.dat
```

Entradas

<Pages.dat >	Lista de paginas en el log file
<SupresedPagesMOD.dat >	Lista de paginas que serán suprimidas en el archivo final
<ClusterMappingMOD2.dat >	Mapeo de páginas a clusters
<RawLogFile.dat >	Archivo log pre procesado

Salidas

<FinalSessionData.dat >	Archivo log final con sesiones de paginas
<TimeStampSession.dat >	Archivo log final con sesiones de paginas junto con su marca de tiempo asociada a cada una

A.2.6. BuildGraph

Dados dos archivos de nodos y ligas se construye un grafo en el formato que ***walrous*** puede leerlo.

Una serie de atributos se asocia a cada nodo para su visualización.

Ejemplo :

```
BuildGrapho  <nodes-file> <links-file> <# users> > graph
```

Entradas	
< nodesfile>	Archivo de nodos
< linksfile>	Archivo de conexiones del grafo
< # users>	Numero de usuario encontrados por las cadenas ocultas de markov
Salidas	
<graph >	Grafo final

A.2.7. GetCimatNodes

De manera exahustiva busca las páginas y ligas del website , únicamente busca paginas dentro del sitio que estén conectadas por alguna liga con las restricción de que busque hasta cierto nivel jerarquico de páginas

Ejemplo :

```
GetCimatNODES <# levels> <nodes-file> <links-file> <origin>
```

Entradas	
<# levels >	Numero de niveles a buscar en el site
<nodesfile >	Nombre del archivo donde se guardaran los nodos y sus nombres
<linksfile >	Nombre del archivo donde se guardaran las ligas y conexiones entre nodos
<origin >	página en donde se comenzará la busqueda
Salidas	
<nodesfile >	Archivo donde se guardaran los nodos y sus nombres
<linksfile >	Archivo donde se guardaran las ligas y conexiones entre nodos

A.3. C++

A.3.1. PageClassifier

Manejador de páginas visual, el cual muestra la página que se va a asignar a un grupo de páginas , se pueden agregar editar grupos ademas de agregar otro log file previamente procesado

Ejemplo :

```
PageClasifier
```

Entradas	
<Pages.dat >	Lista de paginas
<AviableClusters.dat >	lista de clusters o grupos de páginas
Salidas	
<ClusterMappingMOD.dat >	Asignación a pagina a cada cluster
<SupresedPagesMOD.dat >	Lista de paginas que se eliminarán del log file
<AviableClustersMOD.dat>	Lista de clusters

A.3.2. Chmm

Calcula los parámetros de las cadenas ocultas de Markov de acuerdo al modelo propuesto.

Ejemplo :

```
Chmm <USUARIO-DE-CONTRASTE> <RawLogFile> <# de sessiones> <# de tipos de usuario> <# iteraciones >
```

Entradas	
<USUARIO-DE-CONTRASTE>	Usuario de contraste 0 no usuario, 1 Aleatoreo Uniforme, 2 Proporcional al numero de pagi
<RawLogFile>	sesiones extraidas del logfile mapeadas a clusters
<# de sessiones>	Numero de sesiones
<# de tipos de usuario>	Numero de tipos de usuario
<# iteraciones>	Numero de iteraciones
Salidas	
E_Init_Pix.dat.dat	Resultado de la estimacion del vector Π^x
E_Transition_Pix.dat	Resultado de la estimacion de la matriz A^x
E_Responsability.dat	Resultado de la estimacion de la matriz C_{ik}

A.3.3. VisualMatrix

Muestra las matrices de transición obtenidas por las cadenas ocultas de Markov presentado el tiempo esperado de antes de ser absorbida en un estado , estados absorbentes

Ejemplo :

```
VisualMatrix
```

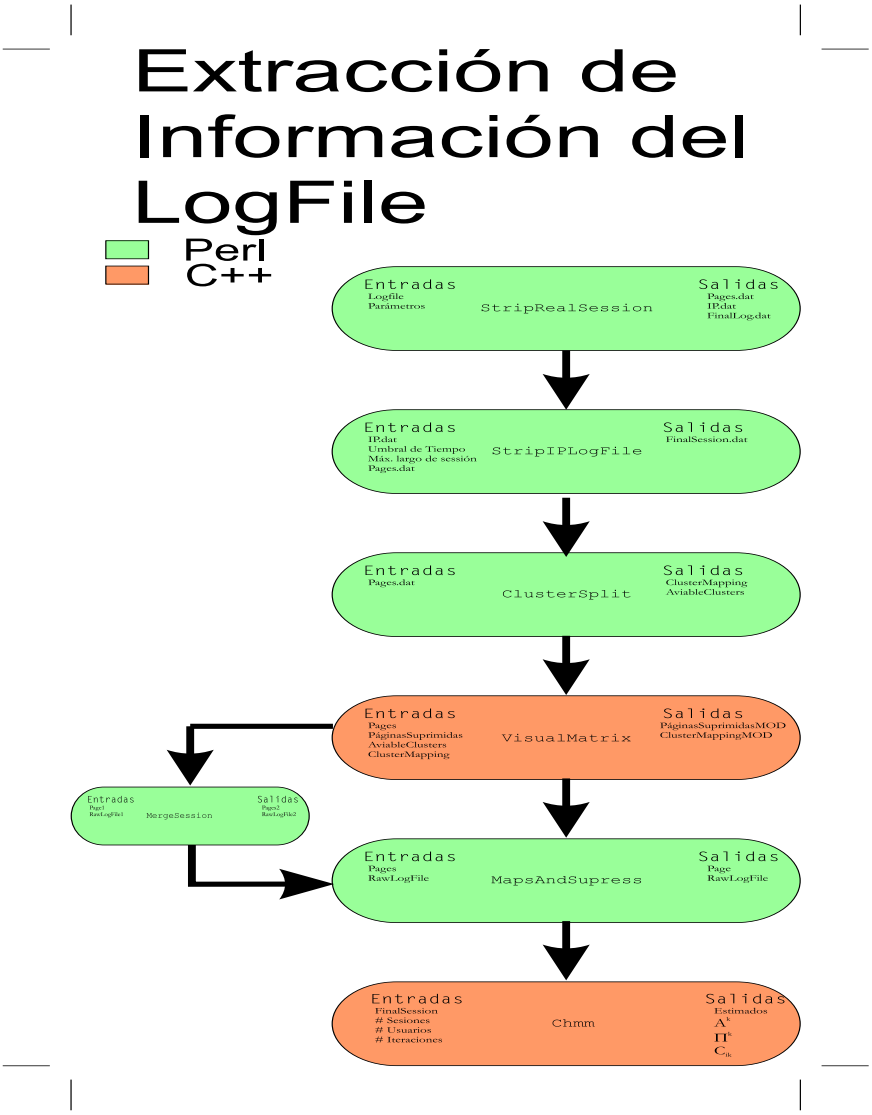


Figura A-1: Diagrame de uso de las herramientas

Figura A-2: Esquema de herramientas Visuales

Apéndice B

Bibliografía Adicional

Análisis de Secuencias: Revisión de técnicas usadas en los últimos años

David de la Rosa H.
davolin@cimat.mx

El siguiente documento es un resumen de las técnicas que se han estado usando en los últimos años para extraer información interesante de un archivo LOG.

Definición 1 *La notación que usaremos a lo largo de todo el documento sera la siguiente:*

Conjunto de n_1 *Usuarios* $U = \{u_1, u_2, u_3 \dots u_{n_1}\}$

Conjunto de n_2 *Direcciones IP únicas* $IP = \{IP_1, IP_2, IP_3 \dots IP_{n_2}\}$

Conjunto de n_3 *Secciones de navegación de cada usuario* $S = \left\{S_j^i\right\}$
Donde S_i^j *es la* $j-esima$ *sesión del usuario* i
˜ *Secuencia de visita del usuario* i *en su* $j-esima$ *sesión* $S_i^j = \{c_1, c_2, c_3 \dots c_n\}$

Conjunto de n_4 *Página* $P = \{p_1, p_2, p_3 \dots p_{n_4}\}$ *diferentes que conforman un sitio Web*

Una vez definida nuestra notación presentaremos una a una las técnicas que se han encontrado para poder dar solución a este problema

B.1. Clustering

Una idea muy intuitiva de este problema es el agrupar a usuarios que tengan características similares. Existen muchas maneras de definir las semejanzas entre los usuario en U por lo que solamente mecionaremos las más relevantes.

B.1.1. Modelamiento por bloques (BlockModeling)

Esta técnica es derivada de la idea de hacer un análisis social de una red o mejor dicho de las interacciones que surgen dentro una red. La red que definimos esta escencialmente conformada por nodos y conexiones entre estos.

La meta de este método es poder resumir el comportamiento social de una red compleja en un modelo donde las unidades "lógicas"[1] o grupos que estan asociados entre si de alguna manera.

Definición 2 *Una Relación en un conjunto dado:*

Dado un conjunto X, una relación R en X es un subconjunto del producto cartesiano $X \times X$. De aquí cualquier elemento $x \in X$ tiene una relación con el elemento $y \in X$ ssi (x,y) es un elemento de R.

Ahora del grupo de los usuarios U definimos la relación entre sus elementos como la matrix binaria R donde cada elemento R_{ij} de esta matriz es uno si existe una relación entre sus miembros y cero de cualquier otra forma.

En este caso estamos hablando que una relación entre los usuarios U_i, U_j si dentro de las sessiones S^i, S^j de estos usuarios comparten un número Θ de páginas en común.

Nuestro propósito es encontrar un grupo de particiones $C = \{c_1, c_2, \ldots c_k\}$ que $\bigcup_i c_i = E$ y $\forall\ i \neq j \Rightarrow c_j \cap c_i = \oslash$ donde E es el conjunto de actores, que pueden ser o el conjunto de usuarios U o el conjunto de las páginas visitadas S_i^j formando asi un cluster.

La forma de poder medir la similitud de los actores deacuerdo a [1, 2] puede ser medida de muchas maneras y deacuerdo a muchos parámetros, teniendo una función multi-objetivo la cual potimizaremos usando como operadores principales el intercambio de actores entre grupos y la asignación de actores a grupos.

Esta técnica es un tipo de aprendizaje supervisado, puesto que de ante mano es necesario indicar cuántas particiones del conjunto de actores se deben de hacer, por lo que necesita un análisis extra para poder determinar el número óptimo de particiones. Además de encontrar las particiones este método nos da una relación entre particiones lo cual es una ventaja.

[1] Refiero a unidades lógicas como bloques de nodos o usuarios a los cuales el conjunto que representan puede atribuirse una propierdad, ie usuarios más frecuentes, páginas relacionadas por contenido , páginas publicadas últimamente etc

B.1.2. Inteligencia artificial y mapas auto organizados

A parte de los métodos de optimización "clásica" existen formas alternativas de resolver el problema de agrupar elementos del mismo tipo de una manera óptima. Inspirados en el comportamiento natural de las hormigas existen técnicas que sintetizan el comportamiento de colonias de estos insectos obteniendo resultados bastante favorables en el sentido de que generan sistemas o agentes adaptables a las diferenctes situaciones que pudieran presentarse a lo largo de la resolución del problema evitando el uso de heurísticas complicadas.

El operador principal del método de agrupamiento(*clustering*) es el cambio de posición de un elemento de nuestros datos a otro lugar, esto se traduce en un agrupamiento del tipo demográfico, es decir una vez que el algoritmo termina de ejecutarse sobre los datos, los clusters serán formados por los elementos vecinos, los cuales fueron movidos de su ubicación original(aleatoria) para ser colocados en el grupo que les corresponde, la manera de visualizar estos grupos generalmente es en un plano en 2D el cual funjirá como nuestra arena o espacio en el cuál podremos apreciar conforme avanza el algoritmo.

La idea que se introduce en [4, 6] es el concepto de ferormonas, y una medida de semejanza entre elementos o actores, ya que un agente una vez que se topa en su camino con un elemento evalua que tan parecido es de sus vecinos, y con una probabilidad alta "recoje" el elemento si es diferente a estos y con propabilidad alta lo deja en caso de que encuentre una alta concentración de ferormonas en el espacio en que está, por último si encuentra un grupo de elementos cuya similitud es mucha con respecto al objeto que está cargando lo dejará en este lugar.

El concepto de ferormonas no es muy complicado, simplemente establece una manera en que los agentes (hormigas) se comuniquen unas con otras indirectamente, evitando el hecho de proveerlos de memoria o histoira de las cosas que cada uno ha hecho, (mejorando la complejidad computacional) la concentración de ferormonas se basa en cuántos elementos hay en el espacio y la circulación de agentes por el mismo espacio, como parámetro adicional la "persistencia" de las ferormonas tiene un parámetro θ en cual indica después de cada cierto tiempo la concentración de ferormonas decae o se disuelve, ayudando a la convergencia del algoritmo, junto con su la fusión de de grupos similares en el mismo cluster.

Las probabilidades de recojer(P_r) y dejar(P_d) un elemento por un agente es la siguiente:

$$P_r = \left(\frac{k_1}{k_1 + f}\right)^2$$

$$P_d = \left(\frac{f}{k_2 + f}\right)^2$$

Donde :

$$k_1 = cte$$

$$k_2 = cte$$

$$f(o_i) = max\left\{0, \frac{1}{s^2} \sum_{o_j \in\ Vecindad_(xxx)(r)} \left[1 - \frac{d(o_i, o_j)}{\alpha}\right]\right\}$$

Donde f da una idea de como el agente persive la cantidad de elementos en su vecindad y $d(o_i, o_j)$ es la distania entre dos objetos en el espacio de las características de cada individuo (diferente al espacio en R^2 donde se visualiza el resultado de la partición del espacio)

A pesar que pudiera parecer no paramétrico el algoritmo es necesario darle valor a unas constantes que manejan el comportamiento de los agentes, lo cual hace muy dependiente el resultado de estas constantes.

Existen otros métodos como este [5] , el cual no utiliza agentes sino usando Programación Lineal Genética (PLG) . Los resultados de estas modificaciones no son muchos ya que el principio es el mismo.

B.1.3. Mapas Auto Organizados Cont.

Una metodología diferente es la que propone [7] en el sentido que las seciones de los usuarios fueron binarizadas tomando en cuenta si visitaba o no una página en específico, pero por la dimensionalidad del problema, se tiene que cada sesión esta inscrita en un hiper cubo de dimensión n_4. Las técnicas de Mapas Auto Organizados (*Self Organized Maps SOM*) son demasiado costosas computacionalmente hablando para dimensiones tan grandes, por lo que se propone el resumir la información en alta dimensionalidad en características, que en principio resumirian la información de la sesión de una manera fiel.

La transacciones en las que se resumen las secciones de cada usuario de ser de un largo de n_4 son obtenidas por el método de *k-means clustering*[7] resumiento así las características en k atributos que usaremos para el SOM.

La metodología propuesta es paramétrica ya que el valor de K es dado por el usuario asi como el número de clases o etiquetas que usaremos para el SOM.

Ahora para obtener el mapa , aplicamos el algoritmo propuesto por [7] el cual esta basado en pesos, cada nueva entrada de nuestros datos será comparada con los pesos de cada etiqueta de nuestro conjunto(número de etiquetas de nuevo es un parámetro), pensando en la idea de una red neuronal, los pesos de cada una de las entradas a las etiquetas de nuestro mapa se entrenan con datos, modificandolos poco a poco dependiendo de que tan cerca esta el nuevo dato, los pesos de la etiqueta más cercana se modifican junto con sus vecinos hasta tener una convergencia evidente o hasta terminar con e número de datos de prueba.

Este método nos provee una manera bastente sencilla de poder agrupar a los usuarios y poder predecir el grupo al que pertenecen los nuevos usuarios, lamentablemente no toma en cuenta la secuencia de navegación que es una característica con mucha información para poder hacer este analisis.

B.2. Patrones de Navegación

En un esfruerzo de extraer de los archivos Log de sevidores Web información útil e interesante , se han planteado diferentes perspectivas de atacar y analisar el problema, en este caso analisaremos un enfoque que engloba varias téncnicas muy similares, o que parten de los mismos supuestos.

B.2.1. Reconstrucción de Sesión

Los datos crudos tal y como son provistos en el archivo Log no son de mucha utilidad , hay que efectuar cierto pre-proceso a estos datos para que puedan ser útiles tanto para su manejo como su interprestación.

Cualquier usuario Ui deja un rastro en el archivo LOG pero no todo su comportamiento no es muy claro o es demasiado obscuro para poder interpretarlo a simple vista, es por eso que es necesitamos interpretar la información de estas entradas del archivo.

B.2.2. Limpieza

Como primer paso hay varios criterios para hacer éste procedimiento, uno es que únicamente estamos interesados en los archivos de contenido, es decir archivos que representen contenido a usuarios, como pag HTML , por lo tanto cualquier otra entrada que no corresponda a estas entradas será removida.

Existen además de imágenes otro tipo de archivos los cuales son removidos JAVAS , FORMAS entre otros que pueden representar alguna información para el usuario , pero por ser un herramienta auxuliar al HTML para desplegar información tambiés son removidos.

B.2.3. Identificación del usuario y Reconstrucción de Sesión

Cada usuario U_i en forma ideal deja un rastro único en su paso por la página Web, debido a que la información contenida en la IP_i es también única. Debido a técnias para optimizar la navegación Web, muchas tecnologias se han desarrollado para dar solución a este problema pero como un efecto, indeseado para nuestros propositos, la unicidad de los usuarios se pierde puesto que existen varios usuarios que cuya huella es la misma IP.

Adicionalmente a la perdida de unicidad de usuarios el uso del cache de los navegadores generan una perdida de información bastante sensible en las secciones del usuario. Diferentes tipos de heurísticas sean propuesto para resolver este problema, basicamente se dedican a dividir sesiones largar y a fusionar sesiones cortas, las heurísticas se pueden dividir en dos clases:

1. Basadas en tiempo: Tomando como referencia a [17] se reconstruyen en base a parámetros como el tiempo promedio de duración de una sesión o el tiempo empleado en una página no debe de exeder cierto umbral. Las reconstrucciones efectuadas son en extremo dependientes de los parámetros, [14, 15]

proponene unas métricas las cuales pretenden dar una idea de que tan fieles son nuestras reconstrucciones.

Las métricas propuestas [14]estan principalmente en función de la cantidad y la precisión de las sesiones recontrustruidas, es necesario tener muy en cuenta que cada heurística pretende disminuir el error atacando el problema en en especifico, pero son suceptibles a la topología del sitio y su tamaño.

2. Basadas en referencias: Se basan en saber si hay concexión directa o no entre las páginas, en caso de no haberlos se asigna la sequuencia a otra sesión. [16]

Una estrategia más inteligente es combinar estas dos formas de manejar los huecos en las sesines, una manera de hacer esto es combinarlas. Otro dato muy útil es el tiempo en el que el usuario pasa viendo una página, esto nos lleva a la idea de poder clasificar nuestras páginas en páginas de contenido y auxiliares, ya que solamente nos cosas que tiene sentido consultar, [13, 16, 17] proponen únicamente utilizar este tipo de sesiones, las cuales llaman sesiones únicamente de contenido, removiendo las ligas que no son de contenido.

Ahora tenemos un problema de clasificación en el cual tenemos como predictores: el número de ligas, tiempo promedio de visita, longitud de texto, número de visitas entre otros.

Ayudados con esta información ahora podemos reducir nuestro problema únicamente con las sesiones de contenido, pero tenemos el problema, que realmente los hábitos de navegación y la idea de secuencias en este tipo de sesiones ha sido removida y los resultados obtenidos pudieran resultar un tanto artificiales con respecto a los verdaderos usos que se le da a la pág.

Combinando esta información podemos tener una mejor reconstrucción de las sesiones del usuario, como por ejemplo el tener una conjunto de sesiones activas, que sería una manera de poder identificar mejor las sesiones, y cada nueva transacción se le asignara dependiendo de si la combinación de referencias que lleva a la página junto con la IP del usuario pueden llevar a discriminar sesiones hasta llegar solamente a una, tomando en cuenta si es una pag de contenido o auxiliar, en caso contrario adicionar una nueva sesion a la lista de sesiones activas.

B.2.4. Procesamiento de Sesiones

Ahora con un conjunto de sesiones S es posible procesarlas deacuerdo a diferentes criterios, como el agruparlos (*clustering*) o buscar relaciones entre ellos.

Una práctica muy común es el análisis de ligas (*link analysis*) en cual busca reglas de asociación basado en la cantidad de elementos repetidos en una secuencia. Las técnicas no difieren mucho de la combinatoria que representa el buscar las parejas en cadenas y estadísticos de sus apariciones. Simplemente se buscan las probabilidades condicionales cuyo valor exeda cierto umbral.

La manera en que se enuncian las reglas encotradas es:

En el evento que U_i visita la página p_i

En $x\,\%$ de los casos el usuario U_i también visita p_j

Esto pasa en un $y\,\%$ de los casos

El segundo renglón forma la regla descubierta con su factor de confianza y el último renglón es el vector de soporte.

El método es paramétrico ya que solamente reportará las reglas que hayan superado el umbral del factor de soporte y confianza o la combinación de los dos. Umbrales muy altos estableceran reglas que son demasiado evidentes y que realmente no son interesantes, en el sentido que no hay incertidumbre en que sean ciertas, intervalos bajos reportan un número humanamente no manejable , para poder interpretarlo de una manera fácil.

B.3. Enfoque Markoviano

Uno de las mejores maneras de conceptualizar las secuencias de visita son las cadenas de Markov y otros enfoques como este, ya que es un enfoque que toma en cuenta el tiempo en el que se efectuan las transiciones entre otras cosas.

B.3.1. Cadenas de Markov

Una manera de conceptualizar una sesión S_i^j es verla como una caminata a lo largo de los nodos del grafo del sitio. Donde existen probabilidades (estimaciones) que nos proveen de los más posibles candidatos a el siguiente estado o paso en esta caminata.

El problema se puede enunciar de la siguiente manera:

Tenemos un la tercia $\langle S, A, \lambda \rangle$ donde S es un vector que nos da el estado actual del sistema A es la matriz de transición de estados y λ es la probabilidad inicial de los estados en S. Algo fundamental de las cadenas de Markov (de primer orden) es que solamente depende del estado anterior estableciendo la probabilidad de pasar al siguiente estado.

$$S(t) = A\ S(t-1)$$

Cada elemento i, j de la matriz de transición A nos indica la probabilidad de ir del estado i al estado j esto es obtenido del conjunto de datos con el que estimaremos cada una de las entradas de la matriz. Donde entrada de esta matriz es una probabilidad condicional.

$$\hat{A}_{i,j} = \frac{C(i,j)}{\sum_x C(i,x)}$$

Donde $C(y,k)$ es el número de veces que en el conjunto de entrenamiento i precede a j. La estimación de λ se hace viendo todos los posibles estados y cuantos de ellos son el estado que nos interesa. Donde $\lambda(i)$ es una probabilidad marginal de encontrarnos al inicio en el estado i.

$$\hat{\lambda}(i) = \frac{C(i)}{\sum_x C(x)}$$

En este caso, podemos hacer estimaciones o predicciones a corto plazo ya que la matrix A esta disenãda para este propósito. Si tenemos un vector $i(t)$ en donde la única entrada no cero del vector es el estado en donde nos encontramos en el estado t entonces.

$$\hat{S}(t) = \hat{A}\ i(t-1)$$

Obteniendo así el estado más probable del sistema a partir la certeza del estado anterior. Una extensión de esto es la obtención de la matriz A^N donde la entrada A^N_{ij} es la probabilidad de $i \to j$ en N estados antes.

Reescribiendo la ecuación principal tenemos :

$$S(t) = \sum_n \frac{1}{n} A\ i(t-1)$$

Que en escencia no es mucho cambio, pero nos da la oportunidad de escribir lo siguiente:

$$S(t) = aA\ i(t-1) + aA^2\ i(t-2) + aA^3\ i(t-3) + aA^4 i(t-4) \ldots aA^n\ i(t-n)$$

Donde cada término es la predicción del estado actual poniendo énfasis en el estado anterior n y a es una constante donde $\sum_n a = 1$ que sirve simplemente para normalizar las probabilidades .

Teniendo en cuenta que el cáculo de A^N parte de que A es estimada, el cáculo de $i \to n_1 \to n_2 \to n_N \to j$ pasando por N páginas o valores , es de un valor diferente que elevar a la potencia N la matriz A, por lo que se propone que se haga directamente de los datos de entrenamiento.

Donde la predicción en $S(t)$ está ahora dada por los n estados anteriores, donde $i(t-n)$ es la secuencia del usuario del que queremos predecir su próxima visita.

El Modelo de Cadenas se amolda muy bien a la manera en que los usuarios utilizan un sitio y la dependencia únicamente de los n estados anteriores es una gran ventaja. Un inconveniente importante es la complejidad computacional que representa el cálculo de A^N ya sea por potencias o estimandola junto con su dimensión n_4

La capacidad de predicción de este método es buena deacuerdo a [19] pero no provee un método para poder interpretar el comportamiento de los usuarios ni poderlos clasificarlos en grupos.

B.3.2. Modelos Markovianos Ocultos

Otro enfoque Markoviano es el uso de Modelos Markovianos Ocultos, los cuales además de las supocisiones que se hacen por las cadenas de Markov, introducen la idea que hay variables que no observamos y que causan un impacto en la manera en que se comporta el sistema.

Como ventaja sobre las cadenas de Markov, podemos generar grupos (*clustering*) por la manera en que se comportan las secuencias de los usuarios. El número de grupos(Θ) es definido por el usuario, y a cada usuario será asignado a alguna categoria o una etiqueta de a que tipo de usuario pertnenece.

Cada página tiene asociado a ella, una categoria en la que particionamos P con etiquetas que son desconocidas de antemano.

Cada entrada de la matriz de transición A_{ij} nos da la propabilidad de que un usuario U_x pase a ser un usuario clase i a un usuario clase j.

Tenemos una matriz de observaciones la cual denotamos como B la cual nos da la categorización de las páginas.

Ahora los estimadores de estas dos matrices son los siguientes;

$$\hat{A}_{ij} = \frac{\sum_i c_{ik} \sum_{j=1}^{n_i} \sum_{t=1}^{T_{ij}-1} \xi_t^{i,j,k}(x, \acute{x})}{\sum_i c_{ik} \sum_{j=1}^{n_i} \sum_{t=1}^{T_{ij}-1} \gamma_t^{i,j,k}(x)}$$

$$\hat{B}_{ij} = \frac{\sum_i c_{ik} \sum_{j=1}^{n_i} \sum_{t=1}^{T_{ij}} \xi_t^{i,j,k}(x, \acute{x})}{\sum_i c_{ik} \sum_{j=1}^{n_i} \sum_{t=1}^{T_{ij}} \gamma_t^{i,j,k}(x)}$$

Donde:

$$\gamma_t^{i,j,k}(x) = P(X_t = x \mid Y^{i,j}, \Theta_k)$$

$$\xi_t^{i,j,k}(x, \acute{x}) = P(X_t = x, X_{t+1} = \acute{x} \mid Y^{i,j}, \Theta_k)$$

]

Donde Θ es el número de grupos propuestos, c_{ij} es la probabilidad de que el usuario i sea de la clase $k \in \Theta$ dado su sesión y el modelo de grupos propuesto.

Bibliografía

[1] Scholer G., *Blockmodeling Techniques for Web Mining*, Dipartamento di Scienze Economiche e Statistiche, Universitá di Trieste, Piazzale Europa , Italy

[2] Batagelj V. , Ferligoj A, *Clustering Realtional Data* University of Ljubljana , Faculty of Mathematics and Physics and Institute of Mathematics , Pgysiscs and Mechanics Slovenia

[3] Ferligoj A., Batagelj V., Mrvay A. (April 2003) ,*Course on Social Network Analysis Block Modeling* University of Ljubljana Slovenia .

[4] Sntosh K., Phoha V. , Selmic R. *Web user cluster and its aplplication to prefetching Using ART Neural Networks* Department of computer Science Lousiana State university

[5] Ajith A., Ramos V. ,*Web Usage Mining Using Artificial Ant colony Clustering and linear Genetic Programming* Department Of Computer Science Oklahoma State University, Technical University of Lisbon Portugal

[6] Ramos V. , Muge F. Pina P. ,*Self-Organized Data Image Retrival as consequence of Interdynamic synergistic realtionships in artificial Ant colonies* , GeoSystems Centre, Technical Univ. of Lisbon Portugal.

[7] Smith K. Ng A. ,*Web page clustering using a self-organizing map of user navigation patterns* School of bussines SystemsMonash Univ. Australia

[8] Ajith A. ,*i-Miner: A Web usage Mining Framework Using Hierarchical Intelligent Systems* Deparment of Computer Science Oklahoma University.

[9] Nakagawa M. , Bamshad Mobasher, *Impact of Site Characteristics on Recomendation Model Based on Association Rules and Sequencial Patterns* School of Computer Science, Telecommunication and Information Systems DePaul University Chicago Ill.

[10] Glover E. J., Tsioutsiouliklis K. Lawrence S. Pennocj D, Flake G. ,*Using web structure for classifying and describing web pages* Computer Scence Department Princeton University.

[11] Buchner A.G Baumgarten M. Anand S.S. ,*Navigation Pattern Discovery from Internet Data* Northen Irelan Knoledge Engineering Laboratory University of Ulser

[12] Batista P. Silva M. ,*Mining web access logs of an Online newspaper* Departamento de Informatica Universidad de Lisboa

[13] Cooley R. Mobasher B. Srivastava J. ,*Data preparation for Mining WWW Broswsing patterns* Department of Computer Science and Engineering University of Minnesota.

[14] Nadjarabshi-Noghani M. Chorbani A. *Improving the referrer-based web log session reconstruction* Faculty of Computer Science University of New Brusnswick, Canada

[15] Berendt B. Mobasher B. Nakagawa M. Spiliopoulou M *The impact of Site Structure and User Enviroment on Session Reconstruction in Web Usage Analysis* Humblod University Berlin

[16] Fu Y. Shih M. Creado M. Ju C *Reorganizing Web Site Based on User Access Patterns* Department of Computer Science University of Missouri-Rolla

[17] Cooley Walker Robert, *Web Usage Mining: Dicovery and Application of Interesting Patterns from Web Data PhD Thesis* , University of Minnesota.

[18] Ypma, A, Heskes T. *Categorization of Web pages and user clustering with mixtures of hidden Markov models* SSN Nijimegen, Geet Grooteplein The Netherlands

[19] Saruka R. *Link Prediction and Path Analysis Using Markov Chains* Yahoo Inc.