UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

ANÁLISIS DEL COMPORTAMIENTO DEL USUARIO EN LA WEB PARA OPTIMIZAR LA ESTRUCTURA DE NAVEGACIÓN DE UN SITIO USANDO ALGORITMOS GENÉTICOS.

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

EVELYN PAMELA ANDAUR ESPINOZA

PROFESOR GUÍA
JUAN D. VELÁSQUEZ SILVA

MIEMBROS DE LA COMISIÓN
PABLO E. ROMÁN ASENJO
SEBASTIÁN A. RÍOS PÉREZ

SANTIAGO DE CHILE
ABRIL 2010

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL
POR: EVELYN ANDAUR ESPINOZA
FECHA: 05/04/2010
PROF. GUIA: JUAN D. VELÁSQUEZ S.

Resumen

El objetivo general del presente trabajo de título, es aplicar algoritmos genéticos en el análisis del comportamiento de navegación de los usuarios en un sitio web, para la extracción de información y conocimiento que luego permita la mejora de la estructura de hyperlinks de un sitio.

El gran crecimiento de Internet, ha llevado a que la mayoría de los negocios tengan su propio sitio web, por lo que la competencia por acaparar más visitas se ha intensificado. Lo anterior ha obligado a las empresas a mejorar continuamente el contenido y estructura de sus sitios, creando, de esta forma, más valor para sus visitantes, lo que luego se traduce en la captura de nuevos compradores. La clave no es sólo tener un portal para mostrar algo al mundo, sino que ir más allá, y adaptarse a los usuarios. Por otro lado, muchas veces navegar y encontrar lo deseado se vuelve un proceder bastante engorroso, pues puede haber un exceso de hyperlinks que pierda al usuario, o una cantidad muy baja que no le permita flexibilidad.

Así, para comenzar esta investigación, se estudió la literatura relacionada al modelamiento de la Web, la aplicación de AG en áreas del web mining y mecanismos de limpieza de registros web.

En base a lo anterior, se desarrollaron tres modelos genéricos para representar los pesos de los hyperlinks, tanto existentes como no existentes, declarando la importancia numérica de mantenerlos o crearlos respectivamente. Considerando los supuestos de que a mayor tiempo de permanencia en una página de la misma sesión, mayor es el interés que el usuario manifiesta en su contenido, y que la frecuencia de uso de las interrelaciones indica una mayor utilidad de las mismas para acceder hacia la información deseada, es que la hipótesis señala que es posible encontrar estructuras de navegación más eficientes, desde el punto vista de la navegabilidad y usabilidad, a partir del comportamiento de los usuarios en el sitio actual.

Cada modelo se basó en una misma esencia probabilística, pero diferentes expresiones y principios matemáticos. El objetivo es realizar los ajustes necesarios a cada uno de ellos para obtener los mejores resultados y poder compararlos entre sí. Al proceder de ese modo, se pudo explorar con mayor amplitud diversas formas de creación de la utilidad de los hyperlinks. Se destaca el concepto de acceso objetivo aquí propuesto, un análisis prospectivo que fija una página y estudia las que le siguen en cada sesión.

Los modelos posteriormente fueron aplicados al mismo sitio web real, correspondiente al del Magíster de Gestión de Operaciones de la Universidad de Chile, el cual logró ser mejorado en utilidad en un 51%, 63% y 29% por los modelos 1, 2 y 3 respectivamente. Para obtener el mejor modelo y su ajuste, se estudia el mencionado porcentaje de mejora de la utilidad, las estadísticas de los resultados obtenidos, índice de navegabilidad y una encuesta de percepción ante los cambios propuestos. Se destacan las mejoras resultantes del modelo 2, el cual se basa en la implementación de acceso objetivo y la construcción de una utilidad lineal.

En conclusión, es posible mejorar la estructura de navegación web usando una adecuada medida del peso de los hyperlinks y la búsqueda optimizada propia de los algoritmos genéticos. Con ello, el identificar la utilidad de los mismos, genera un gran campo de estudio y acción para facilitar el acceso hacia la información requerida en un sitio web. Por otro lado, la aplicación de algoritmos genéticos en estudios de web mining permitió comparar diversas expresiones con relativa facilidad, lo que los sentencia como mecanismos exploratorios efectivos en este campo. Estos resultados en definitiva, van mucho más allá e impactan fuertemente áreas de marketing y tecnologías de información de las empresas, dado el afán de mejorar la experiencia de navegación y responder a las necesidades de los clientes en sus sitios corporativos. Con ello, es posible ofrecer la presente investigación incluso, como estudio de consultoría para la mejora concreta de sitios corporativos que registren el accionar de sus usuarios.

Se recomienda en trabajos futuros, aplicar los modelos propuestos en sitio de mayor tamaño, estudiar otras expresiones de utilidad de los hyperlinks basadas en las aquí expuestas, establecer nuevas restricciones de interés, incorporar procedimientos de web content mining y realizar un análisis automático en la detección de patrones de los principales cambios obtenidos.

Agradecimientos

"No dejes que el pasado te diga quién eres, sino que te diga, ***quien serás****".*

Con bastantes ansias esperaba este momento, momento de agradecer y dar término a una de las etapas más importantes de mi vida, que me permite inaugurar el comienzo de nuevos desafíos, metas y por supuesto, nuevos sueños. Agradezco a quienes tanto me apoyaron en todo este proceso, aquellos que de una u otra manera, me entregaron palabras, gestos, ayuda académica, emocional, etc.… a todos quienes me han permitido llegar a ser quien soy.

Gracias a mis padres, Virginia y William, por brindarme todas las facilidades para estudiar y salir adelante en mi carrera, por enseñarme el valor del esfuerzo y la humildad en los logros, por ser quienes me motivaban a seguir adelante e impulsaron en mí, el ser cada día mejor persona. Son ellos a quienes mucho les debo, y con este logro, este paso hacia delante, espero poder compensar de una u otra manera, todo cuanto han hecho por nosotros. Gracias a mi hermano William, por su paciencia y alegría, por pese a ciertas incomodidades fue capaz de comprender y apoyar todo este proceso, gracias por nuestras largas conversaciones desde el corazón. Gracias a Susana, quien con sus rezos intercedió tanto por mí en momentos difíciles, por ser mi amiga de toda la vida, por ser mi ventanita de mirada al cielo.

Mención aparte y claramente destacada tienen mis profesores. Gracias profesor Juan, mi maestro en todo sentido, imposible es poder escribir cuanto quisiera, pero gracias por guiarme en cada paso académico y espiritual, gracias por su tiempo, preocupación y sabiduría, gracias por enseñarme a creer ante todo en mi misma, en lo que puedo llegar a ser, en mis sueños imposibles. Gracias profesor Pablo, por su paciencia conmigo, por su perseverancia, por estar siempre disponible, por enseñarme tanto, por ser mi agenda, mi verdadero apoyo, y por cada semana constante de trabajo; simplemente, gracias por su entrega e interés por el presente trabajo y tirarme para arriba cuando lo necesitaba. Gracias profesor Sebastián, por estar siempre en el momento indicado, justo y preciso, por ayudarme a organizar y darle curso al presente informe de la mejor manera.

Gracias a mis compañeros del área de tecnología de la información de la Universidad de Chile y a mis pares memoristas, por sembrar en mí, grandes recuerdos y apoyo.

Gracias a Jorge V., por ser mi asesor personal en Matlab, a Eduardo, por enseñarme útiles herramientas, a Jorge T. por preocuparse de mis avances y como yo estaba en cada paso y a Alexis H., por apoyarme cuando más lo necesité y ayudarme a salir adelante en la dificultosa recta final. Solo grandes amigos se preocupan de que cada debilidad, se vuelva una fortaleza.

Gracias a mis amigos de Universidad, a grandes personas que llegué a conocer incluso desde mi primer año de ingreso, y con quienes compartimos las mismas etapas, avances, problemas y logros. Gracias por hacer crecer en mí verdadera amistad y cariño, por ser mis compañeros de batallas a lo largo de todos estos años, y por supuesto, salir victoriosos.

"El sueño de la biblioteca infinita se ha hecho realidad… De hecho, se estima que la pieza promedio de información en la Web hoy en día nunca será vista más que por su productor y sus amigos cercanos, y uno no puede ver más que un porcentaje minimal de todo lo que está publicado." [16]

Índice

Índice de Figuras

Índice de Tablas

Índice de Definiciones

Índice de Ecuaciones

1.- Introducción.

"…las empresas están aprendiendo a modificar sus sitios en la red para crear más valor para sus visitantes y convertirlos en compradores… todo en post de una mejor experiencia y lealtad…"[34].

Con dicha frase, mencionada bajo el contexto de cómo desarrollar competencias centrales hacia una ventaja competitiva sostenible, se muestra a modo inicial la importancia de crear valor en la experiencia en Internet, hoy por hoy una obligación para toda empresa pues le ayuda a neutralizar sus amenazas y explotar sus oportunidades. Ante eso, el saber cómo se comporta un cliente es la ambiciosa meta de todo estudio de marketing, es decir, poder predecir qué es lo que prefiere, los atributos que más valora, sus contenidos de interés, necesidades al momento de comprar, darle facilidad para acceder a su objetivo de búsqueda, etc. Con el auge de la tecnología en los estudios y manejo de datos de manera inteligente en la Web, aparece un nuevo desafío apoyado por Alejandro Zuzenberg, gerente de ventas y operaciones de Google Hispanoamérica: "*la importancia del Marketing en línea tiene, como requisito, adaptarse a cada tipo de usuario*" [1].

En el contexto expuesto, nace ambiciosamente la inquietud de poder construir un modelo de comportamiento del consumidor en la Web, donde mediante los registros web log (gracias a los cuales es posible obtener la ruta de navegación de un usuario), sea posible determinar la estructura web más adecuada. El objetivo entonces, es ser capaces de captar más clientes y aumentar la fidelidad hacia la empresa o institución representante a lo largo del tiempo, controlando los datos generados en la Web pues "*Internet puede destruir ganancias tan fácilmente como puede posibilitarlas*" [33].

1.1.- Antecedentes generales y motivación.

La importancia de crear valor en la experiencia en la Web conlleva una gran oportunidad para toda empresa, dicha experiencia va mas allá de la Web 2.0 (desarrollo web basado en comunidades de usuarios y gama especial de servicios en post de redes sociales [3][16], favoreciendo la colaboración y el intercambio ágil de información entre los usuarios), va hacia la capacidad de adaptarse a cada tipo de usuario en post de hacer más rápida y placentera su búsqueda. Ante eso, el saber cómo se comporta y que busca un cliente en la Web es el negocio de todo estudio de marketing y área tecnológica de una empresa, poder predecir que es lo que prefiere el usuario web de su sitio, los atributos que más valora, sus contenidos de interés, necesidades al momento de comprar, etc. En otras palabras: "*El conocimiento que las compañías desean, es como entender la tendencia de los consumidores en sus sitios web y cómo reaccionar para atraer a esos consumidores a comprar sus productos y servicios*" [37].

Gracias a las técnicas de web mining, es posible estudiar las sesiones de los usuarios (registros de navegación) y con ello, obtener las rutas más recurrentes y las páginas de mayor

[1] **Fuente:** LUN, Domingo 12 de Abril, página 10

interés, suponiendo que ésta variable es dada por un mayor tiempo de permanencia en ellas. Muchos métodos matemáticos de búsqueda y clasificación han sido aplicados para la minería de los registros web, con el fin de obtener información relevante de los mismos. Sin embargo, se hacen necesarias nuevas metodologías de manejo probabilístico, capaces de representar de forma fidedigna y ajustada el comportamiento del usuario en la Web. Un ejemplo de alto nivel en el manejo de metodologías, es el caso del buscador Google, donde queda claro que "*no basta con dar el primer paso en Internet… lo que importa es ser los mejores*" [33].

Por otro lado, el Departamento de Ingeniería Industrial (desde ahora DII), al igual que muchas instituciones, posee su página web representativa cuya dirección es *www.dii.uchile.cl*. En ella es posible acceder a información de toda la gama de programas de pregrado y postgrado (magíster y doctorados) del departamento, además de cursos de especialización y diplomados de post título. Permite además, conocer los proyectos e investigaciones de los académicos del DII, como también acceder directo a sitios relacionados con el departamento, tales como la biblioteca, Facultad de Ingeniería y Ciencias Físicas y Matemáticas, Universidad de Chile, etc. Es en este sitio, donde se procederá con el estudio propuesto, específicamente en el sub sitio asociado al Magíster de Gestión de Operaciones (desde ahora MGO) de dirección *http://www.dii.uchile.cl/~mgo2007/.*

1.2.- Descripción del proyecto.

Con el objetivo de determinar la mejor forma de estructurar un sitio web a través de la visualización común a un sin fin de clientes potenciales, se plantea la idea de cuantificar este comportamiento en base a parámetros del estudio, registros web log y uso de algoritmos genéticos (también indicados como AG) [27].

Los AG son llamados así, porque se inspiran en el proceso de selección natural planteada por Darwin [67] y su base genético-molecular cuyo objetivo es buscar dentro de un espacio de soluciones candidatas la mejor de ellas, representando una metáfora de como se cree que ciertos organismos celulares se han comportando a lo largo de todo el proceso de evolución. Es decir, se busca generar una serie de posibles soluciones al azar (en el presente caso, estructura web óptima) e ir creando sucesivas generaciones (soluciones hijas, mezcla de las anteriores posibles soluciones), dando más importancia y relevancia a las mejores, con el objetivo de buscar dentro de un espacio de hipótesis candidatas la mejor de ellas.

Este tema es de especial interés para las empresas poseedoras de una página web institucional y que están interesadas en consolidar sus transacciones por esta vía (sobre todo en la conversión de usuarios a compradores), lo que se vería favorecido en gran medida al determinar la mejor estructura Web. Para ello, comprender el uso de los AG, sus utilidades, sus procedimientos, alcances, etc., permitirá realizar una aplicación de éstos con los datos e interrelaciones web. En este caso particular, se trabajará con el sitio del Magíster de Gestión de Operaciones (MGO) del DII, y se establecerá una plataforma experimental para converger a la mejor estructura de él.

Se propone el uso de los AG porque son métodos de búsqueda dirigida basada en probabilidad. Se puede demostrar que el algoritmo converge en probabilidad al óptimo, es un método robusto y aplicable a diversas metodologías que van desde la optimización hasta la segmentación, son algoritmos altamente flexibles y exploratorios, y poseen un gran número de configuraciones distintas que se pueden utilizar según las necesidades de cada problema. Ante eso, se vuelve de considerable interés la aplicación de los AG en web mining, donde se posee una amplia gama de datos y espacios de soluciones posibles.

1.3.- Objetivos.

Para poder llevar a cabo el proyecto mencionado, se precisan los fines y metas a cumplir. El objetivo general es el siguiente:

> Aplicar algoritmos genéticos para mejorar la estructura de hyperlinks presente en un sitio web a partir del estudio del comportamiento del usuario mediante registros web log, a fin de hacer más eficiente la navegación de los usuarios en éste. En otras palabras, permitir a los usuarios encontrar el contenido que ellos buscan sin perder la usabilidad requerida en el sitio web.

1.3.1.- Objetivos específicos.

1.- Estudiar el estado del arte de los AG: historia, principios, teorías y aplicaciones.
2.- Revisar literatura asociada a modelamiento de la estructura web, uso de AG en web mining y técnicas de limpieza de registros web.
3.- Analizar descriptiva y analíticamente los registros web log del sitio *www.dii.uchile.cl*, enfocando el análisis en el conjunto de páginas correspondientes al magíster en gestión de operaciones, bajo el dominio *http://www.dii.uchile.cl/~mgo2007/*.
4.- Establecer supuestos y mecanismos de estandarización de los datos.
5.- Diseñar el modelo de adaptación entre la sesionalización del sitio, creación de una medida de importancia representativa de los registros web, y los parámetros y operadores genéticos del algoritmo genético.
6.- Aplicar el AG en la búsqueda de la estructura web óptima.
7.- Analizar los patrones de comportamiento de los usuarios obtenidos del sitio y su influencia en la organización estructural de una página web.
8.- Establecer las mejoras estructurales de un sitio web real.
9.- Concluir sobre el análisis de resultados y expansión de la investigación.

1.4.- Metodología.

La principal metodología dado el contexto de investigación del presente trabajo, se basa en los principios y etapas del método científico[2].

[2]**Fuente:** http://www.nasa.gov/audience/foreducators/plantgrowth/reference/Scientific_Method.html

Observación.

Estudio sobre el estado del arte de los AG, sus procedimientos, alcances y aplicaciones, para poder establecer el algoritmo adaptado a las consultas y manejo de vectores web log, donde se contiene la información de las sesiones de los usuarios, el tiempo en cada página del sitio, la secuencia de hyperlinks realizada, etc. El sitio web de estudio es *http://www.dii.uchile.cl/~mgo2007/*, el cual cuenta con diversos clientes de áreas de pregrado y postgrado, como potenciales estudiantes y estudiantes respectivamente. Además, como parte de la etapa de observación, se formará un marco conceptual en relación a las técnicas actuales sobre el estudio del consumidor y la importancia de éste en la actualidad.

Inducción.

Acción y efecto de extraer, a partir de determinadas observaciones y/o experiencias, el principio particular que poseen. Una vez estudiados los modelos y aplicaciones, se procederá a la confección del algoritmo y su aplicación desarrollada en Matlab [71]. Dicho programa posee un paquete de herramientas especializadas en AG, donde se realizarán las modificaciones y adaptaciones necesarias.

Hipótesis.

Se obtendrá una estructura mejorada del sitio *http://www.dii.uchile.cl/~mgo2007/*, en base a la creación de un modelo que cuantifique la importancia de cada hyperlink tanto existe como no existente, valorando el mantenerlo o crearlo respectivamente. Se podrá reflejar así, el uso del sitio por parte de los usuarios, aplicando AG para su búsqueda optimizada.

Supuestos:

- A mayor tiempo de permanencia en una página web, mayor interés en ella.
- Cada sesión iniciada tiene intenciones de búsqueda (es decir, usuario busca cierta(s) información(es), la cual fue preconcebida o bien, construida en el proceso de navegación).

Principios de una estructura web óptima:

- Maximizar el beneficio de la navegación, disminuyendo la demora en acceder a páginas de interés.
- Disminuir la cantidad de hyperlinks totales.

Experimentación.

Ejecutar el algoritmo ajustando el modelo, con los datos de entrada señalados, canalizando y analizando los resultados obtenidos,

Antítesis.

El uso de variaciones porcentuales en las funciones de utilidad, un índice de navegabilidad y el desarrollo de una encuesta de usabilidad, permitirá constatar la validez a las mejoras propuestas y los resultados de la cuantificación.

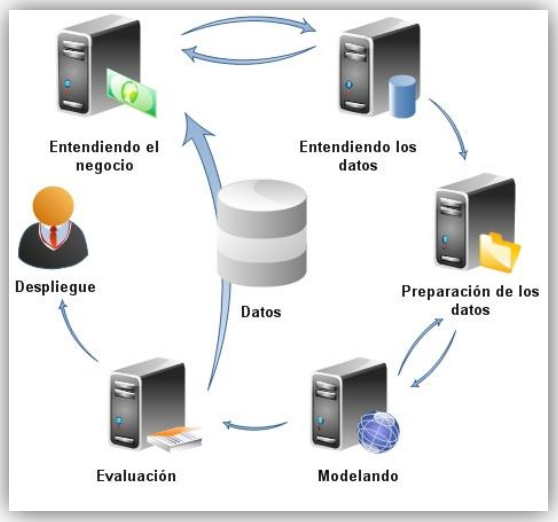
Tesis.

Conclusiones del trabajo realizado, enfocados en las ventajas que proporcionan los AG dentro de las técnicas de web mining y su aplicación concreta al estudio del comportamiento del consumidor.

1.4.1.- Metodología web mining.

CRISP-DM[3] es una metodología estándar y propia de la minería de datos, la cual ve la extracción de ellos como un proceso completo, contemplando la comunicación del problema de negocio a través de la recopilación, gestión y pre-procesamiento de datos, además de la construcción, evaluación e implementación del modelo [44]. Si bien está enfocada principalmente a estudiar temas de negocios, es perfectamente equivalente y ampliable a temas de investigación, considerando las etapas e interrelaciones que se muestran en la Figura 1.

Figura 1: Metodología CRISP-DM.



Fuente: http://www.crisp-dm.org/Process/index.htm

En el caso de la investigación, la etapa "entendiendo el negocio" consiste en ahondar en la problemática a enfrentar y estudiar la literatura relacionada, de modo de dar cuerpo a la hipótesis de trabajo. Se continúa "entendiendo los datos", donde se analiza que característica poseen los registros a ocupar, de donde provienen, como se encuentran almacenados, etc. Luego se preparan los datos, de modo de, considerando un diseño preliminar del modelo, establecer cuál será la mejor estructura de los mismos de manera de darles un orden y organización adecuada. Hasta aquí entonces, se ha cumplido con el proceso de extracción, almacenamiento y procesamiento de los datos. En la etapa "modelando" se da paso a construir el modelo, preparar un sistema para experimentar con él y ejecutar en concreto diversas pruebas. Cumplido lo anterior se da paso a la "evaluación" donde se interpretan los resultados y se realizan los ajustes necesarios. Con ello, se itera para finalmente obtener el despliegue final de resultados.

[3] **CRISP-DM**: Cross-Industry Standard Process for Data Mining

1.5.- Resultados esperados.

Llegar a la estructura óptima del sitio web del DII, mejorando sustancialmente la interacción entre sus páginas, al poseer una cantidad de hyperlinks adecuada en la posición y dirección de destino correcta. Ello será medido en términos simples, en base a las búsquedas mayoritarias de los usuarios (comportamiento). Se espera obtener importantes avances y metodologías de adaptación de este estudio en cualquier sitio web.

En términos específicos, los resultados esperados son:

1.- Diseñar un modelo mediante los web log que permita reflejar el comportamiento del usuario en la Web usando AG.

2.- Encontrar la estructura web óptima del sitio *http://www.dii.uchile.cl/~mgo2007/.*

3.- Establecer los mecanismos de expansión del modelo a otros sitios web, dada la validación del estudio en el sitio del MGO.

4.- Proponer una nueva metodología sobre el estudio del comportamiento del usuario web, mediante la aplicación de técnicas de web mining y AG en decisiones de marketing.

1.6.- Alcances.

Los objetivos del presente trabajo establecen la búsqueda de una cuantificación sobre el comportamiento de los usuarios web mediante ajustes de patrones, iterando sobre cambios y precisiones en el algoritmo para que esto suceda. La idea es determinar la estructura web óptima para que los usuarios se sientan satisfechos en su navegación, al minimizar el tiempo en que tardan en llegar a una página de interés. En otras palabras, establecer una disposición de hyperlinks entre páginas tal, que minimice la cantidad total de ellos y también el costo de tener que visitar muchas páginas para llegar a la deseada por la inexistencia de hyperlinks directos. Cabe recalcar que sólo se considera la estructura de hyperlinks (no contenidos), que no se realiza un contraste con otros algoritmos que puedan dar solución a los modelos propuestos y que únicamente se realiza una prueba unitaria de validación del modelo en un sitio concreto.

1.7.- Contribuciones

La investigación que a continuación será expuesta, fue plasmada en un paper, el cual fue presentado en el primer *workshop en business analytics and optimization BAO 2010.* Corresponde a un seminario realizado por el Instituto Sistemas Complejos de Ingeniería y auspiciado por Conicyt, Iniciativa Científica Milenio, Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, Instituto Chilenos de Investigación Operativa, ICHIO, la Asociación Latino-Iberoamericana de Investigación Operativa ALIO y el Doctorado Sistemas de Ingeniería de la Universidad de Chile. A continuación se expone el título y su abstract en sus versiones originales (inglés).

Best Web Site Structure for Users Based on a Genetic Algorithm Approach

Evelyn Andaur[1], Sebastián Ríos[2], Pablo Román[3*] and Juan Velásquez[4]
Department of Industrial Engineering, University of Chile, Santiago, Chile.
evelynandaur@gmail.com[1], *srios@dii.uchile.cl*[2], *proman@ing.uchile.cl*[3*], *jvelasqu@dii.uchile.cl*[4]

Abstract

A new approach is presented for the evaluation of the web site hyperlink structure, where a web user benefit utility function is proposed based on known assumptions. The aim is to suggest and improve the Web Structure according to a utility criterion. The frequencies of usage and time fraction to click on a hyperlink are used for constructing the web user utility. Researching probabilistic behaviors have allowed quality improvements of the hyperlink structure, using a genetic algorithm. The proposed methodology is tested on a real web site and the results are interpreted using the web user benefits by hyperlinks

2.- Marco Conceptual.

En el presente capítulo se dan los esbozos conceptuales necesarios para la construcción del modelo que se propone, realiza y prueba, con el fin de conocer las herramientas necesarias para su diseño, soporte y validación. Se comienza con una introducción a la Web para comprender su funcionamiento y elementos claves, para luego dar paso a las formas de descubrimiento y análisis de su información útil gracias a las técnicas de web mining. Así, los aspectos relevantes del comportamiento del usuario en la Web, sus representaciones posibles y estudios asociados, también serán reflejados. En la misma línea de lo expuesto, se hace finalmente referencia al funcionamiento de los AG, con un enfoque centrado en sus procedimientos, potencial e incursiones en el manejo de datos web.

2.1.- La Web

…"La Web es hoy un canal masivo para la difusión en todo el mundo y el intercambio de información, las compañías han respondido a estos desafíos de muchas formas de las cuales la más común ha sido la creación de sitios web corporativos que muestran la información de la compañía, como un tipo de 'tarjeta virtual de negocios'" [61].

Una gran revolución en las comunicaciones dado por el cambio en interacciones a distancias y de índole personal, ha provocado el advenimiento de la Web, cuyo impacto ha traspasado fronteras, tipos de usuarios e instituciones cobijadas en sus principios, cambiando el día a día y muchas veces las prácticas de toda sociedad inserta en esta gran red.

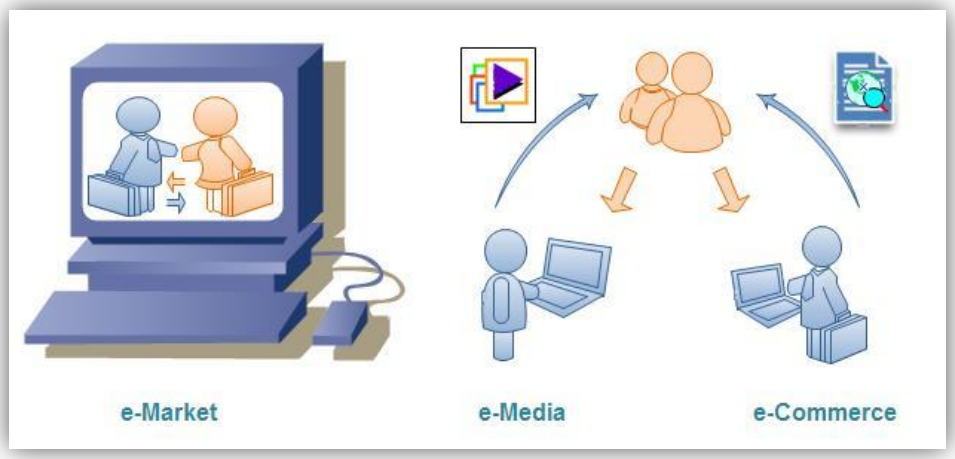
Tim Berners-Lee, investigador británico del Conseil Europeen pour la Reserche Nuclaire CERN (centro europeo para la investigación nuclear), fue quien desde 1989 lideró el desarrollo de la Web. Su idea inicial estaba basada en compartir hipertexto[4] [9] del modo de mejorar el control de la información acerca de aceleradores y experimentos del centro. Su planteamiento consistía en organizar la información usada como una estructura de grafo[5], donde los nodos correspondían a los documentos y los hyperlinks fueran las relaciones entre ellos [44]. Este amplio modo de representación (donde todo tipo de documento podía ser compartido) pasó a tener un desarrollo aun mayor hacia otras organizaciones, al momento establecer que los documentos podían ser distribuidos en una red de computadores (no solo acceso local), donde la red física necesaria para su implementación ya había sido desarrollada: Internet.

Comúnmente Web e Internet se usan como palabras con una denotación equivalente. Sin embargo, poseen diferencias técnicas importantes de destacar. Internet tiene relación con la red física que conecta en base a protocolos de transmisión de datos, mientras que la Web tiene que ver con la arquitectura lógica de la información construida sobre dicha red física. Puede compararse la mencionada diferencia como la que hay entre cerebro y conocimiento, respectivamente [16].

[4] **Hipertexto:** texto que conduce a texto relacionado, ya sea como hipervínculos o referencia cruzada.
[5] **Ver sección** 2.1.3.- Organización y estructura de un sitio web

En la actualidad, nuevas tendencias están en constante desarrollo en base a dichas necesidades, como el nacimiento de la Web 2.0[6] y las redes sociales, la Web 3.0[7], el incremento en el uso e investigación de la Web por el lado de la semántica con el fin de determinar que significa cada dato [16][75], y el aumento de publicidad adaptada y negocios on line[8]. En este nuevo medio ambiente influenciado por la Web, la lucha por retener consumidores y atraer otros nuevos, ha llegado a ser un estudio decisivo para la supervivencia de las compañías en el mercado digital [61] naciendo nuevos escenarios y formas para hacer negocios, las cuales se representan en la Figura 2. Puede ser un e-market al ser un canal de encuentro para vendedores y compradores a través de mercados virtuales con interacciones en él de manera no presencial, puede ser un medio de distribución de contenidos y objetos multimedia: e-media [51], o bien, ser un e-commerce al construir un medio de publicidad accesible y de alcance global de manera de llegar a ser un verdadero canal de ventas de productos y/o servicios.

Figura 2: Esquema e-market, e-media & e-commerce



Fuente: Elaboración propia

Como negocios web, hoy en día es posible encontrar diferentes tipos como B2B (Business to Business, es decir, entre instituciones), B2C (Business to Consumer, desde compañía a usuarios), P2P (Peer to Peer, entre pares) y B2G (Business to Goverment, relación de negocios hacia el gobierno). Cualquier variación en los modelos mencionados requiere de portales que puedan modificar su estructura y contenido en base a los requerimientos de los usuarios [61]. Dichos requerimientos se basan en diversas necesidades. Por un lado se quiere aprender acerca de algo con el fin de adquirir cierta información (necesidad de información), ir a un sitio en particular (necesidad navegacional), querer completar cierta actividad en concreto (necesidad transaccional) [10] o bien, realizar una búsqueda más exploratoria, mezcla de las anteriores (ambigüedades).

[6] **WEB 2.0:** Término para referirse a la segunda generación del desarrollo web, basada en comunidades de usuarios, redes sociales, blogs, sitios colaborativos (wikis), etc., hacia un intercambio de información ágil y colaborativo.

[7] **WEB 3.0:** Término para referirse a la tercera generación del desarrollo web, con la influencia de la inteligencia artificial, web semántica o sistemas 3D, para una manipulación de datos más eficiente al convertir la red en una gran base de datos.

[8] **On line:** Interacciones que se realizan al estar en la Web.

Así, la Web aun está en un crecimiento de rapidez considerable incluyendo e incluso superando a otros medios de comunicación y formas de hacer negocio. Cada vez se realizan estudios más avanzados, con la gran ambición de tener más usuarios y/o clientes.

2.1.1.- Funcionamiento de la Web

¿Qué es lo que permite que dos computadores ubicados en lugares de enorme distancia puedan transmitir, compartir y/o ver la misma información? La respuesta va más allá de la red física que hace esto posible, involucra tres pilares básicos en los que se sustenta la Web [16]: identificadores únicos, lenguaje universal y protocolo de transmisión de datos.

1.- Identificadores únicos

Para poder referenciar y describir objetos, son necesarios identificadores únicos para los mismos. En este sentido aparece la URL[9] como un tipo elemental de los mismos como la localización universal de recursos. En otras palabras, la URL corresponde a la dirección de un objeto en la Web, llamados comúnmente páginas web.

Un URL tiene tres elementos: protocolo, dominio y ruta [44] dispuestos de la forma:

<protocolo>://<dominio>/<ruta-archivo>

<protocolo> corresponde al lenguaje para intercambiar información entre navegador y servidor (Protocolos como: HTTP, FTP[10], etc.), <dominio> es el nombre del servidor web reflejado en el nombre corporativo de acceso, y <ruta-archivo> involucra el camino a seguir para acceder al archivo almacenado en el servidor. Por ejemplo: *http://www.dii.uchile.cl/ index.html*, accede al archivo "index.html" en el servidor "www.dii.uchile.cl". Dicha URL también puede ser escrita sin "index.html" (http://www.dii.uchile.cl/) dado que el navegador automáticamente busca el archivo index si solo el dominio es identificado. Para el caso *http://www.dii.uchile.cl/mba/paginas/profesores.html*, ocurre que se accede al archivo profesores.html ubicado dentro de la carpeta "paginas", que a su vez se encuentra en la carpeta "mba" en el servidor www.dii.uchile.cl.

2.- Lenguaje universal

Para una comunicación universal, es necesario un lenguaje único [16]. Este es HTML[11], lenguaje de marcado cuya principal característica es su simple uso, entendimiento y organización. Permite especificar la disposición y tipo de letra, integrar diagramas y crear hipervínculos (links), el cual, en HTML, se indica como una etiqueta de anclaje ("<", ">") también llamadas "tags", con un atributo href [15], de la forma:

<a href="http:// www.dii.uchile.cl/">

[9] **URL:** Universe Resource Locator
[10] **FTP:** File Transfer Protocol
[11] **HTML:** Hyper Text Markup Language

3.- Protocolo de transmisión de datos

Ante la gran red en constante crecimiento, el envío de información, ha de poseer normas y prácticas reconocidas y respetadas por los involucrados. Ante eso, es que se ha consolidado el protocolo HTTP[12], encargado del transporte de hipertexto para enviar HTML y otros datos en Internet a computadores-clientes, quien accede a ellos mediante un navegador. HTTP se basa en la parte superior del control de protocolo de transporte (TCP[13]), la cual proporciona datos fiables de las corrientes de información para ser transmitida desde un ordenador a otro a través de Internet [15].

Comprendiendo lo anterior, la cadena de eventos que permiten el funcionamiento de la Web se basa en el paradigma de cliente-servidor [61], donde un computador cliente mediante el navegador ejecuta aplicaciones a modo de validar y procesar los requerimientos antes de ser enviados al servidor, el cual realiza las actualizaciones necesarias y envía la información solicitada. Se define:

> **Navegador:** Aplicación en el computador cliente que permite navegar en la Web, proceso cíclico que comprende la realización de peticiones (tanto a sitios Web como a objetos en particular) y el acceso a las mismas. Ejemplo: Mozilla Firefox, Internet Explorer, Crome, Opera, Safari, etc.
> **Servidor:** Aplicación en el computador servidor cuyo input son peticiones recibidas por la red, y cuyo output es enviado por la red al cliente que envió la petición. Dichas peticiones/respuestas están formalizadas en base al Protocolo de transmisión. Ejemplo: Apache, Tomcat, Caucho Resin, PWS, etc.

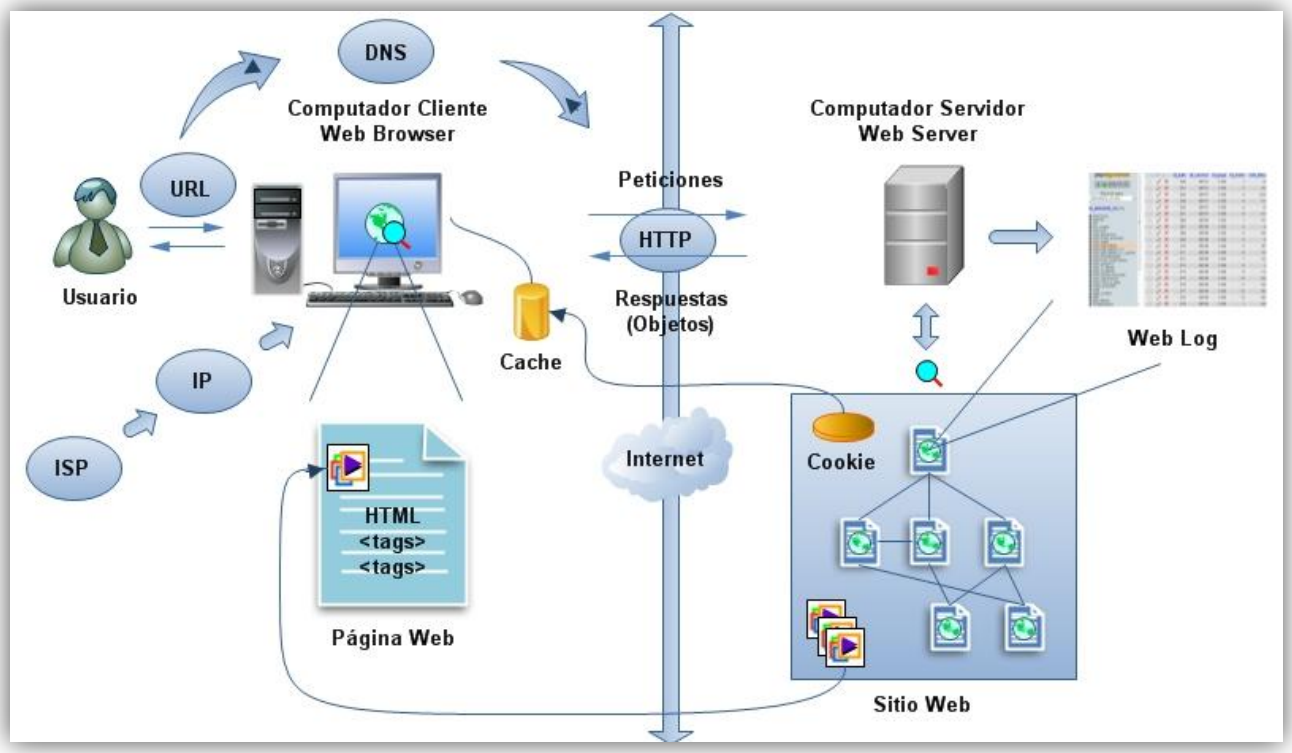
Se destaca en este contexto el concepto de **puerto**[14], asociado a la interfaz necesaria por el protocolo para establecer un camino de comunicación necesario para poder acceder a un programa a través de la red, de modo que diferentes tipos de datos puedan ser enviados y recibidos mediante él. Por ejemplo, el navegador se conecta al puerto 80 del servidor web al momento de solicitar una página, si este número de puerto no se supiera o estuviera bloqueado no podría acceder a ella[15].

La Figura 3, indica las interrelaciones y elementos claves de dicho funcionamiento, basado en el paradigma cliente/servidor [61] donde un cliente solicita servicios y/o aplicaciones a un servidor web quien busca y responde dichas peticiones. La URL es usada por los navegadores web para solicitar documentos desde los servidores e hyperlinks como una referencia hacia otros documentos [44]. En definitiva, para acceder a una página deseada, el usuario habrá de introducir la dirección URL de la misma en el navegador, o bien, hacer clic en un hyperlink que lo dirija a ella. El clic se traduce por el navegador en una solicitud de red en busca de la página de destino utilizando HTTP [15]. Una vez realizada dicha petición (por URL o hyperlink) el servidor devuelve el código HTML de la página, cuyos objetos

[12] **HTTP:** Hyper Text Transfer Protocol
[13] **TCP:** Transport Control Protocol
[14] **Fuente:** http://www.ordenadores-y-portatiles.com/puertos-ip.html
[15] **Fuente:** http://www.sitiosargentina.com.ar/webmaster/cursos%20y%20tutoriales/puerto.htm

(archivos multimedia, de texto, etc.) están identificados por <tags>, los cuales indican cómo debe ser mostrada la página. Así, el navegador pide al servidor los objetos necesarios para completar su despliegue, es decir, se genera más de una petición desde el navegador al servidor web [51][61]. Con lo anterior, logra entregarse el sitio web solicitado, donde ciertas cookies del mismo[16] brindadas por el servidor, se almacenan por ciertos períodos de tiempo en la memoria caché del computador cliente de manera de garantizar servicios más personalizados. Todas las solicitudes realizadas para poder ver el sitio se registran en los archivos de bitácora del servidor, los web log. Gracias a ellos es posible tener la ruta de navegación (qué páginas visitó) del usuario y el tiempo invertido en cada visita.

Figura 3: Funcionamiento de la Web



Fuente: Elaboración propia basada en [61]

2.1.2.- Principios de la Web

El gran principio de la Web puede resumirse en la siguiente frase: "*Todos pueden publicar, todos pueden leer, nadie puede restringir"* [16] Así, se sustenta su estructura, protocolos, lenguajes, documentación, etc. Para mantenerlo, es que se creó el **Consorcio de la Web (W3C)** [74], organización internacional que se propuso impulsar la interoperabilidad y evolutividad de la Web, para un acceso de la misma información desde diversos dispositivos independiente del software de uso, además de la aplicación de estándares en la codificación de caracteres para la universalidad, uniformidad y unicidad en la transmisión de datos (**UNICODE**[17]). En otras palabras, el principio de la Web se basa en "ser, estar, servir". Ello,

[16] **Ver sección**: 2.3.3.- Proceso de sesionalización

[17] **UNICODE:** Codificación estándar de caracteres diseñado para facilitar el tratamiento informático, transmisión y visualización de los mismos, de modo que sean universalmente entendidos, uniformes y únicos.

relacionado a los principales aspectos técnicos de la misma tiene que ver con su arquitectura, ubicuidad y usabilidad respectivamente.

La **arquitectura** tiene relación con la disposición de componentes que le dan sustento, de modo que cualquier computador conectado a Internet pueda conectarse a un servidor identificado por la URL solicitada. Parte de dicho proceso lo realiza el ISP[18], empresa encargada de brindar el servicio de acceso a los datos web (ejemplos en Chile: Telefónica, VTR, Telmex, entre otras). Dicho servicio lo hace traduciendo los URL en direcciones IP, como también brindando un número del mismo tipo como identificador a los usuarios. Para esto, varios servidores de nombres DNS[19] se encargan de traducir el nombre de dominio a dirección IP y viceversa. Al realizar una petición para acceder a una página Web, ésta se envía al servidor DNS local del sistema operativo[20]. El sistema operativo, antes de establecer comunicación, comprueba si la respuesta se encuentra en la memoria caché[21] accediendo a ella, si no se encuentra, la petición se envía a uno o más servidores DNS, cuya información está replicada en un árbol n-ario, con más de una rama para llegar al mismo objetivo. La participación anteriormente descrita en el proceso Web del DNS, ISP, URL, IP y Protocolo puede apreciarse en la Figura 3.

Ubicuidad tiene que ver con el hecho de estar en la Web. Para poder encontrar y ver el sitio, se establecen subconceptos como facilidad de búsqueda y visibilidad respectivamente. Por otro lado, "*la **usabilidad** es un concepto que engloba una serie de métricas y métodos que buscan hacer que un sistema sea fácil de usar y de aprender"* [73] Tiene dos aspectos centrales: el contenido y la estética (la forma, el diseño gráfico). Ambos conceptos se relacionan e impactan entre sí, donde por ejemplo, la falta de una visibilidad adecuada afecta la usabilidad del sitio. En otras palabras, la usabilidad es el atributo cualitativo que indica cuán fácil es usar algún elemento específicamente en la Web, se refiere a cuán rápido los usuarios pueden aprender a navegar con ligereza en un sitio y cuán eficiente llega a ser su proceso de navegación [51].

2.1.3.- Organización y estructura de un sitio web

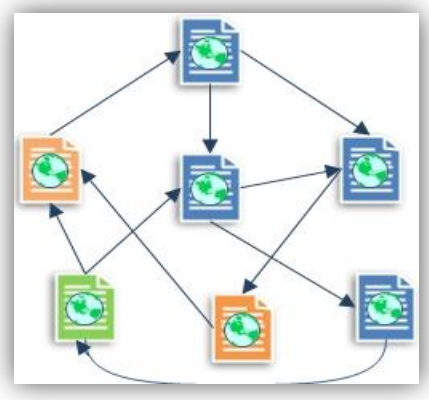
El análisis de **redes sociales** (disciplina de las ciencias sociales) junto a la **teoría de grafos** (disciplina matemática) ha permitido cuantificar los vínculos entre las personas que pertenecen a una red social analizando su estructura y sustento [16]. De acuerdo a los principios de la teoría de grafos, el análisis de redes sociales define a las personas como nodos y las relaciones entre éstas como arcos.

Gracias a ello es posible establecer y estudiar la estructura y organización de un sitio web, ajustando su representación a **grafos dirigidos** como se visualiza en la Figura 4, en el que los nodos serían las páginas (direcciones URL) y los arcos dirigidos los hyperlinks entre

[18] **ISP:** Internet Service Provider
[19] **DNS:** Domain Name System
[20] **Sistema operativo:** Software que permite controlar e interactuar con el sistema, controlando el hardware y dando soporte a otros programas. Ejemplos: Window, Mac OS, Linux, etc.
[21] **Caché:** Sistema especial de almacenamiento de alta velocidad, puede ser dependiente (área reservada en la memoria principal) o independiente (dispositivo de almacenamiento). Otorga acceso rápido a requerimientos reiterativos, basados en mismos datos o instrucciones ejecutadas.

ellas [15][32]. En la Figura 4 se representa un pequeño extracto de todo lo que sería la Web, donde los nodos de un mismo color indican un mismo sitio. Se desprende que la popularidad de una página Web puede ser determinada considerando el número de otras páginas que la apunten [61]. Grafo entonces alude a la representación con nodos y arcos mientras que su conceptualización como dirigido tiene relación con la precisión entre hyperlinks entrantes (in-links) o salientes (out-link) reflejando una dirección de movimiento y en una nivel más alto, la intención de búsqueda.

Figura 4: La Web como grafo dirigido



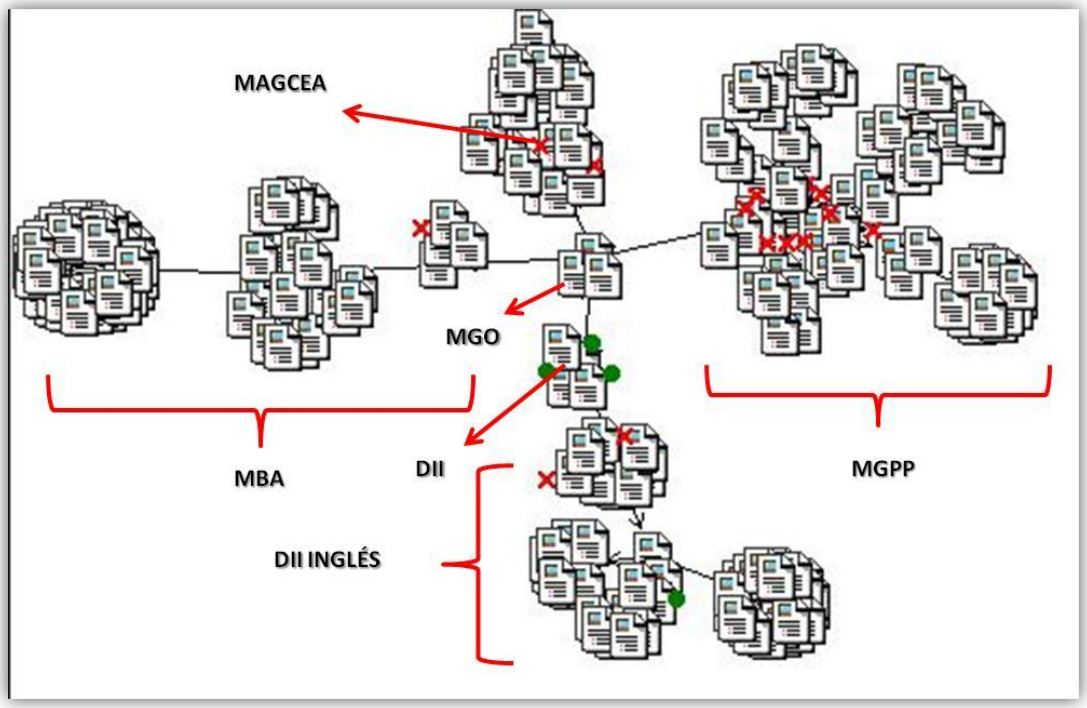
Fuente: Elaboración propia

Cuando una página web contiene un hyperlink apuntando a otra página diferente a la del sitio a la cual pertenece, suele ocurrir que de una u otra manera ambas están relacionadas por contenidos [61]. Además, es posible encontrar conjuntos de sitios con más conexión (hyperlinks) con ciertos sitios que con otros. En base a ello nace el concepto de **comunidad web** [51] correspondiente a aquellos sitios web que poseen más hyperlinks hacia los miembros de la comunidad. Dependiendo de la información contenida en las páginas ciertos nodos apuntarán a otros, y dependiendo de dichas interacciones las páginas tendrán distinta importancia en la comunidad [61]. Esto ha favorecido enormemente los motores de búsqueda web al poder identificar nichos de información común y relevancia implícita. Por otro lado, cada sitio web se construye basándose tanto contenidos como hyperlinks que relacionan, comunican y permiten navegar por sus componentes (páginas web).

Por otro lado, se puede observar la estructura de un sitio web real mediante un crawler[22], rastreador de estructura e interrelaciones. WebSPHINX[23] es un crawler de pequeña escala que mediante un algoritmo de búsqueda de grafos (moviéndose por los hyperlinks de cada página) logra desprender la estructura de un sitio web. Aplicando este software a www.dii.uchile.cl (DII) se obtiene un grafo aproximado que se observa en la Figura 5, cuyas páginas poseen la misma ruta en su URL, relacionada principalmente a la información de magíster como MAGCEA, MGO, MGPP, MBA, etc.

[22] **Ver sección** 2.1.3.- Organización y estructura de un sitio web
[23] **Fuente:** http://www.cs.cmu.edu/~rcm/websphinx/

Figura 5: Estructura web de www.dii.uchile.cl



Fuente: Elaboración propia gracias a estructura generada con WebSPHINX

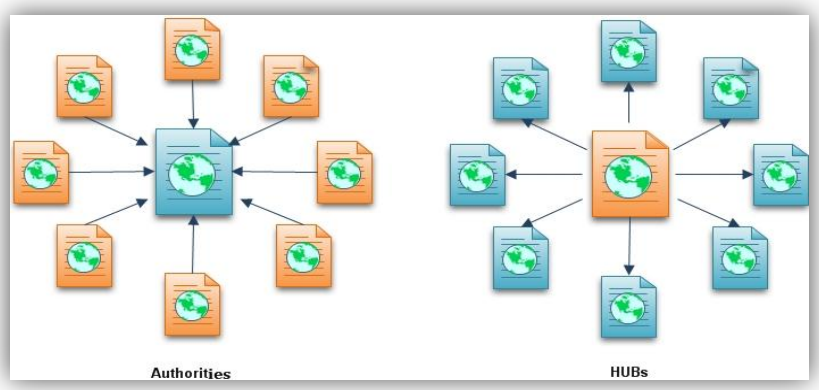
En dicho contexto se desprende la clasificación de páginas web en relación al contenido más relevante y la cantidad de hyperlinks hacia o desde ella, distinguiendo dos tipos [15][44][56][61]:

Authorities o páginas autoritativas:
Páginas apuntadas por muchas otras páginas, siendo un repositorio natural de información buscada por los usuarios. Ante eso su representación es un nodo recibiendo hyperlinks hacia él.

Hubs o páginas centrales:
Páginas que apuntan a muchas otras páginas, estudiarlas permite tener una colección completa de enlaces a páginas valiosas sobre algún tema. Su representación por el contrario, sería un nodo desde el cual parten muchos hyperlinks.

En la Figura 6, se representan ambos tipos de páginas ya sea como repositorios entrantes o salientes de información.

Figura 6: Páginas autoritativas y hubs



Fuente: Elaboración propia

En la línea de **estructura de redes** [51] y su representación vía grafos, es que se hace necesario poder determinar la importancia de un nodo ante los demás, para lo cual algunas medidas son [16]:

1.- **Centralidad:** Cantidad de aristas que conecten a un nodo con el conjunto restante, donde aquel más conectado será más central.

2.- **Cercanía:** Medida de suma de los arcos que conectan un nodo con los demás, reflejado en la capacidad para llegar en pocos pasos a cualquier otro.

3.- **Intermediación:** Número de veces que un nodo aparece en el camino mínimo entre otros dos nodos, es decir, su capacidad de conector para que la red se mantenga unida.

Como característica propia del grafo, usando una analogía geométrica, se la llamará **diámetro** del mismo a la máxima distancia entre dos nodos no conectados directamente entre sí. Un experimento en este contexto[24] se llevó a cabo a mediados de 1960 para estimar el diámetro dentro de EEUU, viendo como un grafo la red de comunicaciones posibles vía correo físico entre distintas personas. Así, los destinatarios de correo eran los nodos y la distancia correspondía al hecho de que la carta le llegara a un destinatario totalmente desconocido fijando un nodo (donde cada uno enviaba un mensaje a un nodo-persona conocida). El valor del diámetro resultó ser de valor 6.

En relación al diámetro de la Web, se hace necesario estudiar la topología del grafo, determinando su conectividad y como efectivamente es posible encontrar la información. S. Lawrence y Giles en [41], estudio de 1998, estimaron el tamaño del grafo web en $8x10^8$ documentos, los cuales estaban en permanente cambio en contenidos, cantidad e interrelaciones. El desafío de poder obtener un completo mapa de la Web ha desencadenado una serie de estudios, donde se destaca [52], el cual mediante la búsqueda de la **ruta más corta** entre dos documentos (definida como el menor número de páginas que deben ser seguidas mediantes sus hyperlinks para llegar de un documento a otro) determinaron su equivalencia con el concepto de diámetro de la Web (*d*), estableciendo la Ecuación 1.

[24] **Fuente:** http://redalyc.uaemex.mx/redalyc/pdf/282/28210402.pdf

Ecuación 1: Diámetro de la Web

$$d = 0{,}35 + 2{,}06 \cdot \log(N)$$

Contemplando el número estimado de documentos web (*N* de $8x10^8$) el diámetro resultó ser de que d_{web}=19, es decir, dos documentos escogidos aleatoriamente en la Web, tienen en promedio 19 clicks entre ellos. Cabe destacar el pequeño diámetro encontrado lo que significa que toda la información está disponible en solo unos pocos clicks.

El **grafo web** está evolucionando constantemente, debido a que sus nodos y arcos son creados y borrados todo el tiempo. Su estructura ha surgido sin dirección guiada, debido a una arquitectura web sin restricciones de publicación, autoría ni hyperlinks. En datos del año 2004, el tamaño de la Web ya pudo ser establecida con una cantidad por sobre los $5{,}2x10^9$ páginas [23]. Ante lo anterior, surgen dos problemas a estudiar: el tamaño y la rápida evolución del grafo web [32].

Un modelo básico para la Web está dada por el grafo $G = (P, L)$ donde P corresponde al conjunto de páginas $P=\{p_1, p_2, \ldots, p_n\}$ siendo n el número total de ellas y L, representa el conjunto de hyperlinks $L= \{l_1, l_2, \ldots, l_m\}$ en un grafo con m hyperlinks totales [61]. Es decir, G representa la estructura web en término de sus páginas y sus hyperlinks. Un camino aleatorio (**random walk**) a través de G es un **proceso estocástico**[25] que visita iterativamente los vértices del grafo G [2], con igual probabilidad de ir a uno u otro, donde:

$$\{ X_t \}_{t \in S} => \forall\ t \in T, X_t \text{ es una variable aleatoria}^{26} \text{ con valores en } S.$$

Puede verse este proceso estocástico en la Web como un conjunto de variables aleatorias, donde $t \in T$ representan los instantes de tiempo y S corresponde al conjunto de estados posibles, en este caso, en el grafo G. De ese modo X_t puede interpretarse como un estado en el grafo en el instante t, el cual visto desde un instante "anterior" a t, es una variable aleatoria [12].

Otro estudio acerca del modelamiento de la Web indica una diferenciación entre una estructura local y global de la Web, la primera como detección de intereses en una comunidad web (Estructura Microscópica) y la segunda como un amplio set de páginas conectadas entre ellas (Estructura Macroscópica) [32].

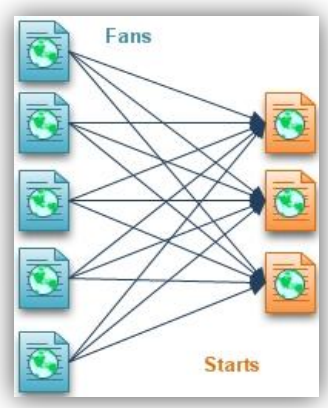
Estructura microscópica:

Se puede observar que hay ciertas páginas tipo A a las que llegan muchos links desde páginas tipo B, donde éstas últimas se dirigen exclusivamente a las primeras. A las páginas tipo A se les llamará *fans*, mientras que a las B, aquellas que reciben todas las conexiones, s*tarts*. La estructura se denomina *Bipartite clique*, cuya representación puede observarse en la Figura 7, que en su esencia es una comunidad centrada en un tópico de interés.

[25] **Proceso estocástico:** Modelos aleatorios presentados con funciones aleatorias cuyo argumento es el tiempo.

[26] **Variable aleatoria:** Su valor numérico está determinado por el resultado del experimento aleatorio

Figura 7: Bipartite clique

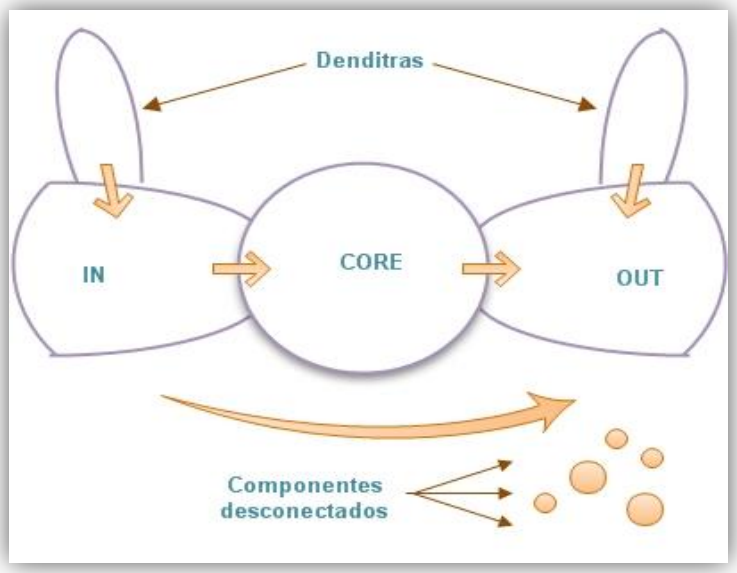


Fuente: Elaboración propia

Estructura macroscópica:

Una estructura que refleja una composición de nodos fuertemente conectados, conteniendo a la mayoría de ellos es *Box-Tie* (Figura 8). Esta estructura tiene cuatro tipos de componentes: core, IN, OUT y dentritas. Core corresponde al nodo más conectado, IN a las páginas desde las cuales se puede acceder al core pero sin ser accesibles desde él, OUT será el set de páginas accesibles desde el core pero que no van a él, mientras que, finalmente, las dendritas son las páginas accesibles desde uno de los tipos de los tres componentes anteriores o acceden a los mismos, sin pertenecer a dichas clasificaciones previas.

Figura 8: Estructura Box Tie



Fuente: [32]

2.1.4.- Clasificación de sitios web

A continuación se propone una serie de clasificaciones útiles para el posterior estudio de un sitio web en particular[27]:

Por su Audiencia
Públicos: Accesible por todo público, sin restricciones de acceso.
Extranet: Limitados a que usuarios accedan a ellos (proveedores, clientes particulares, etc.).
Intranet: Restricción total a una empresa u organización, como redes privadas en general.

Por su dinamismo
Sitios interactivos: Usuario puede influir en el contenido y formas del sitio, es decir, son distintas para cada usuario quienes tienen una participación activa y están ante un sitio personalizado para ellos.
Sitios estáticos: Contenidos y formas a cargo de diseñadores, usuario son pasivos de las modificaciones realizadas.

Por su estructura
Lineal: Sitio organizado en base a secuencias de pasos a seguir, orden cronológico.
Parrilla: Cuenta con páginas asociadas en categorías.
Jerarquía: Contiene una página de inicio que permite ir a las demás páginas ramificadas y asociadas en base a ella y relaciones mutuas.
Web pura: Información con pocas restricciones, hay una libre circulación de ideas.
Web mixta: Mezcla entre estructura jerárquica y web pura, contrapesando las restricciones de la primera y el exceso de información y libertad de la segunda.

Por su apertura
Estructura abierta: Todas las páginas y objetos del sitio están disponibles y pueden accederse desde cualquier punto del sitio web.
Estructura cerrada: Limita acceso a ciertas páginas (incluso solo una página puede llegar a ser visible para todos), en general consideran un registro previo para acceder a las restantes.
Estructura semicerrada: Mezcla de tipos anteriores. Contienen páginas principales y de entrada a diversas secciones.

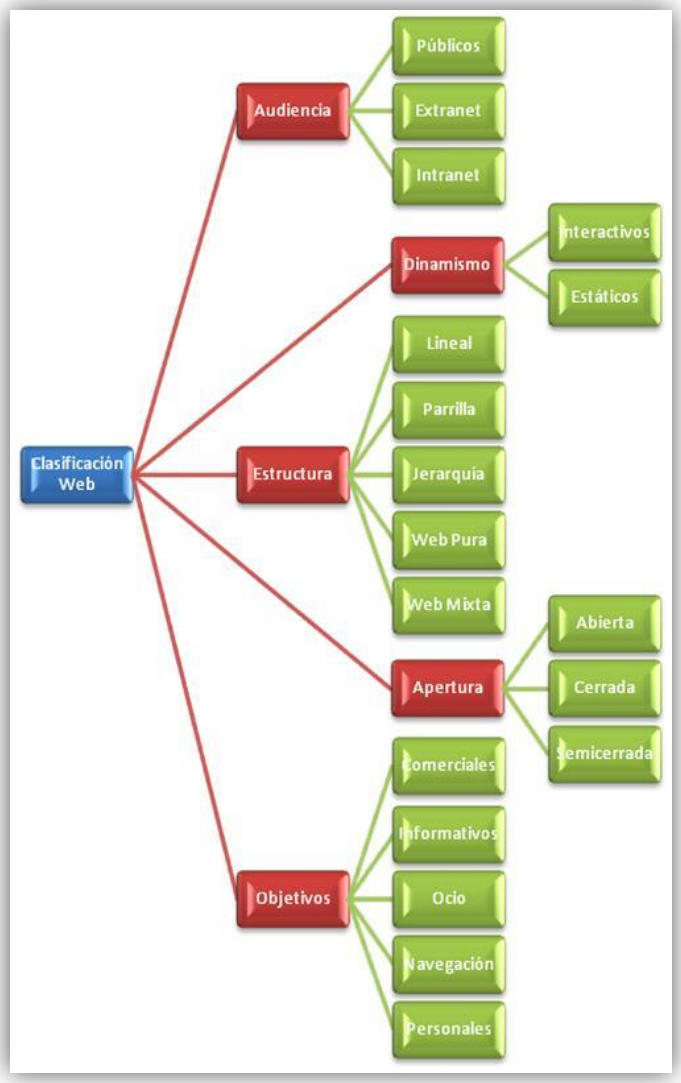
Por sus objetivos
Comerciales: Buscan promocionar los negocios de una empresa con fines económicos. Visitados por clientes, inversores, empleados, competencia y medios de comunicación. Aquí se encuentran: c*orporativos,* los cuales informan sobre la empresa y p*romocionales,* encargados de promocionar productos y/o servicios.
Informativos: Buscan distribuir información.
Ocio: En general tienen finalidad económica pero principalmente buscan sorprender al usuario.
Navegación: Ayudar al usuario a encontrar lo que desea en la Web, como portales y buscadores.

[27] **Fuente**: http://www.lawebera.es/manuales/primeros-pasos/como-empezar/tipos-de-webs.php

Personales: Medios de expresión de su creador ajustado a sus estructuras y contenidos deseados.

Para sintetizar, se expone el detalle anteriormente expuesto en la Figura 9.

Figura 9: Clasificación sitios web



Fuente: Elaboración propia

2.2.- Web mining

"El análisis de la Web es más que una secuencia de clicks, es más que una conversión de tasas, es más que sólo números." [38].

En el último tiempo, los esfuerzos en la investigación de la Web han buscado que cada sitio logre posicionarse ante los demás y alcanzar la fidelidad de los usuarios, para con ello poder convertirlos en potenciales clientes o visitantes continuos. Ante eso, se requieren herramientas que permitan estudiar que han hecho los usuarios para determinar patrones de comportamiento, con el fin de hacer el sitio mucho más agradable según las necesidades del cliente. Aparece entonces el web mining, minería de la Web que pretende encontrar la información que necesita para poder satisfacer la experiencia de navegación de los usuarios [56], ello incluye el descubrimiento de datos, documentos y elementos multimedia, para su posterior análisis.

Data mining corresponde a la disciplina de análisis de un conjunto de datos diversos para encontrar relaciones entre ellos. Dichas metodologías, aplicadas a los datos web más modelos de la misma comprenden el web mining, proceso de descubrimiento de patrones en el contenido, estructura y uso web [15][44][56][61]. En otras palabras:

Data Mining + Datos Web = Web Mining.

Web mining es una disciplina por sí sola, al enfrentarse a datos web que poseen naturaleza propia: heterogéneos, variantes en el tiempo, con muchas dimensiones, sin clasificar, distribuidos, semiestructurados, etc. [1][62].

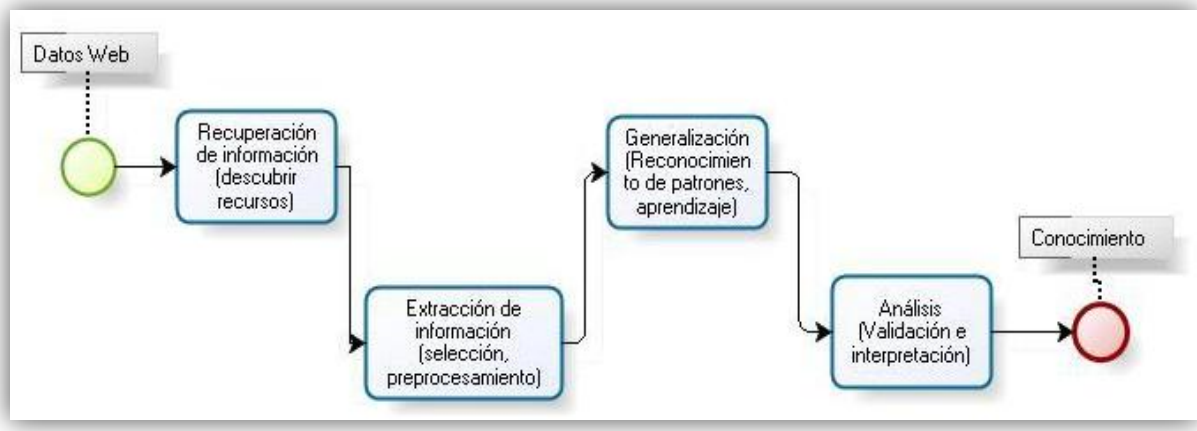
2.2.1.- Componentes y metodologías del web mining

El proceso del estudio web mining puede ser visto como un proceso de 4 tareas (Figura 10), donde la entrada son los datos web. Primero, se recuperan de manera que aquellos que sean irrelevantes no sean considerados, dicha extracción se hace desde bases de datos, páginas HTML o texto libre, es decir, desde fuentes altamente estructuradas, semiestructuradas o sin estructura respectivamente. Luego, se procede con la extracción de información de manera de encontrar un contexto semántico central que permita identificar automáticamente. Como tercera tarea se busca generalizar ciertos patrones encontrados, mediante técnicas de clasificación, clustering y/o reglas de asociación. Finalmente el enfoque es poder entender, visualizar e interpretar los patrones encontrados una vez acontecidas las tareas anteriores, para desembocar así, en conocimiento.

El web mining usa el contenido de los documentos, la estructura de hyperlinks y el uso estadístico para asistir a los usuarios con el fin de que encuentren la información requerida [56]. Así, tres aspectos fundamentales que definen los distintos tipos de datos web corresponden al Contenido (texto, imágenes, sonidos, videos, etc.), estructura (datos que

permiten determinar la estructura, como HTML, XML[28] al contener los hyperlinks) y Uso (datos de las preferencias del consumidor mientras navegan en un sitio web, donde se desprenden aquellos datos más específicos del usuario a veces recolectados, como nombre, intereses, etc.) [62].

Figura 10: Sub tareas del web mining



Fuente: Classification and Learning Using Genetic Algorithms [1]

Dependiendo entonces de los datos estudiados, se da pie a las tres áreas principales de clasificación de los estudios de web mining, representadas en la descripción jerárquica de la Figura 11.

Minería de contenido:
Proceso de análisis y extracción de información desde los contenidos de los documentos web a fin de encontrar cuáles palabras, oraciones y párrafos son más interesantes para los usuarios [51]. Las 2 principales estrategias en esta área corresponden a minar los contenidos web (*web page content mining*) o bien, mejorar el contenido gracias a máquinas de búsqueda (*search result mining*).

Minería de estructura:
Procesos para identificar patrones e interrelaciones en la organización del sitio web, como el ranking de páginas, relaciones de las mismas, popularidad de ellas, etc. mediante el estudio de los hyperlinks y considerando el sitio web como un grafo dirigido.

Minería de uso:
Proceso de obtención del significado de los datos e interpretaciones del comportamiento usando registros web log por ejemplo. Gran importancia cobra la personalización de los sitios web [60].

Las tres técnicas, en su aplicación siguen los siguientes procesos [8] ordenados por orden de ocurrencia: colección de datos, preparación de datos, minería de datos, presentación de los resultados, evaluación e interpretación de los resultados y acción sobre los resultados.

[28] **XML:** Extensible Markup Language. Reglas para codificar documentos electrónicamente, diseñando objetivos de simplicidad, generalidad y usabilidad en Internet.

Figura 11: Representación jerárquica de las áreas del web mining

Fuente: http://gamoreno.files.wordpress.com/2007/08/fa11.jpg

2.2.2.- Técnicas de web mining

Las técnicas de web mining nacieron como resultado de la aplicación de la teoría de data mining en datos web para el descubrimiento de patrones en los mismos [61]. Las técnicas aplicadas sobre todo en la representación del web content mining son [61]:

Clasificación:

El objetivo es asignar cada documento (página web) a una o más categorías existentes. Para ello, es necesario generar un aprendizaje previo de un conjunto de datos ya previamente clasificados, para luego clasificar otro conjunto de datos de la misma naturaleza. El proceso de entrenamiento itera hasta que el umbral de clasificaciones correctas se ha superado al comparar los resultados de la técnica en el primer conjunto de datos con la clasificación real esperada. Esta forma de proceder se le llama *aprendizaje supervisado*, al tener como entrada datos ya categorizados y luego expandir el aprendizaje generado desde ellos.

Clustering:

En base a medidas de semejanza o diferencia, se busca agrupar documentos sin tener clasificación a priori, es decir, se generará un *aprendizaje no supervisado*. Cada clase ha de poseer elementos homogéneos entre ellos, con la máxima heterogeneidad con las otras clases encontradas. La idea por lo tanto, es particionar el conjunto total de datos en clusters determinados por medidas de similitud. Las técnicas de clusters son:

- *Clustering particionado*: Divide el conjunto de elementos en subgrupos no vacíos y sin intersección, de modo de que cada elemento quede en uno de los clusters determinados.
- *Clustering jerárquico:* En base a métodos de unión o separación se busca la descomposición jerárquica. Procediendo de la primera forma llamada *aglomerativa*, se inicia considerando cada objeto como un cluster, para después ir agrupando mediante semejanzas y distancias. Según la segunda, conocida como *divisiva*, se asume el universo de datos como un único y gran cluster, para luego realizar separaciones en el mismo. En ambos caminos, se itera hasta satisfacer la condición de término.
- *Clustering basado en la densidad:* Considera el concepto físico de densidad y utilizan los conceptos de *densidad umbral, cardinalidad* y *radio.* Así, un elemento pertenecerá a un cluster *C* si la distancia (por ejemplo Euclideana) entre él y el centroide de *C* es menor al radio. El clustering se detiene cuando la cardinalidad del cada cluster haya superado la densidad umbral, sino, el centroide es redefinido y se vuelve a iterar.

Reglas de asociación:

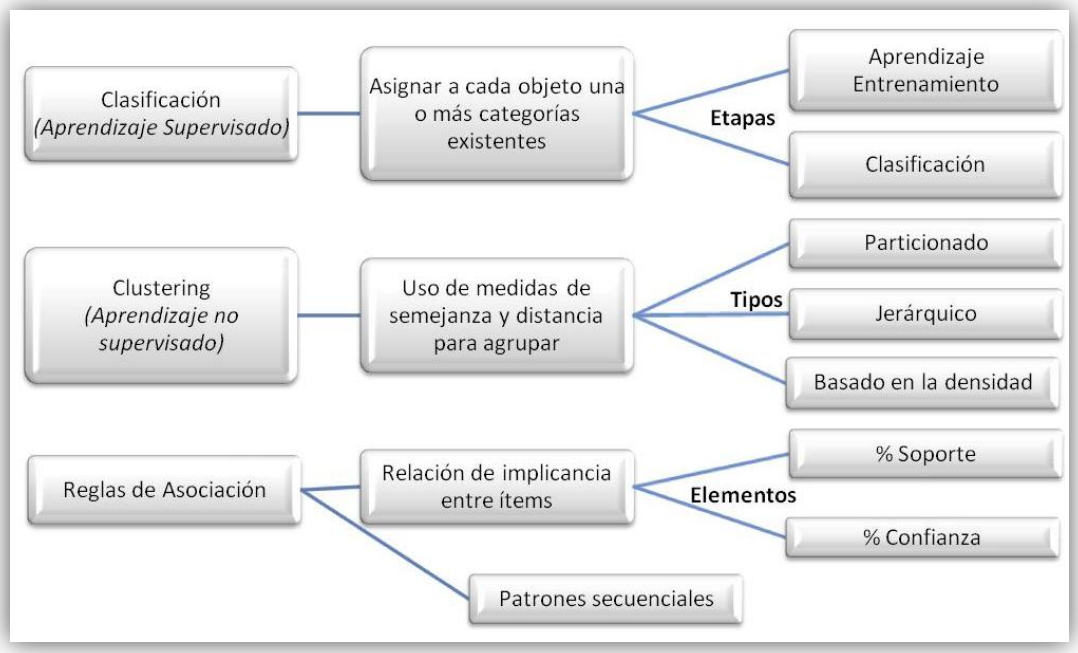
El fin es identificar asociación entre los documentos, en cuanto a correlaciones significativas entre subconjuntos de todo el set de datos. Su formalización viene dada por la expresión: *si* <*X*> *entonces* <*Y*> [61], donde *X* e *Y* corresponden a subconjuntos de ítems sin elementos comunes. El objetivo es desprender en base a las transacciones de ítems realizadas por los usuarios (conjunto de productos comprados) la relación de implicancia entre *X* e *Y*. Dos medidas claves de *X* => *Y* en un conjunto de transacciones *T*, es el soporte α que indica el porcentaje de *X* e *Y* en *T*, y el nivel de confianza β, que corresponde al porcentaje de *X* U *Y* en *T*, es decir, β% de las personas que compran *X* también compran *Y* en un α% de las transacciones totales.

Descubrimiento de patrones secuenciales:

Extensión de las reglas de asociación, cuyo fin es determinar patrones de co-ocurrencia, considerando una secuencia temporal, por ejemplo, un conjunto de páginas visitadas inmediatamente después de otra en numerosas ocasiones [44].

Un resumen de las técnicas anteriormente descritas es posible encontrarlo en la Figura 12, indicando etapas, tipos o elementos correspondientes.

Figura 12: Técnicas del web mining



Fuente: Elaboración propia

2.2.3.- Desafíos y limitaciones del web mining

El incremento dinámico y permanente de la Web, tanto en contenido como en estructura, ha inducido una serie de desafíos que enfrentan los estudios de ella aumentando cada vez más su complejidad. Habrá que superar la carencia de datos fiables, presencia de ruido, la naturaleza dinámica y transiente (estado temporal previo a la estabilización) de los datos, su ambigüedad semántica, requerimiento de personalización, alta redundancia, corrupción de los, entre otros [1]. Ante dichas características las limitaciones no son menores, las cuales van desde la subjetividad en la recuperación de información para determinar aquella verdaderamente relevante, hasta el enfrentar algoritmos a datos variantes en el tiempo, donde la naturaleza heterogénea de ellos llega a exigir métodos de minería de datos distintos por tipo de datos.

Considerando entonces la Web con datos distribuidos, altamente dimensionales y ruidosos, surgen los desafíos a los que el web mining se ha de enfrentar, como el uso de métodos mucho más resistentes y robustos ante datos impuros o fuera de norma (outliers[29]), además de entender, interpretar y visualizar los patrones y tendencias descubiertas durante la generalización de los mismos.

Así, teniendo en cuenta los mencionados desafíos del WM, es que la aplicación de *soft computing* ha desencadenado un gran interés, teniendo hoy en día un gran desarrollo y potencial[30], sobre todos en las diversas e incipientes áreas de aplicación del web mining [56].

[29] **Outliers:** Observación numéricamente distinta al resto de los datos
[30] **Ver sección** 2.5.2.- Origen y definición

2.3.- Procesamiento de datos web

Un web *crawler*, también llamado *spyder* o *robot*, es un programa que analiza cada página, recuperando e indexando[31] información desde ella. Identifica las relaciones (links) con otros documentos web, las cuales sigue para poder así, identificar todas las relaciones entre las páginas [23]. Los pasos iterativos que realiza un crawler podrían entonces indicarse como: acceso, recorrido, recuperación / indexación y salto a la nueva página. Así, puede desprenderse que el comportamiento de un crawler se rige por la estructura de hyperlinks de la Web [61] y ello determina la calidad de la información rastreada y recuperada. Son en definitiva, máquinas de búsqueda usadas incluso por buscadores ampliamente reconocidos como Google o Yahoo! los cuales recopilan las últimas versiones de las páginas y las indexan en sus servidores facilitando la búsqueda [44].

Texto plano, imágenes, hyperlinks, sonidos, videos, etc., son ejemplos de datos que corresponden a contenidos web, los que se vuelven inmensamente relevantes de estudiar, pues un mayor tiempo de permanencia en la página tendría relación directa con el interés que proporciona el contenido [51]. Para recuperar y analizar el contenido de texto plano, **Information Retrieval (IR)** [4] es la disciplina para extraer su información directamente desde el código HTML.

Así, una vez descrita la representación de la Web como grafo y rastreando su estructura y contenido mediante crawlers, para poder acceder a la información requerida es necesario proporcionarla mediante métodos de búsqueda, donde luego se generarán los datos propios de la navegación, para finalmente identificar la secuencia de páginas visitadas y el tiempo invertido en cada una de ellas.

2.3.1.-Algoritmos para rankear páginas

Debido a la amplia cantidad de información al momento de realizar una búsqueda, se vuelve inmensamente útil y necesario poder brindar respuesta de manera ordenada, identificando desde la mejor a la peor de ellas. Esto bien lo han comprendido los buscadores, por lo que han desarrollado e implementado métodos adecuados para proporcionar dicho orden deseado. La idea es poseer un ranking de páginas, lo cual es posible considerando algoritmos como HITS[32] o PageRank [61]. Ambos enfrentan dicho requerimiento en base al análisis de redes sociales y comunidades web[33] [15]. Un esquema que permite contrastar diferencias y similitudes entre ambos algoritmos se precisa en la Tabla 1. A modo general se describe:

[31] **Indexar:** Registrar ordenadamente información para elaborar un índice de manera de obtener resultados de manera mucho más rápida y relevante al momento de realizar una búsqueda.
[32] **HITS:** Hypertext Induced Topic Search
[33] **Ver sección** 2.1.3.- Organización y estructura de un sitio web

HITS:

Algoritmo desarrollado en IBM Almaden[34] por Jon Kleinberg alrededor de 1999. Combina dos factores claves en la determinación de importancia web, la relevancia basada en contenidos y la popularidad basada en la estructura de hyperlinks [44]. Con ello, luego de una consulta, HITS encuentra todas las páginas que contienen dicho texto, y las ordena de acuerdo a su importancia en la comunidad web. Para calcular la relevancia, las páginas son agrupadas y clasificadas como hubs o autoritativas, de modo de identificar las páginas que concentran hyperlinks a otras o bien, son repositorios de información con hyperlinks hacia ella respectivamente [61]. Genéricamente, dada una consulta *q*, los siguientes son los pasos claves en el algoritmo [44]:

1.- Detectar un pequeño set de páginas relevantes O_q llamado set origen.
2.- Expandir R_q considerando las páginas apuntadas por aquellas dentro del set origen y páginas que apuntan hacia el mismo, para así, obtener un set más amplio S_q llamado set secundario.
3.- Obtener la matriz de adyacencia[35] de S_q
4.- Analizar S_q para encontrar las páginas autoritativas y hubs

PageRank:

Algoritmo desarrollado en la Universidad de Stanford por Larry Page[36] y Sergey Brin a fines del año 1996. Se basa en la medición del prestigio de una página en términos de cuan a menudo un usuario la visita en promedio, es decir, no considera solo los hyperlinks, sino también las rutas de navegación[44]. Asume que el **grafo web** es del tipo **conexo**, es decir, que desde cualquier página p_1 habrá una ruta directa hacia una página p_2, en otras palabras, que existe al menos una sucesión de vértices adyacentes sin repetir los nodos, para llegar a p_1 a p_2. Por otro lado, considera un navegador web del tipo **random surfer,** es decir, un usuario siguiendo los hyperlinks en régimen permanente[37] con igual probabilidad de ir de una página a otra [15]. Cabe destacar que el prestigio de una página p_1, depende del prestigio de las páginas que la apuntan. Los pasos del algoritmo se describen como sigue:

1.- Suponer random surfer y obtener la matriz de adyacencia del universo de páginas *U*.
2.- Calcular la cantidad de oulinks (hyperlinks salientes) desde cada página en *U*.
3.- Precisar los parámetros del modelo, como la probabilidad de no visitar una página.
4.- Iterar para obtener la importancia de una página en el grafo en base a su ecuación principal, la cual incluye en el cálculo de la importancia de una página p_1, el prestigio de la página p_2 que la apunta, considerando la probabilidad de pasar o no de p_2 a p_1.

[34] **Fuente:** http://www.almaden.ibm.com/
[35] **Matriz de adyacencia:** Matriz cuadrada de componentes binarios, donde 1 indica la existencia de hyperlinks y 0 la ausencia de él entre dos páginas (columnas y filas de la matriz). Sus dimensiones son de *mxm*, donde *m* el número total de páginas en cuestión
[36] Larry Page, uno de los fundadores de Google.
[37] **Régimen permanente:** Cuando el tiempo (en medición continua) del proceso en cuestión, tiende a infinito.

Tabla 1: PageRank v/s Hits

Page Rank	Hits
Ambos algoritmos se basan en el modelo de la Web como un Grafo	
Fundamentos	
Ranking basado en la estructura de links.	Ranking basado en la estructura de links y en la relevancia basada en contenidos (consulta)
Busca páginas de acuerdo a grado de autoritatividad (Importancia de la página está dada por el número de páginas que la apuntan).	Busca páginas tanto autoritativas como Hubs
Se realiza un ranking previo a la consulta, trabajando con un universo global de páginas.	La consulta es usada para seleccionar un subgrafo relacionado a ella desde la Web, luego realiza el ranking.
Cada página de la Web tiene una medida de prestigio que es independiente de la información necesitada por la consulta.	Una consulta es usada para seleccionar un subgrafo desde la Web con el cual se determinará su prestigio.
Medición de la importancia de cada página	
Define prestigio mediante solo un camino aleatorio sin ser influenciado por una consulta específica	Después de usar la consulta y extraer las páginas asociadas, ignora el contenido, y pasa a ser puramente basado en la estructura.
En la importancia de cada página influye la importancia de las páginas que la apuntan.	Calidad de ser página autoritativa o hub se va propagando en el grafo Web en cada iteración de manera bidireccional
No considera solo los hyperlinks como una relación entre pares de páginas, sino también las rutas de navegación y por ende las páginas en ellas como trayectoria de un usuario	Considera páginas relevantes asociados a la consulta, sin considerar páginas con una alta cantidad de links hacia ella (inlinks)
Rapidez de respuesta	
Mayor rapidez de respuesta pues tiene el ranking predefinido e independiente de la consulta (más fácil de actualizar)	Tiempo de respuesta mas lento, pues los resultados tienen que estar on-line, recuperando primero los documentos relacionados.

Fuente: Elaboración propia basada en [15][44][56][61]

2.3.2.- Datos web

Con el fin de anticiparse a la toma de decisiones, el análisis y entendimiento del comportamiento del usuario (el cual se detalla en la sección 2.3.4) cobra una importancia fundamental, donde se hace necesario seguir en detalle sus acciones, las cuales podrían incluir toda la trayectoria de páginas visitadas, el tiempo gastado en cada una de ellas, los productos escogidos, cantidad de transiciones antes de llegar a concretar una compra, etc. [61]. Así, con el hecho de navegar en la Web se van acumulando una serie de datos que proporcionarán información relevante: los registros web log [48]. Ellos contienen todo el recorrido de los usuarios, descargas asociadas (páginas, objetos multimedia, archivos, entre otros) y tiempo invertido. Estos se generan cuando se realiza la lectura del código HTML de cada página, que llega y se interpreta por el navegador, registrando cada objeto entregado. Es importante indicar que aunque solo una página sea visitada, más de un objeto puede ser registrado [61], entre ellos una alta cantidad de objetos sin información relevante (banner publicitarios, íconos, etc.), aunque dicha relevancia dependerá del estudio que se quiera realizar, por lo que el filtrar y registrar previamente los datos de interés se vuelve una tarea clave.

En la Tabla 2 puede apreciarse un registro web log tradicional y se destacan las filas que determinan la secuencia de páginas visitadas por un usuario único, determinado por su mismo IP y UserAgent. La configuración de éstos depende como el *webmaster*[38] haya organizado el sitio [51], pues corresponden a registros realizados en el servidor. Es posible de todos modos tener el registro de las siguientes entidades principales:

IP Adress: Dirección IP del computador del usuario con cual accede a la Web.
Time: Tiempo en que el servidor responde a la petición.
User-Agent: Nombre de la versión del navegador usado por el cliente.

Tabla 2: Ejemplo de registro Web Log

id	id_session	time	ip	host	uri	agent
414421	5,17539E+16	1254414929	163.247.48.58	www.dii.uchile.cl	/~webmgpp/mgpp_ejecutivo.htm	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1...
414422	5,17539E+16	1254414930	163.247.48.58	www.dii.uchile.cl	/~webmgpp/horario.htm	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1...
414423	9,85626E+15	1254414940	190.46.118.244	www.dii.uchile.cl	/~webmgpp/	Mozilla/5.0 (Windows; U; Windows NT 5.1; es-ES...
414424	9,85626E+15	1254414942	190.46.118.244	www.dii.uchile.cl	/~webmgpp/folletos.htm	Mozilla/5.0 (Windows; U; Windows NT 5.1; es-ES...
414425	5,17539E+16	1254414948	163.247.48.58	www.dii.uchile.cl	/~webmgpp/horario.htm	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1...
414427	1,36356E+16	1254414962	201.236.175.49	www.dii.uchile.cl	/~mgo2007/contacto/	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1...
414428	4,52421E+16	1254414970	190.196.7.40	www.dii.uchile.cl	/~magcea/	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1...
414429	4,52421E+16	1254414971	190.196.7.40	www.dii.uchile.cl	/~magcea/costos_becas.htm	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1...

Fuente: Elaboración propia a partir de registros del sitio www.dii.uchile.cl

Los web log contienen los registros en su orden de ocurrencia y puede observarse que no hay información privada del usuario en particular. Ahora, como paso previo al estudio de los caminos seguidos por cada usuario, se hace necesario previamente considerar los ruidos propios de los datos registrados [51][61]. Las direcciones IP reales de distintos usuarios pueden llegar a verse como una sola, debido por ejemplo al uso de **proxy**[39] o **firewall**[40] por parte de las instituciones, las cuales buscan controlar el acceso a la Web de sus trabajadores, con esto entonces, se tendrán muchas sesiones distintas registradas como una sola. El **asincronismo** de los registros web implica una alta dificultad para determinar a quién pertenece la sesión, resultando en una falta de identificación de los usuarios. Aquí, técnicas de reconstrucción de sesiones o métodos como las cookies[41] son aplicadas. Los problemas de contaminación en los datos son proporcionados por los crawlers por ejemplo, dado que mediante su rastreo en la Web acceden a una gran cantidad de objetos al mismo tiempo, donde cada uno de ellos queda registrado en los logs. Si dichos registros fueran considerados al momento de construir sesiones, éstas serían artificiales pues no representan a un usuario en particular. Otro factor que agrega ruido a los registros, corresponde a la **memoria caché** de cada computador, la cual guarda objetos a los que se accedieron previamente y que pueden visitarse nuevamente usando el botón back del navegador. La mencionada acción produce que no se realice petición al servidor, es decir, el registro de navegación no será almacenado, produciendo inconsistencias al momento de construir la ruta seguida.

Una disciplina útil para diseñar o rediseñar la arquitectura de información de un sitio web, corresponde al análisis de los **search logs**: archivos de servidor que se generar al realizar

[38] **Webmaster:** Persona responsable de un sitio web específico, quien controla el contenido, imágenes, forma, estilo y estructura del mismo.
[39] **Proxy:** Programa o dispositivo que representa a otros. Como servidor, permite el acceso a internet (al disponer de solo un computador conectado) de muchos equipos.
[40] **Firewall:** Elemento de red computacional que controla (al prohibir o permitir) las comunicaciones.
[41] **Ver sección** 2.3.3.- Proceso de sesionalización

una petición de búsqueda en un buscador. Si se hace en la propio sitio (buscador propio de elementos solo en las páginas de él) se le denominan search log interno, mientras que si resultan de buscadores generales como Google, se les llaman search logs externos. *"Los archivos de search logs guardan un auténtico tesoro pues muestran las consultas en su lenguaje natural, aportando información sobre quién, cuándo, y cuánto se busca algo"* [57]. En el ámbito externo la idea es detectar cuáles serían las consultas en las que aparece más veces como respuesta un sitio en particular, y cuales determinan que aparezca en los primeros resultados. En el ámbito interno, se busca poder evaluar los contenidos existentes buscando aquellos más solicitados, para reestructurar parte del sitio e incluso, inferir el momento del año en que conviene más destacar cierta arquitectura o palabras claves pues *"cuanto más prominente o más fácilmente visibles en el sitio web, más probabilidades tendrán de ser encontradas"* [57].

Especial mención se hace al manejo del tiempo, variable categórica en el estudio que aquí se presenta, donde el cálculo del tiempo neto de visita en cada página es elemental. La forma de registro del tiempo puede contener o no la fecha y/o indicar hora, minutos, segundos. Sin embargo, una forma igualmente usada es el **tiempo formato unix**[42], el cual corresponde a la cantidad de segundos transcurridos desde la medianoche UTC[43] del 1 enero de 1970. El viernes 13 de Febrero del año 2009, a las 23:31:30, hora UTC, el tiempo unix se igualó a 1234567890[44]. Así, habiendo fijado el tiempo cero y al tener un incremento en segundos, puede obtenerse el tiempo neto de visita en cada página simplemente restando el tiempo unix de los eventos de inicio y término del fenómeno en cuestión.

2.3.3.- Proceso de sesionalización

Teniendo los registros *web log* se hace necesario ahora estudiar las rutas de cada usuario individualizando sus sesiones. A aquél proceso de reconstrucción de caminos tomados se le llama **sesionalización** [61]. El propósito del mismo, es encontrar las sesiones reales de cada usuario, por ello se han propuesto heurísticas para reconstruirlas desde los web log para asociar los registros a una sesión única durante un período de tiempo. Las técnicas ocupadas pueden ser clasificadas según la estrategia que siguen, ya sea **proactiva** o **reactiva**:

Estrategias proactivas:

Corresponde por ejemplo, a la colocación de un dispositivo de rastreo vía cookies**.** Una vez que el usuario entra al sitio web por primera vez, una cookie es enviada a su computador siendo almacenada en éste. Gracias a eso al ingresar por segunda vez, el usuario será identificado con lo que se puede otorgar una página mucho más personalizada y cercana a los intereses particulares [60]. Puede desprenderse de lo anterior, que el acercamiento a las versiones reales de la sesión vía cookie llega a ser bastante fidedigna, sin embargo, presenta problemas de **privacidad** (pues hay un elemento externo en el computador que detecta las

[42] **Fuente:** http://www.unixtimestamp.com/index.php
[43] **UTC:** Universal Time Coordinated, tiempo universal coordinado. Corresponde al tiempo de la zona horaria de referencia en base a la cual se calculan todas las demás zonas del mundo.
[44] **Fuente:** http://www.1234567890day.com/

trayectorias tomadas y en base eso toma decisiones para cada usuario en particular) e incluso, con legislaciones y debates al respecto[45], además, pueden ser **fácilmente detectadas** y eliminadas por los usuarios.

Usar cookies es uno de los mejores métodos para establecer las trayectorias de los visitantes, identificar a usuarios únicos y las sesiones en general, sin embargo son difíciles de configurar e implementar. Por otro lado, si bien están propensas a ser usadas de manera errada y, en términos de privacidad causan controversias, pueden llegar a mejorar la experiencia en la Web de manera considerable, con un amplio beneficio y acercamiento a los clientes [70]. Los tipos de cookies existentes son:

1.- Persistentes o de sesiones
Una cookie persistente es guardada en el computador cliente y permanece por un período específico siendo capaz de seguir la pista de visitantes que se repiten, mientras que una cookie de sesión no es guardada allí y expira al final de la sesión del visitante.

2.- Primera o tercera parte
Una cookie primera parte es ofrecida desde el sitio y es aceptada por el navegador usuario, mientras que una cookie tipo tercera es proporcionada desde fuera del sitio, por entidades que ofrecen seguimiento y análisis, pero pueden ser bloqueadas.

3.- No personales o personales
Cookies no personales rastrean visitantes y sus sesiones sin obtención de información personalizada, lo contrario ocurre con las personales, que pueden identificar usuarios en particular.

Cabe destacar en este aspecto una plataforma para las preferencias de privacidad *P3P*[46], la cual establece utilizar declaraciones de privacidad estandarizada hecha por el emisor de cookies para administrar su aceptación, de modo que sea lo más amplia posible. De este modo la idea es garantizar el uso de Cookies del tipo primera parte y no personal.

La aplicación de las *cookies* llega incluso a impactar el mundo económico, donde tiene que ver con las empresas que disfrutan del poder de la escasez[47] [33]. Ellas buscan utilizar sofisticados métodos para identificar a sus clientes, sobre todo bajo las facilidades que actualmente la tecnología proporciona. Un ejemplo es el caso de Amazon[48], comerciantes de libros por Internet quienes pueden identificar a sus clientes mediante las *cookies*, localizándolas en el ordenador de cada uno de ellos, obteniendo registros individuales que le permiten establecer mecanismos de diferenciación de precios.

[45] **Fuente:** http://www.delitosinformaticos.com/articulos/100733093469733.shtml
[46] **P3P**: Platform for Privacy Preferences, http://www.w3.org/P3P/P3FAQ.html#What%20is%20P3
[47] Poder que se alcanza cuando un producto o servicio tiene un gran valor y es ofrecido por pocas empresas, de modo contrario dicho servicio o producto bajaría de precio.
[48] **Fuente:** http://www.amazon.com/

Estrategias reactivas:

Solo utilizan la información contenida en los web log y pueden ser heurísticas orientadas a la navegación o bien orientadas al tiempo. La primera tiene relación con la construcción de los caminos posibles en base al supuesto que toda página a la que se accede es alcanzable en base a otras páginas visitadas, de no ser así (si no es posible llegar a una página desde las previas en la misma sesión) otra sesión es inicializada. El otro caso, corresponde a las heurísticas que asumen un tiempo máximo de duración de cada sesión (en general 30 minutos [61]), superando el valor establecido se considera otra sesión.

Tomando amabas estrategias anteriores (tanto proactiva como reactiva), en la Tabla 3 se visualizan los principales mecanismos para la identificación de sesiones, con una breve descripción, su nivel de intervención en la privacidad, además de sus ventajas y desventajas. Para que la sesionalización sea fructífera, se requiere pre-procesar los datos siguiendo los siguientes pasos [48][51]:

1.- **Limpiar los datos:** Quitar por ejemplo los registros que corresponden a objetos multimedia que pertenecen a una página, pues no son pedidos explícitamente por el usuario.
2.- **Limpiar registros no humanos**: Quitar el acceso a las páginas por parte de spider, crawlers y otras máquinas automáticas de peticiones al servidor.
3.- **Identificar usuarios independientes**: Combinar técnicas de direcciones IP-Agent y cookies para identificar a cada usuario.
4.- **Identificar sesión del usuario**: Por cada visita, determinar páginas visitadas y la duración en cada una de ellas. También aquí, se trata de estimar cuando el usuario dejó el sitio web.
5.- **Completar rutas**: Completar rutas incompletas por el uso del botón back, conociendo los caminos posibles que el sitio permite.

Para identificar a un usuario único las formas más comunes son vía dirección IP y agente [48], dado que son elementos necesarios para navegar. Si bien podría considerarse la IP por si sola como identificador, el que solo haya un número limitado disponible hace que se vuelva ineficiente, sobre todo por la acción de los ISP que pueden otorgar dos direcciones distintas a un mismo usuario durante la navegación, pues poseen un conjunto limitado de IP para otorgar a sus clientes.

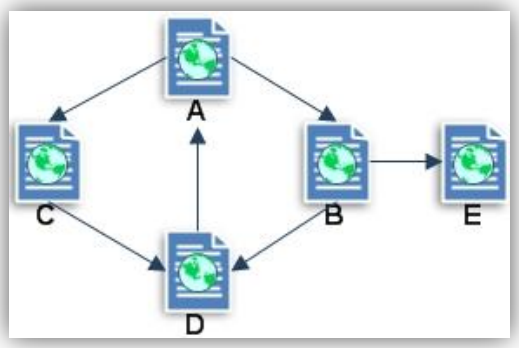
Tabla 3: Mecanismos para la identificación de sesiones

Método	Descripción	Intervención en privacidad	Ventajas	Desventajas
Dirección IP + Agente (Browser)	Asume que cada par único de {Dirección IP, Agente} indica a un usuario único	Baja	Siempre disponible y solo requiere poseer poseer los registros Web, sin tecnología adicional	No garantiza la unicidad y pierde control de los Ips rotativos
Ids de sesiones integradas	Usa dinamicamente las páginas generadas para asociar el ID con cada hiperlink	Baja a Media	Siempre disponible, indendiente de la dirección IP	No puede capturar visitantes repetidos y requiere de un trabajo adicional ante páginas dinámicas
Registro	Registra explícitamente a los usuarios en el sitio	Media	Puede serguir pistas individuales, no solo de navegadores	Muchos usuarios no serán registrados. No es válido antes del registro.
Cookie	Guarda ID en el computador cliente	Media a Alta	Puede seguir la pista de visitantes desde un mismo browser	Puede ser eliminado por los usuarios
Programas Agentes	Programas cargados en el Navegador el cual devuelve los datos de uso	Alta	Uso preciso de los datos para un solo sitio.	Rechazo de los usuarios

Fuente: [8]

Por otro lado, se asume que cada sesión tiene un tiempo máximo de duración. Sesiones con tiempo mayor a dicha cantidad, aunque posean a un mismo usuario definido anteriormente, son consideradas como una sesión completamente distinta. Lamentablemente, ante los problemas mencionados en la Sección 2.3.3.- Datos web, se hace bastante difícil determinar las sesiones únicas de cada usuario, sobre todo por el uso del botón back del navegador, pues faltaría un registro en los web log resultando en una construcción incompleta de la trayectoria total. Por ejemplo, si se tiene un grafo como el de la Figura 13 y por otro lado, el registro de navegación indicada que uno sesión recorrió el camino *A*, *B*, *E*, *D*.

Figura 13: Ejemplo grafo para completar rutas



Fuente: Elaboración propia

Dado que no existe camino posible entre la página *E* y *D*. Lo más lógico es haber presionado el botón back estando en *E* hacia *B*, pues ésta si posee un hyperlink hacia la página *D*. Finalmente, la sesión quedaría: *A*, *B*, *E*, *B*, *D* [49].

[49] Asumiendo que siempre es posible volver a una página aunque no exista hyperlink, presionado back

Finalmente, condiciones mínimas se le pedirán a las sesiones reales para establecerlas como tal [61]. Ante eso, sea *R* el conjunto de registros web log y $W = \{w_1, w_2, \ldots, w_n\}$ un set particular de sesiones reales extraídas de *R*, se le pedirá:

1.- Ser registros que deben seguir una secuencia temporal, considerando el tiempo de respuesta a cada petición (Timestamp) [51], es decir:

$$timestamp(r_i) > timestamp(r_{i-1})$$

2.- Solo objetos de *R* pueden aparecer en *W*

3.- Cada petición en *R* pertenece a solo una sesión (solo un conjunto *W*)

2.3.5.- Nuevas problemáticas y mejoras

A modo de muestreo de la contingencia en aspectos de sesionalización y registros web, se exponen a continuación diversas temáticas relacionadas.

Una nueva problemática para detectar las sesiones se presenta con la reciente extensión de Google para Internet Explorer[50]. El primero tomó la decisión de insertar Chrome[51] en el navegador, lo que ha generado el temor y discusión por parte de investigadores ante la fragmentación, pérdida de control, confusión de los usuarios y por sobre todo, aumento de la dificultad en el manejo de la información.

Sin considerar aún lo anterior en los mecanismos actuales, ante problemas inevitables al momento de identificar sesiones (erradas o con pérdidas) se hace cada vez más necesario potenciar el proceso, contar con un estudio posterior que distinga el comportamiento obtenido con el real y reducir considerablemente el número de errores [61].

La metodología de sesionalización más usada es la orientada al tiempo, sin embargo, no es clara su validación ni precisión. Existen otros enfoques, como los basados en el campo "referer"[52] de los web log, o bien los basados en la semántica [20].

Como forma de optimizar el proceso de sesionalización se han propuesto diversos estudios entre los cuales se destaca el uso de un modelo de programación entera en la creación de un algoritmo para identificar sesiones de los usuarios Web [20][21]. La idea es, tal como la programación heurística[53], agrupar los registros en base al mismo IP y agente, garantizando que las sesiones construidas se basen en la estructura de enlace que la página posee. La programación entera en cambio, en vez de construir las sesiones una a una lo hace simultáneamente y tiene una gran ventaja en lo que corresponde a tiempo de ejecución. En [20] se agrega un modelo de optimización (llamado Bipartite cardinlity matching) obteniendo cerca de la mitad de los errores estándar de las heurísticas tradicionales.

---

[50] **Fuente:** http://www.diarioti.com/gate/n.php?id=24249

[51] **Chrome:** Navegador Web realizado por Google

[52] **Referer:** Dirección de origen desde la cual se realizan las peticiones, puede ser la URL de una página interna o externa con un hyperlink hacia ella.

[53] **Ver sección** 2.3.3.- Proceso de sesionalización

Finalmente, y como estudio del impacto de la estructura web en la calidad de la construcción de sesiones, en [7] se investigó el comportamiento de las cookies y de las heurísticas restantes de sesionalización, estableciendo que dependiendo de las características del sitio, se recomienda una u otra metodología, basándose en términos estadísticos, contenidos, distribuciones, largo de sesiones y en definitiva, los perfiles de los usuarios. Se destaca el considerable impacto de una reconstrucción fidedigna en las aplicaciones predictivas como la personalización web.

2.4.- Comportamiento del usuario en la Web

"Para efectos de marketing es de una importancia estratégica conocer y estudiar modelos que describan el comportamiento de navegación del usuario web" [53]

Con lo anterior y con el objetivo de poder brindar a los usuario un buen servicio (dada su clara implicación en la imagen de la empresa) se desarrolla un área específica para establecer *"el mejor entendimiento de las preferencias del usuario y con ello, poder desarrollar un sitio web más atractivo en estructura y contenido" [61]*. Se desea encontrar la relación directa entre el diseño y el comportamiento del consumidor en el mismo, para comprender la conectividad "virtual" entre las características estructurales de un sitio y la dimensión perceptual. Ello involucra teoría de comunicaciones, estudios de marketing, investigación publicitaria, diseño de interfaces, investigación de sistemas de información, etc. A grandes rasgos, los estudios del consumidor online, están orientados a estudiar el comportamiento basado en el comercio web, el comportamiento de compra, las técnicas interactivas de publicidad, reclamos de los consumidores, entre otros [24], con el fin de poder extraer patrones de interés que favorezcan la toma de decisiones.

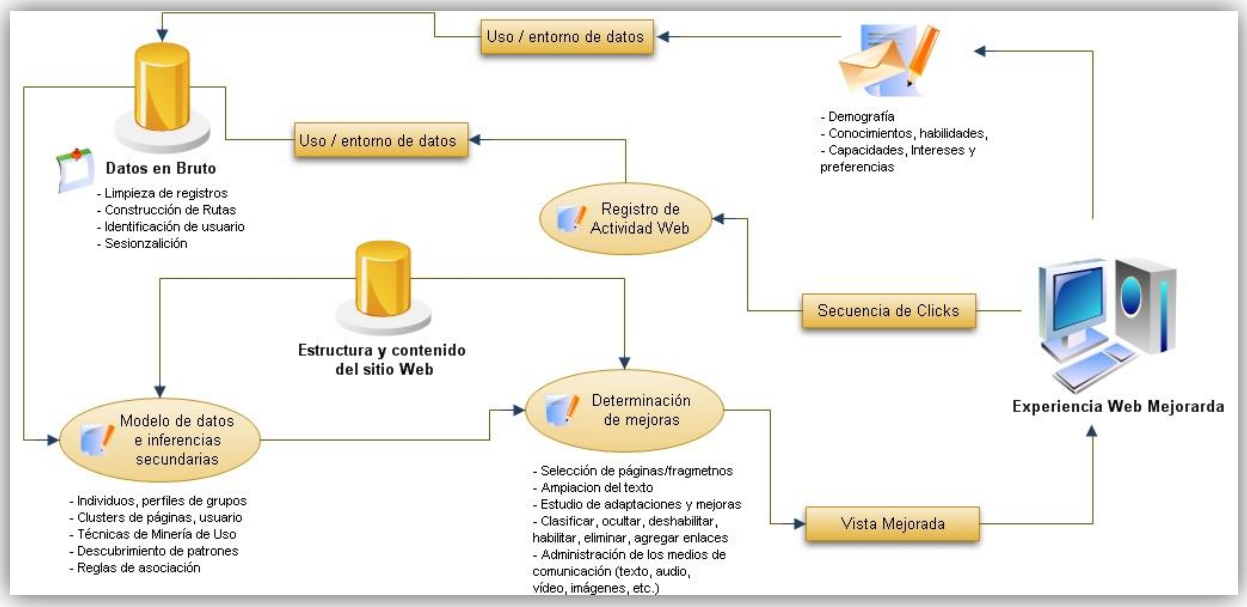
Como ya se ha hecho mención, el análisis del comportamiento del usuario en un sitio, proporciona los elementos claves que permiten establecer cómo mejorarlo, tanto en contenidos como estructura, para de esta manera asegurar una participación exitosa en el mercado digital [62]. Los web log son los archivos que contienen evidencia de la actividad de cada usuario, por lo que son verdadera encuesta electrónica inmensamente amplia y por construcción totalmente anónima [21][60]. Desde el punto de vista de comportamiento del consumidor, el objetivo del análisis de los logs es conocer quienes visitan el sitio, qué necesidades de información tienen, qué información consumen, cómo está estructurada la información, situación de la web en el contexto general, etc. [48].

Luego de que las sesiones han sido determinadas, numerosas técnicas del web mining permiten el descubrimiento de patrones del comportamiento del usuario web. Entre ellas cabe mencionar reglas de asociación, clustering, clasificación, patrones secuenciales, entre otras [20]. Un proceso completo hacia la obtención de patrones de interés que favorezcan la acción, puede observarse mediante los pasos claves en el estudio del comportamiento del usuario en la Web, en la Figura 14.

En forma particular, se destacan dos estudios. Uno de ellos establece una nueva medida de similitud para entender el accionar del visitante [62]. La idea es combinar el

contenido de las páginas web visitadas con la medida de similitud entre secuencia de ellas, de modo de establecer nivel de diferencias entre visitas en general. El segundo estudio, presenta el análisis del comportamiento de navegación del usuario basado en cadenas de Markov a tiempo continuo [53]. Las propiedades markovianas [13] se establecen en la navegación del usuario de una página a otra, con el fin de determinar estimadores de transición de probabilidad, tasas de entrada, salida y permanencia.

Figura 14: Pasos claves en el estudio del comportamiento del consumidor



Fuente: [56]

2.4.1.- Conductas de navegación

Un principio fundamental para lograr participación efectiva en el mercado es entender el comportamiento de de compra de los consumidores [61], pero antes, dado que la fuente de información es la Web, comprender como se genera la navegación ayuda a establecer primeros acercamientos. Así, se deben tener en cuenta los dos modos fundamentales de visita al diseñar la arquitectura del sitio [57]:

Browsing (navegación)
Sigue la estructura interna de enlaces organizados.
Searching (búsqueda)
Realiza una búsqueda ya sea utilizando un buscador interno para navegar en la página o bien externo, como Google, donde puede comenzar su navegación desde distintas páginas del sitio, no necesariamente la principal.

Conociendo lo anterior, la idea será determinar la preferencia por una u otra forma de visita, sobre todo desprendiendo aquella más frecuente en el sitio. Así, se evaluará si el sitio se debe realizar enfocado más para la búsqueda o navegación, ver si el buscador ofrece lo

solicitado en tiempos de respuesta aceptables (Search engine optimization[54] [31]) y ver si los enlaces que se poseen permiten acceder a la información requerida. Históricamente se tiene que "*las búsquedas de información son minoritarias frente a la navegación o las transacciones (descargas, compras o reservas)"* [3].

Ahora, desde un punto de vista macroscópico, es posible contar las estadísticas de uso del sitio web ya sea analizando los web log del servidor, controlando los accesos desde el cliente o bien, utilizando un programa de estadísticas, el cual proporciona informes (en general gratuitos) con la distribución de los visitantes, las páginas y/o documentos visitados, el origen de los usuarios, motores de búsqueda, consultas realizadas, rutas seguidas, tiempo utilizado, velocidad de conexión, sistema operativo, entre muchas otras opciones [48]. Se destaca en este contexto dos herramientas útiles y de fácil instalación: StatCounter[55] con una versión pública y limitado a las 500 últimas visitas, y Google Analytics[56], donde sólo el webmaster tiene acceso a su interfaz.

En [17] se encuentra un método experimental del comportamiento de los usuarios, cuyo fin es obtener patrones de visitas y revisión de las páginas, para proveer de información que permita guiar el diseño de las interfaces de modo que éstas respondan mejor a las actividades de navegación. Para lo anterior se consideran relaciones numéricas entre el número de páginas visitadas, la cantidad de visitas, tasas de revisión, dominancia de algunas comunidades de usuarios, etc.

El principio del "mínimo esfuerzo", conocido también como **ley de Zipf,** indica que el mayor uso de un conjunto recae en pocos elementos, habiendo una amplia gama de ellos. Dicho principio se hace tangible y comprobable en el comportamiento web. Este principio, explicado en relación a las palabras usadas en textos, indica que al ordenarlas de la más frecuente a la menos frecuente, se tiene que $f_i \cdot i = C$, donde f_i es la frecuencia de la *i*-ésima palabra y *C* una constante que depende del texto. Para textos web *i* se eleva a un exponente *k* entre 1 y 2. Usando una escala logarítmica en una gráfica de palabras v/s frecuencia se obtendrá una recta de pendiente negativa [69]. Lo clave es la extrapolación del comportamiento a diversos fenómenos web, como el número de enlaces que salen o llegan a una página, uso de palabras en las consultas a un buscador, etc. [16].

2.5.- Algoritmos genéticos (AG)

..."Como de cada especie nacen muchos más individuos de los que pueden sobrevivir, y como, en consecuencia, hay una lucha por la vida, que se repite frecuentemente, se sigue que todo ser, si varía, por débilmente que sea, de algún modo provechoso para él bajo las complejas y a veces variables condiciones de la vida, tendrá mayor probabilidad de sobrevivir y, de ser así, será naturalmente seleccionado. Según el poderoso principio de la

[54] **SEO:** Search Engine Optimization, abarca una amplia variedad de tareas que mejoran la presencia de un sitio web en los motores de búsqueda.
[55] **Fuente:** http://www.statcounter.com
[56] **Fuente:** http://www.google.com/intl/es_ALL/analytics/

herencia, toda variedad seleccionada tenderá a propagar su nueva y modificada forma"… [67].

Los algoritmos genéticos [45] (desde ahora también AG) pertenecen a la familia de modelos computacionales inspirados en la evolución. Estos algoritmos codifican una solución potencial de un problema específico como un cromosoma contenedor de la estructura de datos (características), para luego, aplicar operadores de selección, cruce y mutación con el fin de preservar la mejor información. Así se construye mejor solución, es decir, el individuo más adaptado al medio. Estos algoritmos suelen ser vistos para fines de optimización restringidos a ciertas áreas de investigación, sin embargo, el rango de problemas en los que pueden ser aplicados es mucho más amplio.

2.5.1.- Historia

Durante millones de años, los organismos vivos, se han transformado para adaptarse al medio y a las nuevas exigencias del mismo, con el fin de sobrevivir y hacerse cada vez más aptos a los cambios de la naturaleza. Esta teoría con ciertas modificaciones hoy en día recibe gran apoyo científico y es la base del estudio del origen de las especies [67]. Su gestor es el famoso naturista inglés Charles Robert Darwin, quien postuló que todas las especies de seres vivos han evolucionado con el tiempo a partir de un antepasado común mediante un proceso denominado selección natural [67].

Por otro lado, en los últimos 40 años, la ciencia ha recibido el apoyo de nuevas metodologías y técnicas de estudio, entre ellas el enorme potencial de la información automática por medio de ordenadores: la Informática. Con ello, se han ido desarrollando diversas aplicaciones de las nuevas y cada vez más consolidadas tecnologías de la información[57] en las ciencias biomédicas, naciendo así la BioInformática: "*Disciplina científica que se interesa por todos los aspectos relacionados con la adquisición, almacenamiento, procesamiento, distribución, análisis e interpretación de información biológica, mediante la aplicación de técnicas y herramientas de las matemáticas, de la biología y de la informática, con el propósito de comprender el significado biológico de una gran variedad de tipos de datos*" [68]. En ella, se diferencian tres líneas de investigación, representadas esquemáticamente en la Figura 15.

1.- BioInformática (Informática → Biología molecular + Genética)

Investigación y desarrollo de la infraestructura y sistemas de información y comunicaciones que requiere la biología molecular y la genética. Ejemplos: Redes y bases de datos para el genoma, microarrays[58], etc.

2.- Biología molecular computacional (Informática + Matemáticas → Biología)

Modelización y simulación para el entendimiento de aspectos de la biología básica. Ejemplo: modelos moleculares.

[57] Técnicas, procesos y dispositivos que se involucran hacia propiedades de almacenamiento, procesamiento y transmisión de datos.

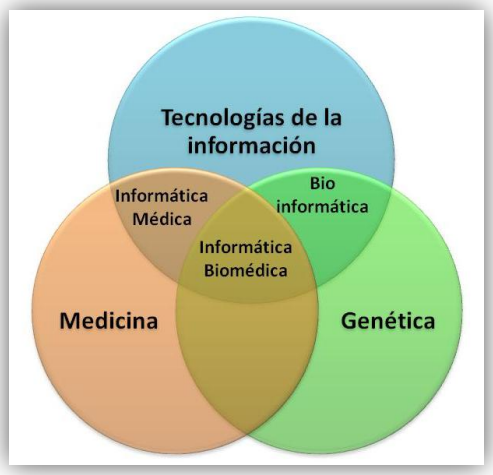
[58] **Microarreglos de ADN** o bien chips de ADN: Superficies empleadas para fijar ADN, al permitir unir una serie de fragmentos del mismo para averiguar la expresión de genes, monitorizándose miles de datos y segmentos simultáneamente.

3.- Biocomputación (Biología→ Computación)

Desarrollo y uso de sistemas computacionales basado en modelos y materiales biológicos. Ejemplos: biosensores, computación basada en ADN, redes neuronales, algoritmos genéticos.

Se destaca no solo el apoyo de la informática hacia la ciencia biológica, sino que además, el evento recíproco por parte de la biología y la medicina en aplicaciones de las ciencias de la información, en la creación de modelos que busquen la resolución, búsqueda y optimización de soluciones de problemas complejos.

Figura 15: Información biomédica



Fuente [68]

2.5.2.- Origen y definición

La computación evolutiva [26], basa sus prácticas en los principios de Darwin elaborando fundamentos, aplicaciones y técnicas heurísticas. Se clasifica como uno de los campos de investigación en *Soft Computing*, como se observa en la Figura 16.

Figura 16: Soft computing

Soft Computing	=	**Computación Evolutiva**	+	**Redes Neuronales Artificiales**	+	**Lógica difusa**

Fuente: Introducción a los Algoritmos Genéticos [26]

Continuando en la línea de la computación evolutiva, sus componentes, influencia y principios se ven reflejados en la Figura 17.

Figura 17: Ecuación evolutiva

Computación Evolutiva	=	**Algoritmos genéticos**	+	**Estrategias de evolución**	+	**Programación Evolutiva**

Fuente: Introducción a los Algoritmos Genéticos [26]

La computación evolucionaria agrupa una serie de técnicas conocidas como algoritmos evolutivos, entre ellos: algoritmos genéticos, programación evolucionaria, estrategias evolutivas y programación genética. En este marco, los AG son métodos adaptativos. Su definición formal viene dada por David Goldberg, consolidador de este mecanismo en su obra en 1975: "*Los Algoritmos genéticos son algoritmos de búsqueda basados en la mecánica de selección natural y de la genética natural. Combinan la supervivencia del más apto entre estructuras de secuencias con un intercambio de información estructurado, aunque aleatorizado, para constituir así un algoritmo de búsqueda que tenga algo de las genialidades de las búsquedas humanas.*" [27].

Una definición más actual es la desarrolla por John Koza en su obra en 1992:*"Un algoritmo genético es un algoritmo matemático que transforma un conjunto de individuos o población (colección de objetos matemáticos representado un individuo), cada uno de los cuales tiene asociado un valor de adaptación, en una nueva población (la siguiente generación) utilizando una serie de operadores basados en los principios darwinianos de supervivencia del más adaptado" [40]*.

2.5.3.- Fundamentos biológicos

Los fundamentos biológicos que sustentan el desarrollo de los algoritmos genéticos, tienen relación con los conceptos que participan en todo el proceso de evolución y que han sido llevados de manera analógica hacia la creación del algoritmo.

Los AG comprenden todo el campo de soluciones potenciales que corresponden a los individuos, donde aquellos con mejor adaptación al medio, es decir, con mejor valor matemático que responda al problema cumpliendo las restricciones, son las que tendrán mayor probabilidad de reproducirse y dejar una descendencia que contenga parte de las mejores soluciones. Así, se va expandiendo el espacio de búsqueda en base a procesos no determinísticos en cada generación. Aquellos aspectos o cambios que favorecen la competitividad son preservados, y aquellos aspectos que debilitan su adaptación son eliminados. Estas características favorables o desfavorables, se almacenan y controlan desde una unidades llamadas genes, que a su vez se agrupan formando conjuntos llamados cromosomas [11].

Aunque no se sabe a ciencia cierta qué información específica codifica cada unidad genética en un individuo, se plantean los hechos aceptados por la comunidad científica sobre la teoría de la evolución [26]:

- La evolución es un proceso que opera fundamentalmente sobre los cromosomas
- El proceso de selección natural señala que aquellos individuos mas adaptados al medio se reproducen más que aquellos.
- En la reproducción es donde tiene lugar la evolución

Para profundizar, se lleva a cabo un orden de los conceptos biológicos raíces en el desarrollo del presente documento con su definición científica [63] y correspondiente analogía algorítmica:

Gen : Unidad indivisible de información hereditaria, que interviene en el desempeño de una función que al final afectará alguna característica del organismo. En AG, los genes corresponderán al conjunto de Bits que determinan un parámetro (característica)

Alelo : Formas alternativas de un mismo gen que regulan las variaciones de las mismas características determinando el estado recesivo o dominante de información genética. En términos algorítmicos, será la unidad elemental que regirá la variación de un mismo estado de la representación del individuo. Corresponde a cada bit de información binaria[59] tomando valores 0 o 1.

Cromosoma : Estructuras del interior del núcleo celular que contiene todas las unidades hereditarias, llamadas genes. Corresponderá por ende, en analogía, a la cadena de valores de todos los parámetros que caracterizan un individuo.

Genotipo : Construcción genética completa del organismo, toda la información indispensable en su composición. Por ello, en AG corresponderá al conjunto de características que conforman a un individuo (solución posible) por completo.

Fenotipo : Expresión física de los genes de un organismo, lo "visible" del genotipo, es decir, en analogía, su valor como solución potencial del problema, la adaptación (función Fitness)

Individuo : Organismo vivo que interactúa con el medio y su grupo de semejantes más cercanos, con la principal característica y capacidad de reproducir su especie. En un AG, corresponderá a la representación de las soluciones potenciales del problema a resolver.

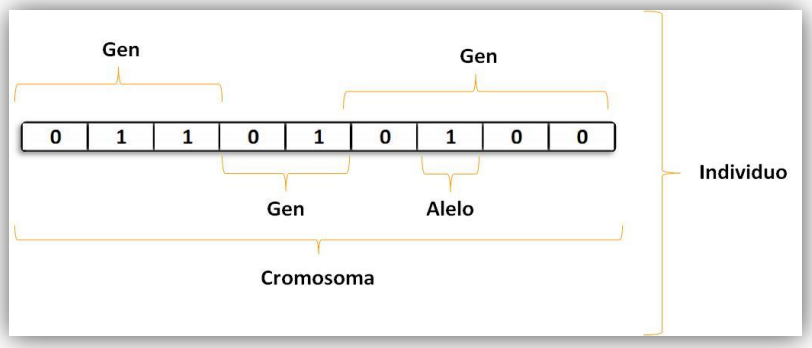
Población : Conjunto de individuos, el cual ha de representar toda la variedad posible de soluciones, en otras palabras, corresponde a la colección de soluciones candidatas.

Ambiente : Corresponde a los factores externos con que interactúa el individuo, condicionando en cuanto a su rigurosidad las adaptaciones necesarias a lo largo del tiempo. En relación al nivel algorítmico, se verá representado por las restricciones y condiciones que ha de cumplir para satisfacer la problemática a resolver.

La Figura 18 muestra un ejemplo de representación binaria de un individuo genético y sus principales componentes. Gen se indica como un grupo de alelos binarios codificadores de cierta información genética, y el conjunto de ellos, determinan el cromosoma de información propio de un individuo.

[59] Cabe destacar que en un comienzo, la representación de un AG viene dada por valores binarios, pero se ha expandido a representaciones enteras, simbólicas, entre otras, donde el alelo igualmente corresponderá a la unidad atómica de la cadena de información.

Figura 18: Individuo genético de representación binaria



Fuente: Elaboración propia

2.5.4.- Codificación del algoritmo

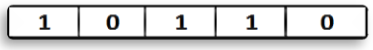
La clave para un buen desarrollo del algoritmo viene dado por una adecuada representación de los individuos, es decir, como se verán representadas las soluciones candidatas en término de una cadena de valores binarios, enteros u otros, además del por qué de dichos valores. Por otro lado, gran relevancia tiene la definición de la función objetivo, que indique que tanto mejor es la nueva solución hija con respecto a sus progenitoras y también en relación a los demás individuos de la población. Citando a D.Mohanraj: *"Una representación efectiva y magnífico fitness de evaluación son las clave del éxito en los AG"* [46].

Por mucho tiempo, y desde los inicios, la principal **representación** de los AG ha sido del tipo binario. Dichos números representan en su conjunto, en términos simples, la presencia o no de una o más características, y pese a que puede complicarse mucho la representación binaria, de una u otra forma siempre es posible. Ante dicha complejidad no menor, son necesarias y útiles otras representaciones de acuerdo al problema a resolver.

❖ **Representación binaria**

Cada alelo tiene sólo dos opciones, ser 0 o 1, tal como muestra la Figura 19.

Figura 19: Representación binaria

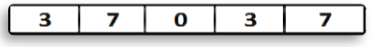


Fuente: Elaboración propia

❖ **Representación entera**

En este caso (Figura 20) cada alelo es un valor entero, depositando información a niveles cuantitativos (dar significados propios a cada número).

Figura 20: Representación entera

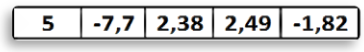


Fuente: Elaboración propia

❖ **Representación real**

Cada alelo es un valor real, tal como se indica en la Figura 21, donde se abarca un aún mayor espacio cuantitativo, aunque agrega una complejidad adicional, pero que dependerá de las características del problema.

Figura 21: Representación real



Fuente: Elaboración propia

Cabe señalar que gracias a los procedimientos del sistema binario[60] es posible llevar a nomenclatura de ceros y unos tanto números enteros como racionales, pero esto depende de las características del problema a resolver para no exceder en complejidad la representación.

La función de evaluación o **fitness,** corresponde a la medida de ajuste y adaptación de los individuos a las restricciones del problema, un mayor valor en su función de evaluación indica mejor adaptación a las características del medio, y con eso el ser más cercano al valor óptimo.

Para resolver un problema de optimización, es necesario minimizar o maximizar una función objetivo, ésta corresponderá entonces a la función fitness. Cabe destacar que al igual que en mecanismos de búsquedas de respuestas óptimas, para poder castigar la función ante el incumplimiento de las restricciones, se procede con lo que se denomina **la penalización**, es decir, dar un valor arbitrariamente muy alto al Fitness si se está buscando minimizar por ejemplo. En la literatura por otro lado, es posible distinguir cuatro tipos de fitness [40]:

❖ **Fitness puro fp(i,t)**
El valor de bondad de un individuo se mide en base a la diferencia entre su valor deseado *e* y su valor obtenido *r***.**

❖ **Fitness estandarizado fe(i,t)**
Trabajo con valor absoluto del fitness anterior para evitar la dificultad para identificar si se está ante un proceso de maximización o minimización.

❖ **Fitness ajustado fa(i,t)**
Proceder con el ajuste estableciendo el cociente *1/(1+ fe(i,t))*. Siempre tomará valores entre 0 y 1, donde mientras más cercano esté el fitness ajustado a la unidad, mayor será su adaptación.

❖ **Fitness normalizado fn(i,t)**
Se introduce ahora la bondad de la solución en relación a las demás soluciones (población restante), representado por el cociente entre el fitness ajustado y la sumatoria total de los fitness ajustados.

Las etapas claves de un algoritmo genético simple (también llamado AG canónico [65]) pueden ser descritas esquemáticamente a nivel de seudocódigo como sigue:

[60] **Fuente:** http://www.binarymath.info/

Entrada:

- Conjunto de M puntos (**tamaño de la población**) del espacio de búsqueda, codificados mediante cadenas finitas de **largo L** sobre un alfabeto finito.
- Definir **número de generaciones G**

Codificación

Algoritmo:

INICIO
  t=0
  Inicializar P(t_0) aleatoriamente
  Evaluar P(t_0)
  Mientras (no condición de término) hacer
    Seleccionar P(t) a partir de P(t-1)
    Cruzar padres (pares en P(t)) con **probabilidad p_c**
    Si se ha producido el cruce
      Mutar descendientes con **probabilidad P_m**
    Evaluar P(t)
    t = t+1
  FIN
FIN

Población inicial

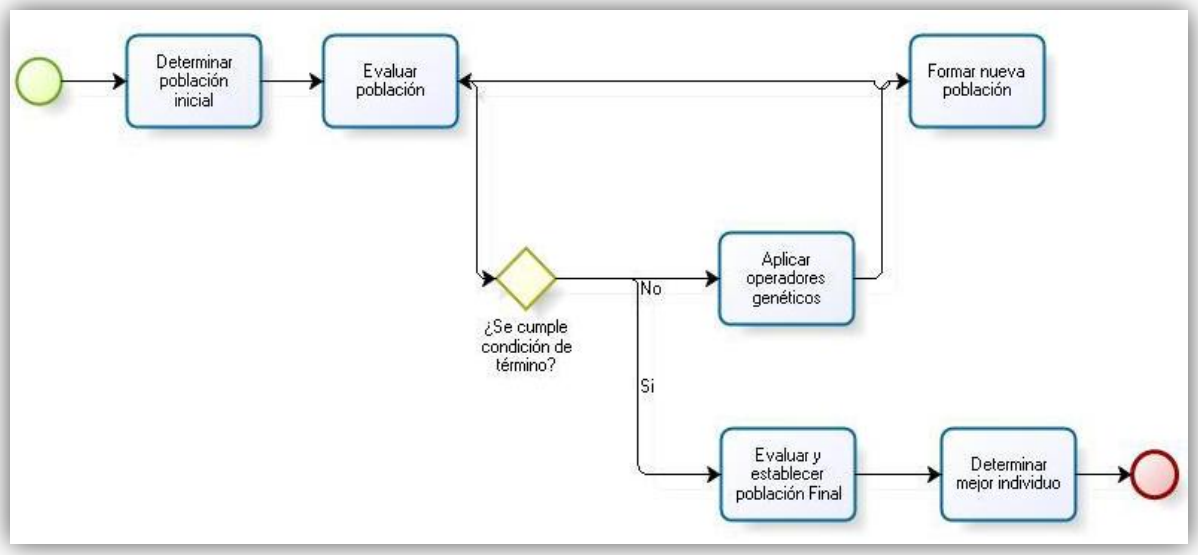
Evaluación

Operadores genéticos

Evaluación

Desde el punto de vista del proceso algorítmico, una vez determinada la representación de las soluciones, el problema se lleva a cabo como se indica en la Figura 22.

Figura 22: Modelo del proceso algorítmico



Fuente: Elaboración propia

2.5.5.- Operadores genéticos

Los siguientes operadores, son aquellas etapas claves en la construcción, desarrollo y convergencia del algoritmo dependiendo de la representación que tenga cada individuo y la adaptación ad hoc al problema a resolver.

Población inicial

Dado que se trabaja sobre poblaciones, para comenzar se ha de poseer evidentemente una población inicial. Para generarla se pueden emplear diversas técnicas como [11]:

- Generación aleatoria de cada valor de los alelos de los individuos.
- Uso de una función ávida (codiciosa en cuanto a su valoración y representación)
- Generar alguna parte de cada individuo y luego aplicar búsqueda local.
- Registro de soluciones reales o conocidas.

Principio de Reproducción

Los algoritmos de reemplazo, los cuales indican cono insertar la nueva población generada por algunas de las técnicas anteriores en la nueva generación: [26][27]

- **Reemplazo de los padres:** Se eliminan los padres, ingresan los hijos
- **Reemplazo de individuos similares:** Un individuos de la población hija reemplazará a un individuo previo con ajuste similar, dentro de algún rango definido.
- **Reemplazo de los peores individuos**: Se escoge en general al 10% peor de los individuos de la población actual para ser reemplazados por la descendencia.
- **Reemplazo aleatorio**: aleatoriamente se seleccionan los individuos a eliminar.

En estos casos, para pasar a la siguiente generación, se deben alcanzar cierto número de selecciones, cruces y mutaciones, los cuales dependen de las probabilidades respectivas de que se concreten las tasas de dichos operadores.

El operador de reproducción además, necesita establecer la diferencia entre cruce y copia, el primero como operador que representa la reproducción sexual, de donde se obtienen individuos (en general 2 hijos por cada 2 padres) con características compartidas de los padres, mientras que el segundo representa la reproducción asexual, donde un subconjunto de individuos de la población actual, pasa a ser parte, sin modificaciones, de la población de la generación siguiente.

Selección

La idea de la selección, es señalar cuáles serían los individuos más propensos a cruzarse, con el fin de dar más posibilidades a aquellos con mejor fitness. La idea no es prescindir de los individuos con bajos fitness pues igualmente tienen posibilidades de ser seleccionados y con ello, capacidad de ofrecer información útil. El número de individuos a

seleccionar, depende del parámetro de tamaño de la población y de aquel que indica cuántos hijos se obtienen a partir de los padres. Hay que tener claro que el operador de selección no genera nuevos individuos, sino que determina cuáles dentro de la población actual dejarán descendencia.

Para llevar a cabo la selección, hay diversas técnicas en la literatura [27][35][45], entre ellas selección proporcional, ranking y torneo.

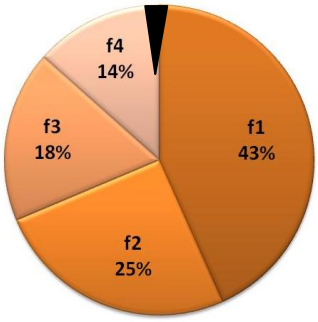
❖ **Selección proporcional**

Bajo este método, la probabilidad que tiene un individuo de reproducirse es proporcional a su valor de función fitness, expresión de la probabilidad de selección del *i*-ésimo individuo en la población *P(t)* en relación a la adaptación relativa con respecto a toda la población. Una vez determinados los valores esperados, se procede con el método de la ruleta o bien, el muestreo estocástico universal.

Método de la Ruleta

Aquí la idea es armar una ruleta de tal modo que el 100% sea el total de la suma de funciones fitness de la población y, partiendo desde un 0 fijado arbitrariamente en la ruleta, se van completando las cifras de cada valor del fitness particular de forma consecutiva, como se representa en la Figura 23, junto a la Tabla 4. Puede apreciarse que el individuo 1 posee mejor valor de la función de evaluación, lo que hace que ocupe mayor porción de la ruleta, y con ello tendrá mayores posibilidades de ser seleccionado. Este método es la base del Holland en relación a que: "*la estrategia óptima de selección consiste en aumentar exponencialmente el número de copias del mejor individuo observado respecto al peor*" [22]. Un alcance a este modelo es que la diferencia entre el número esperado de copias y el real obtenido puede ser muy grande. Para corregir esta situación es que hay un nuevo método llamado muestreo estocástico universal.

Figura 23: Ruleta de selección



Fuente: Elaboración propia

Tabla 4: Datos para la construcción de ruleta simple

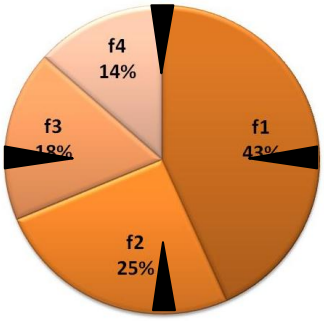
Individuo	Fitness	Valor	Porcentaje
Individuo 1	f1	58	43%
Individuo 2	f2	34	25%
Individuo 3	f3	24	18%
Individuo 4	f4	18	14%
	Total:	**134**	**100%**

Fuente: Elaboración propia

Muestreo estocástico Universal

Como se aprecia en la Figura 24, este es el caso de una ruleta pero con M punteros igualmente separados entre sí alrededor de la ruleta, donde, con un solo giro de la misma se obtienen los M individuos necesarios a cruzar. Este método, sin embargo es sensible a la presencia de un individuo que ocupe la mayor porción de la ruleta, dado que puede llevarse un número demasiado elevado de copias y con eso converger prematuramente a tener dicho individuo como óptimo final al transcurrir las generaciones, siendo que puede ser, más bien, un óptimo local.

Figura 24: Muestreo estocástico



Fuente: Elaboración propia

- **Ranking**

Con el fin de corregir la convergencia prematura por la presencia de superindividuos [22] anteriormente indicada, es que se desarrolla la selección vía método de Ranking. Éste corresponde a ordenar la población en cuanto a su valor de adaptación. El más habitual es al ranking lineal.

- **Torneo**

Este método en general contempla la selección aleatoria de dos factores, un par de individuos y un número entre 0 y 1. Entonces, teniendo a los dos individuos, si el factor 2 es mayor que un nivel dado (comúnmente 75%) se selecciona dentro del par de individuos, el con mejor adaptación, si no es así, se selecciona aquel con peor fitness. Por la sencillez enunciada, cabe destacar que éste método tiene un bajo costo computacional, dado que no requiere de algoritmos de muestreo adicionales, y por ello, no se basa en valores esperados.

Una medida para evaluar el desempeño de los métodos de selección es su presión selectiva [22], la cual se define como *"el tiempo requerido por el mejor individuo para llenar la población con copias de sí mismo (takeover time) cuando no actúa otro operador genético"*. Para valores estándar de sus parámetros, los métodos de selección ordenados de mejor a menor presión selectiva son torneo, ranking y selección proporcional.

Cruce

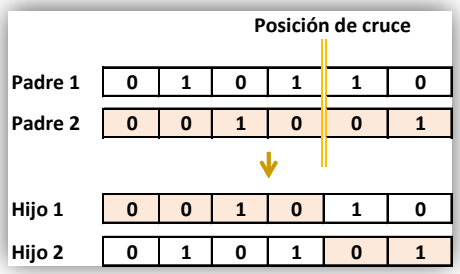
Una vez ya realizada la selección, y con eso la expansión en el espacio de búsqueda hacia los individuos potencialmente mejores, es hora de explorar y trabajar la información almacenada, combinándola adecuadamente, con la idea de aprovechar la alta adaptación de

los padres y con ello, las ventajas de dicha combinación. La forma de cruce determina los distintos tipos existentes, dependiendo de la cantidad y distribución de los puntos de corte.

Cruce de un punto

Método de cruce más sencillo, donde se selecciona una de las posiciones intermedias de la cadena de representación de las soluciones y se intercambian los genes, formando dos (por lo general) nuevos individuos, como se observa en la Figura 25.

Figura 25: Cruce en un punto

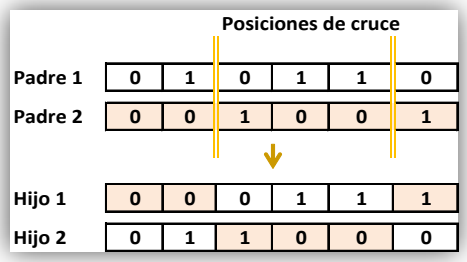


Fuente: Elaboración propia

Cruce de dos puntos

Pequeña modificación en cuanto al tipo de cruce anterior donde ahora, tal como se indica en la Figura 26, son dos puntos de intercambio de información.

Figura 26: Cruce de dos puntos



Fuente: Elaboración propia

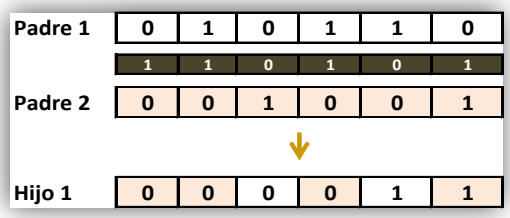
Cruce de n puntos

Extensión de los dos anteriores con “n” puntos de cruce.

Cruce Uniforme/Binomial

Se realiza un test aleatorio intermedio para decidir cuál será el progenitor que aporte su información a la descendencia. Dicho test intermedio, indicado en la Figura 27, puede ser representado por un nuevo individuo con valores binarios, donde 0 indica que la descendencia tendrá la característica del padre 1 en dicha posición y 1, que poseerá la característica del padre 2, tal como se observa en la Figura 27. El inconveniente de este método es que por su construcción es posible obtener un sólo hijo.

Figura 27: Cruce Uniforme



Fuente: Elaboración propia

El operador de cruce opera con probabilidad p_c, lo que permite que haya posibilidad de parejas de padres sin cruce, manteniéndose ellos mismos en la siguiente generación. El algoritmo así, explora un amplio espacio de soluciones y mediante la iteración de operadores de selección y recombinación, va convergiendo hacia las regiones con mayores grados de adaptación.

Mutación

Si con la selección se expande el espacio de búsqueda y con el cruce se combina la información previamente seleccionada, con la mutación se busca no perder la diversidad posible y no converger a un óptimo local, pues, el cambiar cierto número de genes de una solución existente, se está fomentando la variabilidad dentro de una población. Cabe destacar que la tasa de mutación p_m debe ser pequeña, para no tener demasiada diferencia con los padres. Con este operador, se evita la pérdida de diversidad producto de bits que han convergido a un cierto valor para toda la población, lo que no es recuperado por el operador de cruce.

Criterios de detención

Los operadores anteriores se van aplicando hasta que no se cumpla algún criterio de detención previamente definido. Los mecanismos para detener la simulación de más y más generaciones, viene dado por alguna de las siguientes (o un subconjunto de más de dos opciones) alternativas comúnmente aplicadas:

- La diferencia entre los mejores fitness de las poblaciones sucesivas es menor que un pequeño ε definido inicialmente.
- Un 95% de la población de soluciones potenciales posee el mismo valor para un gen, con lo que se dice que la población ha convergido.
- Se alcanzó el número de generaciones máximas especificadas
- Se ha superado el tiempo máximo de ejecución predefinido

Una vez detenida la funcionalidad del algoritmo, la solución será el ajuste del mejor individuo en la última generación.

2.5.6.- Aplicaciones

Los algoritmos genéticos y otros algoritmos como Simulated Annealing y Direct Search, son usados para enfrentar problemas difíciles de resolver por técnicas de optimización tradicionales, incluyendo problemas no bien definidos o difícilmente modelados matemáticamente, además, son usados cuando la función objetivo es discontinua, fuertemente no lineal, estocásticos o tiene poco confiables o indefinidas derivadas.

Con el fin de fomentar la rapidez y potencial computacional de los AG simples, nacen los AG paralelos. En este contexto se encuentran los **algoritmos de nichos paralelos** [22], los cuales obtienen más de una solución cuando no hay solo un valor óptimo, como es el caso de la optimización de funciones con múltiples máximos donde un algoritmo genético tradicional solo convergería a un individuo. Por otro lado, es posible ejecutar algoritmos genéticos en paralelo en distintas computadoras, donde una de ellas coordinará el intercambio, procesamiento, evaluación y limpieza de la población de cada una, esto se le ha denominado **máquina paralela** [54].

El web mining está creando nuevos desafíos hacia los diferentes componentes y tareas del mismo, sobre todo ante el hecho de que la cantidad de información en la Web se está incrementando y cambiando rápidamente sin control [1]. Con el descubrimiento y análisis de la información útil gracias al diseño, búsqueda y obtención de algoritmos que pueden encontrarla, surge la necesidad de incorporar inteligencia artificial dentro de herramientas Web. Así, los AG y sus características han demostrado interesantes resultados en diversas áreas del web mining. Por ejemplo, en búsqueda y recuperación de información, se presenta GSearch [18] como mecanismo de búsqueda basada en AG, también GMiner [50] como una máquina de búsqueda complementaria a las estándar y GCrawler [58], con el objeto de optimizar la búsqueda vía crawlers mejorando el rendimiento de la misma. En optimización de consultas, se presenta un AG para construir perfiles de usuarios, que permite monitorear el comportamiento de navegación usando representación de documentos y consultas vía vectores que permitan representarlas [47]. Otra aplicación relevante de los AG en web mining tiene que ver con la representación de documentos, con el fin de derivar en la mejor de ellas en base a sets de términos indexados [29], como también en relación a la minería distribuida, donde GEMGA (Gene Expression Messy Genetic Algorithm) [36] es el algoritmo de búsqueda evolucionaria paralela que trata los problemas de búsqueda de patrones en un medio ambiente donde tanto los datos como los recursos computaciones son distribuidos.

Cabe destacar además, las investigaciones en relación a la estructura web y el estudio del comportamiento del consumidor, aspectos relacionados directamente con la investigación del presente documento. Aquí se destacan cuatro fuentes fundamentales. En [6] se presenta un simulador del comportamiento del consumidor en general, para simular los efectos de las estrategias de marketing en un contexto de mercado competitivo. Ello fue realizado al implementar un modelo para obtener consumidores virtuales de un mercado real, ajustando las características de ellos mediante algoritmos genéticos y simulación multi- agente. Por otro lado, una combinación de múltiples clasificadores (majority voting, método bayesiano, BKS, borda count y redes neuronales) para la predicción del comportamiento de compra de un

consumidor en los e-commerces llevados a un AG de modo de obtener las mejores combinaciones, se reflejan en [37]. Como tercera fuente en [43] se plantea, una visión de la búsqueda web como un problema de optimización introduciendo en un AG una población de páginas para obtener aquellas más interesantes para el usuario. Finalmente, en [64] se busca construir un sitio que proporciona a sus usuarios la información deseada con la menor cantidad de hyperlinks posibles, analizando la estructura y contenido del mismo, es decir, construir un modelo para optimizar la estructura Web para una navegación más efectiva, usando un AG para precisar los hyperlinks que cumplen con el objetivo.

En definitiva la aplicación de AG en el dominio del web mining provee una perspectiva y dirección potencial para futuras investigaciones [1] sobre todo con miras hacia los nuevos desafíos de proponer modelos de optimización dinámica para hacer frente a las complejas características de la Web (el dinamismo en la información solicitada, hyperlinks existentes, estructura de un sitio, contenidos, etc. [64]) hacia la obtención sitios web adaptativos [61] de comportamiento óptimo.

2.5.7.- Comparación ante métodos tradicionales

Las diferencias de los AG con los métodos tradicionales de optimización se expresan en los siguientes aspectos [22][27]:

1) **Codificación**
   Los AG trabajan con codificaciones de los puntos del espacio de búsqueda en lugar de los puntos como tales. No evolucionan los parámetros en específico, sino que su codificación por completo. He allí la importancia de dicha representación.
2) **Búsqueda paralela implícita**
   En vez de realizar una búsqueda punto por punto, ésta se realiza a partir de un conjunto de ellos de manera simultánea, por lo que se puede hablar de una forma de búsqueda paralela implícita del AG abarcando un amplio rango de soluciones posibles de una sola vez.
3) **Uso directo de la función de adaptación**
   Para optimizar, en general el cálculo de derivadas es la primera metodología inmediata a seguir, sin embargo, no en cualquier función puede aplicarse, y presenta limitaciones en cuanto al tipo de solución final (¿mínimo?, ¿máximo?, ¿inflexión?), sobre todo si se trata de múltiples soluciones posibles. Ante eso, los AG no presentan las limitaciones del método del gradiente, al no exigir derivadas ni otras propiedades a la función objetivo, sino que simplemente ella por sí sola.
4) **Transición probabilística**
   Las reglas de transición entre iteraciones (generaciones) y cumplimiento o no de los operadores genéticos viene dado por probabilidades de ocurrencia, no por reglas determinísticas.
5) **Combinación de técnicas**
   De modo eficiente, los AG exploran áreas en el espacio de búsqueda y aprovechan al mismo tiempo el ir evaluando cada uno de los puntos.

3.- Diseño del algoritmo

Si una compañía desea permanecer competitiva en el mercado digital, necesita un sitio web que ofrezca la información en concreto que los usuarios están buscando, de manera simple y accesible. Sin embargo, la contingencia indica que en muchos casos la estructura del sitio no ayuda a los usuarios a encontrar la información deseada, incluso cuando ésta si existe [61]. Ante eso, con el fin de determinar la mejor estructura web, es que desde ahora se describe un modelo tal que mediante el estudio del usuario del consumidor (gracias a registros web log) permita determinar el mejor grafo para una navegación más eficiente (maximizando la utilidad de los hyperlinks existentes en base a los principios de usabilidad) en el sitio web.

Considerando los datos de entrada que se observan en la Tabla 5, cabe señalar preliminarmente los supuestos y observaciones propios de la extracción de datos y operación de los mismos.

Tabla 5: Datos de entrada para el algoritmo

Datos
Número total de páginas
Número de sesiones totales
Matriz de adyacencia = estructura (existencia de links)
Sesiones con rutas de páginas visitadas
Sesiones con tiempo de permanencia en cada página

Fuente: Elaboración Propia

Supuestos y observaciones

1.- A mayor tiempo de permanencia en una página de una misma sesión, mayor interés tiene el usuario en ella [61].
2.- Sesiones de largo 1 no se consideran.
3.- No se considera la comparación de secuencias totales (clúster de sesiones [61]), dado que con las iteraciones del AG, se van creando y eliminando hyperlinks lo que implica que las secuencias completas ya no sean válidas, es decir, se consideran las relaciones entre pares de páginas y acceso objetivo[61] en cuanto a frecuencias de uso.
4.- No se considera la existencia de más de un hyperlink en una misma página i saliendo a otra j, se tomará como un solo camino presente entre dichas páginas, por principios de diseño, usabilidad y simplificación de los datos.
5.- Los cálculos son construidos en base a los registros de las visitas al sitio considerando cada transición realizada, es decir, se considera el comportamiento de los usuarios para la construcción de sus sesiones, teniendo implícito por lo tanto un cálculo basado en el uso.
6.- La adyacencia es un estado estático en un momento determinado. Ante eso, se sugiere realizar el estudio en sitios con una baja tasa de dinamismo (cambios) en su estructura.

[61] Ver concepto de acceso objetivo en Tabla 6: Conceptos relevantes.

3.1.- Parámetros del algoritmo

Los parámetros son aquellos valores de entrada que no cambian a lo largo de la ejecución del algoritmo y que dependen en gran medida de las etapas a implementar. En el caso de los AG se hace necesario definir:

Tamaño de la los individuos:
En este caso, corresponde a un vector fila de largo *mxm* columnas, donde *m* es el número total de páginas del sitio. Sus valores binarios internos representan la existencia o no de hyperlinks entre páginas[62].

Tamaño de la población:
Indica cuántos individuos habrá en cada generación. Desde el punto de vista algorítmico, tiene que ver con el espacio de las soluciones posibles a estudiar.

Criterio de detención:
Son aquellos criterios por los cuales se establece que el algoritmo termine. Estos pueden ser fijando un número máximo de iteraciones que el algoritmo realice (número máximo de generaciones), indicando un tiempo límite de ejecución del mismo precisando un fitness límite (se detiene el algoritmo al encontrar valores menores o iguales a dicho fitness), detallando una función de tolerancia que considere el cambio acumulado en el valor del fitness a través de las generaciones, o bien, considerando intervalos de tiempo máximo. Así, si no hay mejoras en la función objetivo, se detiene el algoritmo.

Los restantes parámetros corresponden a los pedidos por los operadores genéticos propios de los AG, los cuales dependerán del estudio.

Fitness scaling:
Especificación de la escala a realizar entre todos los individuos de modo de compararlos en base a dicha escala.
Selección:
Elección de los padres de la próxima generación, basado en su valor fitness escalado.
Reproducción:
Forma en que se crearán los hijos, especificando qué individuos tendrán sobrevivencia garantizada, qué porcentaje serán producidos vía cruce y qué porcentaje vía mutación.
Mutación:
Porcentaje de cambio aleatorio en los individuos de la población, el cual provee diversidad genética y explorar un espacio más amplio.
Cruce:
Forma de combinación de los padres, para formar un nuevo individuo.

[62] Mayores referencias del individuo en la sección 3.2.- Representación del individuo

3.2.-Representación del Individuo

Con el objetivo de encontrar una mejor estructura web de un sitio, se establece como individuo genético la matriz de adyacencia $\vec{R}$, matriz binaria contenedora de la existencia de hyperlinks (Definición 1).

$\vec{R}$

- Matriz de *mxm* (siendo *m* el número total de páginas del sitio)
- r_{ij} representa la existencia (valor 1) o no existencia (valor 0) de un hyperlink entre la página *i* y la página *j*, con $i, j \ \varepsilon \ \{1,2,3...,m\}$

Definición 1: Matriz de adyacencia

$$\vec{R} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ ... \\ M \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & ... & M \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & ... & 1 \\ 0 & 1 & 0 & 0 & ... & 1 \\ 1 & 0 & 0 & 1 & ... & 1 \\ 1 & 0 & 1 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & 0 \\ 0 & 1 & 0 & 0 & ... & 1 \end{bmatrix} \end{array}$$

Como requerimientos del programa a utilizar (Matlab [72]), se hace necesario linealizar la matriz para proceder con los operadores genéticos de manera más directa y lógica, como se explicará en los ítems posteriores. Así, con el fin de construir un individuo binario lineal para aplicar el AG se procede:

1.- Largo del individuo

Corresponde al total de datos contenidos en la matriz estructura, es decir, el individuo lineal tendrá un largo *M* igual a *mxm*, siendo *m* el número total de páginas del sitio.

2.- Linealización

Se llamará **linealización** al proceso por el cual se construirá el arreglo lineal (cadena) de representación del individuo genético mediante la matriz estructura, donde el primer bit corresponde a la existencia o no de un hyperlink entre la página 1 y 2, 2º bit: existencia de hyperlink entre 1 y 2, 3º bit: existencia de hyperlink entre 1 y 3, y así sucesivamente. La existencia es indicada en términos binarios: 1= si hay hyperlink, 0=no hay hyperlink.

Desde el punto de vista matemático, se generaliza a continuación la relación entre el individuo lineal y sus componentes matriciales (hyperlinks):

Sea:

C = Individuo genético (cadena de representación)

c_i = Cada bit de la cadena, con $i \in \{1,2,...m\}$

Así, con m=3 páginas se tendría:

c_1	= Existencia de hyperlink de 1 a 1	==>	c_1 :	1→ 1	
c_2	= Existencia de hyperlink de 1 a 2	==>	c_2 :	1→ 2	
c_3	= Existencia de hyperlink de 1 a 3	==>	c_3 :	1→ 3	$m = 3$
c_4	= Existencia de hyperlink de 2 a 1	==>	c_4 :	2→ 1	
c_5	= Existencia de hyperlink de 2 a 2	==>	c_5 :	2→ 2	
c_6	= Existencia de hyperlink de 2 a 3	==>	c_6 :	2→ 3	$2m = 6$
c_7	= Existencia de hyperlink de 3 a 1	==>	c_7 :	3→ 1	
c_8	= Existencia de hyperlink de 3 a 2	==>	c_8 :	3→ 2	
c_9	= Existencia de hyperlink de 3 a 3	==>	c_9 :	3→ 3	$3m = 9$

El ejemplo de la Figura 28 establece que existe hyperlink entre las páginas 1 y 2, 1 y 3, 3 y 1 y finalmente entre 3 y 2.

Figura 28: Ejemplo de representación del individuo genética con m=3 páginas

Bit =>	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
Link =>	1 --> 1	1 --> 2	1 --> 3	2 --> 1	2 --> 2	2 --> 3	3 --> 1	3 --> 2	3 --> 3
Individuo =>	0	1	1	0	0	0	1	1	0

Fuente: Elaboración propia

La generalización se puede apreciar en la Figura 29.

Figura 29: Generalización de individuos

c_1	c_2	c_3	...	c_m	c_{m+1}	c_{m+2}	...	c_{m+m}	c_{2m+1}	c_{2m+2}	...	c_{2m+m}	c_{3m+1}	...	c_{mxm}

Fuente: Elaboración propia

Es decir, cada componente c_i del individuo será:

$$c_i \begin{cases} 0 \text{ si } X_{kj}(i) = 0 \\ 1 \text{ si } X_{kj}(i) = 1 \end{cases}$$

Donde:

$$X_{kj}(i) \begin{cases} 0 \text{ si no existe hyperlink entre página } k(i) \text{ y } j(i) \\ 1 \text{ si existe hyperlink entre página } k(i) \text{ y } j(i) \end{cases}$$

Así, cada bit en la posición *i* del individuo, poseerá implícito una relación entre dos páginas de la forma indicada en la Ecuación 2.

Ecuación 2: Relación entre páginas y posición en el individuo lineal.

$$k = \left\lceil \frac{i}{m} \right\rceil \quad \text{Función cajón superior o función techo.}$$

$$j = \begin{cases} m & \text{si} \quad res(i,m) = 0 \\ res(i,m) & \text{si} \quad res(i,m) \neq 0 \end{cases}$$

Función *res* corresponde a la función residuo.

Se utiliza anteriormente la función cajón o también llamada función techo. Si corresponde a cajón superior se define como aquella función que se aplica a un número real *x*, y devuelve el mínimo número entero *k* superior a *x*, tal como se generaliza en la Ecuación 3. Por otro lado si es cajón inferior, busca el máximo número entero *k* inferior a *x*.

Ecuación 3: Función cajón inferior

$$\lceil x \rceil = \min\{k \in \mathbb{Z} \mid x \leq k\}$$
$$y = \lceil x \rceil : \quad y = \{y : \quad y \in \mathbb{Z} \quad \wedge \quad x \in \mathbb{R} \quad \wedge \quad y - 1 < x \leq y\}$$

La función *res(i,m)* por otro lado, simplemente indica el residuo de la división entre *i* y *m*, por lo que siempre arroja un valor entero. La Ecuación 4 indica la operación que permite obtener el residuo, usando la función cajón inferior.

Ecuación 4: Función residuo

$$res(i,m) = i - m \cdot \left\lfloor \frac{i}{m} \right\rfloor$$

3.3.- Fitness

La función fitness, tal como se indicó en la sección 2.5 Algoritmos Genéticos (AG), es clave para el correcto desarrollo, convergencia y ajuste de los datos reales [45], por lo que una clara definición es indispensable con el fin de obtener lo deseado. Antes de comenzar, se realizan cálculos previos de acuerdo al modelo a desarrollar (una vez ya realizado el proceso de limpieza de los datos detallado en la sección 4.4.- Aplicación del algoritmo), para luego establecer la función de adaptación de los mismos en búsqueda del individuo óptimo.

La Tabla 6 presenta un esquema con los conceptos que serán utilizados al momento de explicar los cálculos efectuados, junto a su explicación y representación esquemática de modo de facilitar el entendimiento de los cálculos posteriores.

Tabla 6: Conceptos relevantes

Concepto	Descripción	Representación
Acceso Directo	Se le llama acceso directo entre i y j al hecho de acceder desde la página i a la j usando el link directo que las une.	
Acceso Objetivo	Se le llama acceso objetivo entre i y j al hecho de acceder desde la página i a la j en algún momento en la sesión con páginas intermedias, es decir, hay una cantidad mayor a 1 de links para acceder desde una página a otra.	
Pasadas	Se le llama número de pasadas entre i y j a la cantidad de veces en que se pasa de i a j	
Posición	Se le llama posición de la página i, al momento en que accede a dicha página en una sesión en particular (1ra página visitada, 2da, 3ra, etc.)	
Largo	Se le llama largo entre i y j a la cantidad de links usados para pasar de i a j	
Factor Tiempo	Corresponde al tiempo neto (en segundos) en la página j en la sesión S, dividido por el tiempo total de la sesión S (sumatoria sobre el tiempo de todas las páginas que pertenecen a dicha sesión)	**Ecuación 5: Factor tiempo** $$f^{S}{}_{j} = \frac{t^{S}{}_{j}}{\sum_{j \in S} t^{S}{}_{j}}$$

Fuente: Elaboración propia

Se destaca la Ecuación 5 de la Tabla 6, allí se busca obtener la verdadera importancia del tiempo invertido en una página, en comparación a los tiempos visitando otras en la misma sesión. Lo anterior bajo el supuesto que *"el tiempo gastado en una página es proporcional al interés que el visitante tiene en su contenido*" [62], en otras palabras, a mayor tiempo en una página de una sesión, mayor fue el interés que su contenido y estructura despertaron en el usuario durante su visita. Ante eso, es necesario calcular el factor tiempo para hacer comparable la importancia entre páginas de distintas sesiones. Por ejemplo, visitar por 5 minutos una página *x* en una sesión de 10 minutos reporta un factor tiempo en *x* del 50%,

mientras que visitar 10 minutos una página *z* en otra sesión de 25 minutos indica un factor tiempo en z del 40%. Es decir, la página *x* tuvo mas importancia que *z*, pese a que *z* tuvo un mayor tiempo bruto (minutos) de visita.

Al comenzar con la construcción del fitness surgen las preguntas: ¿Cómo crear una función que represente una estructura de hyperlinks adecuada para navegar? ¿Cómo caracterizar la utilidad del uso de los hyperlinks? ¿Cómo dimensionar el potencial de crear un hyperlink que no existe y evidentemente, nunca ha sido usado? En la búsqueda de las respuestas más adecuadas a dichas preguntas, se recapitularon conceptos de teoría de grafos [23][32], cadenas de Markov[13], estructura web[5], etc., para poder dar forma a las expresiones en cuestión. El procedimiento desembocó en tres modelos tentativos, de esencia probabilística común, pero diferentes expresiones y principios matemáticos.

3.3.1.- Modelo 1: Caminos mínimos.

El primer modelo propuesto se denomina caminos mínimos pues al momento de realizar ciertos cálculos, usa el concepto de la ruta más corta para acceder desde una página a otra. La idea es construir el peso de cada hyperlink existente en base a su uso y a la importancia del tiempo de visita que tuvo la página de destino de dicho hyperlink (en una misma sesión). Por otro lado, si el hyperlink entre *i* y *j* no existente, su peso será construido en base a la multiplicación de los pesos de los hyperlinks existentes que pertenezcan al camino mínimo necesario para comunicar *i* y *j*.

Cálculos previos (fórmulas)

Primero, se busca calcular la frecuencia de cada hyperlink (v/s transiciones totales) , lo que se observa en la Definición 2.

Definición 2: Tiempo del hyperlink t_{ij}

$$\lambda_{ij} : \frac{\text{Número de veces de pasar de i a j vía acceso directo}}{\text{Número de todas las transiciones posibles desde i vía acceso directo}}$$

λ_{ij} corresponde a la fracción entre el número de veces de pasar de *i* a *j* con el total de veces de uso de hyperlinks que salen desde *i*, por ende, corresponde a la probabilidad condicional de acceder desde *i* a *j*, es decir, la probabilidad estacionaria [55] de uso del hyperlink o bien, la probabilidad de que transición[63] entre *i* y *j*.

Otra medida de representatividad de la importancia de cada hyperlink, corresponderá al factor tiempo en que se estuvo en la página final tal como se observa en Definición 3.

[63] **Ver Apéndice** C.- Conceptos utilizados de teoría de grafos y cadenas de Markov.

Definición 3: Tiempo del hyperlink t_{ij}

t_{ij}	:	Factor tiempo promedio de permanencia en página donde termina el acceso directo (en página *j*) iniciándose en página *i*.

En este contexto entonces, t_{ij} corresponde al valor promedio del factor tiempo en la página destino *i* del hyperlink que parte en *i*. Específicamente, se asume que el tiempo de permanencia en la página destino, tiene que ver con la importancia del hyperlink que la llevó a ella, pues a mayor tiempo en la página de destino, más útil fue el hyperlink que le permitió acceder a ella.

Posteriormente, se construyó el peso de los hyperlinks existentes tal como se define en la Ecuación 6.

Ecuación 6: Peso del hyperlink existente *ij*

$$\pi_{ij} = \lambda_{ij} \cdot t_{ij}$$

Desde ahora a π_{ij} se le llamará peso del hyperlink, pues refleja que tan importante es al depender de la probabilidad de uso y del factor tiempo de la página destino. Será mayor a mayor factor de probabilidad λ_{ij} y mayor, a mayor factor tiempo de permanencia en página destino. Lo anterior se sustenta por proporcionalidad directa

Aplicando la Ecuación 6 en los datos, se desprende que sólo se podrán tener los valores de π_{ij} donde existe hyperlink entre *ij*, pues en dichas transiciones será posible calcular λ_{ij} y $t_{ij,}$ Así, se podrá decidir sobre mantener o borrar hyperlinks existentes de acuerdo a su peso. Por otro lado, también se espera poder crear nuevos enlaces en base al valor que reportaría su posible existencia considerando el comportamiento del consumidor. Lo anterior se realizará calculando los caminos existentes para llegar de una página a otra en el caso de no poseer link directo entre ellas.

Sea P el conjunto de todos los caminos posibles en un grafo y $p \in P$ un camino en particular de largo l_p que va entre *i* y $j_{lp,}$ entonces se define un camino, como se indica en el recuadro de Definición 4.

Definición 4: Camino

$$p = \left\{(i_1, j_1), (i_2, j_2), (i_3, j_3), (i_4, j_4), ..., (i_{l_p}, j_{l_p})\right\}$$

Tal que $j_k = i_{k+1}, \forall k \in [1, l_p - 1]$

Así, para determinar los pesos de hyperlinks no existentes[64], se propone primero calcular la probabilidad de ir de ir de un nodo *i* a un nodo *j* donde no hay hyperlink (λ'_{ij}), tal como se obseva en la Ecuación 7.

[64] Peso de los hyperlinks no existentes también serán llamados pesos potenciales

Ecuación 7: Probabilidad mediante caminos posibles

$$\lambda'_{ij} = \sum_{p \in P_{(i,j)}} \prod_{(k,h) \in p} \lambda_{kh}$$

La Ecuación 7 se sustenta en conceptos de probabilidades [55] en el sentido que para ir de *i* a *j*, puedo tomar ya sea el camino 1 (que consiste en pasar de *i* a *x*, de *x* a *y* y de *y* a *j* por ejemplo), o bien el camino 2 (pasar de *i* a *z* y de *z* a *j* por ejemplo), o una tercera forma, y así sucesivamente. Dicha cantidad de formas posibles está determinada por la existencia de enlaces en todo el grafo, a mayor cantidad de ellos, habrá una mayor cantidad de recorridos disponibles.

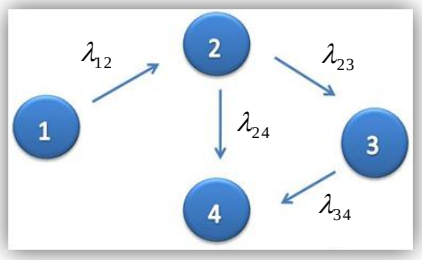
La Ecuación 7 también tiene implícito que mientras un camino posible se haga cada vez más grande (visite más páginas), el valor numérico de la multiplicación de sus probabilidades disminuye (multiplicación de valores menores que uno). Por eso, en vez de tomar todos los caminos posibles, sólo se considerará el **camino mínimo,** el cual corresponde a la menor cantidad de pasos para ir de una página a otra. Lo anterior, se sustenta tanto matemáticamente (en el sentido que los valores resultantes de camino de largo mayor se considerarían despreciables) como también por una lógica de comportamiento, en relación a que la probabilidad de un hyperlink no existente debe reflejar las relaciones posibles en el menor camino para el cálculo del peso potencial. En otras palabras, si el peso potencial de un hyperlink es mejor que el menor camino posible, convendría agregar dicho hyperlink en cuestión, pues introduciría un camino aún menor y de mayor peso al crear el acceso directo. Cabe destacar que a menor cantidad de hyperlinks para llegar a la información deseada, mayor es el beneficio de cada usuario [61]). La ecuacion final para el cálculo de las probabilidades potenciales (λ'_{ij}) se observa en la Ecuación 8.

Ecuación 8: Peso potencial camino mínimo

$$\lambda'_{ij} = \prod_{(k,h) \in p^{\min}} \lambda_{kh}$$

Por ejemplo, al obsevar la figura 30, se desprende que falta el peso directo entre 1 y 4, pues no existe hyperlink. En este caso, el camino mínimo es 1 => 2 => 4 en vez de 1=> 2 => 3=> 4, que es otro camino posible. Con dicho contexto, λ'_{14} quedaría como $\lambda'_{14} = \lambda_{12} \cdot \lambda_{24}$

Figura 30: Ejemplo hyperlinks entre páginas



Fuente: Elaboración propia

Luego, con el cálculo de las probabilidades potenciales λ', se procede a calcular los pesos potenciales, normalizando por fila dichas probabilidades, de modo de mantener la ley de transición (la suma de todas las probabilidades de transición por fila es igual a la unidad[65]). Se asume por ley de los grandes números [55] que el tiempo promedio de la página de destino se mantiene.

Ecuación 9: Pesos potenciales modelo 1

$$\pi'_{ij} = \lambda'_{ij} \cdot t_{ij}$$

La ecuación 9 refleja el cálculo de los pesos potenciales del modelo 1, que se realiza en forma preliminar con todos los pesos no existentes y cuyo camino mínimo sea posible de calcular.

Razonamiento matemático

La fórmula que sustenta los términos de la Ecuación 7, se refleja en la Ecuación 10, donde la probabilidad de ir de *i* a *j*, será la suma de llegar desde *i* a *j* vía un camino de largo 1, de largo 2, de largo 3, etc., siendo *l* el largo del camino. Por ende, la probabilidad de ir de *i* a *j*, estará dada por la suma de todas las probabilidades de acceso entre dichas páginas mediante todos los caminos posibles.

Ecuación 10: Caminos posibles

$$\lambda_{ij} = \sum_{p^{(l)} \in P(i,j), \forall l} \lambda_{ij}$$

Así, para obtener la probabilidad para acceder de *i* a *j*, se restringe la ecuación anterior, considerando solo el largo del camino mínimo, despreciando las probabilidades restantes en un $\vartheta(l)$ tal como se indica en la ecuación 11.

Ecuación 11: Aproximación de probabilidad de transición

$$\lambda_{ij} \approx \lambda_{ij} \mid \{p^{(cam\_\min)} \in P(i,j)\} + \vartheta(l)$$

[65] **Ver sección** C.- Conceptos utilizados de teoría de grafos y cadenas de Markov

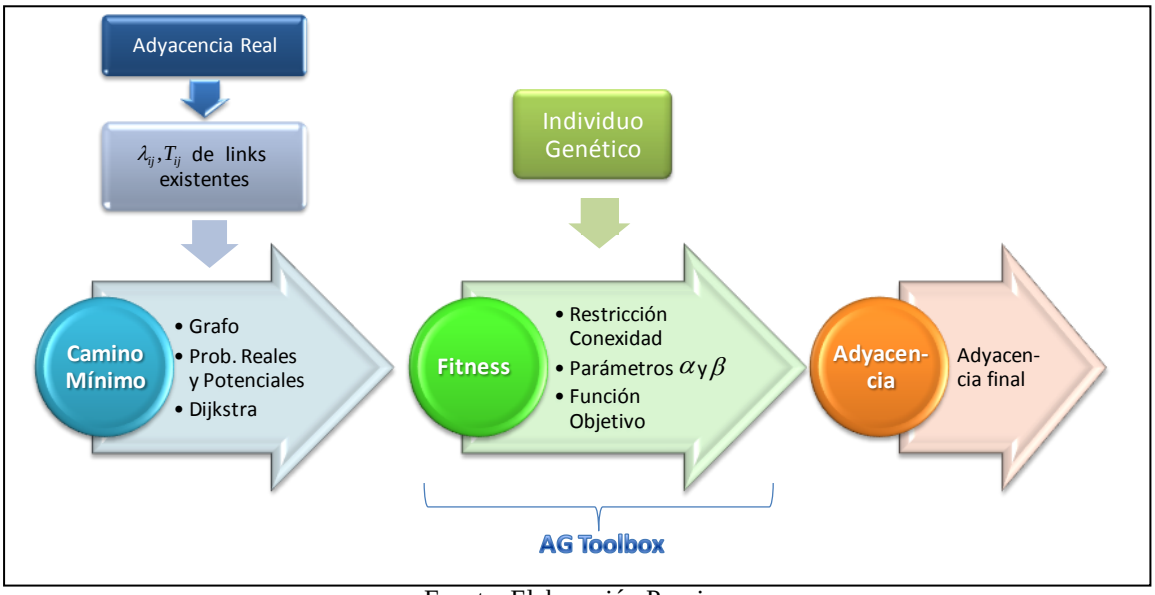
Cabe destacar que el valor de π_{ij}, cabe siempre en la definición de probabilidad, dado que se encuentra en el rango {0,1} y además, permite a partir de los parámetros conocidos, relizar predicciones acerca de los valores que toma la variable aleatoria (probabilidad de acceso entre hyperlinks inexistentes).

Seudocódigo Modelo 1[66]

La función fitness se compone de tres códigos principales, separados por lógica de trabajo pero comunicados entre sí. Recibe como dato de entrada la matriz de adyacencia actual y como variable, cada individuo genético de la población que el algoritmo crea aleatoriamente en base a los operadores genéticos.

La Figura 31 es una representación esquemática del procedimiento de los códigos realizados, cuyo detalle se explica en los puntos que siguen.

Figura 31: Esquema etapas modelo 1



Fuente: Elaboración Propia

1.- Camino Mínimo

El código de camino mínimo, recibe tanto la adyacencia como las probabilidades y tiempo de los hyperlinks existentes. Así, se calculan los pesos de dichos hyperlinks cumpliendo con la Ecuación 6, formando la matriz de pesos P_{mxm}. Luego, se realiza el cálculo de camino mínimo aplicando la función *graphshortestpath(DG,ini,fin)*[72] en Matlab, la cual se basa en el algoritmo de Dijkstra[67] en su proceder. 'DG' corresponde al grafo que indica los hyperlinks existentes de manera dirigida captada del sitio web, del cual puede desprenderse su gráfica gracias a la función *view(biograph(DG,[],'ShowWeights','on'))*[72]; 'ini' señala al grupo de páginas de inicio desde los que se quiere iniciar el recorrido para obtener el camino

[66] Código formal en Matlab en sección 7.5.2. Código fitness en Matlab
[67] **Fuente:** http://www-m3.ma.tum.de/foswiki/pub/MN0506/WebHome/dijkstra.pdf

mínimo, y 'fin', el grupo de nodos hacia donde se quiere llegar desde los anteriores. Obtenidos los caminos mínimos, se van multiplicando las probabilidades asociadas para obtener la probabilidad potencial según la Ecuación 8. Obtenida la probabilidad se calculan los pesos potenciales (Ecuación 9) agregándolos en la matriz *P* que servirá de entrada para la función fitness.

Cabe señalar que el algoritmo de Dijkstra asume que los pesos de los arcos son de valores positivos en una matriz de adyacencia de *mxm*. La complejidad[68] en tiempo es *O(log(P)*L)*, donde *P* es el número de páginas y *L* el número de arcos [72].

2.- Fitness F

Corresponde al código central que es aplicado en el toolbox de los algoritmos genéticos, el cual recibe como entrada los individuos genéticos R^* (lineales) que se van construyendo y entregando en cada iteración. En Matlab se crea esta función indicando:

function *F*= Fitness(R^*)

Los pasos del fitness comprenden:

2.1.- Llevar individuo lineal R^* a matriz cuadrada (pues es lineal como entrada) para mayor comodidad en los cálculo y uso de las funciones matriciales propias de Matlab.
2.2.- Restringir por conexidad mediante el método de matriz laplaciana[69], de modo que siempre haya un camino posible para llegar a cualquier página. Para esto, si no se cumple dicha restricción, se la da a *F* un valor arbitrariamente muy alto[70] (1E7)
2.3.- Multiplicar punto por punto la matriz de R^* por la matriz de Pesos *P* (al resultado de dicha operación se le llamará Beneficio). Cabe destacar que si los pesos son nulos, no afecta el valor final de la operación, pues simplemente implicará sumar 0.
2.4.- Precisar la función objetivo *F*, que corresponderá a la resta de la sumatoria de hyperlinks totales (punto 2.4) con la multiplicación de pesos por R^* (punto 2.4.), es decir, se busca minimizar la cantidad de hyperlinks totales de modo que se maximice el beneficio.

Con lo anterior, una restricción de conexidad estará involucrada y el concepto de penalización al no cumplirla. Finalmente, la ecuación objetivo a maximizar queda expresada en la Ecuación 12.

[68] La complejidad de un algoritmo se mide en cuanto al tiempo de ejecución y/o la cantidad de memoria que requiere del computador para su procesamiento, en términos del número de operaciones que realiza. http://www.lab.dit.upm.es/~lprg/material/apuntes/o/index.html
[69] **Ver apéndice** sección C.- Conceptos utilizados de teoría de grafos y cadenas de Markov
[70] Se le da un valor muy alto pues por default la aplicación de AG en Matlab busca la minimización

Ecuación 12: Función objetivo modelo 1

$$MaxU(R,\pi) = \max\left[\underbrace{\alpha \cdot \sum_{ij,\pi_{ij}\neq 0} R_{ij} \cdot \pi_{ij}}_{\text{Beneficio}} - \underbrace{\beta \cdot \sum_{ij} R_{ij}}_{\text{Hyperinks totales}}\right]$$

3.- Adyacencia Final

Luego de ejecutar la función fitness en la entrada de AG del modo @fitness[71], se obtendrá el individuo resultante deseado. Para poder utilizarlo se exporta como variable en Matlab, donde se encontrará disponible como tabla, lo que es necesario para poder estudiar los cambios generados. Se procede entonces con los siguientes pasos:

3.1.- Exportar matriz resultante del algoritmo.
3.2.- Restar matriz generada con la matriz de adyacencia inicial donde se obtendrá valor 1 cuando un hyperlink fue creado por el algoritmo, -1 si el hyperlink fue eliminado y 0 cuando la situación fue mantenida.
3.3.- Sacar estadísticas de valores anteriores (cantidades) y diferencia porcentual de beneficio y cantidad de hyperlinks.
3.4.- Generar resultados específicos de creación de hyperlinks o eliminación de los mismos con miras a su implementación y posterior pruebas de usabilidad.

3.3.2.- Modelo 2: Tasas de uso lineal.

Con el segundo modelo, se busca evitar la complejidad numérica acontecida por construcción de los pesos potenciales mediante el uso del camino mínimo vía acceso directo, ahora en cambio, se propone la creación de tasas de uso basadas en el concepto de acceso objetivo indicado en la Tabla 6: Conceptos relevantes. En otras palabras, la idea será construir las frecuencias de uso de los hyperlinks, considerando las rutas seguidas para ir de un nodo a otro, independiente de la cantidad de páginas visitadas entre ellas.

Cálculos previos (fórmulas)

La cantidad de veces que se pase desde una página otra es independiente del número de pasos requeridos para ello, se agrupa en un solo gran conteo realizado sobre todas las sesiones, determinando la frecuencia de llegada a la página *j* desde el nodo *i*, tal como se observa en Definición 5.

[71] Forma de llamar la función fitness en la ventana de diálogo (Tabla B. 2: Interfaz AG **en Matlab**) del toolbox de AG en Matlab

Definición 5: Frecuencia de llegada

$$f_a(i,j) = \frac{\text{Cantidad de pasadas desde } i \text{ a } j \text{ vía acceso objetivo y/o directo}}{\text{Cantidad de pasadas desde } i \text{ a cualquier página vía acceso objetivo y/o directo}}$$

La frecuencia de llegada entonces, representa la probabilidad condicional de pasar desde la página *i* a la *j*, al realizar el cálculo sobre todo el universo de sesiones.

Por otro lado, en la Definición 6, se presenta la frecuencia de estadía, que corresponde al factor tiempo condicional de visita a la página de destino.

Definición 6: Frecuencia de estadía

$$f_t(i,j) = \frac{\text{Suma de Factor tiempo en } j \text{ al acceder desde } i \text{ vía acceso objetivo y/o directo}}{\text{Suma del Factor tiempo de las páginas visitadas desde } i \text{ vía objetivo y/o directo}}$$

Con ambos valores, se procede a realizar el cálculo de los pesos como se indica en la Ecuación 13.

Ecuación 13: Pesos potenciales modelo 2

$$\overline{\pi}_{ij} = f_a(i,j) \cdot f_t(i,j)$$

Este nuevo peso lleva implícito de que a mayor frecuencia de llegada y factor tiempo en la página de destino, mayor será la importancia del hyperlink.

Razonamiento matemático

Se puede desprender que numéricamente se dispondrá de los pesos para la mayoría de las relaciones entre todas las páginas, sin embargo, habrá pares de páginas con pesos nulos. Esto significa que ninguna de las sesiones accedió desde *i* hacia *j*, ni siquiera como página de inicio y fin (máxima distancia entre ellas). Ante ello, dicho hyperlink tendrá peso nulo siempre, pues por la gran cantidad de datos desde los que se realizaron los cálculos, se puede establecer que su uso no sucede (ley de los grandes números [12][55]).

Por otra parte, se está en un contexto de probabilidades condicionales, donde finalmente π_{ij} corresponderá al peso condicional de acceso a *j* dado que se está en *i*, incorporando la probabilidad de ocurrencia de la interacción *i-j* y la probabilidad de permanencia en la página de destino.

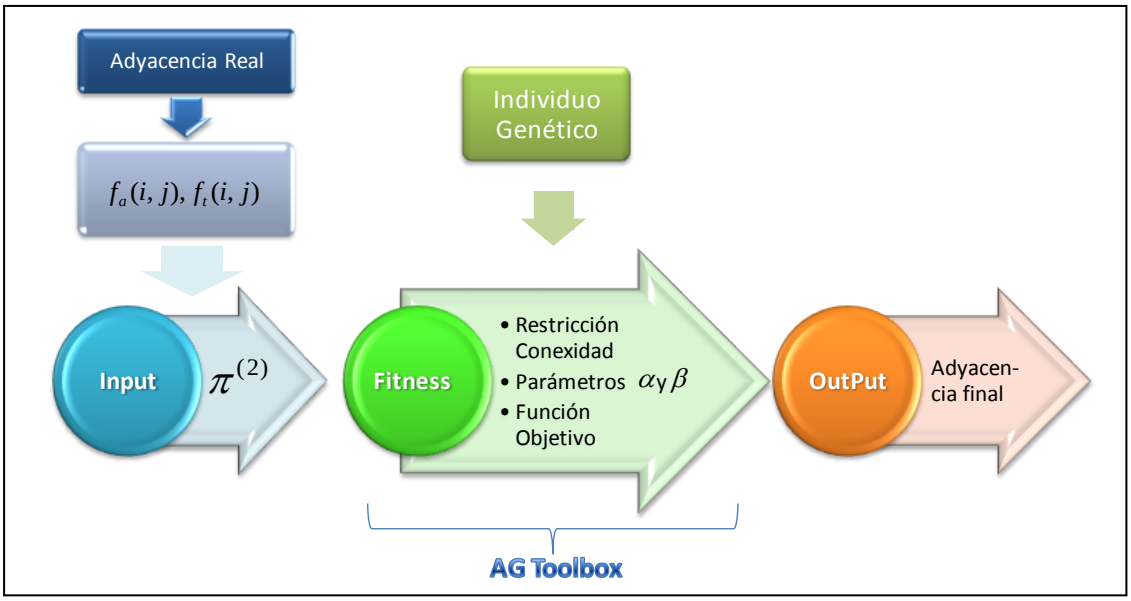
Seudocódigo Modelo 2[72]

Para la construcción del fitness, nuevamente se hace una separación en tres etapas, esta vez llamadas input: carga de datos, fitness: función objetivo y output: estructura final. La

[72] Código formal en Matlab en apéndice E.2. Código fitness en Matlab

Figura 32 esquematiza los principios y etapas fundamentales del modelo 2, donde $\pi^{(2)}$ corresponde a la matriz final de pesos dados por la Ecuación 13.

Figura 32: Esquema etapas modelo 2



Fuente: Elaboración propia

1.- Input

Esta etapa inicial consiste en aplicar la Ecuación 11 de modo de determinar el peso. Cabe señalar que más que el peso de un hyperlink como en el modelo 1, $\overline{\pi}$ es la matriz que contiene el peso de los pasos objetivos entre dos pares de páginas, es decir, el hecho de poseer un valor más alto indica que hay más interés en la comunicación entre ambas páginas existiendo o no existiendo hyperlink.

2.- Fitness

Al igual que en el modelo 2, se construye la función que recibe todas las potenciales soluciones R^* que se van construyendo y entregando en cada iteración, del modo

function F= Fitness(R^*)

Los pasos del fitness en el caso del modelo 2 comprenden:

2.1.- Llevar individuo lineal R^* a matriz cuadrada
2.2.- Restringir por conexidad mediante el método de matriz laplaciana, sino penalizar.
2.3.- Multiplicar punto por punto la matriz de R^* por la matriz de Pesos $\overline{\pi}_{ij}$ (al resultado de dicha operación se le llamará beneficio). Cabe destacar que si los pesos son nulos, al ser una sumatoria, simplemente suma 0 por lo que numéricamente no se alteran los resultados.
2.4.- Hacer la sumatoria sobre R^* para obtener la cantidad de hyperlinks totales.

2.5.- Precisar el valor *F* de la función objetivo indicada en la Ecuación 14, con el fin de maximizar dicho valor.

Ecuación 14: Función objetivo modelo 2

$$MaxU(R,\bar{\pi}) = \max\underbrace{\left[\alpha \cdot \sum_{ij,\pi_{ij}\neq 0} R_{ij} \cdot \bar{\pi}_{ij}\right.}_{\text{Beneficio}} \underbrace{\left. - \beta \cdot \sum_{ij} R_{ij}\right]}_{\text{LinksTotales}}$$

3.- Adyacencia final

Luego de ejecutar la función fitness en la entrada de AG, se obtendrá una vez cumplido algún criterio de detención, el individuo resultante deseado, procediendo exactamente igual a los pasos indicados en el punto 3 (Adyacencia final) del modelo 1.

3.3.2.- Modelo 3: Tasas de uso potencial.

Finalmente el modelo 3, corresponde a una variación del modelo 2, ahora separando por cantidad de hyperlinks necesarios para acceder de una página a otra, es decir, la idea es desprender la cantidad de pasos necesarios, por lo que se tiene la influencia del largo de cada sesión y posición de cada página en ella. Así, se construirán los pesos calculando las frecuencias de paso de *i* a *j* en *k* hyperlinks.

Cálculos previos (fórmulas)

Como se indicó, el modelo 3 es bastante similar al modelo 2, salvo que ahora se disponen de *H* matrices de pesos separadas entre sí, dependiendo de la cantidad *k* de hyperlinks ocupados entre las páginas. Cabe señalar que $k \in IN \cap [1,H]$.

La cantidad de veces que se pasa desde una página a otra entonces, ahora se separa de acuerdo al número de pasos requeridos para ello, determinando la frecuencia de llegada a la página *j* desde el nodo *i* en *k* hyperlinks, tal como se establece en la Definición 7.

Definición 7: Frecuencia de llegada en *k* pasos

$$f_a(i,j)^{(k)} = \frac{\text{Cantidad de pasadas desde i a j vía acceso objetivo y/o directo en k pasos}}{\text{Cantidad de pasadas desde i a cualquier pág. vía acceso objetivo y/o directo en k pasos}}$$

La frecuencia de llegada en *k* pasos, representa la probabilidad condicional de pasar desde la página *i* a la *j* usando *k* hyperlinks, al realizar el cálculo sobre todo el universo de sesiones que realizan dicha acción.

Por otro lado, en la Definición 8, se presenta la frecuencia de estadía, que corresponde al factor tiempo condicional de visita a la página de destino, luego de visitar *k-1* páginas previamente.

Definición 8: Frecuencia de estadía en *k* pasos

$$f_t(i,j)^{(k)} = \frac{\text{Suma de Factor tiempo en j al acceder desde i vía acceso objetivo y/o directo en k pasos}}{\text{Suma del Factor tiempo de las páginas visitadas desde i vía objetivo y/o directo en k pasos}}$$

Con ambos valores, se procede a realizar el cálculo de los pesos como se indica en la Ecuación 15.

Ecuación 15: Pesos potenciales modelo 3

$$\pi^{(k)}{}_{ij} = f_a(i,j)^k \cdot f_t(i,j)^k$$

Es decir, se tendrán tantas matrices tipo $\pi^{(k)}$ como el largo máximo de paso entre una página y otra. π_{ij} (modelo 1) solo contiene la multiplicación de frecuencias de llegada y factor tiempo asociadas a cuando hay un hyperlink entre las páginas, $\overline{\pi}_{ij}$ (modelo 2) multiplica las frecuencias cuando hay 2 hyperlinks entre las páginas y así sucesivamente.

Razonamiento matemático

Se desprende que en el presente modelo habrán muchas matrices $\pi^{(k)}$ sobre todo si el largo máximo de sesiones se hace cada vez más grande. Es por eso que es necesario definir una cantidad *H* máxima menor al largo de la sesión que más páginas visita. En este sentido, la navaja de Occam[73] [25] induce la necesidad de establecer un largo adecuado en vez de todos los largos posibles y por ende evitar el gran costo computacional que eso significa. Dicho concepto se refiere a que han de preferirse teorías más simples a las más complejas, sobre todo en el mundo de la informática ante la creciente complejidad de los algoritmos y sistemas informáticos. Pese a que no se considera como un resultado científico, se ha utilizado como una regla heurística para orientar el desarrollo de modelos teóricos, teniendo resultados favorables y empírica justificación. En el caso del presente modelo, pueden haber *N* sesiones y el 0.02% de ellas pude tener largo mayor a *H*. Si se contempla todos los largos en la ecuación 13, el tiempo computacional crece enormemente, por lo tanto, como las sesiones de largo mayor a *H* es un porcentaje menor (pero representativo) de los datos totales, será mejor fijar el cálculo hasta *H*. De este modo se evita el costo de calcular sobre todos los largos posibles, a fin de que no se pierdan muchos datos (y por ende se mantenga la representatividad de éstos en los datos) y se facilite el cálculo con los que se consideran.

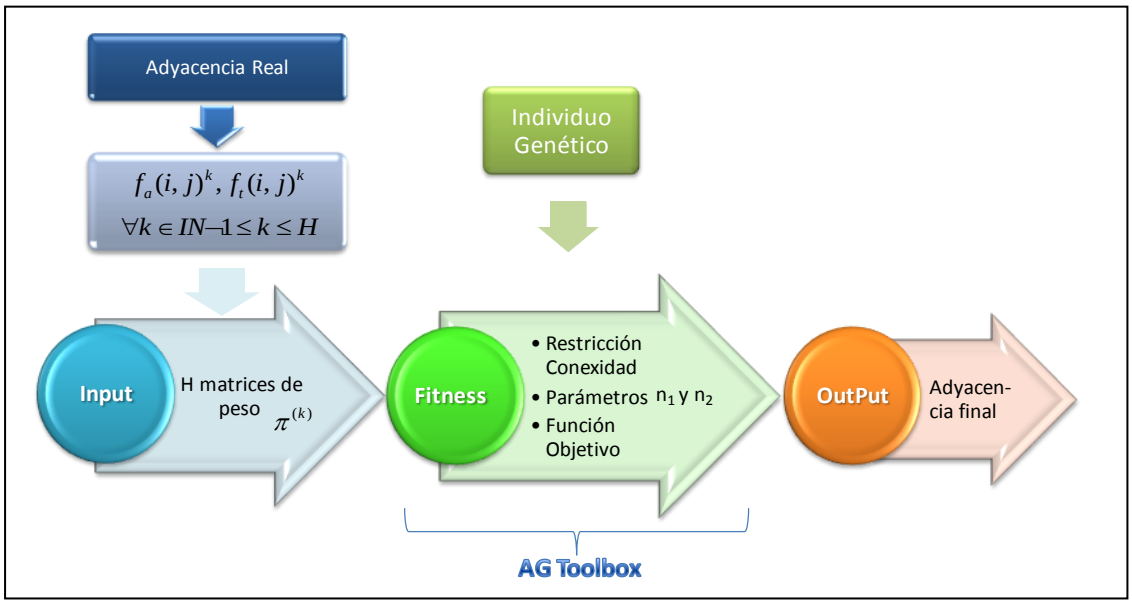
[73] **Fuente:** http://plato.stanford.edu/entries/simplicity/

Cabe destacar que a medida que aumenta *k*, la matriz $\pi^{(k)}$ contendrá cada vez menos valores, lo que es lógico desde el punto de vista de la ley de Zipf[74]: Hay pocas sesiones con largos cada vez mas grandes.

Seudocódigo Modelo 3[75]

El Modelo 3 en general, será muy símil al desarrollado en el modelo 2, a tal punto que se podría decir que será *H* veces el mismo procedimiento, siendo *H* el máximo largo considerado. La Figura 33 muestra el esquema de las etapas principales del modelo 3.

Figura 33: Esquema etapas modelo 3



1.- Input

La entrada para el modelo serán las H matrices $\pi^{(k)}$ con $k \in IN \neg 1 \leq k \leq H$.

$\pi^{(k)}$ contendrà los valores de los pesos entre las páginas cuando *k* hyperlinks se usan para acceder entre ellas.

2.- Fitness

Al igual que en los modelos anteriores, se construye la función que recibe todas las potenciales soluciones R^*. La principal diferencia, además de la entrada de diferentes matrices que separan la cantidad de clicks usados para acceder de una página a otra, corresponde a las restricciones y función objetivo. Una vez más la conexidad es necesaria, pues la matriz final debe permitir el acceso a todas las páginas, sin embargo, se agrega ahora una restricción en cuanto a la cantidad de hyperlinks dentro de cierto rango, de modo de maximizar el beneficio propio del modelo 3. En detalle, los pasos a seguir son:

[74] **Ver sección** 2.1.2.- Principios de la Web
[75] Código formal en Matlab en sección 7.5.2. Código fitness en Matlab

2.1.- Llevar individuo lineal *R* a matriz cuadrada

2.2.- Restringir por conexidad mediante el método de matriz laplaciana, si *R* no cumple, penalizar.

2.3.- Hacer la sumatoria sobre *R* para obtener la cantidad de hyperlinks totales (num_link) y restringir a que dicho valor esté entre los parámetros n_1 y n_2, del modo: $n_1 \leq num\_link \leq n_2$. Si *R* no cumple, penalizar.

2.4.- Definir la matriz potencia de *R*, es decir, *R* elevado a k de la forma: $R^{(k)}$. Dicha operación establece la cantidad de caminos que *R* permite, en otras palabras $R_{ij}^{(k)}$ corresponderá al valor de la fila *i* y columna *j* de la matriz *R* elevada a *k*, donde dicho valor es la cantidad de caminos de largo *k* que existen para ir de *i* a *j*.

2.5.- Multiplicar punto por punto la matriz de potencia *k* de *R* por la matriz de pesos $\pi^{(k)}$ correspondiente. El resultado de dicha operación se le llamará beneficio potencial.

2.6.- Precisar el valor *F* de la función objetivo indicada en la Ecuación 16, con el fin de maximizar dicho valor.

Ecuación 16: Función objetivo modelo 3

$$Max\{U(R)\} = \max\{\sum_k \sum_{ij} R_{ij}^{\ k} \cdot \Pi_{ij}^{\ (k)}\}$$

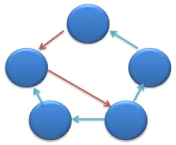
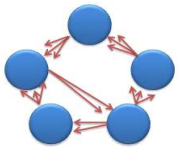
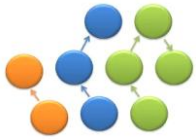
3.- Adyacencia Final

Nuevamente obtener la función fitness en la entrada de AG, exportando su valor y estructura matricial binaria, del mismo modo explicado en los modelos anteriores.

Así, finalmente se da forma al diseño de tres modelos a fin de acercarse con ellos a un mejor ajuste de resultados, estudiar sus ventajas, desventajas y resultados. La Tabla 7 presenta un cuadro comparativo resumido entre ellos. Dicha tabla permite visualizar la codificación del tipo de entrada de cada modelo (símbolo que representa cada método), el principio fundamental que los subyace (camino mínimo, acceso objetivo, acceso objetivo por camino), la fórmula de cálculo de cada peso en concreto, las restricciones propias de los algoritmos [76] (donde sólo el modelo 3 es restringido por número de links), los parámetros que necesitan ser cuantificados y finalmente la función objetivo que busca ser maximizada.

[76] **Ver apéndice** E.2. Código fitness en Matlab

Tabla 7: Resumen modelos

	Modelo 1	Modelo 2	Modelo 3
			
	Camino mínimo	**Lineal**	**Potencial**
Entrada	Matriz de pesos π	Matriz de pesos $\bar{\pi}$	H matrices de peso $\pi^{(k)}$
Principio	Construir peso de links no existentes en base al camino mínimo	Multiplicación de frecuencia de llegada por factor tiempo de permanencia.	Separación de frecuencia de llegada y factor tiempo en relación a la cantidad de links para pasar de un nodo a otro.
Pesos	$\pi_{ij} = \lambda_{ij} \cdot t_{ij}$	$\bar{\pi}_{ij} = f_a(i,j) \cdot f_t(i,j)$	$\pi^{(k)}{}_{ij} = f_a(i,j)^k \cdot f_t(i,j)^k$
Restricciones	Conexidad	Conexidad	Conexidad y número de links
Parámetros	α y β	α y β	n_1 y n_2
Función Objetivo	$\max\left[\alpha \cdot \sum_{ij,\pi_{ij}\neq 0} R_{ij} \cdot \pi_{ij} - \beta \cdot \sum_{ij} R_{ij}\right]$	$\max\left[\alpha \cdot \sum_{ij,\pi_{ij}\neq 0} R_{ij} \cdot \bar{\pi}_{ij} - \beta \cdot \sum_{ij} R_{ij}\right]$	$\max\{\sum_k \sum_{ij} R_{ij}{}^k \cdot \Pi_{ij}{}^{(k)}\}$

Fuente: Elaboración propia

4.- Aplicación en sitio web

Lo más interesante, desde el punto de vista comercial de un sitio web, es poder analizar y obtener patrones de comportamiento de sus usuarios desde las bitácoras de acceso (Logs). Por ejemplo: "*una página que nunca es visitada tal vez no tiene razón de ser, o si páginas muy visitadas no están en los primeros niveles, esto sugiere mejorar la organización y navegación del sitio*" [16]. Al determinar lo anterior, los resultados pueden ser determinados para recomendar servicios o productos ya sea de forma general o para usuarios específicos, lo que conlleva a la una personalización dinámica de los sitios web [60]. De este modo y con el fin de mejorar la estructura de hyperlinks de un sitio mediante los registros de visita del mismo y el modelo propuesto, se procede con la aplicación en un sitio web real.

4.1.- Descripción del sitio

El sitio a estudiar es la página principal del Magíster de Gestión de Operaciones (MGO), sub sitio del Departamento de Ingeniería Civil Industrial (DII) cuya URL corresponde a: *www.dii.uchile.cl/~mgo/* con dirección IP 200.9.100.73. Su página principal se presenta en la Figura 34:

Figura 34: Home *www.dii.uchile.cl/~mgo/*

Fuente: Extracto de la página www.dii.uchile.cl

De acuerdo a la clasificación indicada en la Figura 9, *www.dii.uchile.cl/* y por ende el sub sitio del MGO, es un sitio web del tipo público (pues se tiene acceso a todo su contenido), estático (contenidos y formas a cargo de diseñadores), abierto (acceso a todas las páginas desde cualquier punto), de estructura mixta (jerárquica con enlaces cruzados) y comercial corporativo.

Los usuarios tienen un perfil académico, en el sentido que buscan información asociada a temas universitarios, ya sea de pregrado o postgrado, mallas curriculares, grafo de los académicos, cursos ofrecidos, etc. Los usuarios que visitan el sitio lo hacen tanto vía

searching como browsing[77], es decir, pueden llegar buscando en la Web alguna página en especial o navegando mediante las distintas páginas del sitio directamente. Los usuarios pueden ser clasificados como estudiantes de pregrado, en vías de titulación, estudiantes de algún postgrado en particular, titulados y profesores, provenientes tanto de la Universidad de Chile como de otros recintos educativos tanto nacionales como internacionales.

Actualmente los cambios administrativos en el sitio del MGO se realizan mediante un editor de SODA[78], empresa que realiza páginas Web y esqueletos de ellas facilitando su modificación. La tasa de frecuencia de cambios en las páginas es estacional, pues la cantidad de ellos depende de cada mes (no hay registros de valores numéricos concretos asociados) que principalmente se relacionan a modificaciones por nuevas noticias, egresados, realización de seminarios, congresos y publicaciones. Cabe señalar que si bien el dinamismo del sitio es amplio, el estudio se realizará captando una adyacencia fija en un momento determinado. Lo anterior no afecta significativamente pues la estructura básica se mantiene permanentemente, dado que su renovación implica en gran medida un cambio en contenidos, es decir, en mayor cantidad de cambios sólo afecta la información interna[79].

Como descripción preliminar, en la Tabla 8 se observan las principales rutas corporativas del sitio del DII, donde se destacan los grupos principales de páginas web del sitio, para conocer el contexto específico donde el sitio del MGO se encuentra inserto.

Tabla 8: Rutas principales de www.dii.uchile.cl

Principales rutas de www.dii.uchile.cl	
Ingenieria Civil Industrial	Admisión, Plan de Estudios, Boletines
Boletin	Set de Boletines Informativos de noticias destacadas por fechas
Tutoría	Tutoría para estudiantes
Pregrado	Sitio de Pregrado
Magister/Doctorados	Set de Magíster y Doctorados
MGPP	Magíster en Gestión y Políticas Públicas
MBA	Magíster en Gestión y Dirección de Empresas
MBE	Magíster en Ingeniería de Negocios
MGO	Magíster en Gestión de Operaciones
MAGCEA	Magíster en Economía Aplicada
Eduación Continua	Noticias intitución, programas, diplomas
Diplomas	Diplomas de Postítulos ofrecidos
Investigación	Areas de investigación y proyectos destacados
CEA	Centro de Economía Aplicada
CGO	Centro de Gestión de Operaciones
CEGES	Centrp de Gestión
Publicaciones	Papers por área, libros, artículos, documentos de trabajo
RIS	Revista de Ingeniería de Sistemas
EYG	Economía y Gestión
Información	Misión, Visión, Valores y Fortaleza Institución
Académicos	Profesores jornada completa, media o adjuntos.
Ex Alumnos	Corporación ICI y Oficina de Colocaciones
Galerías	Fotografías anuales

Fuente: Elaboración propia en base a la estructura de www.dii.uchile.cl

[77] **Ver sección**: 2.4.1.- Conductas de navegación
[78] **Fuente:** http://www.thesodastudio.com/
[79] La información descriptiva indicada se obtuvo al entrevistar a la persona encargada de administrar el sitio web, Srta. Julie Lagos, secretaria administrativa del MGO (entrevista Octubre 2009)

Gracias al apoyo del DII, se logró recopilar información de las visitas de los usuarios en todo su sitio Web. El intervalo de tiempo de registro de datos fue de 5 meses, en detalle:

Tiempo primer registro:	2009-05-22 16:59:33
Tiempo último registro:	2009-10-04 16:08:28

Dichas fechas fueron obtenidas gracias a la función de conversión FROM_UNIXTIME()[80], la cual recibe como argumento el tiempo en formato unix (que es el tiempo en que se encuentran los registros) para entregar el formato fecha tradicional.

Considerando la estructura de la Tabla 8, se alcanza una cantidad de **615** páginas y un total de **128.937** registros[81] de visita en todo el sitio del DII. Considerando desde las sesiones de largo 2, todo lo anterior se traduce en un total de **72.242** transiciones[82] entre páginas, lo que corresponde a una o más visitas a uno de los **7.059 hyperlink**s detectados. Gracias a la recuperación autorizada de datos vía cookies y posterior proceso de sesionalización que se explica en detalle en la sección 4.1.2.- Sesionalización, se obtiene un total de **55.692** sesiones, que poseen un largo promedio de 2,3 páginas visitadas. Sin considerar las sesiones de largo 1 (42% de las sesiones), el promedio asciende a 4,19 páginas visitadas por sesión, con un tiempo promedio en cada una de 56 segundos.

Ahora, dada la considerable cantidad de datos con los que se dispone y a que se tiene como por objetivo realizar una primera aproximación de los modelos a estudiar con datos concretos (siempre indicando todo en forma general el algoritmo para cualquier cantidad de datos y expansión posible), es que se decidió reducir el campo de estudio sólo a las páginas relacionadas con el MGO. Las estadísticas comparativas de este grupo de datos versus el total de ellos en el sitio del DII se observan en la Tabla 9. Allí se indica el contraste entre tres grupos de datos. El primero, en cuanto a la cantidad total de registros, páginas y tiempo promedio en cada una. El segundo grupo, compara las sesiones, la cantidad total de ellas, aquellas de largo mayor que uno (Sesiones s/1), el porcentaje de éstas últimas en relación al total de sesiones y finalmente, el largo promedio de las sesiones de largo mayor que 1, que en definitiva son las sesiones a considerar para el estudio. Por último, el tercer grupo compara las transiciones totales y la cantidad total de hyperlinks. La columna porcentaje MGO/DII es el porcentaje de los valores asociados al MGO en relación al sitio total del DII.

[80] **Fuente:** http://dev.mysql.com/doc/refman/5.0/es/date-and-time-functions.html
[81] Entiéndase en este contexto a cada registro como una fila de los web log, donde se identifica cada página visitada por una sesión.
[82] Entiéndase por transición, a todo paso desde una página a otra.

Tabla 9: Estadísticas simples registros MGO v/s DII

		MGO	DII	Porcentaje MGO/DII
1	Registros	7.010	128.937	5%
	Páginas	25	615	4%
	Tiempo promedio (seg.)	59,599	56,226	
2	Sesiones	2.573	55.692	5%
	Sesiones s/1	1.414	23.492	6%
	Porcentaje	55%	42%	
	Largo prom.	5,96	4,19	
3	Transiciones	4.284	72.242	6%
	Links	317	7.059	4%

Fuente: Elaboración propia basada en acceso a Base de Datos

Aplicando la Ecuación 1[83] para obtener el diámetro del grafo completo del sitio del DII tal como se indica en la expresión (4.1), se tiene que el diámetro resulta ser aproximadamente 7, es decir, dos documentos escogidos aleatoriamente del conjunto de páginas totales, tienen en promedio 6 clicks para acceder entre ellos.

$$d = 0{,}35 + 2{,}06 \cdot \log(615) = 6{,}095 \quad \textbf{(4.1)}$$

En base al mismo procedimiento pero asociado exclusivamente a las páginas del MGO, se obtiene un promedio de 3 clicks para acceder entre dos páginas del magíster, tal como resulta en la expresión (4.2).

$$d = 0{,}35 + 2{,}06 \cdot \log(25) = 3{,}229 \quad \textbf{(4.2)}$$

Finalmente, en la Tabla 10, se indican las 25 páginas del MGO, las cuales serán estudiadas en cada modelo, de acuerdo a su importancia (cantidad de hyperlinks de salida y entrada) en la estructura final. En la columna tipo donde dice menú, se indica a aquellas páginas que se encuentran en el menú principal actualmente en la página del MGO, es decir, son visibles desde cualquier página a la que se ingrese.

[83] **Ver sección** 2.1.3. Organización y estructura de un sitio web

Tabla 10: Páginas del MGO

id_page	Página	Tipo
1334	/mgo2007	Menú
1335	/mgo2007/plan_de_estudios	Menú
1344	/mgo2007/sobre_el_magiiister	Menú
1345	/mgo2007/futuro_estudiante	Menú
1346	/mgo2007/costo_y_becas	Menú
1349	/mgo2007/plan_de_estudios/cursos_electivos	
1350	/mgo2007/contenido/plan_de_estudios/cursos_electivos/IN78K_Otoño_2007.pdf	
1370	/mgo2007/egresados/graduados_2006	
1373	/mgo2007/egresados	Menú
1374	/mgo2007/egresados/graduados_2009	
1375	/mgo2007/egresados/graduados_2008	
1376	/mgo2007/egresados/graduados_2007	
1389	/mgo2007/calendario	Menú
1413	/mgo2007/profesores	Menú
1423	/mgo2007/admisiooon	Menú
1459	/mgo2007/egresados/graduados_2005	
1521	/mgo2007/servicios_generales	Menú
1523	/mgo2007/links_de_intereees	Menú
1608	/mgo2007/egresados/graduados_2001	
1609	/mgo2007/egresados/graduados_2002	
1610	/mgo2007/egresados/graduados_2003	
1611	/mgo2007/egresados/graduados_2004	
1613	/mgo2007/contacto	Menú
1700	/mgo2007/contacto/enviado	
2442	/mgo2007/contenido/plan_de_estudios/IN71K.pdf	

Fuente: Elaboración Propia

4.2.- Almacenamiento de los datos

El repositorio de datos se encuentra en el servidor wi.dii.uchile.cl, donde se accede vía PhpMyAdmin[84], herramienta que permite administrar MySQL[85] a través de páginas web vía Internet, creando, editando y eliminando bases de datos, tablas y/o campos mediante sentencias SQL.

La Tabla 11 indica los componentes de los registros web. En ella se indica el 'host', el cual corresponde a una máquina o servidor conectada a una red de ordenadores, el cual puede albergar múltiples sitios web usando diferentes dominios o números de puerto [51]. En este caso, corresponde para todos los casos al dominio www.dii.uchile.cl. Por otro lado, 'query' corresponde a indicadores que prosiguen en la ruta de la URL luego del símbolo '?', son parámetros de formulario para una misma página que posee distinta presentación debido a cambios de los mismos. Finalmente se destaca 'event' con valores solo del tipo IN o OUT; el primero indica que la solicitud realizada fue de acceso a la página y OUT de salida a la misma. Si bien lo anterior da facilidad en al cálculo de los tiempos por páginas, se debe enfrentar la problemática de que no siempre existe un IN y OUT por cada acceso, sino que solo se registra una u otra en algunos casos. Dicho análisis se realiza en la sección 4.2.1.- Sesionalización, indicando la convención tomada al respecto.

[84] Página de acceso a los datos: http://wi.dii.uchile.cl/~proman/dii/
[85] MySQL: Sistema de gestión de bases de datos. http://www.mysql.com/

Tabla 11: Registros de la Web

Dato	Detalle
id_log_session	Número correlativo que indica el orden en que se recibió la solicitud de página
id_session	Identificador de un usuario único, misma sesión
time	Tiempo Unix de acceso a la página
ip	IP del usuario
host	Host al que se está accediendo
uri	Ruta de la página a la cual se está accediendo
event	Evento Capturado (IN entrada, OUT salida)
query	Parámetros de formulario para una misma página pero distinta presentación
agent	Navegador desde el cual se accede a la página

Fuente: Elaboración propia basada en acceso a Base de Datos

En cuanto a los registros propios de la Tabla 11 asociados al MGO, se desprende que los navegadores más usados corresponden a Internet Explorer y Mozilla Firefox, gracias a la detección del agente y el estudio de los mismos, tal como se indica en la Tabla 12.

Tabla 12: Principales navegadores utilizados

Nº	%	User Agent	Browser
608	1,92%	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)	Internet Explorer
504	1,59%	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727)	Internet Explorer
497	1,57%	Mozilla/5.0 (Windows; U; Windows NT 5.1; es-ES; rv:1.9.0.11) Gecko/2009060215 Firefox/3.0.11	Mozilla Firefox
433	1,37%	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727; .NET CLR	Internet Explorer
396	1,25%	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322; .NET CLR 2.0.50727)	Internet Explorer
347	1,10%	Mozilla/5.0 (Windows; U; Windows NT 5.1; es-ES; rv:1.9.0.14) Gecko/2009082707 Firefox/3.0.14 (.NET C	Mozilla Firefox
102	0,32%	Opera/9.64 (Windows NT 5.1; U; es-LA) Presto/2.1.1	Opera

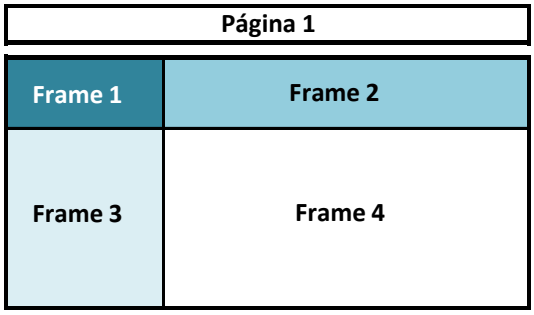
Fuente: Elaboración propia basada en acceso a Base de Datos

4.2.1.- Sesionalización

La sesionalización se realizó vía cookies autorizadas del tipo persistente y P3P[86], para poder tener un ajuste mayor a la construcción de rutas de visita de los usuarios [70]. Estudiando los registros de la Tabla 11, se identificaron dos problemáticas de interés. La primera tiene relación a la carencia de un 'event' IN y luego OUT por cada página, por lo que el cálculo de tiempo dependería de los registros previos y/o posteriores. Por otro lado, se constató el registro de páginas del tipo frame que, en términos de desarrollo de una página Web, corresponden a una forma de estructurarla en distintos espacios independientes los unos de los otros, de modo que en cada espacio se coloca una página distinta que posee su propio lenguaje HTML. Los frames son marcos, una manera de dividir la página en distintos espacios independientes los unos de los otros, de modo que en cada espacio se coloca una página distinta [87]. Una esquematización de la organización de frames formando una página, se presenta en la Figura 35.

[86] **Ver sección**: 2.3.3.- Proceso de sesionalización
[87] **Fuente:** http://www.desarrolloweb.com/articulos/791.php

Figura 35: Organización de los frames



Fuente: http://www.desarrolloweb.com/articulos/791.php

Así, se observan registros de visita por parte de algún usuario a solo una parte de alguna página en particular, como por ejemplo a la barra menú de ella. Estas secciones de páginas son archivos html los cuales son llamados[88] en la página completa para su visualización. Sin un estudio detallado fácilmente podrían confundirse y mezclar como páginas completas. Un ejemplo de ese tipo de registros se observa en la Tabla 13 donde se pierde la información de la página completa que fue visitada, pues corresponde a solo la parte superior de ella.

Tabla 13: Ejemplos de frames detectados

Time	TNETO	ip	host	uri	event
1243032186	**1**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/ts_files/scroll.html	IN
1243032494	**26**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/superior2.html	OUT

Fuente: Elaboración propia basada en acceso a Base de Datos

Para dar solución a la carencia de registros IN y OUT se elaboró una pauta de procedimientos considerando los casos posibles que se describe a continuación:

1.- Caso en que hay registros de IN y OUT por página

El tiempo en cada página (TNETO) se calculará restando el tiempo en que hace OUT a dicha página con el tiempo en que hace IN, ambos indicados en la columna time, tal como se indica en la Tabla 14.

Tabla 14: Caso registro IN y OUT por página

Time	TNETO	ip	host	uri	event
1243032453		200.27.145.92	www.dii.uchile.cl	/~mba/paginas/profesores.html	IN
1243032456	**3**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/profesores.html	OUT

Fuente: Elaboración propia basada en acceso a Base de Datos

2.- Caso en que sólo hay registros de IN o solo OUT por página

Se asume cambios instantáneos, es decir, el tiempo en cada página se calcula como la diferencia entre el time OUT de la página *i*, y el time del evento anterior. Un ejemplo de ello se observa en la Tabla 15.

[88] Desde el punto de vista del código html, una llamada se refiere a indicar el archivo que desea desplegarse, es decir, si una página A es el menú de una página B, entonces B llama a A para que B se visualice con su menú.

Tabla 15: Caso de registro solo IN o OUT por página

Time	TNETO	ip	host	uri	event
1243032222		200.27.145.92	www.dii.uchile.cl	/~mba/paginas/postulacion1.html	IN
1243032246	**24**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/postulacion1.html	OUT
1243032281	**35**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/convenio.html	OUT
1243032301	**20**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/costos1.html	OUT

Fuente: Elaboración propia basada en acceso a Base de Datos

3.- Caso entre registros tipo 1 y 2

Hay un tiempo de "carga" de la página por así decirlo, o una demora que indicaría estar navegando en otro sitio adicionalmente o seguir otras opciones. El cálculo del tiempo es la diferencia entre el evento que comienza y el tiempo en que termina. Es el caso de los 56 segundos que se calcularon en el ejemplo de la Tabla 16, donde se inserta una fila en los registros para calcular los tiempos asociados:

Tabla 16: Caso de registros entre registros IN-OUT

Time	TNETO	ip	host	uri	event
1243032397	**96**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/fuentesfinanciamiento.html	OUT
	56				
1243032453		200.27.145.92	www.dii.uchile.cl	/~mba/paginas/profesores.html	IN
1243032456	**3**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/profesores.html	OUT

Fuente: Elaboración propia basada en acceso a Base de Datos

En la Tabla 17 de identifica una limpieza final. Registros en los que se ha sobrescrito una línea se eliminan, ya sea porque corresponde a tiempo en que se carga la página (casos 1 y 2, Tabla 17) o porque la página tenía sus dos registros IN/OUT, para que no haya redundancia de datos (Caso 3, Tabla 17). Además, como solución a la existencia de frames, también son eliminados dado que no representan una cantidad significativa en el total de registros (caso 4, Tabla 17) y que no corresponden a páginas completas (las páginas completas no se pierden).

Tabla 17: Eliminación de registros adicionales

Time	TNETO	ip	host	uri	event	
1243032397	**96**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/fuentesfinanciamiento.html	OUT	
	56					1
1243032453		200.27.145.92	www.dii.uchile.cl	/~mba/paginas/profesores.html	IN	
1243032456	**3**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/profesores.html	OUT	
	2					2
1243032458		200.27.145.92	www.dii.uchile.cl	/~mba/paginas/perfil.html	IN	3
1243032459	**1**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/perfil.html	OUT	
1243032468	**9**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/porque.html	OUT	
1243032494	**26**	200.27.145.92	www.dii.uchile.cl	/~mba/paginas/superior2.html	OUT	4

Fuente: Elaboración propia basada en acceso a Base de Datos

4.- Situaciones complejas y generalización

El procesamiento de las sesiones se realiza mediante un programa en Java, el cual se agrupa por el identificador común de la sesión y luego calcula los tiempos de las páginas involucradas en cada una de ellas. Para esto, se procederá según las precisiones de la Tabla 18 según el estado de los registros IN o OUT en la columna 'event'.

Tabla 18: Combinaciones posibles para cálculos de tiempos

Combinaciones posibles para cálculo de tiempo

	Página 1		Tiempo descarga	Página 2		Tiempo descarga	Página 3		Tiempo Pág.2	
	(E) IN	(F) OUT		(A) IN	(B) OUT		(C) IN	(D) OUT		
1	1	1		1	1		1	1	B-A	
2	1	1		1	0		1	1	C-A	
3	0	1		1	1		0	1	B-A	
4	1	0		1	1		1	0	B-A	
5	1	1		0	1		1	1	B-F	*
6	0	0		1	1		0	0	B-A	
7	0	1		1	0		0	1	(D-A)/2	
8	1	1		0	0		1	1	x	
9	1	0		1	0		1	0	C-A	
10	0	1		0	1		0	1	B-F	*
11	1	0		0	1		1	0	(B-E)/2	*
12	0	0		1	0		0	0	Ind.	
13	0	1		0	0		0	1	x	
14	0	0		0	1		0	0	Ind.	
15	1	0		0	0		1	0	x	
16	0	0		0	0		0	0	x	

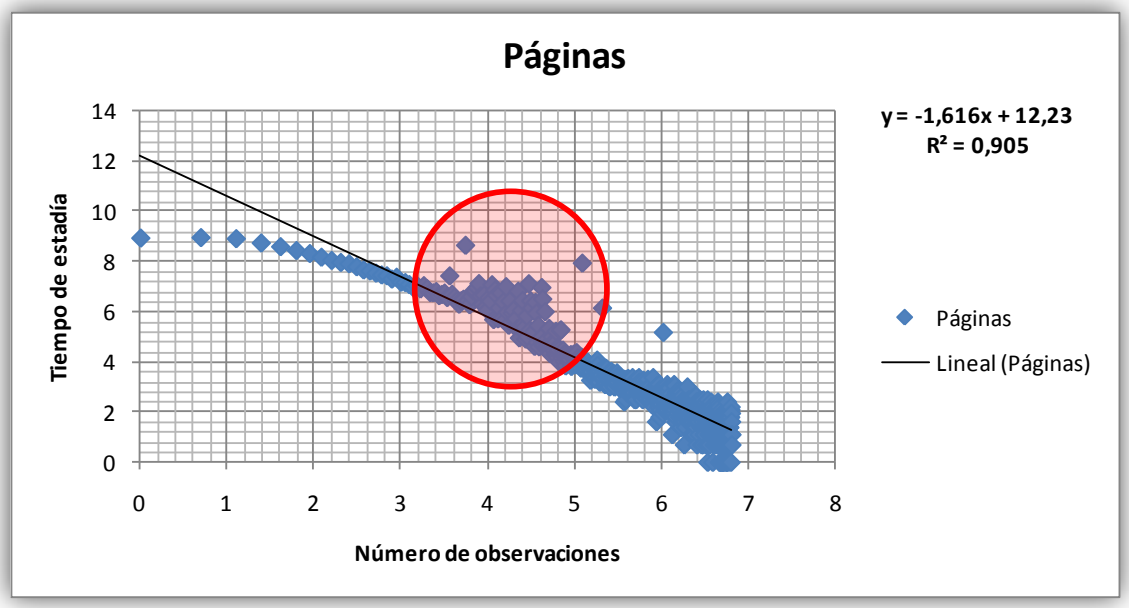
Ind.	Indeterminado por ser página final
x	Nulo
*	Indeterminado si llega a ser página de inicio

Fuente: Elaboración personal

La Tabla 18 indica cómo se lleva a cabo el cálculo dependiendo de la situación presentada, considerando los registros de las páginas previas o posteriores, y por ende, de la situación en cada uno de las 2^4 situaciones posibles.

En la Figura 36 se observa la relación del tiempo de estadía y el número de observaciones para cada tiempo (se cuenta la cantidad de páginas que fueron visitadas una misma cantidad de tiempo), ambos datos en escala logarítmica. La relación lineal anterior logra un ajuste del 90% y viene dada por estudios previos en regularidades en la Web, que van desde la estructura y el crecimiento de la misma a los patrones de acceso en la navegación [28][42]. En esos trabajos se ha demostrado un enfoque basado en los usuarios para obtener modelos que caracterizan las regularidades empíricas del uso de la Web. En particular, se ha formulado la aplicación de la ley de navegación, obteniendo el interesante comportamiento lineal aplicando logaritmos a las variables en estudio. Dicha sentencia estuvo basada en agentes del modelo y validada con conjuntos de datos empíricos de registro web.

Figura 36: Gráfica Tiempo de estadía en página v/s número de observaciones



Fuente: Elaboración propia basada en acceso a Base de Datos

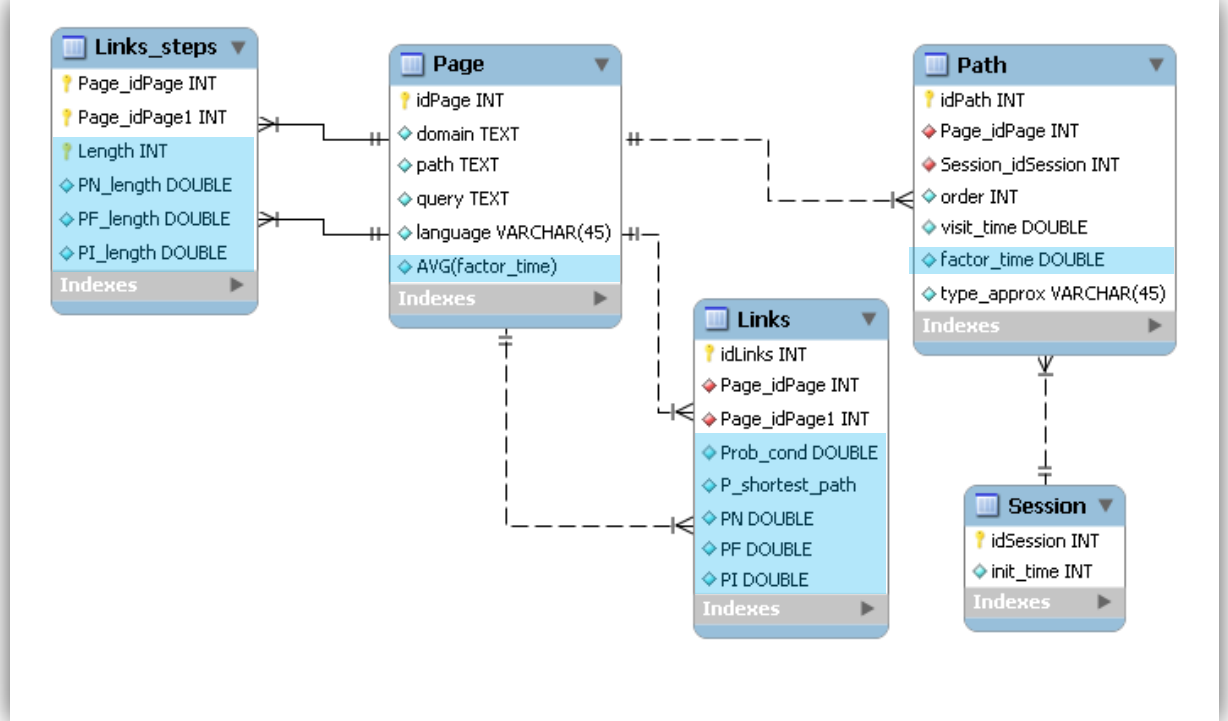
4.2.2.-Modelo de datos

Una vez realizada la sesionalización, se procede a realizar los cálculos que el algoritmo requiere, éstos corresponden a tiempo promedio (AVG(time)), probabilidad condicional (Prob_cond), frecuencia de llegada (PN), factor de estadía (PT), pesos asociados a cada modelo como P_shortest_path, PI y PI_length para modelo 1, 2 y 3 respectivamente, entre otras.

En la Figura 37 se muestra el modelo de datos del repositorio de los mismos en PhpMyAdmin, dispuestos en la tablas correspondientes manteniendo la normalización[89] global, con el fin de evitar la redundancia, problemas de actualización, grandes estructuras de datos, entre otras, favoreciendo la integridad de toda la información en dimensiones más pequeñas y precisas, facilitando el acceso y el compresión de la información.

[89] **Fuente:** http://www.mysql-hispano.org/page.php?id=16&pag=2

Figura 37: Modelo de datos



Fuente: Elaboración propia

La Figura 36 describe el modelo de datos, es decir, la organización en que se éstos se almacenan, el tipo de datos y las restricciones de integridad[90] entre ellos. En cuanto a las relaciones, una sesión puede visitar muchas páginas, y una página puede ser visitada por muchas sesiones, es decir, entre session y page hay una relación *n* a *m*. Así, la tabla que refleja e intermedia dicha relación normalizada *n-m*, es Path, conteniendo como llaves foráneas las llaves primarias de cada tabla asociada (idSession e idPage). Por otro lado, se observan además dos relaciones reflexivas, como lo es en el caso de la tabla links con page y links_steps con pages de igual modo. En el caso de hyperlinks, contiene su propia llave primaria llamada idLinks, y como llaves foráneas los identificadores de las páginas de inicio y fin (Page_idPage y Page_idPage1) respectivamente. En el caso de hyperlinks_steps, su llave principal pasa a estar dada por la combinación de tres atributos: Page_idPage, Page_idPage1 y Length, pues puede ver un acceso desde la misma página de inicio a la final, pero con distinto largo.

En cuanto a los atributos, la **tabla session** contiene el identificador y el tiempo de inicio de cada una. Dicho identificador también se encuentra en la **tabla path** como llave foránea, la cual contiene el registro de una página visitada en la sesión (indicada como Page_idPage), el tiempo en ella (visit_time), el factor tiempo correspondiente (% de tiempo en relación a la sesión total, como se indicó en la Ecuación 5) y el tipo de aproximación (type_approx) realizada para el cálculo del tiempo[91].

[90] **Restricciones de integridad:** Requisitos de referencia entre registros
Fuente: http://www.mysql-hispano.org/page.php?id=27
[91] Dicha aproximación se refiere a cual caso de la Tabla 15 fue enfrentado para el cálculo del tiempo.

El detalle de cada página se encuentra en la **tabla page**, indicando su dominio, ruta, query, lenguaje y el cálculo correspondiente de diversos atributos. El cálculo de posición, tiempo y factor promedio se realiza obteniendo el valor medio de su visita en todas las sesiones, al realizar una consulta que agrupa por página los datos provenientes de la Tabla Path: order, visit_time y factor_time respectivamente.

Por otro lado, cada página tiene hyperlinks de salida (link-in) y de llegada (link-out), lo que se indica en la **tabla Links,** indicando como par ordenado las páginas relacionadas. Los cálculos asociados a cada hyperlink, son primero los correspondientes al modelo 1: su probabilidad total (prob_total) calculada como el cociente entre la cantidad de veces que se usa el hyperlinks y el total interacciones realizadas; su probabilidad condicional, que también es un cociente entre la cantidad de veces que se usa el hyperlink, pero ahora dividido por la cantidad de veces en que la página de inicio sale a otra página cualquiera, y por otro lado, la probabilidad asociada al modelo 1 (p_shortest_path) dada por la Ecuación 8. Los cálculos relevantes para el modelo 2 son PN (frecuencia de llegada, Definición 4), PF (Frecuencia de estadía, Definición 5) y PI (Peso potencial Modelo 2, Ecuación 11). El modelo 3 se encuentra representado por la **tabla Links_step**, dado que cambia la llave primeria, ahora dependiente del largo de la transición entre dos páginas. Los atributos PN_length, PF_length, y PI_length son calculados realizando las consultas adecuadas según Definición 6, Definición 7 y Ecuación 13 respectivamente.

La Figura 36 posee ciertos valores ensombrecidos, los cuales fueron calculados vía consultas en sql pues no se obtienen directamente de los registros web. Ellos son los datos claves para el estudio realizado.[92] El valor de p_shortest_path no es calculado por consultas, sino que corresponde a la aplicación de las Ecuaciones 6 o 9 del modelo 1, donde los valores finales se calcularon en Matlab y se almacenaron posteriormente en la base de datos.

4.3.- Estudios estadísticos

Ya dando paso al estudio de las sesiones, se hace necesario develar el contexto en que la página del MGO se encuentra inserta. Se destaca el porcentaje de uso de las páginas de inicio de todo el DII, es decir, cuáles son aquellas donde en mayor cantidad comienzan los usuarios su sesión en el sitio. La Tabla 19 permite vislumbrar que un 34,4% de las sesiones comienzan en el home (página principal del sitio), un 3,55% visitan en primer lugar la página inicial del Magíster de Gestión y Políticas Públicas, con un 3,31% la página de los diplomas ofrecidos y en séptimo lugar, se observa la página a estudiar (MGO), con un 1,87% de los ingresos como página de inicio.

[92] **Ver apéndice** E.1. Consultas sql.

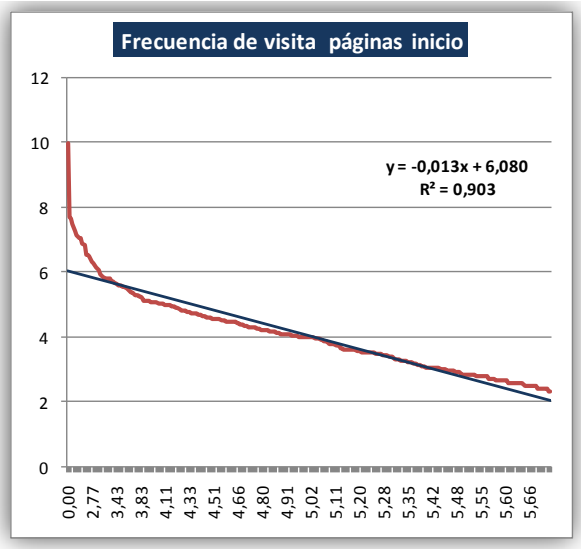
Tabla 19: Páginas de inicio de sesión.

Nº	id_page	Cantidad	Porcentaje	Página
1	1230	21886	34,40%	www.dii.uchile.cl
2	1264	2260	3,55%	www.dii.uchile.cl/webmgpp
3	1252	2104	3,31%	www.dii.uchile.cl/diplomas
4	1364	1772	2,78%	www.dii.uchile.cl/diplomas/pages/menu_inteligencia.htm
5	1248	1451	2,28%	www.dii.uchile.cl/diplomas/pages/menu_proyectos.htm
6	1242	1278	2,01%	www.dii.uchile.cl/diplomas/pages/menu_estrategia.htm
7	1334	1193	1,87%	www.dii.uchile.cl/mgo2007
8	1247	1183	1,86%	www.dii.uchile.cl/ingenieria_civil_industrial
9	1238	1138	1,79%	www.dii.uchile.cl/mba/paginas/home2.htm
10	1271	968	1,52%	www.dii.uchile.cl/magister_doctorados
11	1256	944	1,48%	www.dii.uchile.cl/cea

Fuente: Elaboración propia a partir de consulta[93] en la Base de datos

Se desprende que la visita comienza asociada a temas de postgrado, representando un 51,7% de los accesos a sitios de este tipo sólo indicados en la Tabla 19. El comportamiento establecido se relaciona directamente con la ley de Zipf enunciada en la sección 2.1.2.- Principios de la web, en el sentido que el mayor uso de un conjunto (páginas) recae en pocos elementos, habiendo una amplia gama de ellos. Claramente se aprecia que la página Home tiene una alta frecuencia como página de inicio, lo que es esperado por construcción del sitio y por el principio de Zipf del comportamiento en la Web. En la Figura 38 se grafica el esperado comportamiento lineal usando escala logarítmica en las columnas cantidad y Nº de la Tabla 16, donde se aprecia un considerable ajuste con R^2 de un 90,3%.

Figura 38: Aplicación de la ley de Zipf

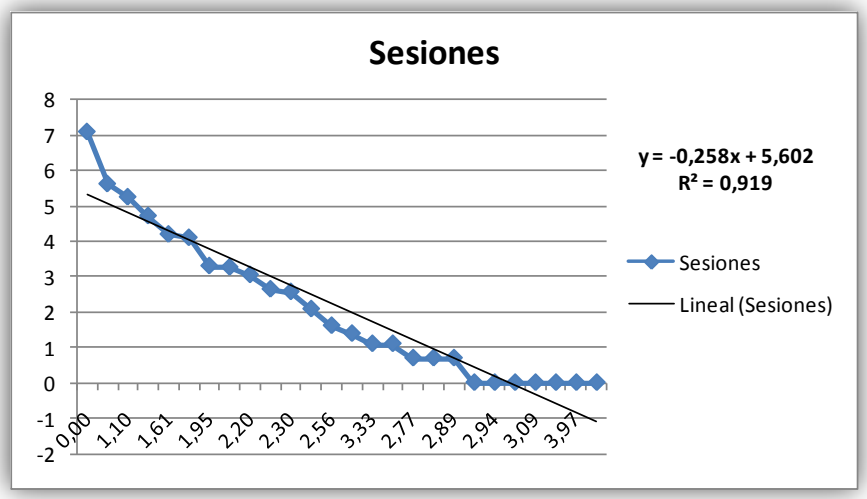


Fuente: Elaboración propia a partir de los datos Tabla 6

Analizando ahora las sesiones que sólo visiten páginas del MGO, se obtiene la gráfica del largo de sesiones versus la cantidad de veces que ello ocurre, dispuesta en la Figura 39, con un ajuste de 91,9%.

[93] Entiéndase consulta en la base de datos a aquella que se realizan vía SQL (Structured Query Language), lenguaje de consulta estructurado que permite desprender datos relevantes desde la base de datos según se desee.

Figura 39: Largo de sesiones versus la cantidad de veces que ocurre dicho largo



Fuente: Elaboración propia basada en acceso a Base de Datos

En cuanto a las páginas específicamente del magíster, resulta interesante analizar aquellas más visitadas. Como se establece en la Tabla 20, aquella más visitada según lo esperado es la de inicio del sitio, con un 28% de los requerimientos. Prosiguen peticiones de acceso al plan de estudio, costos y becas, sobre el magíster, proceso de admisión, egresados y cursos electivos, un reflejo lógico de cómo se estructura la información y como los usuarios buscan información profundizándola en contenidos.

Tabla 20: Páginas más visitadas del MGO

id_page	Ruta	Cantidad	%
1334	/mgo2007	2011	28%
1335	/mgo2007/plan_de_estudios	1151	16%
1346	/mgo2007/costo_y_becas	567	8%
1344	/mgo2007/sobre_el_magiiister	391	5%
1423	/mgo2007/admisiooon	373	5%
1413	/mgo2007/profesores	346	5%
1373	/mgo2007/egresados	343	5%
1349	/mgo2007/plan_de_estudios/cursos_electivos	289	4%

Fuente: Elaboración propia basada en acceso a Base de Datos

A nivel estructural, resulta de interés evaluar las páginas hubs y autoritativas presentes[94]. Realizando una consulta en la tabla hyperlinks, de modo de contabilizar la cantidad de veces que una página va hacia otra (tipo INI) y por otro lado la cantidad de veces en que llegan a ella (tipo FIN), ambas vía acceso directo. La Tabla 21 muestra las páginas con mayor cantidad de hyperlinks de salida siendo plan de estudios y egresados aquellas con 22.

[94] **Ver sección**: 2.1.3.- Organización y estructura de un sitio web

Tabla 21: Páginas con mayor hyperlinks de salida

id_page	Ruta	Nivel Hub
1335	/mgo2007/plan_de_estudios	22
1373	/mgo2007/egresados	22
1346	/mgo2007/costo_y_becas	20
1345	/mgo2007/futuro_estudiante	19
1413	/mgo2007/profesores	19
1344	/mgo2007/sobre_el_magiiister	18
1334	/mgo2007	17
1521	/mgo2007/servicios_generales	17

Fuente: Elaboración propia basada en acceso a Base de Datos

Agrupando ahora para detectar la mayor cantidad de hyperlinks de entrada y por ende el nivel autoritativo de las páginas, se obtiene la Tabla 22, donde la página de egresados nuevamente presenta alto nivel en cantidad de hyperlinks llegando a ella, como también la página de inicio del MGO.

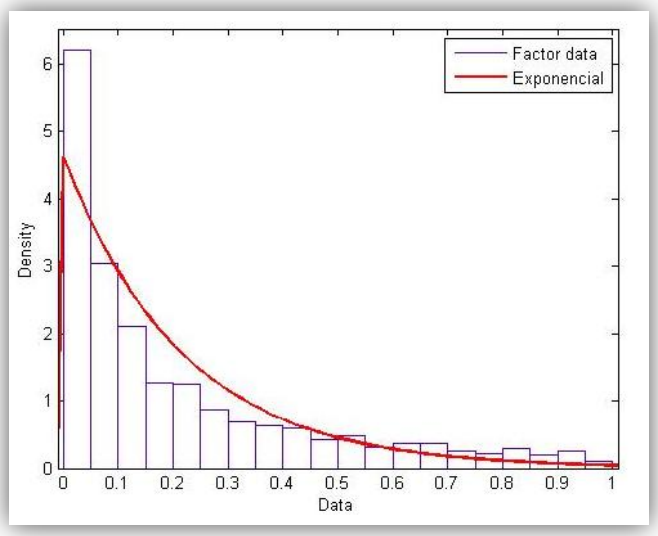
Tabla 22: Páginas con mayor hyperlinks de entrada

id_page	Ruta	Nivel Autoritativo
1373	/mgo2007/egresados	21
1334	/mgo2007	19
1335	/mgo2007/plan_de_estudios	16
1345	/mgo2007/futuro_estudiante	16
1349	/mgo2007/plan_de_estudios/cursos_electivos	16
1374	/mgo2007/egresados/graduados_2009	16
1375	/mgo2007/egresados/graduados_2008	15
1423	/mgo2007/admisiooon	15

Fuente: Elaboración propia basada en acceso a Base de Datos

Continuando en la línea de las páginas en la Figura 40 se observa la distribución de probabilidad del factor tiempo en la página de inicio del MGO, ajustada a una distribución exponencial.

Figura 40: Distribución factor tiempo de www.dii.uchile.cl/ ~mgo



Statistical values

Log likelihood: 813.671
Domain: 0 <= y < Inf
Mean: 0.21554
Variance: 0.0464577

Parameter Estimate Std. Err.
0.00552486

Estimated covariance of parameter estimates:
mu 3.05241e-005

Fuente: Elaboración propia basada en acceso a Base de Datos

Por otro lado, las probabilidades estacionarias de estar en una página *i* (p_i) [55], será calculada contando la cantidad de veces en que se está en la página *i* en relación a estar en cualquier otra. Se hablará de estado estacionario por la gran cantidad de datos recolectados, lo que permite establecer que los mismos ya se encuentran en un régimen permanente[95]. Dichas probabilidades cumplen con la Ecuación 17.

Ecuación 17: Probabilidades estacionarias de estar en una página

$$\sum_{i=1}^{n} p_i = 1$$

Finalmente en la Figura 41 se obtienen las probabilidades de estado, indicando las siete páginas con mayor probabilidad de estar en ellas.

Figura 41: Probabilidades de estado

	Path	Prob.
1	/mgo2007	28,7%
2	/mgo2007/plan_de_estudios	16,4%
3	/mgo2007/costo_y_becas	8,1%
4	/mgo2007/sobre_el_magíster	5,6%
5	/mgo2007/admisiooon	5,3%
6	/mgo2007/profesores	4,9%
7	/mgo2007/egresados	4,9%

Fuente: Elaboración propia basada en acceso a Base de Datos

4.4.- Aplicación del algoritmo

Antes de proceder con el algoritmo, se requiere rescatar la matriz de adyacencia existe. Para esto, se realiza un programa en Java que se comunique con la base de datos en PhpMyadmin, exportando previamente la tabla de transiciones únicas (Tabla hyperlink, Figura 36) condicionada a que ambas páginas visitadas sean subyacentes al MGO[96]. Luego, se crea una base de datos en el servidor local[97] para acceder a ella más directamente considerando el siguiente seudocódigo:

1.- Conectar con la base de datos
2.- Realizar la consulta para acceder a la tabla de hyperlinks de páginas del MGO
3.- Obtener el identificador de la página de inicio y fin en cada registro
4.- Generar código para poblar matriz de adyacencia final:

"update mgo_adyacencia set `"+id_pagefin+"`=1 where id="+id_pageini+";"

[95] **Ver sección** C.- Conceptos utilizados de teoría de grafos y cadenas de Markov
[96] Entiéndase páginas subyacentes al MGO a aquellas que en la ruta de la URL comienzan con /mgo, es decir: www.dii.uchile.cl/mgo/...
[97] El servidor local (localhost), corresponde a la dirección donde se encuentra el servidor de datos propios, es decir, nombre reservado para acceder a sí mismo.

De este modo se busca identificar los pares de hyperlinks relacionados, cuyo valor en la tabla de adyacencia será 1, resultando una matriz como la indicada en la Figura 42.

Figura 42: Trozo de matriz de adyacencia original obtenida

	ADYACENCIA ORIGINAL							
	1	2	3	4	5	6	7	8
1	1	1	1	1	1	0	0	0
2	1	1	1	1	1	1	0	0
3	1	1	1	1	1	0	0	0
4	1	1	1	1	1	0	0	1
5	1	1	1	1	1	0	0	0
6	1	1	1	1	1	1	1	0
7	1	1	1	1	1	1	1	0
8	1	1	1	1	1	0	0	0

Fuente: Elaboración propia

Una vez obtenida la matriz de adyacencia inicial, se linealiza la matriz concatenando las filas en un solo gran vector. Esto se realiza pues Matlab trabaja un con vector lineal como individuo genético. Allí, una vez programado el algoritmo, se realizan una serie de corridas de prueba a modo de análisis de sensibilidad de los parámetros a utilizar. El objetivo es determinar la combinación de parámetros que otorga un mejor resultado dado por valores numéricos en la suma final de los pesos existentes. Así, se evalúan los mejores parámetros de acuerdo a la forma de proceder de cada uno de ellos y el comportamiento esperado en las generaciones para la convergencia. La Tabla 23 contiene los parámetros establecidos.

Tabla 23: Parámetros AG en Matlab

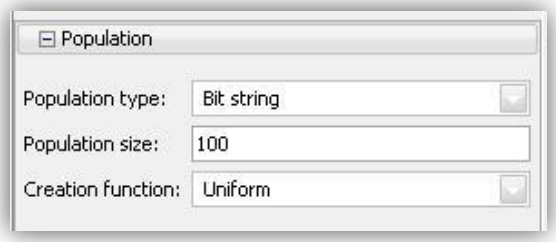
Parámetro / Etapa	Método / Valor
Fitness Scaling	Rank
Selection	Roulette
Crossover fraction	0,8
Crossover	Scattering
Mutation	Gaussian

Fuente: Elaboración propia, recatando los valores establecidas en la interfaz de Matlab

La escala del fitness (Fitness Scaling) se realiza mediante el método rank, el cual clasifica los puntajes en bruto de cada individuo basado en un ranking global de ellos. Una selección vía ruleta, simula una rueda con el área de cada segmento proporcional al fitness. Dicho método fue comparado con el torneo y selección estocástica, donde no se reportaron variaciones en el resultado final. La fracción de cruce por otro lado, especifica la fracción de la próxima generación que será producida vía cruce. Del 20% restante, el 10% de los mejores de la generación previa es mantenido y el otro 10% cae en el proceso de mutación. El proceso de cruce en sí es del tipo disperso, es decir, crea un vector binario el cual selecciona los genes respectivos del padre si su valor es 1, y los genes de la madre si su valor es 0, combinando los genes para formar el hijo. Finalmente, la mutación gaussiana cambia un número aleatorio a cada vector de entrada, cuyo gen afectado es tomado desde una distribución gaussiana

centrada en cero. Así, la generación de la población inicial se considera con un tamaño preliminar de 100, bajo una función de generación aleatoria como se observa en la Figura 43:

Figura 43: Población en Matlab



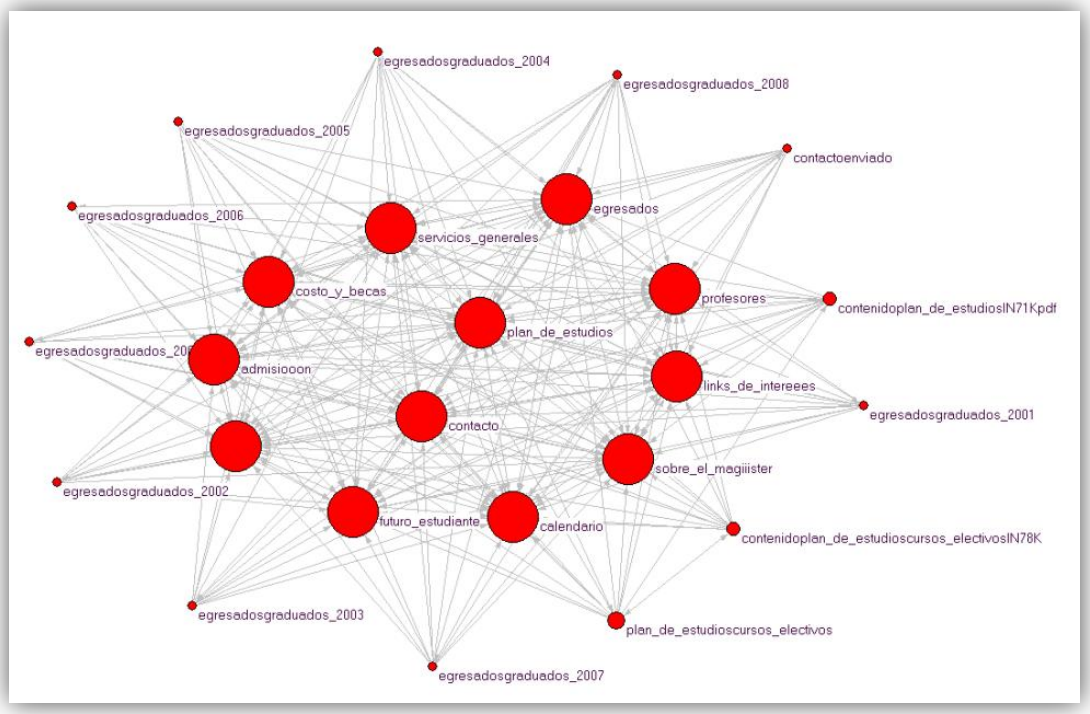
Fuente: Interfaz de Matlab

Los otros operadores claves en la ejecución del algoritmo, tienen que ver con los criterios de detención, que son aquellos que determinan su interacción al momento de encontrar un buen ajuste dado por ciertos parámetros. Ellos corresponden a un número máximo de generaciones, a un factor de tolerancia que indique el grado de diferencia entre el mejor individuo detectado entre las distintas generaciones y a un conteo de las mismas. El primero se fija en 10.000, el segundo se toma como nulo, y la cantidad de generaciones en que el mejor individuo detectado se mantenga (no hay otro individuo que tenga mejor fitness) se fija en 1000. Con éste último valor se está diciendo que si el mejor individuo se mantiene como mejor individuo por sucesivas generaciones, su calidad de óptimo es cada vez mayor.

4.4.1.-Ejecución y resultados

Configurado el fitness representativo de cada modelo y el conjunto de datos de entrada que cada uno de ellos necesita se comienza con la ejecución. A modo de visualización inicial, se desprende gracias al uso del software Pajek [19] el grafo inicial que actualmente conforma el sitio del MGO. La Figura 44 muestra las interrelaciones entre las páginas y su configuración estructural, enfatizando el grado de entrada de hyperlinks hacia las páginas (grado en que se es página Autoritativa, en base a la estructura). Cada nodo está representado por un círculo, donde el tamaño indica el grado de entrada vía hyperlinks, es decir, a mayor tamaño, más hyperlinks le llegan a dicho nodo. Puede observarse un núcleo central que agrupa aquellas páginas con mayores llegadas, ellas son la página principal, plan de estudios, sobre el magíster, futuro estudiante, costos y becas, egresados, calendario, profesores, admisión, servicios generales, hyperlinks de intereses, y contacto. Como valores periféricos se encuentran las páginas restantes, donde se destacan los planes de estudios de ramos, tanto electivos como obligatorios.

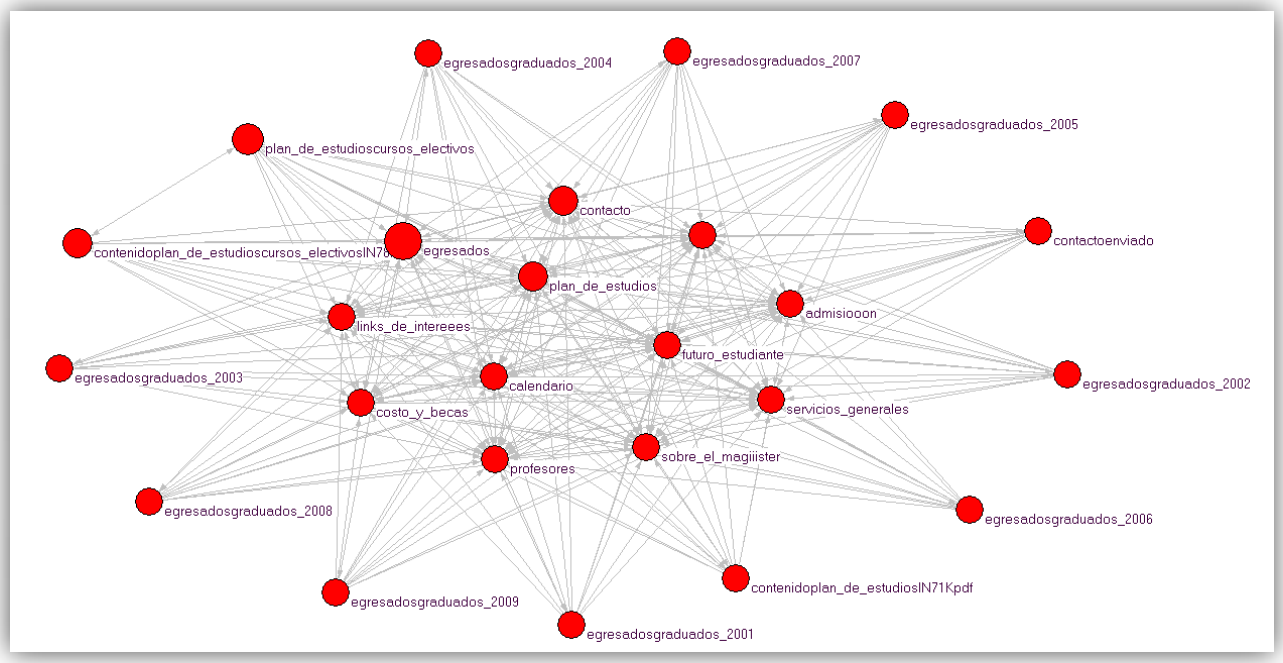
Figura 44: Grafo inicial del sitio MGO: Grado de entrada de hyperlinks



Fuente: Grafo resultante vía uso de Pajek

La Figura 45 por otro lado, muestra el grado de salida de hyperlinks desde las páginas. Es posible observar tamaños semejantes entre los nodos, lo que significa que las páginas poseen un grado de salida relativamente equivalente. De todos modos la página de egresados se destaca por sobre los demás tamaños de nodos, seguida de las páginas plan de estudios, cursos electivos y contacto.

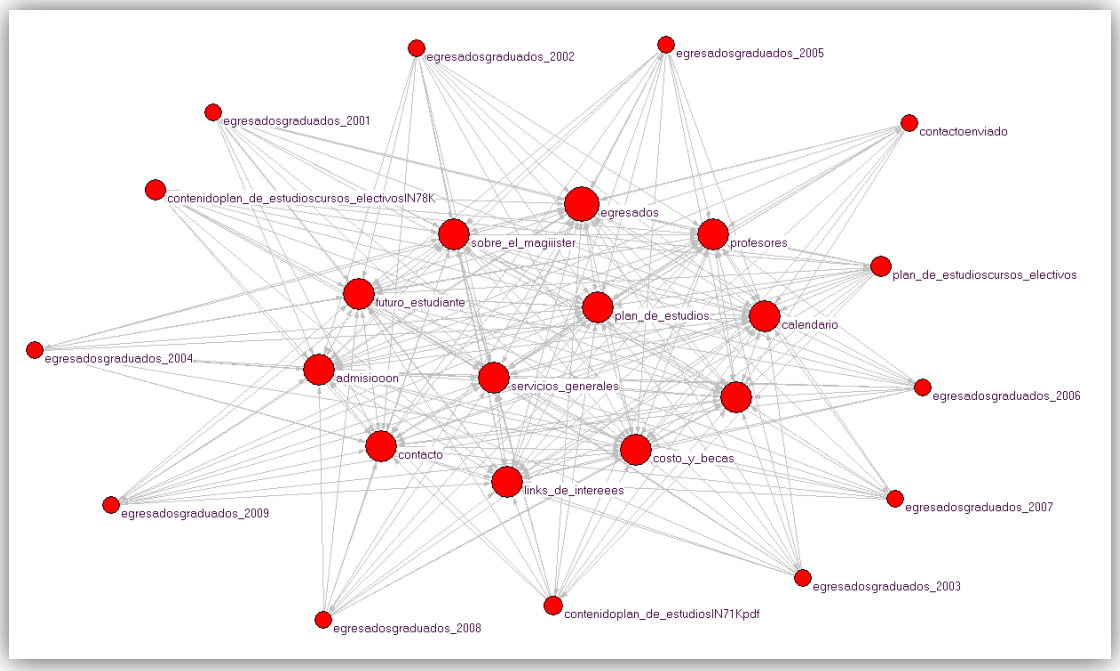
Figura 45: Grafo inicial del sitio MGO: Grado de salida de hyperlinks



Fuente: Grafo resultante vía uso de Pajek

Continuando con la descripción de la matriz inicial se tiene ahora la configuración de total de la importancia de cada nodo, por su nivel de interacción tanto a nivel de entrada de hyperlinks como de salidas. Tal como se observa en la Figura 46 se aprecia un tamaño uniforme entre todos los nodos, donde los centrales corresponden a los hyperlinks del tipo menú y los más exteriores a la red, son aquellos de mayor profundización en el sitio.

Figura 46: Grafo inicial del sitio MGO: Grado de salida de hyperlinks



Fuente: Grafo resultante vía uso de Pajek

El individuo final de cada modelo (matriz de adyacencia) y la precisión de los parámetros asociados se presentan y describen a continuación.

Modelo 1:

Antes de la ejecución se configuran los parámetros requeridos, siendo en este caso los valores de alfa y beta. Su estudio es principalmente empírico y asociado a los valores de los pesos de los hyperlinks, con la idea de hacer comparable el valor numérico del beneficio y del número de conexiones que se observan en la Ecuación 10. Ante eso, se busca que el valor promedio de los pesos π esté en el orden de 10^2, para compararse numéricamente con 317, que es la cantidad inicial de hyperlinks. Así, se obtiene un valor promedio como se observa en la expresión 4.3.

$$\frac{\sum_{ij} \pi_{ij}}{625} = 0{,}00542 \quad (4.3)$$

Con ello entonces, se ve que se requiere multiplicar por 10^5 para alcanzar el orden comparable con la cantidad total de hyperlinks. Como resultado de lo anterior, se fija el valor de *alfa* en 10^5 y de *beta* en 1, como se observa en la Tabla 24.

Tabla 24: Parámetros y resultados modelo 1

Resultados Modelo 1	
alfa	100000
beta	1
Nº de links creados	36
Nº links eliminados	66
Links totales	287
Variación de la Función de utilidad	51%
Variación en el Nº de Links	-9%

Fuente: Elaboración propia

El resultado del algoritmo, arroja un individuo binario, donde 1 es existencia de hyperlinks y 0 es la no existencia. Comparando dicho individuo con la matriz inicial, se obtiene que el nuevo individuo posee 36 hyperlinks creados (antes no existían), 66 hyperlinks eliminados y 523 hyperlinks mantenidos, con una cantidad total de 287 hyperlinks. Comparando la utilidad inicial con la utilidad final, donde R_o es la adyacencia inicial (actualmente existente en el sitio Web) y R_f es la adyacencia final obtenida por el AG, se refleja un aumento de la utilidad en un 51% y una reducción del 9% en la cantidad total de hyperlinks (Ecuación 18). La matriz de pesos π es la obtenida tal como se indica en la Ecuación 9.

Ecuación 18: Variación de la utilidad modelo 1

$$\frac{\pi \cdot R_f - \pi \cdot R_o}{\pi \cdot R_o}$$

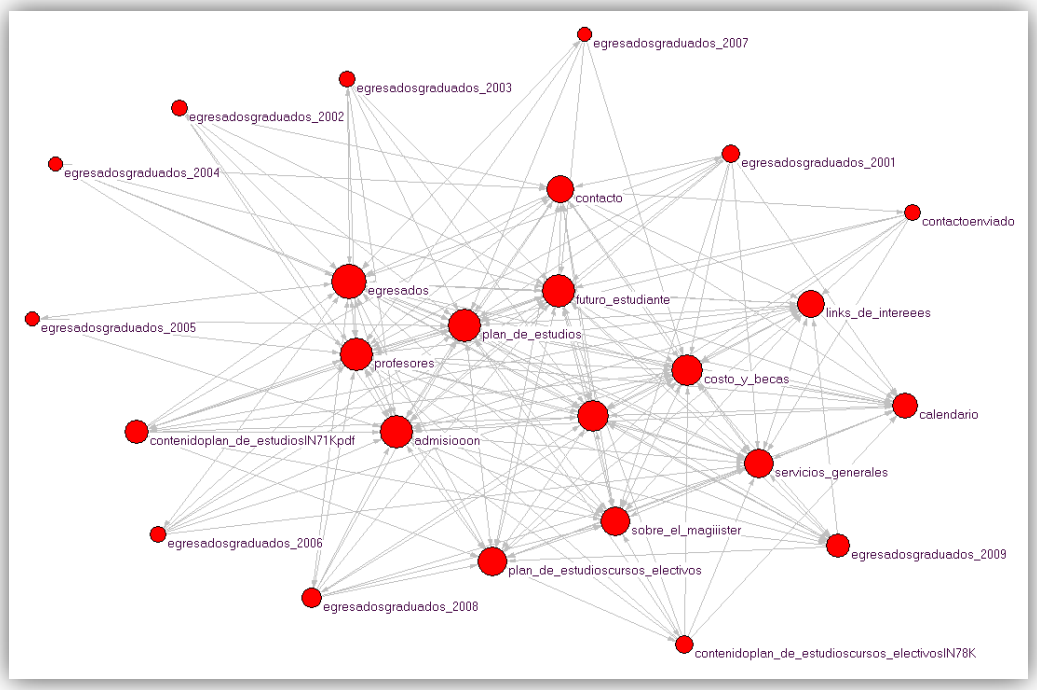
Profundizando en el mejor individuo obtenido, la Figura 47 es la gráfica proporcionada por Matlab del mejor individuo, el cual tiene una representación lineal. La figura representa la densidad de la matriz. Las líneas negras corresponden al valor 1 en el individuo, es decir, la existencia del hyperlink, mientras que las líneas blancas son el valor 0, señalando la no existencia. El eje *X* del gráfico que se ve en la figura, va desde 1 a $M=mxm$, donde m es el número total de páginas. En este caso m=25, por lo tanto, M=625. Así, cada valor desde 1 a M indica una relación entre hyperlinks, de acuerdo al proceso de linealización realizado en la sección 3.2.-Representación del Individuo, específicamente, en la expresión de la Ecuación 2 puede desprenderse la correcta relación asociada a los hyperlinks que corresponden a cada valor del eje *x*.

Figura 47: Densidad de mejor Individuo modelo 1

Fuente: Imagen obtenida desde Matlab una vez terminada la ejecución

El grafo final que representa el mejor individuo y por ende, la nueva estructura web propuesta por el modelo, puede observarse en la Figura 48. Dicho grafo, contiene nodos de mayor o menor tamaño, indicando en esta ocasión, la participación total en la red, es decir, tanto el grado de salida como entrada a los mismos. Así, egresados, plan de estudios, profesores, futuro estudiante, admisión, costos y becas y Home, pasan a ser las páginas de mayor interacción en la estructura resultante, en orden descendiente de importancia.

Figura 48: Grafo final modelo 1



Fuente: Grafo resultante de estructura final Modelo 1 usando Pajek

Modelo 2:

La configuración del modelo 2 requiere ajustar los valores de las constantes alfa y beta, lo que se hace calculando la utilidad dada por la Ecuación 14 de la matriz actual. Así, se busca hacer comparable ambos términos, de modo que el modelo no busque los extremos: minimizar la cantidad de hyperlinks o bien maximizarla de manera independiente; lo que

resultaría en la matriz nula o unitaria respectivamente. Así, los cálculos en la matriz inicial resultaron ser una cantidad de hyperlinks inicial de 317 y un beneficio de 1,39. Se puede ver un orden de magnitud de $3x10^2$ de diferencia, por lo que se estudia empíricamente alfa con valores entre [300;317] y beta entre [0,0139;1,39]. Con el mencionada estudio se fijó *alfa*= 300 y *beta*=0,2 al proporcionar una mejor matriz resultante y adecuada convergencia[98].

Tabla 25: Parámetros y resultados modelo 2

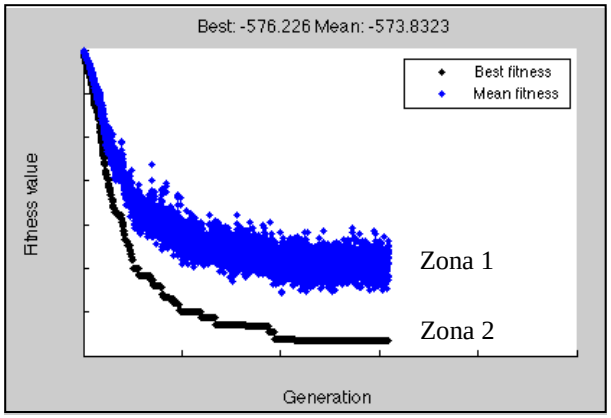
Resultados Modelo 2	
alfa	300
beta	0.2
Nº de links creados	97
Nº links eliminados	135
Links totales	279
Variación de la Función de utilidad	63%
Variación en el Nº de Links	-12%
Iteraciones	1571

Fuente: Elaboración propia

La Tabla 25, indica los resultados del mejor individuo resultante del modelo 2, el cual posee 97 hyperlinks creados, 135 hyperlinks eliminados y 393 hyperlinks en la misma situación (de ellos, 182 hyperlinks de mantuvieron existentes y 211 se mantuvieron como no existentes). La nueva adyacencia contempla 279 hyperlinks. El valor de la función de utilidad fue incrementada en un 52% con un 12% de reducción en la cantidad de hyperlinks. En relación al incremento de la utilidad, se utiliza la misma expresión de la Ecuación 16 pero con la matriz de pesos $\overline{\pi}$, obtenida por medio de la Ecuación 13.

La Figura 49 grafica el mejor valor de la función fitness a través de las generaciones, mostrando la convergencia de los individuos a lo largo de las iteraciones. En la figura, la zona 1 corresponde a la convergencia del fitness medio y la zona lineal decreciente (zona 2) corresponde a la convergencia del mejor individuo en cada generación. El individuo final fue obtenido luego de 1571 iteraciones.

Figura 49: Convergencia del mejor individuo

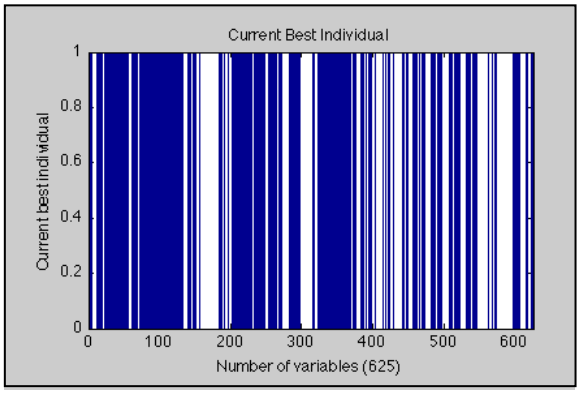


Fuente: Imagen obtenida de Matlab una vez terminada la ejecución

[98] En este contexto una convergencia adecuada tiene relación con no converger hacia valores extremos, es decir, lograr un equilibrio entre la maximización y minimización de hyperlinks.

Profundizando en la adyacencia final obtenida, la Figura 50 es la gráfica proporcionada por Matlab del mejor individuo, representa la densidad de la matriz resultante del modelo 2 el cual tiene una representación lineal.

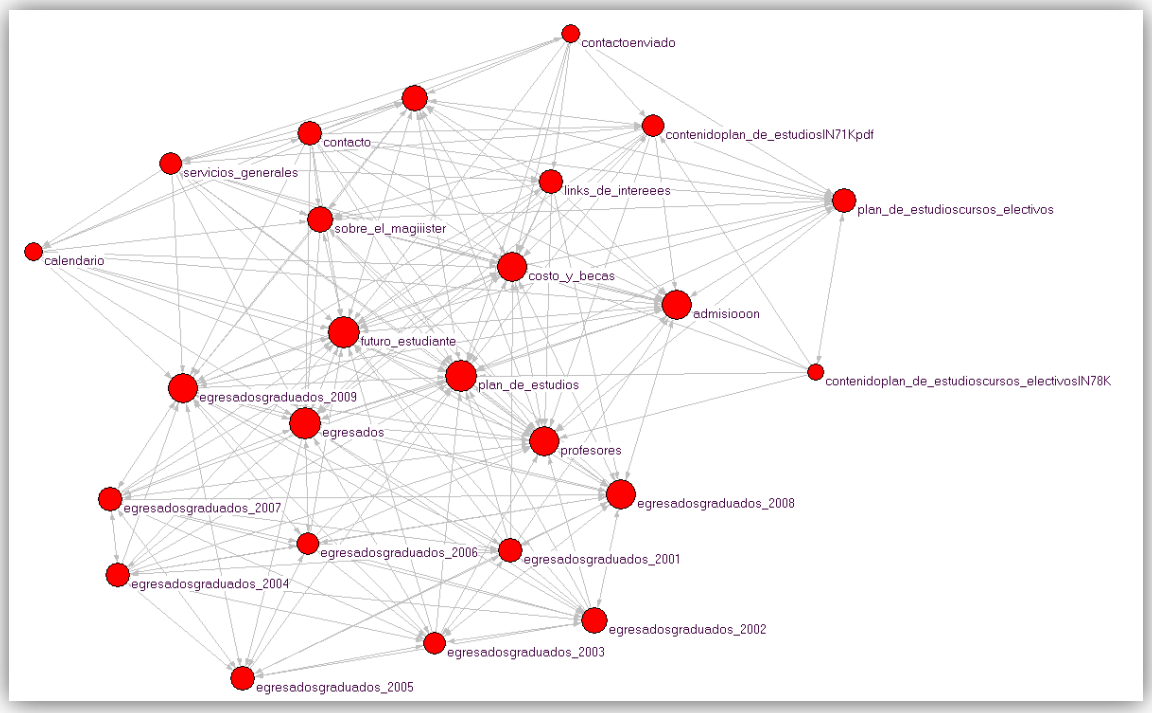
Figura 50: Densidad de mejor individuo modelo 2



Fuente: Imagen obtenida de Matlab una vez terminada la ejecución

El grafo final que representa el mejor individuo, y por ende, la nueva estructura web propuesta por el modelo 2, puede observarse en la Figura 51, la cual posee todas las interrelaciones tanto de entrada como de salida a los nodos e incorpora los tamaños de cada nodo, normalizando el peso representativo que en este caso indica una mayor importancia de las páginas. En este caso, destacan por su importancia plan de estudios, futuro estudiante, egresados, profesores, costos y becas, egresados 2009 y egresados 2008.

Figura 51: Grafo final modelo 2



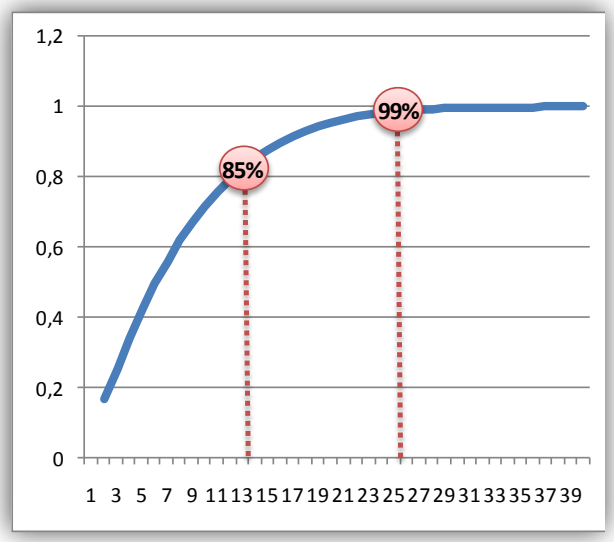
Fuente: Grafo resultante de estructura final Modelo 2 usando Pajek

Modelo 3:

El modelo 3 requiere ante todo, las matrices de pesos separadas por los largos de cada camino de acceso, con lo que se hace requiere definir la cantidad máxima. De acuerdo a los registros obtenidos, se observa que el máximo largo entre dos páginas del MGO es 41, valor registrado solo una vez. Con el objetivo de explorar la cantidad de transiciones que se representan con cada largo, se realiza la gráfica de largo v/s el porcentaje de transiciones contenidas. Considerando los valores de máxima inflexión de la curva de la Figura 51, se aprecia que con un largo de 26 se contiene un 99% de los datos, y con un largo de 14, un 85%; es decir, hay una pérdida de un 1% y 15% de las transiciones entre hyperlinks respetivamente. Por otro lado, calculando el promedio de las transiciones, se obtiene que 244 veces en promedio las transiciones son de largo 14, y 156 veces en promedio son de largo 26. Ante eso, y para evitar aún más la complejidad computacional, 14 serán en definitiva el total de matrices que iniciarán los cálculos (*H=14*)

Una vez creadas las matrices de peso, el modelo 3 requiere ajustar los valores del intervalo entre los que se establecerá que varíe la cantidad de hyperlinks, de modo que no minimice demasiado la cantidad de ellos pero que si busque una menor cantidad de los mismos. En ese sentido es que se fija el valor límite inferior como *n1*=200, pues bajo éste valor no se registran mayor aumento en la función de utilidad al haber una baja cantidad de hyperlinks de alto peso, y un límite superior de *n2*=300, esperando que se reduzca en un mínimo de 5% la cantidad total de hyperlinks (en comparación a los 317 hyperlinks actualmente existentes).

Figura 52: Curva de largos v/s porcentaje total de transiciones consideradas



Fuente: Elaboración propia

La Tabla 26 muestra la precisión indicada de los límites en la cantidad de hyperlinks y los resultados finales al comparar el individuo final con el inicial. Se puede observar que la nueva matriz crea 127 hyperlinks, elimina 144 y mantiene los 354 restantes (de ellos 173 hyperlinks de mantuvieron existentes y 181 se mantuvieron como no existentes). La nueva

adyacencia contiene un total de 300 hyperlinks. El valor de la función de utilidad fue incrementada en un 29% con un 5% de reducción en la cantidad de hyperlinks.

Tabla 26: Parámetros y resultados modelo 3

Resultados Modelo 3	
n1	200
n2	300
Nº de links creados	127
Nº links eliminados	144
Links totales	300
Variación de la Función de utilidad	29%
Variación en el Nº de Links	-5%
Iteraciones	5492

Fuente: Elaboración propia

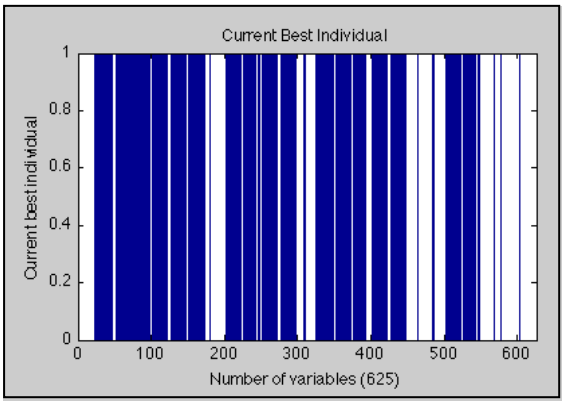
En relación al incremento de la utilidad, la ecuación 19 muestra el cálculo realizado. R_o es la adyacencia inicial (actualmente existente en el sitio Web) y R_f es la adyacencia final obtenida por el AG. Cada una de las 14 matrices de pesos $\pi^{(k)}$, es la obtenida por medio de la Ecuación 15.

Ecuación 19: Variación de la utilidad modelo 3

$$\frac{\sum_k \pi^{(k)} \cdot {R_f}^k - \sum_k \pi^{(k)} \cdot {R_o}^k}{\sum_k \pi^{(k)} \cdot {R_o}^k}$$

La Figura 53 es la gráfica de la densidad del mejor individuo.

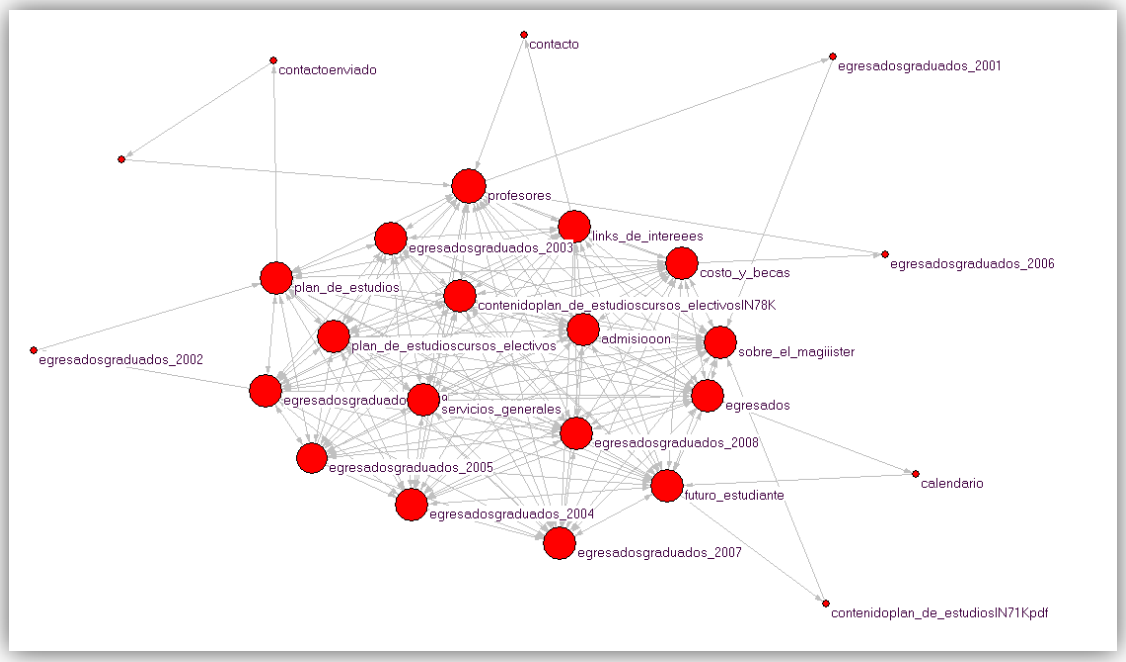
Figura 53: Densidad de mejor individuo modelo 3



Fuente: Imagen obtenida de Matlab una vez terminada la ejecución

La Figura 54 muestra la estructura del mejor individuo obtenido bajo la aplicación del modelo 3. Se aprecia un cúmulo central en la cual muchas páginas cobran vital importancia e interacción. Profesores es la página que mas sobresale, y por el contrario, Home, egresados 2001-2002-2006, contacto y calendario las menos insertas en el cúmulo y por ende en la red de comunicación con las páginas restantes.

Figura 54: Grafo final modelo 2



Fuente: Grafo resultante de estructura final Modelo 3 usando Pajek

5.- Análisis de resultados

Para profundizar en relación a los cambios generados, a continuación se indica la comparación de la matriz inicial con la nueva resultante de cada modelo, donde se estudiarán las áreas principales de modificaciones y mantención, de modo de encontrar los patrones de cambios asociados a cada procedimiento.

Especial hincapié se hace en la acción del webmaster quien, al ser un conocedor de los aspectos del negocio de la página, deberá analizar los cambios propuestos considerando el esqueleto propio del sitio y los intereses de la compañía. Específicamente en el sitio del MGO existen cambios detectados que eliminan hyperlinks hacia la página Home, pues el hyperlink es usado con baja frecuencia, sin embargo, será el webmaster el que tome la decisión de aceptar o rechazar dicha modificación pues por construcción del sitio y aspectos de usabilidad, suele existir el acceso hacia ella desde todas las páginas restantes.

Modelo 1

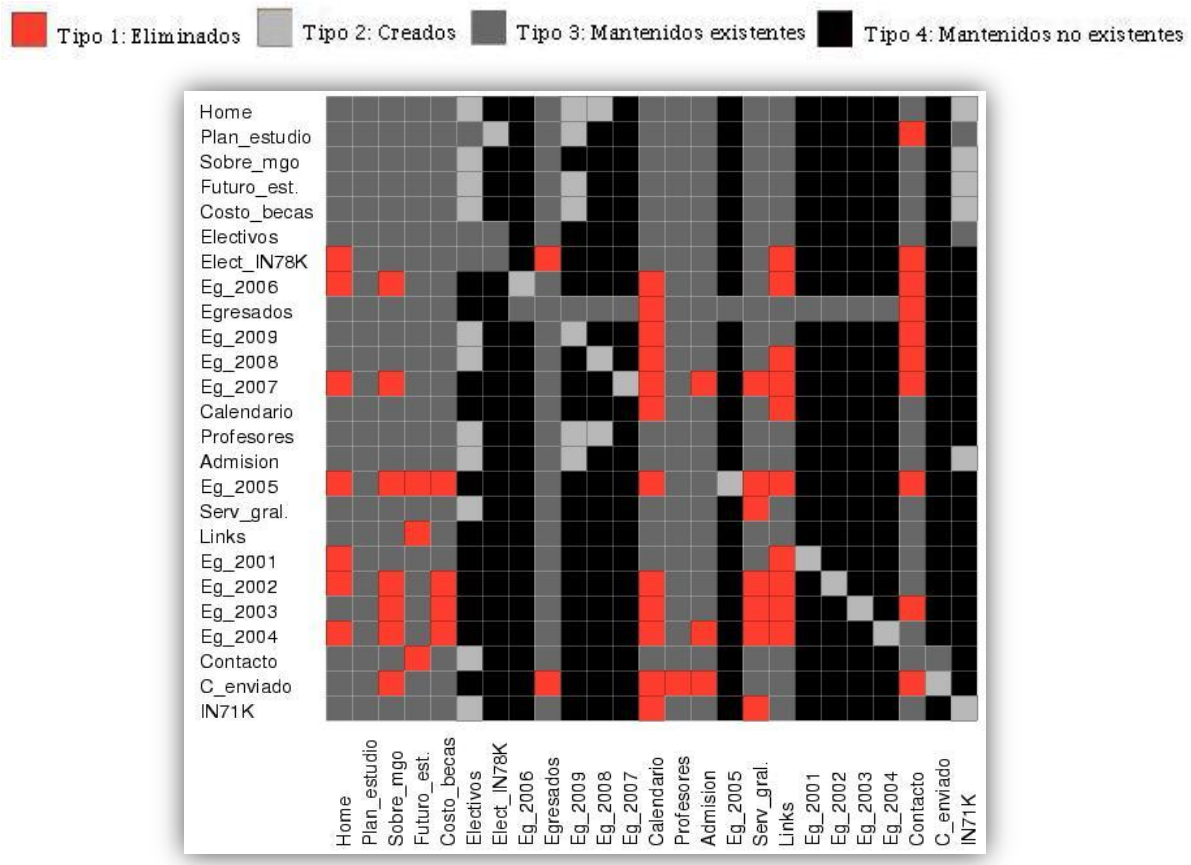
La Figura 55 muestra un recuadro que contiene las modificaciones realizadas por la adyacencia resultante del AG, comparada con la matriz inicial actualmente existente en el sitio web. Los cambios principales se resumen en cuatro acciones: eliminar un hyperlink, agregar un hyperlink, mantenerlo existente o mantenerlo como no existente. Los cuadros tipo 1 (color rojo) significan que el hyperlink fue eliminado, los tipo 2 (color gris claro) que fueron creados, los tipo 3 (gris oscuros) se mantuvieron existentes, y tipo 4 (color negro) se mantuvieron como no existentes. En definitiva, los tipos 2 y 3, son hyperlinks existentes con valor 1 en la matriz binaria, y los tipo 1 y 4, hyperlinks no existentes, con valor 0. Dicha matriz comparativa de las estructuras es una imagen resultante gracias a realizar *block model* en Pajek [19], ingresando una matriz contendedora de la diferencia punto por punto de las matrices binarias final e inicial, condicionando por existencia o no existencia inicial. Blockmodeling es una técnica capaz de detectar la cohesión, centro-periferia, las estructuras, y el ranking, efectiva para pequeñas redes densas la cual se basa en diferentes conceptos estructurales como la equivalencia y posiciones.

Con énfasis en la cantidad de hyperlinks eliminados y creados, se observa una baja cantidad de éstos últimos, los cuales se enfocan principalmente a la creación de 10 hyperlinks de acceso entre ciertas páginas (Home, sobre el magíster, costo y becas por ejemplo) hacia el plan de estudio de cursos electivos (columna 6), también se crearon 7 hyperlinks hacia la página de egresados 2009 y 3 hacia egresados 2008 (columna 10) desde profesores y página Home por ejemplo, y 6 hyperlinks hacia los contenidos específicos del curso IN71K (columna 25). Cabe señalar que también se crean auto-links entre páginas sobre todo entre páginas de egresados.

En cuanto a la eliminación de hyperlinks, ellos se encuentran más dispersos, eliminando hyperlinks hacia las páginas Home, sobre el magíster, futuro estudiante, costos y becas, egresados, calendario, profesores, admisión, servicio general, hyperlinks de interés y contacto. Todas las páginas anteriores cumplen con una misma característica: son páginas del

tipo menú. Los hyperlinks de acceso a ellas son eliminados principalmente desde páginas que no son del tipo menú y tienen que ver con los egresados de distintos años en su mayoría.

Figura 55: Block Model sobre cambios, modelo 1



Fuente: Elaboración propia a partir de Block Model en Pajek

La Tabla 27 muestra ejemplos específicos de modificaciones en los hyperlinks ya sea agregándolos o eliminándolos. Se observa la creación de los hyperlinks 01 y 02, el primero como necesidad de tener un acceso directo desde la página Home a la información de cursos electivos y el segundo, creando un acceso directo entre la página profesores del menú, hacia la información de los últimos graduados. En cuanto a la eliminación se observa un peso nulo, lo que por el procedimiento propio del cálculo, significa que no hay registros de acceso entre dichas páginas. El hyperlink 03 elimina el hyperlinks desde contacto/enviado hacia sobre el magíster (página menú). Se observa además que se eliminan bastantes hyperlinks desde contacto/enviado, lo que la catalogaría como página de término de sesión. El hyperlink 04 por otro lado, elimina el contacto entre hyperlink de interés y futuro estudiante, lo que induce un orden lógico de navegación donde futuro estudiante se visita antes que hyperlinks de interés.

Tabla 27: Ejemplo específicos modificaciones modelo 1

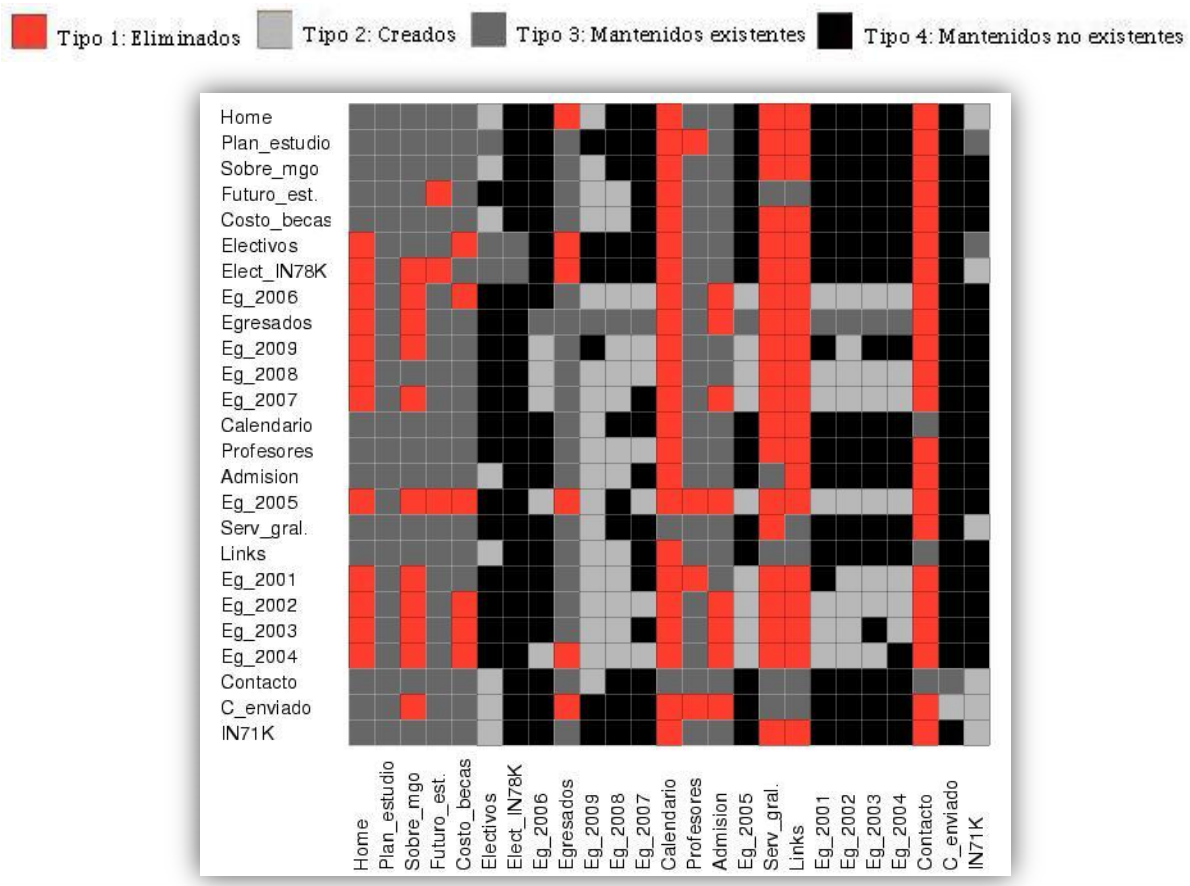
Nº	Acción	Peso	Interpretación
Link 01	Agregado	0.0014	Acceso directo hacia cursos electivos
	Desde	/mgo2007	
	Hasta	/mgo2007/plan_de_estudios/cursos_electivos	
Link 02	Agregado	0.000416	Acceso directo hacia últimos graduados
	Desde	/mgo2007/profesores	
	Hasta	/mgo2007/egresados/graduados_2009	
Link 03	Eliminado	0	Página fin
	Desde	/mgo2007/contacto/enviado	
	Hasta	/mgo2007/sobre_el_magíster	
Link 04	Eliminado	0	Orden de navegación
	Desde	/mgo2007/links_de_interés	
	Hasta	/mgo2007/futuro_estudiante	

Fuente: Elaboración propia

Modelo 2

La figura 56 muestra el block model asociado al modelo 2, construido comparando siempre la matriz final del cada modelo, con la inicial existente en el sitio.

Figura 56: Block Model sobre cambios, modelo 2



Fuente: Elaboración propia a partir de Block Model en Pajek

Enfocando el estudio en la creación y eliminación, se desprenden las características principales asociadas a las modificaciones que el modelo 2 entrega. En relación a las

eliminación, se observan 22 hyperlinks que iban a contacto y que fueron eliminados, es decir, solo quedaría la posibilidad de hacer acceder a contacto desde calendario, hyperlinks de interés y sí misma. Esto puede resultar extraño desde el punto de vista estructural y en post de la idea de facilitar el contacto por parte de los usuarios al tener el hyperlink de acceso siempre disponible, sin embargo, el modelo refleja que su uso no es muy elevado en comparación a las otras interrelaciones, por lo que queda a juicio del webmaster evaluar si seguir esta indicación o mantiene los hyperlinks hacia contacto a modo de formalidad. 20 hyperlinks son eliminadas hacia servicios generales e hyperlinks de interés, lo que también indica un bajo uso de ellas, y un acceso fijo desde otras páginas específicas. Calendario se presenta como una página de poco interés, al tener solo 2 hyperlinks de llegada lo que representa la baja utilidad de acceder a ella, en este contexto, sería interesante evaluar temas de contenido. La eliminación de 11 hyperlinks hacia sobre el magíster indica una secuencia lógica de navegación, donde ésta es visitada más bien un comienzo, no al profundizar[99] en la estructura Web

Con respecto a la creación de hyperlinks, egresados 2009 recibe gran parte de ellos como también egresados 2008 y 2007, además de interrelaciones propias entre egresados de distintos años y la creación de hyperlinks hacia la información directa de los cursos electivos.

La Tabla 28 especifica cuatro ejemplos concretos de creación y destrucción de hyperlinks. Por ejemplo, el hyperlink 01 tiene un alto valor en su peso, resultando que es un nuevo hyperlink creado para economizar el tiempo de navegación del usuario. El descubrimiento de un acceso directo requerido (hyperlinks 01 y 02) introdujo del uso de esos hyperlinks, a fin de mejorar y habilitar la comunicación entre páginas de los estudiantes graduados de diferentes años lo cual no ocurre hoy en día. El caso de la eliminación de un hyperlink como el número 03 fue provocado por su bajo peso junto al poco acceso a la página calendario. Este hyperlink existe porque hay una barra de menú fija, donde calendario se encuentra actualmente. El hyperlink 04 refleja una necesidad de reorganización del menú, donde egresados no debería estar en el menú desde un comienzo, sino que aparecer después una vez realizado el primer click, es decir, pertenecer a algún submenú.

Tabla 28: Ejemplo hyperlinks creados y eliminados modelo 2

Nº	Acción	Peso	Interpretación
Link 01	Agregado	0.045	Acceso directo entre graduados
	Desde	/mgo2007/egresados/graduados_2001	
	Hasta	/mgo2007/egresados/graduados_2002/	
Link 02	Agregado	0.040	Acceso directo entre graduados
	Desde	/mgo2007/egresados/graduados_2009	
	Hasta	/mgo2007/egresados/graduados_2008	
Link 03	Eliminado	1,90E-05	Bajo acceso a Calendario
	Desde	/mgo2007/egresados/ graduados_2001	
	Hasta	/mgo2007/calendario	
Link 04	Eliminado	2,30E-05	Reorganización de Menú
	Desde	/mgo2007	
	Hasta	/mgo2007/egresados	

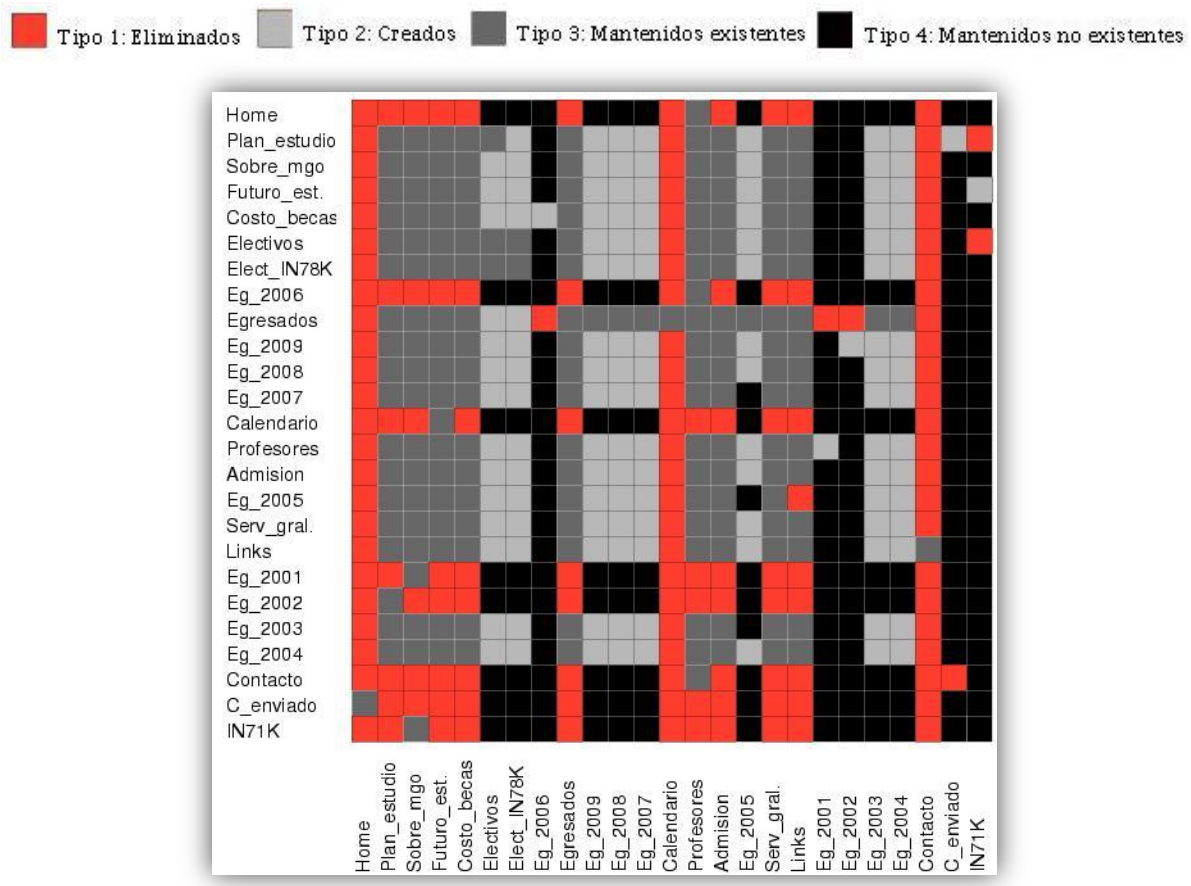
Fuente: Elaboración propia comparando matriz inicial con resultante

[99] Entiéndase por profundizar en la estructura web, al hecho de ir visitando hyperlinks no visibles desde el comienzo, de modo de visitar más páginas avanzado en el orden jerárquico de las mismas

Modelo 3

El modelo 3 por su lado elimina gran cantidad de hyperlinks hacia las páginas Home y contactos, dos páginas que se suelen mantener siempre accesible tal como se observa en la Figura 57.

Figura 57: Block Model sobre cambios, modelo 3



Fuente: Elaboración propia a partir de Block Model en Pajek

Una vez más, el presente modelo indica que la tasa de uso de acceso a ellas por medio de otros hyperlinks no es tan elevada. Calendario nuevamente se reporta como una página de bajo interés y acceso pues solo desde la página egresados sería posible acceder a ella. Egresados 2008, 2001, 2002 y contacto, eliminan gran cantidad de sus hyperlinks hacia páginas del menú, lo que las señalaría como páginas de término (no se accede desde ellas a otras páginas). En este sentido, hacer promoción en ellas, con información de interés (ya sea descuentos, fechas claves de postulación, logros y noticias) sería útil de modo de mantener a los usuarios navegando en el sitio.

La creación de hyperlinks es bastante precisa: se crean hyperlinks directos hacia la información de cursos electivos, hacia los egresados 2009, 2008 y 2007 (últimos graduados), egresados 2005, 2004 y 2003 (graduados en general).

La tabla 29 indica 4 hyperlinks modificados, el primero fue creado como necesidad de acceso directo entre graduados 2008 y graduados 2003. Luego, el hyperlink 02 ejemplifica un caso que requiere acceder hacia la información de los últimos graduados (en este caso, egresados 2008). Hyperlink 03 elimina la relación entre profesores y calendario, dada la baja frecuencia de uso de ésta última y la necesidad de reorganizar el menú de modo de ubicarla adecuadamente. Entre contacto/enviado y costo y becas también se borra la conexión (hyperlink 04), aquí el motivo tiene que ver con la secuencia lógica de navegación (pues no se accede frecuentemente desde contacto/enviado a costos y becas según los registros de la Web), y también con el hecho de que desde contacto/enviado no se registran hyperlinks.

Tabla 29: Ejemplo hyperlinks creados y eliminados modelo 3

Nº	Acción	Peso	Interpretación
Link 01	Agregado	1,24E+15	Acceso directo entre graduados
	Desde	/mgo2007/egresados/graduados_2008	
	Hasta	/mgo2007/egresados/graduados_2003	
Link 02	Agregado	1,27E+14	Acceso directo hacia últimos graduados
	Desde	/mgo2007/costo_y_becas	
	Hasta	/mgo2007/egresados/graduados_2008	
Link 03	Eliminado	6,68E+07	Reorganización del Menú
	Desde	/mgo2007/profesores	
	Hasta	/mgo2007/calendario	
Link 04	Eliminado	0,00E+00	Orden de navegación y página fin
	Desde	/mgo2007/contacto/enviado	
	Hasta	/mgo2007/costo_y_becas	

Fuente: Elaboración propia comparando matriz inicial con resultante

5.1.- Medidas de efectividad

La navegabilidad (facilidad con que se puede visitar una secuencia de páginas) puede ser evaluada de tres maneras, con entrevistas, con análisis de uso y con medidas de navegabilidad [66]. Por ello, se realiza cada uno de esos tipos de análisis en las matrices resultantes de los modelos.

El análisis de uso es lo que sustenta cada modelo, pues el mismo uso impulsa los cambios realizados, tomando diversas medidas (expresiones matemáticas) con el fin de reflejar la importancia de contar con cierto hyperlink o bien, la necesidad de crearlo. Cabe señalar que el análisis de uso ya fue realizado, pues consiste en el porcentaje de variación de la función de utilidad, donde se tuvo mejoras de un 51%, 62% y 29% en los modelos 1, 2 y 3 respectivamente.

5.1.1.- Encuesta

A modo de análisis de la usabilidad y navegabilidad de los cambios resultantes de la aplicación de los modelos, es que se realizó una encuesta a usuarios para obtener sus apreciaciones cualitativas a fin de poder cuantificar. Luego de un análisis exhaustivo de todos los cambios efectuados, es que éstos fueron clasificados en cinco tipos[100]:

[100] Para visualizar los cambios efectuados, **ver sección** 7.4.- Base encuesta

Cambio 1: Sub Menú

Se propone una nueva organización del menú inicial, agregando 2 sub menús en la página de inicio hacia cursos electivos y obligatorios.

Cambio 2: Últimos graduados

Se considera tener un acceso directo en el menú principal para acceder a la información de los últimos graduados (2009-2008)

Cambio 3: Reorganización

Se propone un menú fijo en la página inicial y un menú interno que despliega nuevos hyperlinks al presionar alguna página del Menú Inicial. Menú principal compuesto por las páginas: Sobre el Magíster, Plan de Estudios (cursos obligatorios, cursos electivos), Admisión, Costos y Becas, Profesores y Futuro Estudiante. Menú interior compuesto por lo mismo anterior más Egresados, Servicios Generales, hyperlinks de interés, Contacto y Calendario.

Cambio 4: Entre graduados

Se propone que en cada página de graduados se tenga acceso a los graduados de otros años.

Cambio 5: Promoción

Agregar hyperlink de información de postulación (o noticia destacada, fechas claves, descuentos) en algunas páginas, sobre todo en aquellas páginas detectadas como páginas de término de sesión.

En la clasificación anterior, los cambios 1,2 y 4 tienen que ver con la creación de hyperlinks y los cambios 3 y 5, con la eliminación.

El Universo a estudiar comprende 22 encuestas realizadas. Desde el punto de vista cualitativo, se pide a los usuarios comentar y dar impresiones ante cada cambio estudiado, y por sobre todo, un comentario final a modo de sugerencia global al sitio. Como variables características nominales[101], en la Figura 58 se observa la distribución de los encuestados en relación al tipo de usuario que son clasificados como estudiante MGO, profesor y vías titulación. Este último grupo corresponde a estudiantes de pregrado.

Figura 58: Gráfica de tipo de usuario entrevistado

Fuente: Elaboración propia

[101] **Variable nominal:** Variable cualitativa la cual no puede ser sometida a un criterio de orden jerárquico, como color de ojos, nacionalidad, etc.

La Figura 59 por otro lado, muestra la principal razón de ingreso al sitio, dentro de las que destacan la búsqueda de información general sobre qué temas son abordados por el magíster (35%), sobre la malla, plan de estudios y cursos en especial (22%), novedades en cuanto a noticias y últimas investigaciones (17%), egresados y papers relacionados para ver los avances académicos logrados (13%), y por casualidad (13%), es decir, donde el ingreso no fue previamente planificado.

Figura 59: Gráfica de principal razón de ingreso al sitio

Fuente: Elaboración propia

En cuanto a la frecuencia de acceso al sitio, 9% de los encuestados declararon ingresar 2 veces por semana, 41% entran en promedio 2 veces al mes, y 60% en promedio 2 veces al año. Considerando 12 meses y 52 semanas al año, se tiene una frecuencia de acceso promedio de 17 veces al año por cada usuario. Considerando que la unidad de medida es anual, cabe destacar la baja frecuencia de ingreso al sitio, lo cual permite extrapolar un comportamiento de bajo acceso incluso por parte de los mismos estudiantes del magíster, pues como ellos mismos dicen, todas las dudas pueden ser resueltas en la oficina central del MGO y no del todo por la página. Esto induce a establecer que el sitio requiere una reconstrucción tanto en estructura como en contenidos entregados, de modo de poder resolver las principales inquietudes y entregar lo que los usuarios están buscando en concreto, para con ello aumentar la frecuencia de acceso a la misma.

Desde el punto de vista cuantitativo, se les pide a los encuestados indicar si el cambio les parece muy útil, útil, poco útil o inútil, que numéricamente serán identificados como nota de los cambios desde 0 a 3, donde 0 es inútil (nota mínima) y 3 muy útil (nota máxima).

Para poder llevar a cabo un análisis comparativo de los modelos $m \in \{1,2,3\}$ es que se crea una medida que permita cuantificar los cambios efectuados en relación a la nota que los usuarios encuestados le proporcionaron. En este sentido se buscará ponderar la nota promedio $\bar{x}(i)$ obtenida por cada cambio $i \in \{1,2,3,4,5\}$ con el porcentaje de hyperlinks de cada modelo que tienen que ver con dicho cambio, es decir, $n_m(i)/TC_m$ donde $n_m(i)$ corresponde a la cantidad de cambios del tipo *i* en el modelo *m*, y TC_m a la cantidad total de cambios de eliminación y creación del modelo *m*. Así, se buscará tener la suma ponderada total, lo que se aprecia en la Ecuación 20.

Ecuación 20: Cuantificación de resultados encuesta

$$E(m) = \frac{\sum_i n_m(i) \cdot \bar{x}(i)}{TC_m}$$

Teniendo como base dicho procedimiento, es que se calcula el promedio de la nota obtenida por cada cambio, luego, se calculan los porcentajes de ellos que tiene cada modelo y así, se calcula la nota final ponderada, como se observa en la Tabla 30. Allí, se obtiene una nota final por modelo de 2.1, 2.24 y 2.03 para los Modelos 1, 2 y 3 respectivamente, es decir, reporta más utilidades el modelo 2, sobre todo porque tiene un mayor porcentajes de cambios del tipo 4 que en promedio es el que tiene mejor calificación por parte de los usuarios.

Antes de proceder con los cálculos, se clasifica cada hyperlink modificado (creados y eliminados). En dicho proceso de clasificación no se consideran ciertos hyperlinks que van entre las mismas páginas (autolinks), que eliminan hyperlinks hacia la página inicial (Home) o bien, aquellos que no pudieron ser clasificados. Los autolinks no se consideran dado que por la forma de navegación, siempre es posible actualizar la página en la que se navega, y por ende, acceder desde una misma página a otra, lo que no requiere mayores modificaciones ni estudios pues siempre es posible por construcción. Los hyperlinks hacia Home tampoco se considerarán en el estudio, debido a que igualmente siempre es posible acceder a la página inicial desde cualquier parte en donde uno se encuentre, lo que en general no reportaría mayores modificaciones. Finalmente hay hyperlinks omitidos, los cuales no pudieron ser clasificados dentro de los patrones de cambios detectados.

Cabe señalar que por porcentajes de hyperlinks asociados a cada cambio por modelo (valores en columnas cambio 1, cambio 2, cambio 3, cambio 4 y cambio 5), son calculados con respecto al porcentaje del total clasificado. Por ejemplo, del 100% de los cambios efectuados por el modelo 1 (de los 102 cambios, 81 corresponden a cambios de creación o eliminación) solo un 79% de ellos es clasificado.

Tabla 30: Estudio de cambios clasificados

Modelo\Nota	Total	Autolinks	Home	Omitidos	Total calificado	% de Total
Modelo 1	102	14	7	0	81	79%
Modelo 2	232	8	12	2	210	91%
Modelo 3	271	7	24	17	223	82%

Fuente: Elaboración propia

Considerando entonces la cantidad de hyperlinks indicadas en total calificado por modelo, es que se obtienen los porcentajes y notas por cambio indicados en la Tabla 31, obteniendo la nota final ponderada por modelo.

Tabla 31: Calificaciones de los cambios propuestos vía encuesta

	Cambio 1	Cambio 2	Cambio 3	Cambio 4	Cambio 5	
Modelo\Nota	2,26	1,58	2	2,58	2,21	Nota final
Modelo 1	20%	11%	26%	0%	43%	**2,10**
Modelo 2	6%	8%	19%	30%	38%	**2,24**
Modelo 3	13%	30%	15%	11%	31%	**2,03**

Fuente: Elaboración propia

Se destaca la baja calificación del cambio 2, el cual crea un acceso directo hacia los últimos graduados. Si bien las personas parecen no estar muy sorprendidas, ni necesitar este hyperlink directo, los análisis de uso evidencian que una gran proporción de las personas gusta de ver la experiencia de otros, sobre todos de aquellos egresados más recientes.

Un aspecto no menor y que suele causar discusiones es el cambio 3, en relación a la nueva organización del menú. Por ello, es que se les pide a los encuestados evaluar cada hyperlink propuesto del menú interior, de modo de desprender aquellos mejor evaluados en el menú inicial validando la organización propuesta en el cambio 3[102]. La tabla 32 contiene la calificación promedio de cada página. Aquellas páginas destacadas confirman el contenido del menú principal pues son las de mayor interés, las no destacadas, son las presentadas como menú interno. Cabe señalar que la página profesores no es muy valorada por medio de la encuesta, pero sí se presenta en el menú principal, además, contacto posee una baja tasa de uso efectivo pese a recibir buena evaluación por parte de los usuarios.

Tabla 32: Calificación promedio del menú interior

Página	Nota Promedio
Sobre el Magíster	2,41
Plan de Estudios	2,82
Cursos Obligatorios	2,55
Cursos electivos	2,50
Admisión	2,55
Costos y becas	2,77
Profesores	2,09
Egresados	1,73
Futuro Estudiante	2,14
Servicios Generales	1,36
Links de interés	1,14
Calendario	1,45
Contacto	2,14

Fuente: Elaboración propia

Como impresiones finales de la encuesta y sugerencias hacia el sitio Web de acuerdo a la navegación que en él se desarrolla, destacan menciones para que se incorpore más dinamismo en el sitio web, ya sea con la incorporación de tecnología Ajax[103] para el muestreo

[102] Ver Figura D.4, Sección D.- Base Encuesta
[103] **Ajax**: Técnica de desarrollo web para crear aplicaciones interactivas

de hyperlinks, información, o bien, contar con noticias no estáticas mediante algún banner[104] donde ellas aparezcan, relacionada con los logros de los profesores y estudiantes del magíster. Además, se pide un mejor diseño de la página central, en relación a la imagen principal que no se considera del todo representativa. Por otro lado, se sugiere tener una página de FAQ[105], de modo de poder resolver dudas comunes de manera más directa por parte de los postulantes, estudiantes y profesores, evitando tener que ir a la oficina central del MGO, incluso, contar con una "guía de postulación" o "guía de admisión" se considera bastante útil, sobre todo para alumnos externos, ya sea de otras facultades de la Universidad de Chile, otras universidades o incluso, extranjeros.

5.1.2.- Índice de navegabilidad.

R. Botafogo [66], propone una métrica para identificar propiedades de los hipertextos, donde si bien ignora aspectos de perspectiva de los usuarios, refleja de forma bastante fidedigna el comportamiento específicamente estructural de la adyacencia. En este sentido, la principal medida es la **compacidad**, la cual tiene que ver con la cantidad total de hyperlinks y las referencias cruzadas que ellos producen.

Para el cálculo de la compacidad, se debe calcular el largo del camino mínimo para ir de un nodo a otro, disponiendo dichos valores (largos) en una matriz de distancia *D*. Si no existe algún camino posible entre dos nodos, se completa dicho valor entre *i* y *j*, con una constante de conversión *K*. Dicha constante debe garantizar que será un valor al menos tan alto como cualquier entrada finita de la matriz de distancia. Si bien en general se considera como el valor propio de la cantidad total de nodos n (pues el largo más largo posible en esas circunstancias es *n-1*), considerar *K* como la distancia máxima encontrada en los nodos suele ser bastante representativo y comparable entre matrices. La expresión para la compacidad C_p queda de la forma indicada en la Ecuación 21, siendo *n* el número total de nodos.

Ecuación 21: Compacidad

$$C_p = \frac{(n^2 - n) \cdot K - \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}}{(n^2 - n) \cdot (K - 1)}$$

C_p tendrá valores entre 0 y 1, donde 0 indica un grafo completamente desconectado y 1, un grafo completamente conectado[106]. Una alta compacidad significa que fácilmente se puede visitar cada página, pero la abundancia de hyperlinks puede agobiar a los usuarios y dirigir hacia la desorientación. Por otro lado, con muy pocos hyperlinks, los usuarios podrían llegar a estar igualmente desorientados, dado que necesitan ir a través de una serie de pasos para obtener la información objetivo. Por ende, para una buena navegación, la idea es evitar los extremos, donde valores de compacidad entre 0,3 y 0,8 tendrían un buen ajuste.

[104] **Banner:** Formato publicitario en Internet, a mode de publicidad online, cuyo su objetivo es atraer visitas.
[105] **FAQ:** Frequently Asked Questions
[106] Ver concepto de grafo completo en sección C.- Conceptos utilizados de teoría de grafos y cadenas de Markov

En cuanto el índice de compacidad, se considera la cantidad total de nodos sin contemplar los cambios entrantes o salientes de la página Home, es decir, con n=24. La constante K por otro lado toma el valor 4, el máximo largo registrado en todas las matrices.

Tabla 33: Valores de compacidad

	Compacidad	Variación con respecto a inicial
Original	0,816	
Modelo 1	0,793	-3%
Modelo 2	0,745	-9%
Modelo 3	0,800	-2%

Fuente: Elaboración propia

Considerando lo indicado en el párrafo anterior, la Tabla 33 muestra los valores obtenidos. Se observa que el valor numérico de la compacidad obtenida por todos los modelos es mucho más cercano al rango aceptable de *Cp* (entre 0,3 y 0,8), y logran disminuir el valor de la compacidad original. Cabe señalar que el modelo 2 se incorpora de mejor manera en el rango aceptable al disminuir en un 9% la compacidad de la matriz inicial. Ello indica que contiene un mejor índice de navegabilidad en comparación a las matrices resultantes de los modelos restantes.

La Tabla 34 resume las mejoras de los tres modelos de acuerdo a las medidas usadas, como son porcentaje de aumento de beneficio, nota de calificación por medio de una encuesta e índice de navegabilidad dado por la compacidad. Se agrega además el grado de entrada y salida de hyperlinks por página, indicándose las 5 páginas con mayor cantidad de cada de ellos. Para mayores detalles del grado de centralidad de las páginas, ver sección de apéndices: 7.5.4. Grado de centralidad de nodos.

Tabla 34: Resultados finales por modelo

Resultados Modelo 1	
Variación Utilidad	51%
Variación en Nº de Links	-9%
Iteraciones	1745
Calificación de cambios	2,1
Compacidad	0,793
Páginas Hub	
Egresados	19
Home	16
Plan_de_estudios	15
Futuro_estudiante	15
Costo_y_becas	15
Páginas Autoritativas	
Plan_de_estudios	25
Profesores	24
Egresados	23
Futuro_estudiante	22
Admisión	22

Resultados Modelo 2	
Variación Utilidad	63%
Variación en Nº de Links	-12%
Iteraciones	1571
Calificación de cambios	2,24
Compacidad	0,745
Páginas Hub	
Graduados_2008	16
Contacto	16
Egresados	14
Links_de_intereees	14
Graduados_2007	13
Páginas Autoritativas	
Plan_de_estudios	25
Futuro_estudiante	22
Profesores	21
Costo_y_becas	19
Egresados	19

Resultados Modelo 3	
Variación Utilidad	29%
Variación en Nº de Links	-5%
Iteraciones	5492
Calificación de cambios	2,03
Compacidad	0,8
Páginas Hub	
Contacto	18
Links_de_intereees	18
Graduados_2007	18
Graduados_2002	18
Futuro_estudiante	18
Páginas Autoritativas	
Graduados_2007	20
Profesores	19
Futuro_estudiante	18
Costo_y_becas	18
Egresados	17

Fuente: Elaboración propia

En aspectos generales, la ventaja de todos los modelos es el uso de AG y sus propiedades aptas de optimización, además de una configuración de la función objetivo del tipo fitness puro[107]. Las ventajas de cada modelo se resumen en la Tabla 35.

Tabla 35: Ventajas comparativas de cada modelo

Ventajas		
Modelo 1: Camino mínimo	**Modelo 2: Utilidad lineal**	**Modelo 3: Utilidad potencial**
Base en la teoría de cadenas de Markov y conceptos de probabilidades.	Es posible desprender de él todos los patrones de cambio.	Representatividad del largo de caminos en cada sesión.
Creación de "peso potencial" de los hyperlinks no existentes.	Creación del "acceso objetivo", reflejo de la importancia de cada link existente o no.	Incluye una cantidad representativa de todos los caminos posibles.
Tiempo de ejecución aceptable para la cantidad de datos estudiados en muchas generaciones.	Modelo lineal que podría ser resuelto más rápidamente con otros algoritmos.	Alto tiempo de ejecución en comparación a los otros modelos.
	Tiempo de ejecución aceptable para la cantidad de datos estudiados en muchas generaciones.	

Fuente: Elaboración propia

Por otro lado, la desventaja global es la falta de dinamismo de los modelos, pues no permiten realizar estudios sincrónicos a las visitas en un sitio web, sino que requieren de la extracción de datos en períodos de tiempo fijos. Las desventajas por modelo se observan en la Tabla 36.

Tabla 36: Desventajas comparativas de cada modelo

Desventajas		
Modelo 1: Camino mínimo	**Modelo 2: Utilidad lineal**	**Modelo 3: Utilidad potencial**
Dificultad en aproximación de parámetros.	Valores acumulados, sin detallar por cantidad de hyperlinks necesarios para llegar al objetivo.	Alta cantidad de links que no pudieron ser clasificados
Imprecisión numérica al tratar de determinar una mayor o menor importancia de los pesos.	Manejo empírico de parámetros para obtener el óptimo.	No se consideran los datos asociados a todos los caminos posibles.
Uso de modelo de Markov agregado, sin separar por etapas (cantidad de links entre páginas).		Imprecisión numérica al tratar de determinar una mayor o menor importancia de los pesos.

Fuente: Elaboración propia

[107] **Ver sección** 2.5.4.- Codificación del algoritmo

6.- Conclusiones

El uso de Matlab con los toolbox de algoritmos genéticos y una apropiada interfaz para su aplicación, fue una útil herramienta pues permitió poder enfocarse en la construcción adecuada de mejor fitness representativo, además de facilitar el manejo matricial de las interrelaciones entre páginas y las variables asociadas. Por otro lado, el uso del programa Pajek, enfocado principalmente para el estudio de SNA[108] en redes sociales, agrega considerable valor al momento de evaluar las matrices resultantes de cada modelo, poder comparar y generar análisis de resultados desde un punto de vista más visual y claro.

Dando paso al diseño del algoritmo con la creación de tres modelos representativos de la utilidad de los hyperlinks, de esencia probabilística común, pero diferentes expresiones y principios matemáticos, se pudo explorar con mayor amplitud diversas formas de creación de la utilidad de los hyperlinks, dada su existencia o no existencia. Dicha utilidad es un valor sin precedentes en la literatura y fue construida mediante tres modelos a fin de explorar diversas posibilidades numéricas de representación, con sustento en el análisis de la red de hyperlinks y la creación de un peso adecuado para cada uno de ellos.

Producto de la investigación para obtener la construcción de pesos adecuados, se destacan los conceptos de camino mínimo y acceso objetivo aquí desarrollados. El primero fue la base del modelo 1, el cual contempla teoría de probabilidades y cadenas de Markov, con lo que permitió construir un primer acercamiento hacia la utilidad que podría generar el crear hyperlinks que actualmente no existen. Sin embargo, presenta ciertos errores numéricos al momento de comparar los pesos, propio de una mezcla de procesos de Markov en distintos largos (teoría de Markov por etapas[109]). En base a lo anterior, se da paso al concepto de acceso objetivo, relacionado con el hecho de acceder desde la página *i* a la *j* en algún momento en la sesión con páginas intermedias, cuando hay una cantidad mayor a 1 de hyperlinks para acceder desde una página a otra. En otras palabras, corresponde a una mirada prospectiva en cada sesión, considerando en qué página se está en un momento dado y a cuales se accede posteriormente. Dicho concepto fue la base de los modelos 2 y 3.

Los algoritmos genéticos resultaron tener una útil aplicación en el estudio de la estructura de navegación web, pues en este caso, el beneficio aumenta en un 51%, 63% y 28% en comparación a la actualmente existente, con las nuevas estructuras resultantes del modelo 1, 2 y 3 respectivamente. El exponer porcentajes concretos de aumento de beneficios desde el punto de vista numérico, percepción de los usuarios e índice de navegabilidad, permitió estudiar los resultados entregados por cada modelo a fin de compararlos entre sí. El mejor de ellos resultó ser el modelo 2: utilidad lineal, con un 63% de aumento de beneficio, nota 2,2 en promedio por parte de los usuario en cuanto a los cambios propuestos (de un máximo de 3) y un índice de navegabilidad equilibrado entre exceso y escasez de hyperlink (0,79). De hecho, con la validación concreta de los modelos, sobre todo del modelo 2, es posible expandirlos a otros sitios web de mayor tamaño, cuya estructura de datos y registros estén constituidos de

[108] **SNA:** Social Network Analysis
[109] **Ver apéndice** C.- Conceptos utilizados de teoría de grafos y cadenas de Markov

manera similar a los aquí presentados, pues realizar un proceso de sesionalización fidedigno y sin mayores errores fue clave en el reflejo del uso del sitio por parte de los usuarios.

La Tabla 37 muestra los principales patrones de cambio desprendidos de los modelos, explicando el tipo de cambio y de la acción propuesta en el sitio web. Se hace mención a la adyacencia final resultante del modelo 2, la cual favorece un mejor y mayor acceso a la página de plan de estudios (se cran una alta cantidad de enlaces hacia ella) y hacia la información como futuro estudiante.

Tabla 37: Tipos de cambios y acciones propuestas

Links agregados		
Nº	**Tipo**	**Acción**
1	Entre páginas de graduados	Crear una Barra de Menú
2	Acceso a los últimos graduados	Crear link Directo
3	Acceso a Plan de Estudios	Crear link Directo
Links eliminados		
Nº	**Tipo**	**Acción**
1	Entre páginas del Menú	Agrupar Link
2	Página Final	Hacer promoción
3	Navegación	Decisión del Webmaster

La principal ventaja de usar AG es que al tener una buena codificación, se explora un espacio enorme de soluciones a través de una búsqueda paralela de la mejor de ellas. Por otra parte, los AG siguen un comportamiento probabilístico lo que permite probar distintas adyacencias posibles. Si bien el modelo 2 al ser lineal, puede ser rápidamente resuelto por otros métodos de optimización como Simplex[110], el uso de AG permitió obtener una excelente aproximación, abriendo posibilidad a la aplicación de otros algoritmos más eficientes de acuerdo a la experimentación realizada.

Así, se puede disponer de los modelos expuestos para la mejora de la estructura de navegación de sitios web, lo que impacta tanto en ámbitos de investigación (nuevos conceptos y modelos de representación) como empresariales. En este último aspecto, el incremento de los sitios ha sido inminente sobre todo aquellos corporativos. Este punto es clave entonces para los negocios, pues el sitio pasa a ser la imagen visible por todo el mundo que cada vez recibe mayores herramientas para estar siempre conectado y aumenta el interés ofrecer la mejor experiencia de navegación. Por ello, tener un portal que ofrezca una cantidad equilibrada de hyperlinks a fin de no perderse en el sitio y encontrar la información deseada pasa a ser una obligación, pues con ello se logra que el usuario vuelva al sitio, recomiende y con mayor probabilidad, se haga cliente.

En definitiva, considerando el supuesto de que a mayor tiempo de permanencia en una página de la misma sesión, mayor interés en ella, y que la frecuencia de uso de las interrelaciones indica una mayor utilidad de las mismas para acceder hacia la información

[110] Simplex: Técnica para dar soluciones numéricas a problemas de programación lineal http://www.math.cuhk.edu.hk/~wei/lpch3.pdf

deseada, es que se logró encontrar estructuras de navegación más eficientes desde el punto de la navegabilidad y usabilidad, mejorando en concreto la estructura del sitio http://www.dii.uchile.cl/~mgo2007/, gracias al uso de AG. Con ello, las mejoras se pueden expandir a muchos otros sitios a fin de impactar positivamente en la experiencia de navegación de los usuarios en post de fortalecer y/o hacer crecer el negocio.

6.1.- Trabajo futuro

Cabe señalar que el presente trabajo es experimental, con el objetivo de estudiar las posibilidades existentes y nuevos mecanismos de cálculo, en base a los conceptos de web structure mining y web usage mining.

El trabajo adicional como consecuencia del aquí expuesto debe enfocarse en primera instancia a la implementación concreta de los cambios en la estructura web, estudiando el uso real de los usuarios en un considerable período de tiempo, con porcentajes de mejora, frecuencias de acceso, etc. De relevante interés sería el poder introducir un modelo de comportamiento del usuario en la Web, el cual otorgue registros de sesiones a fin de probar las mejoras y cambios ante la nueva estructura simulando la implementación real. Nuevas restricciones pueden ser analizadas e incorporadas al modelo, como por ejemplo fijar el número de enlaces entrantes o salientes desde páginas de la barra de menú. También se podría restringir la cantidad de links hacia o desde la página Home, o bien, no considerarla en el estudio, dado que en general se tiende a eliminar la mayoría de los accesos a ella. Caso similar al anterior sucede en el caso de la página contacto, donde se podría restringir a que no hayan modificaciones en los hyperlinks que posee.

El modelo propuesto puede ser mejorado en términos de la función de utilidad al agregar contenidos y principios del web content mining. Así, se podrían crear vectores de pesos de las palabras más usadas, desprendiendo los conceptos principales y determinando como se interrelacionan los mismos. Por otro lado, podrían incorporarse métodos que permitan realizar un estudio más dinámico, de modo de alcanzar una configuración del sitio web del tipo adaptativa, personalizada y que contenga sugerencias mientras se navega.

Aplicar otros algoritmos de optimización en los modelos sería de gran utilidad comparativa, a fin de encontrar el de mayor rapidez de ejecución y mejores resultados.

Finalmente, el crear una búsqueda automática de los patrones de cambios propuestos por la estructura resultante sería de gran interés, a modo de poder estudiar grandes sitios corporativos.

Apéndices

A.- Teorema del esquema de los algoritmos genéticos

Para comprender a cabalidad el principio fundamental y esencia de los AG, es necesario enunciar la teoría que sustenta el funcionamiento y convergencia de este procedimiento genético. John Holland desarrolló en su libro [35] *La teoría del esquema.* Se le llamará esquema[111] a aquel patrón de similitud entre las cadenas de representación de los individuos dado por los valores en cada posición, donde el término * indica o bien 1 o 0, en forma indiferente, es decir, se tiene un alfabeto de solo 3 indicadores {0,1,*} para armar las soluciones potenciales. Por ejemplo, ¿Qué hay de común en las siguientes representaciones: 11101, 10101, 10100? Es posible ver que en algunas posiciones los tres casos tienen el mismo valor los cuales se destacan a continuación: **1**1**10**1, **1**0**10**1, **1**0**10**0. Entonces, el esquema que subyace estas tres expresiones podría expresarse del siguiente modo: **1*10*.**

Otra forma de mostrar la utilidad, viene dada por la adaptación de cada individuo, en otras palabras, su función fitness, por ejemplo:

Tabla A. 1: Representación binaria y mejores esquemas

Fitness	Representación
158	10011110
50	00110100
37	00100101
89	01011001
123	01111011

Fuente: Elaboración personal

En la Figura A.1, se indica la representación y su Fitness respectivo, dado en este caso simplemente por la codificación binaria de cada número. Se puede ver en términos comparativos, que los mejores esquemas (Fitness más altos) tienen valores comunes en ciertas posiciones, al igual que los valores con peores Fitness, lo que brinda un acercamiento previo a los mejores y peores esquemas de esta población. En cuanto a los valores de la tabla, el comenzar con 00 indica, en términos generales, un bajo fitness, mientras que por el contrario, el responder al esquema ***11*** posee mejores valores, e incluso ***11*1* es un esquema comparativamente mejor en este ejemplo en particular. Ahora, si éste último esquema es sustancialmente superior al Fitness medio de otros individuos del mismo tipo (largo 8) que conforman el espacio de búsqueda total en generaciones sucesivas, será un buen esquema, determinante para la buena convergencia del algoritmo.

Así entonces, se puede hablar de similitud entre cadenas gracias a los esquemas. Mayores referencias en [27][35][45][65].

[111] Definiciones y teoría aplicada en representaciones binarias

A.1.- Definiciones previas

Para comprender la magnitud de información contenida en los esquemas posibles en toda la población (soluciones potenciales) y la teoría base que sustenta el funcionamiento y convergencia de los AG, se plantean a continuación ciertos cálculos y definiciones relevantes.

❖ Cantidad de esquemas posibles

Todo esquema tiene el mismo largo fijo L de los integrantes de la población. Por ello, y considerando que 3 son los valores posibles para armar un esquema (1,0,*) existirán 3^L esquemas posibles. Ahora, teniendo ya armada una cadena de largo L, ésta tendrá 2^L esquemas que la representan (dado que cada alelo en concreto es 0 o 1). Ante eso al tener una población de M individuos, el número de esquemas representados está entre el mínimo y máximo posible, es decir entre 2^L y $M \bullet 2^L$

❖ Orden de un esquema

Dado un esquema S (Scheme) su orden $\vartheta(S)$ se define como el número de posiciones fijas, es decir aquellas posiciones distintas de *. Por ejemplo en los esquemas ya enunciados anteriormente y uno adicional de ejemplo se tendrían valores como los indicados en la Tabla A.2.

Tabla A. 2: Orden de un esquema

Scheme	Orden
11	2
***11*1*	3
001101*1	7
0*******	1

Fuente: Elaboración personal

❖ Longitud de un esquema

Dado un esquema S (Scheme) su longitud $\delta(S)$ se define como el número de posiciones entre la primera y última distintas de *. En otras palabras, es contar la cantidad de bits entre la primera posición que no sea *, y avanzando en la cadena, la última diferente de *. Se ejemplifica dicho cálculo en la Tabla A.3.

Tabla A. 3: Longitud de un esquema

Scheme	Cálculo	Longitud
11	5 - 4 =	1
***11*1*	7 - 4 =	3
001101*1	8 - 1 =	7
0*******	1 - 1 =	0

Fuente: Elaboración personal

A.2.- Desarrollo de la teoría

·El Teorema del esquema relaciona la calidad de los miembros de un esquema en una generación con el número esperado de miembros en la siguiente generación" [35]. La idea es aplicar el efecto de los operadores genéticos en la supervivencia de un esquema en las iteraciones posteriores y por ende su valor final, lo que permite deducir aquel que será representativo de las mejores soluciones. Para enunciar el teorema en consecuencia, es necesario ir por el efecto de cada operador en los esquemas de representación.

- **Efecto selección:**

Cabe señalar que la selección actúa sobre el conjunto de la población actual, es decir, realiza copias de los individuos de la misma sin comparar con otros individuos que no pertenezcan a ella. La selección realiza un estudio interno en cada población, donde aquellos individuos de mejor fitness tendrán más posibilidades de ser escogidos.

Para comenzar, se tiene en el instante *t* (llámese también generación) una población *P(t)* de tamaño *N*, en ella, se actúa bajo el supuesto que existen m individuos pertenecientes a un esquema S, denotado como *m=m(S,t).* Durante la selección, la probabilidad de que un individuo i sea seleccionado viene dada por:

$$p_i^e = \frac{f_i}{\sum_{j=1}^{M} f_j} \qquad \text{(A.1)}$$

Donde:

p_i^e	=	Probabilidad de selección de individuo *i*
f_i	=	Función fitness del individuo *i*
$\sum_{j=1}^{M} f_j$	=	Suma de los fitness de todos los individuos de *P(t)*

Ahora, el número estimado de individuos del esquema *S* que serán escogidos en *N* selecciones (tamaño de la población) en la Población *P(t)* queda expresado en la ecuación A.2

$$m(S,t+1) = m(S,t) \cdot N \cdot \frac{f(S)}{\sum_{j=1}^{N} f_j} \qquad \text{(A.2)}$$

Donde:

$f(S)$ = Función fitness medio de los individuos del tipo *m(S,t)*

$$f(S) = \frac{f_i}{\sum_{k=1}^{K} f_k} \qquad \text{(A.3)}$$

Siendo *K* el conjunto de individuos que pertenecen al esquema *S* en la generación *t*.

Considerando ahora $\bar{f}$ como el fitness medio de la totalidad de la población:

$$\bar{f} = \frac{\sum_{j=1}^{N} f_j}{N} \qquad \text{(A.4)}$$

la fórmula (A.2) puede ser reescrita de la forma:

$$m(S,t+1) = m(S,t) \cdot \frac{f(S)}{\bar{f}} \qquad \text{(A.5)}$$

Gracias a (A.5), es posible establecer inicialmente, que el crecimiento del esquema viene dado por el cociente entre el fitness medio del esquema y éste a su vez de la población. Con esto se pueden establecer la primera conclusión en la predicción del número de individuos de un esquema entre generaciones:

- ✓ Un esquema con fitness medio mayor que el fitness medio de la población total, tendrán un mayor número de individuos pertenecientes a él, entre cada iteración de *t* a *t+1*. De hecho el crecimiento tiene características exponenciales.
- ✓ Un esquema con fitness medio menor que el fitness medio de la población total, tendrán un menor número de individuos, tendiendo a su desaparición.

Suponiendo ahora, que el fitness medio del esquema es mayor al fitness medio total de la población en un factor *c* del modo:

$$f(S) = \bar{f} + c \cdot \bar{f} \qquad \text{(A.6)}$$

Así, la fórmula (A.5) queda:

$$m(S,t+1) = m(S,t) \cdot \frac{\bar{f} + c \cdot \bar{f}}{\bar{f}} = m(S,t) \cdot (1+c) \qquad \text{(A.7)}$$

Siendo c un valor estacionario a través de las generaciones en *t* queda:

$$m(S,t) = m(S,0) \cdot (1+c)^t \qquad \text{(A.8)}$$

Cabe destacar, que los términos *S, m, f, i,* etc., son expresiones generales, dado que los comportamientos mencionados se llevan a cabo en todo el conjunto de esquemas existentes en la población en forma paralela.

- ❖ **Efecto cruce:**

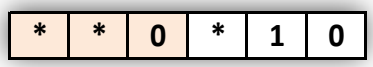
Tomando los individuos seleccionados previamente en la etapa anterior se procede con el cruce para formar la nueva generación. Para comprender mejor el efecto del cruce en la

supervivencia de un esquema, mejor será comenzar con un ejemplo. Suponiendo que se tiene un individuo del tipo 110010, dos esquemas posibles que lo representan vienen dados por *E1 = **0*10* y *E2 = 11*****.

Considerando el cruce en un punto, se analiza en detalle el caso de supervivencia del esquema E1 = **0*10, donde se precisa los efectos de cada uno de las 5 posibilidades de corte total:

Corte 1 (Entre posiciones 1-2)
Este tipo de corte permite la supervivencia del esquema, dado que mantiene en una de sus dos partes (zona en blanco de la derecha) sus valores fijos. Es decir, al menos uno de sus dos hijos pertenecerá a E1.

Corte 2 (Entre posiciones 2-3)
Al igual que en el caso anterior, al menos uno de los dos hijos mantiene los valores fijos por lo que el esquema se mantiene.



Corte 3 (Entre posiciones 3-4)
Esquema se destruye con el cruce, pues se ven separadas sus partes fijas, con lo que un hijo (zona izquierda de corte) tiene el 0, pero no necesariamente el 1 ni 0 en las posiciones 5 y 6 respectivamente. Por otro lado, al lado derecho del cruce se le puede agregar una parte sin el 0 en tercera posición por lo que igualmente se pierde el esquema.

Corte 4 (Entre posiciones 4-5)
Esquema se destruye, análogo a caso anterior.

Corte 5 (Entre posiciones 5-6)
Esquema se destruye, análogo a caso anterior. Ahora teniendo la zona izquierda solo el 0 y 1 fijos en 3° y 5° posición respectivamente, sin contar necesariamente con el 0 en la sexta posición.

En el caso anterior, se pudo apreciar que de 5 posibilidades de corte, el esquema sobrevive en 2, o bien, se destruye en 3 de las 5 ocasiones posibles. Algo indica la intuición en cuanto a la ubicación de los términos fijos.

En cuanto a E2 = 11****:

Corte 1 (Entre posiciones 1-2)
Esquema se destruye, al separar sus posiciones de valores fijos puede ser que cada zona se una a un complemento sin el valor 1 donde se requiere para mantener el esquema.

1	1	*	*	*	*

Corte 2 (Entre posiciones 2-3)
Esquema sobrevive dado que al menos un hijo cuenta ineludiblemente con las características del padre, pues la zona izquierda mantiene las posiciones fijas en las posiciones 1° y 2°. Lo mismo sucede para los cruces 3, 4 y 5.

En el caso de E2, de 5 posibilidades de corte, el esquema sobrevive en 4 ocasiones, o bien, se destruye en 1 de los 5 puntos de corte posibles.

Así entonces, claramente la cercanía o distancia entre los puntos fijos dice bastante al respecto de su sobrevivencia. Notar que mientras δ (E1) =3 y δ (E2) =1, *E1* se destruye en 3 de los 5 puntos de corte y *E2* en 1 de los 5. Ante eso, es posible enunciar la probabilidad de pérdida de un esquema *S* de largo *L*, en general como:

$$p_p = \frac{\delta(S)}{L-1} \qquad \text{(A.9)}$$

Considerando además la probabilidad de cruce p_c y la negación de la probabilidad de pérdida, entonces la probabilidad de supervivencia p_a queda como:

$$p_a \geq 1 - p_c \cdot \frac{\delta(S)}{L-1} \qquad \text{(A.10)}$$

Acumulando a los efectos de selección expresados por (7.5), el efecto de cruce dado por (7.10) al considerar que los efectos son claramente independientes se tiene que la cantidad de individuos del esquema en la generación siguiente viene dado por:

$$m(S,t+1) \geq m(S,t) \cdot \frac{f(S)}{\bar{f}} \cdot \left(1 - p_c \cdot \frac{\delta(S)}{L-1}\right) \qquad \text{(A.11)}$$

Con esto se da paso a la siguiente conclusión:

- ✓ El crecimiento del número de individuos *m* del esquema *S* depende tanto del fitness medio del esquema tanto de la longitud, donde a mayor longitud (asumiendo los otros términos constantes), menor es la probabilidad de supervivencia por lo que *m(S, t+1)* se ve disminuido.

❖ **Efecto mutación:**

Para introducir el efecto de mutación, se necesita establecer una probabilidad de mutación p_m que por lo general es bastante pequeña.

La mutación produce el cambio de un bit de datos de la representación del individuo, por lo que pensando a nivel de esquema, éste no será afectado por la mutación si algunas de

sus posiciones en * es modificada, mientras que por el contrario, si un valor fijo se ve mutado, el esquema no sobrevivirá a la mutación.

Dado que el orden $\vartheta(S)$ de un esquema indica el número de posiciones fijas, es claro establecer que mientras que ninguno de esos términos se vea mutado, el esquema continuará, en términos de ecuación, la probabilidad de supervivencia a la mutación viene dada por la ecuación A.12.

$$(1-p_m)^{\vartheta(S)} \qquad \text{(A.12)}$$

Es decir, la probabilidad de que ninguno de los términos fijo mute, donde la probabilidad de que una posición estable no lo haga es $(1-p_m)$. Dado que p_m como se indicaba suele ser un término muy pequeño por aproximación de pequeñas oscilaciones, la expresión (12) puede expresarse como lo indicado en la ecuación A.13.

$$1-\vartheta(S)\cdot p_m \qquad \text{(A.13)}$$

Finalmente entonces, el número de individuos esperados que pertenezcan al esquema m en la generación siguiente, al pasar por los operadores genéticos viene dado por la unión de las expresiones claves en (A.11) y (A.13)

$$m(S,t+1) \geq m(S,t)\cdot\frac{f(S)}{\bar{f}}\cdot\left(1-p_c\cdot\frac{\delta(S)}{L-1}-\vartheta(S)\cdot p_m\right) \qquad \text{(A.14)}$$

A estas alturas, es posible gracias a la expresión (A.14) contar con la conclusión final del teorema fundamental de los algoritmos genéticos:

- ✓ El crecimiento del número de individuos *m* del esquema *S* depende del fitness medio del esquema, su longitud y su orden. Mientras el esquema tenga un mayor $f(S)$ en relación a $\bar{f}$, tenga un pequeña longitud y tenga un bajo orden, entones dicho esquema aumentará su presencia de manera exponencial en las generaciones que suceden.

B.- Algoritmos genéticos en Matlab

Mención especial se hace de la versión 2.4.1 de las herramientas de algoritmos genéticos y búsqueda directa de Matlab [71]. Aquí, se dispone de herramientas para el uso de algoritmos genéticos, simulated annealing, y la búsqueda directa. Estos algoritmos se ocupan en problemas difíciles de resolver con las tradicionales técnicas de optimización, incluyendo los problemas que no están bien definidas o son difíciles de modelar matemáticamente. Son herramientas complementarias a otros métodos de optimización para ayudar a encontrar buenos puntos de partida, donde incluso es posible utilizar técnicas tradicionales para refinar su solución.

Las funciones de la caja de herramientas, son accesibles a través de una interfaz gráfica de usuario (GUI) o la línea de comandos Matlab, escritas en el lenguaje abierto M. Esto significa que es posible inspeccionar los algoritmos, modificar el código fuente, y crear propias funciones personalizadas. Los AG implementados en Matlab permiten llevar a cabo los siguientes operadores indicados en la Tabla B.1.

Tabla B. 1: Opciones de AG en Matlab

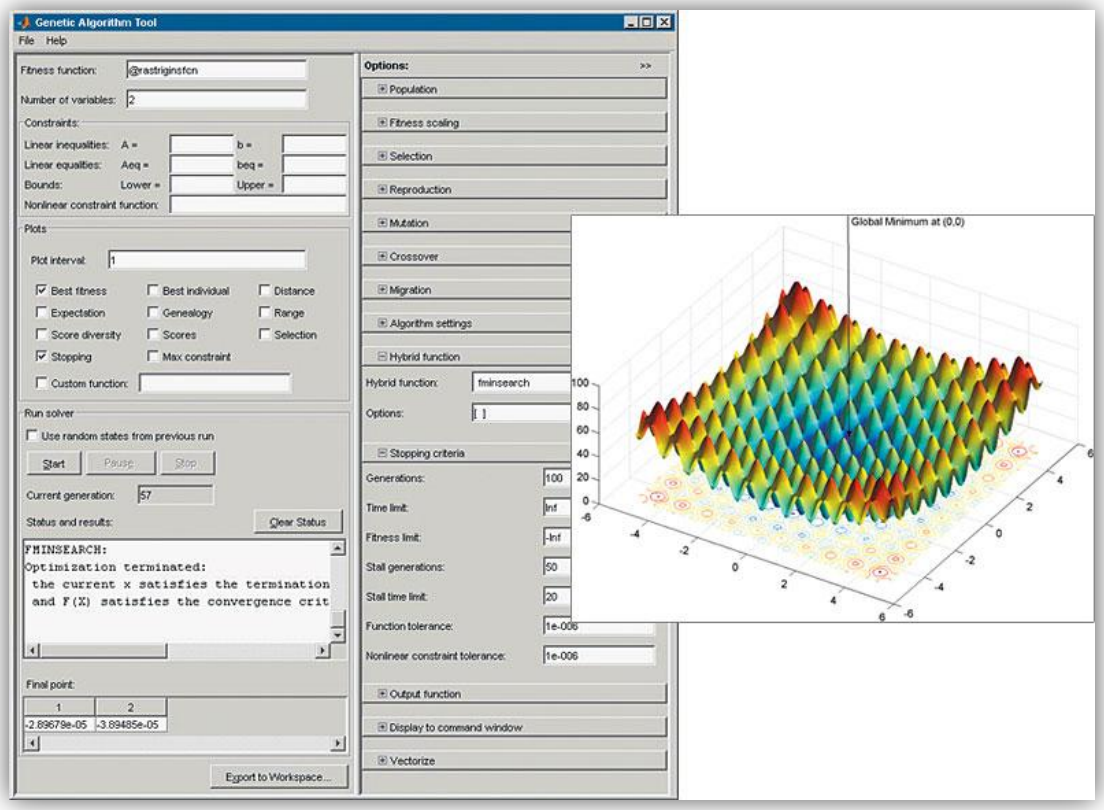
Paso	Opciones
Creación	Uniforme
Fitness	Basada en el ranking, Proporcional, Truncada, Escala lineal, Turno
Selección	Ruleta, Estocástica Uniforme, Torneo, Uniforme
Cruce	Aritmético, Heurístico, Intermedio, Escalado, Un punto, Dos puntos
Mutación	Adaptación posible, Gaussiana, Uniforme
Gráfico	Mejor Fitness, mejor individuo, distancia entre los individuos, expectativas de los individuos, rango, diversidad de la población, el índice de selección, las condiciones de parada

Fuente: [71]

Las herramientas brindadas también le permite precisar: restricciones, tamaño de la población, número de descendencia, probabilidad de cruce y algún proceso de migración entre las subpoblaciones. Proporcionando funciones definidas, es posible personalizar estas opciones del algoritmo y representar el problema en una amplia variedad de formatos de datos. Es posible además fijar los criterios de detención, como ser el tiempo de ejecución del algoritmo, fitness límite, número fijo de generaciones, etc. Por último, es posible vectorizar la función de evaluación para mejorar la velocidad de ejecución.

La interfaz que ofrece Matlab, cuya imagen corresponde a la Figura 59, permite definir el problema en un set de variadas opciones, gestionar la optimización, monitorear el desempeño, cambiar opciones, definir criterio de detención, personalizar barra de herramientas para la creación de algoritmo propio, definir los operadores a utilizar, combinar algoritmos y visualizar gráficas de la optimización como valor de la función, histogramas de calificación, valor del fitness, genealogía, etc.

Tabla B. 2: Interfaz AG en Matlab



Fuente: [71]

C.- Conceptos utilizados de teoría de grafos y cadenas de Markov

En el siguiente apéndice, se exponen términos ocupados en las secciones anteriores en relación a teoría de grafos y cadenas de Markov. Se tratan los conceptos relevantes de dichas temáticas, sin hacer un completo marco conceptual de ellas.

Teoría de grafos: Conexidad

Dado un sitio web *W*, el grafo web de *W* es un grafo dirigido de la forma *G=(N,E)* donde N es el set de nodos que representan todas las páginas en W que pueden ser alcanzadas desde la página Home y $E \subseteq NxN$, es el set de arcos que representan los hyperlinks entre los nodos [66]. Los grafos, cuentan con muchas propiedades y términos propios.

Un grafo será completo cuando todas sus aristas conectan cada par de vértices. El grafo completo de n nodos, tendrá n vértices y $n(n-1)/2$ aristas. Por otro lado, se dice que un grafo es conexo, cuando para cualquier página *i* y *j*, existe al menos una trayectoria posible entre ellas, es decir, existe una sucesión de páginas adyacentes sin repetir los nodos visitados que permitan acceder de *i* a *j*. En otras palabras, en un grafo conexo, será posible acceder a todas las páginas.

Diversas metodologías se han desarrollado para determinar la conexidad de un grafo, destacadas son las correspondientes a búsqueda en anchura y búsqueda en profundidad[112]. La primera se enfoca en recorrer los elementos de un grafo comenzando de un nodo raíz y explorando los vecinos del mismo, para luego desde los nodos vecinos explorar los vecinos adyacentes y así sucesivamente. Profundidad en cambio, va recorriendo todos y cada uno de los nodos que va ubicando de forma recurrente formando el camino completo de manera ordenada, así, cuando llegue a un tope donde no tenga más nodo que visitar, repite el procedimiento desde otro nodo. Una nueva metodología, útil en tiempo computacional y base a matrices, factores relevantes de acuerdo a la construcción de los modelos propuestos, corresponde a la construcción de la matriz de Laplace *L*.[113], también llamada matriz de admisión o matriz de Kirchhoff. Es una representación matricial del grafo, que permite determinar la cantidad de subgrafos existentes. Construyendo la matriz de Laplace, el número de veces en que el valor cero aparece, indica el número de componentes conexos en el grafo, cantidad también representada por los vectores propios de *L*, es decir, si solo resulta un vector propio, el grafo es completamente conexo entre sí, sin subgrafos.

A continuación se detallan las etapas para la construcción de la matriz de Laplace en Matlab, gracias a las funciones que el mismo programa ofrece de modo de generar la restricción de conexidad que se requiere en los modelos propuestos, donde, si se obtiene un vector propio se permite la matriz, sino, se penaliza su existencia. Se indica luego de "//" la codificación es específico llevada a la programación en Matlbab.

[112] **Fuente:** http://www.dma.fi.upm.es/java/matematicadiscreta/busqueda/
[113] **Fuente:** http://mathworld.wolfram.com/LaplacianMatrix.html

0.- Sea Adj la matrix de adjacencia inicial
1.- Obtener matriz simétrica Asim // Asim= or (Adj, Adj')
2.- Extraer matriz diagonal MD // MD= diag(Asim)
3.- Crear una matriz con los elementos de la diagonal // MMD=diag(MD)
4.- Restarle a matriz simétrica, elementos de la diagonal obteniendo Asim2
// Asim2= Asim - MMD
5.- Calcular vector de grado G (suma de filas) // G=sum(Asim2')
6.- Crear matriz diagonal con los grados obtenidos // MG = diag(G)
7.- Obtener matriz laplaciana restando matriz de grado con la adyacencia sin diagonal
// LL= MG-Asim2
8.- Cálculo de vectores propios // V=null(LL)
9.- Cálculo de cantidad de de vectores // C=size(V,2),

Cadenas de Markov: Probabilidades estacionarias [59]

Una cadena de Markov es un proceso estocástico (proceso que evoluciona de manera no determinística a lo largo del tiempo) en la cual la probabilidad de que ocurra un evento depende del evento inmediato anterior, es decir, tiene memoria condicionando las posibilidades de los eventos futuros. Sus conceptos claves tienen que ver con los estados del sistema, definición de transición y una ley de probabilidad condicional que una la probabilidad del nuevo estado en relación a los anteriores. Estado, corresponde a la situación, ya sea cualitativa o cuantitativa del proceso en un momento determinado. **Probabilidad de transición** p_{ij} corresponde a la probabilidad de que el proceso evolucione a *j* en la siguiente transición estando en *i*, y cumplirán la siguiente propiedad (ley de probabilidad condicional):

$$\sum_{j=1}^{n} p_{ij} = 1 \text{ con } p_{ij} \geq 0 \qquad \text{(C.1)}$$

Así, a *P* se le llamará la matriz de transición, cuya suma de valores fila será siempre igual a 1.

Cuando las probabilidades de transición cumplen la probabilidad indicada en la ecuación C.1, se dirá que *P* es una matriz estocástica donde la suma de las filas será siempre igual a 1.

Considerando que las probabilidades de transición son estables en el tiempo, de interés será el poder conocer las probabilidades de transición después de *k* pasos. Es decir, la probabilidad de que el proceso se encuentre en el estado *j* en *k* etapas, dado que antes se encontraba en *i.*

$$p\{E_{t+k} = j \mid E_t = i\} = p\{E_k = j \mid E_0 = i\} = p_{ij}^{(k)} \qquad \text{(C.2)}$$

Visitando un estado intermedio *e* se tiene:

$$p_{ij}^{(k)} = \sum_{e=1}^{n} p_{ie}^{(m)} \cdot p_{ej}^{(k-m)} \qquad \text{(C.3)}$$

Así, es posible obtener las matrices en k pasos a partir de las potencias de la matriz *P*.

$$P^{(k)} = P^{(k-1)} \cdot P = P \cdot P^{k-1} = P^{k-1} \cdot P = P^{k} \qquad \text{(C.4)}$$

Por otro lado, una cadena de Markov [13] se vuelve "estacionaria" cuando el resultado de las ecuaciones deja de ser una probabilidad incierta, es decir, que a partir de este momento se conoce con certeza su resultado. Dichas probabilidades se denominan, probabilidades estacionarias π_{ij} , asociadas al hecho de que el proceso de encuentre en el estado *j*, después de un número elevado de transiciones (alcanzando el régimen permanente), si el sistema comenzó su evolución en el estado *i*.

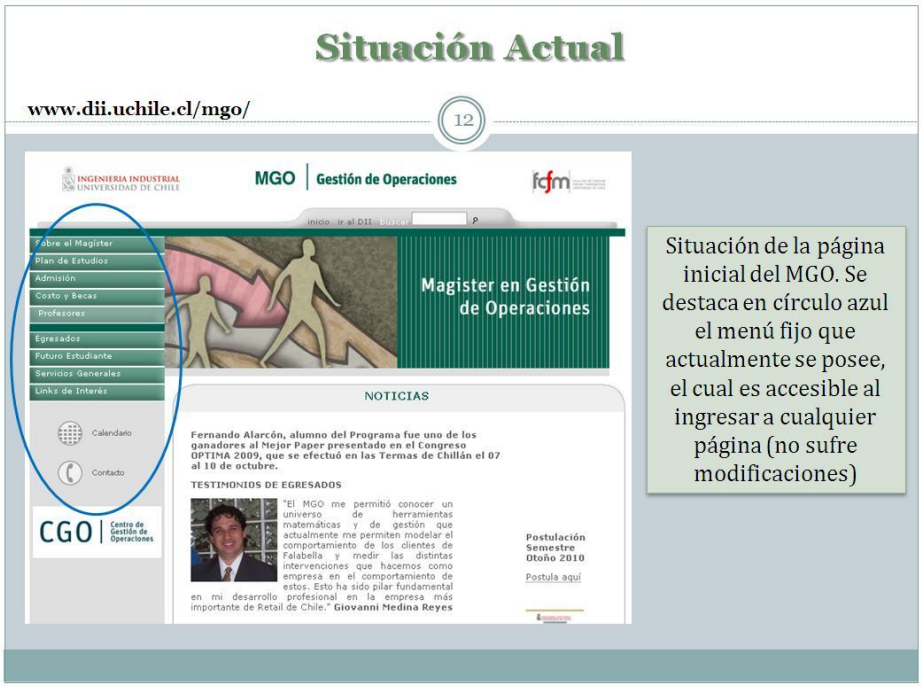
D.- Base encuesta

Se muestra a continuación la encuesta realizada con las imágenes asociadas a cada caso, de modo de hacer explicativo cada cambio en cuestión, donde se le pedía a los usuarios evaluar los cambios observados.

Cambio 1: Sub menú

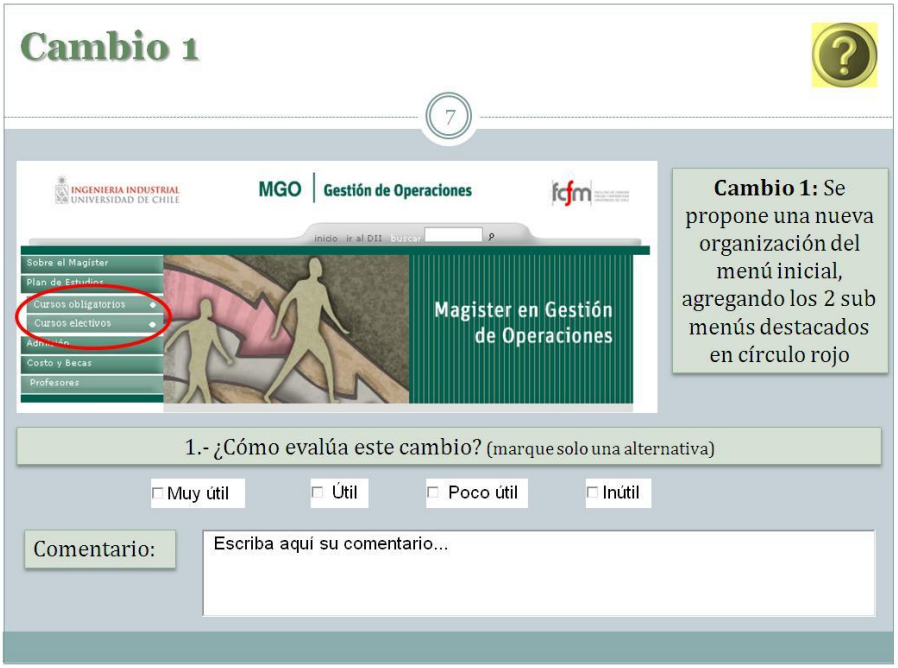
En la Figura 60, se visualiza la página Home actual en el sitio.

Figura D. 1: Encuesta, situación actual



Luego, se muestra al cambio 1 propuesto, tal como se observa en la Figura 61.

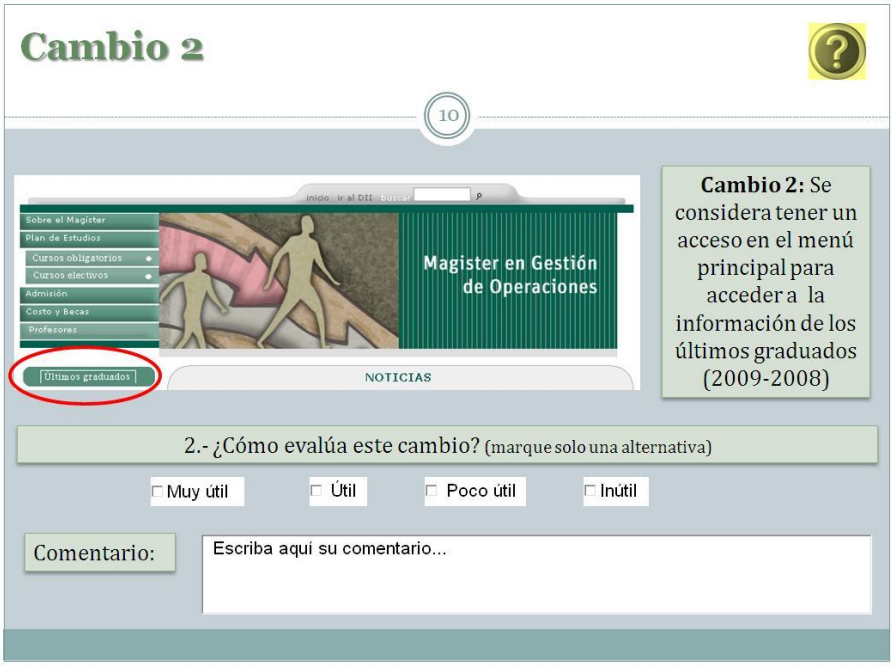
Figura D. 2: Encuesta, cambio 1



Cambio 2: Últimos graduados

Considerando la misma situación actual de la Figura 60, se propone el cambio 2 de la Figura 62.

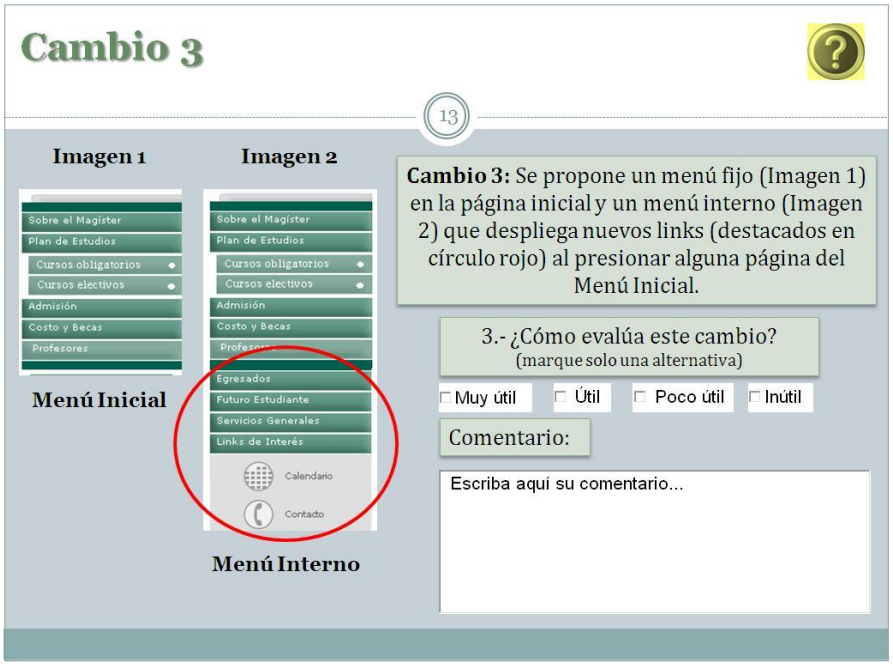
Figura D. 3: Encuesta, cambio 2



Cambio 3: Reorganización

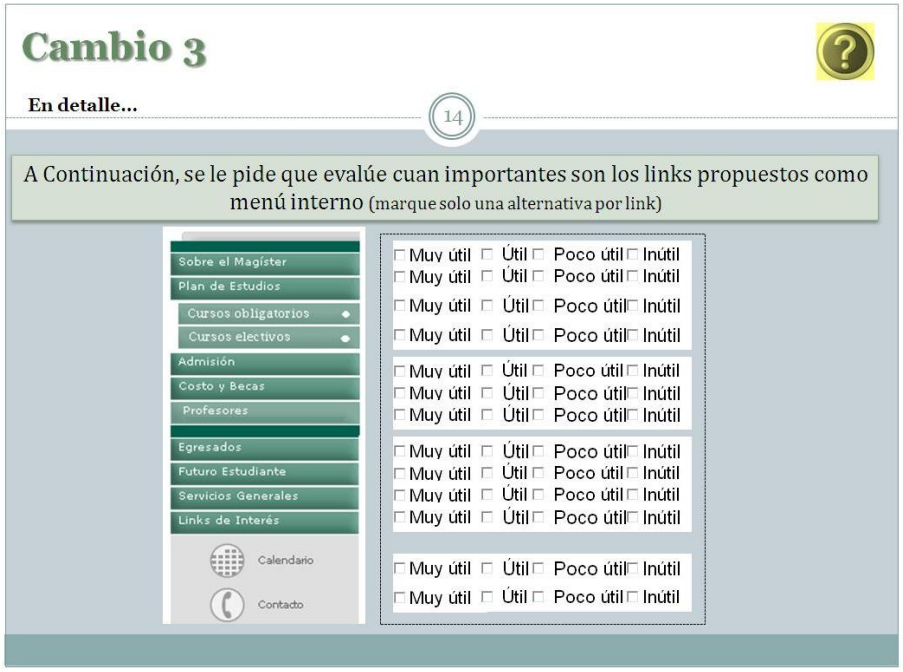
Con énfasis en la disposición y hyperlinks considerados por el menú principal, que se observa en la Figura 60 rodeado de un óvalo, se propone un nuevo cambio observado en la Figura 63.

Figura D. 4: Encuesta, cambio 3



Cabe señalar, que en la imagen 1 de la Figura 63 también contiene el hyperlink Futuro Estudiante, pero se analiza de manera independiente mediante una cuantificación por hyperlink, es decir, como apoyo al modelamiento anterior, se solicita una calificación de cada hyperlink del menú interno (Figura 64)

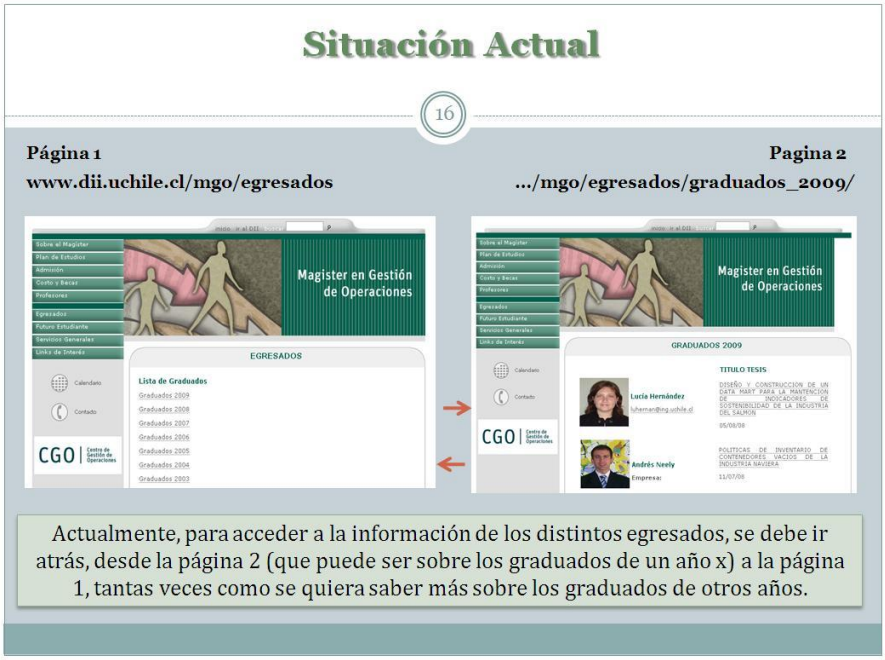
Figura D. 5: Encuesta, evaluación de hyperlinks cambio 3



Cambio 4: Entre egresados

Dado que ahora se estudian hyperlinks no propios de la página de inicio, se explica la situación actual de acceso entre páginas de egresados en la Figura 65.

Figura D. 6: Encuesta, situación actual entre páginas egresados



Explicada la situación actual, se indica el cambio propuesto (Figura 66)

Figura D. 7: Encuesta, cambio 4

Cambio 5: Promoción

Finalmente se indica el último cambio establecido, tal como se observa en la Figura 67.

Figura D.8: Encuesta, cambio 5

E.- Tablas resultados

Se expone a continuación, para mayores detalles, los procedimientos y resultados concretos obtenidos, a nivel de programación y estructuras de datos.

E.1. Consultas sql

Las consultas realizadas se separaron de acuerdo a objetivos, es decir, crear tablas intermedias que permitieran calcular los datos deseados. Se podría haber hecho una consulta sin tablas intermedias por cada caso, pero para efectos de investigación se decidió ir por tablas de modo de poder explorar mejor con los datos antes de la construcción definitiva de los modelos. Para efectos de explicación además, se hace mucho claro de esta manera. Cabe señalar previamente, que todas las consultas de hacen en general a la tabla del total de registros web (path) sin considerar páginas que no pertenezcan al mgo (*http://www.dii.uchile.cl/~mba/*), archivos que no son páginas (*/mgo2007/style/tinymce/jscripts/tiny_mce/plugins/table/row.htm*) ni páginas de administración (*/mgo2007/adminsoda.php*).

Tabla Page
Identificación de páginas solo del mgo

```
(SELECT * from page
WHERE path like "/mgo%")
```

//Tabla Page era dato, proporcionado por los web log.

Tabla Path_std
Selección de factor tiempo (factor_time)

```
SELECT p.visit_time/t1.n  factor_time
FROM path_std p, (
        SELECT id_session, sum(visit_time) n
        FROM path_std
        GROUP BY id_session) t1
WHERE t1.id_session=p.id_session
```

//Tabla Path era dato, proporcionado por los web log. Path_std es Path sin considerar las sesiones de largo1.

Tabla Page
Selección de factor tiempo promedio (AVG(factor_time))

```
SELECT AVG(factor_time)
FROM path_std p
GROUP BY Page_idPage
```

Tabla Links
Selección de probabilidad condicional (Prob_cond)

La probabilidad condicional se obtiene con el cociente entre el total de transiciones entre pares de páginas y la salida total desde la página de inicio.

```
SELECT A.transiciones/B.transpp
FROM ttrans_pp as B, ttrans_links as A
WHERE B.page=A.page_ini
```

Consulta ttrans_links: obtiene todas las transiciones entre las páginas i y j.

```
(SELECT id_pageini, id_pagefin, count( * ) transiciones
FROM
        (SELECT p.Page_idPage id_pageini, pp.Page_idPage id_pagefin, p.visit_time time
        FROM path_std p, path_std pp
        WHERE p.id_session=pp.id_session and p.id_order+1=pp.id_order
        ) t1
GROUP BY id_pageini, id_pagefin
) ttrans_links
```

Consulta ttrans_pp: Obtiene la cantidad de transiciones por página (cantidad de veces que desde una página i se pasa a cualquier otra página)

```
(SELECT t2.page_ini, count(*) transpp
FROM (
        SELECT p.Page_idPage id_pageini, pp.Page_idPage id_pagefin, p.visit_time time
        FROM path_std p, path_std pp
        WHERE p.id_session=pp.id_session and p.id_order+1=pp.id_order
        ) t1
GROUP BY id_pageini
) ttrans_pp
```

Tabla Links
Selección de frecuencias PN, PF y PI

Valor de PN (se actualiza la tabla Links)

```
UPDATE Caminos_mgo A, Caminos_mgo_aux B
SET   A.PN = A.Total_pasadas/B.Sum_pasadas
WHERE A.Page1 = B.Page1
```

Valor de PF (se actualiza la tabla Links)
```
UPDATE Caminos_mgoA, Caminos_mgo_aux B
SET   A.PF = A.Total_factor/B.Sum_factor
WHERE A.Page1 = B.Page1
```

Valor de PI (se actualiza la tabla Links)
```
UPDATE Caminos_mgo A
SET   A.PI = A.PF*A.PN
```

Caminos: obtiene todas las pasadas y tiempo entre las páginas i y j, identificando su largo
```
(SELECT pag1, pag2, L, count( * ) TOTAL_SESSIONES, sum( N )
TOTAL_PASADAS, sum( FP ) TOTAL_FACTOR
FROM
        (SELECT p.id_session s, p.id_page pag1, pp.id_page pag2, count( * ) N, sum(pp.factor_session ) FP
        FROM path_std2 p, path_std2 pp
        WHERE p.id_session = pp.id_session
        AND p.id_order < pp.id_order
        GROUP BY pag1, pag2, s) T
GROUP BY pag1, pag2)
Caminos
```

Caminos_mgo: Caminos desde páginas solo del mgo
```
(SELECT `Page1`, `Page2`,`Largo`, `Total_sesiones`, `Total_pasadas`, `Total_factor`
FROM `Caminos`, mgo_pages_fin mgo1, mgo_pages_fin mgo2
WHERE Page1=mgo1.id_mgo_page and Page2=mgo2.id_mgo_page)
Caminos_mgo
```

Caminos_mgo_aux: Suma de caminos desde una página del mgo de largo1
```
(SELECT Page1, sum(Total_pasadas), sum(Total_factor) FF
FROM Caminos_mgo
GROUP BY Page1)
Caminos_mgo_aux
```

Tabla Links_steps
Selección de frecuencias PN_length, PF_length y PI_length por cada largo X, donde X $\in$ [1,14]

Valor de PN para largo X (se actualiza la tabla Links_step donde lenght =X)
```
UPDATE Caminos_mgoX A, Caminos_mgoX_aux B
SET   A.PN = A.Total_pasadas/B.Sum_pasadas
WHERE A.Page1 = B.Page1
```

Valor de PF para largo X (se actualiza la tabla Links_step donde lenght =X)
```
UPDATE Caminos_mgoX A, Caminos_mgoX_aux B
SET   A.PF = A.Total_factor/B.Sum_factor
WHERE A.Page1 = B.Page1
```

Valor de PI para largo X (se actualiza la tabla Links_step donde lenght =X)
```
UPDATE Caminos_mgoX A
SET   A.PI = A.PF*A.PN
```

Caminos: obtiene todas las pasadas y tiempo entre las páginas i y j, identificando su largo
```
(SELECT pag1, pag2, L, count( * ) TOTAL_SESSIONES, sum( N )
TOTAL_PASADAS, sum( FP ) TOTAL_FACTOR
FROM
        (SELECT p.id_session s, p.id_page pag1, pp.id_page pag2,
        (pp.id_order-p.id_order) L, count( * ) N, sum(pp.factor_session ) FP
        FROM path_std2 p, path_std2 pp
        WHERE p.id_session = pp.id_session
        AND p.id_order < pp.id_order
        GROUP BY pag1, pag2, s, L) T
GROUP BY pag1, pag2, L)
Caminos
```

Caminos_mgo: Caminos desde páginas solo del mgo
(SELECT `Page1`, `Page2`,`Largo`, `Total_sesiones`, `Total_pasadas`, `Total_factor`
FROM `Caminos`, mgo_pages_fin mgo1, mgo_pages_fin mgo2
WHERE Page1=mgo1.id_mgo_page and Page2=mgo2.id_mgo_page)
Caminos_mgo

Caminos_mgoX: Caminos entre páginas solo del mgo de largo X entre ellas
(SELECT `Page1`, `Page2`, `Largo`, `Total_sesiones`, `Total_pasadas`, `Total_factor`
FROM `Caminos_mgo` where Largo=X)
Caminos_mgoX

Caminos_mgoX_aux: Suma de caminos desde una página del mgo de largo1
(SELECT Page1, sum(Total_pasadas), sum(Total_factor) FF
FROM Caminos_mgoX
GROUP BY Page1)
Caminos_mgoX_aux

E.2. Código fitness en Matlab

Se exponen a continuación, los códigos Fitness de los distintos modelos, primero estableciendo los datos de entrada, y luego el Fitness en sí. Cabe destacar que la función objetivo es minimizada por dafault en Matlab, por lo que se pide minimizar –Fitness, es decir, se le antepone un signo negativo de modo que busque maximizar el total. Por eso mismo, es que la penalización es con un valor muy grande, en todos los casos será 1E10. Por otro lado, se señala que todas las importaciones desde Excel, se realizan para introducir los datos de entrada, y cada matriz de Excel, es obtenida vía consultas sql desde la base de datos, considerando las consultas anteriores.

Modelo 1, datos de entrada
Búsqueda de los caminos mínimos entre los hyperlinks no existentes

```
%Definiendo pesos
W = [
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1   1
];

%Definiendo grafo
DG = sparse([
1   1   1   1   1   1   1   1   1   1   1   1   2   2   2   2   2   2   2
2   2   2   2   2   2   2   3   3   3   3   3   3   3   3   3   3   3   3
4   4   4   4   4   4   4   4   4   4   4   4   5   5   5   5   5   5   5
5   5   5   5   5   6   6   6   6   6   6   6   6   6   6   6   6   6   6
6   7   7   7   7   7   7   7   7   7   7   7   7   7   7   8   8   8   8
8   8   8   8   8   8   8   8   9   9   9   9   9   9   9   9   9   9   9
9   9   9   9   9   9   9   9   9   9   10  10  10  10  10  10  10  10  10
10  10  10  11  11  11  11  11  11  11  11  11  11  11  11  12  12  12  12
12  12  12  12  12  12  12  12  13  13  13  13  13  13  13  13  13  13  13
13  14  14  14  14  14  14  14  14  14  14  14  14  15  15  15  15  15  15
15  15  15  15  15  15  16  16  16  16  16  16  16  16  16  16  16  16  17
17  17  17  17  17  17  17  17  17  17  17  18  18  18  18  18  18  18  18
18  18  18  18  19  19  19  19  19  19  19  19  19  19  19  19  20  20  20
20  20  20  20  20  20  20  20  20  21  21  21  21  21  21  21  21  21  21
21  21  22  22  22  22  22  22  22  22  22  22  22  22  23  23  23  23  23
23  23  23  23  23  23  23  23  24  24  24  24  24  24  24  24  24  24  24
24  25  25  25  25  25  25  25  25  25  25  25  25
], [
1   2   3   4   5   9   13  14  15  17  18  23  1   2   3   4   5   6   9
13  14  15  17  18  23  25  1   2   3   4   5   9   13  14  15  17  18  23
1   2   3   4   5   9   13  14  15  17  18  23  1   2   3   4   5   9   13
```

```
14  15  17  18  23  1   2   3   4   5   6   7   9   13  14  15  17  18  23
25  1   2   3   4   5   6   7   9   13  14  15  17  18  23  1   2   3   4
5   9   13  14  15  17  18  23  1   2   3   4   5   8   9   10  11  12  13
14  15  16  17  18  19  20  21  22  23  1   2   3   4   5   9   13  14  15
17  18  23  1   2   3   4   5   9   13  14  15  17  18  23  1   2   3   4
5   9   13  14  15  17  18  23  1   2   3   4   5   9   13  14  15  17  18
23  1   2   3   4   5   9   13  14  15  17  18  23  1   2   3   4   5   9
13  14  15  17  18  23  1   2   3   4   5   9   13  14  15  17  18  23  1
2   3   4   5   9   13  14  15  17  18  23  1   2   3   4   5   9   13  14
15  17  18  23  1   2   3   4   5   9   13  14  15  17  18  23  1   2   3
4   5   9   13  14  15  17  18  23  1   2   3   4   5   9   13  14  15  17
18  23  1   2   3   4   5   9   13  14  15  17  18  23  1   2   3   4   5
9   13  14  15  17  18  23  24  1   2   3   4   5   9   13  14  15  17  18
23  1   2   3   4   5   9   13  14  15  17  18  23
],W);

%Graficar
%h = view(biograph(DG,[],'ShowWeights','on'));

% Nodos de inicio donde no hay hyperlinks
Nodos_ini = [
1   1   1   1   1   1   1   1   1   1   1   1   1   2   2   2   2   2   2
2   2   2   2   2   3   3   3   3   3   3   3   3   3   3   3   3   3   4
4   4   4   4   4   4   4   4   4   4   4   4   5   5   5   5   5   5   5
5   5   5   5   5   5   6   6   6   6   6   6   6   6   6   6   7   7   7
7   7   7   7   7   7   7   7   8   8   8   8   8   8   8   8   8   8   8
8   8   9   9   9   9   10  10  10  10  10  10  10  10  10  10  10  10  10
11  11  11  11  11  11  11  11  11  11  11  11  11  12  12  12  12  12  12
12  12  12  12  12  12  12  13  13  13  13  13  13  13  13  13  13  13  13
13  14  14  14  14  14  14  14  14  14  14  14  14  14  15  15  15  15  15
15  15  15  15  15  15  15  15  16  16  16  16  16  16  16  16  16  16  16
16  16  17  17  17  17  17  17  17  17  17  17  17  17  17  18  18  18  18
18  18  18  18  18  18  18  18  18  19  19  19  19  19  19  19  19  19  19
19  19  19  20  20  20  20  20  20  20  20  20  20  20  20  20  21  21  21
21  21  21  21  21  21  21  21  21  21  22  22  22  22  22  22  22  22  22
22  22  22  22  23  23  23  23  23  23  23  23  23  23  23  23  24  24  24
24  24  24  24  24  24  24  24  24  24  25  25  25  25  25  25  25  25  25
25  25  25  25
];

% Nodos de fin donde no hay hyperlinks
Nodos_fin = [
6   7   8   10  11  12  16  19  20  21  22  24  25  7   8   10  11  12  16
19  20  21  22  24  6   7   8   10  11  12  16  19  20  21  22  24  25  6
7   8   10  11  12  16  19  20  21  22  24  25  6   7   8   10  11  12  16
19  20  21  22  24  25  8   10  11  12  16  19  20  21  22  24  8   10  11
12  16  19  20  21  22  24  25  6   7   8   10  11  12  16  19  20  21  22
24  25  6   7   24  25  6   7   8   10  11  12  16  19  20  21  22  24  25
6   7   8   10  11  12  16  19  20  21  22  24  25  6   7   8   10  11  12
16  19  20  21  22  24  25  6   7   8   10  11  12  16  19  20  21  22  24
25  6   7   8   10  11  12  16  19  20  21  22  24  25  6   7   8   10  11
12  16  19  20  21  22  24  25  6   7   8   10  11  12  16  19  20  21  22
24  25  6   7   8   10  11  12  16  19  20  21  22  24  25  6   7   8   10
11  12  16  19  20  21  22  24  25  6   7   8   10  11  12  16  19  20  21
22  24  25  6   7   8   10  11  12  16  19  20  21  22  24  25  6   7   8
10  11  12  16  19  20  21  22  24  25  6   7   8   10  11  12  16  19  20
21  22  24  25  6   7   8   10  11  12  16  19  20  21  22  25  6   7   8
10  11  12  16  19  20  21  22  24  25  6   7   8   10  11  12  16  19  20
21  22  24  25
];

% 25*25 = 625
```

```
% 625 menos 317 hyperlinks = 308
% A será la matriz con todos los caminos faltantes

A=cell(308,1);
aux0=1;
for i=1:308
    ini=Nodos_ini(i);
    fin=Nodos_fin(i);
        [aux1,aux2,aux3] = graphshortestpath(DG,ini,fin);
        % Primera columna de A contiene las matrices de rutas
        A{aux0,1} = aux2;
        % Segunda columna de A contiene los nodos de inicio y fin
        A{aux0,2} = [ini fin];
        aux0=aux0+1;
end

% Exportar datos Excel archivo 'prob.xls' hoja Pesos_cond'
M=xlsread('Prob3.xlsm','Prob_cond','D3:AB27');
% Estimación de pesos de hyperlinks no existentes
n = size(A,1);
M2 = zeros(n,3);
M3 = M; %Actualizar matriz de probabilidades
for i = 1:n
    if ~isempty(A{i,1})
        M2(i,1) = A{i,1}(1);
        M2(i,2) = A{i,1}(end);
        aux = 1;
        for j = 1:length(A{i,1})-1
            aux = aux*M(A{i,1}(j),A{i,1}(j+1));
        end
        M2(i,3) = aux;
        M3(M2(i,1),M2(i,2))=aux;
    end
end

%Importar matriz de factor tiempo en página destino desde Excel
T=xlsread('Prob3.xlsm','Tiempo_fin','D3:AB27');
```

Modelo 1, Fitness @mem1

```
function perf = mem1(R)
% R     ->  Matriz de adyacencia
% perf  <-  Evaluación de matriz de adyacencia

m = sqrt(l);
if (rem(m,1) ~= 0)
    error('la matriz no es cuadrada')
end

R0 = double(R);   % Convertir números
R1 = reshape(R0,25,25); % Redefinir matriz vector a matriz cuadrada

% ---------------- Restricción - Conexidad ----------------------------

% Sacar matriz simétrica
Asim=or(R1, R1');
% Extraer matriz diagonal MD
MD=diag(Asim);
% Crear una matriz con los elementos de la diagonal
MMD=diag(MD);
% Restarle a matriz simétrica, elementos de la diagonal
Asim2=Asim-MMD;
```

```
% Calcular matriz de grado G (suma de filas)
G=sum(Asim2,2);
% Crear matriz diagonal con los grados obtenidos
MG=diag(G);
% Obtener matriz de Laplace, restando matriz de grado menos diagonal
LL=MG-Asim2;
% Cálculo de vectores propios
V=null(LL);
% Cálculo de cantidad de vectores
C=size(V,2);

if C~=1
    perf=1e10; % Se penaliza con valor muy alto
   %disp('No Conexo')
   return
end

% ----------------- Restricción - Nulidad -----------------------------

    % Suma columna
    if length(find(sum(R1,1)))~=25
        perf = 1e10;
        disp('Hay una página a la que no llegan hyperlinks')
        return
    end
    % Suma fila
    if length(find(sum(R1,2)))~=25
        perf = 1e10;
        disp('Hay una página de la que no salen hyperlinks')
        return
    end

% --------------------------- Funcional -------------------------------

alfa= 100000;
beta= 1;
Num_links = beta*sum(R0);

%Considerar matriz M, ya dispuesta en el Workspace
I = evalin('base','M3');
%Considerar matriz T de factor tiempo, ya dispuesta en el Workspace
T = evalin('base','T');
%Multiplicación por punto
P=I.*T;

Pesos = (R1.*P);
%Minimizar hyperlinks totales y maximizar los pesos de hyperlinks
existentes
perf =  Num_links - alfa*sum(sum(Pesos));
```

Modelo 2, datos de entrada

```
%Exportar matrices de pesos, ya multiplicadas las frecuencias de llegadas y
de factor tiempo desde Excel
P=xlsread('Prob3.xlsm','PI','D3:AB27');
```

Modelo 2, Fitness @mem2

```
function perf = mem2(R)

mm = length(R);
```

```
m = sqrt(mm);
if (rem(m,1) ~= 0)
error('la matriz no es cuadrada')
end

R0 = double(R);   % Convertir números
R1 = reshape(R0,m,m); % Redefinir matriz vector a matriz cuadrada

% ---------------------- Restricción Conexidad ----------------------------

% IDEM A EXPRESION DE RESTRICCIÓN CONEXIDAD ANTERIOR

% ------------------------ Restricción - Nulidad -------------------------

% IDEM A EXPRESION DE RESTRICCIÓN CONEXIDAD ANTERIOR

% ------------------------------ Funcional ---------------------------------

alfa= 300;
beta= 0.2;
Num_links = beta*sum(R0);

% Importar matrices de peso desde Matlab
P = evalin('base','P');

Pesos  = zeros(m,m);

for i = 1:m
    for j= 1:m
        if P(i,j) > 0
            uno=R1(i,j);
            dos=P(i,j);
            %Pesos(i,j) = uno*log(dos);
            Pesos(i,j) = uno*dos;
        end
    end
end

Pesos_sum = sum(sum(Pesos));
perf =  Num_links - alfa*Pesos_sum;
```

Modelo 3, datos de entrada

```
%Exportar todas las matrices de peso por largo de caminos
P1=xlsread('Prob_caminos.xlsm','1','D3:AB27');
P2=xlsread('Prob_caminos.xlsm','2','D3:AB27');
P3=xlsread('Prob_caminos.xlsm','3','D3:AB27');
P4=xlsread('Prob_caminos.xlsm','4','D3:AB27');
P5=xlsread('Prob_caminos.xlsm','5','D3:AB27');
P6=xlsread('Prob_caminos.xlsm','6','D3:AB27');
P7=xlsread('Prob_caminos.xlsm','7','D3:AB27');
P8=xlsread('Prob_caminos.xlsm','8','D3:AB27');
P9=xlsread('Prob_caminos.xlsm','9','D3:AB27');
P10=xlsread('Prob_caminos.xlsm','10','D3:AB27');
P11=xlsread('Prob_caminos.xlsm','11','D3:AB27');
P12=xlsread('Prob_caminos.xlsm','12','D3:AB27');
P13=xlsread('Prob_caminos.xlsm','13','D3:AB27');
P14=xlsread('Prob_caminos.xlsm','14','D3:AB27');
```

Modelo 3, Fitness @mem3

```
function perf=mem3(R)

R0 = double(R);   % Convertir numeros
R1 = reshape(R0,25, 25); % Redefinir matriz vector a matriz cuadrada

% ---------------------- Restriccion Num_links ----------------------------

Num_link = sum(R0);
    if (Num_link <200 || Num_link >300)
            perf=1e6;
            return
    end

% ---------------------- Restricción Conexidad ----------------------------

% IDEM A EXPRESION DE RESTRICCIÓN CONEXIDAD ANTERIOR

% ------------------------ Restricción - Nulidad --------------------------

% IDEM A EXPRESION DE RESTRICCIÓN NULIDAD ANTERIOR

% ----------------------- Funcional -------------------------------------

Largos=cell(14,1);

Largos{1,1} = evalin('base','P1');
Largos{2,1} = evalin('base','P2');
Largos{3,1} = evalin('base','P3');
Largos{4,1} = evalin('base','P4');
Largos{5,1} = evalin('base','P5');
Largos{6,1} = evalin('base','P6');
Largos{7,1} = evalin('base','P7');
Largos{8,1} = evalin('base','P8');
Largos{9,1} = evalin('base','P9');
Largos{10,1} = evalin('base','P10');
Largos{11,1} = evalin('base','P11');
Largos{12,1} = evalin('base','P12');
Largos{13,1} = evalin('base','P13');
Largos{14,1} = evalin('base','P14');

Pesos_fin=zeros(25,25);

for k=1:14
    RR=R1^k;

for i = 1:25
    for j= 1:25
        if Largos{k,1}(i,j) > 0
            uno=RR(i,j);
            dos=Largos{k,1}(i,j);
            %Pesos_fin(i,j) = Pesos_fin(i,j)+uno*(log(dos));
            Pesos_fin(i,j) = Pesos_fin(i,j)+(uno*dos);
            %Pesos(i,j) = uno*dos;
        end
    end
end
end
%Llevar a Workspace matriz de pesos fin
```

```
assignin('base', 'Pesos_caminos_mem4', Pesos_fin);
%Maximizar los pesos de hyperlinks existentes
perf = sum(sum(Pesos_fin));
```

Algoritmo para extraer adyacencia final
(Común para los tres modelos)

```
Prueba0 = evalin('base','optimresults1.x');
Prueba=double(Prueba0);
l = length(Prueba);
m = sqrt(l);
if (rem(m,1) ~= 0)
    error('la matriz no es cuadrada')
end

Ady=reshape(Prueba,m,m);
xlswrite('Prueba1', Ady);
```

E.3. Adyacencia inicial y resultantes

A continuación se muestran las matrices resultantes de cada modelo, con la el grado de autoridad y Hub de cada página, es decir, con la suma por columna y fila de los hyperlinks respectivamente. Primero se muestra la adyacencia inicial existente. En todas las matrices, se destaca con fondo destacado, aquellas interacciones con valor 1.

Tabla E. 1: Matriz adyacencia original

		ADYACENCIA ORIGINAL																								
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	sum
1	Home	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
2	Plan_estudio	1	1	1	1	1	1	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	1	14
3	Sobre_mgo	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
4	Futuro_est.	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
5	Costo_becas	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
6	Electivos	1	1	1	1	1	1	1	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	1	15
7	Elect_IN78K	1	1	1	1	1	1	1	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	14
8	Eg_2006	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
9	Egresados	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	21
10	Eg_2009	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
11	Eg_2008	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
12	Eg_2007	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
13	Calendario	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
14	Profesores	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
15	Admision	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
16	Eg_2005	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
17	Serv_gral.	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
18	Links	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
19	Eg_2001	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
20	Eg_2002	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
21	Eg_2003	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
22	Eg_2004	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
23	Contacto	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	1	0	13
24	C_enviado	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
25	IN71K	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	12
	sum	25	25	25	25	25	3	2	1	25	1	1	1	25	25	25	1	25	25	1	1	1	1	25	1	2	

Tabla E. 2: Adyacencia resultante modelo 1

		ADYACENCIA RESULTANTE MODELO 1																								
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	sum
1	Home	1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	1	1	0	0	0	0	1	0	1	16
2	Plan_estudio	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	0	1	1	0	0	0	0	0	0	1	15
3	Sobre_mgo	1	1	1	1	1	1	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	1	14
4	Futuro_est.	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	1	1	0	0	0	0	1	0	1	15
5	Costo_becas	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	1	1	0	0	0	0	1	0	1	15
6	Electivos	1	1	1	1	1	1	1	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	1	15
7	Elect_IN78K	0	1	1	1	1	1	1	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	10
8	Eg_2006	0	1	0	1	1	0	0	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	8
9	Egresados	1	1	1	1	1	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	19
10	Eg_2009	1	1	1	1	1	1	0	0	1	1	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	12
11	Eg_2008	1	1	1	1	1	1	0	0	1	0	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	11
12	Eg_2007	0	1	0	1	1	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	6
13	Calendario	1	1	1	1	1	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	1	0	0	10
14	Profesores	1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	1	1	0	0	0	0	1	0	0	15
15	Admision	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	1	1	0	0	0	0	1	0	1	15
16	Eg_2005	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	5
17	Serv_gral.	1	1	1	1	1	1	0	0	1	0	0	0	1	1	1	0	0	1	0	0	0	0	1	0	0	12
18	Links	1	1	1	0	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	0	0	11
19	Eg_2001	0	1	1	1	1	0	0	0	1	0	0	0	1	1	1	0	1	0	1	0	0	0	1	0	0	11
20	Eg_2002	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	7
21	Eg_2003	1	1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	7
22	Eg_2004	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	6
23	Contacto	1	1	1	0	1	1	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	1	0	13
24	C_enviado	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	7
25	IN71K	1	1	1	1	1	1	0	0	1	0	0	0	0	1	1	0	0	1	0	0	0	0	1	0	1	12
	sum	18	25	18	22	21	14	3	2	23	8	4	2	13	24	22	2	18	15	2	2	2	2	15	2	8	

Tabla E. 3: Adyacencia resultante modelo 2

		ADYACENCIA RESULTANTE MODELO 2																								
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	sum
1	Home	1	1	1	1	1	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	10
2	Plan_estudio	1	1	1	1	1	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	9
3	Sobre_mgo	1	1	1	1	1	1	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	10
4	Futuro_est.	1	1	1	0	1	0	0	0	1	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	11
5	Costo_becas	1	1	1	1	1	1	0	0	1	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	11
6	Electivos	0	1	1	1	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	8
7	Elect_IN78K	0	1	0	0	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	7
8	Eg_2006	0	1	0	1	0	0	0	0	1	1	1	1	0	1	0	1	0	0	1	1	1	1	0	0	0	12
9	Egresados	0	1	0	1	1	0	0	1	1	1	1	1	0	1	0	1	0	0	1	1	1	1	0	0	0	14
10	Eg_2009	0	1	0	1	1	0	0	1	1	0	1	1	0	1	1	1	0	0	0	1	0	0	0	0	0	11
11	Eg_2008	0	1	1	1	1	0	0	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	0	0	0	16
12	Eg_2007	0	1	0	1	1	0	0	1	1	1	1	0	0	1	0	1	0	0	1	1	1	1	0	0	0	13
13	Calendario	1	1	1	1	1	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	10
14	Profesores	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	11
15	Admision	1	1	1	1	1	1	0	0	1	1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	12
16	Eg_2005	0	1	0	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0	1	1	1	1	0	0	0	9
17	Serv_gral.	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	0	1	0	0	0	0	0	0	1	12
18	Links	1	1	1	1	1	1	0	0	1	1	1	0	0	1	1	0	1	1	0	0	0	0	1	0	0	14
19	Eg_2001	0	1	0	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	1	1	1	0	0	0	11
20	Eg_2002	0	1	0	1	0	0	0	0	1	1	1	1	0	1	0	1	0	0	1	1	1	1	0	0	0	12
21	Eg_2003	0	1	0	1	0	0	0	0	1	1	1	0	0	1	0	1	0	0	1	1	0	1	0	0	0	10
22	Eg_2004	0	1	0	1	0	0	0	1	0	1	1	1	0	1	0	1	0	0	1	1	1	0	0	0	0	11
23	Contacto	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	1	1	0	0	0	0	1	1	1	16
24	C_enviado	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	9
25	IN71K	1	1	1	1	1	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	10
	sum	13	25	14	22	19	11	2	6	19	19	14	8	2	21	17	10	5	5	8	10	8	8	3	2	8	

Tabla E. 4: Adyacencia resultante modelo 3

		ADYACENCIA RESULTANTE MODELO 3																								
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	
1	Home	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
2	Plan_estudio	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	1	0	18
3	Sobre_mgo	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0	17
4	Futuro_est.	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	1	18
5	Costo_becas	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0	18
6	Electivos	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0	17
7	Elect_IN78K	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0	17
8	Eg_2006	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
9	Egresados	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0	18
10	Eg_2009	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	0	0	18
11	Eg_2008	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0	17
12	Eg_2007	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	1	1	0	0	1	1	0	0	0	16
13	Calendario	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
14	Profesores	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0	0	18
15	Admision	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0	17
16	Eg_2005	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	1	0	0	0	1	1	0	0	0	15
17	Serv_gral.	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0	17
18	Links	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	0	0	18
19	Eg_2001	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
20	Eg_2002	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
21	Eg_2003	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	1	1	0	0	1	1	0	0	0	16
22	Eg_2004	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	0	17
23	Contacto	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
24	C_enviado	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
25	IN71K	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	sum	1	18	19	18	17	17	17	1	17	17	17	17	1	20	17	14	17	16	1	1	17	17	1	1	1	

E.4. Grado de centralidad de nodos

Utilizando el programa Pajek, es posible obtener indicadores del grado de entrada, salida o global por cada página. Pajek normaliza la cantidad de nodos entrantes, salientes y totales respectivamente, resultando las tablas que prosiguen.

Centralidad total normalizada

Original

x	Val	Label
1	0.7708333	Home
2	0.8125000	Plan_estudio
3	0.7708333	Sobre_mgo
4	0.7708333	Futuro_est.
5	0.7708333	Costo_becas
6	0.3750000	Electivos
7	0.3333333	Elect_IN78K
8	0.2708333	Eg_2006
9	0.9583333	Egresados
10	0.2708333	Eg_2009
11	0.2708333	Eg_2008
12	0.2708333	Eg_2007
13	0.7708333	Calendario
14	0.7708333	Profesores
15	0.7708333	Admision
16	0.2708333	Eg_2005
17	0.7708333	Serv_gral.
18	0.7708333	Links
19	0.2708333	Eg_2001
20	0.2708333	Eg_2002
21	0.2708333	Eg_2003
22	0.2708333	Eg_2004
23	0.7916667	Contacto
24	0.2708333	C_enviado
25	0.2916667	IN71K

Modelo 1

x	Val	Label
1	0.7083333	Home
2	0.8333333	Plan_estudio
3	0.6666667	Sobre_mgo
4	0.7708333	Futuro_est.
5	0.7500000	Costo_becas
6	0.6041667	Electivos
7	0.2708333	Elect_IN78K
8	0.2083333	Eg_2006
9	0.8750000	Egresados
10	0.4166667	Eg_2009
11	0.3125000	Eg_2008
12	0.1666667	Eg_2007
13	0.4791667	Calendario
14	0.8125000	Profesores
15	0.7708333	Admision
16	0.1458333	Eg_2005
17	0.6250000	Serv_gral.
18	0.5416667	Links
19	0.2708333	Eg_2001
20	0.1875000	Eg_2002
21	0.1875000	Eg_2003
22	0.1666667	Eg_2004
23	0.5833333	Contacto
24	0.1875000	C_enviado
25	0.4166667	IN71K

Modelo 2

x	Val	Label
1	0.4791667	Home
2	0.7083333	Plan_estudio
3	0.5000000	Sobre_mgo
4	0.6875000	Futuro_est.
5	0.6250000	Costo_becas
6	0.3958333	Electivos
7	0.1875000	Elect_IN78K
8	0.3750000	Eg_2006
9	0.6875000	Egresados
10	0.6250000	Eg_2009
11	0.6250000	Eg_2008
12	0.4375000	Eg_2007
13	0.2500000	Calendario
14	0.6666667	Profesores
15	0.6041667	Admision
16	0.3958333	Eg_2005
17	0.3541667	Serv_gral.
18	0.3958333	Links
19	0.3958333	Eg_2001
20	0.4583333	Eg_2002
21	0.3750000	Eg_2003
22	0.3958333	Eg_2004
23	0.3958333	Contacto
24	0.2291667	C_enviado
25	0.3750000	IN71K

Modelo 3

	Val	Label
1	0.0416667	Home
2	0.7500000	Plan_estudio
3	0.7500000	Sobre_mgo
4	0.7500000	Futuro_est.
5	0.7291667	Costo_becas
6	0.7083333	Electivos
7	0.7083333	Elect_IN78K
8	0.0416667	Eg_2006
9	0.7291667	Egresados
10	0.7291667	Eg_2009
11	0.7083333	Eg_2008
12	0.6875000	Eg_2007
13	0.0416667	Calendario
14	0.7916667	Profesores
15	0.7083333	Admision
16	0.6041667	Eg_2005
17	0.7083333	Serv_gral.
18	0.7083333	Links
19	0.0416667	Eg_2001
20	0.0416667	Eg_2002
21	0.6875000	Eg_2003
22	0.7083333	Eg_2004
23	0.0416667	Contacto
24	0.0416667	C_enviado
25	0.0416667	IN71K

Grados de entrada normalizada

Original

x	Val	Label
1	1.0416667	Home
2	1.0416667	Plan_estudio
3	1.0416667	Sobre_mgo
4	1.0416667	Futuro_est.
5	1.0416667	Costo_becas
6	0.1250000	Electivos
7	0.0833333	Elect_IN78K
8	0.0416667	Eg_2006
9	1.0416667	Egresados
10	0.0416667	Eg_2009
11	0.0416667	Eg_2008
12	0.0416667	Eg_2007
13	1.0416667	Calendario
14	1.0416667	Profesores
15	1.0416667	Admision
16	0.0416667	Eg_2005
17	1.0416667	Serv_gral.
18	1.0416667	Links
19	0.0416667	Eg_2001
20	0.0416667	Eg_2002
21	0.0416667	Eg_2003
22	0.0416667	Eg_2004
23	1.0416667	Contacto
24	0.0416667	C_enviado
25	0.0833333	IN71K

Modelo 1

x	Val	Label
1	0.7500000	Home
2	1.0416667	Plan_estudio
3	0.7500000	Sobre_mgo
4	0.9166667	Futuro_est.
5	0.8750000	Costo_becas
6	0.5833333	Electivos
7	0.1250000	Elect_IN78K
8	0.0833333	Eg_2006
9	0.9583333	Egresados
10	0.3333333	Eg_2009
11	0.1666667	Eg_2008
12	0.0833333	Eg_2007
13	0.5416667	Calendario
14	1.0000000	Profesores
15	0.9166667	Admision
16	0.0833333	Eg_2005
17	0.7500000	Serv_gral.
18	0.6250000	Links
19	0.0833333	Eg_2001
20	0.0833333	Eg_2002
21	0.0833333	Eg_2003
22	0.0833333	Eg_2004
23	0.6250000	Contacto
24	0.0833333	C_enviado
25	0.3333333	IN71K

Modelo 2

x	Val	Label
1	0.5416667	Home
2	1.0416667	Plan_estudio
3	0.5833333	Sobre_mgo
4	0.9166667	Futuro_est.
5	0.7916667	Costo_becas
6	0.4583333	Electivos
7	0.0833333	Elect_IN78K
8	0.2500000	Eg_2006
9	0.7916667	Egresados
10	0.7916667	Eg_2009
11	0.5833333	Eg_2008
12	0.3333333	Eg_2007
13	0.0833333	Calendario
14	0.8750000	Profesores
15	0.7083333	Admision
16	0.4166667	Eg_2005
17	0.2083333	Serv_gral.
18	0.2083333	Links
19	0.3333333	Eg_2001
20	0.4166667	Eg_2002
21	0.3333333	Eg_2003
22	0.3333333	Eg_2004
23	0.1250000	Contacto
24	0.0833333	C_enviado
25	0.3333333	IN71K

Modelo 3

x	Val	Label
1	0.0416667	Home
2	0.7500000	Plan_estudio
3	0.7916667	Sobre_mgo
4	0.7500000	Futuro_est.
5	0.7083333	Costo_becas
6	0.7083333	Electivos
7	0.7083333	Elect_IN78K
8	0.0416667	Eg_2006
9	0.7083333	Egresados
10	0.7083333	Eg_2009
11	0.7083333	Eg_2008
12	0.7083333	Eg_2007
13	0.0416667	Calendario
14	0.8333333	Profesores
15	0.7083333	Admision
16	0.5833333	Eg_2005
17	0.7083333	Serv_gral.
18	0.6666667	Links
19	0.0416667	Eg_2001
20	0.0416667	Eg_2002
21	0.7083333	Eg_2003
22	0.7083333	Eg_2004
23	0.0416667	Contacto
24	0.0416667	C_enviado
25	0.0416667	IN71K

Grados de salida normalizada

Original

x	Val	Label
1	0.0416667	Home
2	0.7500000	Plan_estudio
3	0.7083333	Sobre_mgo
4	0.7500000	Futuro_est.
5	0.7500000	Costo_becas
6	0.7083333	Electivos
7	0.7083333	Elect_IN78K
8	0.0416667	Eg_2006
9	0.7500000	Egresados
10	0.7500000	Eg_2009
11	0.7083333	Eg_2008
12	0.6666667	Eg_2007
13	0.0416667	Calendario
14	0.7500000	Profesores
15	0.7083333	Admision
16	0.6250000	Eg_2005
17	0.7083333	Serv_gral.
18	0.7500000	Links
19	0.0416667	Eg_2001
20	0.0416667	Eg_2002
21	0.6666667	Eg_2003
22	0.7083333	Eg_2004
23	0.0416667	Contacto
24	0.0416667	C_enviado
25	0.0416667	IN71K

Modelo 1

x	Val	Label
1	0.5000000	Home
2	0.5833333	Plan_estudio
3	0.5000000	Sobre_mgo
4	0.5000000	Futuro_est.
5	0.5000000	Costo_becas
6	0.6250000	Electivos
7	0.5833333	Elect_IN78K
8	0.5000000	Eg_2006
9	0.8750000	Egresados
10	0.5000000	Eg_2009
11	0.5000000	Eg_2008
12	0.5000000	Eg_2007
13	0.5000000	Calendario
14	0.5000000	Profesores
15	0.5000000	Admision
16	0.5000000	Eg_2005
17	0.5000000	Serv_gral.
18	0.5000000	Links
19	0.5000000	Eg_2001
20	0.5000000	Eg_2002
21	0.5000000	Eg_2003
22	0.5000000	Eg_2004
23	0.5416667	Contacto
24	0.5000000	C_enviado
25	0.5000000	IN71K

Modelo 2

x	Val	Label
1	0.6666667	Home
2	0.6250000	Plan_estudio
3	0.5833333	Sobre_mgo
4	0.6250000	Futuro_est.
5	0.6250000	Costo_becas
6	0.6250000	Electivos
7	0.4166667	Elect_IN78K
8	0.3333333	Eg_2006
9	0.7916667	Egresados
10	0.5000000	Eg_2009
11	0.4583333	Eg_2008
12	0.2500000	Eg_2007
13	0.4166667	Calendario
14	0.6250000	Profesores
15	0.6250000	Admision
16	0.2083333	Eg_2005
17	0.5000000	Serv_gral.
18	0.4583333	Links
19	0.4583333	Eg_2001
20	0.2916667	Eg_2002
21	0.2916667	Eg_2003
22	0.2500000	Eg_2004
23	0.5416667	Contacto
24	0.2916667	C_enviado
25	0.5000000	IN71K

Modelo 3

x	Val	Label
1	0.4166667	Home
2	0.3750000	Plan_estudio
3	0.4166667	Sobre_mgo
4	0.4583333	Futuro_est.
5	0.4583333	Costo_becas
6	0.3333333	Electivos
7	0.2916667	Elect_IN78K
8	0.5000000	Eg_2006
9	0.5833333	Egresados
10	0.4583333	Eg_2009
11	0.6666667	Eg_2008
12	0.5416667	Eg_2007
13	0.4166667	Calendario
14	0.4583333	Profesores
15	0.5000000	Admision
16	0.3750000	Eg_2005
17	0.5000000	Serv_gral.
18	0.5833333	Links
19	0.4583333	Eg_2001
20	0.5000000	Eg_2002
21	0.4166667	Eg_2003
22	0.4583333	Eg_2004
23	0.6666667	Contacto
24	0.3750000	C_enviado
25	0.4166667	IN71K

7.- Bibliografía

[1] S. BANDYOPADHYAY and S.K.PAL. Classification and Learning Using Genetic Algorithms. Springer-Verlag Berlin Heidelberg, 2007

[2] Z. BAR-YOSSEF, A. BERG, S. CHIEN, F. FAKCHAROENPHOL and D. WEITZ. Approximating aggregate queries about Web pages via random walks. Proceeding of the 26th International Conference on Very Large Databases, pp. 535-544, September 2002.

[3] R.A. BAEZA-YATES. Tendencias en minería de datos de la Web. El profesional de la información, Vol. 18, No. 1, pp. 5-10, Enero-Febrero 2009.

[4] R.A. BAEZA-YATES and B. RIBEIRO-NETO. Modern Information Retrieval. Addison-Wesley, 1999.

[5] P. BALDI, P. FRASCONI and P. SMYTH. Modeling the Internet and the Web: Probabilistic Methods and Algorithms. John Wiley & Sons Ltd., 2003.

[6] L. BEN, T. BOURON and A. DROGOUL. Agent-based Interaction Analysis of Consumer Behavior. First International Joint Conference on Autonomous Agents & Multiagent Systems AAMAS, pp. 184-190, 2002

[7] B. BERENDT, B. MOBASHER, M. SPILIOPOULOU, and M. NAKAGAWA. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis . Proceedings of the WebKDD 2002 Workshop, Held at KDD'2002, Edmonton, Alberta, Canada, July 2002.

[8] B. BERENDT, B. MOBASHER and M. SPILIOPOULOU. Web Usage Mining for E-Business Applications, CML/PKDD-2002 Tutorial, 19 August 2002. (http://www.cs.helsinki.fi/events/ecmlpkdd/pdf/berendt-2.pdf)

[9] T. BERNERS-LEE, R. CAILLIAU, A.LUOTONEN, H.F. NIELSEN a.nd A. SECRET. The world wide web. Communications of ACM, Vol 37, No. 8, pp. 76-82, 1994

[10] A. Z. BRODER. A taxonomy of web search. SIGIR Forum, Vol 36, No. 2, pp. 3–10, 2002.

[11] R. CABALLERO F., J. MOLINA L., M. LUQUE G., A. TORRICO G. y T. GÓMEZ N. "Algoritmos genéticos para la resolución de problemas de Programación por Metas Entera. Aplicación a la Economía de la Educación", X Jornadas de la Asociación española de profesores universitarios de Matemática en la economía y en la Empresa. Madrid (España), Septiembre 2002.

[12] R. CALDENTEY and S. MONDSCHEIN. Modelos de Decisión en Ambientes Inciertos. Capítulos 3: Procesos de Poisson. Departamento de Ingeniería Industrial, Universidad de Chile. Enero, 1999.

[13] R. CALDENTEY and S. MONDSCHEIN. Modelos de Decisión en Ambientes Inciertos. Capítulos 4: Cadenas de Markov. Departamento de Ingeniería Industrial, Universidad de Chile. Enero, 1999.

[14] S. CHAKRABARTI, B. E. DOM., S. R. KUMAR, P. RAGHAVAN, S. RAJAGOPALAN, A TOMKINS, D. GIBSON, and J. KLEINBERG. Mining the web's hyperlink structure. IEEE Computer, Vol. 32, No. 8, pp. 60-67, August 1999.

[15] S. CHAKRABARTI. Mining the Web, Discovering Knowledge from hypertext data. Morgan Kaufmann Publishers, Elsevier Science Press, 2008

[16] CIW-CENTRO DE INVESTIGACIÓN DE LA WEB, UNIVERSIDAD DE CHILE. Cómo funciona la Web, Junio 2008

[17] A. COCKBURN and B. MCKENZIE. What do Web users do? An empirical analysis of Web use. International Journal of Human-Computer Studies Vol. 54, No. 6, 903–922, June 2001.

[18] F. CRESTANI and G. PASI, editors. Soft Computing in Information Retrieval: Techniques and Application, Vol. 50. Physica-Verlag, Heidelberg, 2000.

[19] W. DE NOOY, A. MRVAR and V. BATAGELJ. Exploratory Social Network Analysis with Pajek. CUP, January 2005.

[20] R. F. DELL, R., P. ROMAN, and J. D. VELÁSQUEZ. "Optimization Models for Construction and Analysis of Web User Sessions". 11th Informs Computing Society Conference, Charleston, South Carolina, USA, January, 2009.

[21] R.F. DELL, P.E. ROMÁN, J.D. VELÁSQUEZ. Web User Session Reconstruction Using Integer Programming. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Vol. 1, pp.385-388, 2008.

[22] P. ESTÉVEZ. Optimización mediante algoritmos genéticos. Anales del Instituto de Ingenieros de Chile, pp. 83-92, Agosto 1997

[23] M. EVANS and A. WALKER. Using the Web Graph to influence application behavior. Internet Research, Vol. 14, No. 5, pp. 372-378, 2004.

[24] Y. GAO. Web System Design and Onlines Consumer Behavior. Ramapo College of New Jersey, USA. Idea Group Publishing, 2005.

[25] D. GERNERT. Ockham's Razor and Its Improper Use, Journal of Scientific Exploration, Vol. 21, No. 1, pp. 135-140, 2007.
(http://rr0.org/data/2/0/0/7/Gernert_OckhamsRazorAndItsImproperUse/index.html)

[26] M. GESTAL. Introducción a los Algoritmos genéticos. Depto. Tecnologías de la Información y las Comunicaciones Universidad de la Coruña.

[27] D. GOLDBERG. Genetic Algorithms in Search Optimization, and Machine Learning. Addison-Weasley Longman Publishing Co., Inc., Boston, MA, USA. 1989.

[28] B. HUBERMAN, P. PIROLLI, J. PITKOW and R. LUKOSE. Strong regularities in World Wide Web surfing. Science; Vol. 280, No. 5360, pp. 95-97, April 1998

[29] M. D. GORDON. Probabilistic and genetic algorithms for document retrieval. Communications of the ACM, Vol. 31, No. 10, pp. 208-218, 1988.

[30] G. GORMAN. Online Information Review: The international journal of digital information research and use. Web-mining applications in e-commerce and e-services. Vol. 32 No. 2, 2008. (www.emeraldinsight.com)

[31] J. GRAPPONE and G. COUZIN. Search Engine Optimization: An Hour a Day. SYBEX Inc. Alameda, CA, USA, 2006

[32] J. GUILLAUME and M. LATAPY. The web graph: an overview. Proceedings of Algotel 2002 (http://citeseer.nj.nec.com/guillaume02web.html)

[33] T. HARFORD. El economista camuflado: La economía de las pequeñas cosas. Oxford University Press, Inc., Editorial Booket, 2008.

[34] M. HITT, A. IRELAND and R. HASKISSON. Administración estratégica: Competitividad y conceptos de globalización. Thompson International, 5ta Edición, Capítulo 3, pág. 89, 2004.

[35] J. HOLLAND. Adaptation in Natural and Artificial Systems. MIT Press, 1992.

[36] H. KARGUPTA. The gene expression messy genetic algorithm. In Proceedings of the IEEE International Conference on Evolutionary Computation, pp. 631-636, 1996.

[37] E. KIM, W. KIM and Y. LEE. Combination of multiple classifiers for the consumer's purchase behavior prediction. Elsevier Science, 2002.

[38] A. KAUSHIK. Web Analytics: An Hour a Day. Wiley Publishing Inc., 2007

[39] M. KOUFARIS. Applying the Technology Acceptance Model and Flow Theory to Online Consumer Behavior. Information System Research, Vol.13, No. 2, pp. 205-223, June 2002

[40] J. KOZA. Genetic Programming: On the Programming of Computers by Means of Natural Selection (Complex Adaptative System) MIT Press, 1992.

[41] S. LAWRENCE and C. GILES. Nature 400, pp. 107=109, 1999.

[42] J. LIU, S. ZHANG and J. YANG. Characterizing Web usage regularities with information foraging agents. Knowledge and Data Engineering, IEEE, Vol. 16, No. 5, pp. 566 - 584, May 2004

[43] M. H. MARGHNY and A. F. ALI. Web mining based on Genetic Algorithm. In Proceedings of ICGST International Conference on Artificial Intelligence and Machine Learning (AIML-05), Cairo Egypt, December 2005.

[44] Z. MARKOV and D. T. LAROSE. Data Mining the Web: Uncovering Patterns in Web Content, Structure and Usage. John Wiley & Sons, 2007.

[45] M. MITCHELL. An Introduction to Genetic Algorithms. MIT Press, 1999.

[46] D. MOHANRAJ. Soft Computing, Assignment. Genetic Algorithms, 2005.

[47] Z. Z. NICK and P. THEMIS. Web search using a genetic algorithm. IEEE Internet Comput. Vol. 5, No. 2, pp. 18-26, 2001.

[48] J. ORTEGA e I. AGUILLO. Minería del uso de webs. El profesional de la información, Vol. 18, No. 1, pp. 20-26. Enero-febrero, 2009

[49] S. PAL, V. TALWAR, and P. MITRA. Web mining in soft computing framework: Relevance, state of the art and future directions. IEEE Transactions Neural Networks, Vol.13, No. 5, pp. 1163- 1177, 2002.

[50] F. PICAROUGNE, N. MONMARCHE, A. OLIVER, and G. VENTURINI. Web mining with a genetic algorithm. In Eleventh International World Wide Web Conference, Honolulu, Hawaii, pp.7-11, May 2002.

[51] V. REBOLLEDO. Plataforma para la extracción y almacenamiento del conocimiento extraído de los web data. Tesis (Ingeniero Civil Industrial y Magíster en Gestión de Operaciones). Universidad de Chile. 2008.

[52] A- RÉKA, H. JEONG and A. BARABÁSI. Diameter of the Word-Wide Web. Nature, Vol. 401, page 130, 9 September 1999.

[53] P.E. ROMÁN y J.D. VELÁSQUEZ. Cadenas de Markov para modelar la navegación del usuario Web: Inferencia estadística. XIV Congreso Latino- Iberoamericano de Investigación Operativa CLAIO 2008.

[54] F. ROMERO Y V. PARADA. Una máquina paralela para resolución de problemas de Optimización combinatoria. Dpto. Ingeniería Informática, Universidad de Santiago de Chile, 2008.

[55] S. ROSS, Probabilidades y estadística para ingenieros y científicos, McGraw-Hill. 2000.

[56] A. SCIME. Web Mining. Applications and Techniques. Idea Group Publishing, 2005

[57] J. SERRANO COBOS. Combinación de logs internos y externos en la predicción de estacionalidad de búsquedas para el rediseño de webs. El profesional de la información, Vol. 18, No. 1, pp. 11-19, Enero-febrero, 2009

[58] M. SHOKOUHI, P. CHUBAK, and Z. RAEESY. Enhancing focused crawling with genetic algorithms. In International Conference on Information Technology: Coding and Computing (ITCC'05). Vol. 2, pp 503-508, 2005.

[59] A. TORRENTS, J.FONOLLOSA and J.M. SALLÁN. Métodos cuantitativos de organización industrial II. UPC, 2002.

[60] J. D. VELÁSQUEZ. Personalizando la atención del cliente digital. Centro de gestión CEGES. Documentos de trabajo, Series de Gestión, 2007.

[61] J. D. VELASQUEZ and V. PALADE. Adaptive Web site: A Knowledge Extraction from Web Data Approach. IOS Press, 2008.

[62] J. D. VELÁSQUEZ, H. YASUDA, T. AOKI and R. WEBER. A new similarity measure to understand visitor behavior in a web site. IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization. Vol.E87-D, No.2, pages 389-396, February 2004

[63] C. VILLE, E. SALOMON, C. MARTIN, D. MARTIN, L. Berg y P. DAVIS. Biología. Interamericana McGraw Hill, Segunda Edición, 1992.

[64] L. WEN-LONG and L. YE-ZHENG. A Novel Website Structure Optimization Model for More Effective Web Navigation. Workshop on Knowledge Discovery and Data Mining, IEEE Computer Society, pp. 36-41, 2008

[65] D. WHITLEY. A genetic Algorithm Tutorial. Computer Science Department, Colorado State University. Statistics and Computing Vol. 4, pp. 65–85, 1994.

[66] Y. ZHOU, H. LEUNG y P. WINOTO. MNav: A Markov Model-Based Web Site Navigability Measure. Software Engineering, IEEE Transactions, Vol. 33, No. 12, pp. 869 - 890, Dec. 2007

[67] Charles Darwin y su libro: El Origen de las especies http://darwin-online.org.uk/content/frameset?itemID=F373&viewtype= text&pageseq=1 [Consulta: 23 de Agosto 2009]

[68] Fuentes de información genómica y herramientas bioinformáticas básicas http://bvs.isciii.es/bib-gen/Actividades/curso_virtual/Introduccion/ bioinformatica.htm [Consulta: 23 de Agosto 2009]

[69] Ley de Zipf
http://www.nslijgenetics.org/wli/zipf/ [Consulta: 23 de Agosto 2009]

[70] Sesionalización: "Accurate Analytics Require Cookies" By Bryan Eisenberg
http://www.clickz.com/3319891 [Consulta: 7 de Septiembre 2009]

[71] The MathWorks.
http://www.mathworks.com/products/gads/ [Consulta: 23 de Agosto 2009]

[72] The MathWorks. BioInformatic Toolbox Reference
http://www.mathworks.com/access/helpdesk/help/pdf_doc/bioinfo/bioinfo_ug.pdf
[Consulta: 7 de Septiembre 2009]

[73] Usabilidad
http://www.dcc.uchile.cl/~rbaeza/inf/usabilidad.html [Consulta: 7 de Septiembre 2009]

[74] Consorcio de la Web
http://www.w3c.es/ [Consulta: 23 de Agosto 2009]

[75] Web Semántica
http://www.w3c.es/Divulgacion/Guiasbreves/WebSemantica [Consulta: 23 de Agosto 2009]