Capítulo Segundo

Antecedentes

2. Antecedentes

Para comprender claramente la problemática que generan los conjuntos de nodos altamente interconectados para un sistema automático de análisis de redes, es necesario conocer los aspectos básicos de Teoría de Gráficas, la estructura y el funcionamiento fundamental de los sistemas de extracción y análisis de datos, el algoritmo ActiveRank, así como la naturaleza de la red a estudiar, en este caso la World Wide Web, y el segmento de dicha gráfica de nuestro interés, las páginas pornográficas.

A continuación se da una breve descripción de los conceptos necesarios para el claro entendimiento del desarrollo de esta tesis, invitando al lector a profundizar en cada uno de los temas aquí descritos, dado que la extensión de este capítulo está lejos de ser suficiente para cubrir todos los puntos necesarios.

2.1. Conceptos Básicos de Teoría de Gráficas

2.1.1. Definición de Teoría de Gráficas

La Teoría de Gráficas es la disciplina de las matemáticas que permite estudiar todos aquellos fenómenos o conjuntos que pueden ser representados a través de relaciones, implementando un lenguaje claro y homogéneo, así como la base de sus definiciones, propiedades y operaciones, para poder comprender su estructura y composición, así como para realizar y simplificar análisis específicos utilizando métodos y procesos bien establecidos. Entre los diferentes escenarios que se pueden abordar utilizando teoría de gráficas destacan los análisis de redes sociales y de información, eventos epidemiológicos y biológicos, redes de transporte y telecomunicaciones, entre muchos otros.

Una característica importante de las gráficas es su facilidad para ser descritas y operadas matricialmente, lo que combinado con sistemas de cálculo, permite realizar análisis de dimensiones extraordinarias, que antes eran simplemente imposibles. La representación gráfica de las redes también es extremadamente valiosa y simple; consiste en denotar los eventos u actores (nodos) como puntos, su interacción como líneas conectoras, y su relevancia a través de la longitud de esta última, lo que permite obtener un acercamiento visual claro de problemas multifactoriales.

2.1.2. Puentes de Köingsberg

¿Es posible hacer un recorrido por los siete puentes cruzando cada uno sólo una vez? (ver figura 2.1.1.). Esta pregunta común entre los habitantes de la ciudad de Köingsberg podría ser contestada fácilmente por pura observación, sin embargo, ¿Qué pasaría si tuviera más puentes?; conforme aumentáramos el número de posibilidades, un análisis no sistemático se volvería virtualmente imposible. Leonhard Euler resolvería el problema, dando inicio a lo que hoy conocemos como Teoría de Gráficas, al plantear el problema de los siete Puentes de Köingsberg de una manera completamente nueva.

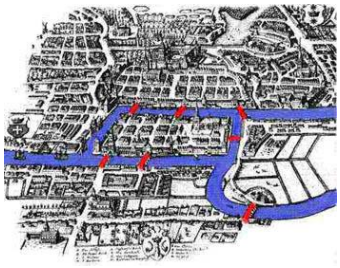


Figura 2.1.1. - Puentes de Köingsberg
[http://www.daviddarling.info/encyclopedia/B/Bridges_of_Konigsberg.html]

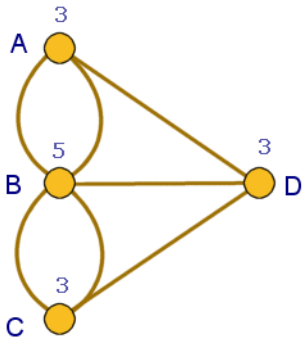


Figura 2.1.2. - Gráfica de los Puentes de Köingsberg
[http://www.infovis.net/imagenes/T1_N137_A6_KonigsGraph.gif]

Euler publicó en 1736 un artículo llamado *Solutio problematis ad geometriam situs pertinentis*, en español, S*olución de un problema relacionado a la geometría de posición*, en el cual señalaba que para que un viaje de ida y vuelta fuera posible, cada cuerpo de tierra debería tener un número par de puentes, o si el viaje inicia en un cuerpo y termina en otro, éstos dos podrían tener un número non de puentes, pero el resto de los cuerpos requeriría forzosamente un número par de puentes. Con lo anterior no sólo demostró que un viaje entre los puentes de Köingsberg era imposible, o el plantear una solución general para cualquier tipo de estructura de nodos interconectada por vínculos, sino que además dio inicio a un nuevo tipo de geometría en el cual la distancia[1] era un factor no significativo[2].

La figura 2.1.2. corresponde a la gráfica generada a partir del problema de los Puentes de Köingsberg; definimos a los cuerpos de tierra interconectados por puentes, como nodos relacionados por vínculos. Como se puede observar, resulta mucho más simple el análisis de la topología de la red; siguiendo con el ejemplo, observamos cuatro nodos, *A*, *B*, *C* y *D*, que representan las 4 porciones de tierra definidas en el mapa de la figura 2.1.1., así mismo, observamos 7 vínculos los cuales representan la abstracción de los puentes

1. Distancia entre dos puntos en un espacio vectorial euclidiano.
2. La distancia, entendida como una medida de cercanía o similitud entre nodos, cobra importancia al trabajar con redes pesadas, en las cuales los vínculos tienen un valor de peso que define su importancia con respecto a otros.

involucrados en dicho problema. Cada uno de los nodos posee tres vínculos, con lo que podemos afirmar que no cumple con el planteamiento general que Euler definiera en 1736.

El estudio de los problemas desde un enfoque de lógica y estructura relacional aumentó significativamente a partir de los inicios del Siglo XX, gracias a los avances matemáticos en el área actualmente conocida como Teoría de Gráficas.

2.1.3. Conceptos básicos

La siguiente información ha sido obtenida y desarrollada siguiendo lo establecido en el capítulo 2.2 del libro *Small Worlds: The Dynamics of Networks between Order and Randomness* de Duncan J. Watts.

NOTA: Notaciones diferentes a las siguientes pueden ser utilizadas siendo antes definidas.

Una gráfica *G* se compone de un conjunto no vacío de elementos llamados *vértices* o *nodos*, y una lista no ordenada de parejas de dichos elementos llamadas *aristas* o *vínculos*; al conjunto de vértices los denotaremos como *V(G)* y a la lista de aristas como *E(G)*; si decimos que *v* y *w* son vértices de *G*, *vw* sería la arista que conecta a dichos dos elementos, pudiendo ser esta relación unidireccional o bidireccional dependiendo del tipo de gráfica.

El número de vértices, es decir, el número de elementos del conjunto *V(G)* determina el *orden* de la gráfica, mientras que la dimensión de *E(G)* determina en sí la *dimensión* o *tamaño* de la gráfica. Los vértices en una gráfica pueden representar cualquier clase de elementos, como puede ser personas, animales, familias o comunidades, documentos, etc., mientras que las aristas, representan la relación entre ellos debido a pertenencia, interacción, alianza, entre otros; cabe destacar que la gráfica y su análisis se ve acotado por los elementos que la conforman.

Existen algunas características básicas aplicables a la mayoría de las gráficas las cuales son:

- *Directividad*: Los vínculos entre los nodos de la red pueden o no tener dirección, dependiendo de la naturaleza del evento o universo que representen.
- *Ponderación*: Una gráfica puede ser ponderada si las aristas entre sus vértices han sido valuadas de acuerdo a un criterio de importancia o cercanía.

- *Multiplicidad*: Se refiere a si existen múltiples aristas entre dos vértices; normalmente se trabaja únicamente con gráficas simples, donde sólo existe una arista por par de vértices, en la cual se denota su característica de directividad, y así mismo, si es el caso, con el peso correspondiente. Múltiples aristas entre la misma pareja de vértices pueden ser condensadas en una sola ponderando la importancia de ésta proporcionalmente al número de enlaces existentes.
- *Dispersión*: Para una gráfica no direccionada (o bidireccional si se quiere ver así), el valor de dimensión máximo *M* de *E(G)* corresponde a:

$$E(G)\max = M = \binom{n}{2} = \frac{n(n-1)}{2}$$

para una gráfica "totalmente interconectada", por lo tanto, la dispersión se da cuando:

$$M \ll \frac{n(n-1)}{2}$$

- *Conexa*: Si cualquier vértice puede ser alcanzado desde otro a partir de seguir un conjunto de aristas finito. En algunos casos se pueden obtener coeficientes de *conectividad* de acuerdo a la proporción de vértices disconexos.

Una *caminata* o *paseo* a través de la gráfica se refiere a la trayectoria de aristas que se debe de recorrer para pasar de un vértice a otro; el *diámetro* de la gráfica corresponde a la caminata más larga, es decir, al número de aristas entre los dos vértices más alejados, lo anterior para gráficas conexas, o subconjuntos conexos de aquellas disconexas.

Uno de los datos estadísticos más importantes de una gráfica es la *longitud característica* (*characteristic path lenght* en inglés), denotada como *L(G)*, se refiere a la longitud típica de una trayectoria entre dos vértices dentro de la gráfica y es la mediana de la media de las trayectorias más cortas para cada pareja de vértices; la distancia no corresponde a una medida euclidiana, sino al número de saltos que se tiene que dar para llegar de uno de los vértices al otro, esta distancia también es conocida como *distancia de Hamming* o *de cuadras*, por su semejanza a medir la distancia en una ciudad a partir de cuantas cuadras se tienen que recorrer.

El *vecindario* de un vértice *v* es la subgráfica que se compone de aquellos elementos relacionados a dicho vértice, excluyendo al vértice en cuestión, se denota como Γ_v. El estudio de los vértices adyacentes puede ser de gran relevancia en algoritmos de calificación de nodos, pues proveen información sobre la importancia de dicho nodo gracias al número y tipo de conexiones que tiene, así como de la posición como concentrador de peso dentro de la red.

Los vecindarios son útiles para la obtención de otra medida estadística extremadamente útil conocida como *coeficiente de agrupamiento* (*clustering coefficient* en inglés), el cual caracteriza que tanto los vértices adyacentes a *v* son adyacentes entre otros del subconjunto. De manera más precisa, sea γ el coeficiente de clustering o agrupamiento del vecindario Γ para un vértice cualquiera:

$$\gamma = \frac{|E(\Gamma)|}{\binom{k}{2}}$$

donde $\binom{k}{2}$ es el número total de posibles aristas en el subconjunto Γ, y $E(\Gamma)$ es la dimensión del subconjunto antes mencionado. A través de los coeficientes de clustering se pueden detectar aquellos nodos y regiones de la red que concentran la mayor cantidad de relaciones, lo que en términos de redes de información y gráficas sociales, representa a los elementos de mayor relevancia en términos de participación en la dinámica del sistema.

Como ya se ha mencionado, la teoría de gráficas permite estudiar de manera simplificada diversos problemas de gran escala y de una amplia variedad de temas; el análisis de redes de información es un escenario perfecto para explotar todas las capacidades de esta disciplina matemática, ha sido el objeto de estudio durante las últimas 2 décadas para el desarrollo de nuevas teorías, actualmente tiene una importancia similar a la que tuvo la investigación de redes sociales en la primera mitad del Siglo XX.

Internet y en particular la WWW, de los cuales se habla más adelante, presentan retos muy importantes en el desarrollo y aplicación de la teoría de gráficas en el estudio de redes debido a su dimensión y complejidad; puede ser utilizada para el análisis de la información, la conformación de su infraestructura, el desempeño de los sistemas, así como los usuarios y su interacción con el contenido disponible, permitiendo por primera vez estudiar el comportamiento global de millones de personas involucradas en un mismo ambiente.

2.2. World Wide Web

2.2.1. Internet

Internet inició con la red de computadoras ARPANET, construida en el periodo comprendido entre los años de 1969 y 1972, durante la Guerra Fría; fue creada para el cálculo de trayectorias de misiles balísticos y ataques nucleares, y para inicios de 1973, ya comprendía a más de 40 centros de cómputo interconectados entre sí. Ese mismo año, Vinton Cerf y Bob Kahn iniciaron el desarrollo de lo que posteriormente sería conocido como el protocolo TCP/IP, el cual consiste en un protocolo en el nivel de la capa de transporte (en términos del modelo OSI) para poder transmitir de manera eficiente información a través de una red de computadoras, lo cual a su vez, facilita la incorporación de nuevos equipos más rápidamente al ser adoptado posterior como un protocolo estándar. El funcionamiento básico del protocolo TCP/IP, es el envío de paquetes de información identificando el destino a través de una dirección única; la capa TCP recibe el mensaje y construye paquetes de una longitud fija, incorporando una serie de encabezados destinados a la identificación del destino, información de estado y datos sobre la codificación del mensaje (con detección y corrección de errores), la cual es leída por otros dispositivos, y dependiendo de la estructura de la red, eventualmente es dirigida y recibida correctamente.

En el año de 1982, TCP/IP fue adoptado como el protocolo estándar de comunicación de la red Internet, lo que permitió la unificación e interconexión de una gran cantidad de redes independientes entre sí a nivel global. La razón de que TCP/IP fuera adoptado como la norma de interconexión de Internet fue su alta eficiencia y fiabilidad; durante el periodo entre su invención en 1973 y 1982, Internet continuaba funcionando bajo la red ARPANET, lo que eventualmente se convirtió en un escenario caótico.

2.2.2. Inicio y fundamentos de la World Wide Web

El concepto de World Wide Web fue desarrollado en 1989 por Tim Berners-Lee, en el Laboratorio Europeo de Investigación Nuclear (CERN), quien desarrollo el primer servidor web (httpd), con el respectivo cliente (navegador y editor) y lenguaje de hipertexto (HTML). La idea central consistía en romper con el esquema convencional de organización jerárquica de la información, planteando un modelo en el cual una aplicación cliente solicitaba a un servidor, a través de una red utilizando un localizador (URL – Uniform Resource Locator), documentos estándar, que al ser interpretados, permitieran elaborar “páginas” con formato, insertando texto y gráficos. El siguiente punto fundamental que Berners-Lee planteó, fue la interconexión de documentos a través de “hipervínculos”, enlaces directos entre páginas, los cuales dieron lugar a la red misma

que actualmente conocemos simplemente con el término “web”.

Mosaic, desarrollado por Mark Andreesen en 1993, consistía en una interface gráfica para la visualización e interacción (“click”) con páginas web; para 1994, cambió su nombre a Netscape Navigator, podía ser utilizado en los tres principales sistemas operativos, y se convirtió en la aplicación líder para la utilización de la WWW. En 1995, Microsoft lanzó su primera versión comercial de Internet Explorer, compitiendo de manera directa contra Netscape. Gracias a los navegadores, la WWW se convirtió en un medio altamente eficiente para el acceso y manejo de información, dando lugar a una amplia variedad de aplicaciones orientadas a la comunicación, y haciendo de dicha red, el sistema más grande del mundo.

2.2.3. Hipervínculos

Los hipervínculos son enlaces entre dos diferentes páginas que permiten ir de una a otra durante la navegación. Inicialmente, eran el único medio para llegar de a una página desconocida, de ahí la importancia en la etapa inicial de la WWW de sistemas de directorio como Yahoo o Excite, los cuales consistían en una gran base de datos de hipervínculos, creada manualmente.

Dado que la interconexión de páginas de Internet conforman una gráfica en sí, todas sus reglas, métodos de análisis y características generales son aplicables, por lo tanto, podemos encontrar dos casos de interconexión, la unidireccional y la bidireccional; si un hipervínculo apunta de la página *A* a la página *B*, pero no de la *B* a la *A* y viceversa, podemos decir que se trata de un enlace unidireccional, mientras que si sí existe, entonces tendremos un enlace bidireccional.

La figura 2.2.1. es una representación de los 6 diferentes tipos de subgráficas que conforman la red WWW desde el punto de vista de Albert-Laslo Barabási, en su libro Linked.

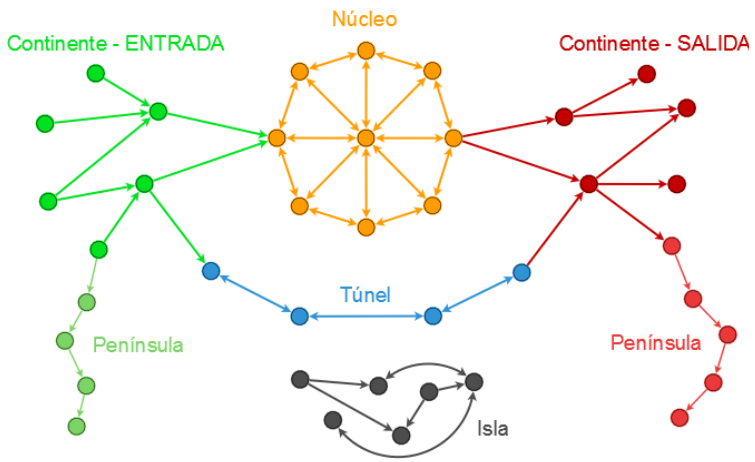


Figura 2.2.1. - Conceptualización de la estructura de la World Wide Web
[BARABÁSI; Linked; p. 166.]

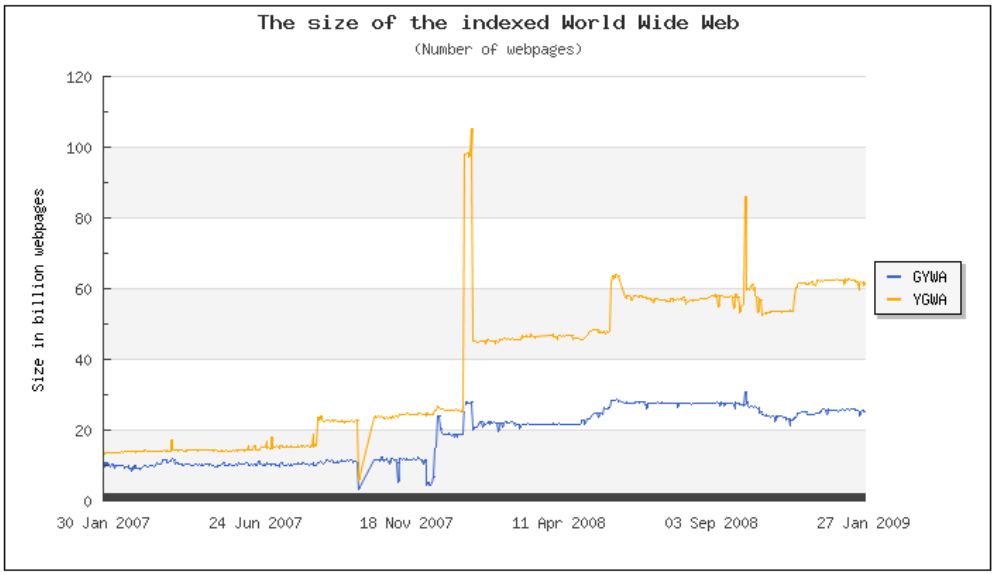
Se puede entender al núcleo como aquellas páginas de Internet que pueden ser accesadas desde muchas otras, las cuales a su vez se encuentran altamente interconectadas entre sí, es decir, mantienen un gran número de vínculos bidireccionales. Los continentes, son extensos grupos de páginas que se encuentran organizadas jerárquicamente, e interconectadas al núcleo gracias a una cantidad relativamente pequeña de elementos, existen continentes cuyo flujo natural va hacia el núcleo, y otros en los cuales, parte del mismo hacia regiones dispersas, se denominan continentes de entrada y de salida respectivamente. Dentro de los continentes, podemos encontrar caminos o rutas hacia páginas poco interconectadas, que se construyen a partir de unos vínculos, a estas regiones se les denomina penínsulas, al ser segmentos del continente poco relacionados al mismo, estructuralmente hablando. Cuando una península interconecta a dos continentes, se le denomina túnel o tubo, ya que es una posible ruta entre los dos continentes, no perteneciente al núcleo; la detección de dichas estructuras suele ser compleja, ya que es necesario contar con una cantidad suficientemente grande de documentos para diferenciar al núcleo y a los continentes, lo que en términos de la WWW representa millones de documentos previamente indexados, y después, la capacidad de cómputo para realizar el análisis estructural. Por último encontramos a las islas, subgráficas compuestas por un bajo número de elementos, las cuales se encuentran totalmente desconectadas del resto del conjunto, contemplando a páginas totalmente disconexas, como podría ser el caso de una página personal en un servidor privado.

2.2.4. Dimensiones de la WWW

El tamaño y crecimiento de la World Wide Web son los más grandes de cualquier sistema construido por el hombre; el número de páginas indexadas supera los 60 billones (60,000 millones), y su crecimiento estimado asciende a un 2% mensual (1,200 millones), sin embargo, existen autores que sostienen que no se conoce más del 15% del total de documentos disponibles en la WWW, y más aún, que su expansión supera la capacidad máxima de indexación (y reindexación) de los sistemas de búsqueda más importantes del mundo.

La siguiente estimación de la dimensión de la WWW indexada es trabajo de Maurice de Kunder, quién en su página *http://www.worldwidewebsize.com/* mantiene un constante monitoreo del número de páginas indexadas por los principales sistemas de búsqueda e indexación, quién utiliza la siguiente metodología para la obtención de sus resultados numéricos y gráficos.

Kunder obtiene la dimensión de la WWW conocida a través de los resultados obtenidos a partir de 4 buscadores (Google.com, Yahoo.com, MSN.com, Ask.com); busca una palabra aleatoriamente seleccionada de una base de datos en cada uno de los sistemas, adiciona los conjuntos de resultados removiendo las repeticiones y hace un conteo, este resultado es multiplicado por el factor estadístico de presencia de la palabra en una muestra de documentos, es decir, en que porcentaje de una muestra de documentos aparece dicha palabra, y así obtiene la dimensión de la web conocida según dicha palabra. Este proceso se repite 50 veces diariamente con distintas palabras, promediando los valores obtenidos, para determinar un valor subestimado del tamaño de la World Wide Web. La gráfica 2.2.2. muestra el crecimiento de la web explorada en los últimos dos años; gráficas complementarias se pueden encontrar en el Apéndice A.



GYWA = Sorted on Google, Yahoo!, Windows Live Search (Msn Search) and Ask
YGWA = Sorted on Yahoo!, Google, Windows Live Search (Msn Search) and Ask

Gráfica 2.2.2. - Crecimiento de la WWW en los últimos 2 años, a partir de mediciones realizadas en los 4 motores de búsqueda más importantes.
[http://www.worldwidewebsize.com]

2.2.5. Análisis de la WWW

El análisis e indexación de toda la información pública disponible a través de la WWW son teórica y prácticamente imposibles bajo los esquemas tecnológicos y metodológicos que se han utilizado hasta el momento, sin embargo, las líneas de investigación para superar el reto de conocer la mayor cantidad de información posible son dignos de estudio; a continuación se describen algunos de los principales puntos que definen la tendencia en el diseño y desarrollo de sistemas de análisis y acceso a la información web, tema que a su vez se desarrolla de manera más particular en la sección 2.3. de este documento.

La primer característica de la WWW que dificulta su análisis es su naturaleza no centralizada, ya que no permite la fiscalización de información, es decir, tener un control estricto de qué, quién, cuándo, cómo y dónde se colocó cierto artículo en la red; desde sus inicios, su planteamiento libre y democrático, la facilidad de publicación de contenido, responsable de su éxito, tuvo también como consecuencia la individualización de los canales generadores del mismo, lo que produjo diferentes efectos nunca antes observados en otros sistemas de ingeniería, como son la diversidad de ubicaciones y fuentes de información, rutas de interconexión, y sobre todo, la velocidad del crecimiento de la red.

Es a través de los vínculos entre páginas web que los sistemas automáticos pueden ir navegando entre ellas, sin embargo, dada la estructura de la WWW representada en la figura 2.2.1., no todas están conectadas entre sí, en realidad, conforman secciones casi independientes, relacionadas por páginas que concentran la mayor cantidad de rutas, por lo que para tener conocimiento de la mayor cantidad de páginas posibles, sería necesario ubicar estos clusters de vínculos dentro de todas las subgráficas de las WWW, lo que resulta sumamente complejo realizar manualmente. En los inicios de Yahoo, el primer sistema de directorio en Internet, la manera de lograr indexar una página era cuando su dueño solicitaba su adhesión al sistema, y luego de un proceso de clasificación humano, quedaba lista para ser accesada a través del sistema; este modelo es importante por dos razones, primero, si todas las páginas fueran declaradas por sus creadores, sería posible tener conocimiento de su existencia, y en segundo término, al ser clasificadas por una persona, el nivel de eficiencia en los sistemas de búsqueda también sería extremadamente alto. Lo anterior es imposible por el volumen de documentos que se generan día con día en Internet, los costos que tendría el equipo de personas clasificando información serían simplemente insostenibles, sin embargo, algunas ideas fueron claves al ser integradas en los sistemas de análisis de información.

Una posibilidad de facilitar el descubrimiento de subgráficas de la red es mediante sistemas de coordinación entre servidores; sistemas automáticos que publican una lista de los servidores en línea y que a su vez, cada servidor tiene una lista pública del contenido que aloja, similar a los *DNS* (Domain Name Server), que vinculan nombres de dominio con direcciones IP. El modelo anterior es utilizado en redes acotadas como intranets o redes de servidores de archivos, y brinda las bases de algunos de los sistemas de gestión documental, así como los modelos *Peer to Peer*; cabe destacar que aún así, existen servidores en línea ocultos, de uso privado o simplemente no accesibles, con lo que obtenemos la conclusión parcial de que es imposible conocer el total de la información disponible en web.

Incluso sin intentar analizar el total de la información, existen retos de ingeniería que surgen casi de inmediato cuando se trata de recopilar información de Internet. En base a mi experiencia profesional y a los primeros sistemas de *crawling* que desarrollé en 2005, el primer problema que se presenta, es el manejo de las bases de datos. Con una PC de características convencionales, y un enlace asíncrono de 2 [Mbps], un sistema de análisis básico, tiene la capacidad de indexar en promedio 250 [ppm] (páginas por minuto), a ese ritmo, son 360,000 páginas en un día, y si cada una tuviera 10 vínculos, la lista de espera de páginas por analizar, en 24 horas, sería de 3,600,000 vínculos, lo que te da suficiente trabajo para 10 días de análisis. Sin embargo, el servidor de base de datos comienza a presentar fallas mucho antes, debido al número de registros y saturación de memoria de consulta, lo que hace necesario la utilización de bases de datos distribuidas y procesos de optimización casi inmediatamente después de haber comenzado nuestro proyecto. Una vez resuelto el problema

de base de datos, el siguiente cuello de botella se presenta en la capacidad del procesador (CPU) y memoria de acceso aleatorio (RAM) del servidor, el cual se ve afectado por la necesidad de procesar el contenido de las diferentes páginas, las cuales son alimentadas de manera constante por programas de recolección operando en paralelo; la necesidad de poder agilizar el procesamiento de toda la información que está siendo recolectada requiere la inserción de sistemas de cómputo distribuido, los cuales sea en tiempo real o en tiempo discreto, deberán filtrar, analizar, estructurar y almacenar al mismo ritmo que los sistemas de minería de datos.

Una vez superados los retos de base de datos y procesamiento, el siguiente problema se presenta en la capacidad del enlace utilizado, en el cual se requiere mayor velocidad de transmisión para poder incrementar el número de páginas descargadas por unidad de tiempo; el análisis que se requiere realizar para comprender el tamaño de enlace requerido es igual al de cualquier otro caso en telecomunicaciones, y la relación entre el crecimiento del enlace y el crecimiento de la capacidad de descarga de páginas se mantiene constante hasta el punto en el que de nueva cuenta, la base de datos, el procesamiento y eventualmente el almacenamiento, comienzan a ser insuficientes en el orden establecido.

Por todo lo anterior resulta evidente que se debe evaluar integralmente la relación entre cada uno de los componentes del sistema para poder crear un sistema lo suficientemente escalable y estable para el procesamiento de información, sin embargo, el círculo de crecimiento planteado anteriormente, conlleva a más retos, como son la cantidad de energía eléctrica requerida por dichos sistemas, el costo de la infraestructura y el espacio, la eficiencia eléctrica de los equipos y la necesidad de sistemas de enfriamiento altamente eficientes; todo comienza a verse como proyectos de ingeniería de gran escala, pero las preguntas que surgen son *¿Qué tan grandes son dichos sistemas?* y *¿Realmente es necesario conocer toda la WWW?*.

Para conocer la dimensión de los sistemas más grandes utilizados para analizar la WWW, podemos tomar como referencia a Google, y describir de manera general las características públicas de los centros de datos utilizados para soportar su motor de búsqueda y aplicaciones. La compañía antes mencionada cuenta con más de 30 centros de datos alrededor del mundo, pero podríamos concentrarnos en los 15 ubicados en los Estados Unidos de América con los cuales desarrolla la mayor parte de su análisis; dado que toda la información referente a la capacidad de cómputo instalada se maneja con gran secrecía, muchos de los datos a continuación descritos son aproximaciones realizadas en base al tamaño de las instalaciones, dimensión de los sistemas de enfriamiento así como la capacidad de alimentación eléctrica por la revista *Harpers*. Cada uno de ellos está equipado con tomas eléctricas que van desde los 50 [MW] hasta los 250 [MW] y su costo unitario se estima cerca de los 600 millones de dólares americanos; el costo de operación del conjunto de centros de datos reportado en el año 2007 fue de 2,400 millones de dólares americanos; sus sistemas de enfriamiento requieren un alto volumen de agua, por lo que los más nuevos han sido construidos cerca de ríos y lagos. Todo lo anterior,

permite que Google actualice el contenido de toda su base de datos en aproximadamente 15 días, sin embargo, esto sólo representa alrededor de 15% de la WWW, que a su vez, parece resultar suficiente para cubrir nuestra necesidad de acceso a la información, lo que nos lleva a pensar directamente en sistemas que no se basen en cubrir toda la red, sino únicamente aquellas secciones que en conjunto, aporten la mayor cantidad de información útil para ciertos grupos sociales.

Wikipedia, una base de contenido enciclopédico generado de manera colaborativa entre todos sus usuarios, es un buen ejemplo de aquellas plataformas que sin indexar información de la WWW, tienen un alto valor en términos informáticos, ya que concentran un alto volumen de información útil para la mayoría de las personas; algunos sistemas automáticos de indexación y análisis se basan en explotar grandes fuentes de información confiable para integrar una amplia red de información, de alto valor, generada con una parte muy pequeña de la WWW; además de que es imposible, no es necesario indexar toda la información de Internet, ya que para resolver un problema en específico, no necesitamos toda la información del universo, sino sólo el conjunto que es relevante para su solución.

2.3. Sistemas de Análisis de Redes de Información

Los sistemas de análisis de redes de información son programas de cómputo que permiten realizar de manera total o parcialmente automática la obtención y procesamiento de grandes volúmenes de información; normalmente son utilizados como plataformas primarias de sistemas de acceso a la información como pueden ser sistemas de búsqueda, indicadores de precio, sistemas de compra-venta automáticos, entre otros.

La arquitectura de los sistemas de análisis o *crawlers* descrita en esta tesis se centra en aquellos que operan recolectando información de Internet a través de seguir los hipervínculos contenidos en cada una de las páginas previamente procesadas; a continuación se describen sus principales características.

2.3.1. Estructura General

Los criterios de operación de los sistemas de indexación y análisis se definen por los objetivos que se quieren cubrir al crear y explotar una base de información particular, ya sea proveniente de la WWW, bases de datos privadas o una mezcla de diversas fuentes, estos definen el rango y profundidad de búsqueda, consolidación de bases de datos, análisis particulares así como los sistemas subsecuentes para su aprovechamiento y mantenimiento, entre otros; todo lo anterior resulta fundamental para poder llevar a cabo un correcto diseño y desarrollo de plataformas de análisis de redes de información.

El funcionamiento básico de un crawler es conceptualmente simple y consiste en analizar una página web almacenada en una lista de espera, analizar su código HTML para obtener las ligas que contiene a otros sitios, y agregarlas a la lista de espera, para volver a repetir indefinidamente el proceso. El diagrama de bloques mostrado en la figura 2.3.1. describe la estructura general de un sistema de indexación y procesamiento de información textual de Internet.

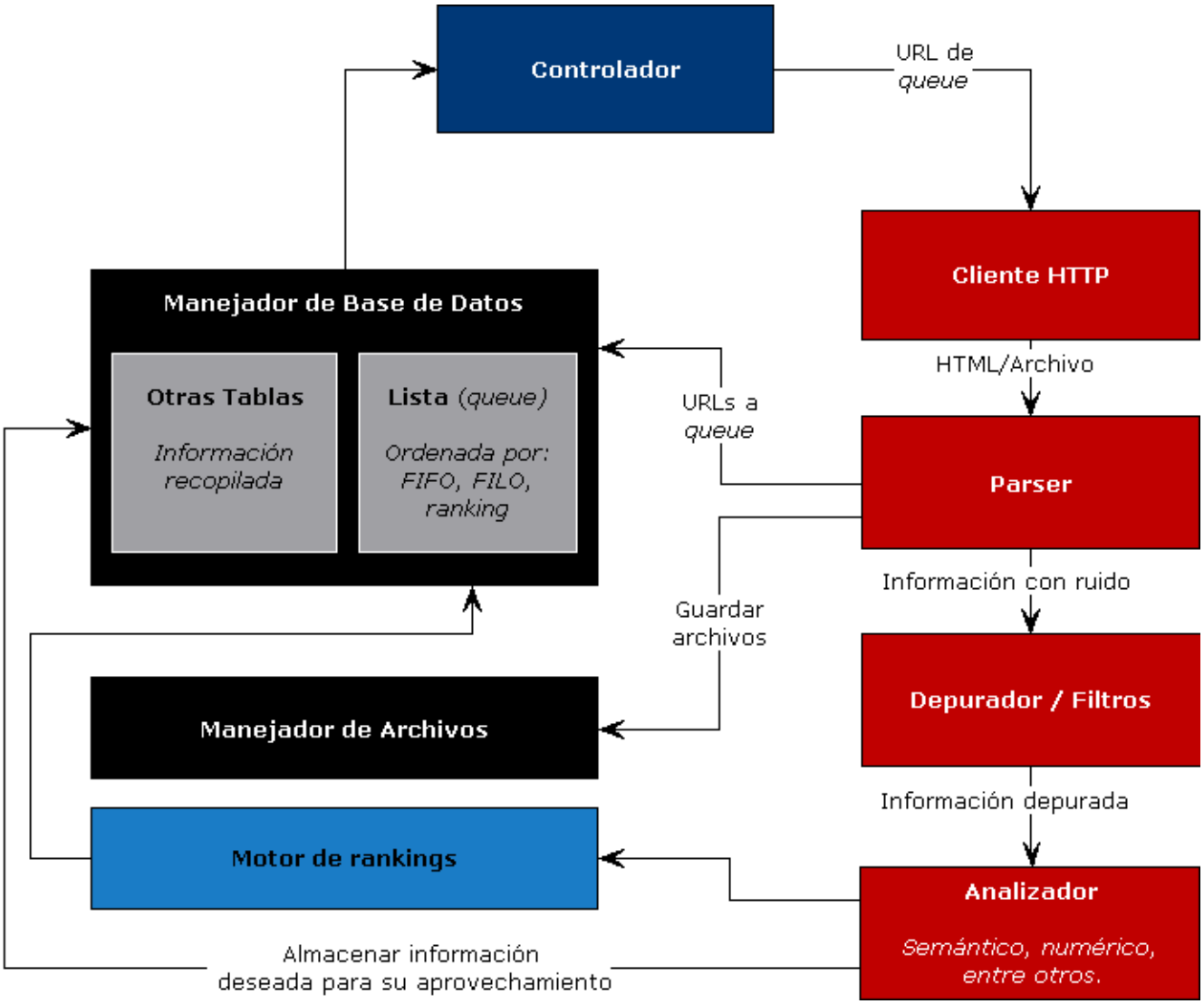


Figura 2.3.1 – Estructura básica de un crawler

- *Queue*: Es una lista que contiene URLs en estado de espera para ser analizados. Existen múltiples criterios para determinar el orden de salida de dichos registros y modifican de manera significativa el rendimiento del sistema, entre ellos destacan FIFO, FILO o LIFO, y por ranking.
    + *FIFO*: Del acrónimo en inglés *First In First Out*, el orden de salida de los registros es el mismo que el orden de entrada de los mismos.
    + *FILO o LIFO*: De los acrónimos en inglés *First In Last Out* o *Last In First Out* respectivamente, también conocida como *Stack*, el orden de salida de los registros es inverso al orden de entrada de los mismos. Tanto FIFO como FILO tienen desempeños similares en el análisis de una muestra aleatoria de páginas de Internet.
    + *Por ranking*: En este caso se utiliza un criterio basado en una calificación determinada normalmente a partir de las características de las páginas donde se encontró dicho URL, dado que no se cuenta información de ésta en el momento que se saca de la queue. Por ejemplo, existe mayor probabilidad de que una página obtenida varias veces sea más importante que una que sólo fue encontrada en una ocasión, u otra opción es analizar primero aquellas provenientes de páginas más importantes que de aquellas menos relevantes para el análisis en cuestión. Una correcta utilización de selección por rankings puede aumentar significativamente el desempeño del analizador.
- *Manejador de Base de Datos*: Es el sistema que se conecta con el motor de base de datos utilizado para el acceso, lectura y escritura de información en las tablas correspondientes. El punto fundamental del motor y del manejador de base de datos es que deben ser plataformas extremadamente eficientes con la utilización de recursos, y tener la capacidad de operar con tablas extremadamente largas; dado que la cantidad de registros es muy grande, siempre será preferible tener más tablas con menos números de columnas que pocas tablas muy anchas, lo anterior debido a que el tamaño del buffer del motor de base de datos es limitado, caben mayor número de resultados mientras más delgados son.
- *Manejador de Archivos*: Se encarga de la escritura de archivos e información recopilada a nivel sistema de archivos, debe tener la capacidad de almacenar correcta y eficientemente una gran cantidad de documentos de tamaño variable. Un punto fundamental es que si se desea almacenar gran parte de la información recopilada, serán necesarios sistemas dedicados de almacenamiento masivo debido al volumen.
- *Controlador*: Es el módulo encargado de gestionar todos los procesos del sistema de análisis; su

función más importante es la de sincronizar los diferentes procesos y mantener el control de hilos en analizadores con capacidad de llevar procesos multihilos. En caso de existir módulos adicionales para el control estadístico y la obtención de indicadores de desempeño, así como interfases de control, y en general cualquier método de lectura y escritura de datos para utilizar el crawler como sistema secundario en otra plataforma, estaría conceptualmente dentro del controlador.

- *Cliente HTTP*: El Cliente HTTP es el subsistema encargado de establecer la conexión entre el servidor host y remoto, y descargar el contenido necesario correspondiente a la URL que está siendo analizada. Su principal característica es que debe administrar de manera eficiente el canal de comunicación disponible para maximizar la transferencia de información a través del enlace; los más avanzados procuran trabajar a bajo nivel enviando y recibiendo paquetes directamente a través de conexiones abiertas por *sockets*.
- *Parser*: El sistema de parseo es el encargado de desglosar e interpretar correctamente las diferentes secciones del documento obtenido por el Cliente HTTP. Normalmente separa las estructuras de HTML del resto del texto, de donde obtiene los URLs contenidos en los hipervínculos, y transfiere el resto de la información a la siguiente etapa, lo que se puede considerar como la primara etapa de filtrado. El parser es normalmente el módulo encargado de la corrección y homogenización de las URL obtenidas, así como de su inserción dentro de la queue para ser analizadas posteriormente.
- *Depurador/Filtros*: Es el módulo encargado de implementar múltiples procesos de filtrado y corrección de información, como ejemplo podemos encontrar la remoción de palabras comunes o estructuras regulares predefinidas, sustitución de errores ortográficos, entre otros; esta etapa es muy importante gracias a que disminuye el procesamiento de información no relevante, así como permite obtener mejores resultados en la mayoría de los procesos de análisis estadísticos de estructuras semánticas.
- *Analizador*: El módulo de análisis es en realidad el centro de todo el sistema, debido a que realiza la obtención y preprocesamiento de la información de interés para el usuario o sistema maestro; puede incluir una amplia gama de análisis a través de algoritmos semánticos, numéricos, gráficos, entre otros. Dependiendo de la arquitectura de la plataforma, en algunas ocasiones los módulos de análisis o procesamiento de información se encuentran como sistemas independientes a los de indexación de contenido, en estos escenarios se considera que existe una actividad de minería de datos y posteriormente una de análisis de información; la única

diferencia radica en los tiempos que se guardan entre cada una de las etapas, y en algunas ocasiones, cuando sólo se puede indexar la información durante un periodo determinado de tiempo, es mejor tener un sistema cuyo objetivo sea descargar la mayor cantidad de información en el menor tiempo posible, y posteriormente, cuando no existe la posibilidad de realizar minería, realizar el resto de los procesos de análisis.

- *Motor de rankings*: El motor de rankings es un bloque opcional de la estructura genérica de un analizador de redes de información, y consiste en un sistema externo el cual es alimentado por datos resultantes de los análisis, y que a su vez retroalimenta al sistema brindándole capacidad de mejorar algunos puntos como el orden de alimentación de registros de la queue como ya se mencionó anteriormente. A su vez, algunos motores de ranking pueden ser aprovechados por otros sistemas, como en el caso de ActiveRank, donde todo el sistema de crawling tiene como finalidad alimentar a dicho motor para que este brinde una funcionalidad adicional en otras plataformas como pueden ser sistemas de búsqueda o autoorganizadores de información.

En el Apéndice B de este documento podemos encontrar el diagrama de flujo de un sistema básico de análisis de redes de información, el cual comprende los pasos y validaciones más importantes en el proceso de exploración, indexación y procesamiento.

Entre otros puntos importantes en el diseño, desarrollo e implementación de sistemas de análisis de redes de información, y en concreto la WWW, se encuentra el criterio para la selección de los puntos iniciales a partir de los cuales se iniciará la exploración de la red; para sistemas focalizados, lo más importante es definir claramente el universo de páginas que se desea atacar, como por ejemplo, aquellas que se encuentren dentro de un dominio determinado, o que pertenezcan a una categoría específica (ej. *.edu*, *.com*, etc.), esto resulta fundamental debido a la velocidad con la que se puede salir de dichos límites si nos dedicamos a seguir a todos los posibles; en muchos casos, resulta extremadamente útil el aprovechar sistemas externos para determinar el conjunto de páginas de interés, como pueden ser sistemas de búsqueda, analizando aquellos documentos resultantes de una consulta a sus motores; dentro del diseño de sistemas de *crawling*, está el lograr altos grados de automatización y estabilidad, ya que involucran una muy alta cantidad de errores de tipo sintáctico inmersos en la información obtenida, y que resultan en muchos casos imposibles de corregir.

La simplicidad, eficiencia y elegancia en el diseño de sistemas de análisis de redes de información está directamente relacionado a la experiencia, y en términos personales, 5 años aún resultan pocos para la cantidad de retos a los que uno se enfrenta al desarrollar estos proyectos.

2.3.2. Integración y procesamiento de información

La principal función de un sistema de crawling es recopilar información determinada de una amplia variedad de fuentes, ya sea que se trate de explorar de manera abierta la WWW o que se esté analizando una porción muy específica de la misma, sin embargo, dicho contenido no mantiene una estructura homogénea en ninguno de los dos escenarios, por lo que es necesario realizar procesos que logren consolidar dicho contenido. Los procesos de integración de información son aquellos cuya función es detectar duplicidad, errores de parseo, homogenizar formatos, y muchas veces realizar preprocesamiento de información para prepararla para su explotación o consumo final; esta etapa es fundamental para lograr altos niveles de eficiencia (y así poder reducir costos de operación) y normalmente resulta ser tan compleja como las etapas de exploración y minería.

Un ejemplo muy claro son los sistemas automáticos de compra-venta los cuales recopilan catálogos de una variedad de plataformas de comercio electrónico, parsean la información de precio e identifican los productos, y si se cumple cierta regla de operación, como por ejemplo, que su costo esté por debajo de cierto límite, notifican a un operador humano para completar la transacción, e incluso, cuando los mercados lo permiten, realizan la operación de manera automática; entre los principales retos de integración de información a los que se ven expuestos esta clase de sistemas están el identificar que dos registros diferentes pueden tratarse del mismo producto, deben de identificar el tipo de moneda y convertir al tipo de cambio correcto para homogenizar la divisa y poder compararla contra un patrón, entre muchos otros aspectos que son fundamentales para el correcto funcionamiento del sistema. En muchas ocasiones, más en sistemas cuyo objetivo es el procesamiento de información textual como son documentos escritos o páginas de web, la integración de datos está estrechamente ligada al sistema de rankings con el cual opera la plataforma de análisis de redes de información, como es el caso de esta tesis; operan comparando el documento analizado contra una referencia establecida, y si la calificación de relación o ranking cumple una regla establecida se decide de manera automática la clasificación del registro y se procede a su procesamiento de manera automática.

Como procesamiento de información se engloba cualquier operación que se realice con la información disponible, y su variedad es tan amplia como nuestra mente y los recursos informáticos disponibles nos lo permitan.

Además del procesamiento e integración de información propios para lograr aprovechar el contenido recopilado, en la mayoría de los sistemas de crawling existen procesos automáticos y semiautomáticos de entrenamiento que son otra parte fundamental de las plataformas en cuestión; entre los principales destacan los procesos de *aprendizaje supervisado*, *aprendizaje semisupervisado* y *aprendizaje por máxima entropía*, de los cuales se puede obtener información en literatura técnica sobre sistemas de minería de datos.

2.3.3. Infraestructura de los sistemas de análisis de información

Retomando la conversación realizada en el capítulo 2.2.5. *Análisis de la WWW* de esta tesis, el principal reto tras el desarrollo de un crawler es lograr hacerlo estable y rentable, esto último depende directamente de la cantidad de infraestructura necesaria para abordar el problema, y normalmente está limitado únicamente por el presupuesto con el que se cuenta. La figura 2.3.3. representa un modelo base para la plataforma de un crawler genérico de mediano alcance.

Los principales puntos de escalabilidad que se deben tener en cuenta van directamente relacionados a la arquitectura distribuida del sistema de crawling, y normalmente existen tres puntos críticos a considerar, los cuales son el ancho de banda y velocidad de transmisión de datos disponibles para las conexiones a páginas web, la capacidad de procesamiento y la velocidad de acceso y manejo de registros en los sistemas de base de datos; existen una amplia variedad de algoritmos de distribución de carga que pueden ser aplicados, ya sean plataformas de procesamiento en paralelo, sistemas distribuidos o gestores de balanceo de carga; a continuación se describen los principales puntos que deben ser cuidados.

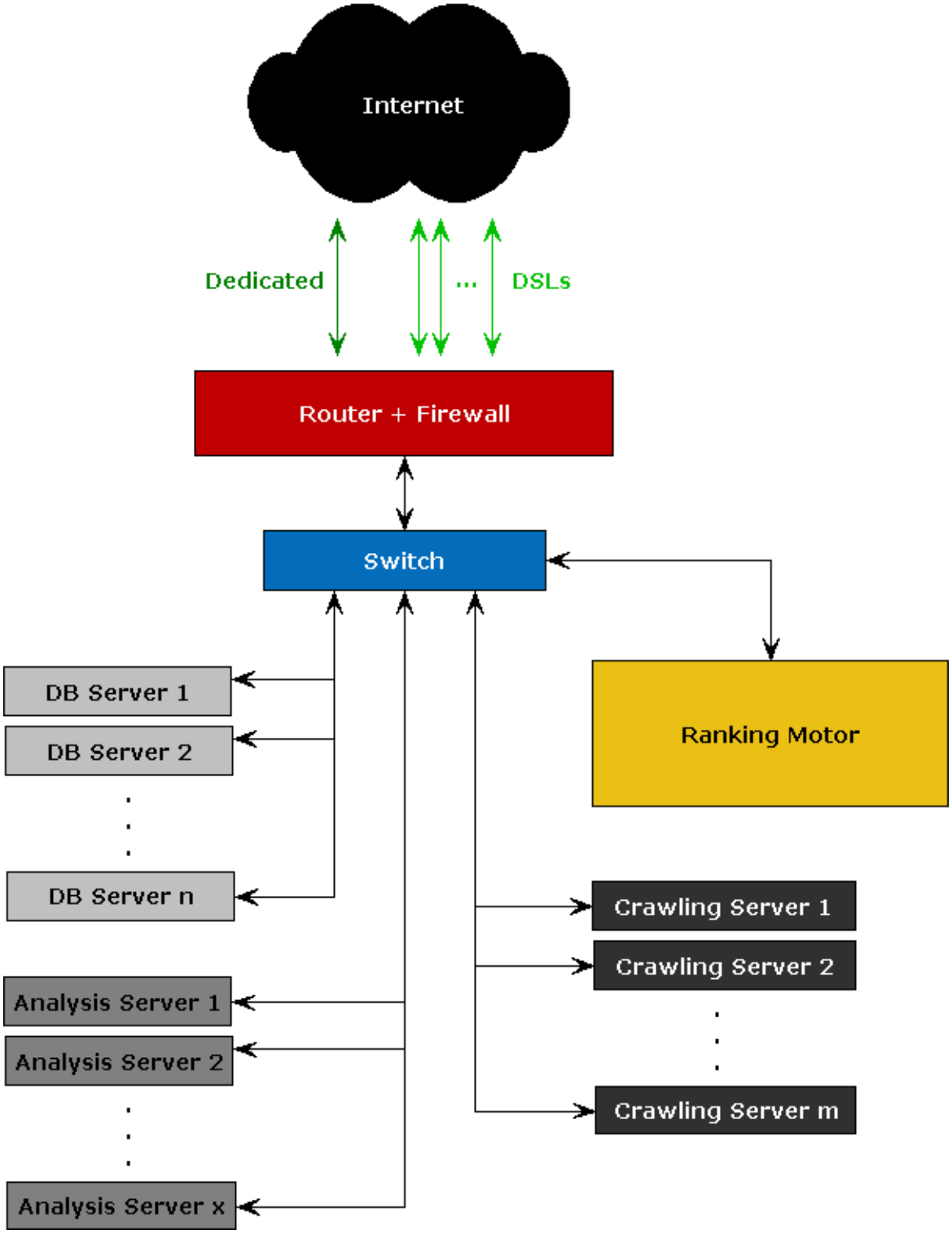


Figura 2.3.3 – Plataforma base para un sistema de crawling

Para los sistemas de base de datos, algunos de los factores más importantes que se deben considerar son los siguientes:

- *Índices y estructura de tablas*: El correcto diseño de la estructura de la base de datos es crucial para lograr los mayores índices posibles de eficiencia y velocidad. En sistemas de crawling, el número de registros normalmente es extremadamente alto, por lo que tablas delgadas (pocas columnas) son preferibles ya que caben más registros en el buffer después de haber realizado una consulta, así mismo, la utilización de tablas relacionales, que básicamente relacionan identificadores numéricos son preferibles a aquellas que relacionen tipos de datos de mayor tamaño por la misma razón. El que una tabla esté indexada quiere decir que el motor ha creado índices sobre todos los datos de alguna columna específica a partir de la cual se desarrollan búsquedas, lo que hacen es ordenar sobre estructuras de árbol tipo B los datos contenidos en la tabla, lo que agiliza de manera significativamente todas las búsquedas y consultas de información en la tabla.
- *Sistemas distribuidos*: Las plataformas de bases de datos distribuidas consisten en sistemas que dividen las tablas entre un número dado de servidores coordinados por un motor central, los cuales permiten operar de manera simultánea secciones de la base de datos, incrementando la capacidad de manejo de información, así como incorporar redundancia, entre otras características. La distribución de bases de datos puede elevar considerablemente el costo de infraestructura así como la complejidad de la programación de los sistemas; pueden ser utilizados sistemas existentes y genéricos que incorporan técnicas ampliamente desarrolladas, o también se pueden generar subsistemas de control de bases de datos al interior de nuestro sistema, que administren y regulen la carga de diferentes servidores de base de datos sin recurrir directamente a un sistema de base de datos distribuida; el caso anterior funciona bien cuando las diferentes secciones de la base de datos se pueden trabajar de manera total o parcialmente independiente.
    + *Subsistemas gestores de carga:* Un subsistema de gestión de carga puede ser aplicado a cualquier etapa de nuestro sistema de análisis, en bases de datos, son sistemas que determinan a qué servidor enviar cierta consulta a través de una regla, como podría ser dividir en dos servidores todas las consultas a partir de la letra con la que inicia el

dominio de la página en cuestión, es decir, en el primer servidor tendríamos toda la información de dominios que iniciaran con las letras A-L y en el segundo todos aquellos de la M-Z; evidentemente la complejidad de estos sistemas es muy superior, y deben ser diseñados para satisfacer las necesidades específicas de nuestros sistemas, pero pueden brindar buenos resultados a un costo menor en sistemas de mediana escala.

Los enlaces entre los diferentes dispositivos de nuestro sistema (hablando en términos de infraestructura) y con los servidores externos que contienen la información que deseamos obtener a través de conexiones HTTP, FTP, entre otras, son uno de los puntos más importantes e impactan de manera directa tanto el desempeño de la aplicación como el costo de operación de los sistemas.

- *Red local*: La implementación de redes locales de alta velocidad es actualmente técnica y económicamente accesible para cualquier persona o empresa dedicada a las tecnologías de la información; en sistemas de análisis de redes de información distribuidos, una red local Gigabit Ethernet o 10 Gigabit Ethernet sobre cable UTP CAT 6 es difícilmente saturable por la comunicación interna entre módulos de minería, almacenamiento y procesamiento de información, así mismo, su costo de implementación y mantenimiento es relativamente bajo, y el índice de estabilidad que presentan es muy alto, permiten establecer redundancia de manera sencilla y accesible, lo que las convierte en una buena solución para plataformas de mediano alcance.
- *Uplinks/Downlinks a Internet*: Uno de los principales cuellos de botella en los sistemas de crawling se presenta con la saturación de los enlaces de salida y principalmente de entrada (descarga) de información, y no precisamente por un tema técnico o tecnológico, sino por un aspecto económico, ya que son proporcionados por proveedores de servicios de Internet (*ISPs*) y sus costos pueden ser extremadamente elevados si se desean altas tasas de transmisión, lo que limita de manera directa el desempeño de nuestro sistema y reduce dramáticamente su rentabilidad como proyecto. Existen dos clases principales de enlaces que pueden ser utilizados, enlaces dedicados y enlaces a través de *ISDNs* (Redes Digitales de Servicios Digitales) entre los que destacan los enlaces DSL, ambos casos sin importar si se trata de enlaces simétricos o asimétricos e independientemente del medio y protocolos de transmisión.
    + *Enlaces Dedicados*: Se trata de enlaces, normalmente simétricos, en los cuales el canal de transmisión está garantizado (con índices de disponibilidad superiores a 99.5%) y dedicado de manera exclusiva, esto quiere decir que la porción del canal pagada estará

disponible y con la capacidad acordada en todo momento para un usuario específico. Las principales tecnologías de transmisión son enlaces de microondas punto a punto y por fibra óptica, para enlaces inferiores a un E1 puede ser utilizado inclusive par trenzado de cobre. La utilización de estos enlaces es extremadamente costosa, y normalmente es utilizada para los servicios provistos a los clientes de la compañía que realiza el análisis y no para la minería de datos en sí; otra clase de enlaces pueden ser utilizados para la recolección de información, simplificando la redundancia y permitiendo reducir en gran medida los costos de operación.

+ *Enlaces DSL (Línea de abonado digital)*: Son enlaces provistos sobre las líneas telefónicas digitales y normalmente presentan un bajo costo, así mismo, no son enlaces dedicados, es decir el canal está compartido por un conjunto de usuarios, y para el caso de ADSL, se trata de enlaces asimétricos, donde la velocidad del enlace de bajada (*downlink*) es 3 o 4 veces superior a la del *uplink*, sin embargo, esta característica favorece a dichos enlaces como buenos candidatos para ser utilizados por sistemas de crawling, ya que mayoritariamente se encontrarán descargando contenido. El utilizar varios enlaces DSL resulta ser un modelo escalable, eficiente y estable para satisfacer las necesidades de descarga de información de los sistemas de análisis de redes a un costo razonablemente bajo. La decisión de qué tecnología y proveedores utilizar debe ser tomada a partir de un análisis de factibilidad técnica y económica, pero las ventajas son sustanciales para la actividad que queremos desarrollar.

Las unidades de procesamiento intervienen en todos los sistemas involucrados, sin embargo, en este punto hablaremos de los aspectos más importantes que se desarrollan durante el análisis de la información recopilada a través de una plataforma de crawling.

- *Procesamiento en paralelo*: La enorme cantidad de información recopilada por los sistemas de crawling debe de ser procesada para obtener los resultados deseados y facilitar su utilización; convencionalmente existen dos modelos de procesamiento de gran escala, aquellos que se enfocan a resolver pocas operaciones de gran complejidad, y los que están destinados a resolver una alta cantidad de cálculos sencillos, este último escenario es el más común en los sistemas de nuestro interés, y la manera más eficiente de ser atacado es mediante la utilización de múltiples unidades de cómputo independientes, coordinadas a través de un sistema central que regula la carga de trabajo de cada una de ellas; el modelo anterior permite encontrar un buen balance

entre escalabilidad y costo, ya que las unidades de cómputo pueden estar distribuidas en casi cualquier clase de equipo terminar, como podrían ser todas las estaciones de trabajo de una oficina, operando en determinados momentos como un gran sistema de cómputo en paralelo.

Todo lo anteriormente descrito es de gran relevancia debido a que el crecimiento que puede presentar un sistema de crawling se comporta normalmente de manera exponencial. La utilización de *multithreading* en los módulos de conexión y procesamiento pueden acelerar en gran medida el desempeño del sistema; el diseño de la infraestructura que soportará la operación de la plataforma deberá ser diseñada evaluando integralmente todos los puntos descritos en esta sección.

2.4. Marco General de la Pornografía en Internet

Antes de entrar en detalle acerca de la estructura del contenido pornográfico en Internet, es fundamental entender la razón que explique su existencia, principalmente los aspectos económicos que sustentan dicha industria, sin dejar de lado un breve análisis sobre su legislación y ética.

2.4.1. Definición de pornografía

El término pornografía proviene del griego "πορνογραφία", donde *porne* significa "prostituta", y *grafía* "descripción", "descripción de una prostituta", sin embargo, siguiendo la línea de su definición etimológica, el concepto de pornografía es todo aquel contenido visual y auditivo que describe actos o imágenes sexuales con la intención de excitar. Un punto importante a denotar es que en la Antigua Grecia, la palabra pornografía era en realidad inexistente.

La presencia de la pornografía es tan antigua como el hombre mismo; los registros más antiguos, con un fin diferente a la excitación, son estatuillas prehistóricas que se presume tenían la intención de representar deidades o figuras místicas, relacionadas principalmente con la fertilidad de la tierra y la mujer; en China, India y Grecia, existen imágenes en templos y construcciones, decoradas con elementos iconográficos claros sobre su carácter sexual, fechados alrededor del 2500 A.C..

A partir del siglo XIX, con técnicas modernas de reproducción gráfica, en particular la fotografía, se dio inicio a una nueva industria dedicada a la comercialización de imágenes, inicialmente mujeres posando desnudas, que en base a publicaciones de carácter regular y de distribución masiva, hacían de ésta actividad un negocio muy redituable. En 1953, inició la publicación y comercialización de Playboy, que a la fecha, vende

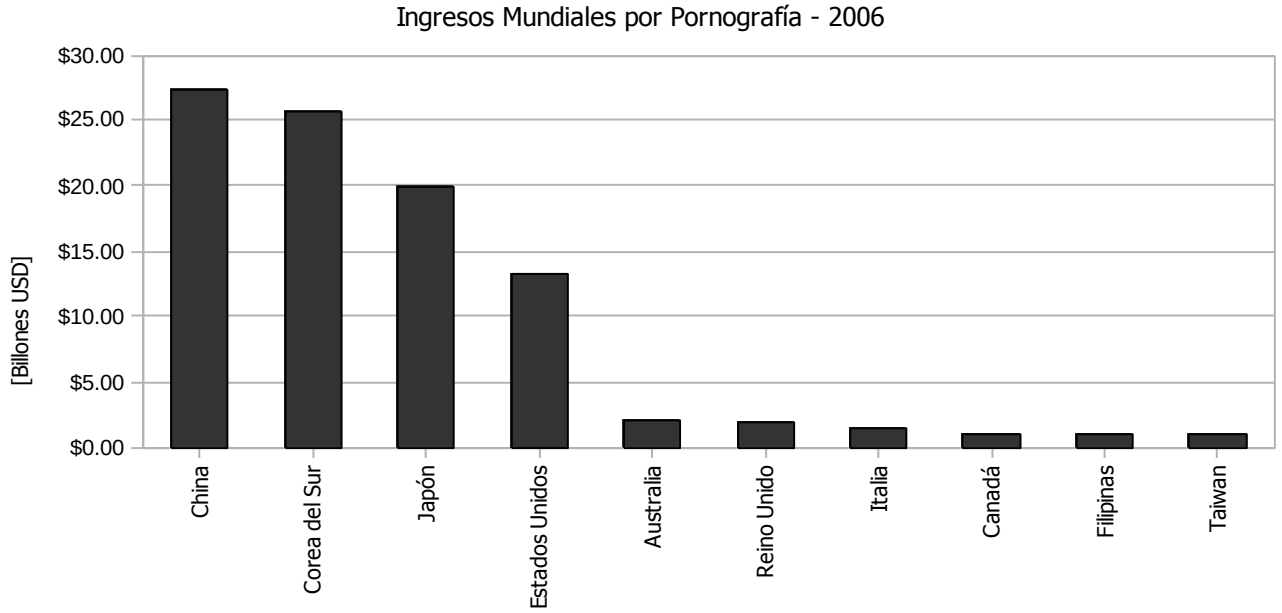
aproximadamente 5 millones de copias al año. A mediados de la década de 1970, gracias al desarrollo de sistemas de reproducción y videograbación accesibles, las películas y videoclips de tipo erótico iniciaron su presencia en lo que ya era una sólida industria de publicaciones impresas. Hasta este punto, existía la posibilidad de controlar, en cierta medida, su venta, y así limitar el acceso a menores de edad.

En 1994, con el inicio de lo que hoy conocemos como World Wide Web, la creación de páginas pornográficas, inicialmente de contenido fotográfico, y más recientemente multimedia, gracias a los avances en técnicas de compresión de video y el significativo aumento del ancho de banda en las redes, disparó todo indicador imaginable, desde la cantidad de dinero recaudado por la industria y el número de empresas multimillonarias dedicadas a ella, hasta desafortunadamente, los índices de acceso a contenido pornográfico por parte de menores de edad, la creación y distribución de pornografía infantil o adolescente, videograbaciones de violaciones y delitos sexuales, entre otros; la lista de elementos legalmente cuestionables, distribuida a través de Internet, es extremadamente amplia.

2.4.2. Estadísticas del perfil económico y demográfico de la pornografía en Internet

La información tabular correspondiente a las gráficas descritas a continuación está disponible en el Apéndice C de este documento, donde también se pueden obtener datos complementarios, acerca de la utilización y distribución de la pornografía en Internet, para 3 de los principales grupos en una sociedad.

¿Qué tan grande es la industria de la pornografía? Es una pregunta controversial entre un gran número de analistas financieros. Para muchos, es una de las industrias más rentables y sólidas a nivel internacional, para otros, en los últimos años ha presentado en realidad graves problemas económicos, poniendo en crisis a muchas de las empresas más representativas de este medio. La gráfica 2.4.1. muestra a los 10 países que reportaron los mayores ingresos por pornografía, de manera conjunta, dicha industria recaudó en el 2006 más de 60 billones (60,000 millones) de dólares americanos a nivel mundial; por otra parte, Forbes, una institución de análisis y calificación financiera ampliamente reconocida, señaló que en realidad, la industria de la pornografía no alcanza los 5 billones (5,000 millones) de dólares anuales. Existen diferentes puntos de vista acerca de por qué resulta tan diferente una cifra de otra, errores causados por la utilización de indicadores especulativos, si se consideran ventas en Internet o no, etc., sin embargo, sin establecer cuál de los casos es el correcto, dado que resulta irrelevante para la idea central de esta tesis, la conclusión que podemos obtener es que la industria pornográfica es inmensa, y sobre todo, está ampliamente difundida, teniendo uno de los más altos índices de consulta en medios digitales, si no es que el más alto.

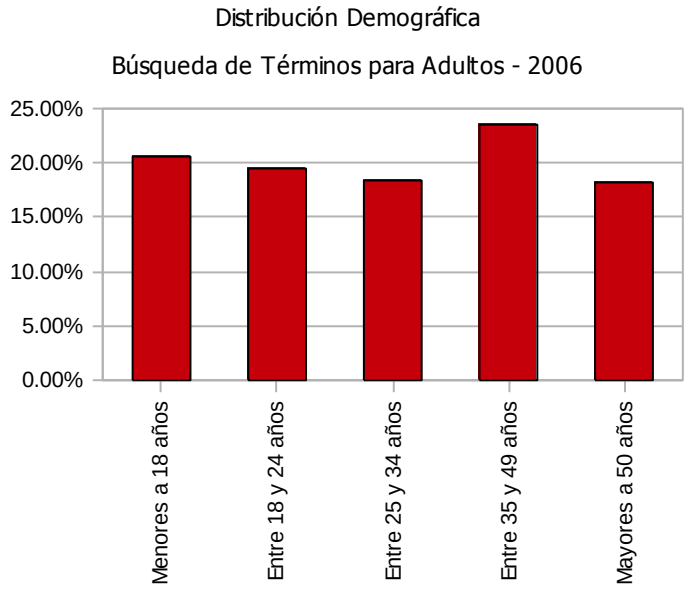


Gráfica 2.4.1. - Gráfica de los 10 países que reportan los mayores ingresos por pornografía en el año 2006.

Entrando un poco más en contexto, la gráfica 2.4.2. representa la distribución de búsquedas de términos para adultos realizadas en el 2006, dividiendo a la población en 5 grupos representativos. Lo sorprendente de estos datos es que nos muestran una distribución muy uniforme entre los diferentes segmentos, cercanos todos al 20%, cuya interpretación directa es que tanto el interés, como la capacidad, al menos de realizar búsquedas, no se limita a ningún segmento de la población, y tomando en cuenta la alta efectividad y simplicidad de los sistemas de acceso a la información, más del 95% tiene éxito en sus consultas, sin importar su edad.

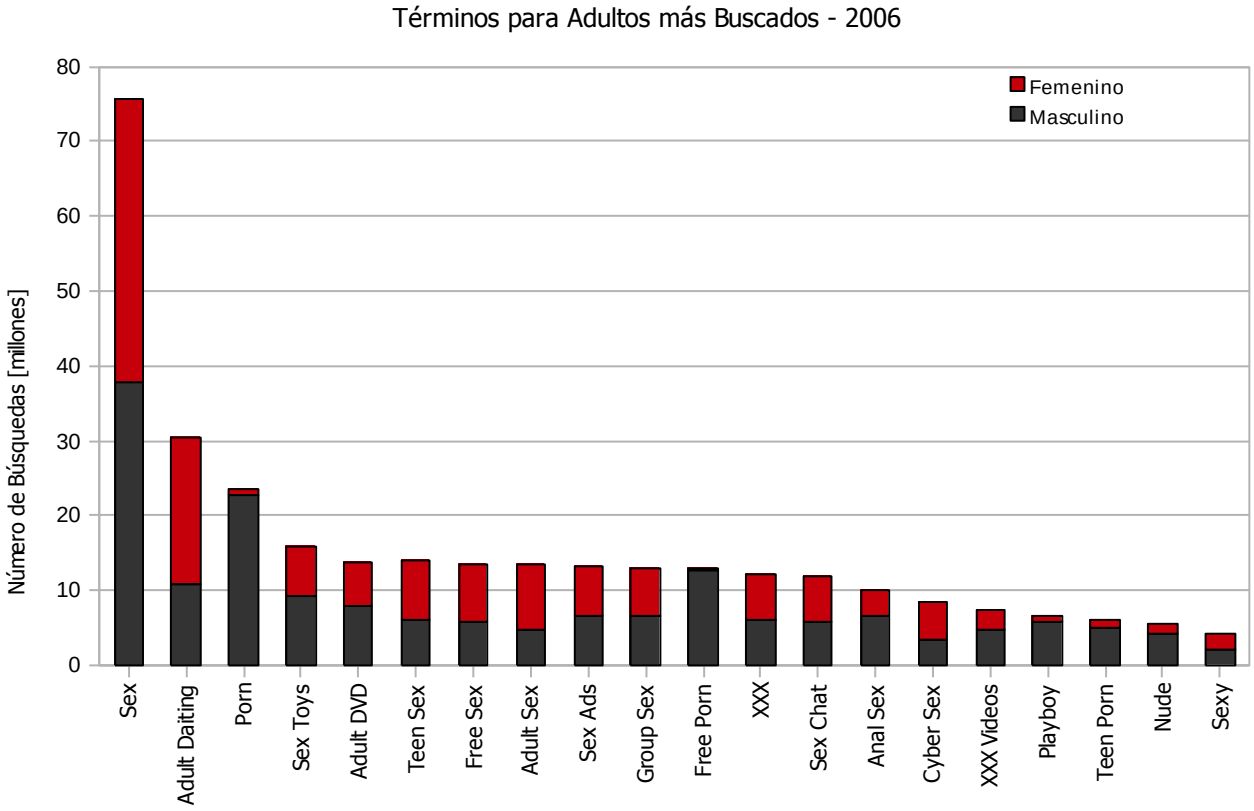
Una de las maneras más efectivas para detectar de manera preliminar contenido para adultos dentro de una página de Internet, es realizar un análisis semántico, buscando como mínimo la aparición de una o varias palabras claves relacionadas de manera directa con dicho tipo de información; es importante mencionar que no basta con encontrar dichos términos, ya que por ejemplo, en un artículo científico sobre sexualidad o reproducción, existe una gran posibilidad de encontrar oraciones similares a otras contenidas en páginas pornográficas, lo que afecta de manera significativa la cantidad de falsos verdaderos en nuestro sistema de clasificación, es decir, documentos que se clasifican de manera errónea. Para solucionar el problema anteriormente mencionado existen una amplia variedad de opciones complementarias, cuyo fin se puede describir en general como el contextualizar a los documentos, a partir de sus relaciones y orígenes, para tener

una idea más clara sobre su intención y naturaleza, nosotros en particular, utilizaremos el algoritmo ActiveRank como medio para el aprovechamiento de la estructura de la red de información.



Gráfica 2.4.2. - Distribución demográfica de búsquedas de términos para adultos en el año 2006.

La gráfica 2.4.3. muestra los 20 términos relacionados con contenido para adultos más buscados en el año 2006, así mismo, la división por color de cada barra representa el porcentaje entre consultas originarias de una persona de sexo masculino, para el color gris, y de sexo femenino para el color rojo. La importancia de estos datos en la realización de la presente tesis es que nos proporciona una idea clara sobre algunos de los conceptos o palabras que tendríamos que buscar de manera inicial dentro del contenido de un documento para realizar una primera aproximación acerca de su naturaleza; el segundo punto fundamental que podemos obtener de su análisis es la confirmación de que la pornografía es en realidad consultada tanto por hombres como mujeres, con una distribución 50% - 50%.



Gráfica 2.4.3. - Distribución demográfica por género de los términos para adultos más buscados en el año 2006.

El método que utilizaremos para encontrar los términos (palabras y conceptos), el cual es descrito con detenimiento más adelante, consiste en realizar un análisis semántico a una muestra representativa de páginas pornográficas, obtener una lista de palabras ordenadas por número de repeticiones, y realizar una selección manual de aquellas que se consideren relacionadas a contenido limitado para adultos, crenado así una lista o diccionario que será utilizado como referencia por los sistemas automáticos.

2.4.3. Estructura del contenido pornográfico en Internet

Basados en la tabla A.4. contenida en el Apéndice A de este documento, podemos señalar que la pornografía es el tema de mayor consulta a través de Internet, concentrando el 25% de las búsquedas realizadas, y ocupando el 8% del total de correos electrónicos enviados cada día, mayoritariamente en forma de correo no deseado. Se estima que existen alrededor de 420 millones de sitios dedicados a la creación de contenido pornográfico, sin embargo, el número de páginas es mucho mayor, ya que cada uno de estos sitios puede contener miles de páginas anidadas además de una cantidad inimaginable de vínculos a otros sitios gemelos.

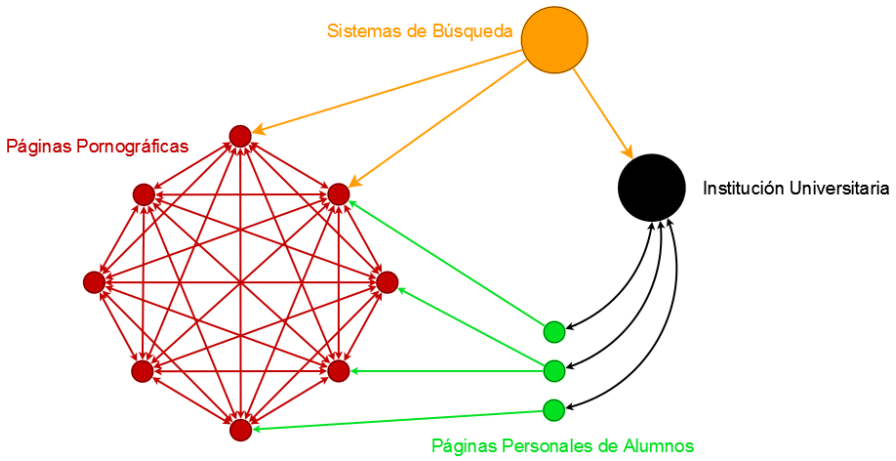


Figura 2.4.4. - Estructura general para la red de sitios pornográficos y su interconexión con otros sistemas.

Retomando la conceptualización de la WWW descrita en la sección 2.3.3. de esta tesis, la figura 2.4.4. nos da una idea clara de cómo las páginas pornográficas constituyen un continente de salida dentro de la web, que además se encuentra altamente interconectado en su interior; es posible accesar a dicho contenido a través de un gran número de caminos, como sistemas de búsqueda, o las páginas personales de alumnos dentro de una institución universitaria, sin embargo, una vez dentro de un sitio pornográfico, nunca encontraremos un vínculo a una página documental, de noticias, gobierno, ni cualquier otro tipo que no sea en sí pornográfica, lo que representa un reto tecnológico para los sistemas automáticos de análisis de redes de información, y que es justamente el tema central de la propuesta descrita en este documento.

El estudio de las plataformas que sustentan a la industria de la pornografía en Internet es por sí mismo un tema muy interesante; la conformación de los sitios para adultos difiere de manera sustancial con el común denominador. Se basan en sistemas de gran escala, los cuales concentran el contenido y lo integran a través de cualquier número de interfaces gráficas, por lo que cada compañía puede producir una gran cantidad de “marcas”, todas basadas en la misma información, y para incrementar el tiempo de navegación de un usuario dentro de su plataforma, se crean millones de vínculos entre estos sitios, generando redes con un grado de interconexión cercano al total, concluyendo, todas las páginas pornográficas están interconectadas entre sí con un grado no mayor a 3.

2.4.4. Legislación y ética de la pornografía en Internet

La legislación de Internet ha sido un tema ampliamente discutido en los últimos años; existen diferentes puntos fundamentales que deben de ser mencionados para lograr comprender de manera general la situación actual. Al ser Internet un sistema global, en el cual se pueden desarrollar un gran número de actividades desde cualquier lugar, utilizando herramientas que pueden residir en cualquier otro, la legislación toma un contexto internacional que ningún otro sistema ha tenido hasta el momento; hablar de reglamentos y leyes aplicables a los usuarios de Internet se refiere lograr que todos los países del mundo declaren e implementen las mismas reglas y penalizaciones, así como homologar criterios de seguridad e identificación, comercio y economía, entre otros, donde se ubican también qué las actividades y contenido legalmente permitido en la red. Todo lo anterior resulta actualmente imposible si se trabaja desde el punto de vista descrito, intentando realizar una unificación global de leyes, sin embargo, Internet si resulta ser parcialmente legislado, en el momento que una de las actividades virtuales, tiene un impacto directo sobre la realidad.

La mayoría de las leyes de casi todos los países del mundo fueron generadas antes de la creación de Internet, por lo que ni siquiera estaba considerada su existencia; uno podría pensar que lo anterior descarta posibilidad alguna de implementarlas en un mundo completamente virtual, y sí, sería así si fuera completamente virtual, pero la realidad es que Internet es utilizado para intercambiar información y realizar actividades que tienen un impacto directo en nuestra vida cotidiana, y es en ese punto, donde algunas reglas hacen sentido. Un buen ejemplo es el comercio electrónico, que haciendo uso de la WWW permite entablar una transacción comercial muy fácilmente, sin embargo, aquí termina la parte que tiene que ver con Internet, una vez que ingresamos los datos de la tarjeta de crédito y presionamos el botón que confirma nuestra orden, se inicia un proceso que incluye desde políticas de transacciones internacionales, hasta exportación, transporte y entrega de una mercancía determinada, que evidentemente está sujeta a todos los procedimientos de comercio internacional

que se han venido llevando a cabo desde hace más de 100 años. Otro caso claro es la información protegida por las leyes internacionales del derecho de autor, que al ser digitalizada y distribuida en Internet, viola un gran número de disposiciones internacionales sobre la reproducción y explotación de contenido propietario, con lo que las personas que distribuyen o utilizan dichos recursos, están cometiendo un acto jurídicamente penalizable. El problema con el último ejemplo mencionado, y que aplica de manera similar a la mayoría de la información disponible en Internet, es la gran dificultad para ejercer dichas leyes; debido a su facilidad de reproducción y transmisión, un documento o archivo puede ser publicado y descargado de manera indetectable, o desde un número de fuentes tan amplio, que resulta técnica y económicamente imposible de legislar, a lo anterior hay que añadir que dichas fuentes pueden encontrarse físicamente alojadas en servidores colocados en países con preferencias fiscales y jurídicas, que hacen mucho más complejo su aprovechamiento. Uno de los casos sobre legislación en Internet más importantes que se han llevado a cabo ha sido el de las principales disqueras contra Napster, donde demandaron al sistema de distribución de contenido por infringir leyes de derechos de autor, ganando la demanda; el caso pudo ser ganado debido a que el sistema centralizaba la información en servidores propiedad de la compañía; los sistemas a los que dio lugar Napster son conocidos como *Distributed Peer to Peer*, donde el sistema realiza la búsqueda entre el conjunto de documentos residentes en las computadoras de los usuarios, haciendo imposible una demanda contra la compañía propietaria de la plataforma, ya que se clasifica como un motor de búsquedas, que además no concentra la información.

La pornografía en Internet resulta ser un tema de legislación extremadamente complejo, teniendo como principales características las siguientes:

a) La generación o producción de contenido pornográfico, en el contexto de la industria de entretenimiento para adultos, está legislada por las leyes del país donde se produzca, en términos generales, las personas participantes deben ser mayores de edad según el país involucrado, realizándolo por voluntad propia y en pleno uso de sus facultades mentales. En Internet, gran parte del contenido comercializado no es producto de la industria antes mencionada, sino grabaciones caseras o "productoras independientes" de menores de edad siendo sexualmente explotados, de adolescentes en fiestas bajo los efectos del alcohol y drogas, hasta actos criminales de secuestro y violación.
b) La comercialización de contenido pornográfico a través de Internet funciona de manera general a través de suscripciones prepagadas, que brindan acceso a determinado tipo de material multimedia. Es importante mencionar que existe una amplia gama de sitios pornográficos de acceso público y gratuito, y que en realidad, son la minoría aquellos que no permiten la visualización de ningún tipo de contenido sin previa identificación y pago. La conclusión es que

en Internet la pornografía es abierta y gratuita, sin embargo, a través de suscripciones que permiten “acceso ilimitado”, es una industria que recauda cifras multimillonarias.

c) La publicación de contenido pornográfico a través de Internet se lleva a cabo utilizando servidores ubicados en países con prerrogativas fiscales, así como reglamentación en la cual consideran confidencial la información contenida en dichos sistemas, que en realidad se interpreta como que no se fiscalizará por ningún motivo los documentos o archivos contenidos dentro de los sistemas.
d) La identificación o autentificación de los usuarios de sitios pornográficos es simplemente inexistente; en el mejor de los casos, las personas deben de leer y acordar las condiciones del servicio, entre las que destacan que está únicamente dirigido a mayores de 18 años, y que en caso contrario deberán abandonar la página; resulta evidente que esta clase de identificación o limitación no es suficiente para evitar que un menor de edad simplemente continúe el proceso e ingrese a la página.
e) El modelo de información distribuida utilizado por los sistemas de contenido pornográfico dificulta enormemente la fiscalización del contenido existente, ya que prácticamente elimina la posibilidad de destruir información ilegal, ya que no se sabe ni cuantas copias existen ni en dónde están.
f) La facilidad de reproducir la información disponible en Internet, y en este caso pornografía, permite que con el mismo contenido, se generen una infinidad de posibles escenarios donde encontrarlo, como pueden ser servidores en universidades, instituciones públicas, entre otras, que complican aún más su legislación, pues cada una de ellas tiene reglamentos internos sobre el acceso y depuración de sus sistemas.
g) El acceso a contenido pornográfico a través de Internet normalmente asocia otros problemas como son virus, spam, y violaciones graves a la privacidad de las personas, que afectan de manera directa a personas y empresas tanto moral como económicamente.

Los puntos anteriormente mencionados nos dan una referencia general de la dimensión y diversidad de áreas que abarca la legislación del contenido pornográfico en Internet; de acuerdo a altísimo porcentaje que representa del contenido disponible a través de la red, aunado a los intereses económicos que representa, pensar en que dicha información es candidata a ser fiscalizada, resulta totalmente irreal, lo que genera un grave problema, ya que no existe manera de que dicha industria contara con un código de conducta y un marco de ética profesional bien definidos, y sobre todo, bien implementados.

Los principales problemas con la pornografía en Internet no es que exista, o que personas con la intención, la edad, la capacidad y los recursos accedan a ella intencionalmente, ya que al final todo se reduciría al derecho y libre albedrío de las personas a decidir consumir dicha clase de entretenimiento a través de un medio digital en base a sus paradigmas éticos. Las verdaderas aberraciones se dan cuando aquellas personas que no quisieran o no deberían estar expuestas a dicha información se ven afectadas de manera directa por sistemas invasivos que perjudican desde su seguridad hasta sus recursos informáticos, al recibir volúmenes increíbles de correo electrónico no deseado, utilizar su conexión de internet para descargar imágenes o *banners* publicitarios en sus equipos de forma no autorizada; cuando dicho contenido es ubicado o descargado utilizando recursos de instituciones públicas o privadas las cuales no desean ni permiten que su infraestructura sea inundada de dicha información; y por último, cuando la pornografía en Internet deja de ser simplemente un negocio de entretenimiento audiovisual para adultos, y se convierte en una plataforma para la difusión de contenido producto de un sinfín de situaciones, entre las que destacan crímenes de abuso y explotación sexual.

Dado que es imposible tomar y eliminar todo el contenido (de cualquier clase) que no se desee en Internet, y que además dicho acto sería una contradicción de todos los principios que sustentan su propio modelo, existen métodos de contención que permiten proteger a las instituciones y usuarios de toda la información que es accedida a través de su infraestructura, sea una PC o una red corporativa. Dichos sistemas consisten en filtros que incorporan diferentes etapas de detección, que básicamente permiten o no que desde una red o equipo determinados se pueda visualizar cierta clase de información. El trabajo desarrollado en esta tesis permitirá conocer las posibilidades del algoritmo ActiveRank como método de detección de contenido, y junto a los sistemas de análisis de información también estudiados, existirá la posibilidad de utilizar parte de esta tecnología como base para aplicaciones de seguridad informática de alta eficiencia para la sociedad en general.

Para concluir las ideas descritas en este subcapítulo, la ética de la industria de la pornografía a través de Internet es inexistente, y además, es ingobernable, debido a la diversidad de fuentes y complejidad en temas legislativos que incorpora, sin embargo, existen los medios para aislar dicho contenido de los entornos sociales y tecnológicos (que en la teoría de la “sociedad de la información” son uno mismo) a través del uso de tecnología de detección y filtrado de contenido, en vez de un modelo que considerara la eliminación del 12% del contenido disponible en la WWW, que es simplemente imposible.