# UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

## FACULTAD DE INGENIERÍA

**"ANÁLISIS DEL ALGORITMO ACTIVERANK COMO MÉTODO DE DETECCIÓN AUTOMÁTICA DE CONTENIDO DENTRO DE REDES DE INFORMACIÓN"**

**TESIS**

QUE PARA OBTENER EL TÍTULO DE:
**INGENIERO EN TELECOMUNICACIONES**

PRESENTA:
**FERNANDO LUEGE MATEOS**

DIRECTOR DE TESIS:
**DR. MIGUEL MOCTEZUMA FLORES**


**MÉXICO, D.F.** **2010**

*"Por mi raza hablará el espíritu"*

Esta tesis está dedicada con todo mi amor, respeto y admiración a mi familia.

Mamá y Papá, gracias por apoyarme y creer incondicionalmente en mí, gracias por enseñarme a dar cada paso, sin ustedes nada sería posible, nada sería realidad.

Mikel, gracias por ser mi mejor amigo, por ayudarme en todo momento con tu tranquilidad, sabiduría y gran corazón.

Nenita bonita totita, gracias por ser mi mejor amiga, por hacerme sonreír todos los días con tu alegría, carisma e intensidad.

Gracias por ser mi inspiración.

Fernando

## Agradecimientos

Quiero extender un agradecimiento especial:

A Asaf Paris Mandoki, quién ha sido un gran amigo y el mejor equipo de trabajo desde hace más de 10 años.

A Rafael Peña Miller, con quién desarrollé las bases de lo que posteriormente se convertiría en el Algoritmo ActiveRank.

A todo el equipo de personas que trabajan en Ondore y que sin su participación el desarrollo de esta tesis habría sido imposible, en especial a Rodrigo Saucedo, Saidel López, Jorge Mendieta y Alexandre Kouznetsov.

A Adriana Sosa, Memo García y Genaro Cruz, por todo el apoyo a lo largo de la carrera y por los grandes momentos que hemos compartido.

Al Dr. Miguel Moctezuma, por dirigir mi tesis y a mis sinodales, el Dr. Victor García, el M.I. Jesús Reyes, el M.I. Federico Vargas y al Fis. Juan Velázquez.

Y al resto de mis grandes amigos, con los que comparto mi vida tanto personal como profesional día con día.

Gracias.

Fernando

# Índice

# Capítulo Primero

## Definiciones

# 1. Definiciones

## 1.1. ¿Cómo leer esta tesis?

El presente trabajo de investigación está dirigido primordialmente a personas con formación matemática avanzada y conocimientos básicos de sistemas computacionales e Internet; sus áreas afines son Ingeniería en Telecomunicaciones, en Sistemas, Informática o Ciencias de la Computación, sin embargo, público en general podrá entender claramente el fin de la presente tesis al leer con detenimiento cada una de sus secciones.

El primer capítulo *Definiciones* presenta el contexto general de la investigación y el glosario de términos necesarios para su correcta interpretación. Personas con conocimientos sobre Teoría de Gráficas, Sistemas e Internet pueden leer la sección 1.2. y dejar para consulta las secciones 1.3., 1.4. y 1.5.

El segundo capítulo *Antecedentes* da a conocer de manera detallada el marco general sobre el cual se desarrolla el resto de la investigación. Personas con conocimientos detallados sobre teoría de gráficas y estructura de Internet, en particular de la WWW, pueden pasar directamente a la sección 2.3. y 2.4.

El tercer capítulo *Algoritmo ActiveRank* presenta a detalle el algoritmo utilizado a lo largo de la presente investigación, y su claro entendimiento es clave para la correcta comprensión de las secciones experimentales, análisis de resultados y conclusiones. Personas carentes de sólida formación matemática podrán leer únicamente la sección 3.1., siendo esta suficiente para comprender qué hace ActiveRank.

El cuarto capítulo *Trabajo Experimental* plantea a detalle los protocolos de los procesos experimentales así como sus resultados.

El quinto capítulo *Análisis de Resultados* explica cada uno de los resultados obtenidos en la sección anterior y sus implicaciones con respecto al fin de la presente tesis, de igual forma, formación matemática es requerida para la comprensión de los resultados numéricos y algoritmos de evaluación de eficiencia planteados ahí.

El sexto capítulo *Conclusiones* plantea los puntos finales sobre el desarrollo de la presente tesis, las principales contribuciones y el trabajo futuro propuesto para continuar los estudios de esta línea de investigación.

Por último, existen cuatro *Apéndices* que presentan información adicional referenciada a lo largo del documento, así como la sección de *Referencia Documental* donde se enumeran cada una de las fuentes bibliográficas y digitales utilizadas para el desarrollo de esta tesis.

### 1.2. Contexto

La capacidad con la que contamos hoy en día para generar y compartir información, a través de medios digitales y redes globales, es simplemente ilimitada; de igual manera, los retos de ingeniería que presentan la recopilación, el análisis, el manejo y la explotación de dichos documentos son increíblemente altos. Los sistemas automáticos de exploración y análisis de redes de información contienen un desafío particularmente complejo; deben clasificar y perfilar los documentos para poder explotar, depurar y mejorar los sistemas de acceso a la información, así como los propios de análisis. A continuación se describe uno de los muchos escenarios en los cuales los algoritmos, y en particular ActiveRank, toman gran importancia dentro de los mismos sistemas de análisis de información, y que servirá como marco de referencia a lo largo del desarrollo de esta tesis.

Dentro de una red, se pueden presentar estructuras cíclicas que dificultan su análisis mediante sistemas automatizados, por lo que existe la necesidad de contar con métodos que sean capaces de detectar y manejar el comportamiento del analizador cuando se presentan dichas estructuras. Un ejemplo para entender claramente el efecto de las estructuras cíclicas en una red, es aquella formada por las páginas pornográficas en la red denominada World Wide Web; se trata de un conjunto de nodos altamente interconectados entre sí, y poco o nada interconectados hacia el exterior de su núcleo, es decir, podemos encontrar un vínculo para llegar a ellos, pero no un vínculo para salir de ellos, por lo que un sistema automático que entrara a dichas estructuras, no sería capaz de seguir escaneando otras secciones de la red, se vería atrapado si no fuera por su capacidad de detectar y manejar dicha situación.

Así mismo, existen algunos tipos de virus (en términos de computación, se refieren a programas automáticos que sin el consentimiento del usuario, desarrollan una acción maliciosa dentro de su equipo), que al infectar un servidor web, generan miles de páginas y vínculos hacia contenido pornográfico; el problema antes descrito, frecuente en universidades, plantea el reto de generar herramientas que nos permitan identificar y remover dichos enlaces, protegiendo no solo los sistemas de las instituciones, sino a aquellos usuarios que navegando una red de información que debería ser segura, se ven expuestos a contenido indeseado.

La presente investigación analiza el desempeño del algoritmo ActiveRank como sistema de clasificación de información, y plantea las bases para su utilización en la detección de contenido pornográfico y estructuras cíclicas en los procesos de indexación y clasificación de redes de información.

### 1.3. Teoría de Gráficas

**Nodo:** Es un punto terminal o de intersección dentro de una gráfica; se trata de la abstracción para representar un sujeto individual o colectivo dentro de la red. Gráficamente se denota como un punto en el plano.

Sinónimos: vértice, agente *[en contexto de redes sociales]*.

**Vínculo:** Es la unión entre dos nodos. El vínculo $( i , j )$, es aquel que inicia en el nodo $i$ y termina en el nodo $j$; puede ser direccional o no direccional (bidireccional).

Sinónimos: enlace, relación, arista.

**Gráfica:** Es un conjunto de nodos interconectados por vínculos.

Sinónimo: red.

**Subgráfica:** Se define como la gráfica generada por un subconjunto de nodos conexos y sus vínculos correspondientes. Esencialmente, cualquier elemento de la red es en sí una subgráfica.

Sinónimos: subred, subgrupo.

**Gráfica Planar:** Gráfica cuya totalidad de nodos y vínculos pueden ser colocados sin intersecarse sobre un plano.

Sinónimos: estructura/red bidimensional.

**Gráfica No Planar:** Gráfica cuya totalidad de nodos y vínculos no pueden ser colocados sin intersecarse sobre un plano.

Sinónimos: estructura/red multidimensional.

**Diada:** Es la interconexión de 2 nodos a través de un vínculo; se considera la expresión mínima de una red.

**Triada:** Corresponde a la interconexión de 3 nodos a través de 3 vínculos; es una estructura planar totalmente interconectada.

**Grupo:** La unión de subgráficas a partir de un criterio determinado, como pueden ser procesos de agrupación *[clustering]*.

**Grado:** Es el número de vínculos entre dos vértices conexos cualesquiera de la gráfica.

Sinónimo: distancia *[distancia de Hamming (Block distance)]*.

**Diámetro:** Es la distancia entre los dos elementos de la gráfica más lejanos entre sí, por consiguiente, corresponde al valor de grado máximo presente en la red.

**Ranking:** Coeficiente numérico que expresa la similitud entre dos elementos de la red. Puede ser interpretado también como el peso del vínculo, o en otros casos, la importancia de un elemento en relación a otros, dependiendo del contexto.

Sinónimos: peso, distancia *[ranking como una medida de cercanía o similitud]*.

**Topología:** Es la forma o conformación estructural que adopta una gráfica, o subconjunto de la misma. El análisis de la topología de una gráfica es fundamental para entender su dinámica, así como en el planteamiento de operaciones sobre la misma, como pueden ser procesos de agrupamiento.

Sinónimo: estructura.

**Clustering:** Es el proceso de reordenamiento de la red, que tiene por objetivo crear grupos de elementos afines según un criterio dado, normalmente en base a un valor de distancia.

Sinónimos: agrupación.

### 1.4. Computación y Redes

**Servidor:** Es la combinación de hardware y/o software que tiene como fin brindar un *servicio* a un *cliente*. Usualmente se trata de la plataforma para aplicaciones que a través de un protocolo establecido, se comunican con otro programa *cliente* a través del cual se desarrolla una tarea en particular.

**Virus:** Existen diferentes clases; en términos generales se trata de un programa automático que se copia e instala de manera indeseada y que conlleva al perjuicio de la información y el equipo del usuario, ya sea

por ataques como robo de información, suplantación de identidad, destrucción de información y recursos, entre otros.

**Protocolo:** Es un método establecido para que dos equipos de cómputo se comuniquen.

**Crawler:** Se trata de un programa automático que tiene la capacidad de explorar y analizar la WWW a través de la extracción de los vínculos contenidos en cada una de las páginas a las que accede.

Sinónimos: araña *(spider)*, robot.

**Paquete:** Un paquete es una unidad formateada de información transmitida a través de una red de computadoras por conmutación de paquetes.

**Socket:** Un socket es el punto terminal de un flujo bidireccional en un proceso de comunicación utilizando un protocolo de Internet en una red de computadoras. Se puede ver como el canal de comunicación que se establece entre el servidor y la aplicación cliente para la obtención del contenido deseado.

### 1.5. Internet y World Wide Web

**Wiki:** Es un tipo de software para trabajo colaborativo que permite crear páginas de Internet de manera conjunta entre un grupo de personas a través de un navegador web. Su importancia radica en la veracidad y pluralidad de la información que se genera al interior de estos sistemas.

# 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.

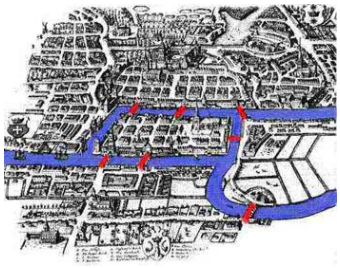


Figura 2.1.1. - Puentes de Köingsberg
*[http://www.daviddarling.info/encyclopedia/B/Bridges_of_Konigsberg.html]*

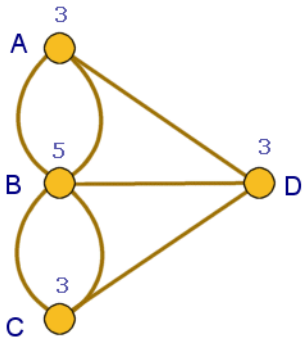



Figura 2.1.2. - Gráfica de los Puentes de Köingsberg
*[http://www.infovis.net/imagenes/T1_N137_A6_KonigsGraph.gif]*

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 $\Gamma_v$. 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 $\gamma$ el coeficiente de clustering o agrupamiento del vecindario $\Gamma$ 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 $\Gamma$, 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.

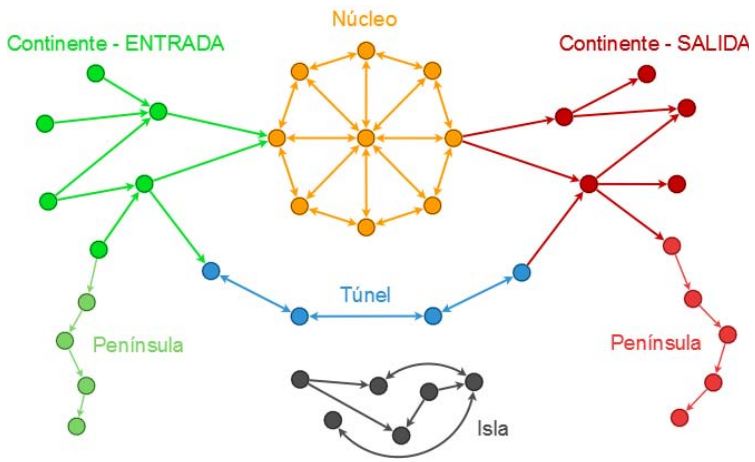



Figura 2.2.1. - Conceptualización de la estructura de la World Wide Web
*[BARABÁSI; Linked; p. 166.]*

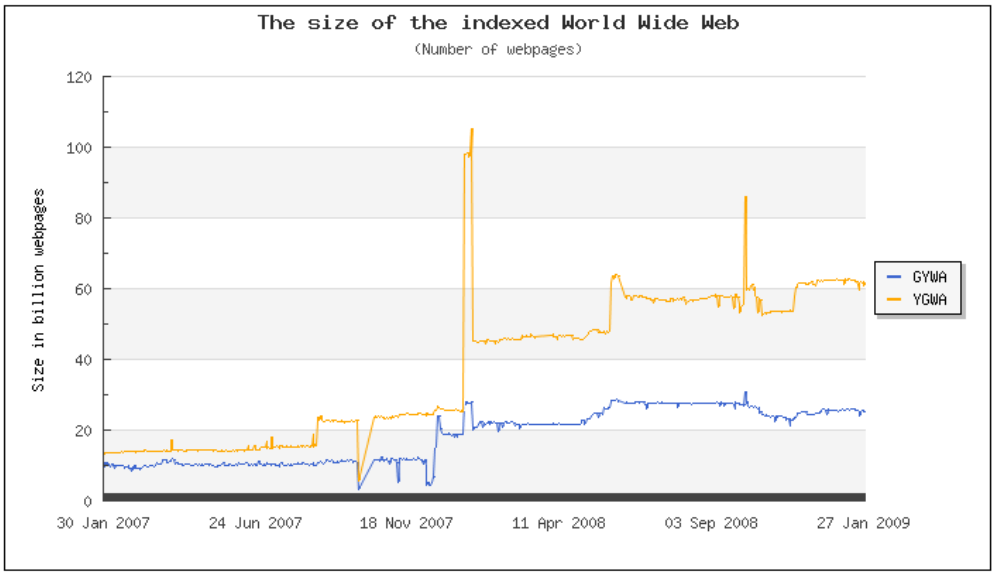
Se puede entender al núcleo como aquellas páginas de Internet que pueden ser accesadas desde muchas otras, las cuales a su vez se encuentran altamente interconectadas entre sí, es decir, mantienen un gran número de vínculos bidireccionales. Los continentes, son extensos grupos de páginas que se encuentran 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.

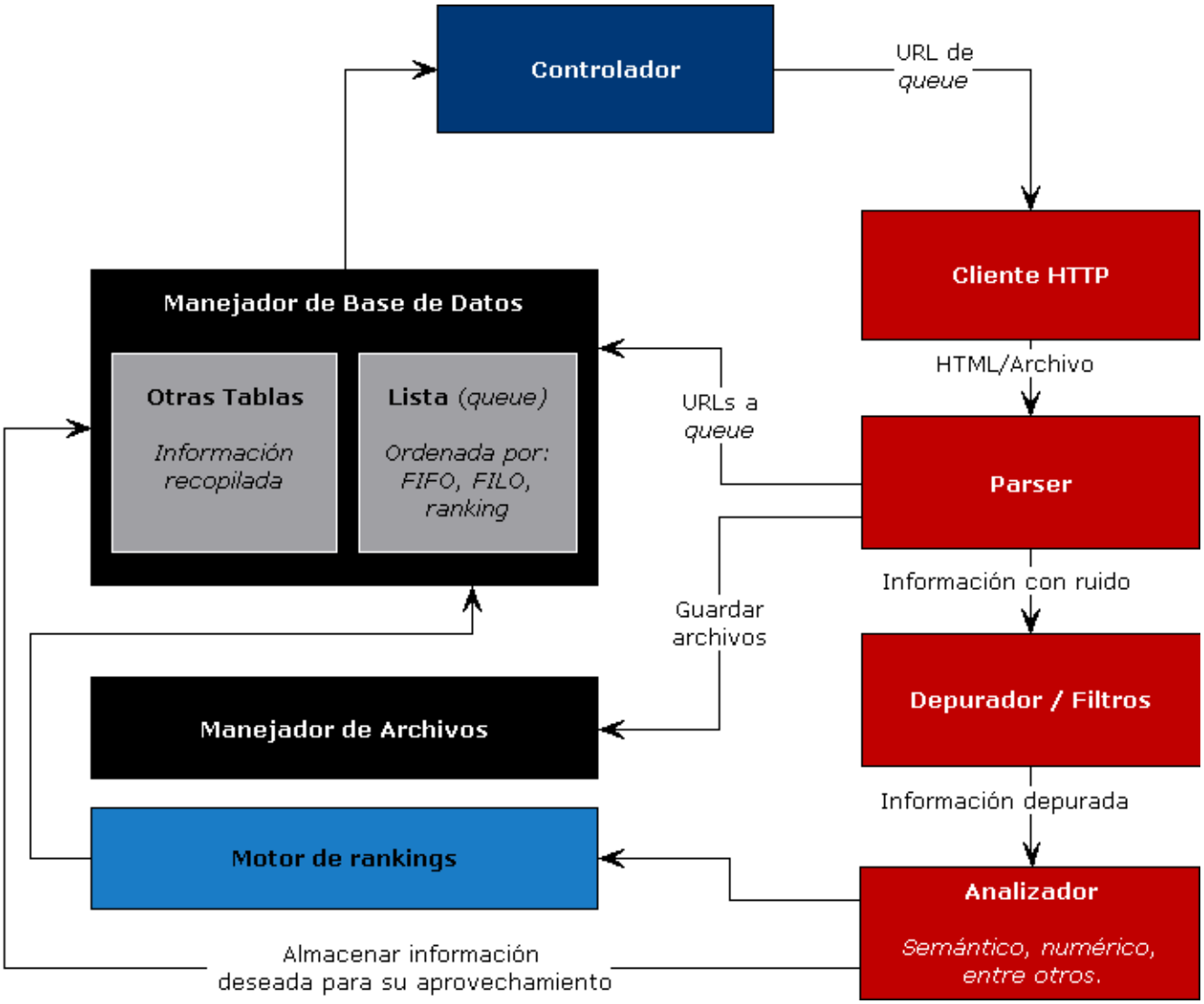



Figura 2.3.1 – Estructura básica de un crawler

- *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.

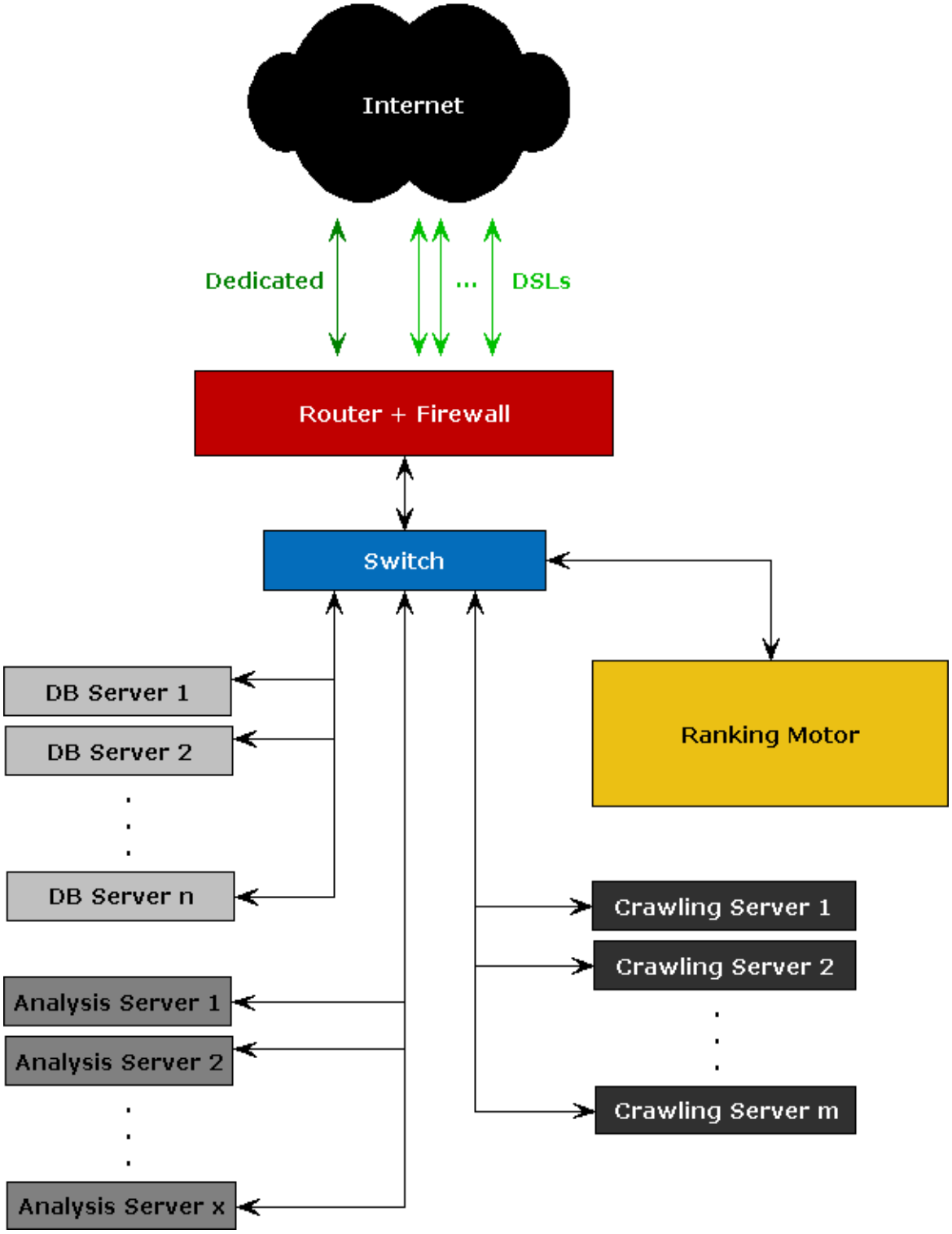



Figura 2.3.3 – Plataforma base para un sistema de crawling

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

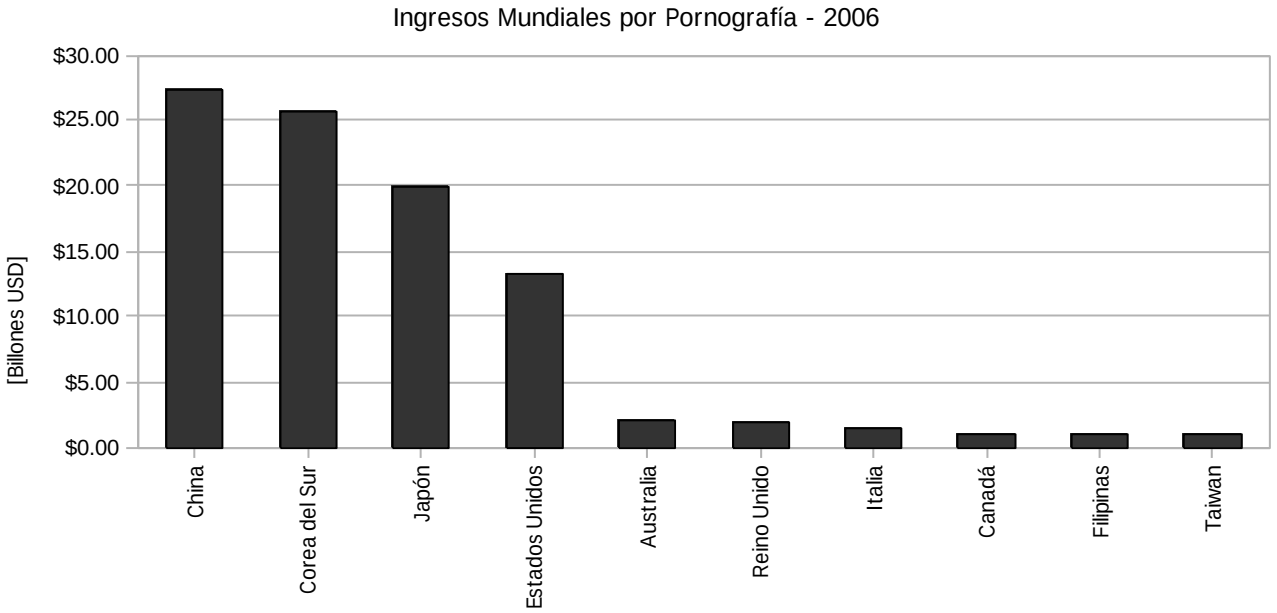
aproximadamente 5 millones de copias al año. A mediados de la década de 1970, gracias al desarrollo de sistemas de reproducción y videograbación accesibles, las películas y videoclips de tipo erótico iniciaron su presencia en lo que ya era una sólida industria de publicaciones impresas. Hasta este punto, 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.

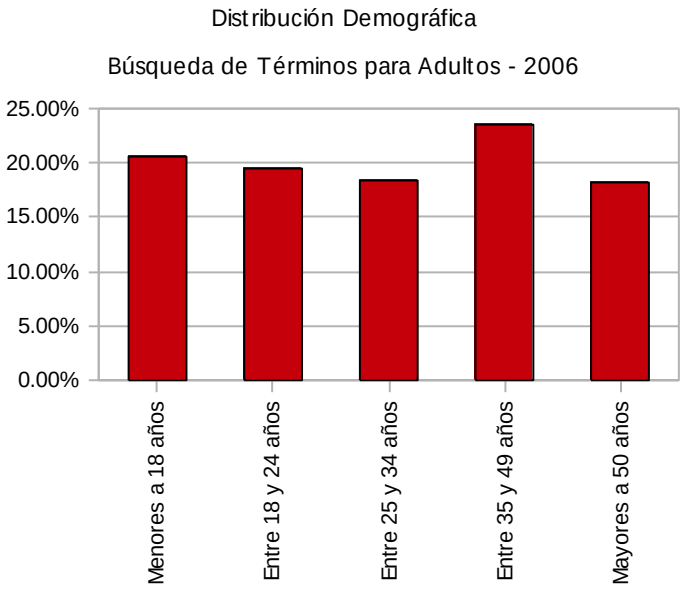



Gráfica 2.4.1. - Gráfica de los 10 países que reportan los mayores ingresos por pornografía en el año 2006.

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.

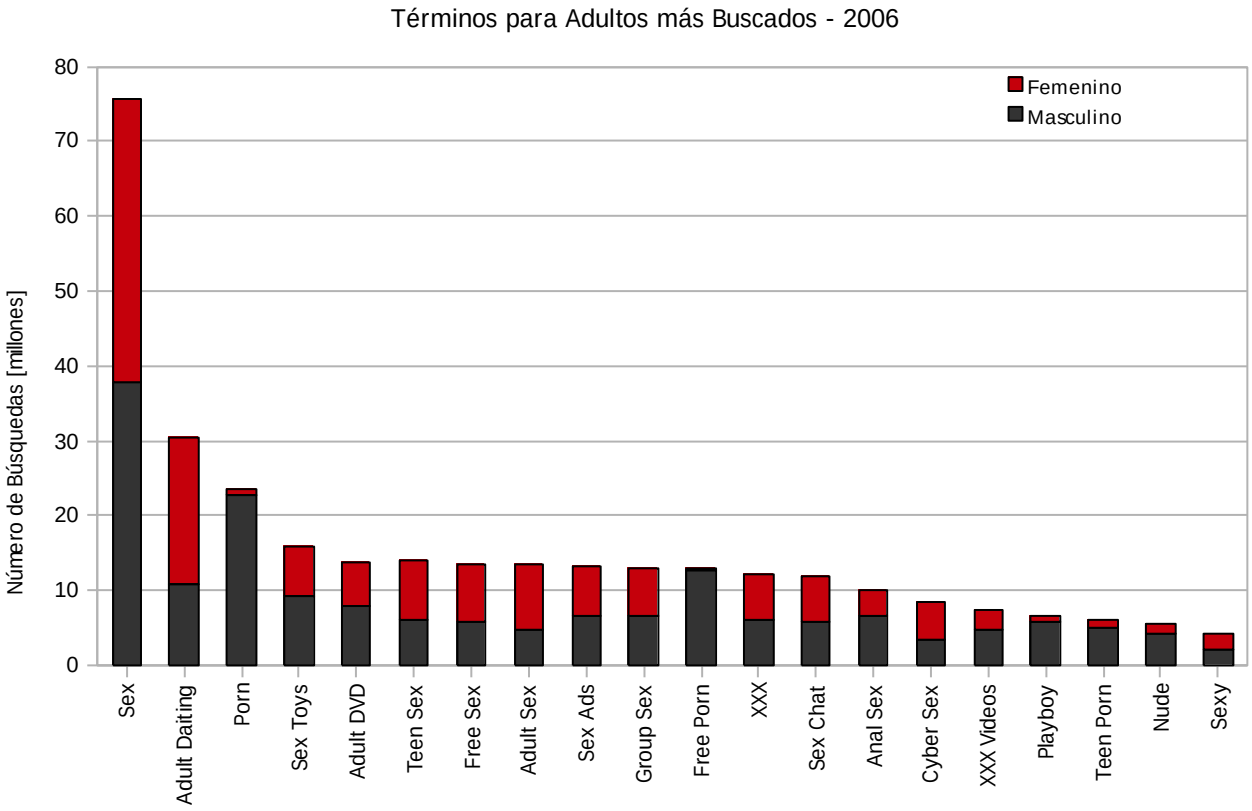
Una de las maneras más efectivas para detectar de manera preliminar contenido para adultos dentro de una página de Internet, es realizar un análisis semántico, buscando como mínimo la aparición de una o varias palabras claves relacionadas de manera directa con dicho tipo de información; es importante 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.

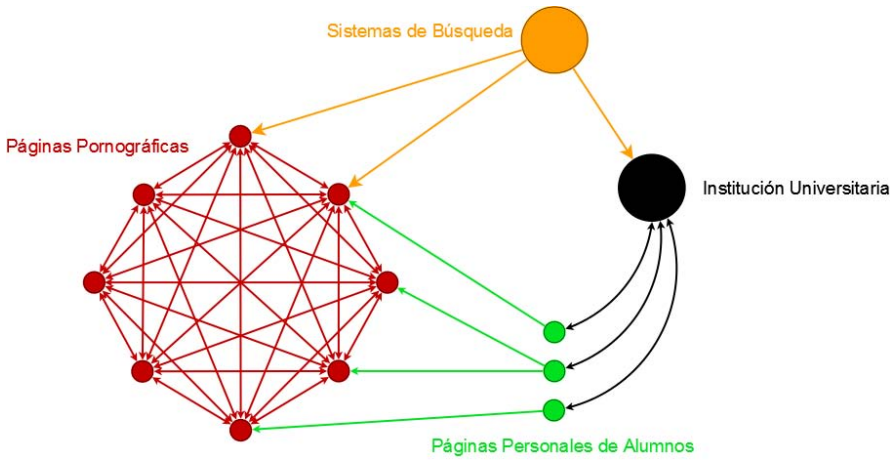



Figura 2.4.4. - Estructura general para la red de sitios pornográficos y su interconexión con otros sistemas.

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.

# Capítulo Tercero

## Algoritmo ActiveRank

# 3. Algoritmo ActiveRank

El algoritmo *ActiveRank* es propiedad de Ondore S.A. de C.V., se encuentra patentado en diferentes países y protegido por las leyes del derecho de autor y propiedad industrial; el uso que se le ha dado en el transcurso de esta investigación ha sido con intereses puramente académicos con permiso de la sociedad propietaria; es importante destacar que cualquier uso de la información o métodos aquí detallados puede tener consecuencias penales para los involucrados.

Los estudios preliminares del algoritmo fueron desarrollados por Fernando Luege Mateos y Rafael Peña Miller en Enero de 2005 como un método para autoorganizar información de bases de datos documentales en el proyecto no académico *Infoteca.org*; posteriormente Fernando Luege continuó los estudios y consolidación de la tecnología y en 2007 los derechos de patente fueron adquiridos por Ondore S.A. de C.V., donde Asaf Paris Mandoki contribuyó en gran medida a la implementación a nivel comercial de la tecnología. Ondore es una empresa totalmente mexicana especializada en sistemas de análisis de información y sistemas de búsqueda basados en teoría de gráficas.

## 3.1. Descripción general

ActiveRank es un método automático de clasificación, calificación y relación de información basado en gráficas, el cual construye una red dinámica a partir de relaciones numéricas, semánticas, conceptuales, etc., entre elementos de diferentes conjuntos estructuralmente semejantes. Mediante las propiedades topológicas de la red, se genera un vector de relaciones para cada elemento del conjunto, el cual permite analizar la relación entre ellos, mejorar y facilitar los procesos de clustering, obtener una medida de *ranking* dinámico e individual, analizar patrones de comportamiento, así como mantener un modelo autoevolutivo de la red a través de la interacción y retroalimentación de sus elementos y otras características.

En su expresión más básica, ActiveRank permite que dos conjuntos de información se relacionen a través de características comunes o por su interacción, midiendo y modificando dichos valores automáticamente, para construir una red dinámica; la medida de ranking que se obtiene entre dos elementos de la red se interpreta como la afinidad o cercanía de estos mismos, operando el conjunto de rankings se puede organizar la información en términos de su relevancia desde el punto de vista de uno de sus elementos. Dado que el algoritmo es genérico, se puede observar como el primer conjunto a un grupo de documentos textuales, y al segundo como el universo de usuarios o personas que interactúan con dicha información; la relación generada a partir de la

interacción de los usuarios con los documentos perfilan de manera automática a ambos conjuntos, y las diferentes medidas de ranking permiten determinar que documentos son de mayor interés para cada uno de los usuarios, así como conformar grupos de usuarios afines de manera automática, lo que resulta en la base de un sistema de redes sociales automático.

En resumen, algunos de los principales puntos del algoritmo ActiveRank son los siguientes:

- Genera una red de información con alto grado de autoorganización a partir de conjuntos de información no relacionados a partir de la interacción de sus elementos.
- A partir de la topología de la red es posible obtener medida de la similitud y relación de los elementos, utilizable en la selección y agrupación por relevancia y afinidad (clasificación).
- La red generada puede ser operada con todas las herramientas que provee la Teoría de Gráficas.
- Es integrable a cualquier plataforma de procesamiento de información en un esquema de "caja negra", recibiendo información estadística del sistema primario y devolviendo listas sobre el orden y la relevancia de la información de interés.
- El algoritmo puede ser implementado en cualquier lenguaje de programación, y su alcance estará limitado únicamente por los recursos de cómputo disponibles.

### 3.2. Operaciones básicas

Una red construida a través de ActiveRank presenta un alto grado de abstracción, donde los nodos pueden ser cualquier entidad de información, objeto o sujeto, y los vínculos la relación existente entre ellos, dada a partir de la interacción con el sistema y calculada por el propio algoritmo; la complejidad del sistema determinará la naturaleza de la implementación de los métodos a continuación descritos.

Para simplificar la explicación del algoritmo se describirá la implementación del mismo en un sistema de administración de información, donde un usuario interactúa con un conjunto de documentos que han sido manualmente relacionados a una lista de categorías.

Sea la gráfica $G$ conformada por los siguientes conjuntos:

$U(G) = \{u_1\}$ , los usuarios del sistema

$C(G) = \{c_1, c_2, c_3\}$ , las categorías de información disponibles

$D(G) = \{d_1, d_2\}$ , los documentos disponibles en el sistema

y las siguientes relaciones iniciales:

$V(G) = \{ v_1(u_1, c_1),$
$v_2(u_1, c_3),$
$v_3(d_1, c_1),$
$v_4(d_1, c_2),$
$v_5(d_2, c_2),$
$v_6(d_2, c_3) \}$ , todas bidireccionales y del mismo peso.

Seleccionando al conjunto *C* como los nodos comunes entre los conjuntos *U* y *D*, podemos expresar la gráfica *G* de forma matricial como se muestra a continuación:

| | $c_1$ | $c_2$ | $c_3$ |
|---|---|---|---|
| $u_1$ | 1 | 0 | 1 |
| $d_1$ | 1 | 1 | 0 |
| $d_2$ | 0 | 1 | 1 |

Tabla 3.2.1. – Gráfica *G* considerando al conjunto *C* como nodos comunes de *U* y *D*

donde las filas de la matriz ahora representan vectores de relación entre los elementos de los conjuntos *U* y *D* con los elementos del conjunto *C*.

Se define un vector ActiveRank $a_x$ como $a_x = \left(a_x{}^0, \ldots, a_x{}^{M-1}\right)$, donde $a_x^i$ corresponde al valor de peso o relación entre el elemento $a_x$ y el elemento *i* del conjunto de nodos al que se relaciona, en nuestro ejemplo, categorías. Se considera que el elemento $a_x$ está relacionado a la categoría *i* si $a_x^i > 0$. Todos los vectores ActiveRank se encuentran normalizados. Un vector $x = (x^0, \ldots, x^{M-1})$ se encuentra normalizado si $\sum_{j=0}^{M-1} x^j = 1$.

Continuando con nuestro ejemplo, la matriz de vectores ActiveRank $u_1$, $d_1$, $d_2$ sería:

| | $c_1$ | $c_2$ | $c_3$ |
|---|---|---|---|
| $u_1$ | 0.5 | 0 | 0.5 |
| $d_1$ | 0.5 | 0.5 | 0 |
| $d_2$ | 0 | 0.5 | 0.5 |

Tabla 3.2.2. – Matriz de vectores ActiveRank en su estado inicial

y su representación gráfica se puede apreciar en la figura 3.2.1.

La dinámica de la red se da a partir de la interacción entre dos elementos de la red; siguiendo el ejemplo propuesto, cuando un usuario descarga o utiliza un documento, ActiveRank redistribuye el peso de las relaciones que ambos comparten para incrementar la semejanza entre los perfiles de ambos elementos; lo anterior puede ser realizado unidireccionalmente o bidireccionalmente, es decir, que el perfil del usuario se vea afectado por el del documento, viceversa o ambos.

Sea $k \in (0,1]$ un parámetro arbitrario que define la velocidad de redistribución de peso, la operación de interacción del vector ActiveRank *u* sobre *a* se define como:

$$a^j_{\ new} = a^j + k\left(u^j - a^j\right)$$ , para cada elemento del vector *a*.

Definimos el *ranking* entre dos elementos de la red a partir de sus vectores ActiveRank *a* y *u* como:

$$\rho(a,u) = 1 - \tfrac{1}{2}\sum_{j=0}^{M-1}\left|a^j - u^j\right| \qquad \text{donde} \qquad \rho(a,u) \in [0,1] \qquad \text{y} \qquad \rho(a,u) = \rho(u,a)$$

Definimos a la matriz de rankings *R* como:

$R[i,j] = \rho(a_i, a_j)$, siendo $a_i$ y $a_j$ cualesquiera dos elementos de la red relacionados al mismo conjunto.

La matriz de rankings es cuadrada, de simetría triangular, con todos sus elementos de la diagonal principal unitarios. Para fines prácticos se puede trabajar con la sección triangular superior o inferior indistintamente, y suprimiendo la diagonal principal debido a que su interpretación directa es que la similitud de un elemento contra él mismo es 1, es decir, son idénticos ya que se trata del mismo elemento.

La figura 3.2.1. denota la estructura inicial de la gráfica en términos de sus vectores ActiveRank, y en línea punteada se pueden apreciar los valores de rankings entre los elementos de la red. La tabla 3.2.3. corresponde al valor inicial de la matriz de rankings.

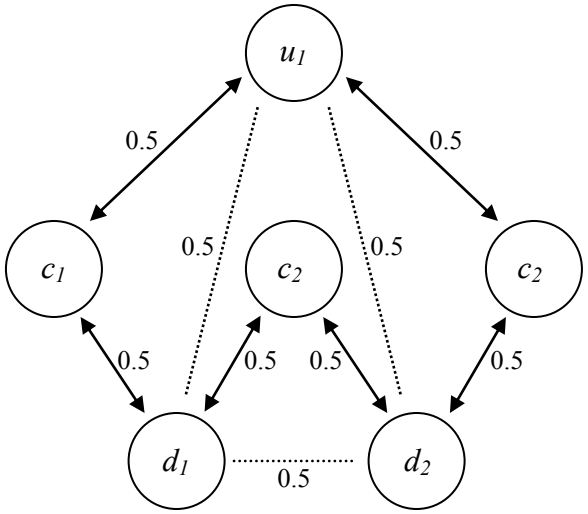



Figura 3.2.1. – Gráfica *G* en su estado inicial con medidas de ranking entre los elementos en línea punteada

| | $u_1$ | $d_1$ | $d_2$ |
|---|---|---|---|
| $u_1$ | 1 | 0.5 | 0.5 |
| $d_1$ | 0.5 | 1 | 0.5 |
| $d_2$ | 0.5 | 0.5 | 1 |

Tabla 3.2.3. – Matriz de rankings de ActiveRank para la red en su estado inicial

Siguiendo con el ejemplo, las tablas 3.2.4 y 3.2.5 muestran la evolución de la gráfica *G* después de una y tres interacciones respectivamente entre los elementos $u_1$ y $d_1$, con un coeficiente de interacción $k$=0.1. Lo anterior debe ser interpretado como que con cada interacción se redistribuirá el 10% de la diferencia de pesos entre los dos vectores ya sea en enlaces previamente existentes o en vínculos generados como consecuencia de dicho evento; el proceso de interacción lleva a una convergencia entre los dos vectores que interactúan, esto es que si dos vectores interactúan de manera repetida, cada vez serán más similares, lo que se puede observar en la tabla 3.2.6. tras el incremento del valor de ranking entre los elementos $u_1$ y $d_1$. Otra observación importante es que tras la modificación de un vector, todos los valores de ranking relacionados a este se ven afectados, a pesar de que bajo ciertas situaciones de simetría pudiera darse el caso de que se algunos valores se mantuvieran constantes (como en el caso de nuestro ejemplo); tras la modificación de un vector es necesaria la actualización de la matriz *R*.

| | $c_1$ | $c_2$ | $c_3$ |
|---|---|---|---|
| $u_1$ | 0.5 | 0.05 | 0.45 |
| $d_1$ | 0.5 | 0.45 | 0.05 |
| $d_2$ | 0 | 0.5 | 0.5 |

Tabla 3.2.4. – Matriz de vectores ActiveRank después de una interacción bidireccional de $u_1$ con $d_1$ con $k = 0.1$

| | $c_1$ | $c_2$ | $c_3$ |
|---|---|---|---|
| $u_1$ | 0.5 | 0.122 | 0.378 |
| $d_1$ | 0.5 | 0.378 | 0.122 |
| $d_2$ | 0 | 0.5 | 0.5 |

Tabla 3.2.5. – Matriz de vectores ActiveRank después de tres interacciones bidireccionales de $u_1$ con $d_1$ con $k = 0.1$

| | $u_1$ | $d_1$ | $d_2$ |
|---|---|---|---|
| $u_1$ | 1 | 0.744 | 0.5 |
| $d_1$ | 0.744 | 1 | 0.5 |
| $d_2$ | 0.5 | 0.5 | 1 |

Tabla 3.2.6. – Matriz de rankings de ActiveRank para la red en su estado inicial

En la figura 3.2.2. se puede apreciar claramente el cómo cambió la estructura de la red después de que el usuario interactuara con uno de los documentos, incrementando el valor de ranking entre los dos elementos involucrados, y generando nuevas relaciones que antes no existían. Este ejemplo debe ser extrapolado a cualquier número de elementos u conjuntos pertenecientes a la red.

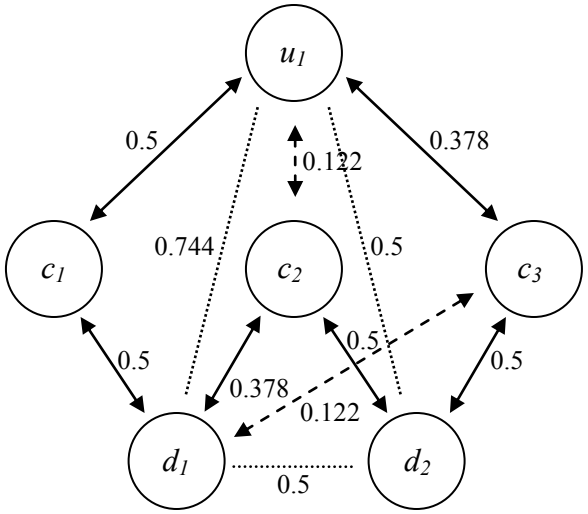



Figura 3.2.2. – Gráfica *G* después de 5 interacciones entre $u_1$ y $d_1$ con medidas de ranking entre los elementos en línea punteada y nuevas aristas generadas por la interacción

Todos los conceptos y metodologías anteriormente descritas pueden ser utilizadas variando los puntos de vista; por ejemplo, si en el escenario del caso estudiado anteriormente, en vez de existir categorías, existiera un segundo conjunto de usuarios no relacionados al conjunto *U*, se podrían considerar como elementos comunes los documentos con los que interactúan, y calcular los valores de ranking entre los elementos del conjunto *U* y los del nuevo grupo, o bien, se pueden cambiar documentos por productos comerciales de cualquier tipo, y entonces, considerando a los usuarios como elementos comunes entre los productos, realizar un estudio de la relación o similitud de dichos elementos utilizando la matriz de ranking obtenida.

### 3.3. Red de información generada a partir de ActiveRank

La gráfica de vectores ActiveRank permite incrementar de manera significativa los alcances de cualquier sistema de información gracias a diferentes elementos, como la construcción de relaciones no consideradas o inexistentes entre los elementos de la red así como su medición y ajuste automático, esto último ayuda a que métodos convencionales de teoría de gráficas como pueden ser los procesos de clustering operen de mejor manera al tener una red que ha sido desarrollada a partir de la interacción de sus elementos. Otro punto importante a destacar es que permite la incorporación de múltiples clases de elementos en una sola estructura, es decir, permite integrar redes que de otra manera serían estudiadas independientemente, pero que en la realidad, se encuentran relacionadas de maneras complejas y difíciles de determinar por métodos convencionales.

La estructura de la red de ActiveRank es aprovechada al implementar sobre ella cálculos como la matriz de rankings, procesos de clustering, patrones de subgráficas, entre otros, y su valor radica en su simplicidad y escalabilidad, teniendo la capacidad de ser integrada a sistemas de gran escala. Puede ser utilizada para generar y detectar perfiles de intereses de usuarios, como analizador semántico, numérico y estadístico, entre otros. Dado que es un sistema autoregulado por la interacción, su eficiencia y exactitud crece significativamente mientras más elementos contenga la red, o para fines prácticos, el sistema donde haya sido implementado; siguiendo con nuestro ejemplo, la diversidad de usuarios y documentos en un sistema de análisis de información produce un mejor comportamiento en el perfilamiento, clasificación y explotación de los recursos.

La figura 3.2.3. es la representación gráfica de la red de categorías de Infoteca.org donde el diámetro de los nodos denota su importancia para el usuario en cuestión, es decir, a partir de la medida de ranking de dicho nodo con respecto al usuario se obtiene la relevancia del mismo. A su vez, las relaciones entre categorías fueron generadas de manera dinámica utilizando los documentos como nodos comunes entre los usuarios y las categorías y su valor de relación como el ranking entre los elementos del mismo conjunto, un acercamiento alterno pero análogo al planteado como ejemplo anteriormente.

Gracias a que la red de ActiveRank puede ser trabajada con un alto grado de abstracción, es posible implementar el sistema sin siquiera saber de qué tipo de información o elemento de la red se está trabajando en un momento dado; la tecnología de ActiveRank desarrollada por Ondore simplemente recibe información estadística y entrega listas de relaciones ordenadas a partir del ranking según lo requiera el sistema primario, lo que otorga un alto grado de seguridad informática, pues toda la información contenida en la red no es humanamente traducible.

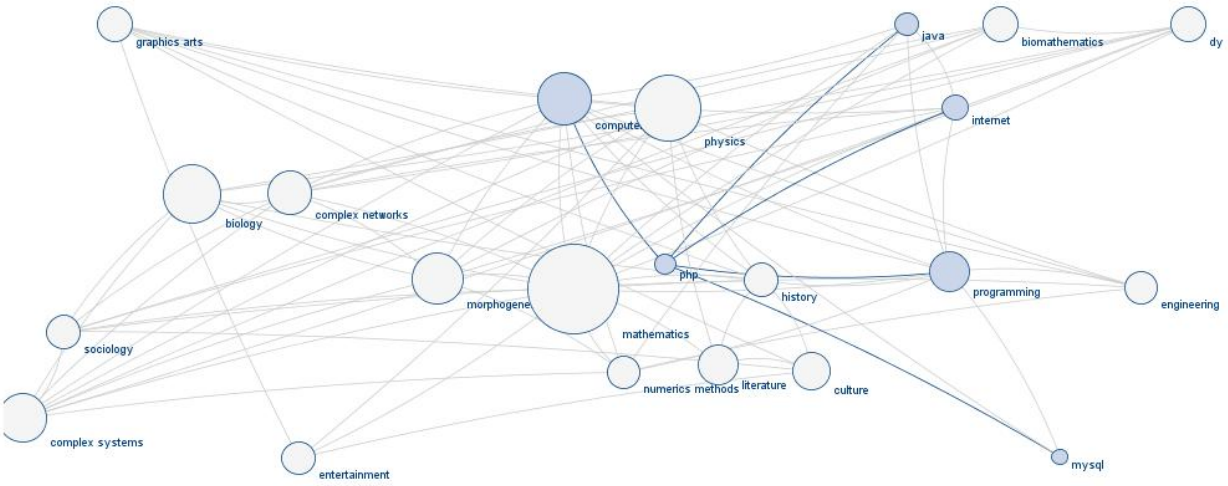



Figura 3.2.3 – Red de categorías en Infoteca.org con diámetros proporcionales a su relevancia para el usuario – 2005 *[Luege/Peña]*

### 3.4. Manejo de información utilizando ActiveRank

La red de información generada a partir de ActiveRank es operable a través de los métodos convencionales de teoría de gráficas, sin embargo, la diferencia radica en cómo han sido construidas las relaciones a través de ellas y que a partir de la matriz de rankings se pueden construir subredes interpretables y analizables de manera independiente; como se aprecia en las figuras 3.2.1. y 3.2.2., las medidas de rankings pueden ser consideradas como valores de peso para generar nuevos vínculos entre nodos no conexos, y de esta manera, analizar el comportamiento de nuevas subgráficas.

Condensando una gráfica de estructura análoga a la del ejemplo planteado en la sección 3.2. de este documento, tres usuarios ($u_1$ a $u_3$) y tres documentos ($d_1$ a $d_3$) mantienen un valor de ranking entre ellos denotado como la línea punteada en la figura 3.2.3.[3]; si quisiéramos generar una gráfica de usuarios, lo único necesario sería considerar los valores de ranking ahora como valores de peso no normalizados, generar los vectores ActiveRank al normalizar las relaciones de cada usuario en forma vectorial, y obtener la medida de ranking entre estos elementos gracias a las nuevas relaciones con el conjunto alterno ($d_1$ a $d_3$). La figura 3.2.3. presenta la evolución de la gráfica $A$ al normalizar los valores de ranking entre los elementos en cuestión para generar vectores ActiveRank.

3. Para este ejemplo los valores de ranking han sido generados de manera aleatoria, pero es importante destacar que se obtendrían siguiendo la metodología explicada en la sección 3.2. de este documento.

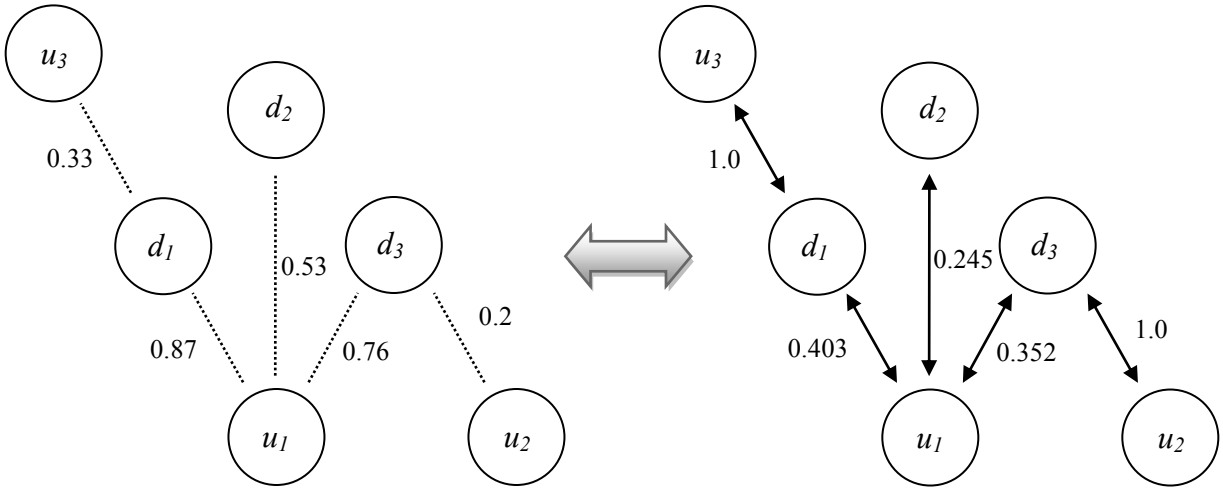



Figura 3.2.3. – Evolución de una gráfica *A* al pasar de rankings a vectores ActiveRank de su conjunto de usuarios.

Y ahora se obtiene una nueva matriz de rankings de usuarios para la nueva subgráfica, al considerar el conjunto alterno como elementos comunes entre los nodos antes mencionados. La tabla 3.2.7. expresa la relación entre usuarios.

| | $u_1$ | $u_2$ | $u_3$ |
|---|---|---|---|
| $u_1$ | 1 | 0.352 | 0.403 |
| $u_2$ | 0.352 | 1 | 0 |
| $u_3$ | 0.403 | 0 | 1 |

Tabla 3.2.7. – Matriz de rankings entre usuarios de la gráfica *A*.

Y su representación gráfica sería como se muestra en la figura 3.2.4.

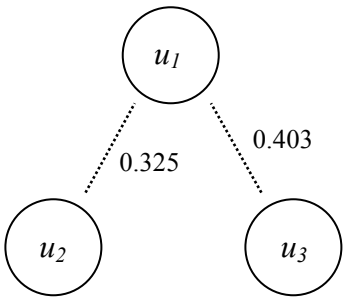



Figura 3.2.4. – Subgráfica de usuarios de la gráfica *A*.

Ahora puede ser fácilmente apreciable el alto valor que tuvo el utilizar ActiveRank en el manejo y procesamiento de la red en cuestión. Si tuviéramos que recomendarle al usuario 2 $u_2$ otro usuario para compartir información según sus intereses, ordenando de mayor a menor el resto del subconjunto a partir de sus medidas de ranking con el usuario 2 en este caso, obtendríamos una lista con los mejores candidatos para resolver el problema; agrupando a los usuarios bajo este mismo criterio a partir de una operación convencional de clustering, podríamos generar de manera automática grupos sociales que comparten intereses comunes, cuando en ningún momento se tuvo información sobre estas características, fue construida de manera automática.

De manera análoga se podría generar la gráfica de documentos, y su matriz de rankings sería la expresada en la tabla 3.2.8.

| | $d_1$ | $d_2$ | $d_3$ |
|---|---|---|---|
| $d_1$ | 1 | 0.725 | 0.725 |
| $d_2$ | 0.725 | 1 | 0.792 |
| $d_3$ | 0.725 | 0.792 | 1 |

Tabla 3.2.8. – Matriz de rankings entre documentos de la gráfica $A$.

Y su gráfica sería aquella representada en la figura 3.2.5.

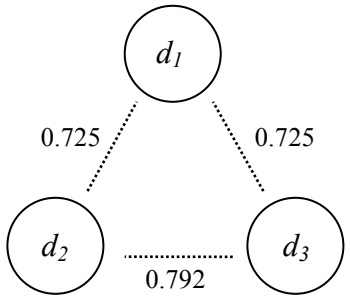



Figura 3.2.5. – Subgráfica de documentos de la gráfica $A$, donde casualmente el ranking entre los elementos $d_1 - d_2$ y $d_1 - d_3$ tienen el mismo valor.

De la misma forma que con los usuarios, ahora podemos utilizar la nueva matriz para agrupar por afinidad los diferentes documentos de la gráfica original, ya fuera a través de un proceso de clustering convencional al considerar el ranking entre documentos de nueva cuenta como el peso del vínculo entre cada

uno de ellos[4], o a través del ordenamiento de los valores de ranking para un documento en particular.

Una de las primeras implementaciones de ActiveRank como algoritmo gestor de información y motor de búsqueda fue realizada en el 2005 en el proyecto Infoteca.org *[Luege/Peña]*; la idea básica era relacionar manualmente un conjunto de documentos y usuarios a un universo finito de categorías de información (*matemáticas*, *ciencias sociales*, *ingeniería*, *computación*, etc.), a partir de la interacción entre usuarios y documentos, la estructura de la red se modificaría automáticamente para reclasificar todos los documentos y usuarios, obteniendo mejores perfiles de ambos conjuntos. Al considerar desde otro punto de vista a los documentos (podría haberse hecho de igual manera considerando a los usuarios) como elementos comunes de las categorías, fue posible generar una red de categorías, utilizando los valores de ranking como el peso de la relación entre ellas; la figura 3.2.6. denota la estructura de la red de categorías obtenida tras realizar el experimento anteriormente descrito. Si bien se puede apreciar que para este punto ActiveRank permitió la depuración automática de las clasificaciones inicialmente propuestas, así como la generación de una nueva red entre elementos de un conjunto dado, su valor no sería representativo si no pudiéramos constatar su correcto funcionamiento, y una de las pruebas realizadas fue la agrupación de nodos a partir de un algoritmo de clustering convencional sobre la red generada, cuyo resultado gráfico puede ser apreciado en la figura 3.2.7. donde las categorías similares fueron agrupadas.

4. Es importante denotar que en este punto ya no es necesario normalizar de nuevo los valores de relación entre los elementos, ya que se realizará una operación de clustering convencional donde no es indispensable dicha preparación previa.


Figura 3.2.6. – Red de categorías desagrupada en Infoteca.org – 2005 *[Luege/Peña]*

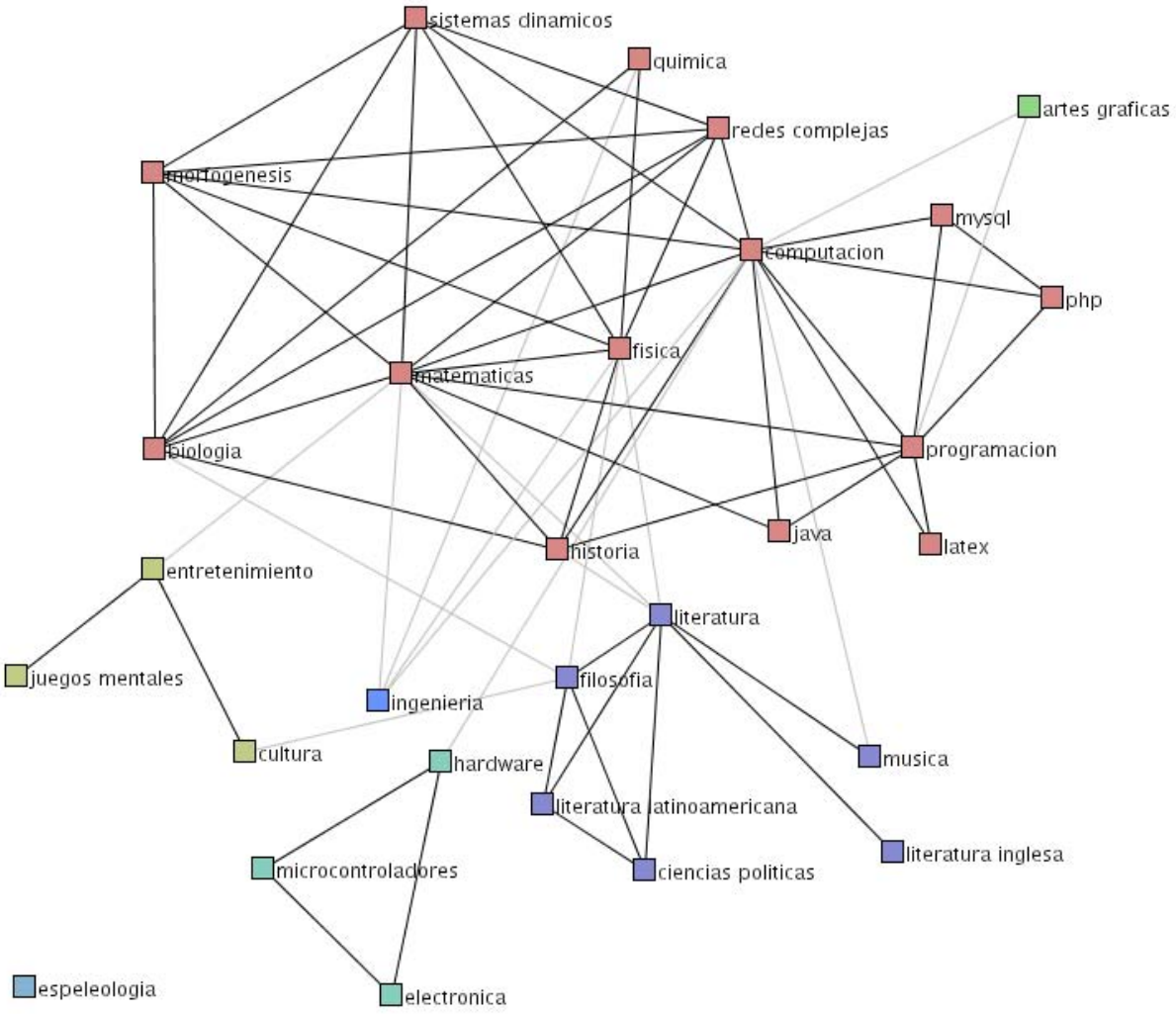



Figura 3.2.7. – Red de categorías agrupadas en Infoteca.org – 2005 *[Luege/Peña]*

Como conclusión a las ideas desarrolladas en este capítulo lo más importante de la utilización del algoritmo ActiveRank en el análisis y conformación de redes que relacionen diferentes conjuntos es que a través del ordenamiento de la matriz de rankings pueden ser calificadas de forma cualitativa y cuantitativa las relaciones y elementos más importantes de la gráfica, generar y desglosar la red es subconjuntos de interés, así como crear una estructura dinámica estudiable desde la amplia perspectiva de teoría de gráficas. La principal observación es el cómo los valores de rankings pueden ser interpretados como pesos de vínculos para la condensación y creación de nuevos vectores ActiveRank según sea la conveniencia; de igual forma, cada uno de los vectores de la matriz de rankings puede ser graficado, característica que será ampliamente explotada en el análisis de los resultados obtenidos más adelante.

# Capítulo Cuarto

## Trabajo Experimental

# 4. Trabajo Experimental

El centro de esta tesis es estudiar el comportamiento del algoritmo ActiveRank como método para la clasificación y detección automática de contenido; a continuación se describe la estructura de los sistemas de análisis de información utilizados, las condiciones y entornos de análisis, así como la metodología utilizada para la exploración de redes de información y la obtención de resultados.

## 4.1. Descripción de sistemas de análisis de información y escenarios analizados

### 4.1.1. Arquitectura de sistemas y escenarios analizados

Retomando los conceptos definidos en la sección 2.3. de este documento, a continuación se enumeran las características de los sistemas de Ondore S.A. de C.V. utilizados en el desarrollo experimental de esta tesis:

- *Ondore Analyzer v4.0.2*: Este sistema es la base tecnológica utilizada por Ondore para la exploración y recolección de información en la WWW, se encuentra totalmente integrado con los módulos de análisis basados en el algoritmo ActiveRank, y se compone de los siguientes subsistemas:

  Sistemas Primarios

  + *HTTPManager*: Permite la gestión y utilización de los enlaces a bajo y medio nivel entre la aplicación y los servidores web que hospedan la información objetivo; el resultado que provee es directamente el código fuente de la página web en cuestión.
  + *DBManager*: Es el subsistema gestor de base de datos y permite la correcta y segura lectura escritura de las tablas contenidas en dichas bases de datos a través de sus múltiples métodos. Una característica importante es que su arquitectura permite integrar una amplia variedad de tecnologías de base de datos, la utilizada en este experimento es MySQL, así como realizar una gran cantidad de validaciones para guardar la fiabilidad de los datos almacenados, lo cual resulta fundamental en un sistema de mediana escala como es el caso.

+ *ContentAnalyzer*: Compuesto a su vez de una amplia variedad de subsistemas, es el núcleo de la aplicación y realiza en primer término el procesamiento de la estructura de vínculos e información HTML contenida en la página así como el primer filtrado y depuración de la información textual, posteriormente realiza análisis semánticos y estadísticos de la información, y si se está utilizando ActiveRank, alimenta al motor del algoritmo para generar los vectores iniciales para cada uno de los documentos.

  Sistemas Secundarios

+ *ContentIntegrator*: En caso de ser uno de los objetivos del análisis, a través de este sistema es posible almacenar y reconstruir documentos, imágenes, y en términos generales cualquier tipo de contenido incorporado a las páginas analizadas. Resulta extremadamente útil si se desea centralizar una base de datos documental a partir de contenido web.
+ *Snapper*: Permite realizar un *snapshot* de la página web analizada, desde términos gráficos (obteniendo una imagen de la pantalla que un usuario vería en su navegador) hasta generar una copia fiel de la estructura de directorios y archivos de dependencia para generar una imagen *cache* del contenido.

La configuración utilizada para los experimentos realizados en esta tesis ha sido utilizando únicamente los sistemas primarios con ActiveRank, no se ha almacenado ningún tipo de documento relacionado dado que en la mayoría de los casos sería la concentración de contenido pornográfico que no tiene relevancia alguna para los objetivos de esta tesis.

- *Ondore DARE (Distributed ActiveRank Engine) v2.0*: Es el sistema de cómputo distribuido que soporta el algoritmo ActiveRank; en términos generales consiste en un sistema central de control y múltiples clientes de procesamiento, incorpora redundancia de datos y cálculos, así como una arquitectura de caja negra para su fácil integración con cualquier otro sistema. La descripción detallada del funcionamiento de este sistema es un tema de estudio por sí solo.
- *Ondore CustomGrapher v1.0*: Es un sistema genérico de graficación a través del cual se analiza visualmente el comportamiento de la distribución de los valores de la matriz de ranking; es posible trabajar con otros sistemas de procesamiento matricial como MatLab integrando el motor de ActiveRank para facilitar la obtención de datos.

Los escenarios utilizados para el análisis del desempeño del algoritmo ActiveRank en esta tesis se pueden dividir en dos grupos, el primero corresponde la exploración de conjuntos antagónicos de páginas web para la obtención de la matriz de rankings que los relacione y nos permita comparar el comportamiento de cada uno de los valores de ranking, con respecto a una referencia, a fin de encontrar un valor divisor o clasificador entre los conjuntos, en este caso de páginas web pornográficas contra sitios de Internet sobre religión católica y judía y un tercer conjunto conformado por páginas de Wikipedia. En el segundo escenario, utilizando los valores de referencia obtenidos en la etapa anterior, se analizará de manera abierta el dominio unam.mx (podría ser cualquier otro) con el fin de encontrar contenido pornográfico automáticamente en un proceso posterior al análisis del contenido. En ambos casos se realizará una revisión estadística de la eficiencia del algoritmo al clasificar contenido pornográfico la cual será utilizada durante el proceso para ajustar los valores de referencia explicados más adelante.

### 4.1.2. Infraestructura para el procesamiento y almacenamiento de información

A continuación se describen las características generales de la infraestructura utilizada en la realización de esta tesis gracias al apoyo de Ondore.

- *Red local y enlaces a Internet*:
  - + Enlace de entrada/salida a Internet conformado por 2 líneas ADSL de 4 [Mbps] (ISP: Telmex) balanceadas a través de un router Soekris modelo 5501 conectado a través de un puerto 100Mbps Ethernet al switch.
  - + Red local Gigabit Ethernet; cableado estructurado con cable UTP categoría 6, todos los equipos utilizando puertos Gigabit Ethernet incorporados directamente en la tarjeta madre conectados todos a un switch Gigabit Ethernet administrado de marca Dell modelo PowerConnect 2748 en topología de estrella.
- *Analyzer*:
  - + *Servidor central de análisis (1 unidad)*:

    Procesador: Intel Core 2 Duo 2.8 [GHz] 64 [b]

    RAM: 2 [GB] + SWAP 1 [GB]

    HD: 120 [GB]

    SO: Linux Ubuntu 8.10 (Intrepid) kernell 2.6.27-14-generic

    Tomcat: v6.0.18

Apache: v2.2.9

JAVA: v1.6.0_10

MySQL: v14.12 distribución 5.0.67

- *Distributed ActiveRank Engine*:
  - + *Servidor central (1 unidad)*:

    Procesador: Intel Core 2 Duo 2.8 [GHz] 64 [b]

    RAM: 2 [GB] + SWAP 4 [GB]

    HD: 65 [GB]

    SO: Linux Debian 2.6.26-1-amd64

    Apache: v2.2.9

    JAVA: v1.6.0_12

    MySQL: v14.12 distribución 5.0.51a
  - + *Servidores clientes (4 unidades)*:

    Procesador: Intel Core 2 Duo 2.8 [GHz] 64 [b]

    RAM: 2 [GB] + SWAP 1 [GB]

    HD: 120 [GB]

    SO: Linux Ubuntu 8.10 (Intrepid) kernell 2.6.27-14-generic

    Tomcat: v6.0.18

    Apache: v2.2.9

    JAVA: v1.6.0_10

    MySQL: v14.12 distribución 5.0.67

### 4.1.3. Función de ActiveRank en el proceso de análisis e indexación

La función de ActiveRank en su actual implementación es generar los perfiles de cada documento indexado por el sistema Analyzer y generar la matriz de rankings correspondiente a la relación entre todos los documentos analizados. En este caso, la red generada por ActiveRank es una gráfica que relaciona de manera directa dos conjuntos de nodos, conformados por los documentos, y a que palabras se encuentran relacionados ponderando este último valor a partir del número de apariciones de la misma dentro del documento. Para la obtención de la matriz de rankings, las palabras serán consideradas los nodos comunes entre los documentos para así poder obtener la relación entre los últimos.

Se trabajará con una versión simplificada de rankings de ActiveRank donde el intervalo de valores es *[0,2]* donde *0* representa máxima similitud, y *2* representa ninguna similitud.

Como ya se ha mencionado, el proceso de clasificación consiste en obtener un valor de ranking de referencia a partir del cual poder tomar una decisión sobre si un nuevo documento pertenece o no a un conjunto determinado; en el presente trabajo, una vez generada la matriz de rankings completa entre dos ó tres conjuntos antagónicos, en este caso documentos pornográficos vs católicos, se selecciona un documento de referencia (ej. una página pornográfica) y se obtiene la fila de rankings de esta referencia, cuya interpretación es la similitud de este elemento contra todos los demás; la hipótesis consiste en que la medida de ranking de dicha referencia con respecto a sus semejantes sería alta (denotando similitud), y de menor magnitud para todos aquellos diferentes. A través de un proceso estadístico de ajuste posteriormente descrito, se obtendrá un valor de ranking que le permita al sistema clasificar información (en este caso pornográfica) en posteriores análisis de redes de información, ya sea la red de una institución limitada a través del nombre de dominio o una exploración abierta de la WWW.

## 4.2. Metodología de exploración, procesamiento y análisis de información

A continuación se describen los protocolos experimentales realizados para la etapa de calibración (obtención de vector de referencia y valor de ranking clasificador) y la etapa de prueba en un subconjunto acotado de la WWW.

Etapa 1 – Creación del vector de referencia y valor de ranking clasificador

1. Se selecciona manualmente un conjunto inicial de páginas pornográficas que se entiende son claramente de dicha clase y además concentran una gran cantidad de vínculos a otras de su mismo tipo y se insertan como valores iniciales en una nueva instancia del sistema Analyzer.
    a) *http://www.xxxvogue.net/*
    b) *http://www.sunporno.com/multi/*
    c) *http://www.peepingtom.com/*
    d) *http://www.jennymovies.com/*
    e) *http://www.porncity.net/*
    f) *http://www.bunnypost.com/*
    g) *http://xxxdessert.com/*
    h) *http://www.rawthumbs.com/*

i) *http://www.lovefuckk.com/*
j) *http://www.lamalinks.com/*
k) *http://www.twilightsex.com/*
l) *http://www.galleries4free.com/*
m) *http://www.redhothoneys.com/*
n) *http://www.sleazyland.com/*
o) *http://www.annasdungeon.com/*
p) *http://www.jasminerouge.com/*
q) *http://www.gigagalleries.com/*
r) *http://www.movieisle.com/*

Es importante identificar los vectores ActiveRank asociados a los registros iniciales ya que serán utilizados en el paso 3.

2. Se inicia el sistema Analyzer en su configuración básica (no se almacena ninguna clase de contenido multimedia) hasta cubrir una cuota de 3000 documentos analizados y sus respectivos vectores ActiveRank; en ese momento se detiene el sistema Analyzer.
3. Se suman (operación convencional de suma vectorial en geometría euclidiana) los vectores ActiveRank asociados a los registros iniciales del paso 1 y su resultado se normaliza, este nuevo vector ActiveRank es insertado manualmente al DARE y se identifica como el vector pornográfico de referencia que será utilizado más adelante.
4. Se repiten los pasos 1, 2 y 3 con un nuevo conjunto de registros iniciales descritos a continuación correspondientes a información sobre religión católica y judía, de nueva cuenta es importante poder diferenciar posteriormente el nuevo conjunto de información a analizar, un método simple es ubicar el identificador inicial y final de los vectores ActiveRank generados.
    a) *http://www.regnumchristi.org/english/*
    b) *http://www.catholic.net/*
    c) *http://www.vatican.va/phome_en.htm*
    d) *http://www.disciples.org/*
    e) *http://www.cofe.anglican.org/*
    f) *http://www.anglicancatholic.org/*
    g) *http://www.jewishencyclopedia.com/view.jsp?artid=52&letter=N*
5. Se repiten los pasos 1, 2 y 3 con un nuevo conjunto de registros iniciales descritos a

continuación correspondientes a información enciclopédica de temas variados seleccionados por una persona seleccionada aleatoriamente ajena al desarrollo de esta tesis; se aplican las mismas condiciones de identificación que en el paso 4.

a) *http://en.wikipedia.org/wiki/Budapest*
b) *http://en.wikipedia.org/wiki/Chill-out_music*
c) *http://en.wikipedia.org/wiki/WWII*
d) *http://en.wikipedia.org/wiki/Midnight_sun*
e) *http://en.wikipedia.org/wiki/Michael_Jackson*
f) *http://en.wikipedia.org/wiki/Train*
g) *http://en.wikipedia.org/wiki/Novel*
h) *http://en.wikipedia.org/wiki/Psycology*
i) *http://en.wikipedia.org/wiki/Eiffel_Tower*
j) *http://en.wikipedia.org/wiki/Mexican*

6. Utilizando el sistema DARE, se obtienen las filas de la matriz de rankings de ActiveRank correspondientes a los valores de rankings de los 3 vectores de referencia contra todos los demás.
7. Generar 3 gráficas de dispersión de puntos para cada una de las filas obtenidas en el punto 7 donde sea posible comparar de manera visual los niveles de ranking de cada uno de los 3 conjuntos de fuentes de información con respecto a cada uno de los 3 vectores de referencia.
8. Obtener el valor promedio de los valores de ranking asociados al conjunto de documentos pornográficos con respecto al vector pornográfico de referencia. En términos generales:

$$\bar{r}_x = \frac{1}{N}\sum_{i=1}^{N} \rho\left(v_{ref}, v_i\right),$$

donde, $N$ es el número de elementos del conjunto dado de documentos.

$v_{ref}$ es el vector de referencia.

$v_i$ es el i-ésimo vector del conjunto dado.

$\rho\left(v_{ref}, v_i\right)$ es el ranking entre el i-ésimo vector y el de referencia.

$\bar{r}_x$ es el valor promedio de valores de ranking asociados al conjunto dado de documentos con respecto al vector de referencia.

9. Se repite el paso 8 para obtener los valores promedio de ranking de los conjuntos de documentos católicos y enciclopédicos con respecto al vector pornográfico de referencia.
10. Se define como valor inicial de umbral de clasificación el valor medio entre el valor promedio de ranking del conjunto de documentos pornográficos con respecto al vector pornográfico de referencia y el valor más alto de los dos obtenidos de los valores promedio de ranking de los conjuntos católico o enciclopédico.[5] Lo anterior se expresa como:

$$r_{umbral} \begin{cases} \dfrac{\left|\bar{r}_{porno} + \bar{r}_{cat\,ólico}\right|}{2} & si \quad \bar{r}_{cat\,ólico} < \bar{r}_{enciclop\ édico} \\ \dfrac{\left|\bar{r}_{porno} + \bar{r}_{enciclop\ édico}\right|}{2} & caso\ contrario \end{cases}$$

donde, $\bar{r}_{porno}$ es el valor promedio de valores de ranking asociados al conjunto de documentos pornográficos con respecto al vector pornográfico de referencia.

$\bar{r}_{cat\,ólicos}$ es el valor promedio de valores de ranking asociados al conjunto de documentos católicos con respecto al vector pornográfico de referencia.

$\bar{r}_{enciclop\ édicos}$ es el valor promedio de valores de ranking asociados al conjunto de documentos enciclopédicos con respecto al vector pornográfico de referencia.

$r_{umbral}$ es el valor umbral de ranking de clasificación.

En teoría para todos los casos, el valor de ranking promedio correspondiente a los vectores del conjunto de información dado con respecto a su propio vector de referencia será menor que el valor de ranking promedio de cualquier otro conjunto de vectores contra el mismo vector de referencia. En el contexto de esta tesis lo anterior se describe como:

$$\bar{r}_{porno} < \bar{r}_{cat\,ólicos} \quad \text{y} \quad \bar{r}_{porno} < \bar{r}_{enciclop\ édicos}$$

11. Se realiza un proceso de evaluación manual de la eficiencia de clasificación del sistema como se describe a continuación:

5. Recordemos que estamos trabajando con una versión simplificada de la operación de ranking del algoritmo ActiveRank cuyo intervalo es [0,2] y su interpretación se encuentra invertida a la descrita en el capítulo 3 de esta tesis. La selección del menor valor de los rankings promedios de los conjuntos católico o enciclopédico representa el peor caso al analizar su similitud con documentos pornográficos.

11.1. Se obtiene una muestra de aleatoria del 2%[6] del tamaño del universo compuesto por los 3 conjuntos de información.

11.2. El sistema divide la muestra automáticamente en dos secciones, documentos *pornográficos*, para aquellos cuyo ranking con respecto al vector pornográfico de referencia se encuentre por debajo o en el umbral de clasificación, y como *no pornográficos* para aquellos que se encuentren por encima del mismo.

11.3. Se evalúa manualmente el número de falsos verdaderos en la clasificación *pornográficos*/*no pornográficos*. Si se disminuye el valor del umbral, la eficiencia de clasificación como *pornográficos* será muy alta, pero la eficiencia de clasificación como *no pornográficos* se verá afectada; en el caso contrario, cuando el umbral tiene un valor demasiado alto, la clasificación de *no pornográficos* será muy buena pero la de *no pornográficos* decrecerá significativamente.

12. Se ajusta el valor de umbral y repetir el paso 11 hasta obtener una eficiencia de clasificación homogénea entre *pornográficos*/*no pornográficos* y superior al 90%[7] en ambos casos.
13. En este punto se concluye la etapa 1 con la obtención de un vector pornográfico de referencia y su valor de umbral asociado para la correcta clasificación de contenido como *pornográfico*/*no pornográfico*. Se realizan conclusiones sobre el desempeño de la tecnología.

Etapa 2 – Análisis del dominio *unam.mx*

1. En una nueva instancia del sistema Analyzer, se inserta el conjunto inicial de páginas referentes al dominio *unam.mx*.
   a) *http://www.unam.mx*
2. Se inicia el sistema Analyzer en su configuración básica (no se almacena ninguna clase de contenido multimedia) hasta cubrir una cuota de 5000 documentos analizados y sus respectivos vectores ActiveRank; en ese momento se detiene el sistema Analyzer. El

6. El tamaño de la muestra puede ser reducido o incrementado posteriormente dependiendo de la eficacia del método de selección del valor inicial del umbral de clasificación.
7. Se determina 90% como un valor mínimo de eficiencia aceptable para un sistema de clasificación automático de información.

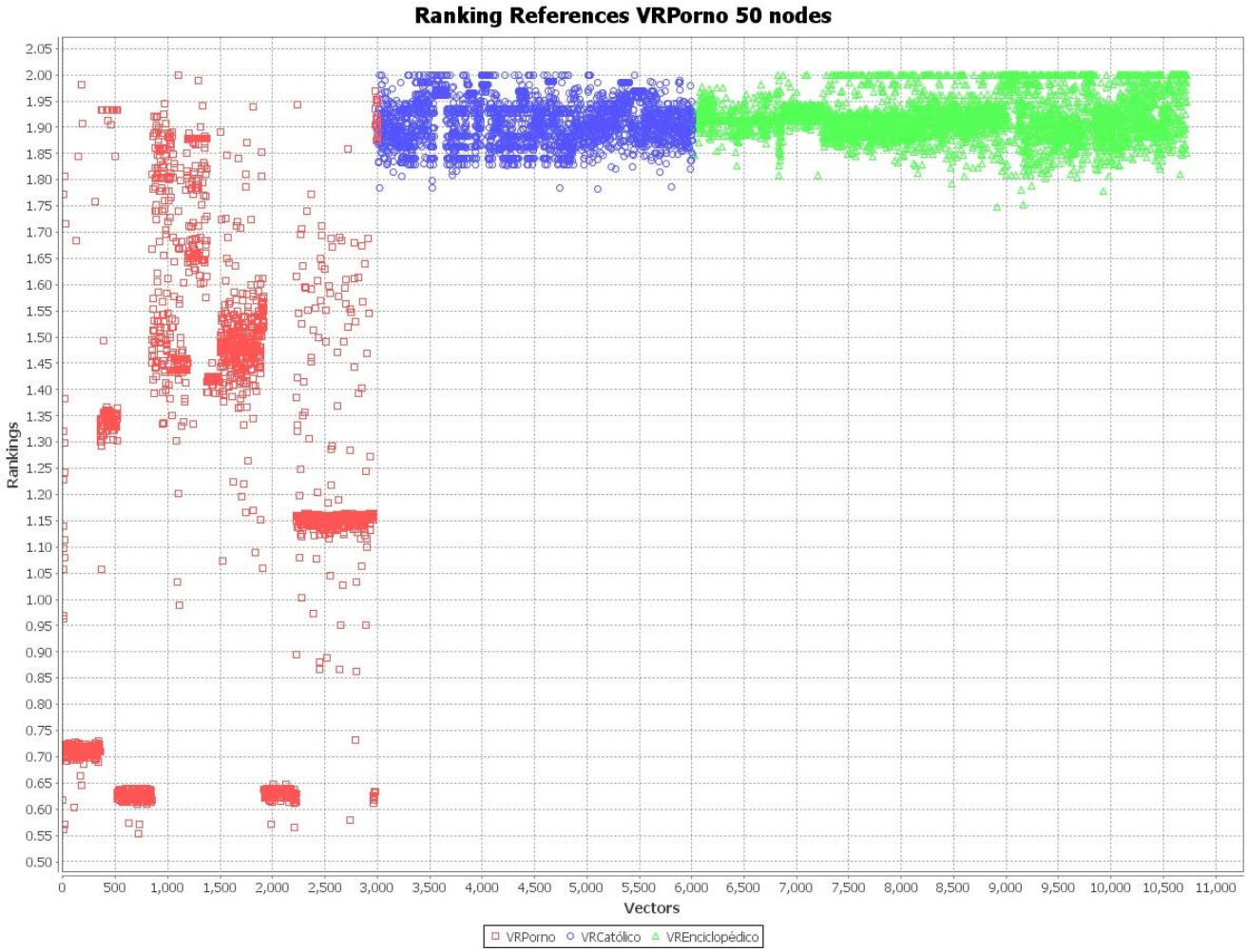
nivel de indexación para sitios pertenecientes al dominio *unam.mx* será de 3, y sólo serán analizados dominios bajo *unam.mx*.

3. Se inserta manualmente al DARE el vector pornográfico de referencia obtenido en la etapa 1.
4. Utilizando el sistema DARE, se obtiene la fila de la matriz de rankings de ActiveRank correspondiente a los valores de ranking del vector pornográfico de referencia contra todos los demás.
5. Se realiza un proceso de evaluación manual de la eficiencia de clasificación del sistema como se describe a continuación:
    5.1. Se obtiene una muestra aleatoria representativa del universo compuesto por los documentos obtenidos que presenten un ranking mayor al umbral obtenido en la etapa 1.
6. Evaluar manualmente el número de falsos verdaderos en la clasificación *pornográficos*/*no pornográficos*.
7. Concluir sobre la presencia de contenido pornográfico o rutas al mismo desde el dominio *unam.mx* y el funcionamiento del algoritmo ActiveRank como sistema de detección de contenido pornográfico en redes de información.

## 4.3. Resultados Obtenidos

A continuación se despliegan gráficamente las filas de la matriz de rankings obtenidas tras el desarrollo experimental, así como algunas tablas sobre información estadística importante para la realización del análisis de resultados. Información adicional se puede encontrar en el Apéndice D de la presente tesis.

### Resultados Etapa 1




Gráfica 4.3.1. - Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector de referencia pornográfico contra todos los demás.

La gráfica 4.3.1. presenta el ranking del vector de referencia pornográfico contra todos los elementos de la red ActiveRank; resulta fácilmente apreciable la diferencia en la distribución de valores de similitud entre los documentos del conjunto pornográfico con respecto a los pertenecientes a los otros dos entornos.

A continuación, la tabla 4.3.1. presenta los valores máximos, mínimos y la media de la distribución de valores de ranking para cada entorno de información analizado con respecto al vector de referencia porno.

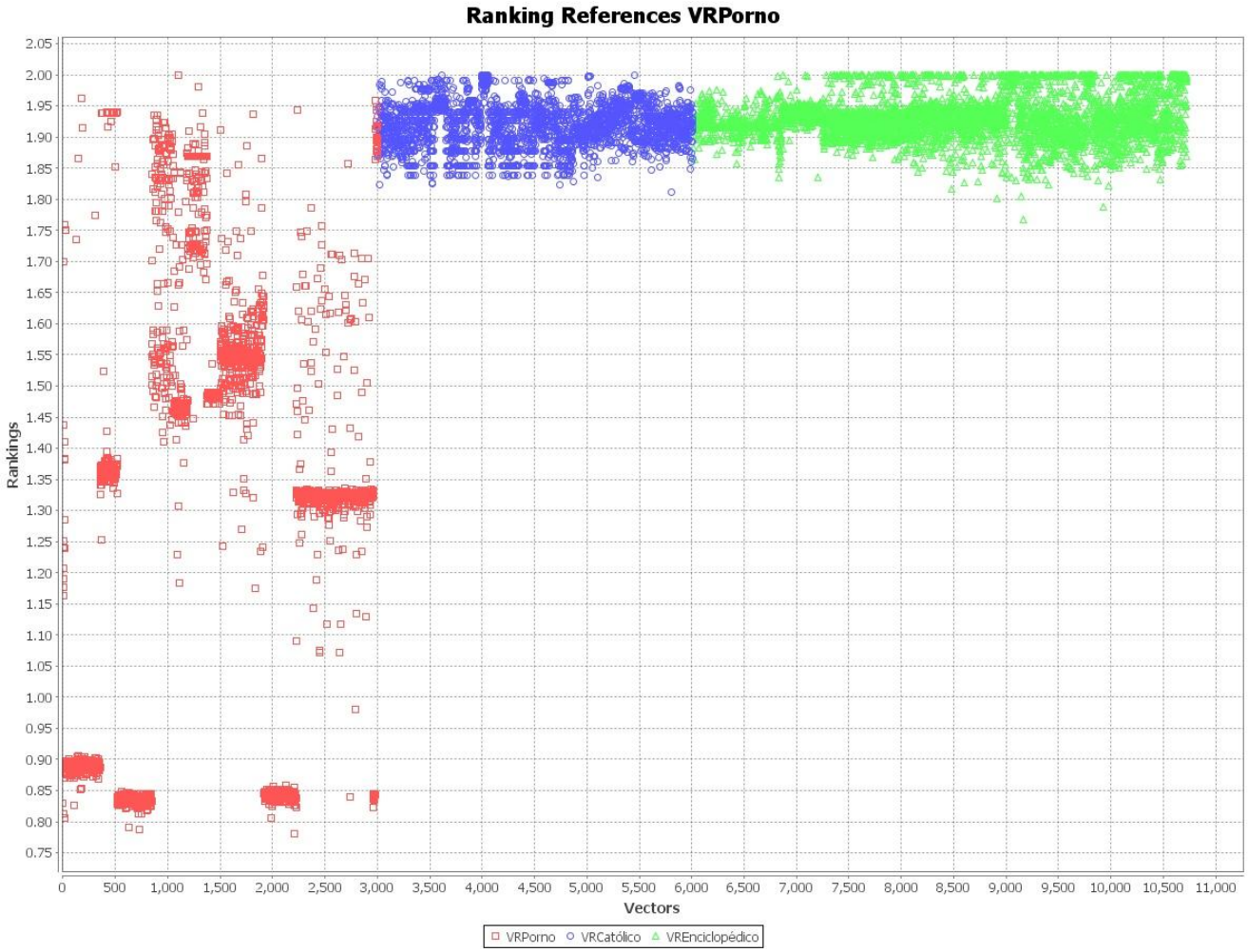
Tabla 4.3.1. – Distribución de valores máximos, mínimos y medios para cada entorno de información con respecto al vector de referencia pornográfico.

| | $\bar{r}$ | $\bar{r}_{min}$ | $\bar{r}_{max}$ |
|---|---|---|---|
| **Entorno Pornográfico** | 1.1543 | 0.5539 | 2 |
| **Entorno Católico** | 1.9101 | 1.7816 | 2 |
| **Entorno Enciclopédico** | 1.9206 | 1.7477 | 2 |

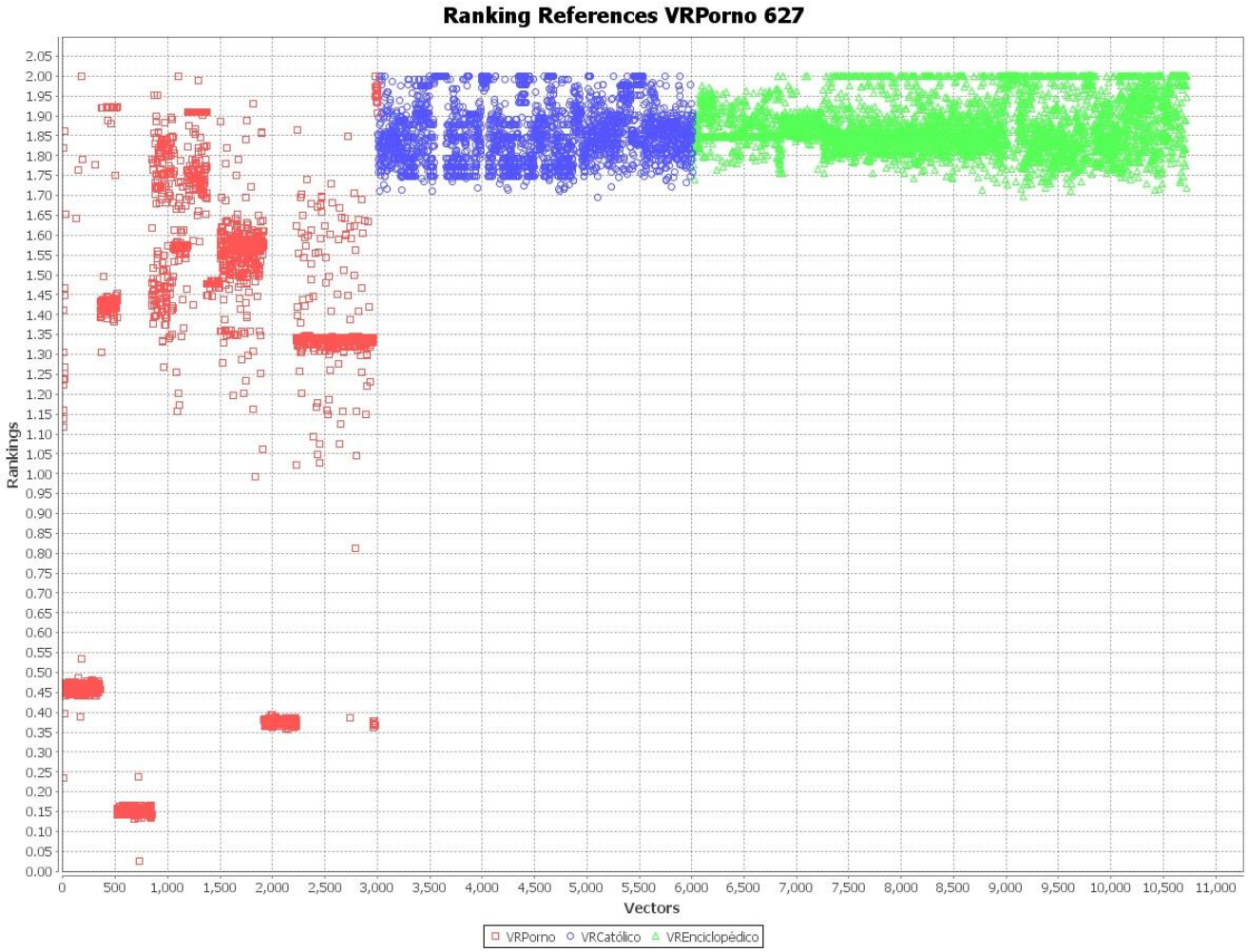
La red ActiveRank estudiada en la presente tesis está conformada por vectores con 50 componentes cada uno; al sumar los diferentes vectores para crear el de referencia, se debe de truncar su longitud a la misma cantidad de nodos que el resto de la red. En la gráfica 4.3.2. podemos observar la fila de la matriz de rankings del vector de referencia pornográfico sin haber sido limitado a los 50 nodos más importantes; el efecto que genera dicha disparidad en el número de componentes entre vectores es la disminución del rango de valores de ranking que se pueden alcanzar, esto debido a que se aumenta la probabilidad de nodos no coincidentes para cualquier par de elementos de la red; lo anterior se puede estudiar como un decremento en la resolución del sistema. Existe un punto óptimo en el número de componentes de cada vector de ActiveRank dependiendo del escenario y la naturaleza del sistema que se quiera analizar.

De manera adicional se obtuvieron las gráficas de las filas de la matriz de rankings correspondientes a 3 vectores aleatoriamente seleccionados pertenecientes al conjunto de documentos pornográficos; comparando las gráficas 4.2.3 a 4.2.5. con lo obtenido en la gráfica 4.2.1. podemos comprobar que el comportamiento del vector de referencia con respecto a cualquier elemento claramente descriptivo de la naturaleza del conjunto cumple los mismos principios, lo que agrega certeza al proceso previamente desarrollado.

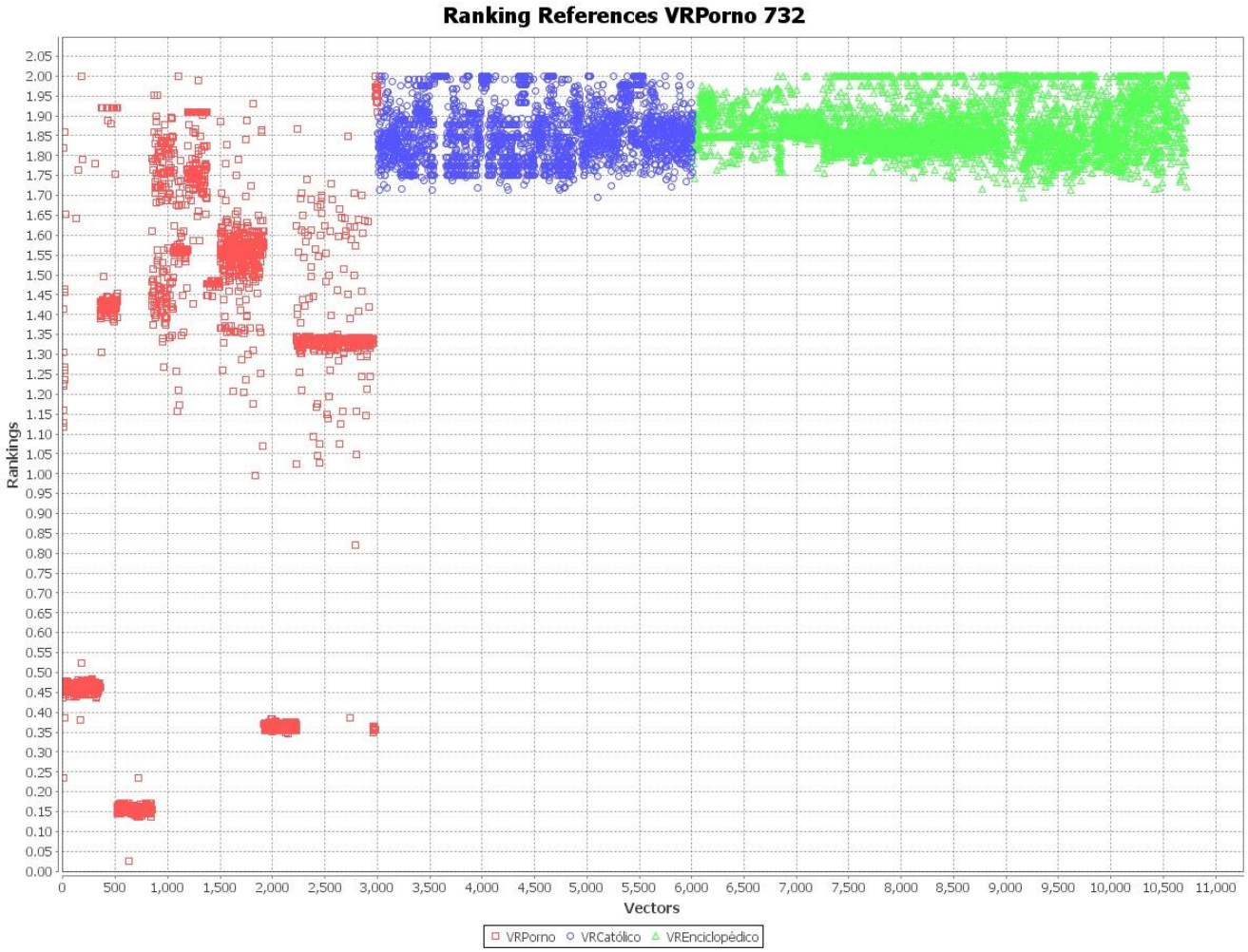
El análisis de la eficiencia del proceso de clasificación automático de información se encuentra más adelante en el capítulo 5 de la presente tesis, y no en la sección de resultados, con la intención de facilitar su comprensión.




Gráfica 4.3.2. - Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector de referencia pornográfico, sin longitud acotada, contra todos los demás.




Gráfica 4.3.3. - Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector 627 del conjunto pornográfico contra todos los demás.

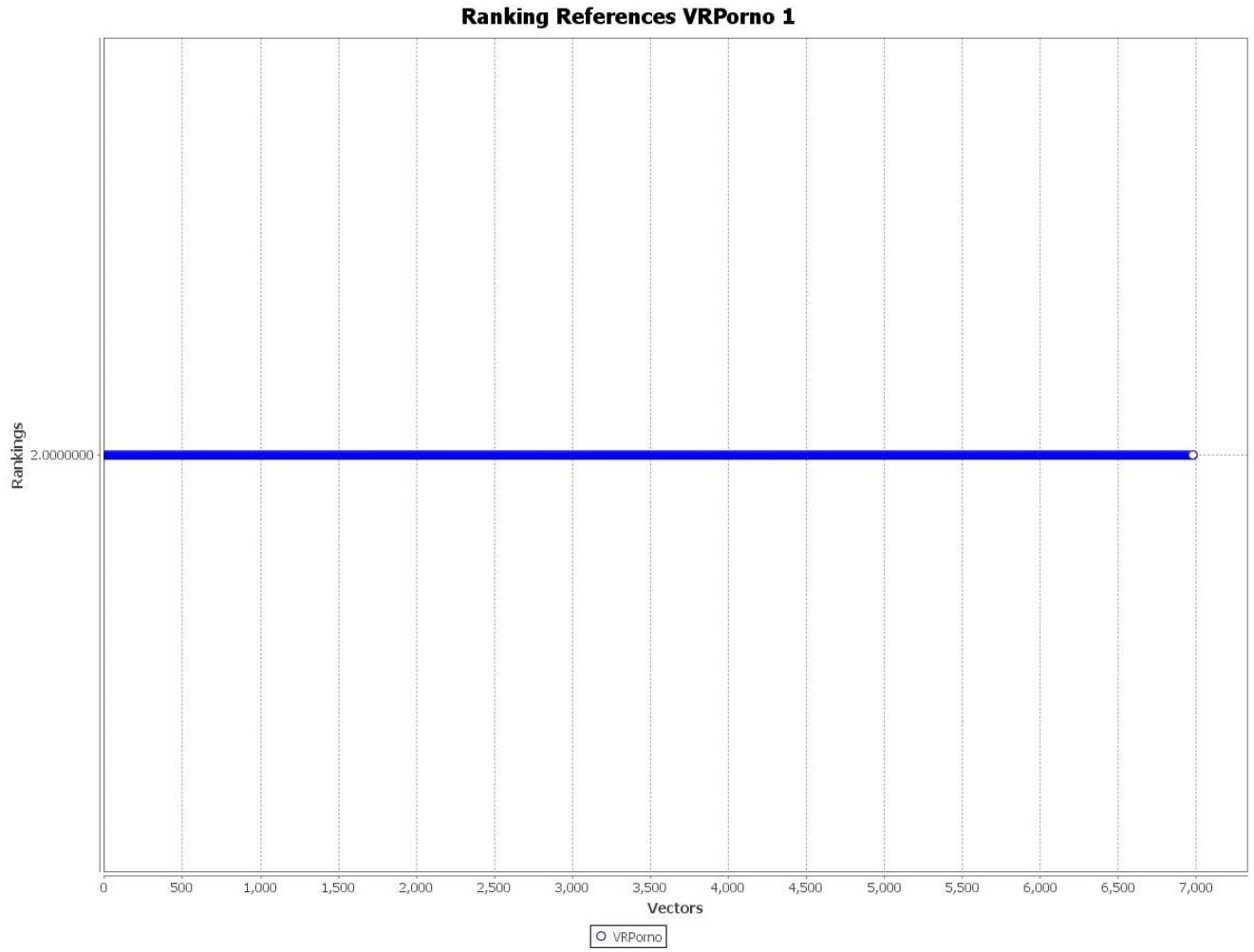



Gráfica 4.3.4. - Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector 732 del conjunto pornográfico contra todos los demás.


Gráfica 4.3.5. - Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector 2005 del conjunto pornográfico contra todos los demás.

**Resultados Etapa 2**

A continuación se pueden observar los resultados del análisis de las páginas escaneadas siguiendo el procedimiento especificado en la sección 4.2. de esta tesis.




Gráfica 4.3.6. - Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector de referencia pornográfico contra todo el conjunto de páginas bajo el dominio unam.mx

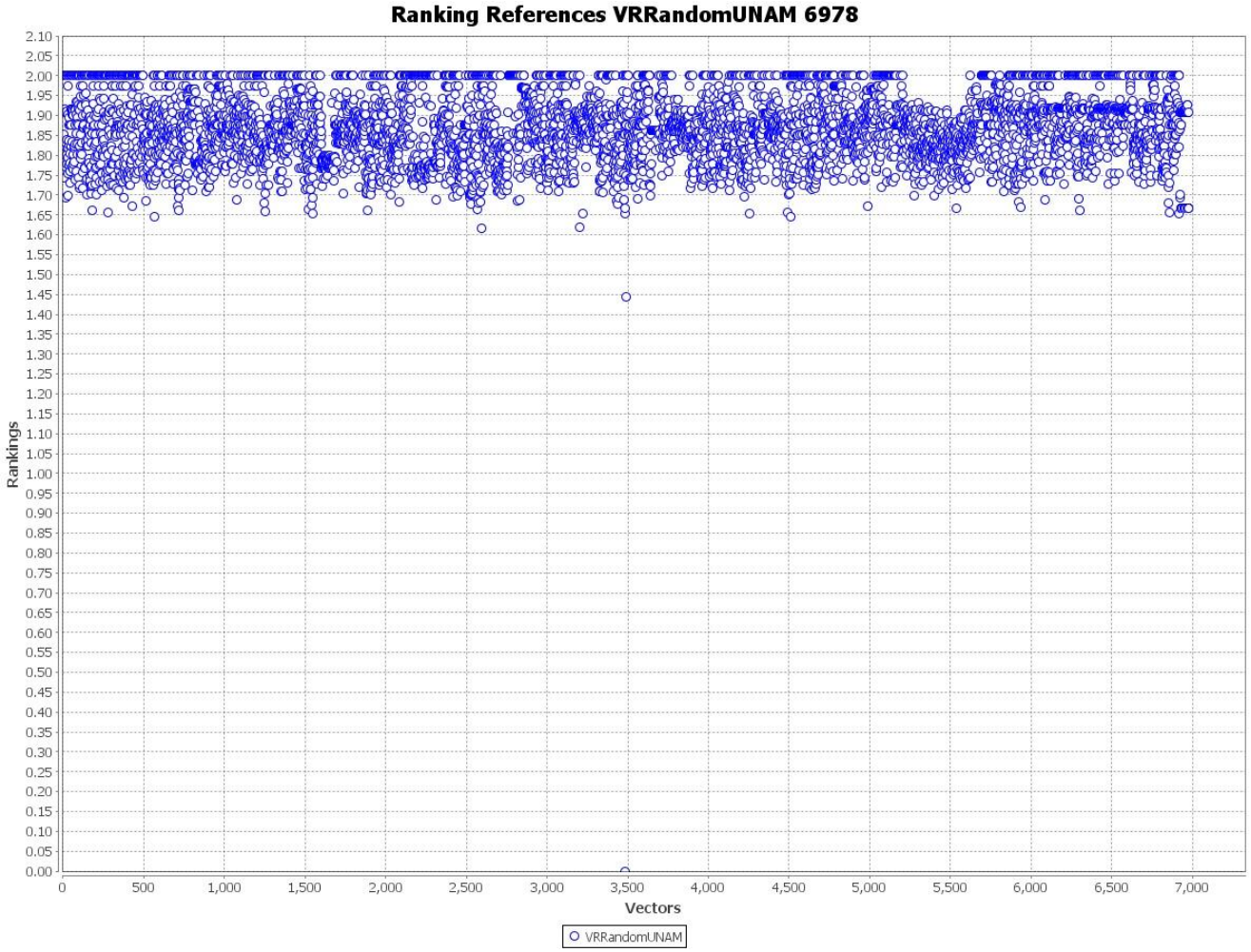
Como podemos observar, el ranking del vector de referencia pornográfico contra cualquiera de los elementos pertenecientes al conjunto de páginas bajo el dominio *unam.mx* es *2*, lo que es consecuencia de que el vector de referencia no comparte ningún nodo con cualquiera de los elementos del grupo; lo anterior es un punto delicado porque se puede deber a 3 situaciones muy específicas:

1. El sistema DARE presenta un error y no está generando la matriz de rankings correspondiente, regresando por *default* un valor de ranking *2* (*nulo*, en la versión extendida del algoritmo ActiveRank).
2. El vector de referencia no es lo suficientemente extenso o no contiene los nodos adecuados para ser considerado válido. La situación más común es una conjunción entre estas dos características; el vector de referencia contiene un número insuficiente de nodos que a su vez no son compatibles u óptimos para el universo que se está analizando ya sea por el idioma, tecnicismo, regionalismo de las expresiones, etc..
3. En realidad todo el universo no contiene ninguna página similar al vector de referencia, en este caso, ninguna página pornográfica.

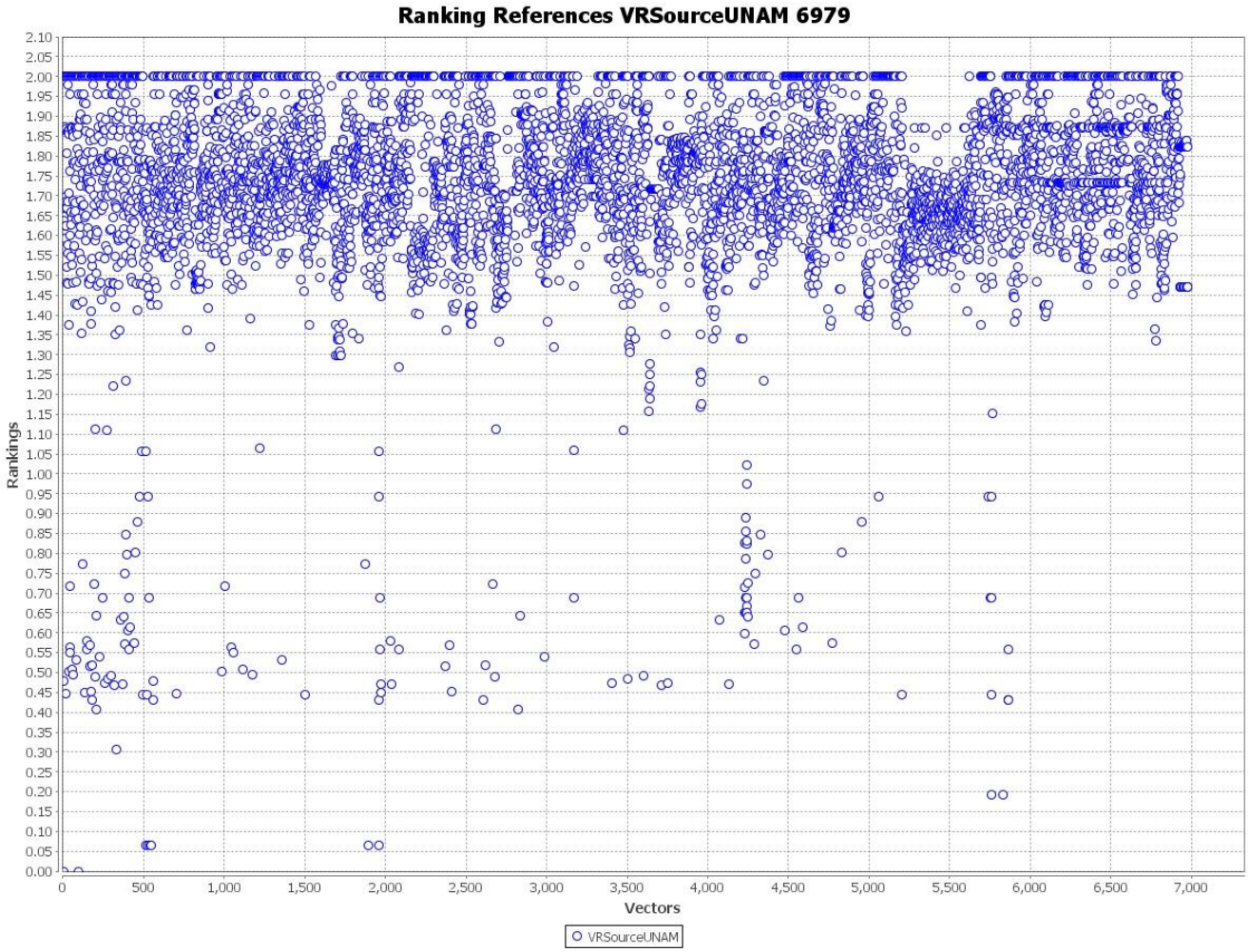
Con la intención de comprobar el buen funcionamiento del sistema en el cálculo de rankings y profundizar un poco más en el comportamiento y características propias del algoritmo ActiveRank, se generaron algunas pruebas adicionales que se muestran a continuación.

**Resultados Etapa 2 Extendida**

La primer prueba adicional que se realizó tuvo como fin comprobar que el sistema DARE hubiera funcionado correctamente durante la generación de la gráfica 4.3.6., lo cual se llevó a cabo generando dos gráficas adicionales correspondientes a los vectores de las páginas *http://www.unam.mx/* y de otra aleatoriamente seleccionada contra el resto del universo, analizando en primer lugar que existieran valores de ranking diferentes de *2* y que aquellos valores de la fila iguales a *2* o muy cercanos se trataran de páginas con el mismo contenido, es decir, diferentes URLs que despliegan la misma información. A continuación, las gráficas 4.3.7. y 4.3.8. despliegan los resultados obtenidos.




Gráfica 4.3.7. - Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector 6978, aleatoriamente seleccionado, contra todo el conjunto de páginas bajo el dominio unam.mx




Gráfica 4.3.8. - Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector de la página http://www.unam.mx contra todo el conjunto de páginas bajo el dominio unam.mx

De las dos gráficas anteriormente mostradas podemos comprobar el correcto funcionamiento del sistema, y que el resultado mostrado en la gráfica 4.3.6. es correcto y no un error en la plataforma de ActiveRank.

El siguiente punto a evaluar fue la validez del vector pornográfico de referencia que estaba siendo utilizado, revisando manualmente los nodos que lo conformaban. Se detectó que todos los nodos pertenecían a palabras en inglés, y que los términos pornográficos en páginas pornográficas en español, así como su propia estructura eran significativamente diferentes, por lo que existía la posibilidad de que páginas con contenido pornográfico en español estuvieran siendo mal clasificadas. Éste problema puede ser resuelto de dos formas; aumentando la extensión del vector de referencia ingresando palabras descriptivas de contenido pornográfico en

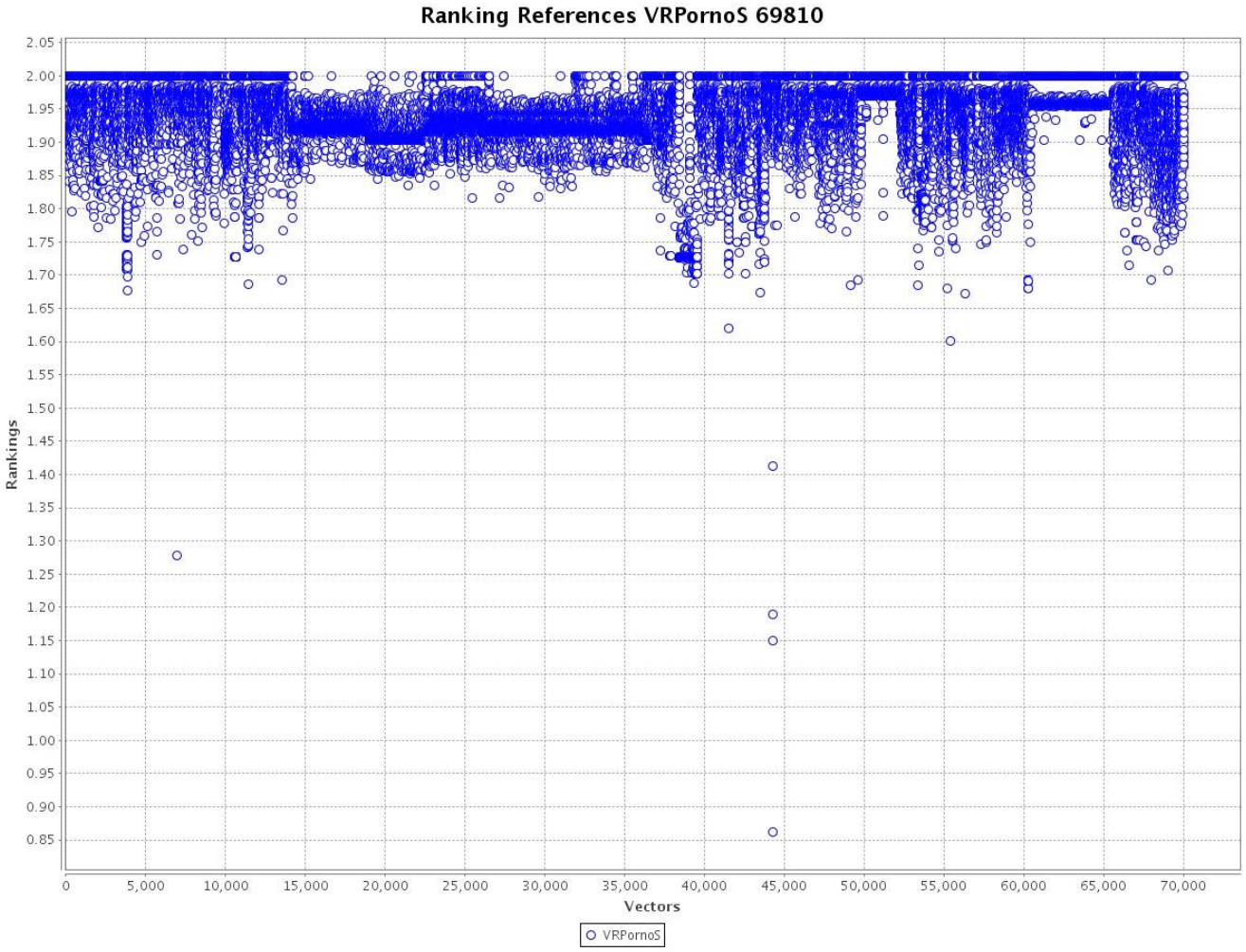
español al vector existente, o ingresar estas mismas palabras como un vector independiente y realizar la clasificación a través de una combinación lineal de los valores de ranking obtenidos para cada documento con respecto a estos dos vectores de referencia; este último formato fue el adoptado para las pruebas subsecuentes ya que incrementa la flexibilidad en el análisis del comportamiento del sistema así como en escenarios donde la segmentación es crítica.

El vector pornográfico de referencia utilizando fuentes en español se generó repitiendo el procedimiento especificado en la sección 4.2. de esta tesis utilizando como fuentes iniciales las siguientes páginas pornográficas:

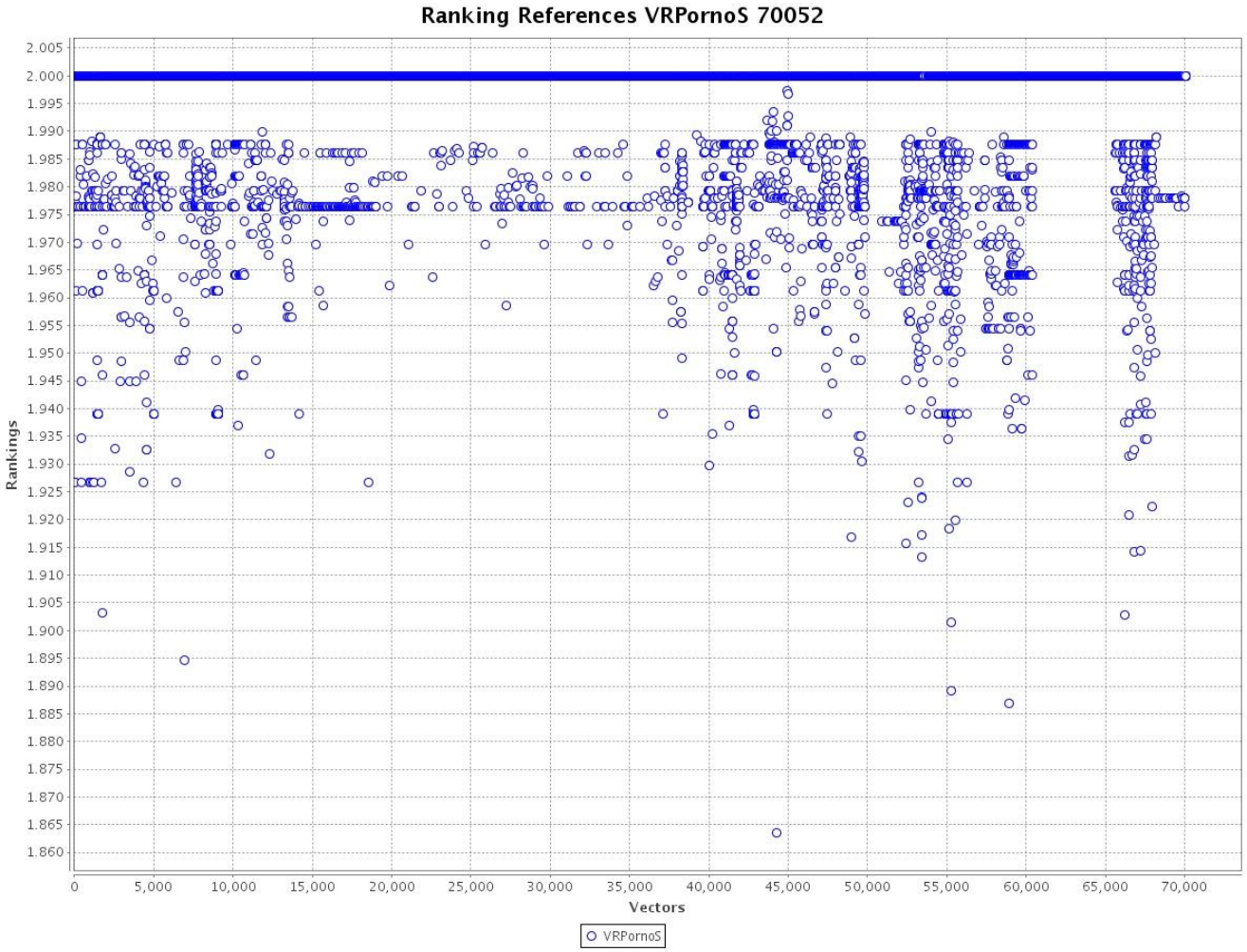
a) *http://www.pornomexico.net/*
b) *http://www.macizorras.com/*
c) *http://www.viendosexo.com/*
d) *http://www.iberporno.com/*

De igual forma, con la intención de mejorar el análisis desarrollado, se incrementó el número de páginas analizadas de 5000 a 75000, eliminando la restricción de que pertenecieran al dominio *unam.mx* y limitando únicamente el nivel de escaneo para abarcar hasta vínculos de segundo nivel partiendo de cualquier página de la universidad. Este aumento significativo en el número de documentos a analizar arrojó información adicional sobre el desempeño en redes de mediana escala de todos los sistemas utilizados, misma que fue utilizada para realizar mejoras en el diseño e implementación de los mismos.

Las gráficas 4.3.9. y 4.3.10. representan las filas de la matriz de rankings correspondientes a los vectores pornográficos de referencia en español e inglés respectivamente contra todos los demás; el análisis de estas nuevas gráficas es desarrollado en el capítulo 5 de ésta tesis.




Gráfica 4.3.9. - Gráfica de dispersión de la fila de la matriz de rankings correspondiente al vector pornográfico de referencia en español contra todo el conjunto de páginas analizadas.




Gráfica 4.3.10. - Gráfica de dispersión de la fila de la matriz de rankings correspondiente al vector pornográfico de referencia en inglés contra todo el conjunto de páginas analizadas.

# Capítulo Quinto

## Análisis de Resultados

# 5. Análisis de resultados

Retomando parte del proceso experimental descrito en la sección 4.2. de la presente tesis, a continuación se presenta el análisis y resultados obtenidos sobre el desempeño y eficiencia del algoritmo ActiveRank como método para la clasificación automática de información.

## 5.1. Acerca de la interpretación visual de la matriz de rankings ActiveRank

Debido al enorme tamaño de la matriz de rankings generada por ActiveRank, su manipulación e interpretación en forma tabular puede ser extremadamente compleja en aplicaciones de una escala relativamente pequeña, una alternativa útil en el proceso de revisión y evaluación es representar la matriz en una gráfica tridimensional, donde los ejes *x* y *y* representan la combinación de los elementos *i* y *j*, y el eje *z* representa el valor de ranking *R(i,j)*. Cuando se trata de analizar el comportamiento del ranking de un elemento contra todos los demás se puede graficar bidimensionalmente la fila de interés de la matriz, y de esta forma analizar gráficamente los resultados obtenidos, método ampliamente utilizado en la presente tesis.

## 5.2. Proceso de clasificación manual de la muestra

La siguiente tabla despliega la evaluación manual del tipo de página para cada uno de los vectores pertenecientes a la muestra del 2% como se especificó en la sección 4.2. de esta tesis. Se muestra el identificador del vector, seguido por el ranking de dicho elemento contra el vector pornográfico de referencia, y por último VERDADERO si la página es pornográfica, y FALSO si el documento no es pornográfico. La clasificación se realizó abriendo el URL de cada uno de los documentos especificados en un navegador web y observando su contenido.

Tabla 5.2.1. – Clasificación manual de la muestra generada aleatoriamente mostrando el identificador del vector, su valor de ranking con respecto al vector pornográfico de referencia y la condición booleana para la premisa de pertenencia al conjunto de interés.

| id_vector | ranking | real type |
|---|---|---|
| **108** | 0.716147 | VERDADERO |
| **119** | 0.702886 | VERDADERO |
| **151** | 0.701142 | VERDADERO |
| **210** | 0.715083 | VERDADERO |
| **246** | 0.715083 | VERDADERO |
| **257** | 0.716821 | VERDADERO |
| **409** | 1.3208 | VERDADERO |
| **407** | 1.34164 | VERDADERO |
| **427** | 1.33924 | VERDADERO |
| **502** | 1.34816 | VERDADERO |
| **499** | 1.32817 | VERDADERO |
| **556** | 0.629866 | VERDADERO |
| **538** | 0.637565 | VERDADERO |
| **738** | 0.627553 | VERDADERO |
| **724** | 0.626547 | VERDADERO |
| **811** | 0.6284 | VERDADERO |
| **775** | 0.62727 | VERDADERO |
| **849** | 1.49506 | VERDADERO |
| **999** | 1.88974 | VERDADERO |
| **987** | 1.85913 | VERDADERO |
| **1091** | 1.4369 | VERDADERO |
| **1126** | 1.4369 | VERDADERO |
| **1125** | 1.79746 | VERDADERO |
| **1358** | 1.68373 | VERDADERO |
| **1392** | 1.4142 | VERDADERO |
| **1386** | 1.4142 | VERDADERO |
| **1333** | 1.87804 | FALSO |
| **1533** | 1.48021 | VERDADERO |
| **1641** | 1.47048 | VERDADERO |
| **1612** | 1.3867 | VERDADERO |
| **1623** | 1.48021 | VERDADERO |
| **1565** | 1.50475 | VERDADERO |
| **1869** | 1.50129 | VERDADERO |
| **1833** | 1.4481 | VERDADERO |

| id_vector | ranking | real type |
|---|---|---|
| **3369** | 1.90752 | FALSO |
| **3826** | 1.82897 | FALSO |
| **3825** | 2 | FALSO |
| **3838** | 1.88287 | FALSO |
| **3811** | 1.92608 | FALSO |
| **3683** | 1.93589 | FALSO |
| **4081** | 1.87545 | FALSO |
| **4046** | 1.98123 | FALSO |
| **3973** | 1.86834 | FALSO |
| **3996** | 2 | FALSO |
| **3934** | 1.86464 | FALSO |
| **4117** | 1.88798 | FALSO |
| **4126** | 1.90143 | FALSO |
| **4213** | 1.86834 | FALSO |
| **4165** | 1.85013 | FALSO |
| **4264** | 1.88273 | FALSO |
| **4261** | 1.87517 | FALSO |
| **4273** | 1.8582 | FALSO |
| **4915** | 1.89964 | FALSO |
| **4896** | 1.8607 | FALSO |
| **5027** | 1.92712 | FALSO |
| **5113** | 1.89659 | FALSO |
| **4726** | 1.93094 | FALSO |
| **4774** | 1.93589 | FALSO |
| **4802** | 1.88287 | FALSO |
| **4847** | 1.85729 | FALSO |
| **5595** | 1.92103 | FALSO |
| **5542** | 1.86182 | FALSO |
| **5220** | 1.93082 | FALSO |
| **5235** | 1.86099 | FALSO |
| **5232** | 1.91361 | FALSO |
| **5344** | 1.90752 | FALSO |
| **5373** | 1.94568 | FALSO |
| **5270** | 1.93589 | FALSO |

| id_vector | ranking | real type |
|---|---|---|
| **8150** | 1.87968 | FALSO |
| **8085** | 1.91733 | FALSO |
| **8116** | 1.8298 | FALSO |
| **7977** | 1.9046 | FALSO |
| **7984** | 1.87789 | FALSO |
| **7985** | 1.88842 | FALSO |
| **7894** | 1.91733 | FALSO |
| **7803** | 2 | FALSO |
| **8807** | 1.93589 | FALSO |
| **8791** | 2 | FALSO |
| **8765** | 1.84861 | FALSO |
| **8720** | 1.93589 | FALSO |
| **8895** | 1.91733 | FALSO |
| **9065** | 1.98961 | FALSO |
| **9010** | 1.94178 | FALSO |
| **9206** | 1.98123 | FALSO |
| **9180** | 1.94336 | FALSO |
| **9148** | 1.98931 | FALSO |
| **8274** | 1.91064 | FALSO |
| **8306** | 1.88734 | FALSO |
| **8208** | 1.9323 | FALSO |
| **8372** | 1.90354 | FALSO |
| **8551** | 1.93597 | FALSO |
| **8468** | 1.89977 | FALSO |
| **8496** | 1.89828 | FALSO |
| **8697** | 1.9046 | FALSO |
| **8687** | 2 | FALSO |
| **8585** | 1.9046 | FALSO |
| **9806** | 2 | FALSO |
| **9863** | 1.90342 | FALSO |
| **9868** | 1.87689 | FALSO |
| **9943** | 1.93302 | FALSO |
| **10003** | 1.91733 | FALSO |
| **10144** | 1.94517 | FALSO |

| | | |
|---|---|---|
| **1793** | 1.48021 | VERDADERO |
| **2021** | 0.630998 | VERDADERO |
| **1970** | 0.637044 | VERDADERO |
| **1966** | 0.638366 | VERDADERO |
| **2208** | 0.638558 | VERDADERO |
| **2300** | 1.15942 | VERDADERO |
| **2296** | 1.15183 | VERDADERO |
| **2111** | 0.626995 | VERDADERO |
| **2085** | 0.630548 | VERDADERO |
| **2174** | 0.636286 | VERDADERO |
| **2477** | 1.15031 | VERDADERO |
| **2505** | 1.55085 | VERDADERO |
| **2365** | 1.16183 | VERDADERO |
| **2692** | 1.15393 | VERDADERO |
| **2669** | 1.16103 | VERDADERO |
| **2653** | 0.950767 | VERDADERO |
| **3033** | 1.89384 | FALSO |
| **2875** | 1.1562 | VERDADERO |
| **2929** | 1.1628 | VERDADERO |
| **2939** | 1.16344 | VERDADERO |
| **2899** | 1.46988 | VERDADERO |
| **3319** | 1.88287 | FALSO |
| **3220** | 1.93303 | FALSO |
| **3094** | 1.82897 | FALSO |
| **3126** | 1.92103 | FALSO |
| **3107** | 1.88287 | FALSO |
| **3582** | 1.97198 | FALSO |
| **3504** | 1.83989 | FALSO |
| **3410** | 1.90732 | FALSO |
| **3337** | 1.94917 | FALSO |
| **3347** | 1.88157 | FALSO |

| | | |
|---|---|---|
| **5956** | 1.90864 | FALSO |
| **5984** | 1.82101 | FALSO |
| **5932** | 1.87654 | FALSO |
| **6120** | 1.92021 | FALSO |
| **6020** | 1.91733 | FALSO |
| **6026** | 1.91239 | FALSO |
| **5716** | 1.89266 | FALSO |
| **5749** | 1.91629 | FALSO |
| **5859** | 1.93082 | FALSO |
| **5762** | 1.94568 | FALSO |
| **6564** | 1.91066 | FALSO |
| **6634** | 1.91066 | FALSO |
| **6647** | 1.89322 | FALSO |
| **6603** | 1.91066 | FALSO |
| **6445** | 1.94262 | FALSO |
| **6526** | 1.91066 | FALSO |
| **6332** | 1.91066 | FALSO |
| **6314** | 1.91066 | FALSO |
| **6346** | 1.89547 | FALSO |
| **6165** | 1.93313 | FALSO |
| **7059** | 1.91024 | FALSO |
| **7125** | 1.92712 | FALSO |
| **7147** | 1.92712 | FALSO |
| **6914** | 1.92712 | FALSO |
| **6930** | 1.94178 | FALSO |
| **6746** | 1.91066 | FALSO |
| **7617** | 1.9196 | FALSO |
| **7391** | 1.88672 | FALSO |
| **7273** | 1.91881 | FALSO |
| **7249** | 1.92712 | FALSO |
| **7212** | 1.92712 | FALSO |

| | | |
|---|---|---|
| **10231** | 2 | FALSO |
| **10197** | 1.94056 | FALSO |
| **10181** | 2 | FALSO |
| **9360** | 2 | FALSO |
| **9435** | 2 | FALSO |
| **9441** | 1.88524 | FALSO |
| **9462** | 2 | FALSO |
| **9459** | 1.91733 | FALSO |
| **9470** | 1.88287 | FALSO |
| **9488** | 1.93209 | FALSO |
| **9557** | 1.89669 | FALSO |
| **9625** | 1.86434 | FALSO |
| **9628** | 2 | FALSO |
| **9609** | 1.9046 | FALSO |
| **9613** | 1.84604 | FALSO |
| **9652** | 1.90247 | FALSO |
| **9670** | 1.89019 | FALSO |
| **9722** | 2 | FALSO |
| **9707** | 1.89948 | FALSO |
| **10445** | 1.83683 | FALSO |
| **10481** | 2 | FALSO |
| **10467** | 1.96184 | FALSO |
| **10426** | 1.97059 | FALSO |
| **10311** | 1.92922 | FALSO |
| **10359** | 1.89234 | FALSO |
| **10263** | 1.89773 | FALSO |
| **10297** | 1.94508 | FALSO |
| **10700** | 1.86661 | FALSO |
| **10499** | 1.89885 | FALSO |
| **10531** | 1.92296 | FALSO |

### 5.3. Medición de la eficiencia del proceso de clasificación

El proceso de clasificación automática, como se ha mencionado a lo largo de la presente tesis, consiste en discriminar a través del valor de ranking si un documento o información pertenece o no a un conjunto determinado; el método para lograr determinar el umbral óptimo de clasificación consiste en un

proceso de maximización de la eficiencia de clasificación en una muestra representativa del universo. Siguiendo el desarrollo experimental planteado en la sección 4.2. de la presente tesis, el valor inicial para el umbral de clasificación corresponde al valor medio entre $\bar{r}_{porno}$ y $\bar{r}_{católico}$ expresados en la tabla 4.3.1., y que corresponde a $r_{umbral\ inicial} = 1.53224907$; la tabla 5.3.1. describe para cada uno de los elementos de la muestra su identificador, el valor de ranking de dicho elemento contra el vector de referencia pornográfico, su pertenencia o no pertenencia al conjunto determinado manualmente como valor de control, su pertenencia o no pertenencia según el algoritmo de clasificación automática con respecto al valor $r_{umbral\ inicial}$, y la matriz binaria de calificación de *Verdadero-Verdadero*, *Falso-Verdadero*, *Verdadero-Falso* y *Falso-Falso* para cada una de las hipótesis de clasificación automática, la cual se construye a partir del siguiente algoritmo presentado en pseudocódigo:

SI *el documento es autoclasificado como pornográfico* Y *el documento es pornográfico* ENTONCES: *Verdadero-Verdadero*

SI *el documento es autoclasificado como pornográfico* Y *el documento no es pornográfico* ENTONCES: *Falso-Verdadero*

SI *el documento es autoclasificado como no pornográfico* Y *el documento no es pornográfico* ENTONCES: *Verdadero-Falso*

SI *el documento es autoclasificado como no pornográfico* Y *el documento es pornográfico* ENTONCES: *Falso-Falso*

NOTA: Si el algoritmo anterior se desea programar con estructuras *IF* anidadas, la condición inicial deberá ser respecto a la evaluación manual, y en término secundario la comparación de autoclasificación, de lo contrario se excluirá una rama completa del árbol (observar con condiciones de frontera).

SI *el documento es pornográfico* ENTONCES:
    SI *el documento es autoclasificado como pornográfico* ENTONCES: *Verdadero-Verdadero*
    SI NO: *Falso-Verdadero*
SI NO:
    SI *el documento es autoclasificado como pornográfico* ENTONCES: *Falso-Falso*
    SI NO: *Verdadero-Falso*

Tabla 5.3.1. – Tabla que despliega los resultados del análisis y proceso de autoclasificación de la etapa 1 del proceso experimental.

| | | $r_{umbral\ inicial}$ | 1.53224907 | CLASIFIED AS PORN | | CLASIFIED AS NOT PORN | |
|---|---|---|---|---|---|---|---|
| **id_vector** | **ranking** | **real type** | **auto clasification** | **TrueTrue** | **FalseTrue** | **TrueFalse** | **FalseFalse** |
| **108** | 0.716147 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **119** | 0.702886 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **151** | 0.701142 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **210** | 0.715083 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **246** | 0.715083 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **257** | 0.716821 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **409** | 1.3208 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **407** | 1.34164 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **427** | 1.33924 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **502** | 1.34816 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **499** | 1.32817 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **556** | 0.629866 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **538** | 0.637565 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **738** | 0.627553 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **724** | 0.626547 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **811** | 0.6284 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **775** | 0.62727 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **849** | 1.49506 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **999** | 1.88974 | VERDADERO | FALSO | 0 | 0 | 0 | 1 |
| **987** | 1.85913 | VERDADERO | FALSO | 0 | 0 | 0 | 1 |
| **1091** | 1.4369 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **1126** | 1.4369 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **1125** | 1.79746 | VERDADERO | FALSO | 0 | 0 | 0 | 1 |
| **1358** | 1.68373 | VERDADERO | FALSO | 0 | 0 | 0 | 1 |
| **1392** | 1.4142 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **1386** | 1.4142 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **1333** | 1.87804 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **1533** | 1.48021 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **1641** | 1.47048 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **1612** | 1.3867 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **1623** | 1.48021 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **1565** | 1.50475 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **1869** | 1.50129 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **1833** | 1.4481 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **1793** | 1.48021 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2021** | 0.630998 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **1970** | 0.637044 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **1966** | 0.638366 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2208** | 0.638558 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2300** | 1.15942 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2296** | 1.15183 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2111** | 0.626995 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2085** | 0.630548 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2174** | 0.636286 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2477** | 1.15031 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2505** | 1.55085 | VERDADERO | FALSO | 0 | 0 | 0 | 1 |
| **2365** | 1.16183 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2692** | 1.15393 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2669** | 1.16103 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2653** | 0.950767 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **3033** | 1.89384 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **2875** | 1.1562 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2929** | 1.1628 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2939** | 1.16344 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **2899** | 1.46988 | VERDADERO | VERDADERO | 1 | 0 | 0 | 0 |
| **3319** | 1.88287 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3220** | 1.93303 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3094** | 1.82897 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3126** | 1.92103 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3107** | 1.88287 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3582** | 1.97198 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3504** | 1.83989 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3410** | 1.90732 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3337** | 1.94917 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3347** | 1.88157 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3369** | 1.90752 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3826** | 1.82897 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3825** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3838** | 1.88287 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3811** | 1.92608 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3683** | 1.93589 | FALSO | FALSO | 0 | 0 | 1 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **4081** | 1.87545 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4046** | 1.98123 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3973** | 1.86834 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3996** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **3934** | 1.86464 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4117** | 1.88798 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4126** | 1.90143 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4213** | 1.86834 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4165** | 1.85013 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4264** | 1.88273 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4261** | 1.87517 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4273** | 1.8582 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4915** | 1.89964 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4896** | 1.8607 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5027** | 1.92712 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5113** | 1.89659 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4726** | 1.93094 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4774** | 1.93589 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4802** | 1.88287 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **4847** | 1.85729 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5595** | 1.92103 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5542** | 1.86182 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5220** | 1.93082 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5235** | 1.86099 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5232** | 1.91361 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5344** | 1.90752 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5373** | 1.94568 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5270** | 1.93589 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5956** | 1.90864 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5984** | 1.82101 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5932** | 1.87654 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6120** | 1.92021 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6020** | 1.91733 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6026** | 1.91239 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5716** | 1.89266 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5749** | 1.91629 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5859** | 1.93082 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **5762** | 1.94568 | FALSO | FALSO | 0 | 0 | 1 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **6564** | 1.91066 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6634** | 1.91066 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6647** | 1.89322 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6603** | 1.91066 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6445** | 1.94262 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6526** | 1.91066 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6332** | 1.91066 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6314** | 1.91066 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6346** | 1.89547 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6165** | 1.93313 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7059** | 1.91024 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7125** | 1.92712 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7147** | 1.92712 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6914** | 1.92712 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6930** | 1.94178 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **6746** | 1.91066 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7617** | 1.9196 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7391** | 1.88672 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7273** | 1.91881 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7249** | 1.92712 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7212** | 1.92712 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8150** | 1.87968 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8085** | 1.91733 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8116** | 1.8298 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7977** | 1.9046 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7984** | 1.87789 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7985** | 1.88842 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7894** | 1.91733 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **7803** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8807** | 1.93589 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8791** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8765** | 1.84861 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8720** | 1.93589 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8895** | 1.91733 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9065** | 1.98961 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9010** | 1.94178 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9206** | 1.98123 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9180** | 1.94336 | FALSO | FALSO | 0 | 0 | 1 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **9148** | 1.98931 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8274** | 1.91064 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8306** | 1.88734 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8208** | 1.9323 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8372** | 1.90354 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8551** | 1.93597 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8468** | 1.89977 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8496** | 1.89828 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8697** | 1.9046 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8687** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **8585** | 1.9046 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9806** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9863** | 1.90342 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9868** | 1.87689 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9943** | 1.93302 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10003** | 1.91733 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10144** | 1.94517 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10231** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10197** | 1.94056 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10181** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9360** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9435** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9441** | 1.88524 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9462** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9459** | 1.91733 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9470** | 1.88287 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9488** | 1.93209 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9557** | 1.89669 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9625** | 1.86434 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9628** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9609** | 1.9046 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9613** | 1.84604 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9652** | 1.90247 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9670** | 1.89019 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9722** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **9707** | 1.89948 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10445** | 1.83683 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10481** | 2 | FALSO | FALSO | 0 | 0 | 1 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **10467** | 1.96184 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10426** | 1.97059 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10311** | 1.92922 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10359** | 1.89234 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10263** | 1.89773 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10297** | 1.94508 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10700** | 1.86661 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10499** | 1.89885 | FALSO | FALSO | 0 | 0 | 1 | 0 |
| **10531** | 1.92296 | FALSO | FALSO | 0 | 0 | 1 | 0 |

| | | | | |
|---|---|---|---|---|
| **Suma Columnas** | **48** | **0** | **141** | **5** |

Una vez obtenida la matriz de aciertos en el proceso de autoclasificación se procede a calcular los coeficientes de eficiencia del proceso, a partir del número total de repeticiones de cada caso para un valor de umbral determinado; dichos datos se presentan en la última fila de la tabla 5.3.1..

Los índices de eficiencia se calculan como sigue:

$$E_{\in} = \frac{\sum VerdaderoVerdadero}{\sum VerdaderoVerdadero + \sum FalsoVerdadero}$$

$$E_{\notin} = \frac{\sum VerdaderoFalso}{\sum VerdaderoFalso + \sum FalsoFalso}$$

Donde $E_{\in}$ representa la eficiencia del método al clasificar elementos como pertenecientes al conjunto deseado, y $E_{\notin}$ la eficiencia del método al clasificar elementos como no pertenecientes al conjunto deseado, ambos utilizando un valor de umbral determinado.

Con $r_{umbral\ inicial} = 1.53224907$ , los dos valores de eficiencia obtenidos fueron:

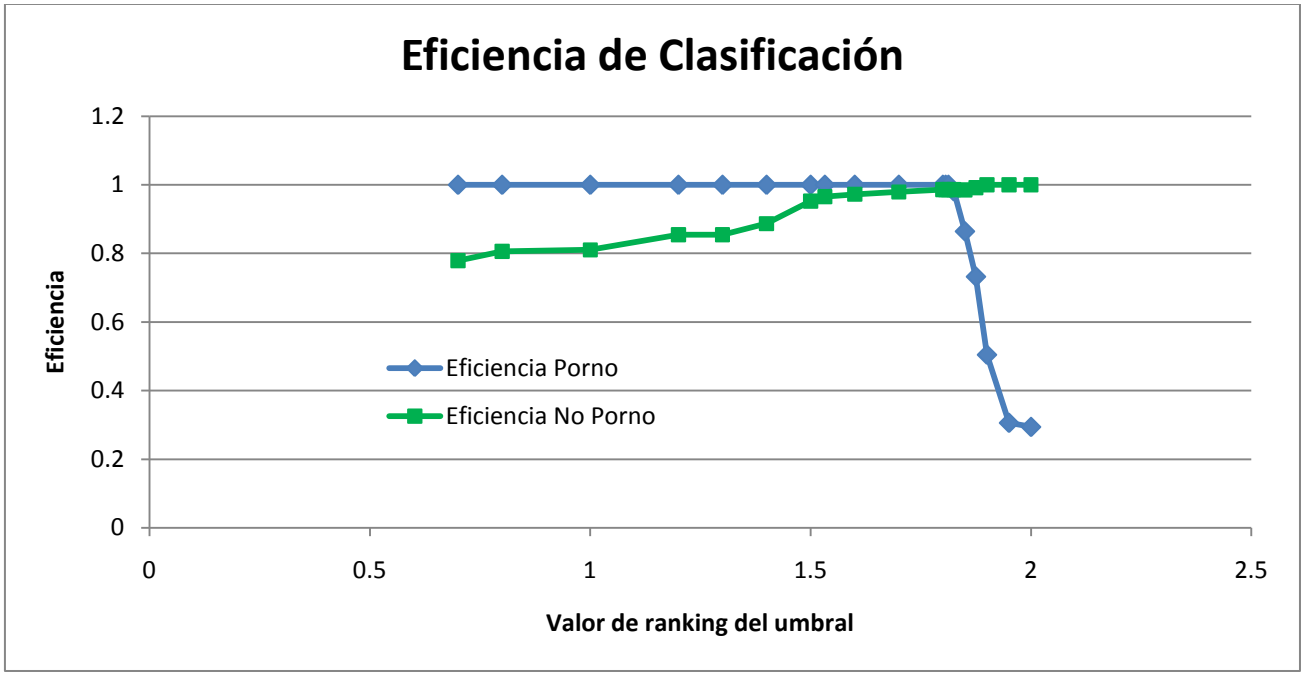
$$E_{\in} = 100\% \quad y \quad E_{\notin} = 96.57\%$$

Para realizar un análisis más detallado del comportamiento de la eficiencia en el proceso de autoclasificación, y que además permite plantear las bases de un sistema semiautomático para la obtención del valor de ranking umbral óptimo de mejor manera, se repitió el procedimiento planteado anteriormente para diferentes valores de umbral, inicialmente divididos en intervalos de 0.1, y aproximando por mitades el punto óptimo (se encuentra el intervalo que presenta máxima eficiencia, se divide en dos mitades iguales, se calcula de

nuevo la eficiencia en el subconjunto que presenta el mejor desempeño y se repite hasta obtener convergencia o un resultado consistente); este algoritmo puede ser integrado en un sistema de cómputo que facilite la clasificación manual de la muestra, y posteriormente, de forma automática, encontrara el nivel óptimo del umbral de clasificación; a continuación, la tabla 5.3.2. muestra los resultados obtenidos tras el desarrollo del análisis.

Tabla 5.3.2. – Tabla que muestra el comportamiento de los índices de eficiencia para diferentes valores de umbral.

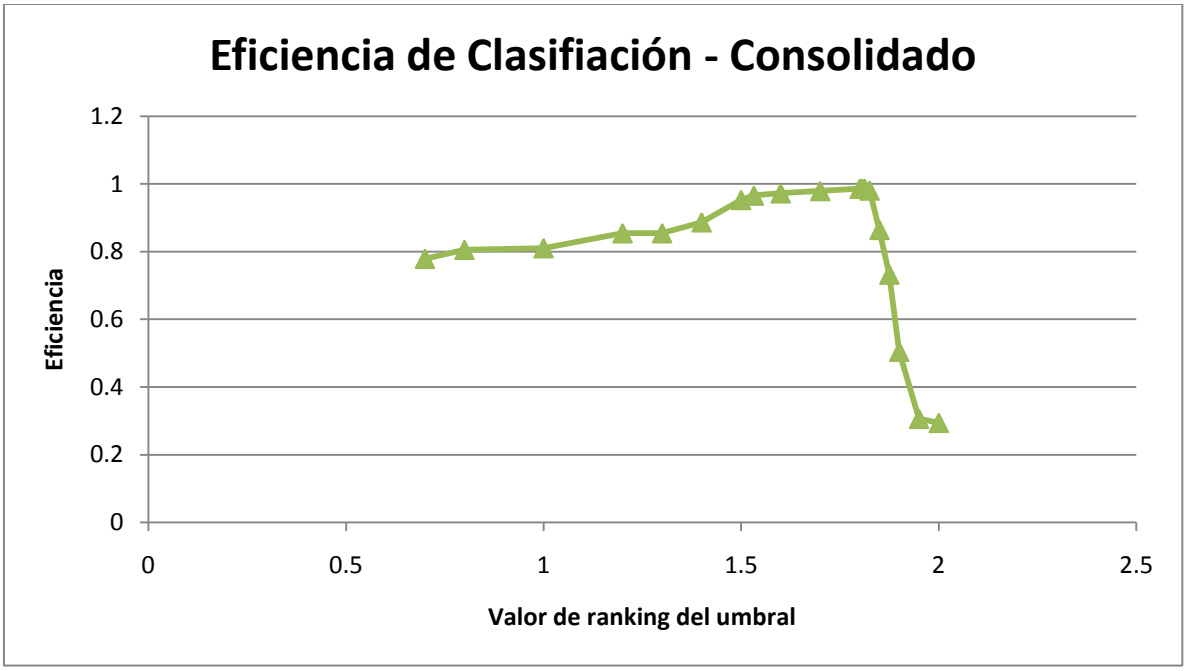
| $r_{umbral}$ | $E_{\in}$ | $E_{\notin}$ |
|---|---|---|
| **0.7** | 1 | 0.779005525 |
| **0.8** | 1 | 0.805714286 |
| **1** | 1 | 0.810344828 |
| **1.2** | 1 | 0.854545455 |
| **1.3** | 1 | 0.854545455 |
| **1.4** | 1 | 0.886792453 |
| **1.5** | 1 | 0.952702703 |
| **1.53224907** | 1 | 0.965753425 |
| **1.6** | 1 | 0.972413793 |
| **1.7** | 1 | 0.979166667 |
| **1.8** | 1 | 0.986013986 |
| **1.80625** | 1 | 0.986013986 |
| **1.8125** | **1** | **0.986013986** |
| **1.825** | 0.980769231 | 0.985915493 |
| **1.85** | 0.86440678 | 0.985185185 |
| **1.875** | 0.732394366 | 0.991869919 |
| **1.9** | 0.504761905 | 1 |
| **1.95** | 0.306358382 | 1 |
| **2** | 0.294444444 | 1 |




Gráfica 5.3.1. – Comportamiento de los índices de eficiencia de clasificación con respecto a la variación en el valor de ranking de umbral.

En la tabla 5.3.2. y la gráfica 5.3.1. se puede apreciar un punto de inflexión entre los coeficientes de eficiencia en la clasificación del contenido como perteneciente y no perteneciente en $r_{umbral} = 1.8125$, donde a su vez ambos índices presentan sus valores máximos. Los altos valores en la eficiencia cuando el umbral tiene su valor óptimo se verán afectados de manera directa por características del conjunto de información como pueden ser la diversidad de contenidos y una segmentación menos precisa, lenguajes y estructuras gramaticales regionales, entre otras. Una prueba interesante que queda fuera de los alcances de esta tesis sería incrementar periódicamente el número de conjuntos polares en el conjunto de información e ir evaluando y comparando su desempeño en la clasificación de cada uno de ellos extrapolando los protocolos planteados en el presente documento.

Si se consolida cada pareja de coeficientes de eficiencia por valor de umbral considerando la magnitud más pequeña como la final podemos observar la curva descrita en la gráfica 5.3.2. y representar más claramente la curva de máxima eficiencia en el proceso de clasificación.




Gráfica 5.3.2. – Curva consolidada de eficiencia de clasificación con respecto a la variación en el valor de ranking de umbral.

### 5.4. Elementos que afectan la eficiencia en el proceso de clasificación

Como pudimos observar en las secciones 5.2. y 5.3. de la presente tesis, el algoritmo ActiveRank es un método viable y de alta eficiencia para la clasificación de contenido en ambientes con conjuntos de información antagónicos, sin embargo, diferentes características pueden afectar de manera significativa su desempeño; a continuación se describen algunas de las más importantes detectadas tras el desarrollo de esta investigación.

1. *Extensión del universo y técnicas de acotamiento de procesos de exploración de la WWW*

Los sistemas de crawling utilizados en la presente tesis pueden ser acotados para discriminar entre páginas pertenecientes a dominios específicos así como limitarse a seguir vínculos hasta cierto nivel de profundidad; esto último, si no está correctamente planteado tiene un impacto directo sobre los procesos subsecuentes, incluyendo la utilización de ActiveRank como plataforma de clasificación automática; un ejemplo sería intentar discriminar páginas cuyo contenido estuviese vinculado estrechamente a una fuente y al mismo tiempo excluir esa fuente en el conjunto de documentos analizados; a pesar de que con ActiveRank se pudiera

clasificar, nunca encontraríamos dicho contenido no por el algoritmo sino por su exclusión en las etapas iniciales de la conformación del universo de información.

2. *Aumento en el número y diversidad de subconjuntos de información*

Relacionado de manera directa al punto anterior, el expandir la cantidad de información recolectada de la WWW trae como consecuencia inmediata la diversificación y el aumento de cúmulos o *clusters* de información en el universo generado; esto dificulta el proceso autoclasificación basado en ActiveRank dado que reduce el intervalo entre el valor de ranking umbral y los valores de ranking medios de los diferentes subconjuntos. Se puede entender como que si se quisieran segmentar dos subconjuntos cuyos valores de ranking medios son muy similares, la eficiencia máxima obtenida tras la clasificación automática sería extremadamente baja. El aumento de subconjuntos de información no afecta el desempeño si el segmento que se desea separar es antagónico a todos los demás, de manera semejante al escenario de estudio en la etapa 1 experimental de la presente tesis, donde el conjunto de páginas de Wikipedia y católicas presentan una alta similitud en términos de ActiveRank entre ellas, pero una gran diferencia con respecto a un tercer conjunto antagónico, en este caso las páginas pornográficas.

3. *Conformación de los vectores de referencia*

Dado que el proceso de autoclasificación basado en ActiveRank utiliza el ranking contra vectores de referencia como criterio de discriminación, afectaciones en la generación de los mismos altera significativamente el desempeño de este proceso. Los principales puntos de falla se encuentran en la incorrecta selección de documentos de referencia para la conformación del vector, en fallas en el proceso de filtrado y depuración del contenido textual del documento, así como en errores sistemáticos producidos por fallas en el parseo, codificación o interpretación de la información digital. Un ejemplo de este tipo de fallas fue lo que nos llevó a desarrollar la extensión de la etapa experimental 2; al operar con un vector de referencia en inglés cuando el universo está conformado por documentos en español inhabilita de manera inmediata la capacidad de discriminación debido a que ninguno de los nodos del vector de referencia es compartido por cualquier otro elemento, o desde otro punto de vista, el vector de referencia se encuentra excluido del universo de información.

4. *Dimensión de los vectores de ActiveRank*

Otro punto fundamental que afecta de manera directa la precisión de ActiveRank es la dimensión de los vectores con los que se trabaja. A pesar de que el algoritmo como tal no limita de manera alguna la extensión o la necesidad de homogenizar el número de componentes de cada vector, existe un punto en el que la expansión

de los vectores no aumenta la eficacia del sistema debido a que su peso relativo es extremadamente pequeño, por lo tanto, es ventajoso económicamente reducir la dimensión de todos los vectores a una longitud óptima, sin embargo, si la dimensión es demasiado pequeña, la información no considerara produce errores importantes en el sistema al otorgar medidas de ranking inconsistentes a la relación o similitud real entre los elementos de la red de información.

### 5.5. Análisis del dominio “unam.mx”

El análisis extendido de las páginas pertenecientes y relacionadas hasta en tercer grado a aquellas bajo el dominio *unam.mx* se desarrolló con éxito superando la cifra de los 70,000 documentos. Tras corroborar el correcto funcionamiento de todos los sistemas incluyendo el proceso de autoclasificación, la interpretación directa de las gráfica 4.3.10. es que no existe ningún documento con contenido pornográfico en inglés, sin embargo, en la gráfica 4.3.9. podemos observar que hay siete documentos con un valor de ranking inferior a 1.65 con respecto al vector de referencia pornográfico en español, los cuales se revisaron manual e individualmente para ser catalogados; ninguno de ellos pertenece al conjunto de interés, en realidad, son páginas que despliegan errores cuyo contenido es extremadamente escaso, en términos generales solo presentan los encabezados del documento. Tras este análisis más detallado, se encontró una deficiencia en el sistema de crawling la cual permitía que documentos con muy pocos nodos fueran ingresados al sistema ActiveRank, y cualquiera de ellos que compartiera unos cuantos nodos, que adicionalmente correspondían a palabras ambivalentes y no representativas del conjunto, obtendría un valor de ranking cercano a 0. Para corroborar adicionalmente este resultado se hizo una búsqueda manual exhaustiva dentro de la base de datos para intentar detectar direcciones que apuntaran a contenido de éste tipo, sin embargo, de acuerdo a lo esperado, no se encontró ninguna fuente relevante.

No hay contenido pornográfico en el segmento de información procesado en la presente tesis.

Un estudio interesante que queda fuera de los alcances de la presente tesis sería analizar sin acotar el escaneo la WWW partiendo desde diferentes fuentes académicas, institucionales y otras seleccionadas aleatoriamente y observar que tan rápido, en términos de distancia de Hamming, se alcanza una página pornográfica.

# Capítulo Sexto

## Conclusiones

# 6. Conclusiones

## 6.1. Conclusiones

1. El algoritmo ActiveRank es un método efectivo y de alta eficiencia en la clasificación automática de contenido dentro de una red de información conformada por múltiples subconjuntos o categorías siempre y cuando se cumplan las siguientes condiciones:
    a) La información por clasificar pertenezca a un conjunto antagónico al resto del universo, y sea posible determinar su pertenencia a través de operaciones binarias (pertenece o no pertenece). Dicho antagonismo puede presentarse de diferentes formas, siendo las más comunes por diferencia de idioma y temática.
    b) Cuando el vector o los vectores de referencia han sido correctamente generados y son realmente representativos del conjunto que se quiere discriminar.
    c) Cuando la dimensión de los vectores ActiveRank es óptima y la pérdida de información por nodos no significativos no altera de manera importante el comportamiento del sistema y en su caso la matriz de rankings ActiveRank.
    d) Cuando los sistemas de filtrado en el proceso de crawling funcionen correctamente retirando información inválida, inconsistente o con errores, y permitiendo el paso de aquella que resulta fundamental para el correcto perfilamiento de un documento dado.
    e) Cuando la información disponible en red pueda ser parseada correctamente y se excluyan errores de codificación, comunicación, interpretación o formato.
2. El proceso de clasificación a través de ActiveRank es una característica adicional en su utilización como motor de relación de información y administración en sistemas de análisis de redes de información, lo que permite disminuir los costos operativos en dichos sistemas al utilizar el mismo núcleo tecnológico para múltiples propósitos.
3. El algoritmo ActiveRank como método automático de clasificación de información es económicamente rentable debido a que no requiere incrementar la infraestructura necesaria para desempeñar dicha tarea, aumentando de manera linealmente su consumo de poder de cómputo con respecto el aumento de nodos en la red, lo cual a comparación del aumento cuadrático en el procesamiento de información cuando se utiliza como motor de relación es significativamente menor.

4. La eficiencia del algoritmo ActiveRank como método automático de clasificación puede disminuir dramáticamente debido a:
    a) Errores en el sistema de crawling que inserten vectores ActiveRank de dimensiones menores al resto de los elementos; lo anterior produce un error sistemático en ActiveRank debido a que dichos elementos tienen mayor probabilidad de obtener una medida de ranking de mayor cercanía por el hecho de que con menos nodos coincidentes se incrementa dicha magnitud, y que además podrían ser no representativos.
    b) El aumento del número de conjuntos que conforman el universo de información analizada, y la reducción de los intervalos entre los valores de ranking medios de cada conjunto, lo que trae como consecuencia directa una reducción en los posibles valores del ranking del umbral de clasificación, que en conjunto con la dispersión de los rankings dentro de los diferentes conjuntos reducen significativamente la capacidad de discriminar la pertenencia de un documento dentro de un conjunto determinado.
    c) Otros errores no predecibles en la operación de los sistemas y subsistemas involucrados en el proceso de indexación, procesamiento y la plataforma DARE, de esto la importancia de mantener un constante y estricto monitoreo de su funcionamiento.
5. En el análisis desarrollado para los 70,000 documentos pertenecientes o accesibles a través de vínculos dentro del dominio *unam.mx* no se encontró contenido pornográfico, utilizando vectores de referencia en inglés y en español, además de un análisis manual de la base de datos recolectada.

### 6.2. Contribuciones

El desarrollo de la presente tesis permitió encontrar nuevas aplicaciones de la tecnología desarrollada por la empresa Ondore, en particular del algoritmo ActiveRank, así como detectar y plantear mejoras a cada uno de los sistemas utilizados en el proceso. Lo anterior ayudará a continuar las diferentes líneas comerciales y de investigación, que a su vez impactará positivamente al generar nuevos empleos en nuestro país y seguir desarrollando oportunidades para los profesionales y la economía de México.

En términos académicos, la presente tesis abarca un gran número de temas relevantes en la investigación y desarrollo de sistemas de análisis de información, y podrá ser utilizada como referencia para futuros trabajos y proyectos al interior de la Universidad Nacional Autónoma de México.

### 6.3. Trabajo futuro

Durante el desarrollo de esta tesis se encontraron diversos puntos que podrían ser de interés futuro tanto en la explotación de la tecnología utilizada, como en la comprensión de algunos fenómenos poco convencionales que podrían no ser triviales y derivar en nuevas líneas de investigación del algoritmo ActiveRank; a continuación se describen en términos generales para sentar base de trabajos futuros.

- *Velocidad y probabilidad de encontrar una página pornográfica en subgráficas de la WWW*

A través de la utilización de la tecnología de análisis de información de Ondore en una implementación muy similar a la desarrollada en esta tesis, es posible detectar el momento y ubicación de una página pornográfica al iniciar una exploración de la WWW desde un punto aleatoriamente seleccionado; repitiendo este experimento un número de veces suficientemente grande sería posible encontrar la velocidad, distancia y probabilidad promedio de llegar a una página web pornográfica al iniciar desde cualquier punto de la WWW. Como primera aproximación, se deberían encontrar valores altos para la velocidad y bajos para la distancia, es decir, que partiendo de cualquier página de la WWW se puede llegar en pocos saltos a una página pornográfica. Es posible seleccionar los puntos de inicio de análisis dentro de subgráficas específicas como podrían ser páginas de universidades, de gobierno, personales, redes sociales, etc., con el fin de obtener un mejor rango de valores para los diferentes tipos de escenarios.

- *Efectos estadísticos de la normalización de un vector*

Durante las primeras etapas del desarrollo del algoritmo ActiveRank se estudió el comportamiento de cada una de sus operaciones en vectores generados de manera aleatoria. Si se genera un vector de dos dimensiones cuyas componentes han sido seleccionadas aleatoriamente del intervalo [0,1] con una distribución uniforme (todos los valores entre 0 y 1 tienen la misma probabilidad de ser obtenidos), este puede ser representado como un punto en el plano unitario del espacio coordenado de 2 dimensiones, al incrementar el número de vectores todos quedarían repartidos uniformemente en dicho plano; análogamente, si habláramos de vectores de 3 dimensiones generados con el mismo procedimiento, estos quedarían uniformemente distribuidos en el cubo unitario del espacio coordenado de 3 dimensiones.

Al normalizar los vectores la condición que se hace cumplir es que la suma de sus componentes sea igual a 1, lo que genera una dependencia o relación limitativa entre las variables, modificando la función de

distribución de cada una de ellas, que a su vez condensa los puntos en una estructura de 1 dimensión menor a la que originalmente tenían. Lo anterior es fácilmente apreciable en las figuras a continuación mostradas, donde para 2 variables, los puntos uniformemente distribuidos en el plano y cubo unitario quedan restringidos a un segmento de línea y plano respectivamente. [8]

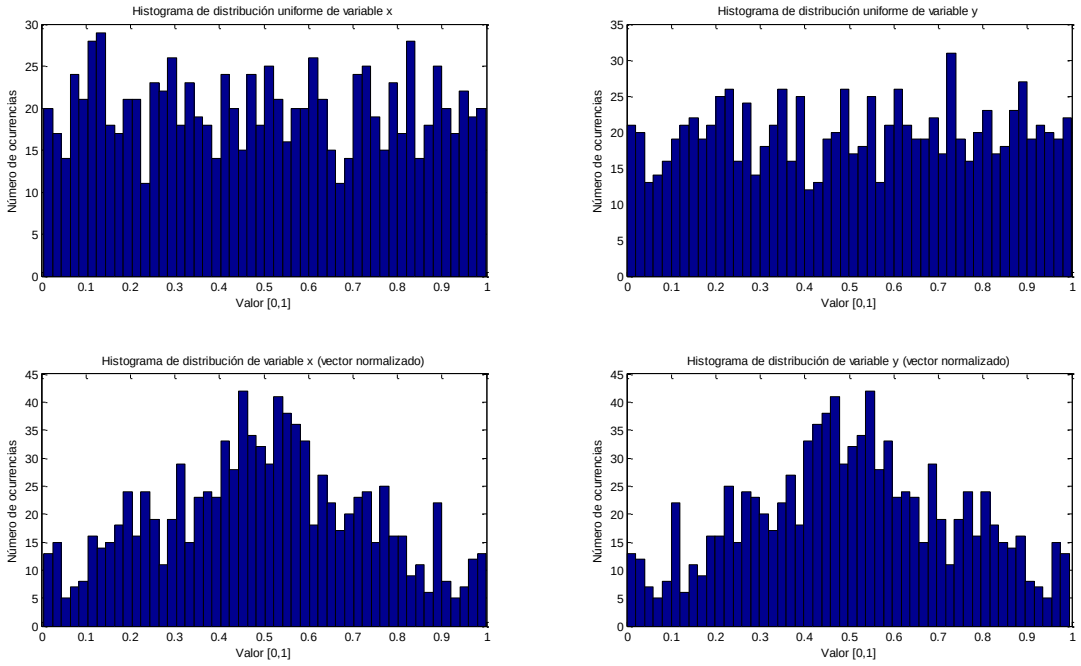



Figura 6.3.1. – Histograma de valores obtenidos por las 2 variables de 1000 vectores de 2 dimensiones generados aleatoriamente antes y después de la normalización.

5. A partir de mi apreciación y sin haber desarrollado ninguna comprobación, la función de distribución parece cambiar de uniforme a distribución chi, la cual es una generalización de la distribución de Rayleigh para *n* variables. Lo anterior se basa en que para el escenario de 2 dimensiones, la nueva distribución parece ser normal, sin embargo, al incrementar el número de dimensiones, la campana se desplaza hacia los valores inferiores, lo que resulta lógico dado que es más probable que todas las variables tengan valores chicos, a que todas tengan valores grandes debido a la relación que produce la condición de normalización.

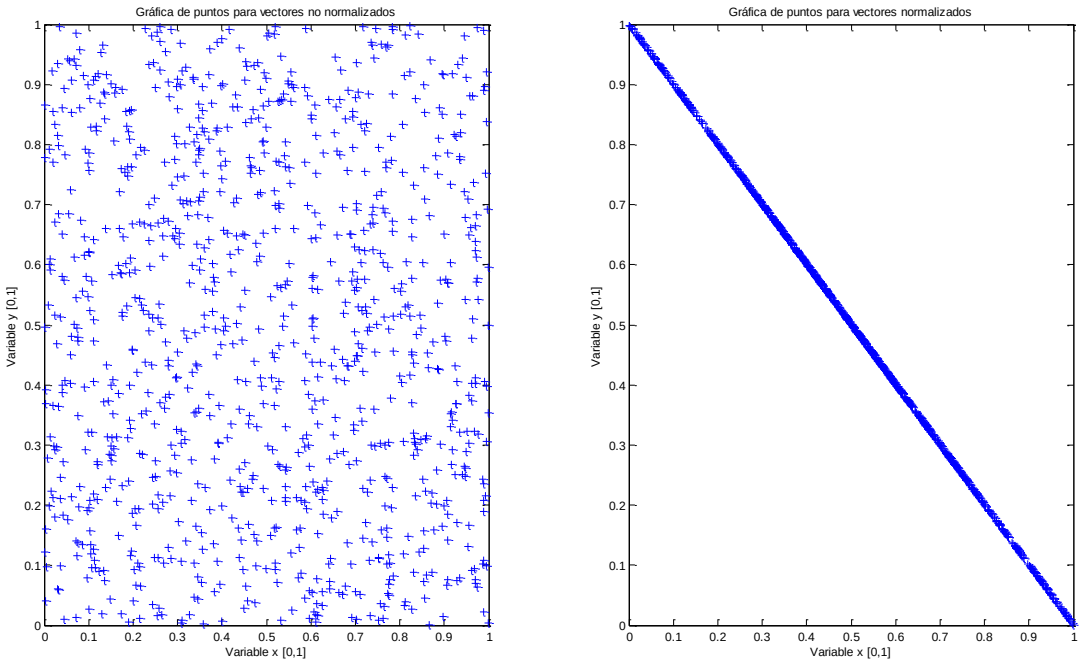



Figura 6.3.2. – Distribución de 1000 vectores de 2 dimensiones generados aleatoriamente antes y después de la normalización.

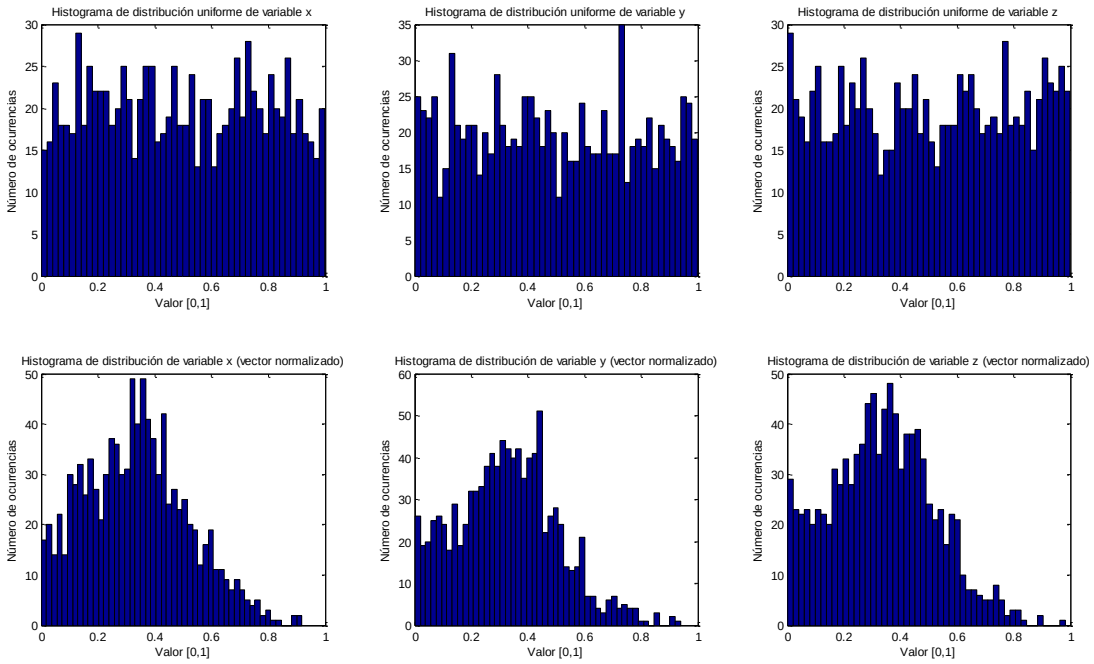



Figura 6.3.3. – Histograma de valores obtenidos por las 3 variables de 1000 vectores de 3 dimensiones generados aleatoriamente antes y después de la normalización; se puede apreciar el desplazamiento lateral de la curva en la distribución de valores después de la normalización.

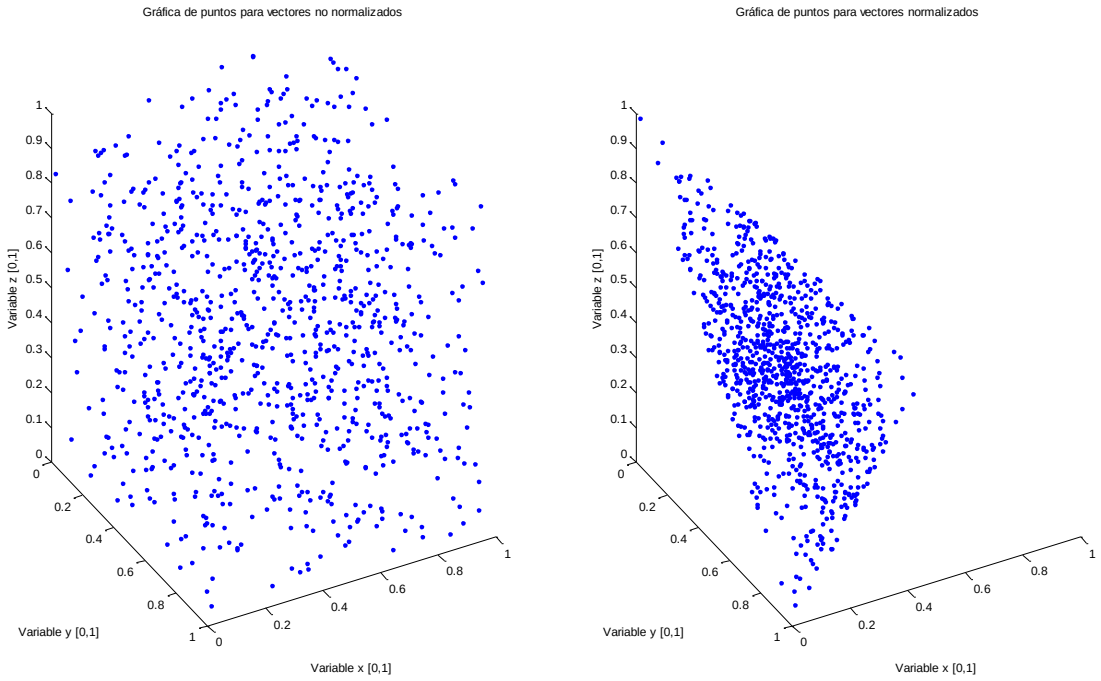



Figura 6.3.4. – Distribución de 1000 vectores de 3 dimensiones generados
aleatoriamente antes y después de la normalización; se puede apreciar como los puntos quedan limitados
a un plano triangular concentrando la densidad en la sección donde las 3 variables tienen un valor aproximado de 0.4

El tema central de investigación no es en realidad el por qué se modifica la función de probabilidad de cada una de las componentes, dado que se conoce que se trata de un efecto de correlación causado por la operación de normalización, sino el impacto que puede tener en la pérdida de información estructural que se da al colocar un punto de *n* dimensiones en una estructura de *n-1* dimensiones.

- *Patrones lineales en la matriz de rankings de ActiveRank*

Una observación importante realizada durante el desarrollo del algoritmo ActiveRank fue la presencia de patrones lineales al representar en una escala de 256 grises la matriz de rankings (ver figura 6.3.5.) de un universo de vectores generados de manera aleatoria contra sí mismo. La interpretación de los patrones lineales observados en la figura antes mencionada es que existe una relación de similitud homogénea entre un vector dado y todos los demás, lo que se opone en gran medida al hecho de que cada uno de los vectores fue generado de manera completamente independiente por lo que la relación entre ellos también debería encontrarse

uniformemente distribuida en intervalo de valores entre 0 y 1.

Si se observa con detenimiento la imagen, y su contraparte numérica, también se obtienen pocos valores extremos, es decir, la mayor concentración de calificaciones se da para los valores cercanos a 0.5, por esto es que se observa un tono homogéneo de gris en la mayor parte de la figura; la diagonal principal blanca se debe a que el ranking de un vector contra sí mismo es en todos los casos igual a 1.

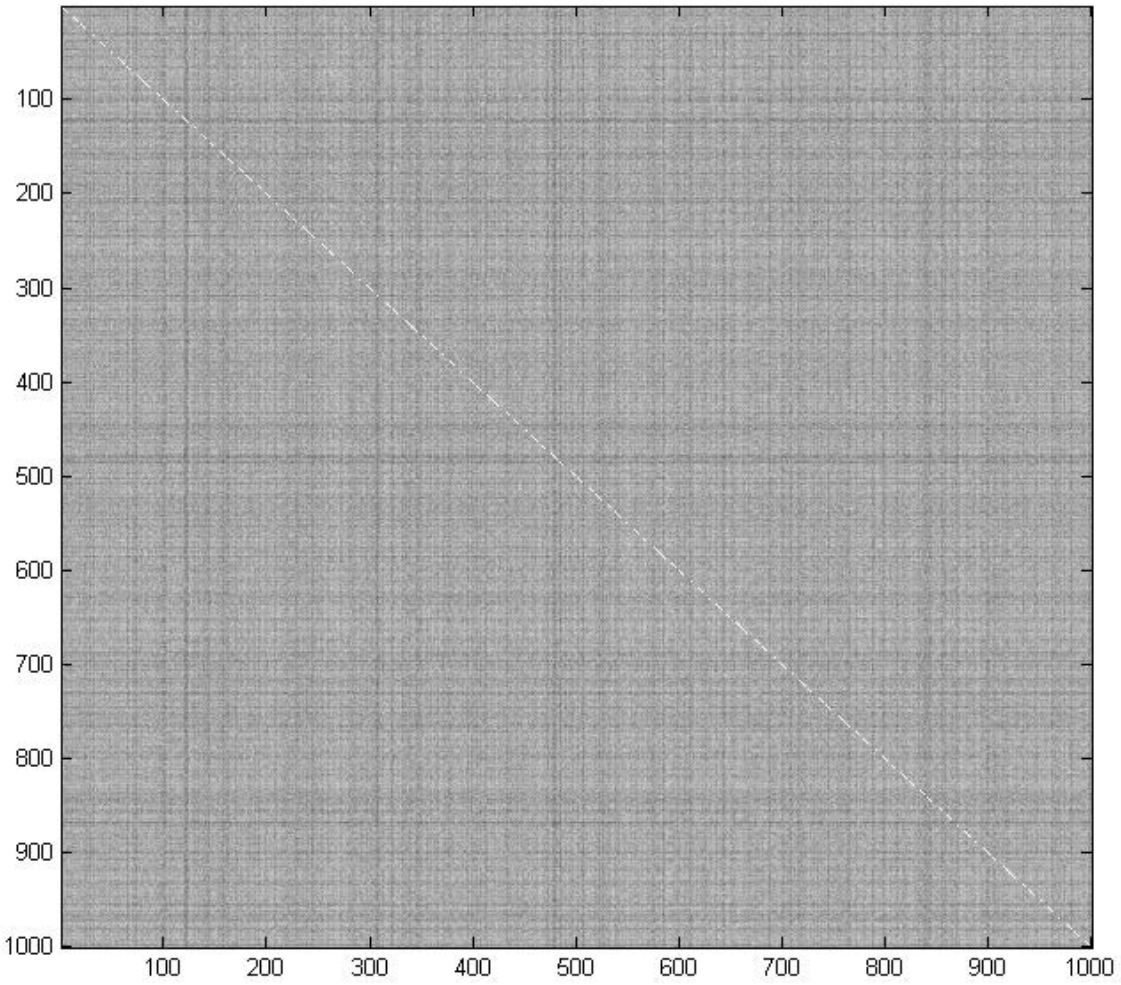



Figura 6.3.5. – Patrones lineales detectados al representar la matriz de rankings de un universo de vectores generados aleatoriamente contra sí mismo en una escala de 256 tonos de gris.

El trabajo futuro en este aspecto es, al igual que el punto anterior, analizar con detalle el efecto de cada una de las operaciones que se llevan a cabo en el algoritmo ActiveRank y determinar si la afectación en la distribución estadística de los valores que produce la normalización es causante de este tipo de correlaciones en

la matriz de rankings, o si son eventos independientes que deben ser analizados separadamente.

- *Implementación de algoritmos de aceleración y mejoras en general en los sistemas de análisis de información de Ondore*

Por último, uno de los principales puntos de desarrollo futuro encontrados tras la realización de esta tesis fue la amplia gama de mejoras en el desempeño de los sistemas de análisis de información que pueden ser implementadas, no porque actualmente sean considerados ineficientes, sino por la gran cantidad de oportunidades de mejora existentes, entre las que destacan:

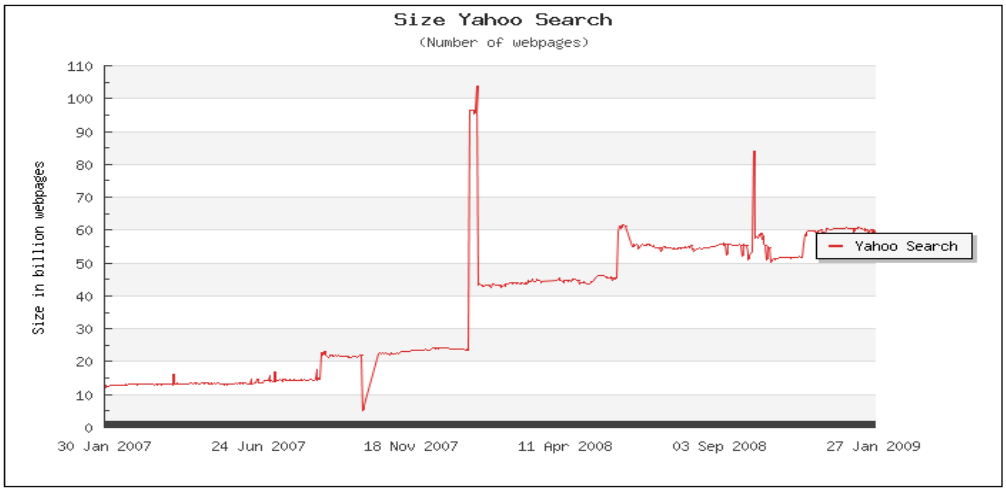
- Sistemas de ruteo de peticiones web entrantes/salientes y servidores proxy que ayuden a reducir el tiempo de conexión entre la aplicación de análisis y las páginas deseadas, permitiendo la incorporación de múltiples servidores de análisis coordinando la exploración.
- Plantear un modelo de cómputo paralelo análogo al utilizado por el sistema DARE (Distributed ActiveRank Engine) en los sistemas Analyzer, de tal manera que fuera posible el mantener mayores ritmos de procesamiento a un costo técnica y económicamente accesibles.
- Trabajar en la simplificación de consultas a bases de datos, así como la revisión de la correcta utilización de índices en las tablas y consideración de algoritmos de aceleración de consultas como podría ser BWA (*Business Warehouse Accelerator*); desincorporar los servidores de base de datos de los servidores donde están instalados los sistemas Analyzer para aumentar la escalabilidad de los procesos al poder utilizar servidores NAS (*Network Attached Storage*) con plataformas de bases de datos distribuidas.
- Seguir analizando el potencial del algoritmo ActiveRank al ser utilizado en tiempo real y desarrollar métodos simplificados para su aceleración. En el ámbito del procesamiento en paralelo del sistema DARE, es posible continuar extensivamente la investigación de su viabilidad operando en infraestructura de arquitectura mixta (tanto en software como en hardware), utilizando plataformas convencionales para cómputo distribuido, entre otras.
- Expandir la funcionalidad y usabilidad de las interfases gráficas actualmente disponibles para mejorar su aprovechamiento, simplificando la navegación y la obtención de resultados, así como la evaluación de los diferentes indicadores estadísticos sobre el desempeño y eficiencia de los sistemas; la creación de *backpanels* para la administración, configuración y depuración de las plataformas de análisis de información pueden reducir significativamente el costo que tienen estas actividades al disminuir el tiempo que se invierte en estos procesos.

# Apéndices

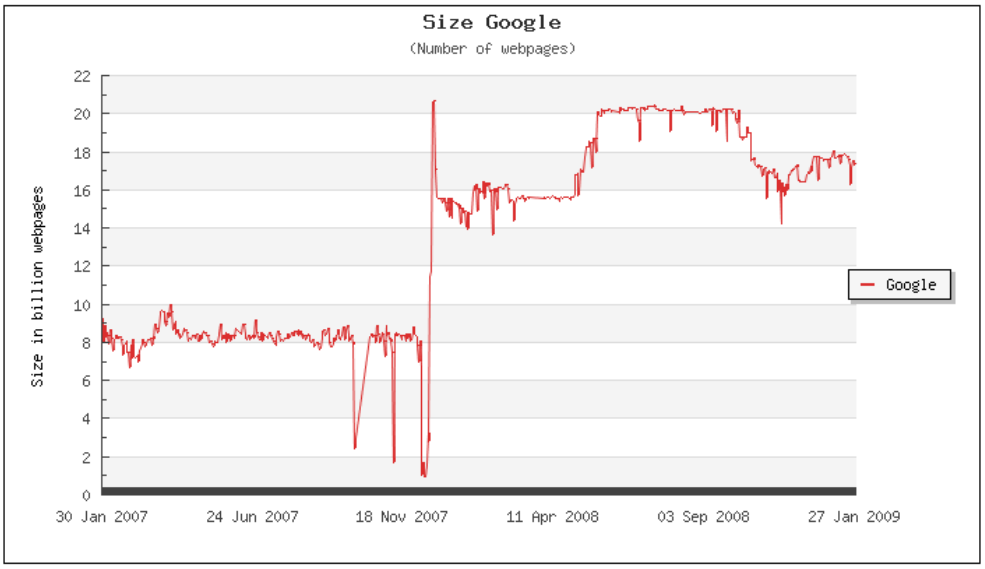
# Apéndice A

## Dimensiones de la World Wide Web

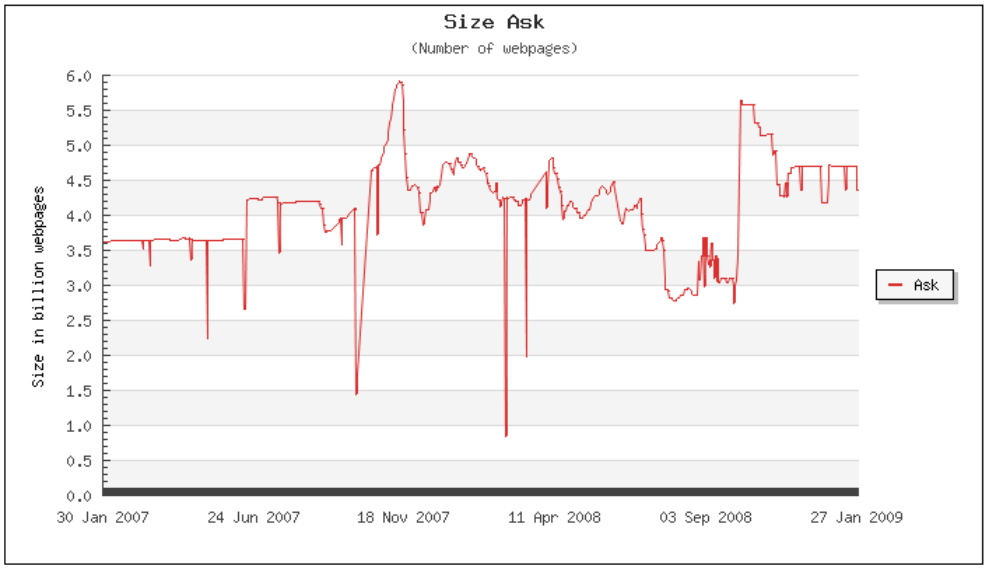
Las diferentes gráficas que a continuación se muestran han sido obtenidas del sitio web *http://www.worldwidewebsize.com*; el contenido ha sido seleccionado con la intención de introducir únicamente información necesaria para la comprensión de las ideas correspondientes a las dimensiones de la WWW.




Gráfica A.1. – Número de páginas indexadas por Yahoo a Enero de 2009




Gráfica A.2. – Número de páginas indexadas por Google a Enero de 2009




Gráfica A.3. – Número de páginas indexadas por Ask a Enero de 2009

# Apéndice B

## Diagrama de flujo de un sistema básico de análisis de la WWW

A continuación se describe de manera general el proceso de análisis de información mediante un diagrama de flujo representado en las figuras B.1. a B.3.; los bloques en azul representan aquellas actividades básicas para poder explorar la WWW, mientras que aquellas en rojo deben ser interpretadas como los procesos llevados a cabo para cumplir un fin específico, en este caso, un análisis semántico de la información textual.

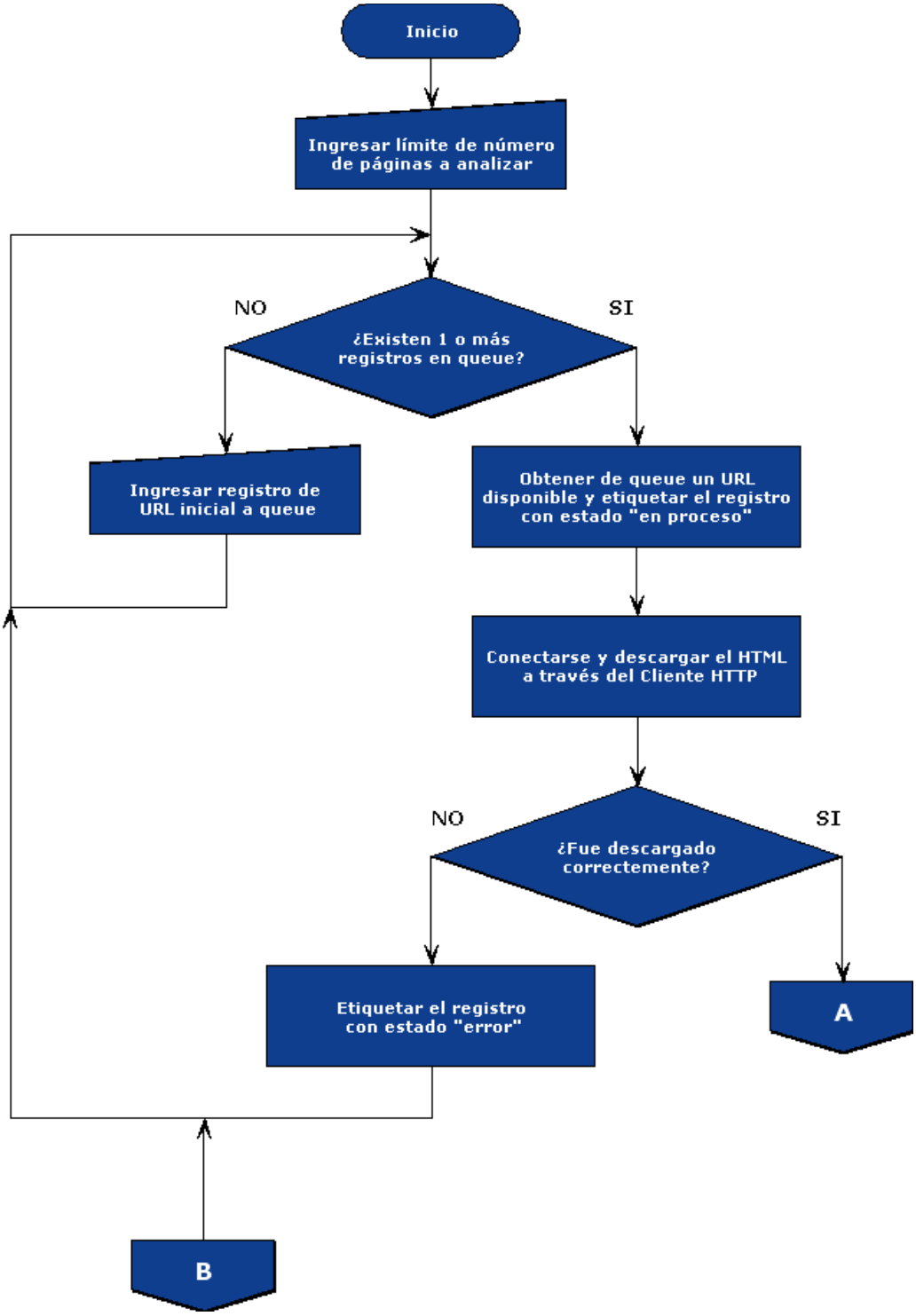



Figura B.1. – Diagrama de flujo de un crawler básico

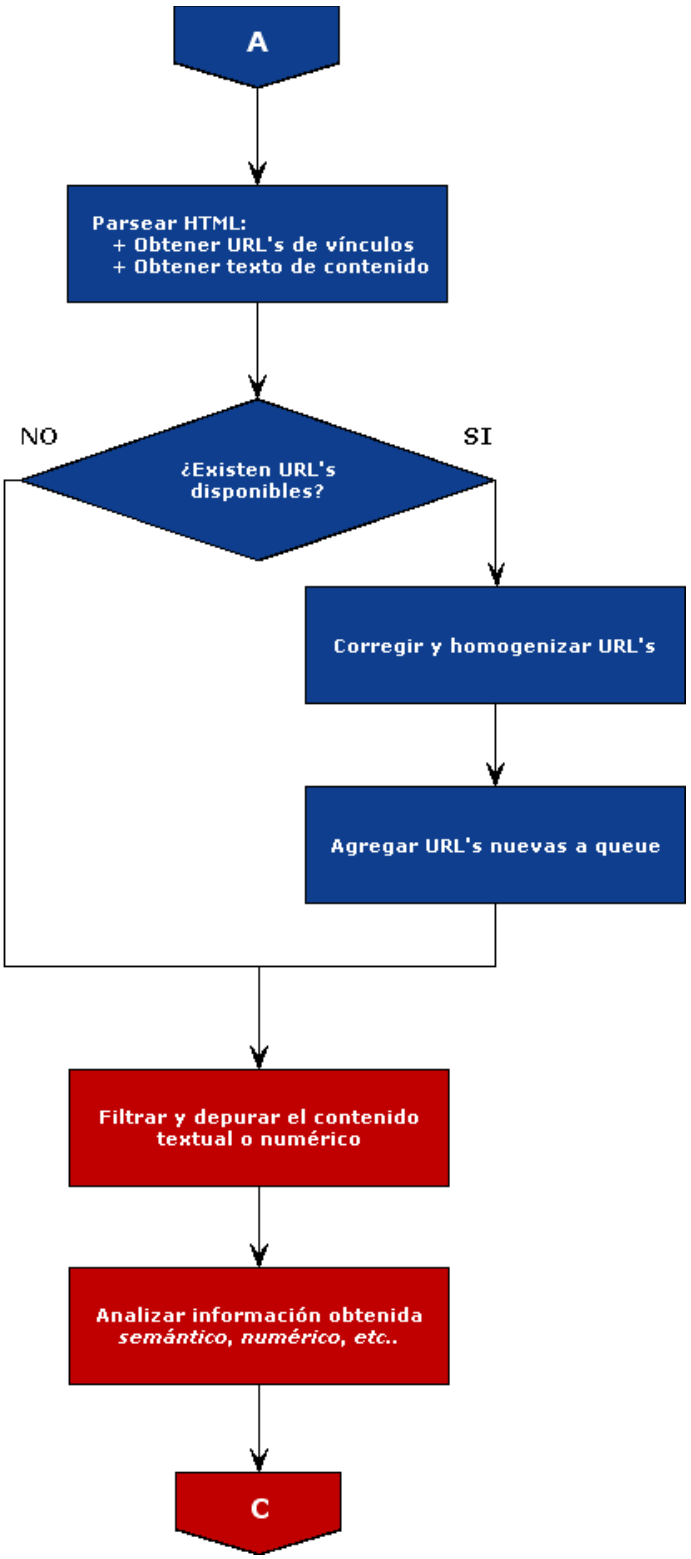



Figura B.2. – Diagrama de flujo de un crawler básico

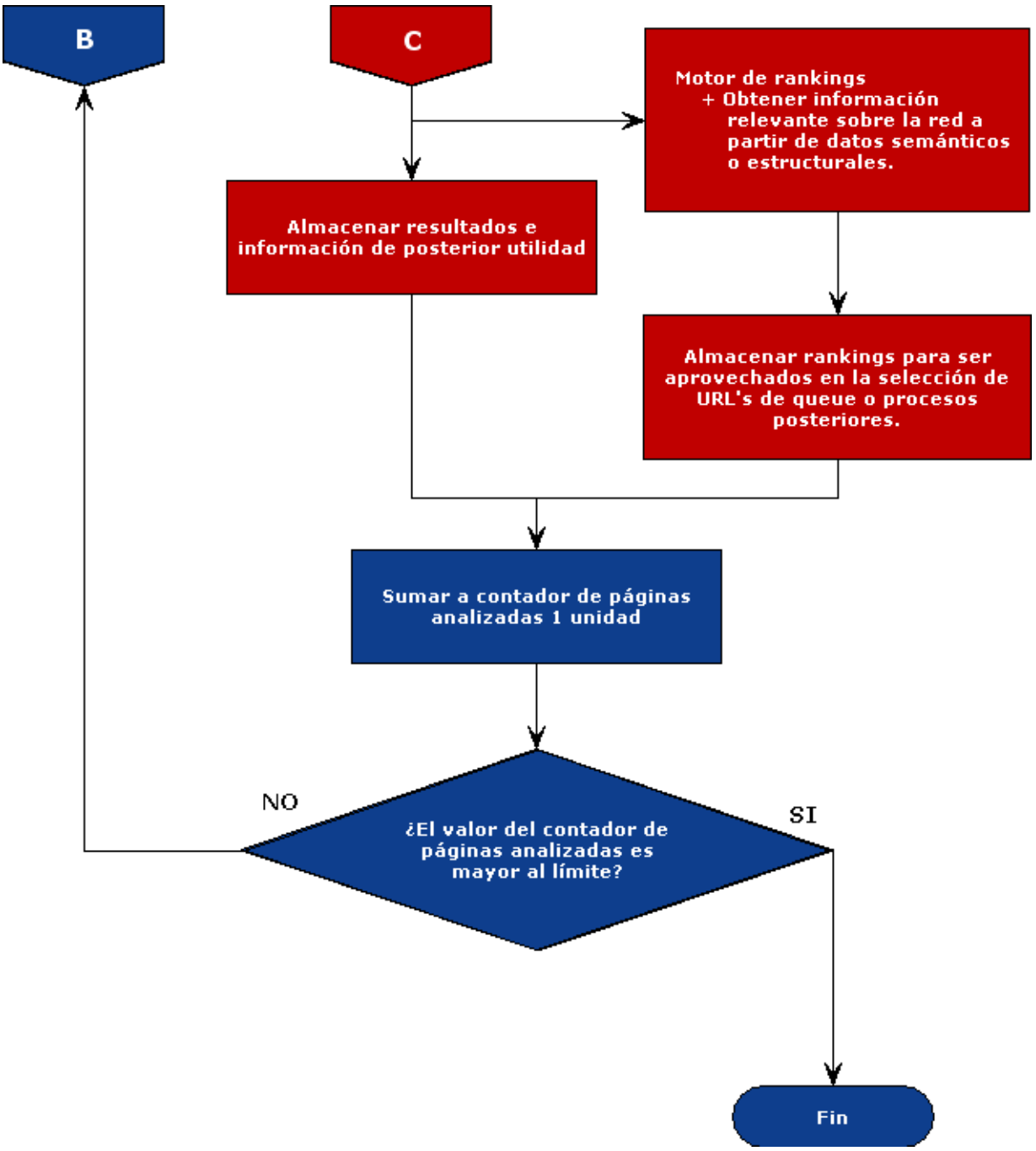



Figura B.3. – Diagrama de flujo de un crawler básico

# Apéndice C

## Información estadística sobre la pornografía en Internet – Año 2006

Las tablas que a continuación se muestran han sido obtenidas del sitio web *http://internet-filter-review.toptenreviews.com/internet-pornography-statistics.html*; el contenido ha sido traducido y depurado para facilitar la comprensión de los datos, lo anterior sin intención de manipular o sesgar los resultados.

Tabla C.1. - Ingresos mundiales de la industria dedicada a la comercialización de la pornografía en el año 2006.

**Ingresos Mundiales por Pornografía – 2006**

| País | Ingresos [Billones USD] | Notas |
|---|---|---|
| China | $27.40 | Incompleto |
| Corea del Sur | $25.73 | |
| Japón | $19.98 | |
| Estados Unidos | $13.33 | |
| Australia | $2.00 | |
| Reino Unido | $1.97 | |
| Italia | $1.40 | |
| Canadá | $1.00 | |
| Filipinas | $1.00 | |
| Taiwan | $1.00 | Incompleto |
| Alemania | $.64 | Incompleto |
| Finlandia | $.60 | Incompleto |
| República Checa | $.46 | Incompleto |
| Rusia | $.25 | Incompleto |
| Holanda | $.20 | |
| Brazil | $.10 | Incompleto |
| Otros 212 paises | - | No Disponible |

Tabla C.2. - Distribución demográfica aproximada en la búsqueda de términos para adultos en el año 2006.

**Distribución Demográfica Aproximada en Búsqueda de Contenido para Adultos – 2006**

| | Distribución Aproximada |
|---|---|
| Menores a 18 años | 20.50% |
| Entre 18 y 24 años | 19.50% |
| Entre 25 y 34 años | 18.30% |
| Entre 35 y 49 años | 23.50% |
| Mayores a 50 años | 18.20% |

Tabla C.3. - Términos más buscados de contenido para adultos con división por género en el año 2006.

**Términos más Buscados de Contenido para Adultos – 2006**

| Término de Búsqueda | Peticiones de Búsqueda en 2006 | División por Género | | Páginas Web que Contienen el Término [Millones] |
|---|---|---|---|---|
| | | Masculino | Femenino | |
| Sex | 75,608,612 | 50.00% | 50.00% | 414 |
| Adult Daiting | 30,288,325 | 36.00% | 64.00% | 1.4 |
| Adult DVD | 13,684,718 | 58.00% | 42.00% | 1.82 |
| Porn | 23,629,211 | 96.00% | 4.00% | 88.8 |
| Sex Toys | 15,955,566 | 58.00% | 42.00% | 2.65 |
| Teen Sex | 13,982,729 | 44.00% | 56.00% | 2.1 |
| Free Sex | 13,484,769 | 44.00% | 56.00% | 2.42 |
| Adult Sex | 13,362,995 | 36.00% | 64.00% | 1.58 |
| Sex Ads | 13,230,137 | 50.00% | 50.00% | 0.28 |
| Group Sex | 12,964,651 | 50.00% | 50.00% | 2.07 |
| Free Porn | 12,964,651 | 97.00% | 3.00% | 2.74 |
| XXX | 12,065,000 | 50.00% | 50.00% | 181 |
| Sex Chat | 11,861,035 | 50.00% | 50.00% | 2.21 |
| Anal Sex | 9,960,074 | 67.00% | 33.00% | 2.95 |
| Cyber Sex | 8,502,524 | 41.00% | 59.00% | 1.24 |
| XXX Videos | 7,411,220 | 64.00% | 37.00% | 1.44 |
| Playboy | 6,641,209 | 86.00% | 14.00% | 43.2 |
| Teen Porn | 6,130,065 | 82.00% | 18.00% | 1.97 |
| Nude | 5,487,925 | 77.00% | 23.00% | 71.3 |
| Sexy | 4,344,924 | 50.00% | 50.00% | 198 |

Tabla C.4. - Información estadística sobre la pornografía en Internet en el año 2006.

**Información Estadística sobre la Pornografía en Internet – 2006**

| | |
|---|---|
| Páginas web pornográficas | 420 millones |
| Búsquedas diarias sobre pornografía en buscadores | 68 millones (25% del total de búsquedas en Internet) |
| Correos electrónicos enviados diariamente con contenido pornográfico o sexual | 2.5 billiones (8% del total de correos electrónicos) |
| Usuarios de Internet que accesan a contenido pornográfico | 42.70% |
| Usuarios con exposición a contenido sexual no deseado | 34.00% |
| Correos electrónicos diarios con contenido sexual por usuario | 4.5 por usuario de Internet |
| Descargas (Peer To Peer) mensuales de contenido pornográfico | 1.5 billiones (35% del total de descargas) |

Tabla C.5. - Número de páginas pornográficas por país en el año 2006.

**Páginas Pornográficas por País – 2006**

| País | Páginas Pornográficas [millones] |
|---|---|
| Estados Unidos | 244.66 |
| Alemania | 10.03 |
| Reino Unido | 8.51 |
| Australia | 5.66 |
| Japón | 2.7 |
| Holanda | 1.88 |
| Rusia | 1.08 |
| Polonia | 1.05 |
| España | 0.85 |

Tabla C.6. - Información demográfica sobre el acceso y consumo de contenido pornográfico en Internet en el año 2006.

**Información Demográfica y Pornografía en Internet – 2006**

| Adolescentes | |
|---|---|
| Edad de primera exposición a contenido pornográfico | 11 años |
| Grupo consumidor de pornografía en Internet | 35 - 49 años |
| Grupo de adolescentes entre 15 y 17 años con experiencia sexual | 80.00% |
| Grupo de adolescentes entre 8 y 16 años que han accesado a contenido pornográfico en Internet | 90% (la mayoría al realizar tarea escolar) |
| Grupo de adolescentes entre 7 y 17 alos que proporcionan su dirección de casa abiertamente | 29.00% |
| Grupo de adolescentes entre 7 y 17 alos que proporcionan correo electrónico abiertamente | 14.00% |
| **Mujeres** | |
| Mujeres que mantienen en secreto su acceso a contenido pornográfico | 70.00% |
| Mujeres con adicción a la consulta de contenido pornográfico | 17.00% |
| Proporción de mujeres:hombres en salas de chat virtuales | 02:01:00 |
| Relación mujeres:hombres que accesan a sitios pornográficos | 01:03:00 |
| Mujeres que accesan a sitios pornográficos | 9.4 milliones (mensualmente) |
| Mujeres que admiten accesar a contenido pornográfico en el trabajo | 13.00% |
| Hombres | |
| Hombres que admiten accesar a contenido pornográfico en el trabajo | 20.00% |
| Adultos estadounidenses que acceden regularmente a sitios pornográficos | 40 milliones |
| Hombres con adicción a la consulta de contenido pornográfico | 10.00% |

# Apéndice D

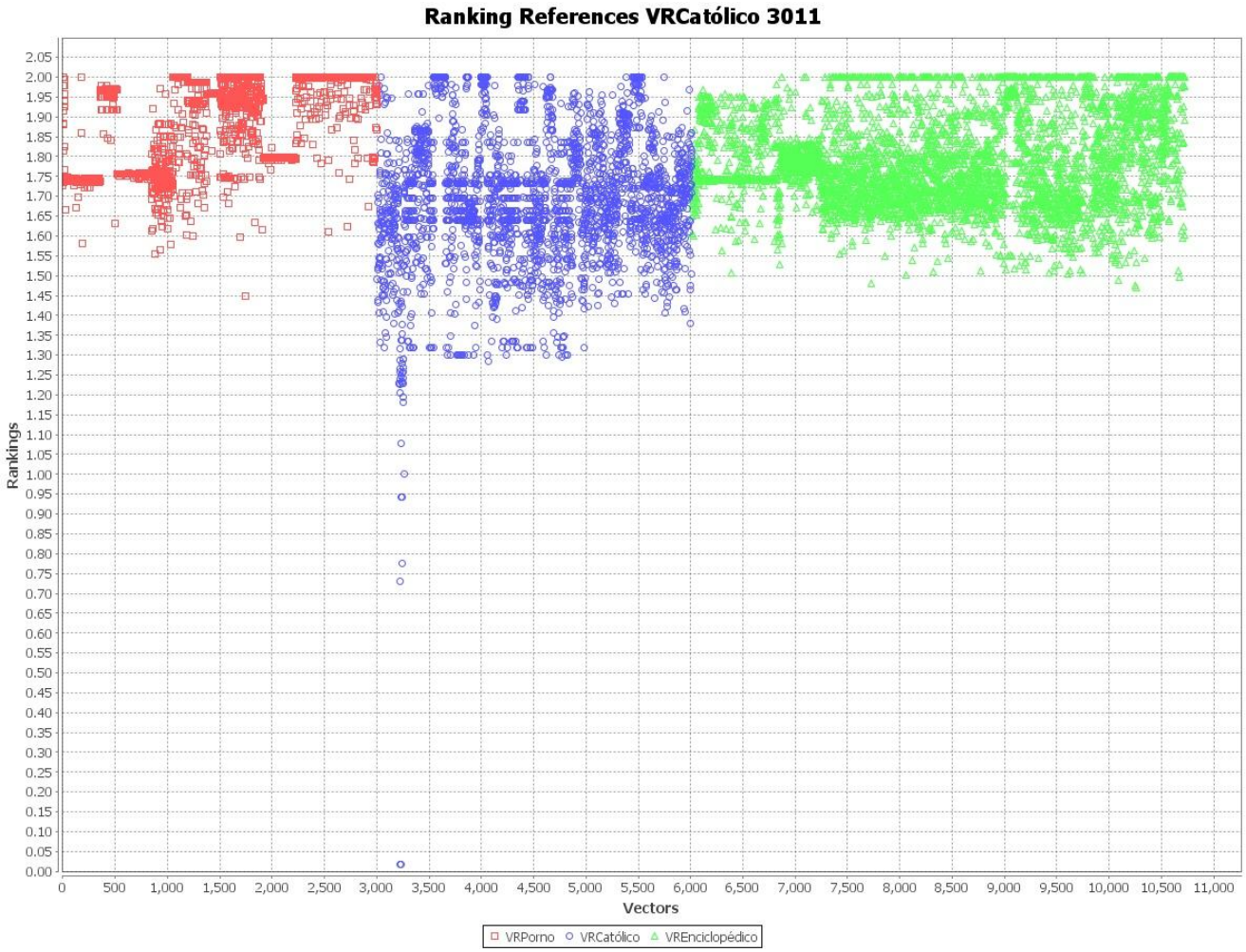
## Resultados complementarios

A continuación se muestran algunas gráficas complementarias utilizando procedimientos análogos a los planteados en el desarrollo experimental de esta tesis con los conjuntos de información alternos; así mismo, se puede generar el mismo análisis aplicado en la clasificación de contenido pornográfico y obtener conclusiones particulares para cada caso.

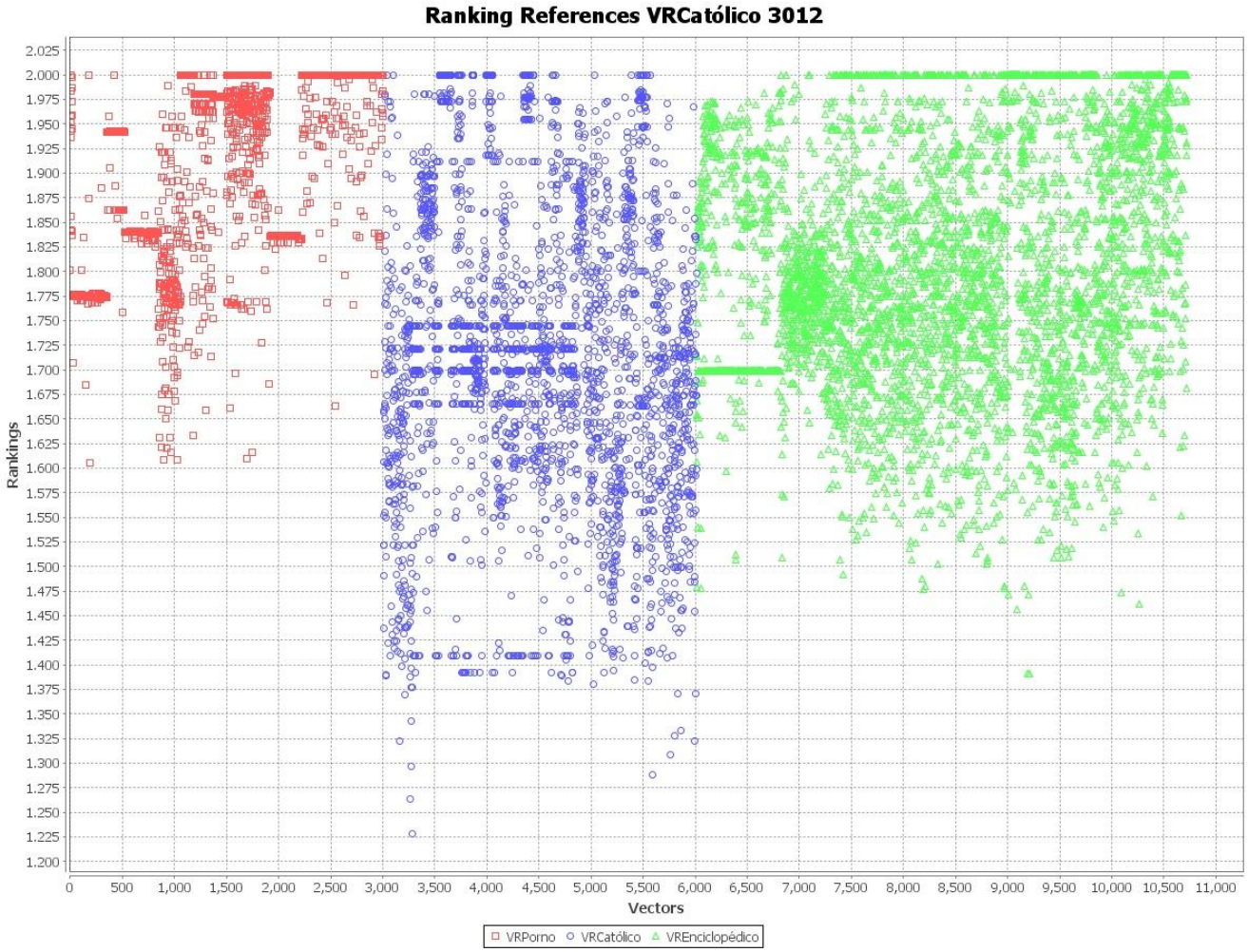

Gráfica D.1. – Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector de referencia católico contra todos los demás.

Tabla D.1. – Distribución de valores máximos, mínimos y medios para cada entorno de información con respecto al vector de referencia católico.

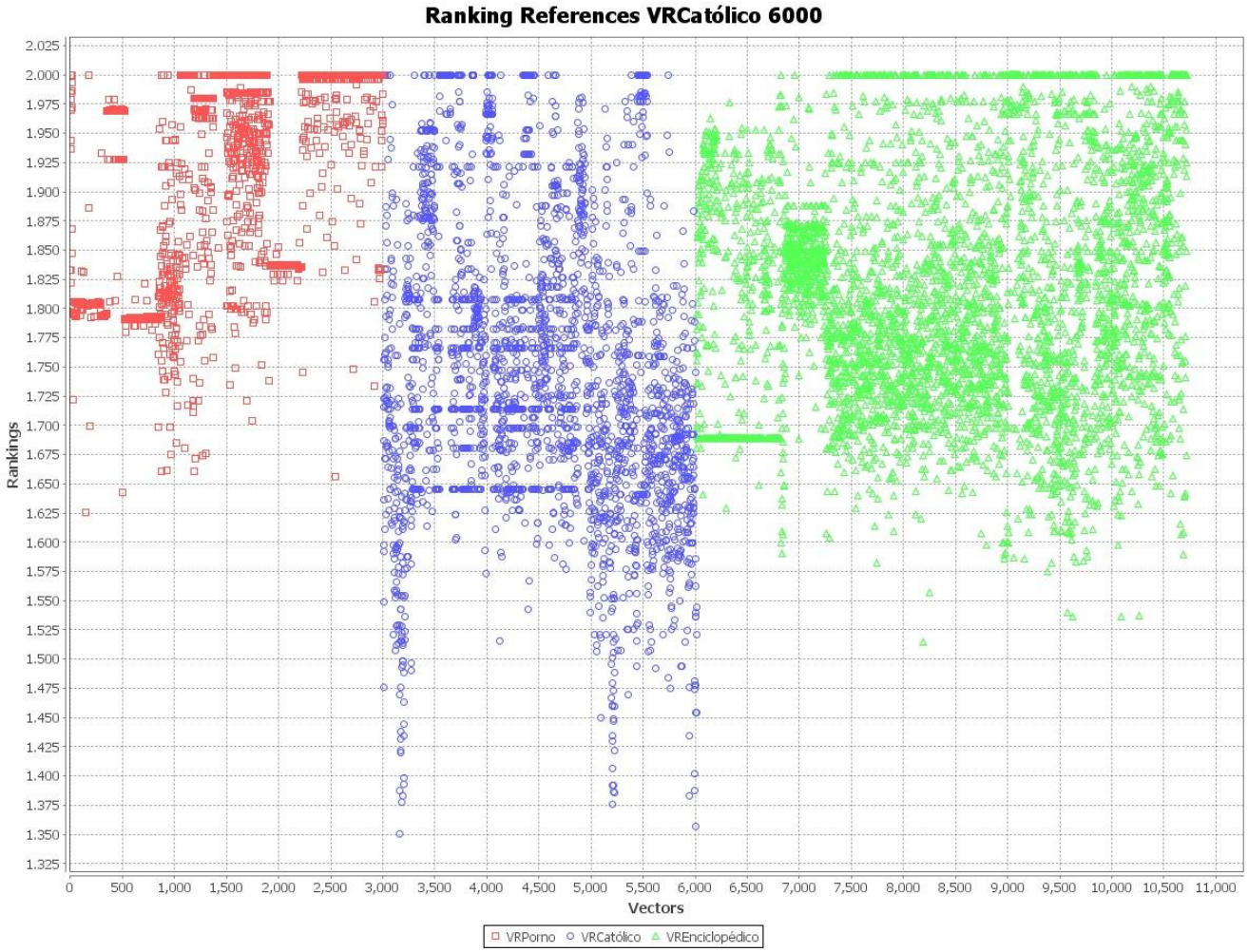
| | $\bar{r}$ | $\bar{r}_{min}$ | $\bar{r}_{max}$ |
|---|---|---|---|
| **Entorno Pornográfico** | 1.8846 | 1.5082 | 2 |
| **Entorno Católico** | 1.5867 | 1.0885 | 2 |
| **Entorno Enciclopédico** | 1.7831 | 1.0885 | 2 |



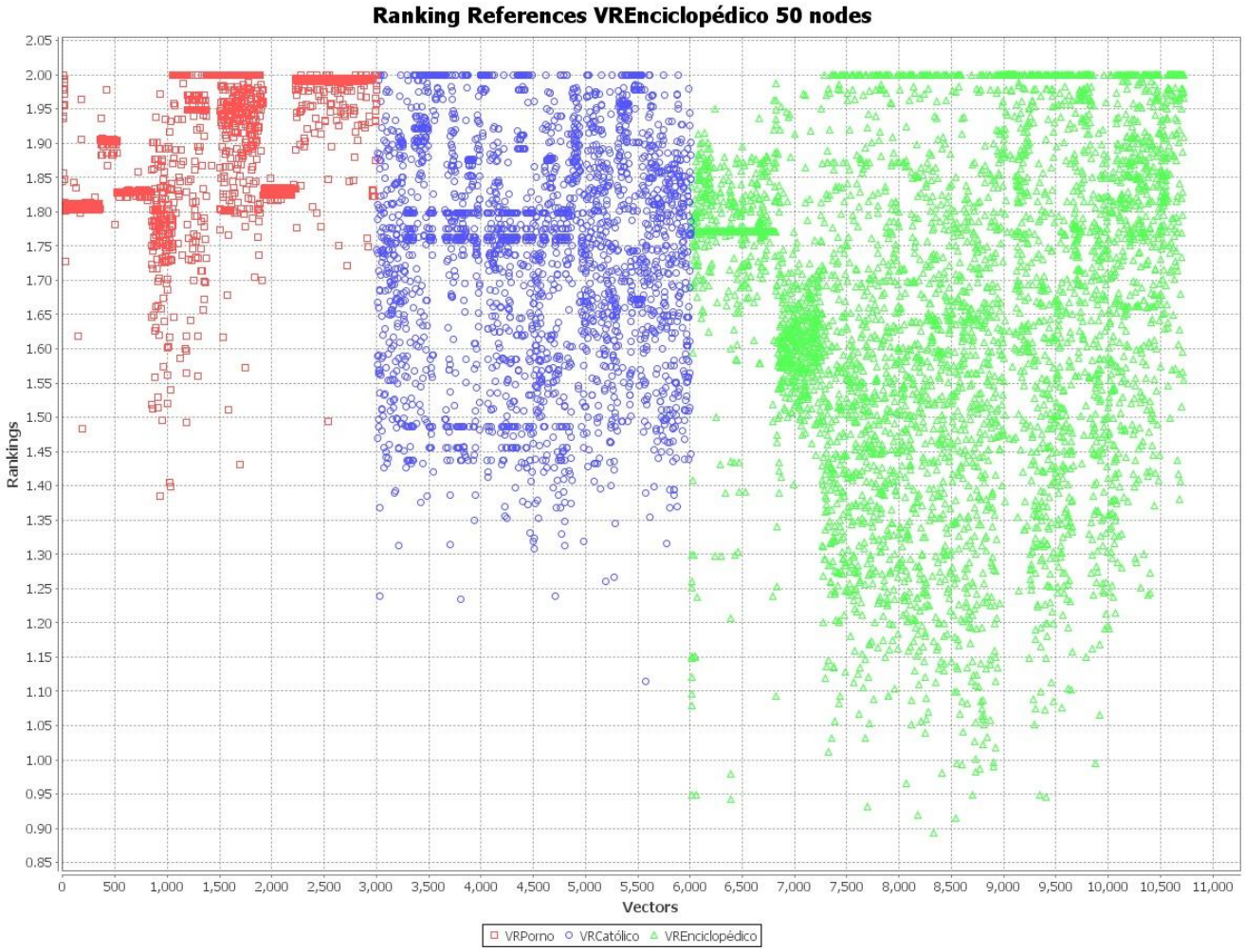

Gráfica D.2. – Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector 3011 del conjunto católico contra todos los demás.




Gráfica D.3. – Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector 3012 del conjunto católico contra todos los demás.



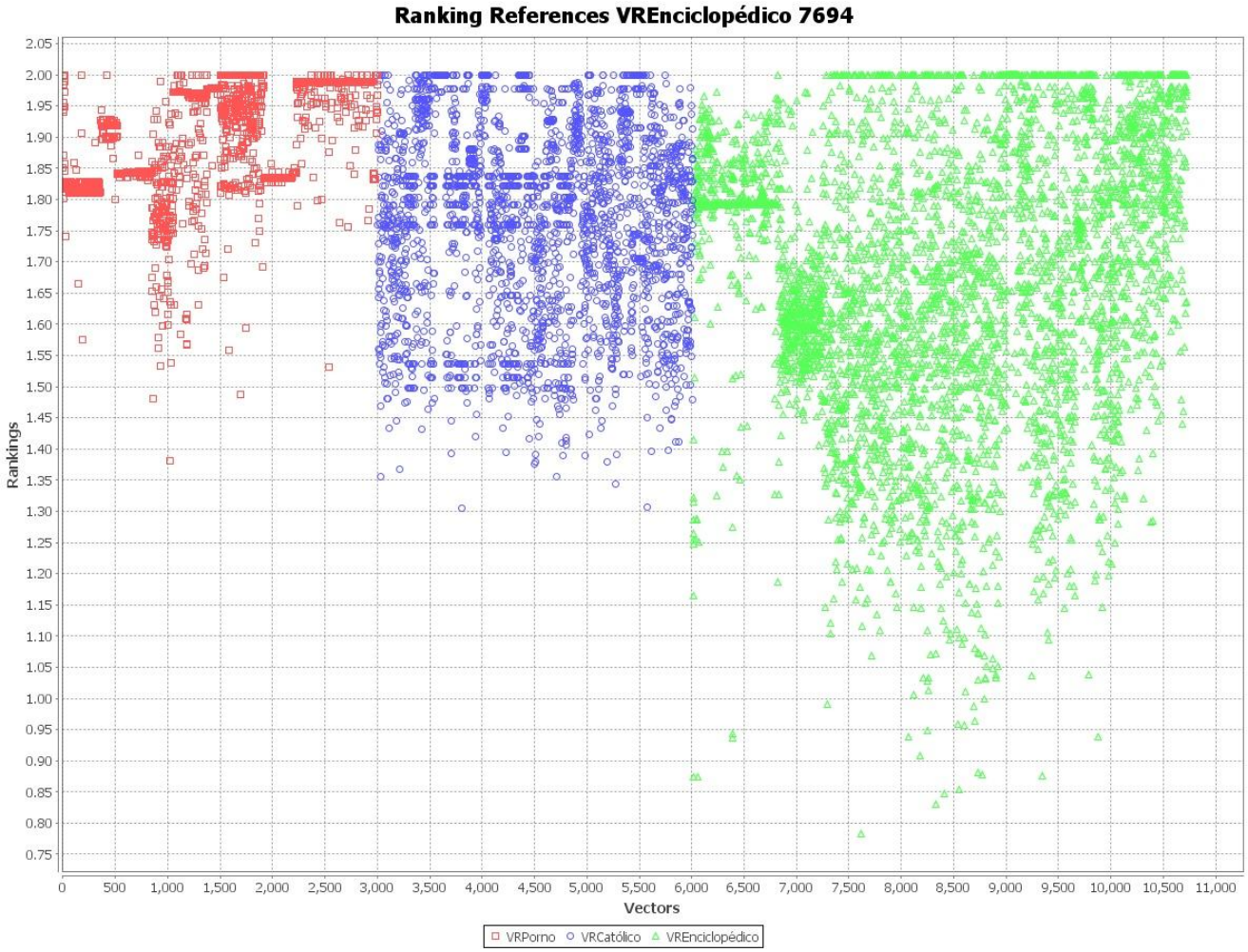

Gráfica D.4. – Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector 6000 del conjunto católico contra todos los demás.



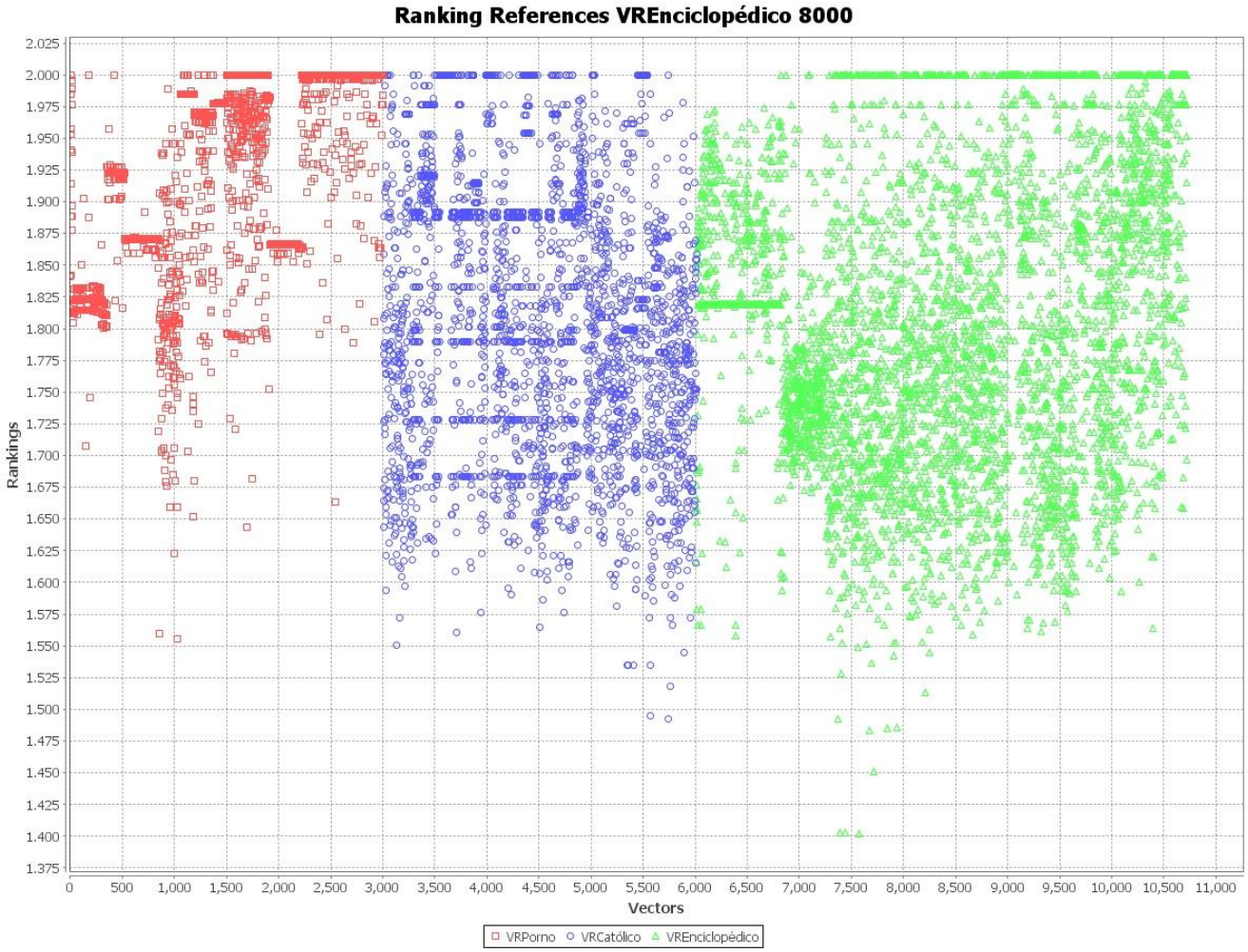

Gráfica D.5. – Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector de referencia enciclopédico contra todos los demás.

Tabla D.2. – Distribución de valores máximos, mínimos y medios para cada entorno de información con respecto al vector de referencia enciclopédico.

| | $\bar{r}$ | $\bar{r}_{min}$ | $\bar{r}_{max}$ |
|---|---|---|---|
| **Entorno Pornográfico** | 1.8998 | 1.3853 | 2 |
| **Entorno Católico** | 1.7556 | 1.1149 | 2 |
| **Entorno Enciclopédico** | 1.5670 | 0.8939 | 2 |




Gráfica D.6. – Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector 7694 del conjunto enciclopédico contra todos los demás.




Gráfica D.7. – Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector 8000 del conjunto enciclopédico contra todos los demás.


Gráfica D.8. – Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector 8237 del conjunto enciclopédico contra todos los demás.

Las gráficas D.1. a D.8. fueron obtenidas al calcular los rankings de cada uno de los vectores especificados contra todos los demás elementos inmersos en el sistema; su análisis puede ser desarrollado siguiendo los procedimientos establecidos en el capítulo quinto de la presente tesis. A partir de una rápida interpretación, las observaciones más importantes son las siguientes:

- Mientras más semejantes sean los conjuntos de información entre sí, menor será la separación generada a través del ranking con respecto a un mismo vector de referencia, por lo tanto, mayor será la dificultad para clasificar dicho contenido, o en otras palabras, se tendrá una eficiencia máxima muy por debajo de los resultados obtenidos con escenarios donde los conjuntos de información son básicamente polares.
- Los mejores resultados serán obtenidos al trabajar con el vector de referencia perteneciente al conjunto de información más diferente con respecto al resto presentes y seleccionando la hipótesis más adecuada, según el objetivo del sistema en particular.

# Referencia Documental

## Referencia Documental

1. **BARABÁSI**, Albert-László; *Linked*; Plume; United States of America; 2003; P.p. 294.
2. **CHAKRABATI**, Soumen; *Mining the Web: Discovering Knowledge from Hypertext Data*; Morgan Kaufmann; United States of America; 2003; P.p. 345.
3. **DATE**, C. J.; *Introducción a los Sistemas de Bases de Datos*; Traductor: I.Q. Sergio Luis María Ruiz Faudón; 7° Edición; Prentice Hall Pearson Education; México; 2001; P.p. 936.
4. **DE KUNDER**, **MAURICE**; Geschatte grootte van het geïnderxeerde World Wide Web; Tilburg University; Duchland; 2008; P.p. 63.
5. **DREYFUS**, Stuart E., **LAW**, Averill M.; *The Art and Theory of Dynamic Programming*; Academic Press; United States of America; 1977; P.p. 284.
6. **GRAMA**, Ananth, **GUPTA**, Anshul, **KARYPIS**, George, **KUMAR**, Vipin; *Introduction to Parallel Computing*; 2nd Edition; Pearson Addison Wesley; England; 2003; P.p. 636.
7. **HEATON**, Jeff; *Introduction to Neural Networks with Java*; 2nd Edition; Heaton Research Inc.; United States of America; 2009; P.p. 442.
8. **LANGVILLE**, Amy N., **MEYER**, Carl D.; *Google's PageRank and Beyond: The Science of Search Engine Rankings*; Princeton University Press; United Kingdom; 2006; P.p. 224.
9. **LIU**, Bing; *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*; University of Illinois at Chicago; Springer; United States of America; 2007; P.p. 532.
10. **LOTON**, Tony; *Web Content Mining with Java: Techniques for Exploiting the Wolrd´s Biggest Information Resource*; Wiley; England; 2002; P.p. 305.
11. **LUEGE**, Fernando; *Method and system of classifying, ranking and relating information based on networks*; USPTO Patent Application Number US 20080249966A1; October 9, 2008; P.p. 11.
12. **SCHRENK**, Michael; *Webrobots, Spiders and Screen Scrapers: A Guide to Developing Internet Agents with PHP/CURL*; No Starch Press; United States of America; 2007; P.p. 306.
13. **WASSWERMAN**, Stanley, **FAUST**, Katherine; *Social Network Analysis: Methods and Applications*; Serie: Structural Analysis in the Social Sciences; 15th Edition; Cambridge; United States of America; 2007; P.p. 825.
14. **WATTS**, Duncan J.; *Small Worlds: The Dynamics of Networks between Order and Randomness*; Princeton University Press; United Kingdom; 1999; P.p. 262.

15. http://en.wikipedia.org/wiki/Chi_distribution
16. http://en.wikipedia.org/wiki/Packet_(information_technology)
17. http://en.wikipedia.org/wiki/Rayleigh_distribution
18. http://en.wikipedia.org/wiki/TCP
19. http://en.wikipedia.org/wiki/URL
20. http://en.wikipedia.org/wiki/Wiki_software
21. http://en.wikipedia.org/wiki/World_Wide_Web
22. http://en.wikipedia.org/wiki/Zipf%27s_law
23. http://ilk.uvt.nl/events/dekunder.html
24. http://internet-filter-review.toptenreviews.com/internet-pornography-statistics.html
25. http://mathdl.maa.org/mathDL/1//?pa=content&sa=viewDocument&nodeId=1310&bodyId=1452
26. http://people.hofstra.edu/geotrans/eng/ch1en/meth1en/ch1m2en_2ed.html
27. http://people.hofstra.edu/geotrans/eng/ch1en/meth1en/ch1m3en_2ed.html
28. http://www.centiq.co.uk/products/bwa.html
29. http://www.datacenterknowledge.com/archives/2008/03/27/google-data-center-faq/
30. http://www.daviddarling.info/encyclopedia/B/Bridges_of_Konigsberg.html
31. http://www.daviddarling.info/encyclopedia/T/traveling_salesman_problem.html
32. http://www.forbes.com/2001/05/25/0524porn.html
33. http://www.infoteca.org/
34. http://www.infovis.net/imagenes/T1_N137_A6_KonigsGraph.gif
35. http://www.thefreedictionary.com/pornography
36. http://www.worldwidewebsize.com/