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.[1] 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%[2] 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%[3] 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.

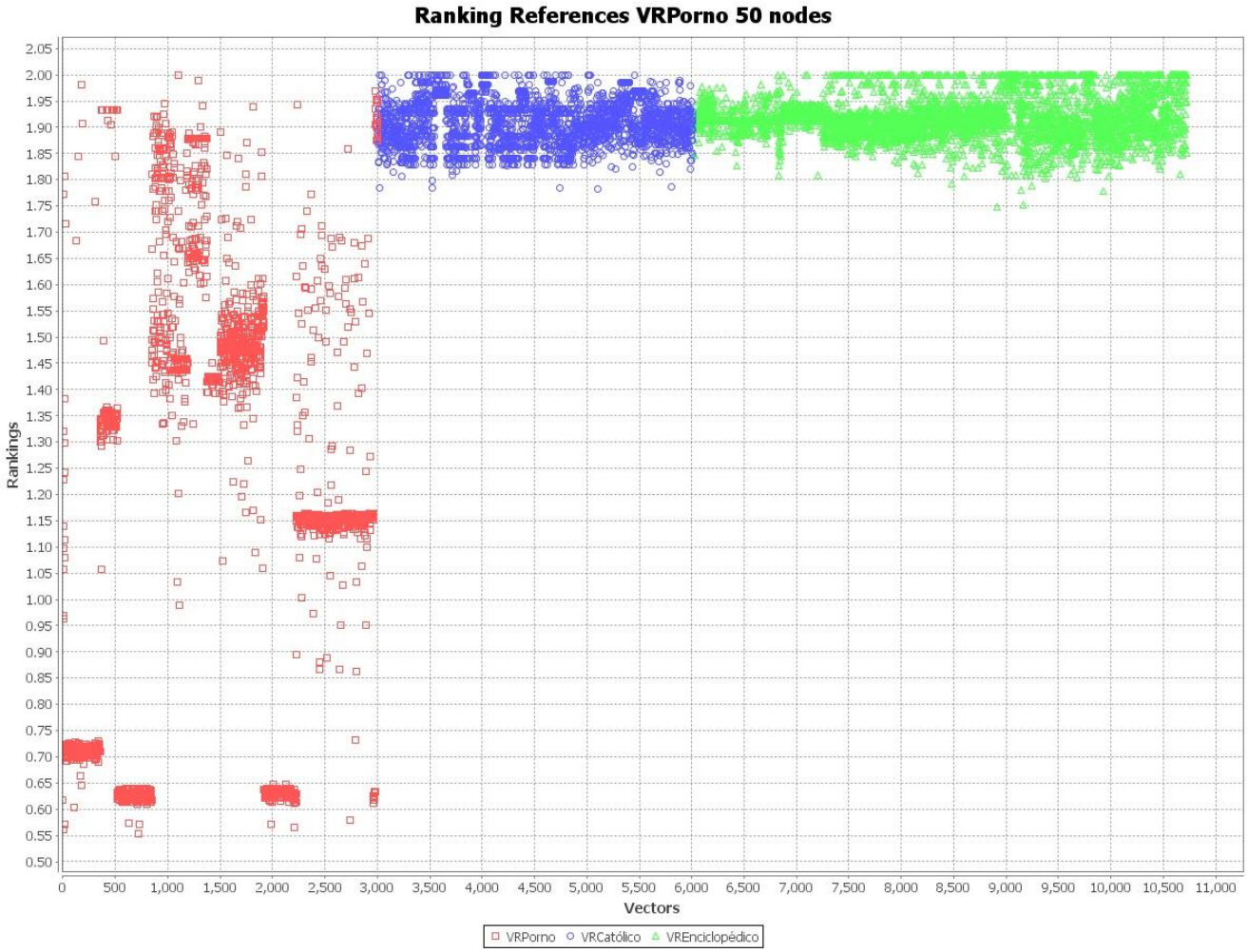
nivel de indexación para sitios pertenecientes al dominio *unam.mx* será de 3, y sólo serán analizados dominios bajo *unam.mx*.

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.

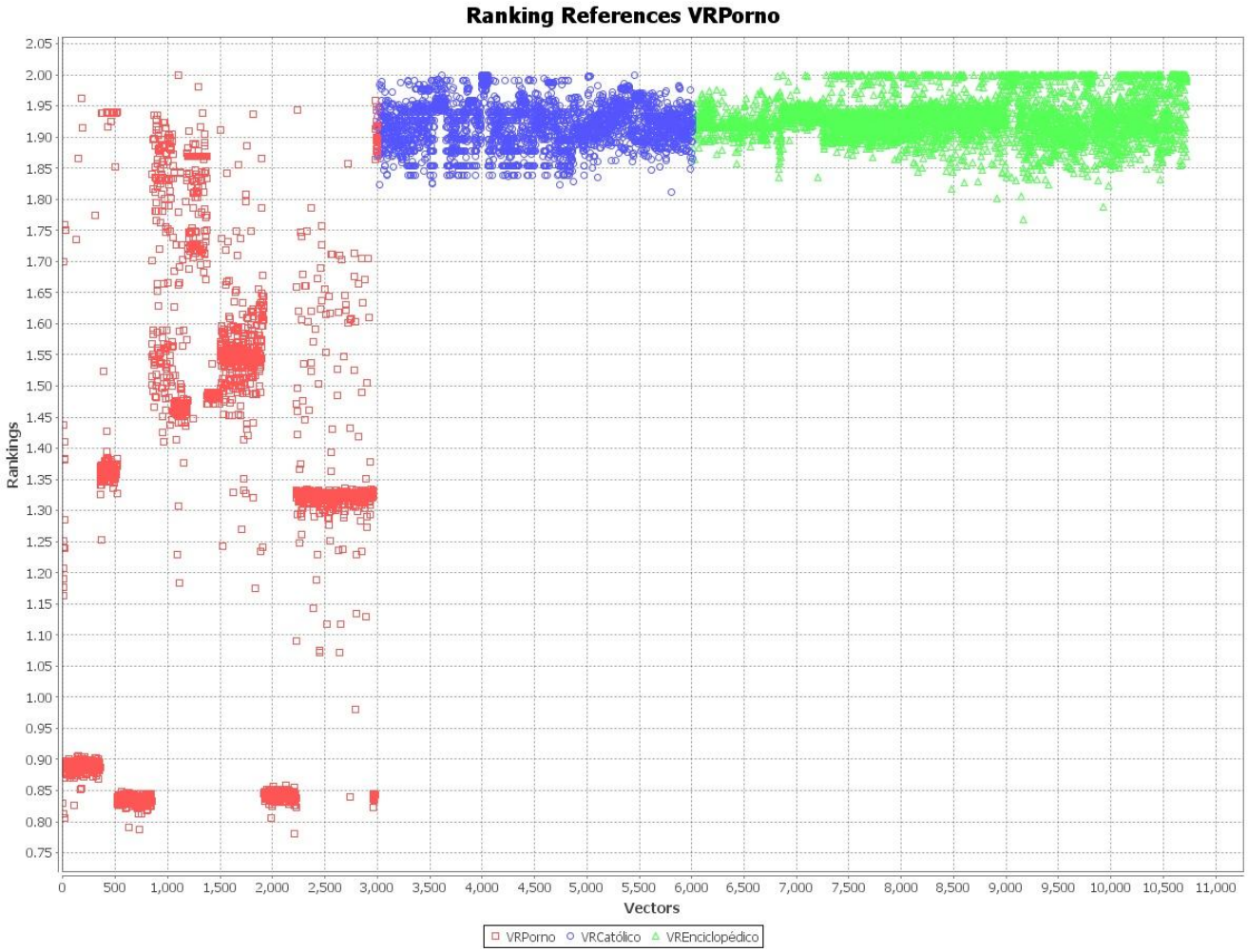
Tabla 4.3.1. – Distribución de valores máximos, mínimos y medios para cada entorno de información con respecto al vector de referencia pornográfico.

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

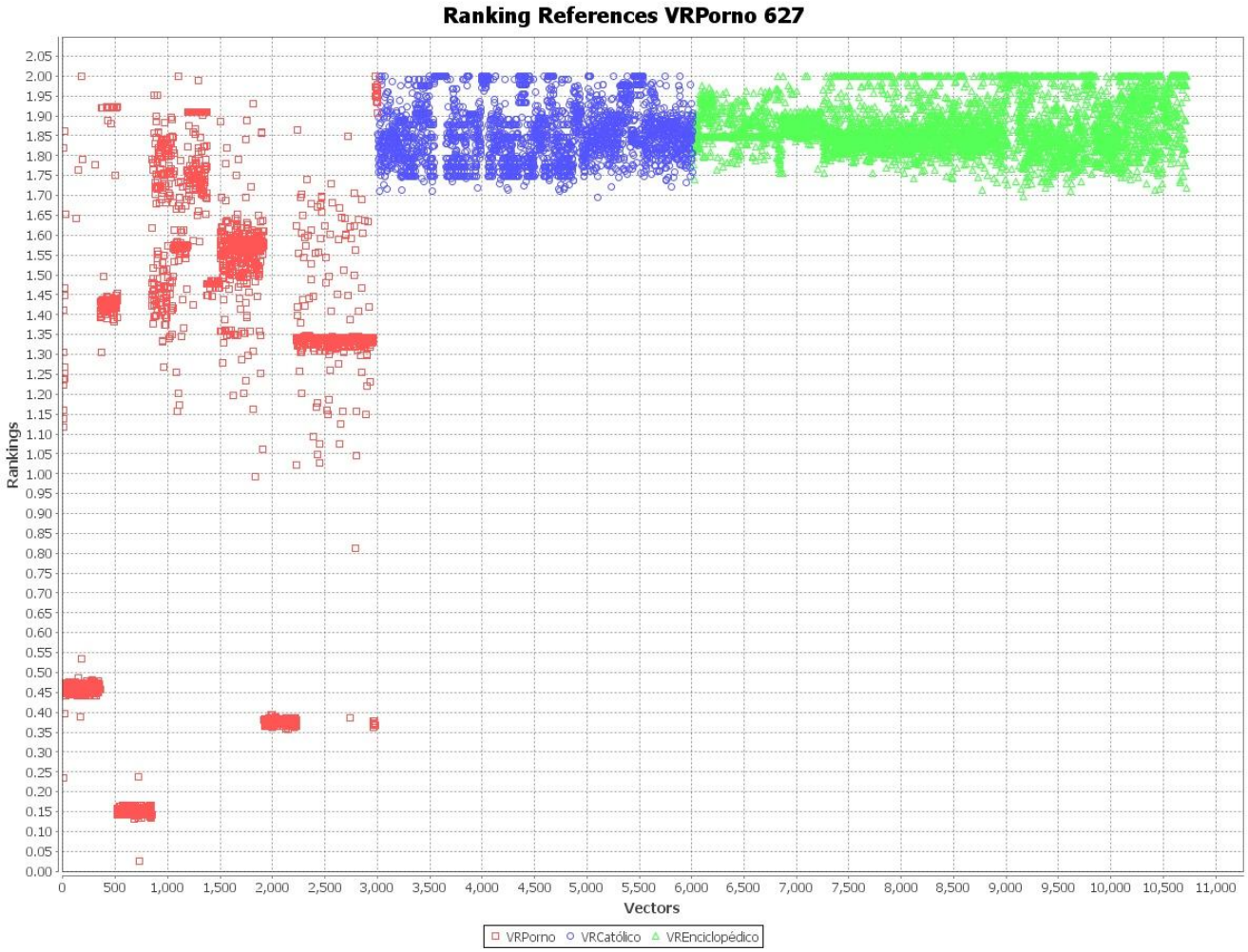
La red ActiveRank estudiada en la presente tesis está conformada por vectores con 50 componentes cada uno; al sumar los diferentes vectores para crear el de referencia, se debe de truncar su longitud a la misma cantidad de nodos que el resto de la red. En la gráfica 4.3.2. podemos observar la fila de 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.

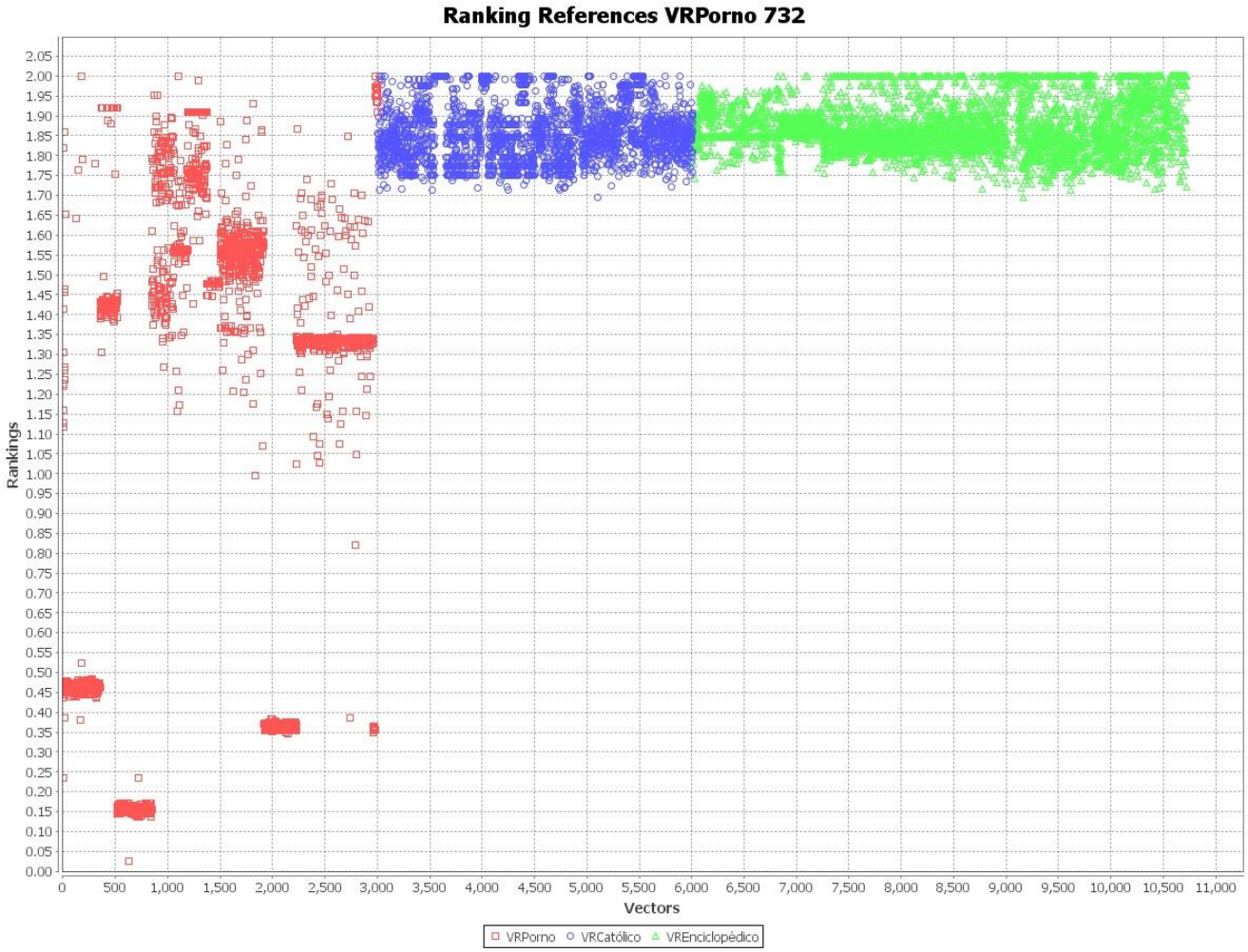
El análisis de la eficiencia del proceso de clasificación automático de información se encuentra más adelante en el capítulo 5 de la presente tesis, y no en la sección de resultados, con la intención de facilitar su comprensión.



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.

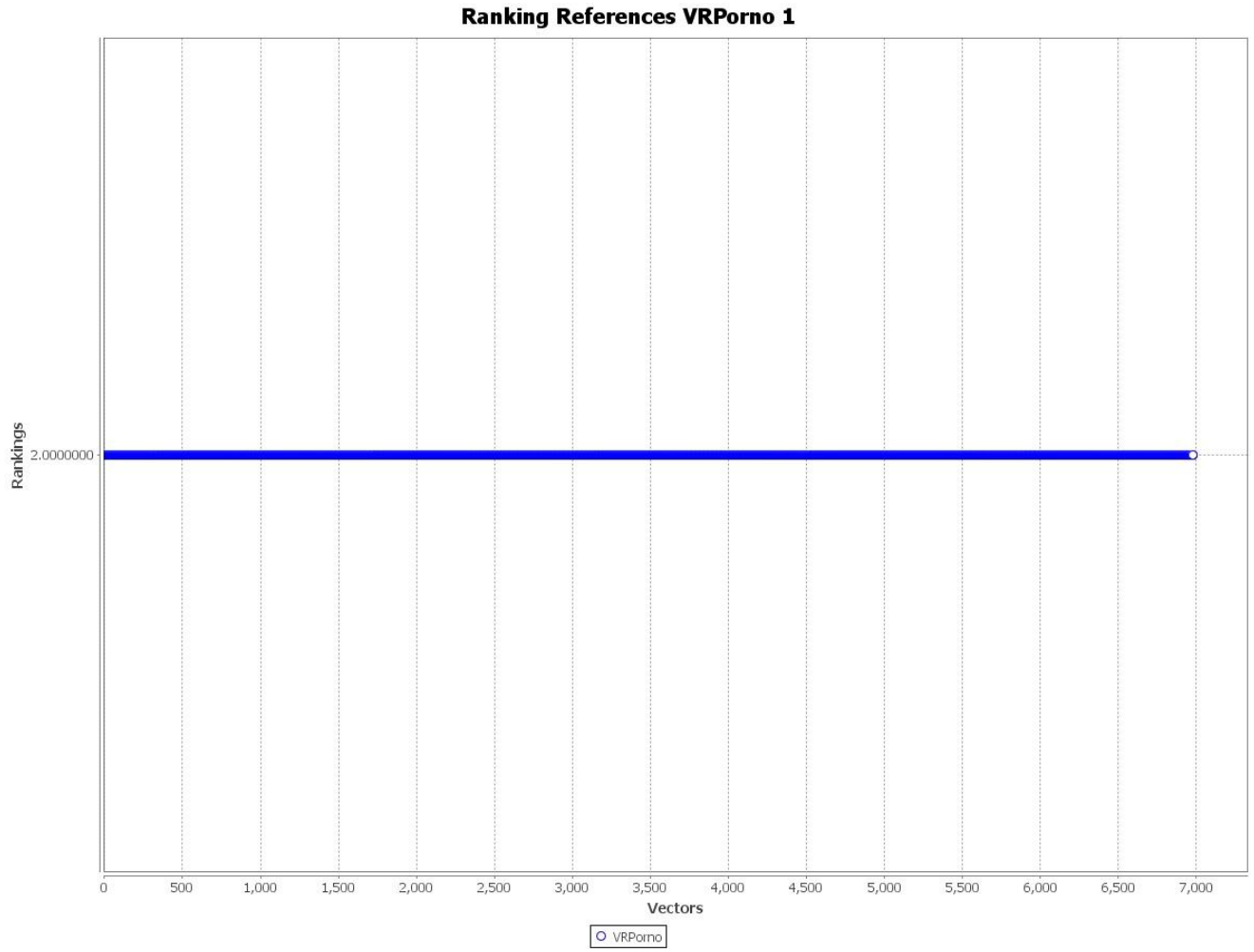


Gráfica 4.3.4. - Gráfica de dispersión de la fila de la matriz de ranking correspondiente al vector 732 del conjunto pornográfico contra todos los demás.

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

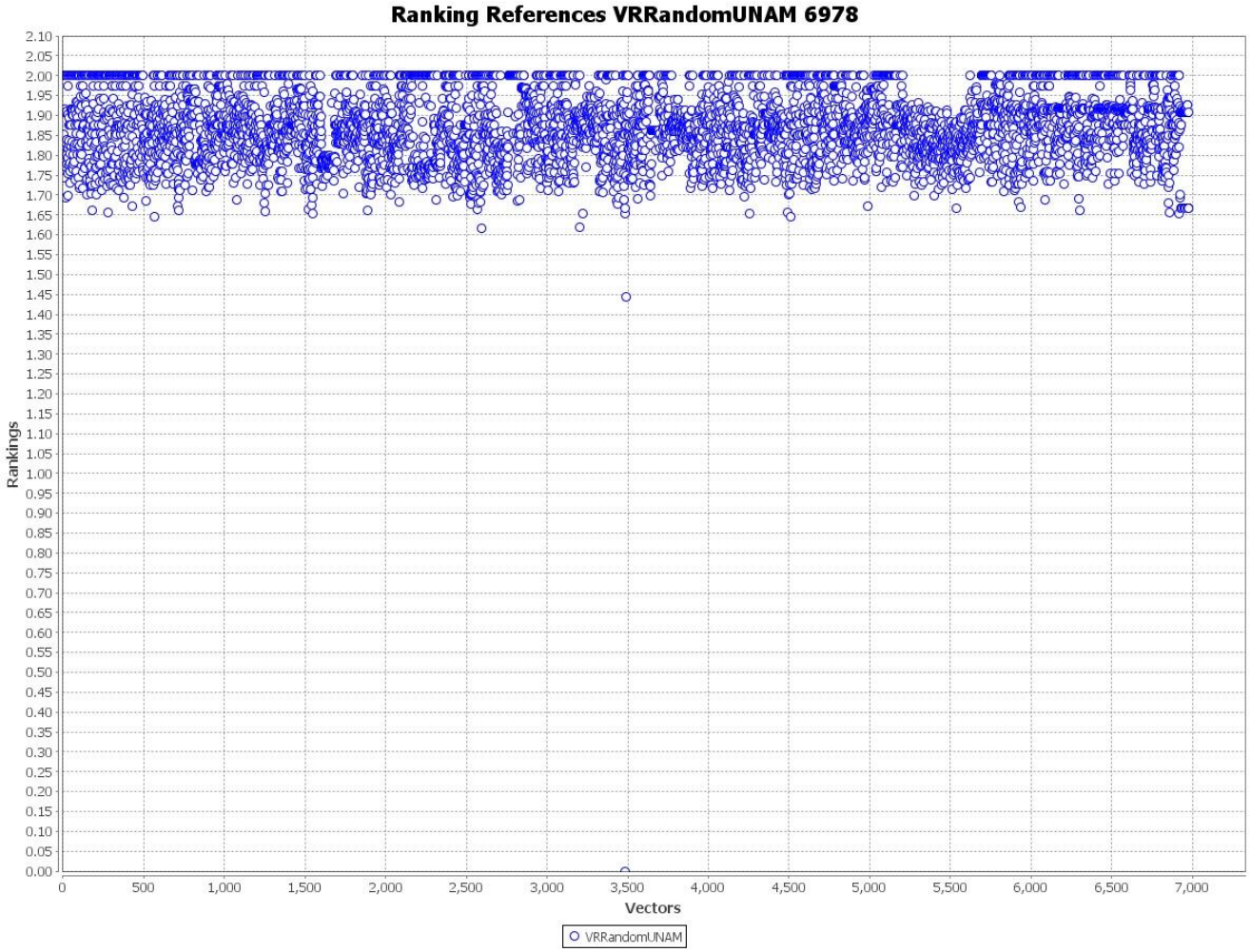
Como podemos observar, el ranking del vector de referencia pornográfico contra cualquiera de los elementos pertenecientes al conjunto de páginas bajo el dominio *unam.mx* es *2*, lo que es consecuencia de que el vector de referencia no comparte ningún nodo con cualquiera de los elementos del grupo; lo 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.

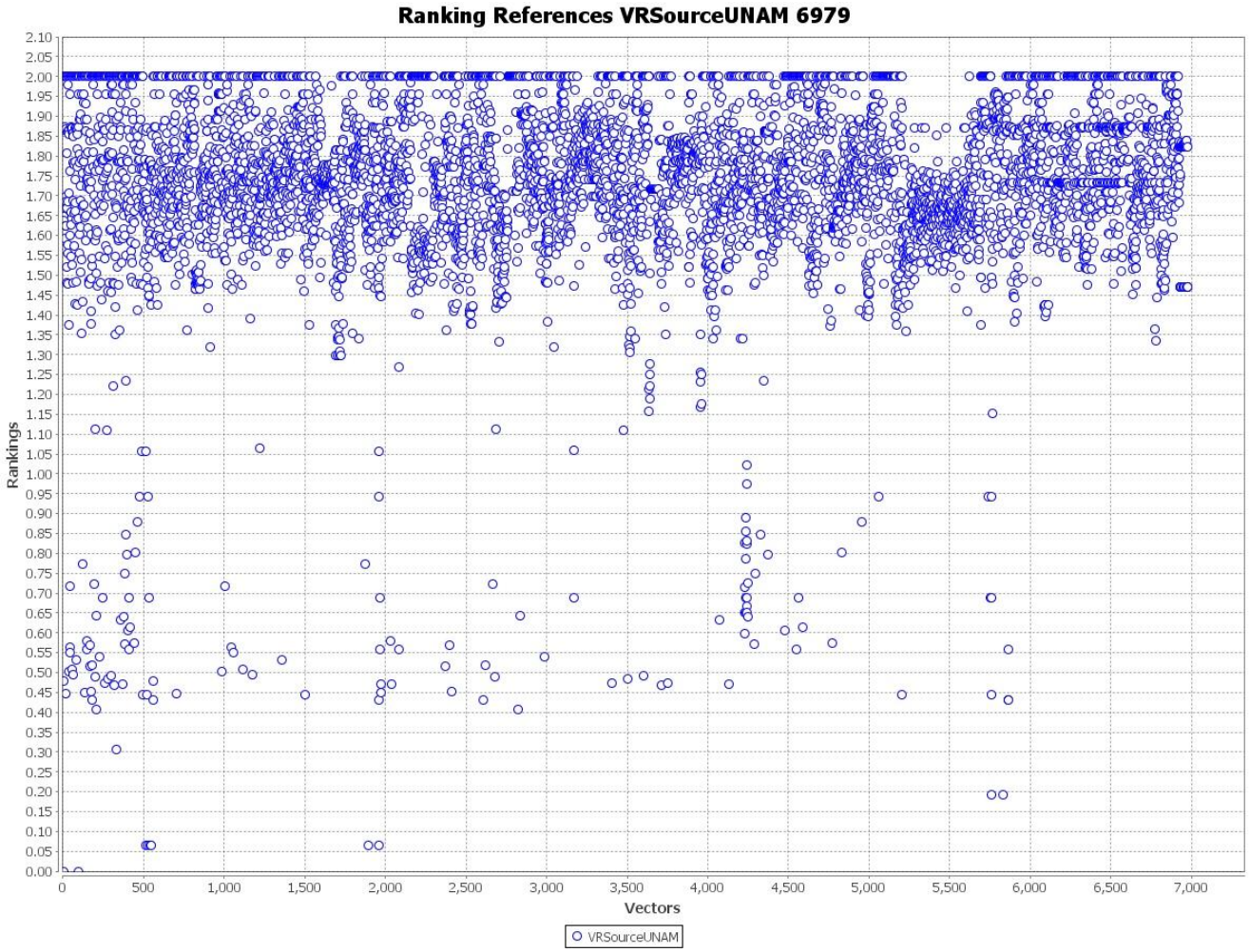
Con la intención de comprobar el buen funcionamiento del sistema en el cálculo de rankings y profundizar un poco más en el comportamiento y características propias del algoritmo ActiveRank, se generaron algunas pruebas adicionales que se muestran a continuación.

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

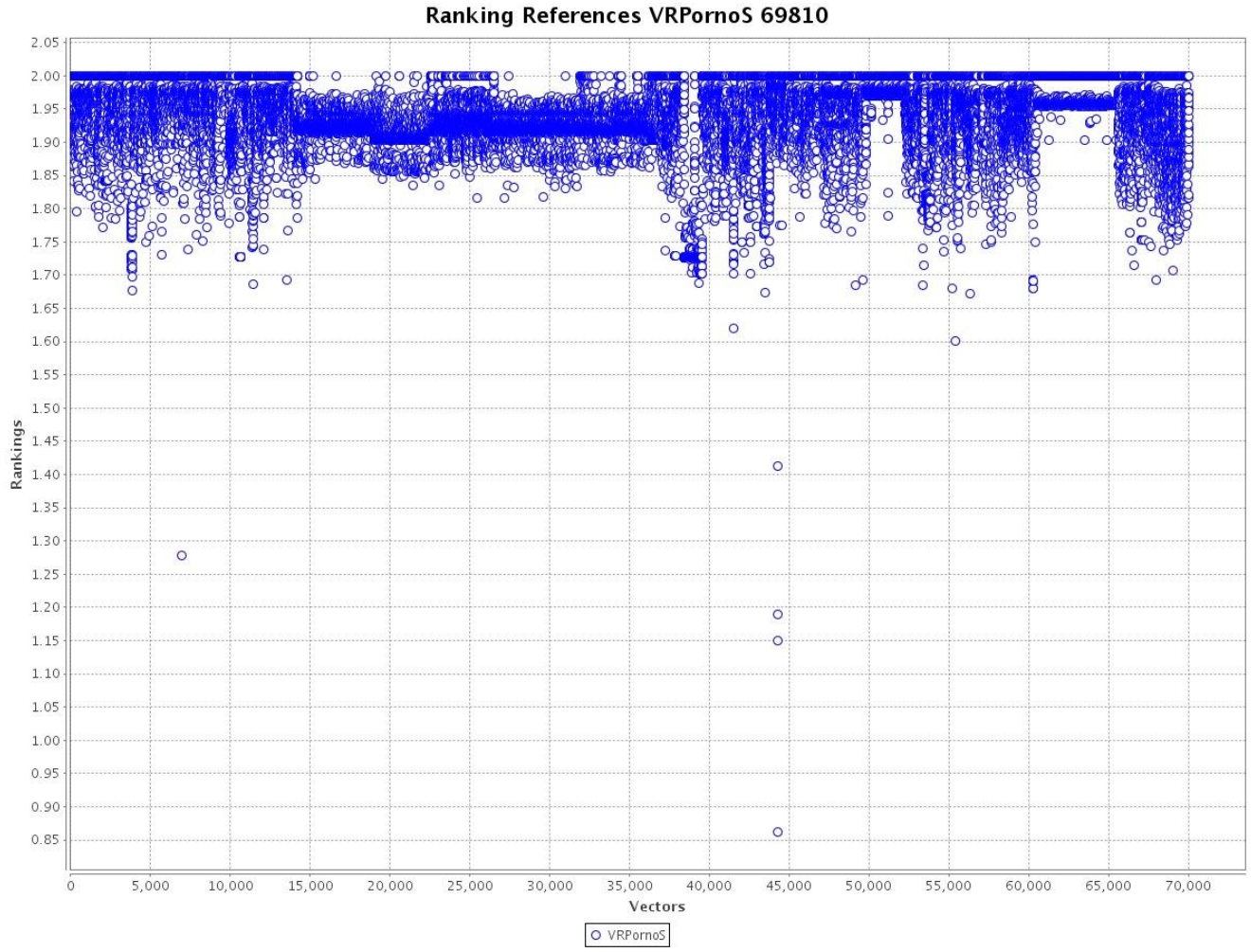
español al vector existente, o ingresar estas mismas palabras como un vector independiente y realizar la clasificación a través de una combinación lineal de los valores de ranking obtenidos para cada documento con respecto a estos dos vectores de referencia; este último formato fue el adoptado para las 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:

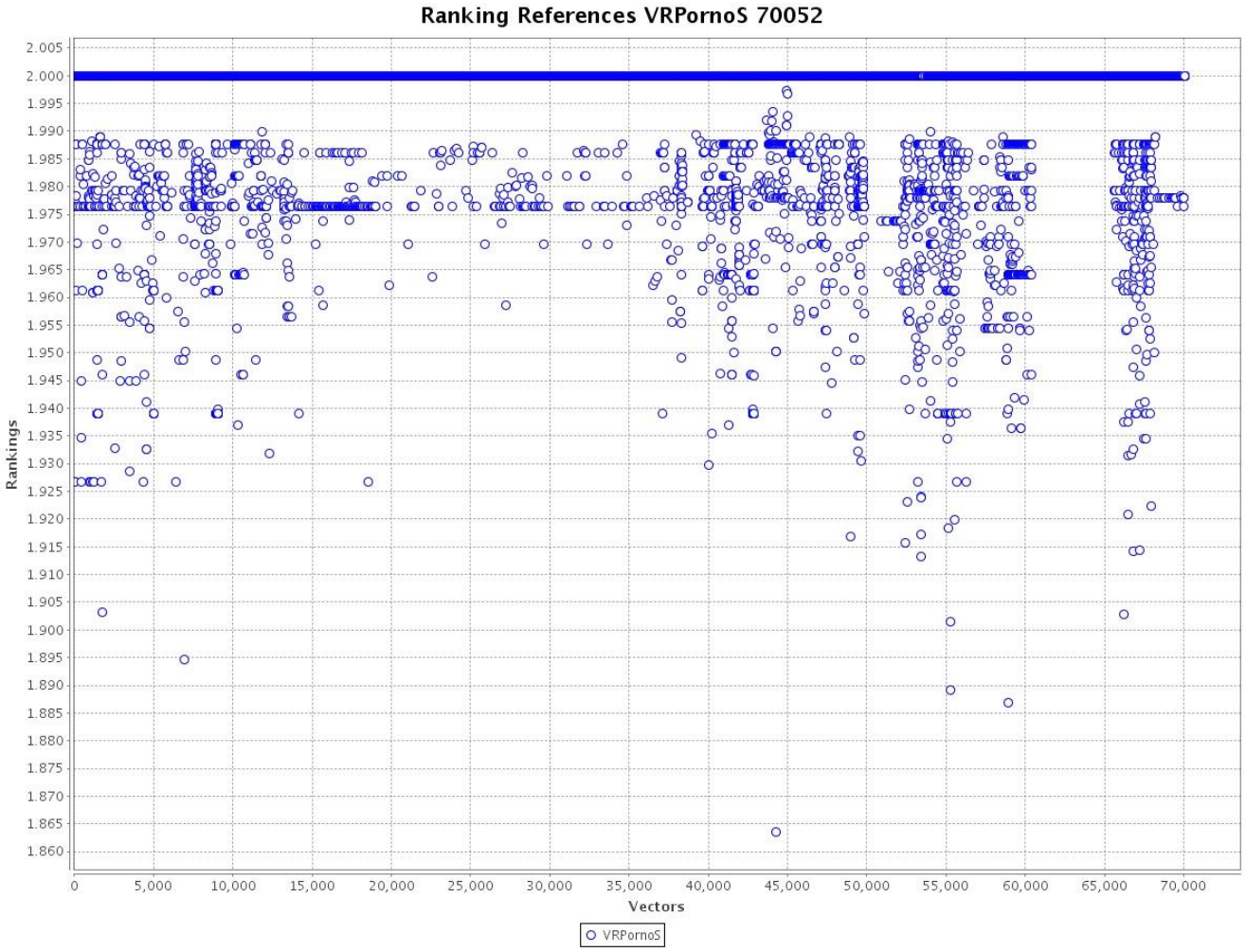
a) *http://www.pornomexico.net/*
b) *http://www.macizorras.com/*
c) *http://www.viendosexo.com/*
d) *http://www.iberporno.com/*

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.