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.