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 $a_x = (a_x{}^0, \ldots, a_x{}^{M-1})$, 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.

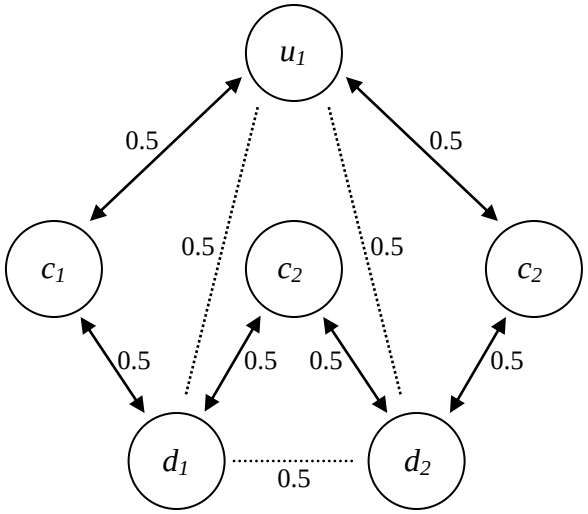

Figura 3.2.1. – Gráfica *G* en su estado inicial con medidas de ranking entre los elementos en línea punteada

	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.

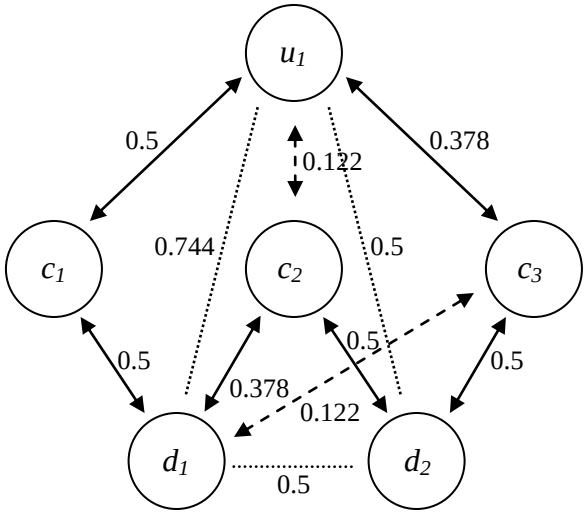

Figura 3.2.2. – Gráfica *G* después de 5 interacciones entre u_1 y d_1 con medidas de ranking entre los elementos en línea punteada y nuevas aristas generadas por la interacción

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.

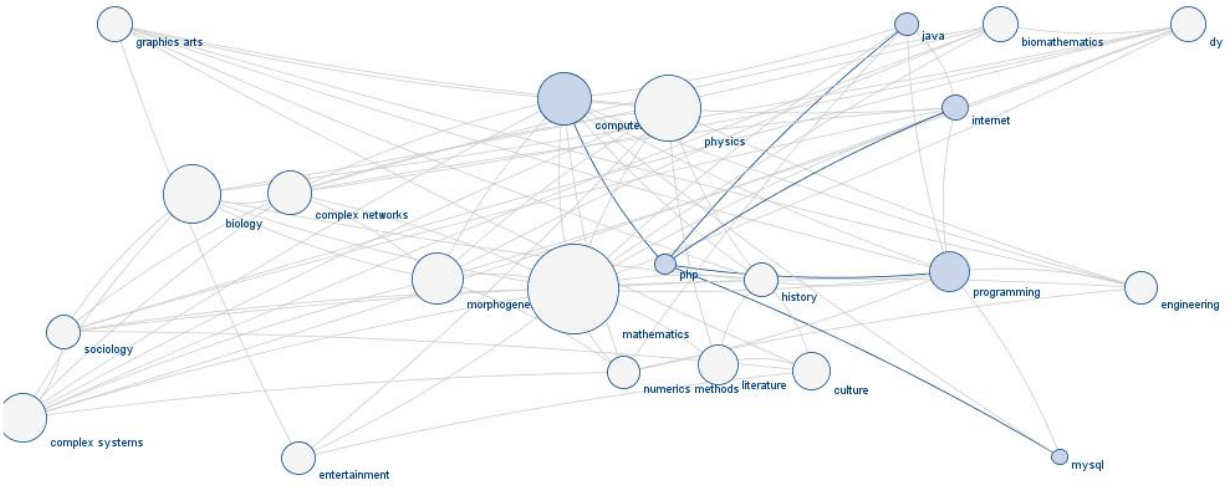

Figura 3.2.3 – Red de categorías en Infoteca.org con diámetros proporcionales a su relevancia para el usuario – 2005 *[Luege/Peña]*

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.[1]; 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.

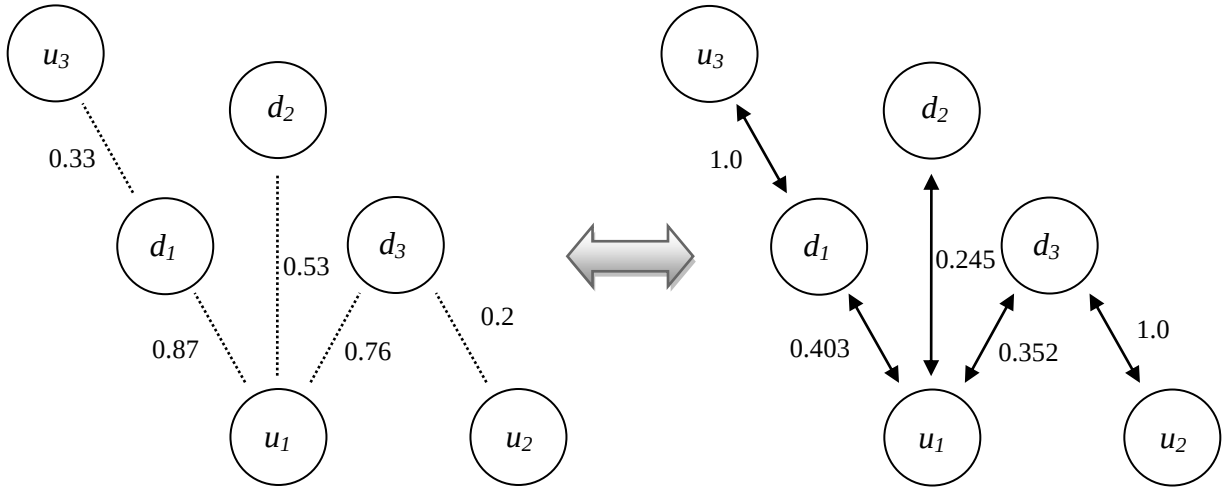

Figura 3.2.3. – Evolución de una gráfica *A* al pasar de rankings a vectores ActiveRank de su conjunto de usuarios.

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.

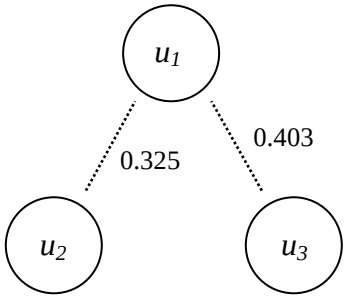

Figura 3.2.4. – Subgráfica de usuarios de la gráfica *A*.

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.

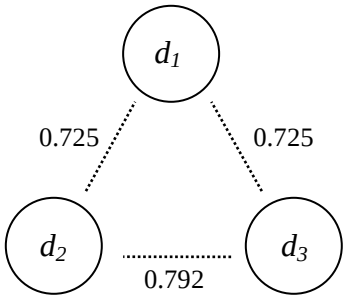

Figura 3.2.5. – Subgráfica de documentos de la gráfica *A*, donde casualmente el ranking entre los elementos $d_1 - d_2$ y $d_1 - d_3$ tienen el mismo valor.

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[2], 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]*

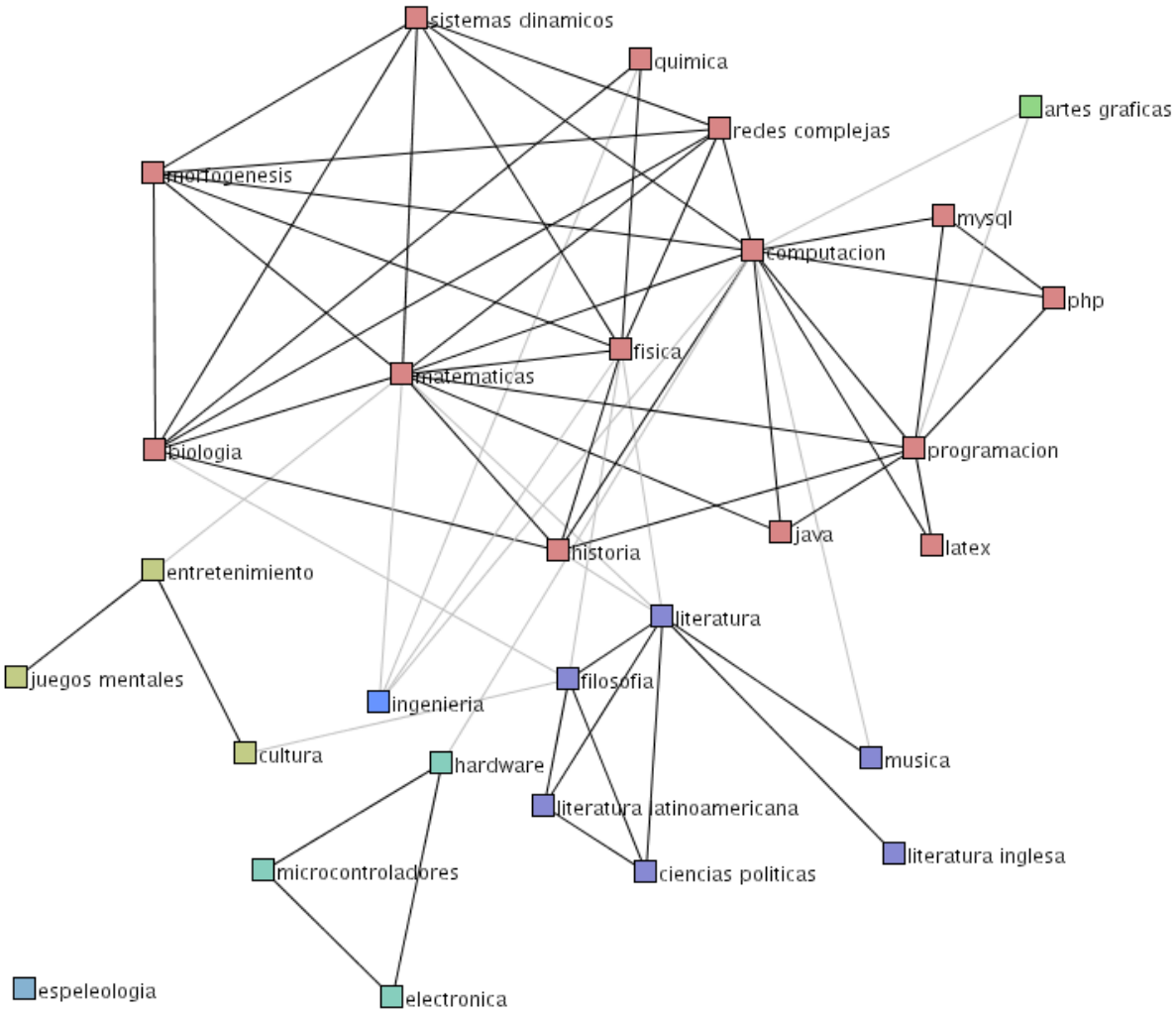

Figura 3.2.7. – Red de categorías agrupadas en Infoteca.org – 2005 *[Luege/Peña]*

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.