Un algoritmo genético para la conformación de grupos de individuos distantes en redes sociales

Eduardo Zamudio[1,2], Luis Berdún[1,2], y Analía Amandi[1,2]

[1] ISISTAN *Research Institute* - Fac. de Ciencias Exactas - UNCPBA - Tandil, Argentina
[2] CONICET, Consejo Nacional de Investigaciones Científicas y Técnicas - Argentina
eduardo.zamudio@isistan.unicen.edu.ar, {lberdun, amandi}@exa.unicen.edu.ar

Resumen La conformación de comisiones es una actividad en la que se requiere seleccionar un conjunto de candidatos a partir de una población de individuos. Estos individuos y sus relaciones pueden representarse mediante una red social. Trabajos previos relacionados con el análisis de redes sociales se enfocan en la identificación de la importancia de los nodos que conforman esas redes, o bien en las propiedades estructurales de las mismas. Este trabajo propone una alternativa para la selección de un conjunto de individuos que presenten menor relación entre sí a partir de una red social, considerando esto como un criterio de independencia entre los mismos. El enfoque propuesto es evaluado con un caso de estudio en el cual se propone en primer lugar, la construcción de una red social en base a datos de coautoría de publicaciones de investigadores y mismo lugar de trabajo, la cual es analizada para determinar las distancias entre cada par de individuos de la red, y en segundo lugar, la optimización de las distancias entre los posibles candidatos mediante un algoritmo genético.
Palabras clave: algoritmo genético, red social, selección, candidato, grupo

1. Introducción

Al momento de conformar una comisión, una característica deseable es la independencia de los miembros candidatos. Cómo seleccionar estos candidatos en base a criterios objetivos puede resultar complejo, ya sea en la definición misma de estos criterios, como en el análisis de la población.

La aplicación del algoritmo genético (AG) sobre una red social se presenta como alternativa al problema de la selección de un conjunto de candidatos sobre una población en la que no interesa el orden de éstos dentro de la comisión, aunque sí interesa que no se repitan los candidatos. Este problema, que en una aproximación incial podría representarse como una combinación, tendría como consecuencia un espacio de búsqueda que fácilmente podría resultar en dimensiones no adecuadas para su tratamiento computacional (mucho menos manual). Esto último debido a que una combinatoria de los individuos de la población tomados en grupos podría generar resultados exponenciales de soluciones posibles, a las cuales se les debería aplicar previamente una función de distancia

entre los miembros candidatos, con el objetivo de generar un ranking de soluciones. Como alternativa a este enfoque (que implicaría una revisión exhaustiva de todas las soluciones posibles al problema) podría plantearse la generación aleatoria de n soluciones posibles a las que se le aplicara la función de distancia, y posteriormente se rankearían los resultados. Nuevamente, este enfoque no sería adecuado debido a que presentaría la subjetividad de la elección aleatoria de los miembros de la comisión, quedando ésta determinada por la probabilidad de cada candidato de ser elegido para la conformación de la comisión, y la probabilidad conjunta de los miembros de la comisión. Esta probabilidad solo expresa la posibilidad de que un candidato sea elegido en la comisión. Si consideramos cuál es la probabilidad de conformar la mejor comisión, dicho valor es inversamente proporcional al número de alternativas.

Si la población es considerada como un conjunto de individuos que se encuentran relacionados entre sí, es posible determinar cuáles de estas relaciones pueden resultar relevantes para analizar la independencia entre ellos. Teniendo en cuenta estos elementos (individuos, y relaciones entre individuos) podemos representar la población como una red social, y en consecuencia aplicar técnicas de análisis sobre la misma que permitan seleccionar un conjunto de candidatos en base a criterios establecidos.

En este sentido, la construcción de una red social requiere un conjunto de datos mediante los cuales se puedan representar los actores, las relaciones entre los actores, el tipo de red, y el objeto de análisis.

En [9] se presenta una clasificación de las notaciones utilizadas para representar redes sociales. Estas representaciones se clasifican en: grafos, sociomatrices (matrices de adyacencia), o representaciones algebraicas. La elección de la notación se encuentra relacionada con el tipo y las características de la red que se pretende representar, así como también por los medios para analizarla.

La notación usualmente preferida para la representación y análisis por computadora de redes sociales son las sociomatrices, debido a que éstas utilizan matrices para representar las relaciones entre cada par de nodos.

Las técnicas actuales de análisis de redes sociales concentran su esfuerzo en la identificación de la cantidad o valor de sus relaciones, el rol de un nodo en la red, o la importancia de dicho nodo en la estructura de la red.

El objetivo de este trabajo consiste en seleccionar un conjunto de individuos representativos de una población con menor relación entre sí. Para ello, se propone primero, la representación de esta población como una red social para determinar las relaciones entre estos individuos, y luego la aplicación de un AG con el objetivo de determinar el grupo de n candidatos más distantes para la conformación de una comisión.

El trabajo está organizado de la siguiente manera. La Sección 2 describe la construcción de la red social y sus características. La Sección 3 describe la implementación del AG. La Sección 4 presenta las condiciones de ejecución del AG y sus resultados. La Sección 5 presenta los trabajos relacionados. Finalmente, la Sección 6 presenta las conclusiones y trabajos futuros.

2. Contrucción de la red social

Este trabajo definió como objetivo la selección de un conjunto de candidatos de una población que presenten menor relación entre sí, para lo cual se optó por la construcción de una red social que permita analizar las relaciones entre los individuos de la población.

Una red social se compone de individuos (representados como actores) y enlaces (entre los actores), donde generalmente interesa analizar las relaciones entre individuos y los grupos que conforman [9].

En particular, nos interesa la construcción de la red social por la capacidad de representar criterios de análisis a partir de sus relaciones. Para clarificar este concepto se propone a modo de ejemplo, la construcción de una red de investigadores relacionados mediante la coautoría de publicaciones, y su pertenencia al mismo lugar de trabajo. En esta red, los investigadores representan los candidatos para la conformación de comisiones, y las relaciones representan los criterios de análisis para determinar la distancia entre cada par de investigadores en la red.

Como se mencionó anteriormente, las relaciones entre los actores determinan lo que puede analizarse de la red. En este caso, el objetivo del análisis es determinar la distancia entre un grupo de actores. Para ello, se propone representar las relaciones como sociomatrices, las cuales establecen en forma binaria las relaciones entre actores (por ejemplo, si dos investigadores han publicado juntos, o bien, si trabajan en el mismo lugar). La Fig. 1 muestra la relación *publicó con* de un conjunto de 5 investigadores (A,B,C,D,E), mediante un grafo y una sociomatriz, en la que se representan relaciones de tipo binaria (está presente o no), no dirigidas (no interesa el sentido de la relación), y no reflexivas (un investigador no publica consigo mismo). Estas mismas características están presentes en la relación que representa si dos investigadores trabajan en un mismo lugar.

Una vez definidas las sociomatrices que corresponden a las relaciones entre los actores, se propone la construcción de una sociomatriz que incorpore los criterios definidos, es decir, todos los tipos de relaciones definidas. Este procedimiento consiste en realizar un solapamiento mediante la aplicación del operador lógico *OR* para las celdas de posiciones iguales de las matrices de criterios en

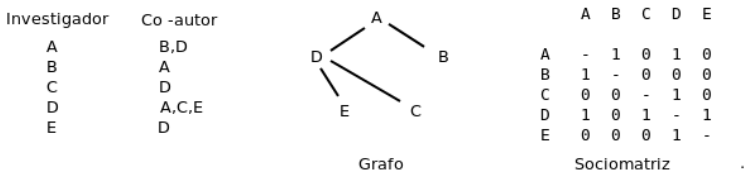

Figura 1. Entradas de la relación *publicó con* representada mediante un grafo y una sociomatriz

una matriz de criterios acumulados. Suponiendo que se tienen dos matrices de criterios $C1$y $C2$, las celdas de estas matrices están representadas por su posición en la fila i y la columna j, y la matriz de criterios acumulados (CA) está dada por $CA = C1\,OR\,C2$, donde cada celda $ca_{i,j}$ perteneciente a CA está dada por $ca_{i,j} = c1_{i,j}\,OR\,c2_{i,j}$ para $\forall i, j \in S$ donde $i \neq j$, y S representa el conjunto de investigadores. La Fig. 2 presenta la aplicación del operador lógico sobre las matrices de criterios del ejemplo y en consecuencia la generación de la matriz de criterios acumulados.

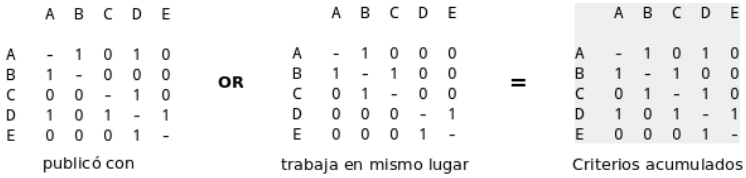

Figura 2. Generación de la sociomatriz de criterios acumulados a partir de sociomatrices de relaciones entre investigadores mediante la aplicación del operador lógico OR.

Este trabajo propone que la menor relación entre cada par de individuos puede obtenerse alternativamente en base a la distancia entre éstos. En consecuencia es necesario determinar la distancia entre cada par de nodos en la red, para lo cual se utilizaron las métricas de camino más corto (o comunicación geodésica) [2], y distancia (longitud del camino más corto) sobre la matriz de criterios acumulados.

Un requisito previo a la determinación de las distancias es que debe garantizarse que la red esté conectada, es decir, que cada nodo de la red sea alcanzable desde cualquier otro nodo de la red. Esto puede determinarse mediante una matriz de alcanzabilidad, la cual se obtiene a partir del producto de matrices [9].

La distancia entre cada par de nodos se representa mediante una matriz de distancia, generada a partir de la potenciación [9] de la matriz de criterios acumulados. Esta matriz de distancia contiene los datos de entrada de un algoritmo cuyo objetivo es conformar una comisión en la cual los integrantes se encuentren minimamente relacionados. Particularmente se trabajó con un AG, en el cual la entrada es una función que pretende ser optimizada para determinar el grupo de individuos que presenten menor relación entre sí, o lo que es lo mismo, mayor distancia entre sus miembros.

3. Algoritmo genético

Un AG es un tipo de algoritmo evolutivo que puede ser considerado como un método de optimización de funciones [8]. Si bien no existe un AG definitivo, es posible adaptar uno a partir de un conjunto de representaciones y operadores que pueden ser adecuados a las necesidades particulares de una aplicación. El elemento con el que trabaja el AG es el cromosoma, el cual contiene la información genética representada por la disposición y valor de sus genes.

Con objeto de seleccionar un conjunto de individuos de la red social, se ha definido una función de relación ad-hoc, la cual establece la suma acumulada de las distancias entre todos los candidatos de la comisión conformada. Consecuentemente se ha diseñado un AG, con el fin de obtener soluciones aproximadas a una mejor solución mediante la maximización de dicha función aplicada a un conjunto de n candidatos.

El AG presenta los siguientes componentes: representación, definición de la función de fitness, mecanismo de selección de padres, operadores de cruce y mutación, y selección de supervivientes. A continuación se presenta cada uno de estos componentes.

Representación

La representación del problema requiere definir cómo se compone un cromosoma. Este trabajo se basó en permutaciones de un vector de enteros (cromosoma) donde cada elemento del mismo representa una referencia a un investigador (gen), y en cuyo vector están incluidos todos los investigadores de la población. Es decir, un cromosoma tiene tantos genes como investigadores contenga la población. La inclusión o exclusión de los candidatos en la comisión está dada por un vector de igual tamaño al vector de investigadores, haciendo corresponder las posiciones de ambos vectores para indicar la conformación de comisiones. De esta manera, el conjunto de candidatos propuestos para la conformación de una comisión presenta ocurrencias únicas de los candidatos, es

decir, un candidato no puede repetirse en una misma comisión. El orden en que se presentan dichos candidatos dentro de la comisión no tiene relevancia para el problema. La Fig. 3 presenta en forma gráfica dicha representación en vectores de longitud n.

El problema requiere determinar el tamaño de las comisiones (el número de candidatos que conformarán la comisión), de modo que el tamaño de la comisión debe ser establecido de antemano, para lo cual es definido el *vector de comisión*.

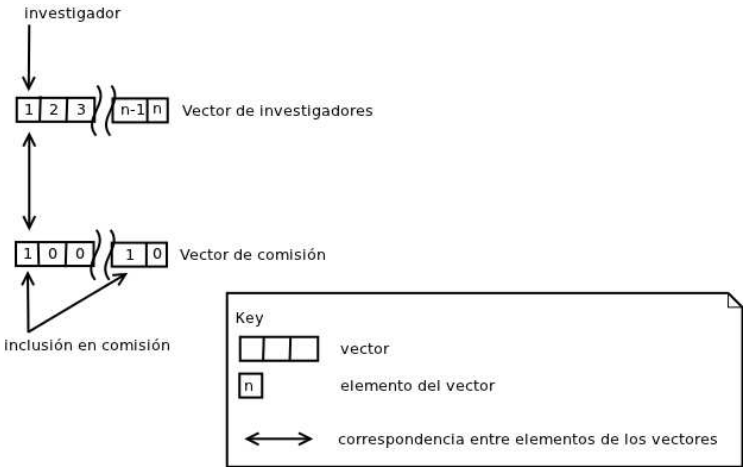

Figura 3. Representación del algoritmo genético mediante un vector de investigadores, que contiene todos los investigadores de la población, y un vector de comisión, que indica los elementos del vector de investigadores incluidos en la comisión.

La codificación de la permutación queda representada mediante el elemento i, el cual denota el evento que sucede en ese lugar en la secuencia. Por ejemplo, para cuatro investigadores $[I1, I2, I3, I4]$, la permutación $[2,1,4,3]$ queda representada por $[I2, I1, I4, I3]$.

Fitness

La función de fitness tiene por objetivo determinar el valor de una solución, el cual se busca optimizar de acuerdo al problema planteado. Para este trabajo, donde se requiere maximizar las distancias entre los investigadores del grupo candidato, se desarrolló una función de modo ad-hoc representando la suma

acumulada de las distancias entre cada par de candidatos dentro de una comisión (solución individual), quedando expresada dicha función de la siguiente manera:

$$f = \sum d(i, j)$$

$$\forall i, j \,/\, i \neq j \text{ y } i, j \in S$$

Donde d es la función de distancia entre dos miembros, i y j son miembros de la comisión, y S representa el conjunto de investigadores.

Como se estableció previamente, es condición de la construcción de la red que ésta debe estar conectada, lo que implica que todos los miembros de cualquier comisión presentan una distancia entre sí distinta de ∞.

Selección de padres

La información genética de las nuevas generaciones se obtiene a partir de padres, los cuales son cromosomas (soluciones) de la generación inmediatamente anterior. Con este fin, es necesario definir una estrategia de selección de padres, para la cual se puede utilizar alguno de los mecanismos adecuados al tipo de problema que se pretende resolver.

Los mecanismos de selección utilizados en este trabajo incluyen los algoritmos: *Stochastic Universal Sampling* (*SUS*) debido a que deben seleccionarse varios padres de una población; y *Tournament*, ya que en ambos casos se pretende obtener resultados sin conocer el fitness global.

Cruce

La información genética de la nueva generación es determinada por la información genética de los padres de los cuales desciende. Este proceso de recombinación genética se realiza mendiante mecanismos de cruce. Por ejemplo, teniendo dos cromosomas indicando soluciones distintas, el cruce implica que la descendencia obtendrá información genética de ambas soluciones.

Con el objetivo de mantener la permutación válida, se utilizaron los operadores de recombinación para permutaciones *Partially Mapped Crossover* (*PMX*), y *Order Crossover* (*OX*). Ya que *PMX,* al ser un algoritmo diseñado para problemas de adyacencia, resulta aplicable al problema presentado. Por su parte, *OX* está diseñado para problemas de orden, sin embargo, la información del orden del segundo padre puede resultar beneficiosa en la generación de un nuevo cromosoma.

Mutación

Otro mecanismo de recombinación genética utilizado en este trabajo es la mutación, la cual implica alterar los genes de un cromosoma. En permutaciones, como es el caso del ejemplo de la representación de los investigadores, la mutación implica alterar la disposición de los valores presentes en el vector solución de la nueva generación.

Los operadores de mutación seleccionados para ser aplicados en la descendencia fueron *Swap Mutation* e *Insert Mutation*, ya que ambos son operadores aceptados para mantener la permutación válida.

Selección de supervivientes

La selección de supervivientes utilizó el mecanismo *Steady-state* con objeto de no perder los individuos con mejor fitness en las sucesivas generaciones.

4. Caso de estudio

Se ha diseñado un caso de estudio en el cual se presentan investigadores candidatos a conformar una comisión, con objeto de demostrar que dichos candidatos presentarán menor relación entre sí, lo que puede asumirse como un criterio de independencia en la actuación de estos investigadores dentro de una comisión.

Los datos utilizados para la experimentación fueron obtenidos de la base de datos publicada por The Auton Lab [1], parte del School of Computer Science de Carnegie Mellon University, la cual contiene información de coautoría de trabajos presentados en el Neural Information Processing conference (NIPS). Este conjunto de datos fue reducido a los primeros 1001 registros con objeto de facilitar su procesamiento, y cuya selección incluyó 720 investigadores, 443 publicaciones, la cual se enriqueció con 11 lugares de trabajo que fueron asignados en forma aleatoria a los investigadores.

Como se mencionó previamente, el enfoque de aplicación del AG para la determinación de los miembros de la comisión se presenta como una alternativa a otros métodos, como podría ser una combinatoria de 720 individuos tomados en comisiones de 5 integrantes, lo que resultaría en ${}_{720}C_5 = 1,590145128\times10^{12}$ soluciones posibles; o como la selección aleatoria de un conjunto de soluciones posibles donde la probabilidad de elección de cada candidato estaría dada por $P\left(\frac{1}{720}\right) = 0{,}1388889 * 10^{-2}$, y la probabilidad conjunta de la comisión de 5 candidatos estaría dada por $\sum_{n=0}^{4} P\left(\frac{1}{720-n}\right) = 0{,}2779712 * 10^{-2}$. La probabilidad de encontrar un óptimo asumiendo la existencia de m posibles conformaciones

de comisiones óptimas, es igual a $m/_{720}C_5$, lo que equivale a un valor próximo a 0.

Con objeto de evaluar la solución propuesta mediante el enfoque del AG, se planteó un caso experimental en el que se tuvieron en cuenta los siguientes aspectos:

- Parámetros de ejecución:
  - Tamaño de población: El número de soluciones candidatas en cualquier punto de tiempo se calculó mediante P/n, donde P representa el conjunto de todos los investigadores, y n el tamaño de las comisiones. En el caso de estudio se utilizó $P = 720$ y $n = 5$.
  - Probabilidad de cruce: Se seleccionó el valor 0,7, tomado del rango $[0{,}6;0{,}9]$.
  - Probabilidad de mutación: Se seleccionó el valor 0,15, tomado del rango $[0{,}01;0{,}15]$.
  - Condición de corte: Se establecieron 25 generaciones como límite de corte.
- Configuraciones: Se tomaron 8 configuraciones distintas, como resultado de las posibles combinaciones de los mecanismos adoptados en este trabajo (selección (2), operadores de cruce (2), y operadores de mutación (2)). Adicionalmente, se optó por el mecanismo de selección Steady-state. La Tabla 1 presenta estas configuraciones.
- Ejecuciones: Se realizaron 40 ejecuciones, correspondientes a 5 ejecuciones por cada configuración propuesta, luego se promediaron los resultados de las ejecuciones, y posteriormente se calculó la desviación estándar (σ) de las mismas.

Configuración	Selección		Cruce		Mutación	
	Tournament	*SUS*	*PMX*	*OX*	*Swap*	*Insert*
1	X		X		X	
2	X		X			X
3	X			X	X	
4	X			X		X
5		X	X		X	
6		X	X			X
7		X		X	X	
8		X		X		X

Tabla 1. Configuraciones del algoritmo genético

Resultados

La Tabla 2 presenta el ranking de las ejecuciones donde se indica para cada configuración: el fitness de las ejecuciones, el fitness promedio, y la desviación estandar. El fitness logrado en las ejecuciones se encuentra en el intervalo [11;20], a partir de lo cual se plantea la hipótesis de que el fitness global se aproxima a 20 para este caso de estudio. Con objeto de corroborar dicha hipótesis, se amplió el número de generaciones a 50, 100, 250, 500, y 1000. La Fig.4 presenta el fitness promedio de 5 ejecuciones para las 8 configuraciones de acuerdo a la cantidad de generaciones indicada, donde se evidencia que con el incremento de generaciones en la ejecución del algoritmo, el fitness promedio se aproxima a 20.

Configuración	Distancia de acuerdo al Nº de ejecución					Distancia promedio[3]	σ
	1	2	3	4	5		
2	17	12	18	12	**20**	15,8	3,6332
5	12	11	19	12	19	14,6	4,0373
3	12	18	12	12	12	13,2	2,6833
7	12	12	18	12	12	13,2	2,6833
4	12	12	11	12	18	13,0	2,8284
6	12	**20**	10	12	11	13,0	4,0000
1	12	17	12	12	11	12,8	2,3875
8	12	12	12	11	12	11,8	0,4472

Tabla 2. Ranking de ejecuciones (se resaltan los máximos obtenidos en las ejecuciones)

5. Trabajos relacionados

Algunos trabajos presentes en la literatura se enfocan en la determinación de la importancia de un nodo en la red, como PageRank [7], AuthorRank [5] (en un ámbito de coautoría de trabajos académicos). Otros trabajos identifican métricas

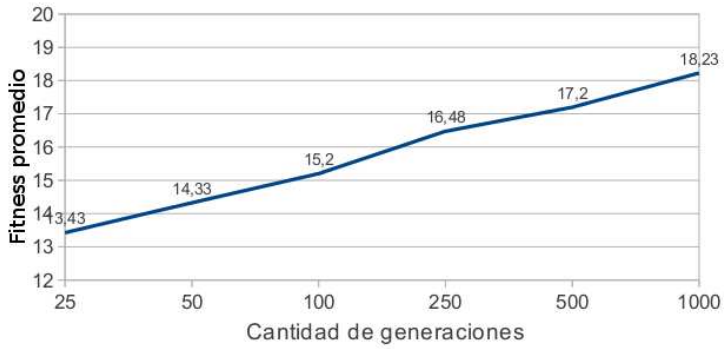

Figura 4. Fitness promedio obtenido a partir de las configuraciones de acuerdo a la cantidad de generaciones

de centralidad de los nodos de una red como *degree, closeness, y betweenness* [3]. En [6] se propone un modelo que generaliza las métricas de centralidad en redes donde se incorporan el número de relaciones y el peso de éstas en las métricas propuestas originalmente en [3].

Otros trabajos se enfocan en la identificación de vecinos más cercanos en la red. En [4] se define una red de coautoría de trabajos académicos para medir la cercanía de los autores y de este modo recomendar, en base a la proximidad, trabajos de otros autores. Por otra parte, en [10] se definen dos medidas de fortaleza del camino, aplicable en grafos valuados.

6. Conclusiones

Este trabajo presenta dos aportes principales al problema tratado. El primero de ellos es la utilización de redes sociales para la representación de relaciones entre individuos que permita deteminar los más distantes entre sí. Esta propuesta resulta innovadora debido a que, como se menciona anteriormente, la mayoría de los estudios se enfocan en la detección del grado de centralidad de los nodos, y no en tratar de resolver el problema en cuestión. El segundo aporte consiste en el método de identificación de la distancia entre estos individuos mediante la implementación de un algoritmo genético. Este trabajo demuestra que la utilización de un algoritmo genético, junto con la definición de una red social, resulta una alternativa posible y adecuada frente a otros enfoques (como la selección manual, o la selección aleatoria de miembros) cuando se tiene por objetivo la conformación de grupos de candidatos que presenten menor relación entre sí. Esto se da gracias a que la definición de la red social permite la generación de matrices de distancia, las cuales son utilizadas por el algoritmo genético en la

búsqueda de la maximización de una función que determine las distancias entre los miembros de una comisión candidata.

Este enfoque se encuentra inicialmente limitado en la definición de la red social, ya que exige que la misma se corresponda con un grafo conectado, es decir, que exista al menos un camino entre cada par de nodos de la red. En otros contextos, donde la red presente otras características como relaciones valuadas o dirigidas, o redes que presenten grupos desconectados, será necesario evaluar alternativas a la función de relación propuesta.

El mismo enfoque presentado en este trabajo podría mejorar su desempeño si se desarrollara una función de fitness que no se encuentre tan afectada por valores extremos, por ejemplo utilizando las variables estadísticas de media o mediana en lugar de una sumatoria de las distancias.

Trabajos futuros tienen por objetivo ampliar las dimensiones de la red en términos de cantidad de nodos, y tanto cantidad como tipo de relaciones entre los nodos, con objeto de analizar la escalabilidad del enfoque propuesto en este trabajo. Adicionalmente, se pretende analizar la conformación de grupos de candidatos mediante otras técnicas de inteligencia artificial.

Referencias

1. SBNS datasets. http://www.autonlab.org/autonweb/17433.
2. Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, March 1977. ArticleType: research-article / Full publication date: Mar., 1977 / Copyright © 1977 American Sociological Association.
3. Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1978.
4. San-Yih Hwang, Chih-Ping Wei, and Yi-Fan Liao. Coauthorship networks and academic literature recommendation. *Electronic Commerce Research and Applications*, 9(4):323–334, July 2010.
5. Xiaoming Liu, Johan Bollen, Michael L. Nelson, and Herbert Van de Sompel. Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6):1462–1480, December 2005.
6. Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, July 2010.
7. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. http://ilpubs.stanford.edu:8090/422/, November 1999.
8. J. E. Smith and Agoston E. Eiben. *Introduction to Evolutionary Computing*. Springer, October 2008.
9. Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, November 1994.
10. Song Yang and David Knoke. Optimal connections: strength and distance in valued graphs. *Social Networks*, 23(4):285–295, October 2001.