Revista Cubana de Ciencias Informáticas
Vol. 12, No. 4, Octubre-Diciembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 115-127
http://rcci.uci.cu

Tipo de artículo: Artículo original
Temática: Reconocimiento de patrones
Recibido: 26/10/2017 | Aceptado: 05/10/2018

Predicción de la evolución de comunidades en redes sociales

Prediction of community evolution in social networks

Armando Díaz Matos[1*], **Reynaldo Gil Pons**[2], **Reynier Ortega Bueno**[3]

[1]armando.diaz@cbiomed.cu
[2]rey@cerpamid.co.cu
[3]reynier.ortega@cerpamid.co.cu

[*]Autor para correspondencia: armando.diaz@cbiomed.cu

Resumen
Muchos investigadores se han volcado en la tarea de analizar y modelar el comportamiento de las redes sociales, debido al auge que han tomado. Son varias las tareas llevadas a cabo como parte de su análisis. Dentro de ellas destacan por su importancia, la descripción y predicción de la evolución de las comunidades que conforman la red. Esta última es tratada desde la perspectiva de las distintas formas que tienen las comunidades en la red; analizando su comportamiento en la evolución. En este trabajo se propone un método para la predicción de la evolución de comunidades en redes sociales basado en subgrafos frecuentes. Finalmente nuestra propuesta es comparada con un enfoque recientemente reportado en la literatura, obteniendo resultados similares.

Palabras claves: análisis de redes sociales, análisis dinámico de las redes, evolución de comunidades, predicción de la evolución de grupos.

Abstract
Nowadays, the fast growth of Social Networks have caused that many researchers have taken the challenge of analyzing and modelling their behavior. There are many tasks in Social Networks Analysis. One of the most important is the prediction of the behavior of communities that form the network. This last task is now analyzed with a new perspective where the behavior of the community is defined by its shape. We propose a method for predicting the evolution of communities based on frequent subgraphs. Then we compare the new way of describing the communities with a recent approach in the literature, obtaining similar results.

Keywords: *community evolution, dynamic social network analysis, prediction group evolution, social network analysis*

Introducción y Trabajos Relacionados

La expansión y múltiples formas de conexión a Internet han provocado que en los últimos tiempos las redes sociales se hayan convertido en unos de los servicios de mayor demanda. Una red social se define como una red

de interacciones o relaciones, donde los nodos consisten en actores y las aristas representan las interacciones o relaciones que se establecen entre dichos actores (Aggarwal, 2011). Debido a la relevancia de los procesos sociales el análisis de redes sociales es ampliamente utilizado en diversos campos como sociología, en el análisis de los patrones de relaciones y la naturaleza de los enlaces (Roman, 2012); epidemiología, en la influencia y determinación de la salud (Berkman et al., 2014); comunicación por correo electrónico, en el análisis de la topología de la red (Kossinets and Watts, 2006); telefonía celular para mejorar la experiencia del cliente dentro de la industria de las telecomunicaciones (Pinheiro, 2011); economía, en la búsqueda de empleo (Calvó-Armengol and Zenou, 2005), entre otros.

Las redes sociales generan un inmenso flujo de información sobre múltiples temas, los cuales representan opiniones de grupos. La forma en que los usuarios son agrupados dentro de una red social es denominada comunidad, la cual está conformada por un conjunto de usuarios que comparten propiedades en común (Fortunato, 2010). El desarrollo de las comunidades no siempre es el mismo, debido al aumento o disminución de la cantidad de usuarios, el surgimiento, cambio o eliminación de relaciones entre éstos, etc. Estas son dinámicas por su evolución en el tiempo, proceso en el que intervienen una gran variedad de factores que afectan tanto a la red como a las comunidades que la componen (Coleman et al., 1964; Freeman, 2004; Moody and White, 2000).

La detección de comunidades dentro del análisis de redes sociales permite agrupar a los usuarios siguiendo un determinado criterio. Actualmente existen varios métodos relacionados con el propósito de detectar comunidades, como son los métodos basados en modularidad (Newman and Girvan, 2004), métodos estocásticos (Choi et al., 2012) y métodos de agrupamiento heterogéneos (Asur et al., 2009).

La identificación de las distintas transformaciones que ocurren con el paso del tiempo en las comunidades de la red social, es el objetivo de los algoritmos de descripción de la evolución. Existen múltiples trabajos donde se aborda la tarea. En (Gliwa et al., 2012) se presenta el método *SGCI* (Identificación de Cambios en Grupos Estables) capaz de identificar la evolución de algunas de las comunidades en la red. Posteriormente (Gliwa et al., 2013), los autores realizan algunas modificaciones al método en cuanto a las medidas para el establecimiento de la semejanza entre las comunidades. Otro trabajo que identifica y describe la evolución de comunidades en una red social es (Bródka et al., 2013), donde se presenta el método *GED* (Descubridor de la Evolución de Grupos).

Una de las primeras propuestas para la descripción de la evolución de las comunidades es abordada en (White et al., 1976), donde se observa el tiempo y las relaciones establecidas, limitándose al análisis de los grupos de usuarios más densos. Por otro lado, en (Hu and Wang, 2009) se considera la forma de las comunidades y utilizan un modelo de difusión para dar seguimiento. En (Zhao et al., 2012) se analiza la dinámica de la red a diferentes niveles: usuario, comunidad y red; para determinar cuántos usuarios son influenciados por el proceso, permitiendo analizar la evolución a distintas escalas.

Una de las estrategias más empleadas por la comunidad científica para modelar los cambios que ocurren dentro de una red dinámica, consiste en dividir la misma en distintos marcos temporales representados por grafos. En dichos marcos se identifican las comunidades y los eventos críticos que tienen lugar entre ellas (Palla et al., 2007; Asur et al., 2009; Bródka et al., 2013; Gliwa et al., 2012).

La tarea de predicción de evolución de comunidades en redes sociales, se realiza considerando un conjunto de características estructurales, de centralidad o cambios temporales que describen a la comunidad, así como, su historial evolutivo. Sin embargo, los distintos trabajos de investigación reportados en la literatura no analizan las formas más comunes en que los grupos de usuarios se relacionan dentro de la comunidad o sea las subgrafos frecuentes derivados de la estructura de grafo de las comunidades.

El objetivo de este trabajo es estudiar el impacto que tienen la representación basada en subgrafos frecuentes en la tarea de predicción del comportamiento evolutivo de las comunidades. La propuesta consiste en una nueva representación basada en éstos, en donde su frecuencia es analizada como un rasgo de la comunidad.

Subgrafos Frecuentes en las Comunidades

Las comunidades que conforman a una red social se componen de las estrechas relaciones que se establecen entre los distintos usuarios que la conforman. La manera en que se relacionan estos son las que condicionan la forma que tiene la comunidad. Éstas son representables mediante un grafo es por esta razón que las distintas subformas que constituye a una comunidad pueden verse como subgrafos. Los subgrafos frecuentes asociados a una comunidad representan las formas más comunes en que grupos de usuarios se relacionan dentro de la misma. Con esta representación se analiza la influencia que tienen las relaciones entre pequeños grupos de usuarios en el comportamiento de la comunidad.

La extracción de subgrafos frecuentes es una tarea que generalmente consiste de dos pasos. Primeramente se generan los subgrafos candidatos y como segundo paso se comprueba la frecuencia de cada uno de éstos. La mayoría de los estudios que se concentran en esta tarea, tiene como principal objetivo optimizar la primera etapa debido a que el segundo paso involucra pruebas de isomorfismo cuya complejidad computacional es excesivamente alta siendo un problema NP-completo (Huan et al., 2004; Schreiber and Schwöbbermeyer, 2005; Wernicke, 2006; Grochow and Kellis, 2007; Kashani et al., 2009; Andrés Gago Alonso, 2009; Omidi et al., 2009).

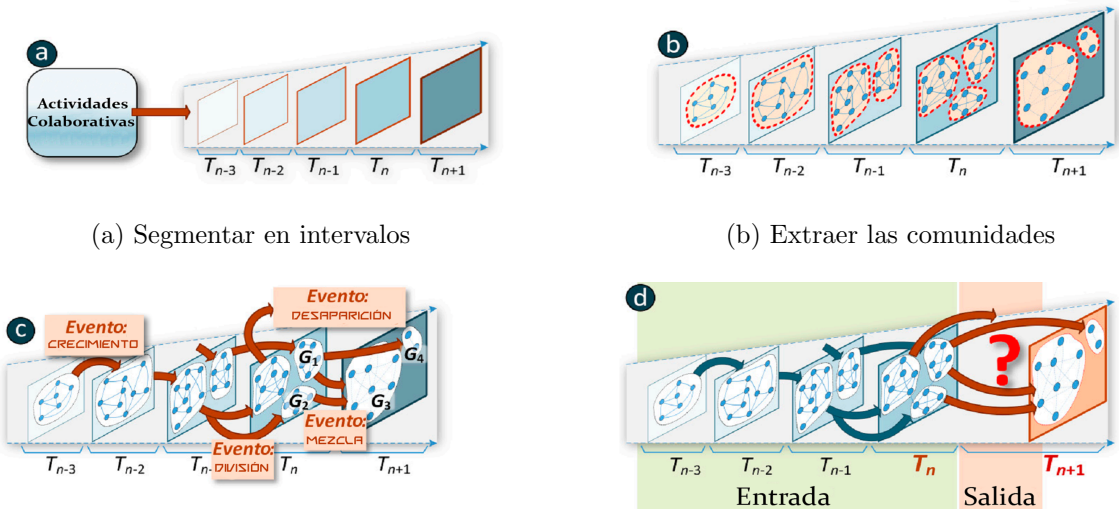
Existen varios métodos para la extracción de subgrafos frecuentes. En especial para la tarea de generación de candidatos existen dos estrategias básicas: basados en Apriori y basados en Crecimiento de Patrones. Después de realizar un estudio sobre esta tarea se decidió usar la herramienta *gtrieScanner*[1], construida como parte del trabajo desarrollado en (Ribeiro and Silva, 2010), para la extracción de los subgrafos frecuentes.

[1]Herramienta que usa la estructura de datos g-trie para contar la ocurrencia de un subgrafo en un grafo.

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 4, Octubre-Diciembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 115-127
http://rcci.uci.cu

Descripción General de la Propuesta

La predicción de la evolución de comunidades en una red social es una tarea que presenta un alto grado de complejidad y se compone de cuatro etapas, como se puede observar en la figura 1. En la primera etapa se recolectan los datos de la red social analizada u otra similar, con la frecuencia con que se desea realizar la predicción. Luego son extraídas las distintas comunidades que conforman a la red para cada intervalo de tiempo. Seguido, se detectan los distintos eventos o transformaciones que sufren las comunidades al pasar de un intervalo de tiempo al siguiente. Finalmente, cada una de las comunidades son agrupadas por el evento que les ocurre al pasar al próximo intervalo de tiempo y modeladas a través de una colección de subgrafos con sus frecuencias, para ser usadas en un modelo de clasificación supervisada.



(a) Segmentar en intervalos

(b) Extraer las comunidades

(c) Describir la evolución

(d) Predecir la evolución

Figura 1. Etapas en la predicción de la evolución de comunidades.

Modelado del Problema

Teniendo una colección de objetos $O = \{O_1, O_2, ..., O_n\}$, en donde O_i representa la $i - ésima$ comunidad de la colección. Cada objeto es modelado como un grafo $G_i = (V_i, A_i)$, donde V_i representa el conjunto de usuarios y A_i el conjunto de relaciones que se establecen entre los usuarios.

Para cada objeto se analizan los subgrafos frecuentes asociados, obteniéndose un conjunto de formas $X = \{X_1, X_2, ..., X_k\}$, donde k es el número de subgrafos frecuentes de la colección. Por consiguiente, para cada objeto de la colección O, se tiene que $X_i(O_j)$ es un valor que representa la frecuencia del $i - ésimo$ subgrafo del conjunto X para el $j - ésimo$ objeto de la colección O, de no existir toma valor 0. La representación o interpretación de un objeto se realiza a través de un vector de rasgos $I(O_j) = (X_1(O_j), X_2(O_j), ..., X_k(O_j))$.

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 4, Octubre-Diciembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 115-127
http://rcci.uci.cu

Cada una de las clases, a la cual pertenece una comunidad, está en correspondencia con los eventos del algoritmo de descripción de la evolución utilizado. En este trabajo son empleados el *GED* y el *SGCI* como algoritmos de descripción de la evolución. Para el uso del algoritmo de descripción *GED* se tienen las clases: *Constante, Disolución, Mezcla, División, Crecimiento y Reducción*. Por otro lado con el algoritmo de descripción *SGCI* se tienen las clases: *Constante, Disolución, Mezcla, División, División-Mezcla(Div/Mez), Adición, Eliminación y Cambio de Tamaño(Redimensión)*. De manera general, se define $C = \{c_1, c_2, ..., c_K\}$ como el conjunto de clases del problema, donde c_i representa la $i - ésima$ clase.

Cada uno de los datos recopilados en las etapas anteriores son utilizados bajo un modelo de reconocimiento de patrones con clasificación supervisada. En este modelo las comunidades son los distintos objetos de la colección y los subgrafos frecuentes son los rasgos que los distinguen. Los distintos eventos que ocurren con una comunidad al pasar de un intervalo de tiempo al siguiente, son las distintas clases del modelo.

Experimentación y Resultados

Para la realización de los experimentos se utilizaron dos tipos de redes sociales distintas; una red de coautoría (DBLP colección que forma parte de las colecciones públicas disponibles en (Ley, 2002)) y otra de comentarios de usuarios (Facebook, disponible en (Viswanath et al., 2009)). Estas forman parte de colecciones de datos estándares usadas dentro del análisis de redes sociales. Cada una de ellas presentan comportamientos distintos en cuanto a la forma en que los usuarios se relacionan y comparten la información, por lo que su evolución en el tiempo suele ser distinta.

Como parte de la descripción de las comunidades son usados los subgrafos frecuentes como rasgos asociados a cada una de ellas. Para ello se utilizó la herramienta *gtrieScanner*, la que necesita ajustar varios parámetros tales como el tamaño de los subgrafos, si son o no dirigidos y el método de búsqueda de los subgrafos. En nuestra propuesta se emplearon subgrafos frecuentes de tamaños 3 y 4 para cada una de las colecciones, los subgrafos frecuente son dirigidos y el método empleado para la búsqueda es ESU[2] (Wernicke, 2006).

Para los experimentos fueron usados los mismos algoritmos de descripción de la evolución, método de extracción de comunidades y las características estructurales expuestas en (Saganowski et al., 2015), las cuales son: *tamaño, densidad, cohesión, liderazgo, reciprocidad* y para *grado de entrada, grado de salida, grado total, intermediación, cercanía y vector propio* fueron calculados el promedio, el mínimo, el máximo y la suma de cada uno de los valores asociados a cada vértice de la comunidad. Las comunidades son descritas usando dos grupos de características, una basada en los subgrafos frecuentes (SF) y la otra en un el grupo de características estructurales mencionadas anteriormente (CE). Dentro del análisis de los subgrafos frecuentes, cuando su tamaño es superior al de la comunidad son utilizados los de tamaño inferior y en caso de no existir se utiliza la propia comunidad.

[2]Es un método de búsqueda y conteo de un subgrafo en una colección.

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 4, Octubre-Diciembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 115-127
http://rcci.uci.cu

Se experimentó con tres clasificadores diferentes los cuales se encuentran implementados en la plataforma Weka, utilizando sus configuraciones por defecto (ver tabla 1)

Nombre en Weka	Nombre del Clasificador
Bagging(REPTree)	Bootstrap aggregating(Breiman, 1996)
J48-C4.5 decision tree	C4.5 decision tree(Quinlan, 1993)
RandomForest	Random forest(Breiman, 2001)

Tabla 1. Clasificadores usados en Weka.

Para evaluar la calidad del método se realizó de manera separada para *GED* y *SGCI* una *validación cruzada* con muestreo estratificado y con tamaño de partición igual a 10.

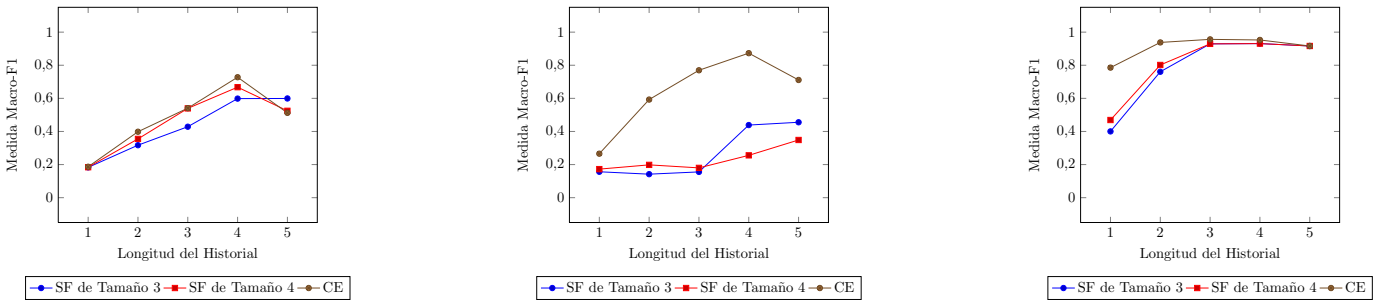
Experimentos con el algoritmo de descrición GED en la colección DBLP

En la tabla 2 se muestran los resultados de aplicar el algoritmo de descripción de la evolución *GED* en la colección DBLP y de analizar el historial evolutivo. Como se puede apreciar existe un desbalance entre las clases algo que tiende a disminuir a medida que crece el historial de la evolución. Las clases *Mezcla* y *División* son las que menor muestras presentan.

Eventos/Longitud de Historial	1	2	3	4	5
Constante	915	149	39	11	5
Disolución	23160	1884	366	87	19
Mezcla	108	33	8	4	4
División	64	70	15	7	3
Crecimiento	977	207	60	32	15
Reducción	947	225	77	28	21

Tabla 2. Cadenas de eventos ocurridos usando el algoritmo de descripción de la evolución *GED* en la colección DBLP.

A continuación son mostrados los resultados de evaluar el comportamiento de los clasificadores mencionados en la tabla 1 en función de la medida *Macro-F1*, utilizando las representaciones basadas en subgrafos frecuentes y las basadas en características estructurales.



(a) Comparación de Macro-F1 utilizando el clasificador Bagging

(b) Comparación de Macro-F1 utilizando el clasificador J48

(c) Comparación de Macro-F1 utilizando el clasificador RandomForest

Figura 2. Resultados de Macro-F1 obtenidos por clasificadores en DBLP usando *GED*

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 4, Octubre-Diciembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 115-127
http://rcci.uci.cu

En la figura 2 los mejores resultados para la estrategia de subgrafos frecuentes son alcanzados por el RandomForest, luego el Bagging y finalmente el J48. Para el caso de las características estructurales los mejores resultados alcanzados por los clasificadores solo varían de la propuesta anterior en el orden del J48 que obtiene mejores resultados que el Bagging. De manera general, la estrategia basada en características estructurales obtiene mejores resultados, en especial, cuando la longitud de las cadenas evolutivas es pequeña, donde son más notables las diferencias con respecto a los resultados alcanzados analizando los subgrafos frecuentes.

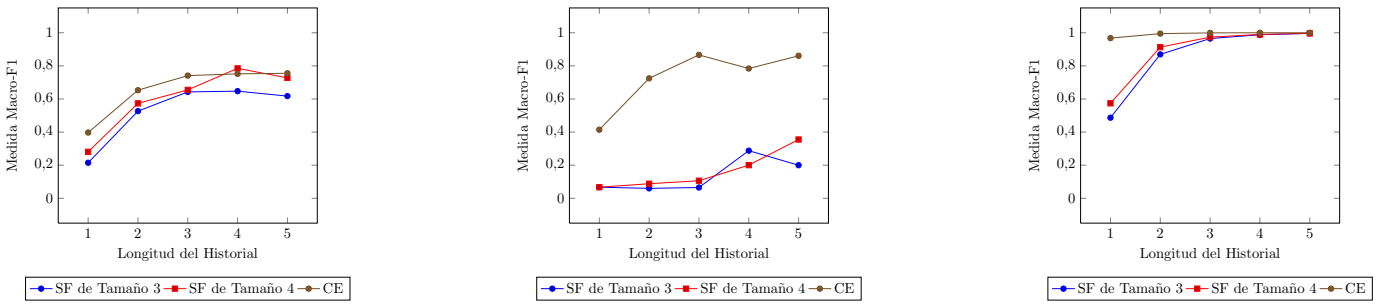
Experimentos con el algoritmo de descrición SGCI en la colección DBLP

En la Tabla 3 se muestran los resultados de aplicar el algoritmo de descripción de la evolución *SGCI* en la colección DBLP y analizar las cadenas evolutivas. Las clases se encuentran desbalanceadas y los eventos *División-Mezcla*, *Adición* y *Eliminación* son pocos comunes. A continuación son evaluados los resultados de la predicción utilizando la medida de calidad F1.

Eventos/Longitud de Historial	1	2	3	4	5
Constante	590	567	406	224	125
Disolución	384	439	466	427	173
Mezcla	269	192	171	92	62
División	212	195	146	100	73
División-Mezcla(Div/Mez)	20	20	14	9	6
Adición	5	4	3	2	2
Eliminación	2	6	2	2	2
Cambio de Tamaño(Redimensión)	860	656	455	224	139

Tabla 3. Cadenas de eventos con el algoritmo de descripción *SGCI* en la colección DBLP.

A continuación son mostrados los resultados de evaluar el comportamiento de los clasificadores mencionados en la tabla 1 en función de la medida *Macro-F1*.



(a) Comparación de MacroF1 utilizando el clasificador Bagging

(b) Comparación de MacroF1 utilizando el clasificador J48

(c) Comparación de MacroF1 utilizando el clasificador RandomForest

Figura 3. Resultados de Macro-F1 obtenidos por clasificadores en DBLP usando *SGCI*

En la figura 3 se puede apreciar la enorme diferencia que existe entre la representación basada en formas frecuentes con respecto al uso de características estructurales con el clasificador J48, mientras que con el resto de los clasificadores son similares los resultados es especial cuando incrementa la longitud del historial de la

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 4, Octubre-Diciembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 115-127
http://rcci.uci.cu

evolución. Los mejores resultados son alcanzados con el clasificador RandomForest para ambas representaciones, en el caso del Bagging con el uso de subgrafos frecuentes de tamaño 4 e historial de la evolución igual a 4 se superan los resultados alcanzados para la representación basada en características estructurales.

Los mejores resultados alcanzados en la colección de DBLP fueron conseguidos por el algoritmo de descripción de la evolución *SGCI* y usando RandomForest como clasificador a pesar de existir dos clases desbalanceadas y con pocas muestras. Aunque con el empleo del algoritmo de descripción *GED* se alcanza buenos resultados, en especial, cuando la longitud del historial evolutivo crece. El clasificador con mejor desempeño fue el RandomForest, luego el J48. La representación basada en subgrafos frecuentes de dimensión 4 alcanzan mejores resultados que los de tamaño 3, lo que se debe al incremento del número de posibles formas existentes.

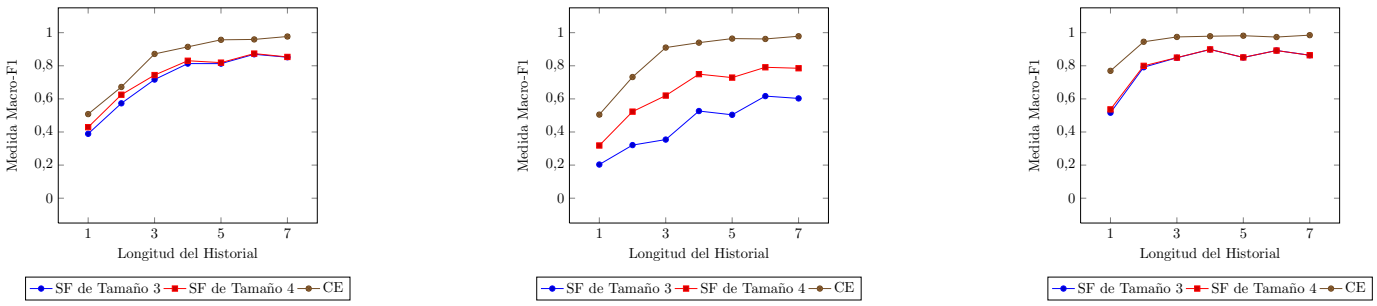
Experimentos con el algoritmo de descrición GED en la colección Facebook

En la Tabla 4 se muestran los resultados de aplicar el algoritmo de descripción de la evolución *GED* en la colección Facebook y analizar las cadenas evolutivas. Esta red social presenta mayor estabilidad en el tiempo por lo que los historiales son de mayor longitud, el evento menos frecuente en la colección es la *División* y seguido está *Mezcla*. El evento con más muestras es la *Disolución* y seguido se encuentra el evento *Constante*. La representación basada en subgrafos frecuentes de dimensión 4 alcanzan mejores resultados que los de tamaño 3, lo que se debe al incremento del número de posibles formas existentes.

Eventos/Longitud de Historial	1	2	3	4	5	6	7
Constante	5294	1388	994	1929	1448	2050	2050
Disolución	10098	6266	4493	11267	9896	11892	11892
Mezcla	1625	812	1249	2378	2608	3290	3290
División	554	10762	25883	18141	43221	37813	37819
Crecimiento	2466	1015	1119	1392	1946	2173	2173
Reducción	2564	2025	1419	3058	2673	3903	3903

Tabla 4. Cadenas de eventos ocurridos bajo *GED* en Facebook.

A continuación son mostrados los resultados alcanzados por los clasificadores mencionados en la tabla 1:



(a) Comparación de MacroF1 utilizando el clasificador Bagging

(b) Comparación de MacroF1 utilizando el clasificador J48

(c) Comparación de MacroF1 utilizando el clasificador RandomForest

Figura 4. Resultados de Macro-F1 obtenidos por clasificadores en Facebook usando *GED*

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 4, Octubre-Diciembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 115-127
http://rcci.uci.cu

En la figura 4 se puede apreciar que los mejores resultados son alcanzados con el RandomForest, tanto para el uso de subgrafos frecuentes como de características estructurales. Como segundo mejor clasificador se encuentra el J48 para el uso de características estructurales mientras que para el uso de subgrafos frecuentes como representación el segundo mejor clasificador es el Bagging. Para los subgrafos frecuentes de tamaño 4 se alcanzan mejores resultados debido a que existen una mayor cantidad de formas y distinguen mejor a las comunidades.

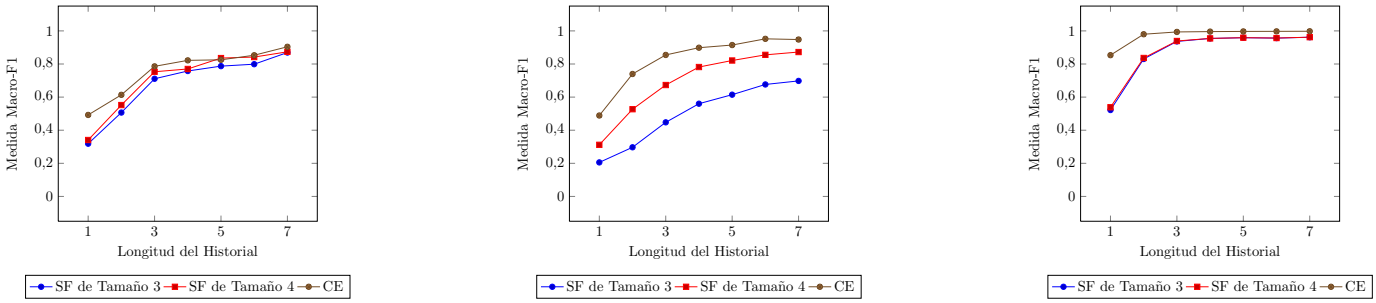
Experimentos con el algoritmo de descrición SGCI en la colección Facebook

En la tabla 5 se muestran los resultados de aplicar el algoritmo de descripción de la evolución *SGCI* en la colección Facebook y analizar las cadenas evolutivas. Los eventos *Eliminación* y *División-Mezcla* son los menos frecuentes en la colección y el más frecuente es *Redimensión* seguido de *Constante*.

Eventos/Longitud de Historial	1	2	3	4	5	6	7
Constante	2217	1597	1344	992	784	672	647
Disolución	1669	1525	1792	1839	1490	1289	1202
Mezcla	1207	830	710	732	694	657	675
División	661	1138	1162	1083	1165	1165	1108
División-Mezcla(Div/Mez)	137	76	67	40	32	27	22
Adición	385	197	176	202	156	188	196
Eliminación	54	260	243	197	210	230	276
Cambio de Tamaño(Redimensión)	2493	2656	2273	1877	1634	1630	1586

Tabla 5. Cadenas de eventos ocurridos bajo *SGCI* en Facebook.

A continuación son mostrados los resultados de evaluar el comportamiento de los clasificadores mensionados en la tabla 1 en función de la medida *Macro-F1*.



(a) Comparación de Macro-F1 utilizando el clasificador Bagging

(b) Comparación de Macro-F1 utilizando el clasificador J48

(c) Comparación de Macro-F1 utilizando el clasificador RandomForest

Figura 5. Resultados de Macro-F1 obtenidos por clasificadores en Facebook usando *SGCI*

En la figura 5 los mejores resultados son alcanzados por el clasificador RandomForest, como segundo mejor clasificador se encuentra el Bagging para el uso de subgrafos frecuentes mientras que con el empleo de

características estructurales está el J48 como segundo mejor clasificador. A medida que crece el historial de la evolución incrementa la calidad de los clasificadores lo que se debe a que se tiene más información de la comunidad. Con el uso de subgrafos frecuentes de tamaño 4 se obtienen mejores resultados que los de tamaño 3 debido a que existen más formas y discriminan mejor a las comunidades.

Para la colección de Facebook los mejores resultados con alcanzados con el algoritmo de descripción *SGCI* y el clasificador RandomForest, esto ocurre porque este algoritmo desecha las comunidades no estables incrementando la precisión ya que el comportamiento de estas comunidades es difícil de predecir, por lo que introducen problemas a la hora de ser clasificadas. La diferencia entre los resultados alcanzados por el algoritmo de descripción *GED* y el *SGCI* no son significativas ya que no superan el 5 %. El segundo mejor clasificador para la colección se mantuvo idéntico para el uso de uno u otro algoritmo de descripción de la evolución, siendo el J48 cuando son analizadas las características estructurales y el Bagging para el uso de subgrafos frecuentes. La representación basada en subgrafos frecuentes de dimensión 4 alcanzan mejores resultados que los de tamaño 3, lo que se debe al incremento del número de posibles formas existentes.

Conclusiones y Recomendaciones

En este trabajo se propuso una nueva forma de representar las comunidades basada en los subgrafos frecuentes. Además se realizaron distintos experimentos con las colecciones DBLP y Facebook, utilizando el *GED* y el *SGCI* como algoritmos de descripción de la evolución y los clasificadores RandomForest, Bagging y J48. El empleo de subgrafos frecuentes como representación de las comunidades alcanzó mejores resultados con el clasificador RadomForest, seguido está el Bagging, luego el J48. Los resultados alcanzados con el uso de características estructurales son superiores a los de subgrafos frecuentes, mas la diferencia entre ambas no sobrepasa el 10 %, desapareciendo cuando incrementa la longitud del historial de la evolución.

Como trabajos futuros se experimentará con subgrafos frecuentes de mayor dimensión, lo cual puede arrojar mejores resultados al caracterizar una mayor variedad de formas dentro de la comunidad. La combinación de la representación basada en subgrafos junto con el uso de las características estruturales es otras de las direcciones que queda pendiente y que podría alcanzar mejores resultados en la tarea de predicción del comportamiento evolutivo de las comunidades en redes sociales.

Referencias

Charu C Aggarwal. *An introduction to social network data analytics*. Springer, 2011.

José Eladio Medina Pagola Andrés Gago Alonso, Jesús Ariel Carrasco Ochoa. Minería de subgrafos conexos frecuentes en colecciones de grafos etiquetados. Technical report, Computer Science DepartmentNational Institute of Astrophysics, Optics and Electronics and Data Mining Department Advanced Technologies Application Center, 2009.

Sitaram Asur, Srinivasan Parthasarathy, and Duygu Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(4):16, 2009.

Lisa F Berkman, Ichiro Kawachi, and M Maria Glymour. *Social epidemiology*. Oxford University Press, 2014.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Piotr Bródka, Stanisław Saganowski, and Przemysław Kazienko. Ged: the method for group evolution discovery in social networks. *Social Network Analysis and Mining*, 3(1):1–14, 2013.

Antoni Calvó-Armengol and Yves Zenou. Job matching, social network and word-of-mouth communication. *Journal of urban economics*, 57(3):500–522, 2005.

David S Choi, Patrick J Wolfe, and Edoardo M Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012.

James Samuel Coleman et al. Introduction to mathematical sociology. *Introduction to mathematical sociology.*, 1964.

Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

Linton Freeman. The development of social network analysis. *A Study in the Sociology of Science*, 2004.

Bogdan Gliwa, Stanislaw Saganowski, Anna Zygmunt, Piotr Bródka, Przemyslaw Kazienko, and Jaroslaw Kozak. Identification of group changes in blogosphere. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 1201–1206. IEEE, 2012.

Bogdan Gliwa, Piotr Bródka, Anna Zygmunt, Stanisław Saganowski, Przemysław Kazienko, and Jarosław Koźlak. Different approaches to community evolution prediction in blogosphere. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1291–1298. ACM, 2013.

Joshua A Grochow and Manolis Kellis. Network motif discovery using subgraph enumeration and symmetry-breaking. pages 92–106, 2007.

Haibo Hu and Xiaofan Wang. Evolution of a large online social network. *Physics Letters A*, 373(12):1105–1110, 2009.

Jun Huan, Wei Wang, Jan Prins, and Jiong Yang. Spin: mining maximal frequent subgraphs from graph databases. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 581–586. ACM, 2004.

Zahra RM Kashani, Hayedeh Ahrabian, Elahe Elahi, Abbas Nowzari-Dalini, Elnaz S Ansari, Sahar Asadi, Shahin Mohammadi, Falk Schreiber, and Ali Masoudi-Nejad. Kavosh: a new algorithm for finding network motifs. *BMC bioinformatics*, 10(1):318, 2009.

Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *science*, 311(5757): 88–90, 2006.

Michael Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. In *International symposium on string processing and information retrieval*, pages 1–10. Springer, 2002.

James Moody and Douglas R White. Social cohesion and embeddedness: A hierarchical conception of social groups. *Unpublished manuscript, Ohio State University*, 2000.

Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):26–113, 2004.

Saeed Omidi, Falk Schreiber, and Ali Masoudi-Nejad. Moda: an efficient algorithm for network motif discovery in biological networks. *Genes & genetic systems*, 84(5):385–395, 2009.

Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Quantifying social group evolution. *Nature*, 446 (7136):664–667, 2007.

Carlos Andre Reis Pinheiro. *Social network analysis in telecommunications*, volume 37. John Wiley & Sons, 2011.

J Quinlan. C4. 5: Programs for machine learning. c4. 5-programs for machine learning/j. ross quinlan, 1993.

Pedro Ribeiro and Fernando Silva. G-tries: an efficient data structure for discovering network motifs. pages 1559–1566, 2010.

Caterina G Roman. *Social Networks, Delinquency, and Gang Membership*. PhD thesis, The Urban Institute, 2012.

Stanisław Saganowski, Bogdan Gliwa, Piotr Bródka, Anna Zygmunt, Przemysław Kazienko, and Jarosław Koźlak. Predicting community evolution in social networks. *Entropy*, 17(5):3053–3096, 2015.

Falk Schreiber and Henning Schwöbbermeyer. Mavisto: a tool for the exploration of network motifs. *Bioinformatics*, 21(17):3572–3574, 2005.

Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, August 2009.

Sebastian Wernicke. Efficient detection of network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(4), 2006.

Harrison C White, Scott A Boorman, and Ronald L Breiger. Social structure from multiple networks. i. blockmodels of roles and positions. *American journal of sociology*, 81(4):730–780, 1976.

Xiaohan Zhao, Alessandra Sala, Christo Wilson, Xiao Wang, Sabrina Gaito, Haitao Zheng, and Ben Y Zhao. Multi-scale dynamics in a massive online social network. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 171–184. ACM, 2012.