UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

RECOMENDACIÓN DE CONTACTOS PARA LA OPTIMIZACIÓN DE REDES SOCIALES

Claudia Lucio Sarsa
Tutor: Pablo Castells Azpilicueta

JUNIO 2017

RECOMENDACIÓN DE CONTACTOS PARA LA OPTIMIZACIÓN DE REDES SOCIALES

AUTOR: Claudia Lucio Sarsa
TUTOR: Pablo Castells Azpilicueta

Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio de 2017

Resumen

En el campo del análisis de las redes sociales, dos de los ámbitos más populares son el estudio de la difusión de la información y de la recomendación de contactos. El primero orientado recientemente al marketing viral y el segundo dado al auge que han tenido los sistemas de recomendación en los últimos años y como mejora de la experiencia del usuario en la red. En este Trabajo de Fin de Grado se estudia la aplicación de los algoritmos de maximización de la difusión como recomendación de contactos para optimizar la difusión de información a través de una red social, centrándose en el caso particular de Twitter.

Para verificar la eficacia de estos algoritmos, aplicados como recomendación no personalizada, se compararon los resultados contra los obtenidos aplicando métricas de topologías de redes, también aplicados como recomendación no personalizada, y contra algoritmos típicos de recomendación personalizada de contactos. Se analizan los resultados desde tres perspectivas distintas:

1. Los resultados de los propios modelos de maximización de la difusión.
2. Utilización en el marketing viral.
3. Aplicados a la recomendación de contactos.

En el primer punto, se prueban los algoritmos clásicos de difusión de información: el modelo de la cascada independiente y el modelo del umbral lineal, con el objetivo de medir cuántos nodos de un grafo pueden ser activados comenzando con un conjunto pequeño de nodos semillas activados inicialmente.

En el segundo punto, se prueban los algoritmos de difusión y los de métricas de topologías de redes y se realiza una simulación de marketing viral, donde los nodos semilla seleccionados por cada algoritmo difunden el mismo mensaje cada uno y se mide la velocidad y la propagación en la red.

En el tercer punto, aparte de los algoritmos mencionados anteriormente también se prueban unos algoritmos clásicos de recomendación personalizada de contactos. Los usuarios semilla de cada algoritmo son aplicados como recomendación no personalizada al resto de usuarios del grafo, excepto los algoritmos de recomendación personalizada, que para cada usuario recomiendan unos usuarios de manera personalizada. En estas simulaciones múltiples usuarios tienen varios tuits que propagar. Se medirá la velocidad, la cantidad de información nueva y la propagación.

Como conclusión, los resultados de la experimentación demuestran que los algoritmos de difusión de información, aplicados como recomendación de contactos no personalizada, pueden ser utilizados para optimizar la información que fluye a través de una red, aunque también tienen sus desventajas frente a otros algoritmos como es el tiempo necesario (NP-hard) para encontrar los usuarios semillas.

Palabras clave

Maximización de la difusión de información, Recomendación, Red social, Topología de red, Velocidad, Propagación.

Abstract

In the field of social network analysis, two of the most popular areas are the study of the information diffusion and friend recommendation system. The first one focuses on viral marketing recently and the second one to improve the user experience in the network due to the rise of recommendation systems in recent years. In this Bachelor Thesis, we study the use of diffusion maximization algorithms as friend recommendation systems to optimize the information diffusion through a social network, focusing on Twitter.

To verify the effectiveness of these algorithms, applied as a non-personalized recommendation, the results are compared with those obtained by applying metrics of network topologies, also applied as a non-personalized recommendation, and against typical personalized recommendation algorithms. The results are analyzed from three different perspectives:

1. The results of the diffusion maximization models.
2. Their use in viral marketing.
3. Applied as friend recommendation systems.

Firstly, the classical information diffusion algorithms are tested: the independent cascade model and the linear threshold model, with the objective of measuring how many graph nodes can be activated starting with a small set of activated seed nodes.

Secondly, the diffusion algorithms and network topology metrics are tested and a viral marketing simulation is performed, where the seed nodes selected by each algorithm spread the same message and the velocity and propagation are measured.

Thirdly, apart from the algorithms mentioned above, we also test classic personalized recommendation algorithms. The seed users of each algorithm will be applied as a non-personalized recommendation to the rest of the users of the graph, except the algorithms of personalized recommendation which for each user will recommend some users in a personalized way. In these simulations, multiple users will have several tweets to propagate. The speed, the amount of new information and the propagation will be measured.

In conclusion, the experimentation results demonstrate that information diffusion algorithms applied as non-personalized recommendation systems can be used to optimize information flowing through a network, although they also have their disadvantages compared to other algorithms such as the amount of time needed (NP-hard) to find the seed users.

Keywords

Information diffusion maximization, Recommender system, Social network, Network topology, Speed, Propagation.

Agradecimientos

En primer lugar, agradezco a mi tutor Pablo por haberme propuesto este Trabajo de Fin de Grado, así como por toda su ayuda para enseñarme lo necesario para realizar este trabajo y para resolver las dudas que iban surgiendo. Asimismo, le doy las gracias a Javier, por ayudarme con la plataforma de pruebas y explicarme las cosas básicas de su funcionamiento.

Por último, gracias a mis amigos de la facultad, con los que he compartido muchos momentos, especialmente en las prácticas de los laboratorios. También a mi familia, por su apoyo durante toda la carrera.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. Introducción

1.1 Motivación

En el campo del análisis de redes sociales, uno de los principales objetivos es el estudio de cómo las ideas, es decir, la información se propaga por una red. El problema de la difusión de información viene motivado desde diferentes áreas, entre las que destacan:

- En marketing, con el objetivo de que los clientes adquieran un nuevo producto [1]. Típicamente, se usa el marketing directo, donde la decisión de promover un producto a una persona se basa únicamente en las características de la misma, o el marketing de masas, donde el producto va destinado a un grupo o segmento de la población. En ambos métodos no se suele tener en cuenta el efecto que producen los clientes de un mercado sobre otros clientes, pero existe una gran cantidad de mercados donde las opiniones de los clientes influyen considerablemente sobre los demás. El marketing viral hace uso de esta característica y se centra en promover sus productos sobre los clientes con mayor influencia en el mercado [3]. Si trasladamos esta idea a las redes sociales, el objetivo sería encontrar esos usuarios cuya influencia haría que la información se propagase a más usuarios y/o más rápidamente.

- En epidemiología, para poder limitar la propagación de una enfermedad o virus [2]. Aunque en este campo el objetivo sería encontrar a las personas que, en caso de que se infectaran, contagiarían a un gran número de gente para así vacunarlas y evitar esa situación, en la maximización de la difusión de información sería buscar el efecto contrario, encontrar a esas personas que fueran las que primero se "contagiarían" para propagar la información, que haría el papel de enfermedad.

- En sociología ya se estudiaban los lazos sociales y la manera en la que fluye la información entre individuos incluso antes de que existieran las redes sociales online, mediante la observación de redes sociales de la vida real [4].

Las redes sociales se basan fundamentalmente en relaciones entre usuarios, éstas pueden ser unidireccionales como ocurre en Twitter donde los usuarios siguen a unos usuarios y son seguidos por otros, o bidireccionales como en el caso de Facebook donde las personas se siguen mutuamente.

Relaciones bidireccionales (Grafo no dirigido)

Relaciones unidireccionales (Grafo dirigido)

Figura 1-1: Tipos de relaciones

Una táctica para que aumente el número de relaciones en una red es la utilización de algoritmos de recomendación de contactos para que los usuarios descubran nuevas personas con las que conectarse. Al aumentar el número de relaciones se consigue que haya mayor cantidad de información fluyendo y que ésta se difunda más rápidamente al haber una gran cantidad de caminos disponibles. Con todo esto se logra que la red sea más eficiente como medio de comunicación de información entre personas.

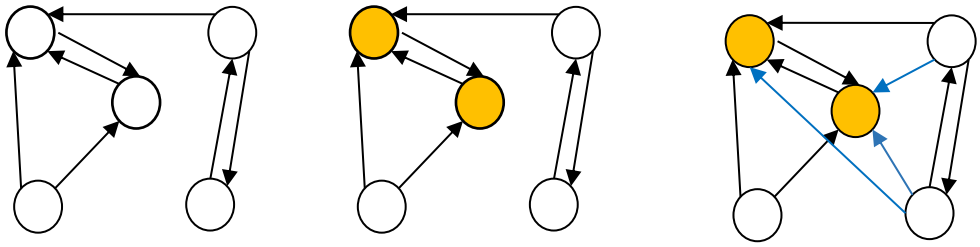
1.2 Objetivo

El objetivo principal de este trabajo es optimizar las redes sociales desde el punto de vista de maximizar la velocidad de difusión y la cantidad de información que fluye por ellas combinando las dos estrategias siguientes:

La primera son los algoritmos de maximización de la difusión de información que se usan para encontrar el menor número de usuarios desde los cuales la información se propaga al mayor número posible de usuarios.

La segunda es la recomendación de contactos. En el caso de redes sociales donde las relaciones son unidireccionales como en Twitter, de donde se han obtenido los datos para crear la red que se usa para las simulaciones de esta memoria, cuantos más seguidores tenga un usuario, mayor será el número de personas a las que llegarán sus ideas y, de la misma manera, a cuanta más gente siga, mayor cantidad de ideas de diferentes personas le llegarán. Por eso mismo, usaremos la recomendación para crear nuevas relaciones que incrementarán la difusión de ideas.

Juntando estos dos frentes, usaremos los algoritmos de maximización de la difusión para encontrar usuarios semilla que posteriormente introduciremos como recomendación no personalizada de contactos para el resto de usuarios de la red con el objetivo de crear nuevos arcos que den lugar a nuevos caminos por donde la información se propagará más rápidamente y llegará a más usuarios de la red.



Paso 1. Red social (grafo dirigido).

Paso 2. Elección de usuarios semilla.

Paso 3. Creación de nuevas relaciones.

Figura 1-1: Combinación de ambas estrategias.

Para comprobar la efectividad de esta combinación se compararán los resultados con los proporcionados por otros algoritmos basados en las topologías de las redes sociales (equivalentes a grafos a nivel estructural), como la distribución de los enlaces y la cohesión y centralidad de los nodos. Asimismo, también se compararán los resultados con los de

algoritmos típicos de recomendación de contactos, tanto recomendación no personalizada como personalizada.

1.3 Organización de la memoria

La memoria consta de los siguientes capítulos:

- En el capítulo 2 se explican algoritmos de cada uno de los campos que se compone este trabajo: difusión de información, métricas de topologías de redes y recomendación de contactos.

- En el capítulo 3 se elabora un breve resumen de las simulaciones realizadas y se determinan los algoritmos elegidos para las pruebas.

- En el capítulo 4 se explica la plataforma de pruebas y los datos utilizados, se describen con mayor detalle las simulaciones y se exponen las métricas utilizadas.

- En el capítulo 5 se muestran los resultados de las simulaciones y se elabora un análisis de los mismos.

- En el capítulo 6 se exponen las conclusiones de este trabajo y se plantean las tareas pendientes que habría que hacer en el futuro.

2. Estado del arte

A continuación, se desarrollan algunos algoritmos de cada ámbito de los que se compone este trabajo.

2.1 Difusión de información en redes sociales

Como se explicó previamente, el campo de la difusión de información utiliza técnicas de una gran variedad de ciencias. En este capítulo, se detallarán cuatro tipos de modelos de difusión de información: modelo de la cascada independiente, modelo del umbral lineal, modelos epidémicos y el comportamiento gregario.

Todos estos modelos tienen tres componentes en común [5]:

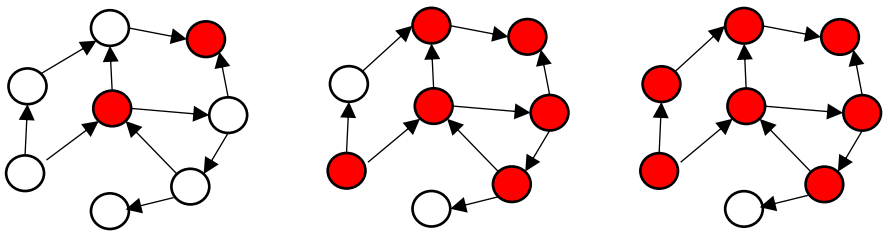
- Emisor(es): un emisor o un pequeño grupo de emisores comienzan la difusión de información.
- Receptor(es): un receptor o conjunto de receptores recibe la información difundida. Normalmente este grupo es mucho mayor que el de emisores.
- Medio: es el canal por el que la información se propaga.

2.1.1 Modelo de la cascada independiente

Hay diversas variantes de este modelo, en este apartado discutiremos la versión detallada por D. Kempe et al. en [7]. El modelo hace las siguientes suposiciones [5]:

- La red se representa como un grafo dirigido. Los nodos son los emisores y receptores y los enlaces muestran los canales de comunicación entre ellos. Un nodo sólo puede influenciar a los que están conectados con él.
- Los nodos sólo pueden estar activos o inactivos. Un nodo activo significa que éste ha decidido adoptar el comportamiento, la información, la innovación o la decisión.
- Cuando un nodo está activado puede activar a sus nodos vecinos. En cada enlace existe un peso que actúa de probabilidad de que el nodo activo active a su vecino inactivo.
- Los nodos pueden cambiar de inactivos a activos, pero no viceversa.

El modelo comienza con un conjunto semilla de nodos activos y finaliza cuando ya no se pueden activar más nodos.



1. Conjunto inicial de nodos activos.

2. Activamos vecinos (seguidores en este caso).

3. No se pueden activar más nodos. El modelo termina.

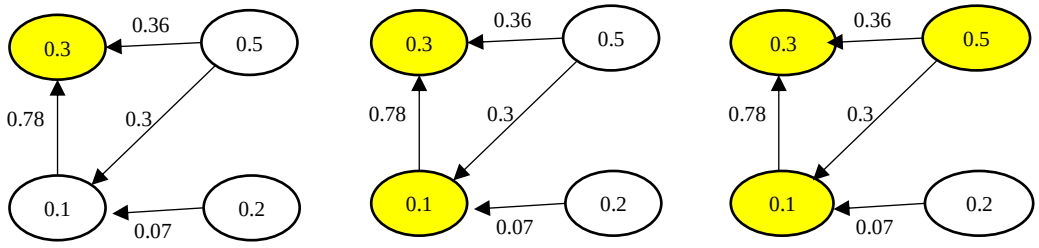
Figura 2-1: Simulación del modelo con probabilidad 1 en todos los enlaces.

2.1.2 Modelo del umbral lineal

En este modelo, un nodo ν puede ser influenciado por su vecino ω según el peso $b_{\nu,\omega}$ del enlace que los une. Cada nodo tiene un umbral $\theta_\nu \in [0,1]$, por lo que un nodo ν se activa cuando la suma de los pesos de los enlaces a sus vecinos activos es mayor o igual que el umbral del nodo ν [7], es decir:

$$\nu \text{ se activa si } \theta\nu \leq \sum_{\omega \text{ vecino activo de } \nu} b\nu,\omega \quad \forall\nu \in N$$

Al igual que en la cascada independiente, una vez que un nodo ha sido activado no puede cambiar a estado inactivo. El modelo comienza con un conjunto de nodos activos y finaliza cuando ya no se activan más nodos.



1. Conjunto inicial de nodos activos.

2. Activamos vecinos (seguidores) comprobando los umbrales

3. No se pueden activar más nodos, el modelo para.

Figura 2-2: Simulación del umbral lineal.

2.1.3 Modelos epidémicos

Una epidemia es una enfermedad que se extiende ampliamente en una población. El proceso consta de un patógeno (la enfermedad que se propaga), un conjunto de huéspedes (humanos y animales, entre otros) y un mecanismo de contagio (aire, agua, etc.) [5].

El modelo de la cascada independiente está relacionado con los modelos epidémicos pero lo que distingue a estos últimos es que tienen un alto nivel de incertidumbre y los individuos no eligen ser infectados o no. En los siguientes modelos detallados se supone que no hay información disponible sobre la red y que el proceso por el que el huésped se contagia es desconocido. Estos modelos pueden ser aplicados a redes sociales donde el proceso de decisión tiene una componente de incertidumbre o es ambiguo para el analista [5].

Asumimos que tenemos una población donde se ha esparcido una enfermedad. Cada individuo de esa población puede estar en uno de los siguientes estados [5]:

- Susceptible: ocurre cuando un individuo puede ser potencialmente infectado por otros individuos ya infectados de la población.
- Infectado: un individuo en este estado puede infectar a otros individuos susceptibles.
- Curado (o Muerto): pueden ser individuos que se han recuperado de la enfermedad y por tanto tienen inmunidad parcial o total hacia ella, o que han muerto a causa de la infección.

Se distinguen cuatro modelos epidémicos [5]:

1. ***Modelo SI***: el modelo más básico, los individuos susceptibles se infectan y ya no pueden curarse. Hay una probabilidad β de establecer contacto entre cada par de individuos. Se asume que cuando un individuo infectado se encuentra con otro susceptible, hay una probabilidad 1 (puede generalizarse a cualquier otro valor) de que éste último se infecte. El modelo finaliza cuando toda la población se ha infectado.

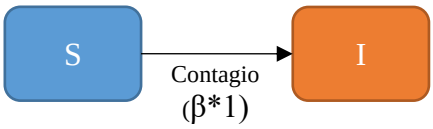


Figura 2-3: Transiciones del modelo SI.

Se ha observado que este modelo tiene un crecimiento parecido al del VIH/sida entre los años 1980 y 1995 en EE. UU, teniendo en cuenta que no todos los individuos de la población estadounidense son susceptibles de tener VIH.

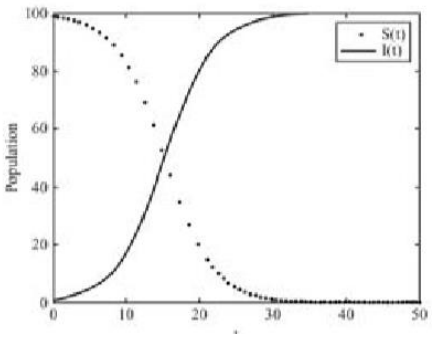


Figura 2-4: Simulación del modelo SI [5].

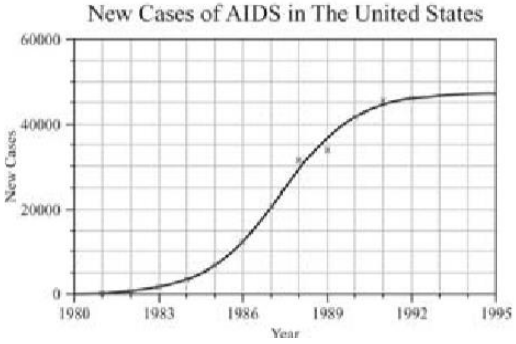


Figura 2-5: Crecimiento de la población infectada por VIH en EE. UU [5].

2. ***Modelo SIR***: añade más detalle al modelo SI, incluyendo el estado de recuperación de la enfermedad. Aparte de la probabilidad β nombrada anteriormente, existe una probabilidad γ de que un individuo infectado se recupere de la infección en un periodo de tiempo Δt.

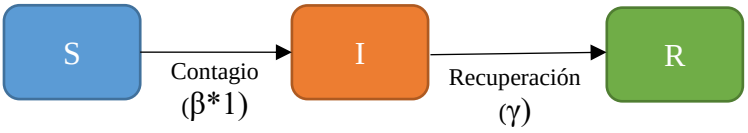


Figura 2-6: Transiciones del modelo SIR.

3. ***Modelo SIS***: es igual que el modelo SI, pero añadiendo que los individuos infectados que se han recuperado de la enfermedad pueden volver a ser susceptibles.

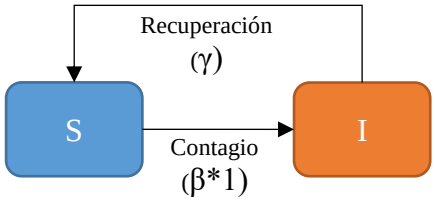


Figura 2-7: Transiciones del modelo SIS.

4. ***Modelo SIRS***: extiende el modelo SIR. En este modelo se asume que, pasado cierto tiempo, los individuos que se han recuperado pierden inmunidad y se convierten en susceptibles de nuevo. Se añade un nuevo parámetro λ que es la probabilidad de perder la inmunidad de un individuo curado.

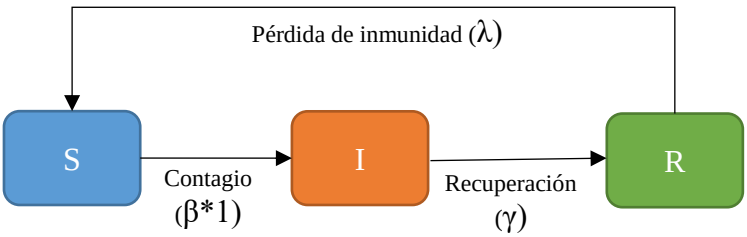


Figura 2-8: Transiciones del modelo SIRS.

2.1.4 Comportamiento gregario

El comportamiento gregario ocurre cuando los individuos observan las acciones de *todos* los demás y actúan de forma alineada con ellos sin planificación previa. Este comportamiento se ha observado en rebaños y manadas de animales, eventos deportivos, manifestaciones, etc. [5].

A diferencia de los modelos de la cascada independiente y del umbral lineal, donde las decisiones se toman basándose en sus vecinos inmediatos (dependencia local), en el comportamiento gregario las decisiones se basan en todos los individuos de la red (dependencia global) [5].

Se puede utilizar un modelo bayesiano para explicar el comportamiento gregario como se muestra en el libro "Social Media Mining. An Introduction" [6].

2.2 Métricas de topologías de redes

La topología de la red es determinante para poder definir el comportamiento de la misma ante diferentes fenómenos, por lo que su análisis nos ayuda en su comprensión. Por ejemplo, de la misma manera que en una red de carreteras la forma en que se han distribuido las autovías determina cómo de rápido y eficiente se desplazarán los viajantes hasta sus destinos, en una red social ocurre lo mismo ya que la manera en que los usuarios están conectados influye en cómo las personas son informadas o pueden propagar sus ideas.

A continuación, se exponen métricas basadas en la centralidad y la cohesión.

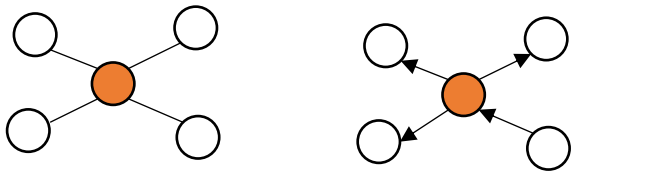
2.2.1 Centralidad

La centralidad de un nodo determina la importancia de éste dentro de un grafo.

2.2.1.1 Grado

Denotado como g(*u*), siendo *u* un nodo, es el número de conexiones que tiene un nodo de un grafo. En grafos dirigidos distinguimos dos tipos de grado:

- Grado entrante (indegree): sólo consideramos las conexiones que entran al nodo.
- Grado saliente (outdegree): sólo consideramos las conexiones salientes del nodo.



g(naranja) = 4

indegree(naranja) = 1
outdegree(naranja) = 3

Figura 2-9: Ejemplo del grado de un nodo.

2.2.1.2 Betweenness

Mide la centralidad de un nodo basándose en el número de caminos mínimos entre dos nodos que pasan por este nodo. Se puede representar como:

$$B(u) = \sum_{s \neq u \neq t \in V} \frac{\sigma_{st}(u)}{\sigma_{st}} \qquad [8]$$

donde σ_{st} es el número total de caminos de distancia mínima del nodo *s* al nodo *t*, y $\sigma_{st}(u)$ es el número de esos caminos mínimos que pasan por el nodo u.

2.2.1.3 Closeness

Se calcula como la suma de las longitudes de los caminos de distancia mínima de un nodo al resto de nodos de la red. Cuanto más central sea un nodo, más cercano es para los demás nodos.

Normalmente, al hablar de closeness nos referimos a su forma normalizada, es decir, la longitud promedio de los caminos de distancia mínima:

$$C(u) = \frac{N-1}{\sum_{v \in V} d(u,v)} \qquad [9]$$

Indica una posición de influencia por la rapidez para llegar a los demás nodos.

2.2.1.4 PageRank

El algoritmo PageRank modela las probabilidades de que un caminante aleatorio en un grafo se encuentre en cada nodo del mismo en un instante dado [12].

En el cálculo de PageRank, para cada uno de los nodos del grafo, se tiene en cuenta la cantidad de enlaces entrantes en cada nodo, la relevancia de los nodos de los cuales provienen esos enlaces y finalmente se valora inversamente el número de enlaces salientes de los nodos de los cuales provienen las aristas entrantes nombradas anteriormente. Por tanto, este algoritmo queda definido por la siguiente función que asigna a cada usuario su valor de PageRank:

$$P(u) = \frac{r}{|V|} + (1 - r) \sum_{\substack{v \to u \\ v \in V}} \frac{P(v)}{\#out(v)}$$

Donde *#out(v)* es el número de aristas que salen del nodo *v* y $r \in (0,1)$ es el factor de teleportación que determina la probabilidad del caminante aleatorio de trasladarse a un nodo cualquiera del grafo, independientemente de que exista o no un enlace que una el nodo en el que el caminante se encuentra en ese momento con el nodo al que se teleporta.

2.2.1.5 HITS

El algoritmo de HITS (Hypertext Induced Topic Search), también llamado hubs y autoridades, es un algoritmo basado en el análisis de enlaces. Clasifica los nodos según sus aristas entrantes y salientes [13]:

- Autoridad: nodo con muchas aristas entrantes.
- Hub: nodo con muchas aristas salientes.

Por tanto, HITS puede ser definido por las siguientes funciones:

$$authority(u) = \sum_{\substack{v \to u \\ v \in V}} hub(v)$$

$$hub(u) = \sum_{\substack{u \to v \\ v \in V}} authority(v)$$

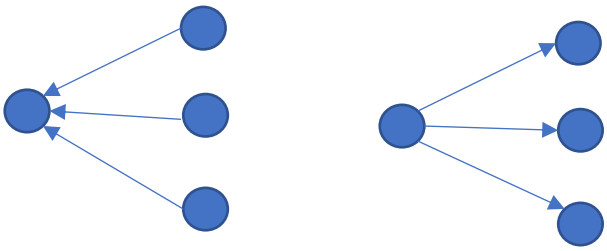


Figura 2-10: Representación de autoridad y hub

2.2.2 Cohesión

2.2.2.1 Coeficiente de clustering local

Muestra como de completo es el entorno de un nodo. Se basa en la noción del cierre triádico que explica que, si dos usuarios tienen un amigo en común, la probabilidad de que estos dos estén conectados aumenta: "los amigos de mis amigos son mis amigos".

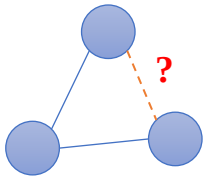


Figura 2.11: Cierre triádico.

Determina en qué medida los vecinos de un nodo están conectados. Se calcula como:

$$C(u) = \frac{nº\ de\ conexiones\ entre\ vecinos\ de\ u}{nº\ de\ conexiones\ posibles\ entre\ vecinos\ de\ u} \in [0,1]$$

Donde el número de conexiones posibles entre vecinos de *u* es,

$$\begin{cases} g(u)(g(u)-1)\ si\ el\ grafo\ es\ dirigido \\ \dfrac{g(u)(g(u)-1)}{2}\ si\ el\ grafo\ es\ no\ dirigido \end{cases}$$

2.2.2.2 Coeficiente de clustering global

Representado por C(G), es la probabilidad de que dos nodos tomados al azar con un amigo en común estén conectados. Refleja la cohesión global de entornos en la red. Su valor depende de cómo se hayan formado las relaciones en la red, si se han formado de manera aleatoria será bajo o, si han sido por amistad, similitud o popularidad, será alto.

Se calcula de la siguiente manera:

$$C(G) = \frac{nº\ de\ caminos\ cerrados\ de\ longitud\ 2}{nº\ de\ caminos\ de\ longitud\ 2}$$

2.3 Recomendación de contactos

La recomendación de contactos consiste en sugerir a un usuario otros usuarios con los que podría establecer una relación basándose en la similitud entre ambos, el número de amigos en común, etc. Se pueden hacer dos tipos de recomendación:

- No personalizada: no se tienen en cuenta ni las características del usuario a quien va dirigida la recomendación ni las de los usuarios susceptibles de ser recomendados.

Es una recomendación válida para cualquier usuario de la red. Ejemplo: recomendar a todos los usuarios el usuario con el mayor número de seguidores.

- Personalizada: se basa en características del usuario objetivo y de los susceptibles de ser recomendados. Ejemplo: recomendar un usuario a otro porque tienen muchos amigos en común.

2.3.1 Popularidad

Se basa en encontrar los usuarios más populares en una red, es decir, aquellos que tienen mayor número de amigos o de seguidores en las redes dirigidas.
Es equivalente al grado entrante de un nodo y se puede usar para ofrecer una recomendación no personalizada.

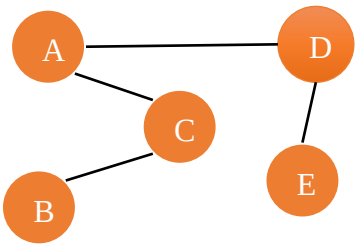
2.3.2 FOAF

FOAF son las siglas de "Friend Of A Friend", es decir, "Amigo De Un Amigo", mide la probabilidad de que dos personas que tienen un amigo en común sean presentadas por éste.

Verifica la correlación entre el número de amigos comunes de *x* e y en tiempo *t,* y la posibilidad de que *x* e *y* se conozcan en un futuro. Se puede utilizar como recomendador personalizado y se calcula como:

$$FOAF(x, y) = |N(x) \cap N(y)|$$

Donde *N(x)* es el conjunto de amigos de *x* y *N(y)* el conjunto de amigos de *y*.



N(A) = {C, D}
N(B) = {C}

FOAF(A, B) = C = 1

Figura 2-12: Ejemplo de FOAF.

2.3.3 Coeficiente de Jaccard

El coeficiente de Jaccard, una métrica muy usada en el campo de la recuperación de información, mide la probabilidad de que *x* e *y* tengan una característica *f*, seleccionada aleatoriamente del conjunto de características de *x* e *y* [11].

Si tomamos las características como vecinos, entonces esta métrica mide la proporción de amigos de *x* que también son amigos de *y*, y viceversa, y así determina la similitud entre *x* e *y*. Se calcula como:

$$J(x, y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|}$$

Donde *N(x)* es el conjunto de amigos de *x* y *N(y)* el conjunto de amigos de *y*.

En recomendación se utiliza para dar una recomendación personalizada, proponiendo a cada usuario un conjunto de usuarios similares a él.

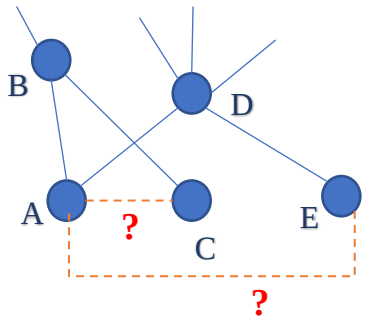
2.3.4 Coeficiente de Adamic/Adar

Refina la métrica de simplemente contar las características comunes dando más peso a las características más raras. El coeficiente de Adamic/Adar formaliza la intuición de que características más raras ofrecen mayor información [10,11], por ejemplo, dos documentos que comparten la palabra "ciudad" son probablemente menos similares que dos documentos que comparten la palabra "Madrid".

El cierre triádico es un mecanismo frecuente por el que se forman nuevas relaciones en las redes sociales, es decir, cuando *x* e *y* son presentados por un amigo común *z*. Entonces, una persona no popular (que no tiene muchos amigos) tiene mayor probabilidad de presentar un par de amigos que no se conocen entre ellos.

El coeficiente de Adamic/Adar se puede utilizar como recomendación personalizada, y se calcula de la siguiente manera:

$$Adamic(x,y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{\log |g(z)|}$$



$$N(A) \cap N(C) = B$$

$$Adamic(A,C) = \frac{1}{\log|g(B)|} = \frac{1}{\log 3} = 2.09$$

$$N(A) \cap N(E) = D$$

$$Adamic(A,E) = \frac{1}{\log|g(D)|} = \frac{1}{\log 5} = 1.43$$

Figura 2-13: Ejemplo de Adamic/Adar.

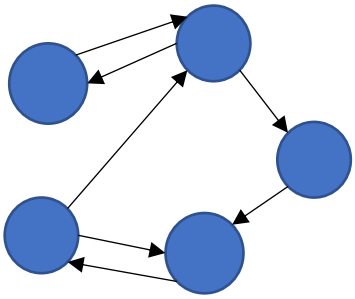
2.3.5 Algoritmo aleatorio

Algoritmo trivial donde se selecciona un número de usuarios al azar para recomendar a cada usuario. Se trata de una recomendación no personalizada, sólo se tiene en cuenta que los usuarios no estén previamente conectados con los que se recomienda.

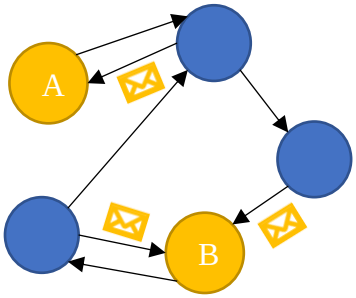
3. Selección de algoritmos

Dada la gran amplitud de algoritmos disponibles para la experimentación, vamos a filtrar cuales algoritmos nos interesan y cuales no basándonos en qué información o enfoque interesantes nos pueden aportar cada uno. Primero, resumimos brevemente el procedimiento que se seguirá durante la experimentación:

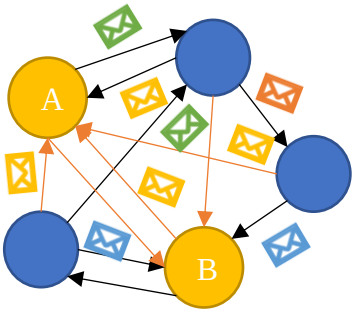
1. Se construye el grafo dirigido, a partir de datos reales de Twitter, donde la manera en que la información se propaga es equivalente a la de esta plataforma: cuando alguien publica información, ésta llega a todos sus seguidores.
2. Se eligen, mediante un algoritmo, los usuarios semilla desde los cuales se inicia la propagación de información de una pieza de información (1 tuit) por cada semilla. Se mide la cantidad y velocidad de la información.
3. Los usuarios semilla seleccionados previamente pasan a ser usados como recomendación no personalizada para todos los nodos del grafo. Se ejecuta una simulación de un grafo de Twitter donde los nodos tienen tuits que enviar y se utilizan las métricas mencionadas previamente, pero habiendo añadido previamente los enlaces creados con la recomendación no personalizada.
4. Se comparan los resultados obtenidos en el punto anterior con los obtenidos en una simulación con recomendación personalizada.



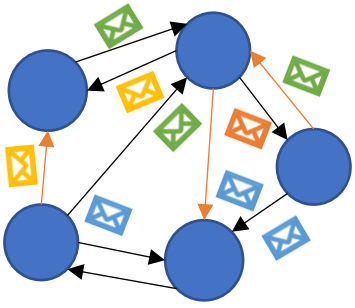
Paso 1: se construye el grafo.



Paso 2: búsqueda de semillas y propagación de 1 tuit (el mismo) por semilla.



Paso 3: los nodos tienen un número variado de tuits que enviar. Se añade recomendación no personalizada con las semillas obtenidas y se ejecuta la simulación.



Paso 4: Se añade recomendación personalizada y se ejecuta la simulación.

Figura 3-1: Representación de la experimentación.

Ahora que tenemos establecidos los pasos a seguir podemos empezar con la selección de algoritmos.

3.1 Difusión de información

De los cuatro tipos diferentes de modelos de difusión de información: la cascada independiente, el umbral lineal, los modelos epidémicos y el comportamiento gregario, se han elegido los dos primeros debido a las siguientes razones:

- Son los modelos básicos y más nombrados en la literatura de la difusión de información.
- Ambos modelos representan dos aspectos diferentes de las interacciones sociales. El modelo de la cascada independiente se centra en la interacción y la influencia individual entre amigos en una red social, mientras que el modelo del umbral lineal se basa en el umbral en la propagación de influencia, si por ejemplo un número suficiente de amigos compra un producto o usa una nueva red social, nos puede llevar a realizar la misma acción [14].
- Los modelos epidémicos están altamente relacionados con el modelo de la cascada independiente por lo que aportan poca información nueva con respecto a éste y para el análisis no nos interesa esa componente de incertidumbre que tienen.
- Los individuos del comportamiento gregario disponen de información a nivel global y tienen constancia de cómo se comportan el resto de individuos y actúan según todos ellos lo cual no se asemeja al modelo típico de las redes sociales donde los individuos solo conocen la información proveniente de sus vecinos.

El problema de optimización para seleccionar los nodos más influyentes en ambos modelos de difusión es NP-hard. Para la cascada independiente y el umbral lineal se han programado su aproximación codiciosa, propuesta por Kempe et al., que obtiene una aproximación del 63% de lo óptimo [7].

En cuanto al modelo de la cascada independiente, se aclara que los usuarios activan a sus seguidores y no viceversa. Por otro lado, las activaciones dependen del número aleatorio generado para cada par de nodos, si el número generado es menor o igual que el valor del peso de la arista, el vecino será activado. Debido a esto es difícil determinar el número de nodos que serán activados al final, ya que dependerán de los números aleatorios generados. Pero se convirtió en determinista al generar los números aleatorios al comienzo del modelo para toda la red. Así, dado un conjunto inicial de nodos activados, podemos calcular el número total de nodos activados al final del proceso.

En el modelo del umbral lineal, un nodo se activa cuando la suma de los pesos de los enlaces que conectan con los nodos activados a los que sigue es mayor o igual que su umbral. Asimismo, dado que la aproximación codiciosa original para encontrar las semillas iniciales era excesivamente lenta, se optó por añadirle una pequeña modificación y ejecutar el algoritmo codicioso sólo en los 2**k* nodos con mayor grado, siendo *k* el número de semillas que se quieren obtener.

3.2 Métricas de topologías de redes

Todas estas métricas se usarán para encontrar usuarios semilla, tanto para la simulación de propagación de un tuit por semilla como para ser utilizados como recomendación no personalizada.

De las métricas de centralidad se han seleccionado:

- Grado entrante (indegree): aunque es la más simple, nos aporta información sobre la importancia del nodo en la red.
- Betweenness: los nodos con alto betweenness tienen una posición de influencia por su papel en el paso de información.
- PageRank: proporciona indicios sobre la importancia de los nodos basándose en la estructura de enlaces de la red.
- HITS: nos centraremos en las autoridades ya que en Twitter para la propagación de información es importante el número de seguidores.

No se usará closeness ya que en las redes naturales las distancias suelen ser muy cortas por lo que la métrica variará poco entre los nodos. Además, es una métrica inestable a pequeños cambios en el grafo, por ejemplo, con sólo conectar un enlace a un nodo muy central se dispara el valor.

De las métricas de cohesión se ha seleccionado únicamente el coeficiente de clustering local por ser una métrica individual, por lo que se pueden clasificar los nodos según el valor de la métrica, a diferencia del coeficiente de clustering global. Se seleccionarán como semillas aquellos nodos con bajo clustering, ya que un valor bajo indica una posición ventajosa en la transmisión de información mientras que un valor alto sólo nos proporcionaría redundancia en la comunicación.

3.3 Recomendación de contactos

Los siguientes algoritmos se utilizan para las simulaciones con recomendación:

- Para recomendación no personalizada se usan popularidad, es equivalente a indegree, y el algoritmo aleatorio, que se usará de referencia en las simulaciones.
- Para recomendación personalizada se usan FOAF y los coeficientes de Jaccard y Adamic/Adar porque, aunque los tres se basen en amigos comunes y el cierre triádico, cada uno proporciona un punto de vista distinto.

4. Experimentación

4.1 Plataforma de pruebas

Para ejecutar las simulaciones se ha usado un conjunto de programas en Java proporcionados por el Information Retrieval Group de la Universidad Autónoma de Madrid.
Estos programas contienen ya implementados algoritmos de recomendación y métricas de topologías de redes, a excepción del algoritmo de betweenness y todo lo relacionado con los modelos de difusión de información que fueron codificados aparte en Java.

4.1.1 Conjunto de datos

El conjunto de datos originales se trataba de un grafo y unos tuits obtenidos de Twitter. La red se componía de 10.000 usuarios, 230.000 relaciones y más de 2 millones de tuits.

Debido a la cantidad de memoria necesaria para ejecutar las simulaciones y el tiempo computacional de algunos algoritmos, se optó por codificar un programa en Java que creara subgrafos con tuits a partir del grafo y de la base de datos de tuits originales con el objetivo de disminuir el tiempo de ejecución para poder realizar un mayor número de simulaciones y reducir el coste de memoria RAM. El nuevo grafo que se usa en las pruebas se compone de 3.000 usuarios, 206.000 relaciones y 30.000 tuits.

4.2 Configuración

En este apartado se explican con mayor detalle las diferentes simulaciones y métricas realizadas.

4.2.1 Métricas

En cada simulación se miden varias o todas de estas métricas en cada iteración:

- Cantidad de nueva información recibida en esa iteración. Se calcula como la suma de la información nueva que le ha llegado a cada usuario dividido por el número de usuarios de la red.
- Velocidad: información total recibida en todas las iteraciones hasta la actual sin contar repetidos, es decir, si a un usuario le llega el mismo tuit desde dos usuarios diferentes sólo contará como uno.
- Información para ser propagada y/o repropagada en esa iteración. Es la suma de todos los tuits que tienen que ser propagados en esa iteración.

4.2.2 Simulaciones

4.2.2.1 Modelos de difusión de información

Esta simulación se realizará sólo con los dos algoritmos de difusión de información: la cascada independiente y el umbral lineal.

Se buscarán las semillas con la aproximación codiciosa y posteriormente se ejecutarán los modelos como se describieron en el capítulo del estado del arte. Para la cascada independiente se utiliza el grafo de 3.000 nodos mencionado anteriormente, pero para el umbral lineal, debido a que necesitaba un tiempo excesivamente largo para simular el modelo, se probó con un grafo más pequeño de 500 nodos. Se mide cuántos nodos son activados al finalizar la simulación.

4.2.2.2 Simulación de marketing viral

En esta simulación primero se buscan los usuarios más influyentes que serán los que comiencen la difusión propagando el mismo tuit cada uno, el resto de usuarios no disponen de ningún tuit. El objetivo es ver a cuántos usuarios alcanzan ese tuit y en cuántas iteraciones. En cada iteración cada usuario enviará el tuit (si lo tiene) a sus seguidores y finalizará la simulación cuando no se propague ningún tuit.

Se calculará la influencia de cada usuario basándose en la valoración del algoritmo utilizado. Los algoritmos que se usan para la simulación son: cascada independiente, umbral lineal, indegree, Betweenness, PageRank, HITS (autoridades), coeficiente de clustering local y aleatorio.

4.2.2.3 Difusión de información con recomendación no personalizada

Primero se obtendrán las semillas iniciales como se hizo en la simulación de marketing viral y esas semillas pasarán a ser recomendadas al resto de nodos de la red y se conectarán con todos aquellos con los que antes no tenían enlace.
Una vez realizado esto, se ejecutará una simulación de difusión de tuits, pero con estos nuevos enlaces de recomendación no personalizada. En cada iteración se propaga un tuit propio y un tuit de los recibidos, con una probabilidad p de elegir propagar un tuit proveniente de un enlace recomendado y una probabilidad $1 - p$ de que sea de un enlace del grafo original. Se ha elegido que esa probabilidad p sea de 0.5.

Para la recomendación, los usuarios son clasificados según el valor que les determine el algoritmo utilizado. Los algoritmos que se usan para la simulación son: cascada independiente, umbral lineal, indegree (equivale a popularidad), Betweenness, PageRank, HITS (autoridades), coeficiente de clustering local y aleatorio.

4.2.2.4 Difusión de información con recomendación personalizada

Para cada usuario de la red se buscan k usuarios para recomendarle usando un algoritmo de recomendación personalizada. Una vez hecho esto, se ejecuta una simulación como la descrita en el punto anterior.

Los algoritmos que se usan para la simulación son: FOAF, coeficiente de Jaccard y coeficiente de Adamic/Adar.

5. Resultados y comparativas

Para abreviar, a los modelos de la cascada independiente y del umbral lineal se les denotará por sus siglas en inglés: ICM y LTM respectivamente.

5.1 Modelos de difusión de información

En los modelos de difusión siempre se busca un número pequeño de semillas (K) que sean las que inicien y maximicen la difusión de información. En las pruebas se seleccionado un número de semillas que va del uno al cinco ya que las redes no eran de gran tamaño.

El primer modelo, ICM, ha sido probado con dos grafos, uno de 3000 nodos y otro de 500 nodos. El segundo modelo, LTM, sólo con el de 500 nodos. Para cada modelo, grafo y número de semillas se han realizado varias ejecuciones y se han calculado la media de nodos totales activados al finalizar y la desviación típica.

Tabla 5-1: Nodos activados por ICM para el grafo de 3000 nodos.

	K=1	K=2	K=3	K=4	K=5
Media	14,56	20,4	31,2	32,4	45,6
Desviación	10,08745095	9,922860369	11,50900269	11,76813022	19,17811252

Tabla 5-2: Nodos activados por ICM para el grafo de 500 nodos.

	K=1	K=2	K=3	K=4	K=5
Media	2,04	4	6	7,9	9,2
Desviación	0,213200716	0	0	0	0,707106781

Tabla 5-3: Nodos activados por LTM para el grafo de 500 nodos.

	K=1	K=2	K=3	K=4	K=5
Media	20,24	30,6	29,73333333	28,7	49,2
Desviación	8,719136043	11,82058864	8,9957662	4,595891885	16,26960356

Como se puede observar, para el grafo de 500 nodos LTM obtiene resultados mucho mejores que ICM, pero tiene la desventaja de su elevado tiempo de ejecución, ya que, independientemente del número de semillas iniciales, las ejecuciones tuvieron que ser abortadas para el grafo de 3000 nodos.

5.2 Simulación de marketing viral

Como se hizo en la simulación anterior se buscó un conjunto de semillas iniciales, los usuarios más 'influyentes', para que sirvieran de inicio para que todos difundieran el mismo mensaje y poder comprobar a cuántos usuarios alcanza. El conjunto de semillas se eligió desde una a cinco semillas.

Independientemente del número de semillas y el algoritmo siempre se alcanzaban 2935 nodos, pero no en el mismo número de iteraciones como se puede ver en la tabla 5-4. La siguiente gráfica muestra la velocidad de propagación, que equivale a cuántos nodos han recibido ese mensaje. Se ha elegido mostrar los resultados para el conjunto de cinco semillas iniciales ya que todos los conjuntos de semillas llegaban al mismo número de nodos que han recibido el mensaje, pero éste fue el más rápido. El resto de gráficas para las diferentes pruebas puede encontrarse en el anexo A.

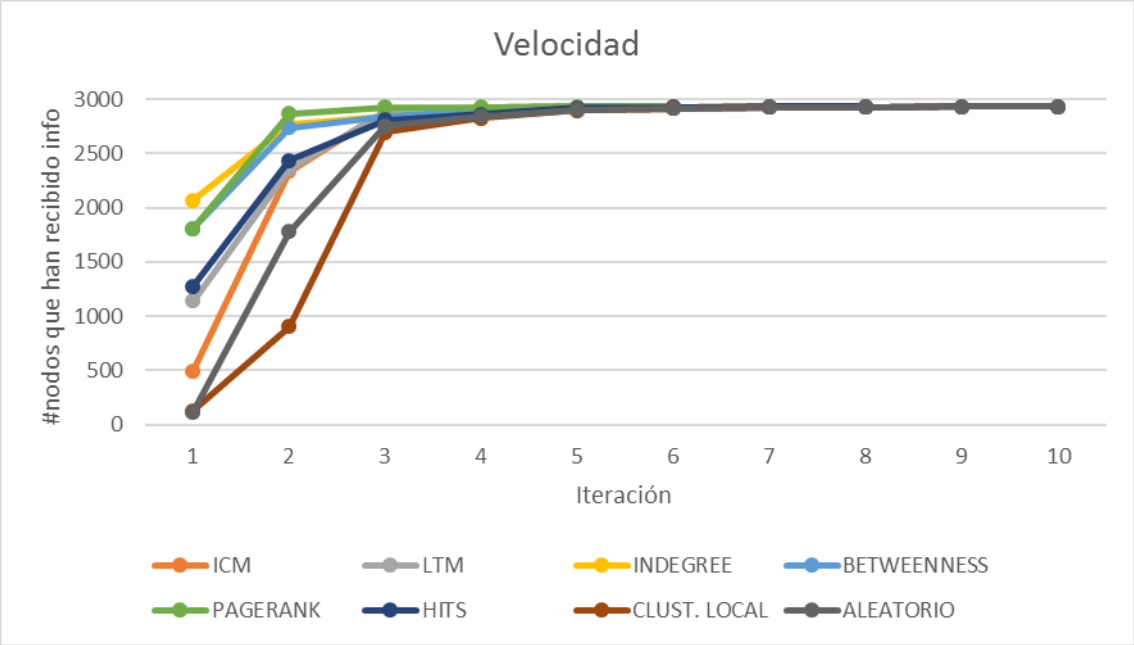


Figura 5-1: Gráfica resumen de la velocidad de propagación por algoritmo.

Observando la gráfica de la figura 5-1 podemos apreciar como con las semillas elegidas por grado entrante consigue que un gran número de nodos reciban la información en las primeras iteraciones, pero rápidamente es alcanzado por el resto de algoritmos. A pesar de que es el algoritmo que mejores resultados obtiene al principio no es el más rápido ya que como se puede ver en la tabla 5-4, necesita más iteraciones que otros algoritmos como PageRank o ICM.

Resulta interesante descubrir que LTM, que había obtenido mejores resultados que ICM en las simulaciones individuales (ver tablas 5-2 y 5-3) y es mejor en las primeras iteraciones de esta simulaión, queda rezagado e ICM consigue informar al mismo número de usuarios, pero casi en la mitad de iteraciones que LTM. Además, ICM ha sido el único algoritmo en conseguir llegar a los 2935 usuarios en siete iteraciones y con sólo tres usuarios semilla como se puede ver en la gráfica de la figura A-1 del anexo A. Otros algoritmos que, a pesar de comenzar en las primeras iteraciones con valores altos, acaban en un número alto de iteraciones en comparación con los más rápidos son Betweenness e HITS.

Cabe destacar que el algoritmo que empieza más lento es el coeficiente de clustering local que es superado incluso por el aleatorio. Esto puede ser debido a que en el afán de buscar valores de clustering bajos que nos dieran nodos que fueran posiciones ventajosas en la transmisión de información, estos valores fueran demasiado bajos y se obtuvieron nodos no tan valiosos como se esperaba.

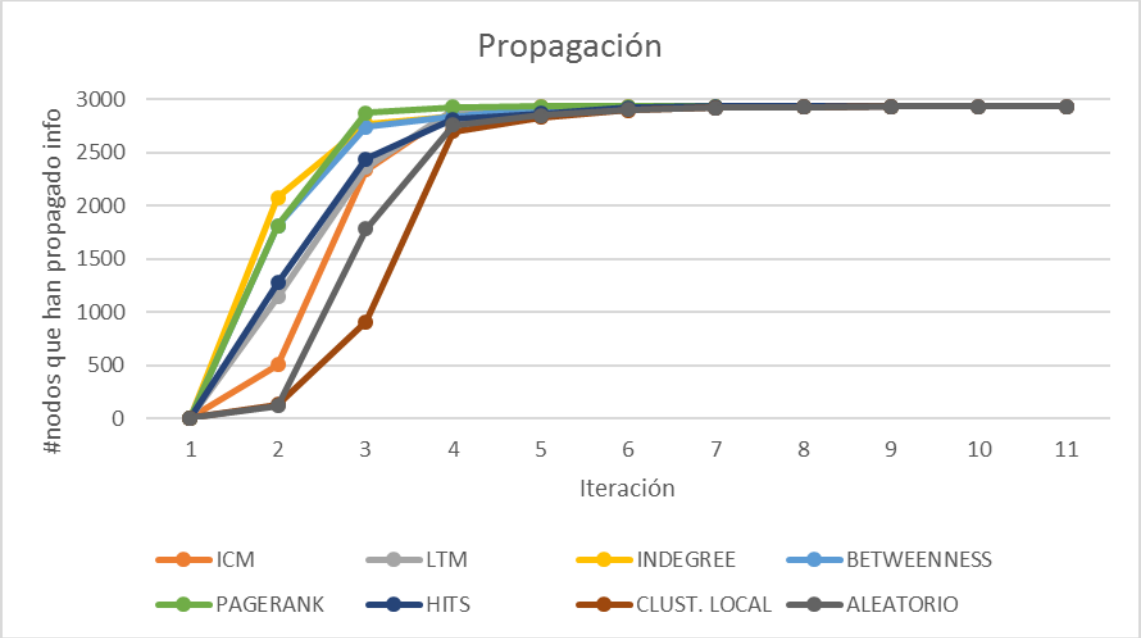


Figura 5-2: Gráfica resumen de la propagación de información por algoritmo.

En la figura 5-2, se muestran los resultados que representan la suma acumulativa de nodos que han propagado el mensaje en cada iteración. Los resultados son parecidos a los obtenidos en la gráfica de la figura 5-1.

Tabla 5-4: Iteraciones necesarias para acabar la simulación por algoritmo.

Algoritmo	**ICM**	**LTM**	**Indegree**	**Betweenness**	**PageRank**	**HITS**	**Clust. Local**	**Aleatorio**
Iteraciones	7	11	9	11	7	9	11	11

Tabla 5-5: Número de seguidores promedio de las semillas seleccionadas.

Algoritmo	ICM	LTM	Indegree	Betweenness	PageRank	HITS	Clust. Local	Aleatorio
Seguidores promedio	232	465	1393	1069	1009	1174	56	47

Para acabar el análisis de esta simulación, se calculó el número de seguidores promedio de los usuarios semilla más influyentes de cada algoritmo, véase la tabla 5-5. Cabe destacar que ICM, a pesar de que la media de seguidores promedio es un número relativamente bajo, ha sido el algoritmo más rápido incluso que indegree que parecía el candidato a ser el ganador por el hecho de empezar la difusión sobre aquellos individuos con más seguidores. También ha sido capaz de equiparar en iteraciones a PageRank, cuyos usuarios semilla tienen unas cinco veces más de seguidores promedio que los de ICM.

Estos resultados nos aportan un dato muy interesante desde el punto de vista del marketing viral: para hacer llegar información a una cantidad de gente no es necesario ir directamente a por aquellas personas que a simplemente vista parecen más influyentes por el número de seguidores, sino que existe la opción de recurrir otros individuos que no tengan tanta influencia directa pero que son puntos clave en la hora de la transmisión de la información. Esto puede ser útil para el marketing que hacen en las redes sociales aquellas empresas que no disponen de los recursos necesarios para costear una campaña publicitaria a grandes dimensiones ni para conseguir que uno o varios de los usuarios más influyentes de la red con muchos seguidores les publicite un producto, ya que mediante un análisis de la red se podrían encontrar usuarios no tan conocidos ni con un gran número de seguidores pero que con una cantidad pequeña de dinero u ofreciéndoles una muestra del producto aceptarían publicitarlo y dárselo a conocer a sus seguidores.

5.3 Difusión de información con recomendación de contactos

Al igual que en las simulaciones anteriores, se buscó un conjunto de semillas iniciales, los usuarios más 'influyentes'. Sin embargo, en vez de que sirvieran de inicio para que todos difundieran el mismo mensaje, esta vez se usaron para ser recomendados al resto de usuarios del grafo con los cuales no estuvieran ya conectados, en caso de que fueran algoritmos que se pudieran usar como recomendación no personalizada. Pero en el caso de Adamic/Adar, FOAF y Jaccard, se utilizó una recomendación personalizada en la cual a cada usuario se le recomendaban k usuarios diferentes, donde k es el número de semillas. El conjunto de semillas se eligió desde una a cinco semillas.

En esta simulación, los usuarios propagan cada iteración un tuit propio y un tuit de los recibidos de otros usuarios, con probabilidad p de elegir un tuit proveniente de un enlace recomendado y con probabilidad $1 - p$ uno proveniente de un enlace del grafo original, p tomará el valor de 0,5.

Se ha decidido mostrar los resultados de las simulaciones con cinco semillas ya que era el valor que obtenía mejores resultados en todos los algoritmos. Se pueden ver el resto de gráficas para los diferentes valores de semillas en los anexos B y C.

Si observamos las gráficas de las figuras 5-3 y 5-5 podemos ver que los algoritmos se clasifican en dos tipos:

1. Algoritmos que difunden más rápido la información, pero la cantidad de información recibida a lo largo de la red es menor.
2. Algoritmos que difunden más lentamente la información, pero la cantidad de información recibida a lo largo de la red es mayor.

Ambos tipos tienen sus ventajas y sus inconvenientes que analizaremos a continuación.

Empecemos con los algoritmos del grupo 1: más rápidos pero menor cantidad de información total al final de la simulación. En este grupo se encuentran los siguientes algoritmos: ICM, LTM, grado entrante, betweenness, PageRank, HITS y el coeficiente de clustering local. Las semillas obtenidas han sido aplicadas en todos los algoritmos como recomendación no personalizada.

Se consideran más rápidos porque acaban en menos iteraciones y en cada iteración propagan más información que los del otro grupo, como se puede ver en la figura 5-5. El que más información total propaga de este grupo es LTM, seguido de cerca por betweenness y grado entrante. Los tres necesitan prácticamente el mismo número de iteraciones para finalizar la simulación (véase tabla 5-4). El más lento es el coeficiente de clustering local, aunque tiene a su favor que no ha necesitado tantas iteraciones como ICM, por ejemplo. Si se hubieran usado unos usuarios con un coeficiente no tan bajo tal vez este algoritmo podría haber obtenido mejores resultados. En el punto intermedio se encuentran HITS, ICM y PageRank, siendo ICM el que más iteraciones ha necesitado, pero ha propagado más información, mientras que PageRank ha sido lo opuesto, ha sido el que menos iteraciones ha necesitado, pero también el que menos información ha propagado de los tres.

A pesar de que estos algoritmos han sido los que han conseguido acabar en menos iteraciones y propagando una mayor cantidad de información, sobre todo en las primeras iteraciones (véase figura 5-4), no han obtenido tan buenos resultados en cuanto a cantidad de información total recibida a lo largo de la red (véase figura 5-3). De estos siete algoritmos, ICM ha sido el que más información total ha logrado hacer llegar al conjunto de usuarios de la red, mientras que PageRank ha sido el que menos.

Este tipo de algoritmos podría usarse para encontrar usuarios a los que publicitar promociones o productos que sólo estén disponibles por un tiempo limitado, ya que interesa que la publicidad se propague rápido, aunque al final llegue a un menor número de usuarios.

A continuación analizamos a los algoritmos del grupo 2: más lentos pero mayor cantidad de información total al final de la simulación. En este grupo se encuentran los siguientes algoritmos: Adamic/Adar, FOAF, Jaccard y el algoritmo aleatorio. Las semillas obtenidas han sido aplicadas en todos los algoritmos como recomendación personalizada, a excepción del aleatorio que es no personalizada.

Se les denomina lentos porque necesitan más iteraciones para finalizar y en cada iteración propagan menos información (véase gráfica de la figura 5-5). El que más información propaga es Adamic/Adar, seguido por FOAF y Jaccard, pero éste último necesita menos iteraciones que Adamic (véase tabla 5-4). Al basarse estos tres algoritmos en los amigos comunes y el cierre triádico, parece lógico que obtengan resultados parecidos, aunque la técnica de Adamic/Adar de valorar los usuarios no populares le ha servido de ventaja frente a los otros dos. El que menos información propaga es el algoritmo aleatorio; sin embargo, FOAF es el que acaba la simulación con más iteraciones.

Aunque estos algoritmos han sido los que han propagado información de manera más lenta y con las simulaciones más largas, tienen de ventaja que al acabar han conseguido que la red reciba mayor cantidad de información que los algoritmos del grupo 1 como se puede observar en la gráfica de la figura 5-3. El que más información total ha acumulado es el aleatorio; esto es debido a que, al elegir las recomendaciones de manera aleatoria, los usuarios seleccionados están más dispersos y se puede llegar a más sitios de la red, a diferencia de los algoritmos del grupo 1, donde en todos el grado o el número de caminos juegan un papel importante, y de los otros algoritmos del grupo 2, que se centran más en las relaciones de amigos comunes. El que menos información ha acumulado es Jaccard, pero no ha sido el que ha tenido la simulación más larga de los cuatro.

Estos algoritmos podrían ser usados en campañas publicitarias donde fuera prioritaria la cantidad de usuarios que vieran el anuncio, sin importar el tiempo necesario para ello, ya que como se puede ver en las gráficas de las figuras 5-4 y 5-5 propagan la información a un ritmo más lento.

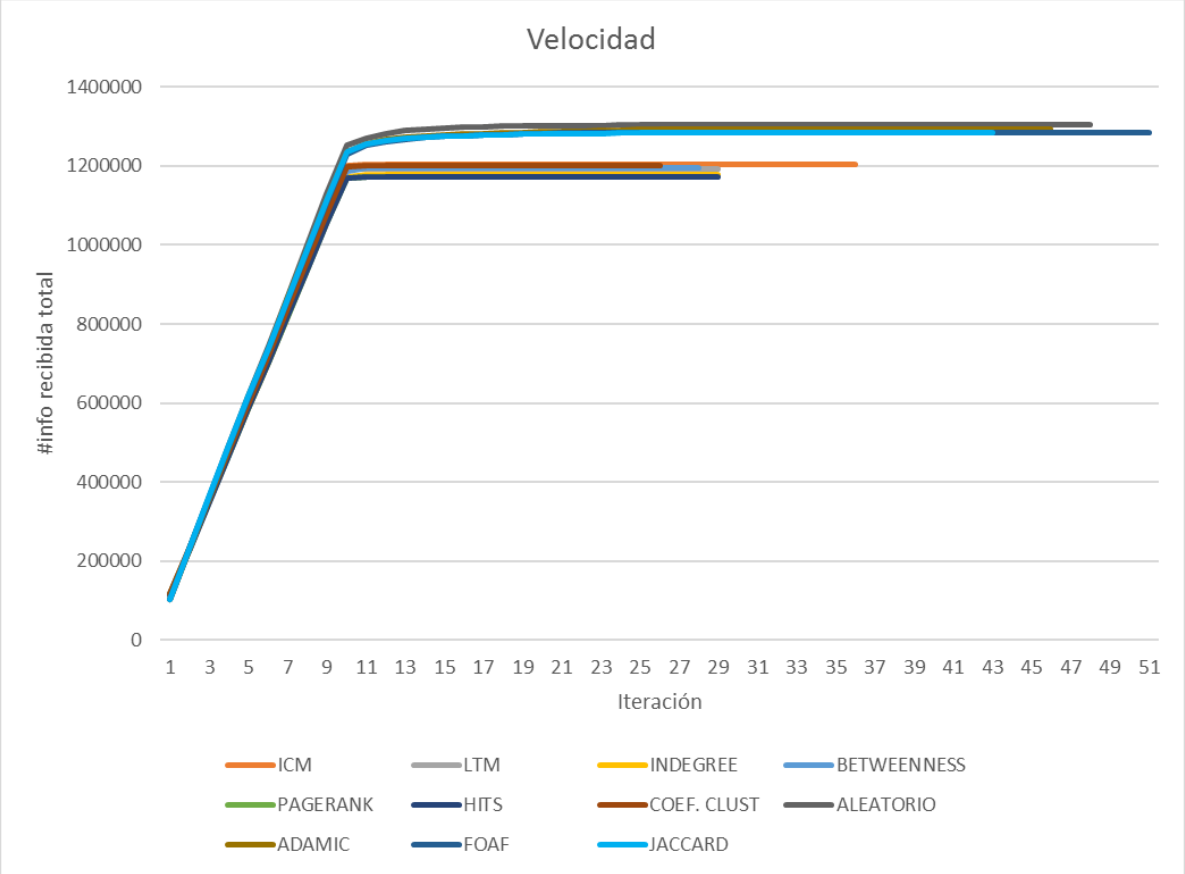


Figura 5-3: Gráfica resumen de la velocidad de propagación.

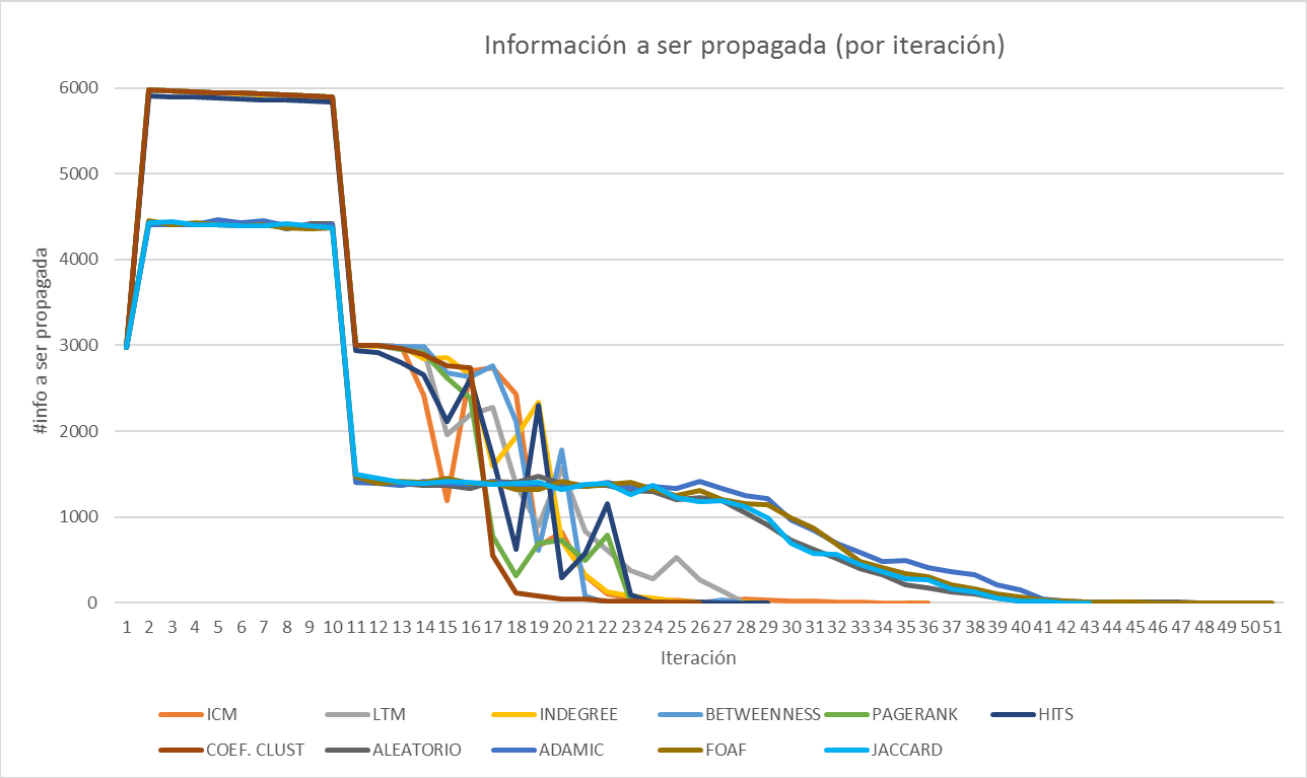


Figura 5-4: Gráfica resumen de la información a ser propagada por iteración.

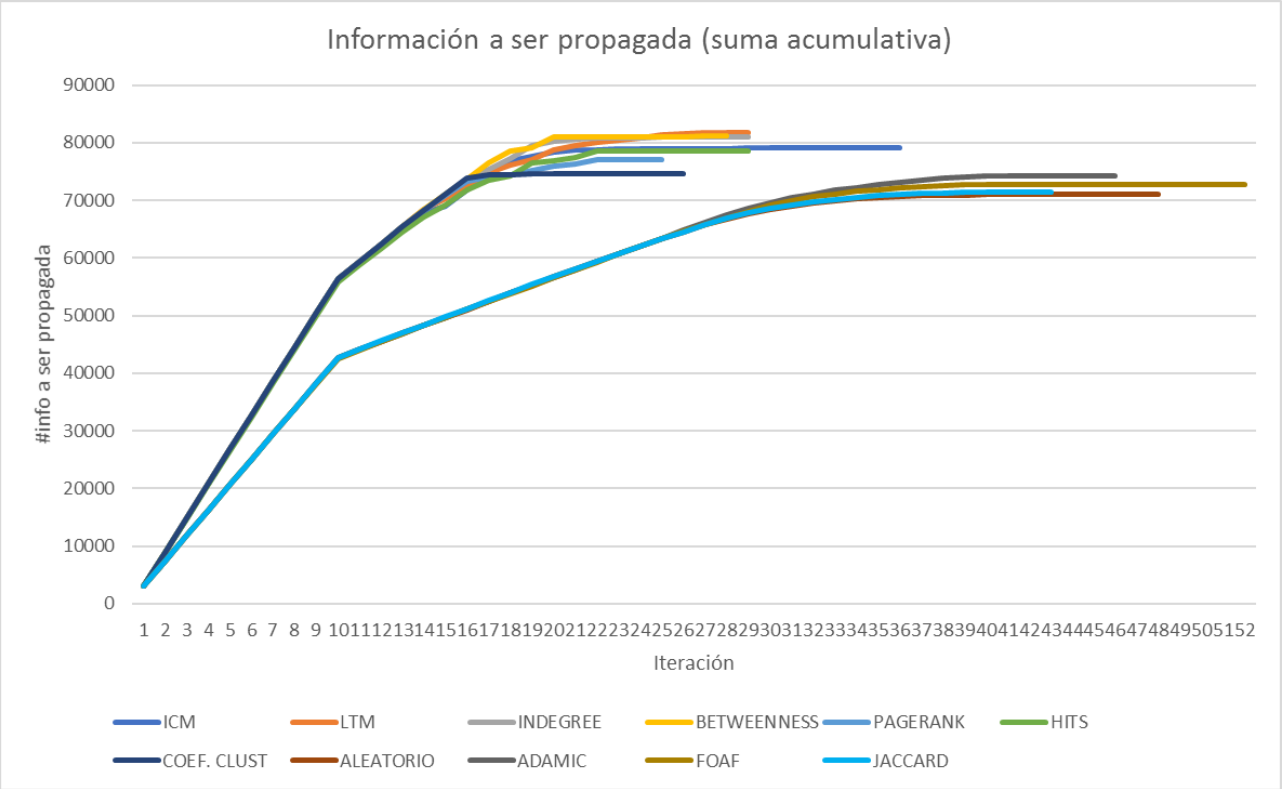


Figura 5-5: Gráfica resumen de la información a ser propagada (suma acumulativa).

Tabla 5-4: Iteraciones necesarias para acabar la simulación por algoritmo.

Algoritmo	ICM	LTM	Indegree	Betweenness	PageRank	HITS	Clust. Local	Aleatorio	Adami/Adar	FOAF	Jaccard
Iteraciones	36	29	29	28	25	29	26	48	46	51	43

Aparte de analizar el número de iteraciones necesarias y la cantidad de información recibida y propagada, también podemos analizar la cantidad de información nueva recibida de media para cada usuario. Como podemos ver en la gráfica de la figura 5-6, en las primeras iteraciones todos los algoritmos obtienen resultados parecidos y es a medida que avanza la simulación donde se aprecian más diferencias. Los algoritmos rápidos, los que hemos clasificado como grupo 1, tiene una caída más brusca de la cantidad de información nueva ya que finalizan antes, mientras que los algoritmos del grupo 2, como tienen más iteraciones, se observa cómo tienen una disminución de la cantidad más suave al final.

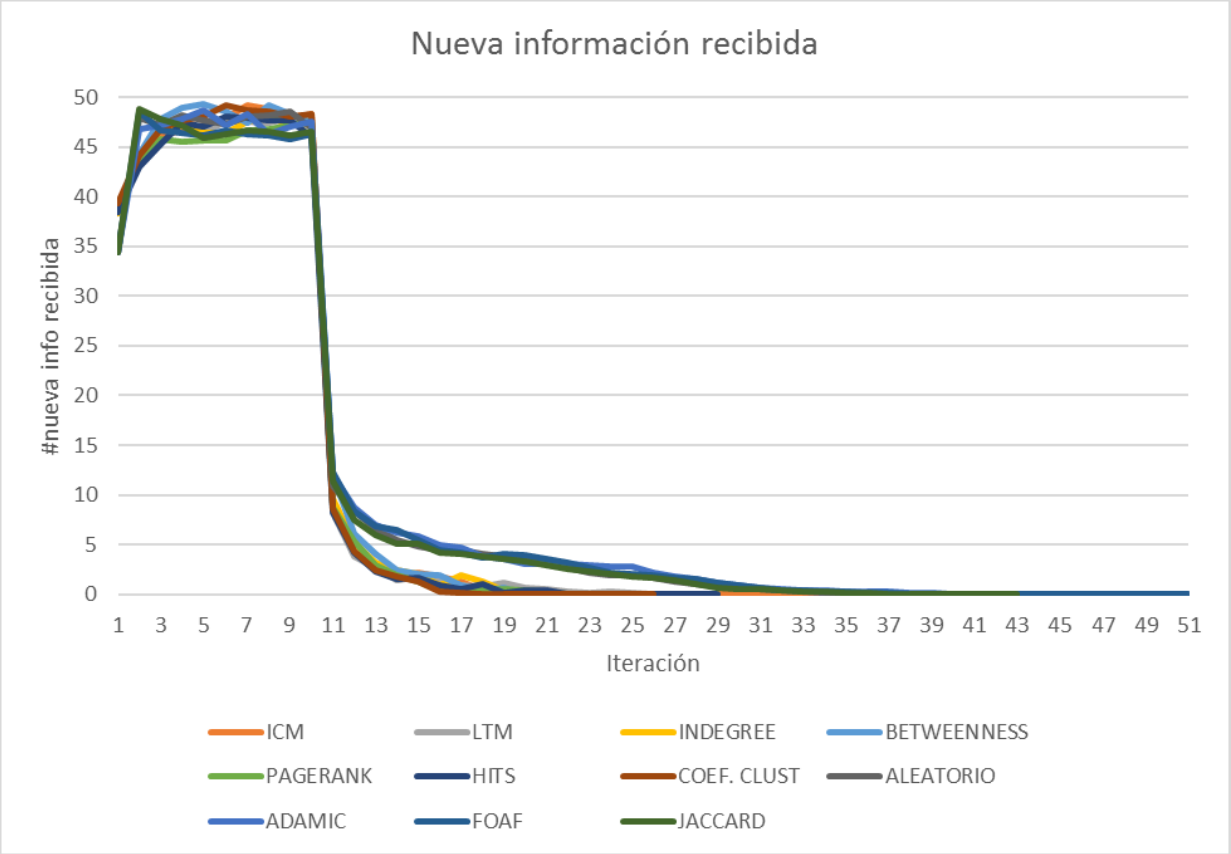


Figura 5-6: Gráfica resumen de la nueva información recibida por iteración.

Como resumen se podría decir que los algoritmos usados como recomendación no personalizada proporcionan una difusión más rápida que los usados como recomendación personalizada, pero estos últimos consiguen que la cantidad de información final que ha recibido la red sea mayor.

Si buscamos una difusión rápida destaca LTM y en cuanto cantidad de información final destacan el algoritmo aleatorio y el coeficiente de Adamic/Adar. Si quisiéramos un punto intermedio que tuviera ambas características, ICM aplicado como recomendación no personalizada sería el algoritmo apropiado.

6. Conclusiones y trabajo futuro

6.1 Conclusiones

Tanto la maximización de la difusión de la información como la recomendación de contactos son dos ámbitos muy presentes en las redes sociales, tanto en su análisis como en su aplicación en el mundo real. Las grandes compañías de servicios de redes sociales como Twitter buscan la maximización de la difusión de la información ya sea como una manera de ofrecer un mejor servicio hacia sus usuarios como desde un punto de vista de conseguir dinero mediante la publicidad, pero también hacen uso de la recomendación de contactos para mejorar la experiencia del usuario y lograr aumentar las dimensiones de la red. Por ello, este trabajo de fin de grado explora la posibilidad de usar algoritmos de maximización de la difusión de la información aplicados como recomendación de contactos para estudiar si se optimiza la difusión de información a través de una red social.

Para comprobar la efectividad de estos algoritmos se compararon los resultados contra los obtenidos aplicando métricas de topologías de redes como recomendación y contra algoritmos propios de recomendación de contactos personalizada. Se pueden analizar los resultados desde tres frentes distintos:

1. Los resultados de los propios modelos de maximización de la difusión.
2. Utilización en el marketing viral.
3. Aplicados a la recomendación de contactos.

En el primer frente, se probaron los dos algoritmos típicos del ámbito de la maximización de la difusión: el modelo de la cascada independiente (ICM) y la variación del modelo del umbral lineal (LTM). El segundo modelo destacó en las pruebas siendo el que más nodos activaba del grafo, sin embargo, el tiempo necesario para ejecutar el modelo es demasiado alto y sólo se pudo probar con un grafo pequeño, mientras que la cascada independiente no tuvo problemas para ser probada con cualquier grafo.

En el segundo frente, el marketing viral, se compararon los resultados que obtenían todos los algoritmos como seleccionadores de usuarios semilla para difundir el mismo mensaje y poder comprobar a cuantos nodos se propagaba. Ningún algoritmo obtuvo peor resultados que el aleatorio, a excepción del coeficiente de clustering local, cuya explicación podría ser que los valores de los nodos elegidos eran demasiado bajos. Aunque todos los algoritmos consiguieron llegar al mismo número de usuarios, destaca ICM por haber necesitado pocas iteraciones, por elegir usuarios semilla que no tienen un grado entrante excesivamente elevado y por ser el único que ha conseguido los mismos resultados con las mismas iteraciones pero con sólo tres semillas.

Por último, en la aplicación como recomendación de contactos, se apreció que los resultados de los algoritmos se podían clasificar como difusión rápida pero menor información total o difusión lenta pero mayor información total. Los dos algoritmos de difusión, usados como recomendación no personalizada, están incluidos en el primer tipo, por lo que, si los comparamos con el algoritmo aleatorio, ofrecen una difusión que acaba en pocas iteraciones y donde un mayor número de nodos propagan en cada iteración. Un buen uso de este tipo de red resultante sería para anunciar productos que están disponibles por un tiempo límite, ya que interesa que la información se propague lo más rápidamente posible.

En las pruebas destacaron LTM por ser el que más nodos propagaron en total (pulsos de información) e ICM por ser el que consiguió que más información recibiera la red de los algoritmos clasificados como rápidos.

Por tanto, se puede concluir que los algoritmos de maximización de la difusión aplicados como recomendación de contactos optimizan la difusión de información, especialmente desde el punto de vista de los pulsos de información.

6.2 Trabajo futuro

Aunque se ha demostrado en base a los resultados generales que estos algoritmos pueden ser aplicados como recomendación de contactos para la optimización de la difusión de la información, ambas aproximaciones codiciosas de estos algoritmos tendrían tiempos de ejecución enormes (NP-hard) si fueran aplicados a redes grandes. Kempe et al. ya experimentaron este hecho con redes de 30.000 nodos, ya que para encontrar 50 semillas necesitaban días [14]. Por tanto, un trabajo futuro sería encontrar versiones de estos algoritmos en la literatura que apliquen modificaciones para obtener algoritmos más óptimos.

Otro trabajo futuro sería probar estos algoritmos con redes más grandes, usando servidores y programación en paralelo para agilizar la búsqueda de semillas.

Por último, resultaría interesante realizar un estudio para averiguar qué porcentaje de las relaciones creadas mediante la recomendación se crean en la realidad para estudiar si tienen una componente sociológica o no.

Referencias

[1] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2010.

[2] M. E. J. Newman. Spread of epidemic disease on networks. Phys. Rev. E,66:016128, Jul 2002.

[3] M. Richardson and P. Domingos. "Mining Knowledge-Sharing Sites for Viral Marketing". Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, January 2002, pp. 61-70.

[4] G. Haralabopoulos, I. Anagnostopoulos (2015) On the Information Diffusion Between Web-Based Social Networks. In: Benatallah B. et al. (eds) Web Information Systems Engineering – WISE 2014 Workshops. WISE 2014. Lecture Notes in Computer Science, vol 9051. Springer, Cham.

[5] R. Zafarani, M. A. Abbasi, H. Liu, "Social Media Mining. An Introduction". Cambridge University Press. Abril 2014, pp. 217-251.

[6] R. Zafarani, M. A. Abbasi, H. Liu, "Social Media Mining. An Introduction". Cambridge University Press. Abril 2014, pp. 222-224.

[7] D. Kempe, J.M. Kleinberg, and ´E. Tardos, Maximizing the spread of influence through a social network, Proceedings of the ninth ACM SIGKDD international conference on Knowledge Discovery and Data Mining, ACM, 2003, pp. 137–146.

[8] Wikipedia. Betweenness Centrality. https://en.wikipedia.org/wiki/Betweenness_centrality

[9] Wikipedia. Closeness Centrality. https://en.wikipedia.org/wiki/Closeness_centrality

[10] Adamic, L. Adar, A. Friends and Neighbours on the Web. Social Networks Journal, 2003.

[11] Liben-Nowell, D., Kleinberg, J. The Link Prediction Problem for Social Networks, CIKM, 2003

[12] Page, L. & Brin, S., 1998. The anatomy of a large-scale hypertextual Web search engine. Brisbane, Australia, s.n.

[13] P. Patel, K. Patel. A Review of PageRank and HITS Algorithms. International Journal of Advance Research in Engineering, Science & Technology(IJAREST), ISSN(O):2393-9877, ISSN(P): 2394-2444, Volume 2, Issue 1, January- 2015.

[14] W. Chen, Y. Yuan, L. Zang. Scalable Influence Maximization in Social Networks under the Linear Threshold Model. ICDM 2010.

Glosario

Algoritmo codicioso Algoritmo que, para resolver un problema, elige la opción óptima en cada paso con la esperanza de obtener una solución general óptima.,16.

Arco Véase enlace., 2.

Arista Véase enlace., 9.

Enlace En teoría de grafos, es la relación entre dos vértices de un grafo., 2.

Epidemiología Parte de la medicina que estudia el desarrollo epidémico y la incidencia de las enfermedades infecciosas en la población., 1.

Facebook Es una corporación estadounidense y un sitio web de redes sociales con sede en Menlo Park, California., 1.

Grafo Conjunto de objetos denominados vértices o nodos unidos por enlaces llamados aristas o arcos que representan las relaciones binarias entre elementos de un conjunto., 1.

Grafo dirigido Es un grafo donde las aristas tienen un sentido definido., 1.

Grafo no dirigido Es un grafo donde las aristas no tienen sentido, son relaciones simétricas., 1.

Inmunidad Estado de resistencia, natural o adquirida, que poseen ciertos individuos o especies frente a determinadas acciones patógenas de microorganismos o sustancias extrañas., 6.

Marketing Conjunto de principios y prácticas que buscan el aumento del comercio, especialmente de la demanda., 1.

Nodo Es la unidad fundamental de la que están formados los grafos., 2.

NP Acrónimo en inglés de “nondeterministic polinomial time” (“tiempo polinomial no determinista”). Es el conjunto de problemas que pueden ser resueltos en tiempo polinómico por una máquina de Turing no determinista., 35.

NP-hard Un problema que es al menos tan complejo como NP (pero no necesariamente en NP)., 16.

Red social Plataforma digital de comunicación global que pone en contacto a gran número de usuarios., 1.

Similitud Semejanza., 11.

Sociología Ciencia que trata de la estructura y funcionamiento de las sociedades humanas., 1.

Tuit Denominado en inglés tweet, mensaje digital que se envía a través de la red social Twitter y que no puede rebasar un número limitado de caracteres., 15.

Twitter Es un servicio de noticias en línea y redes sociales donde los usuarios publican e interactúan con mensajes, "tuits", restringidos a 140 caracteres., 1.

Usuario semilla En maximización de difusión e influencias, usuario que es seleccionado debido a ciertas características que lo clasifican como valioso para maximizar la difusión o propagar influencias., 2.

Vértice Véase nodo., 35.

Anexos

A Gráficas de las simulaciones de marketing viral

Gráficas de la velocidad (cantidad de información recibida total) por algoritmo. K es el número de semillas utilizado.

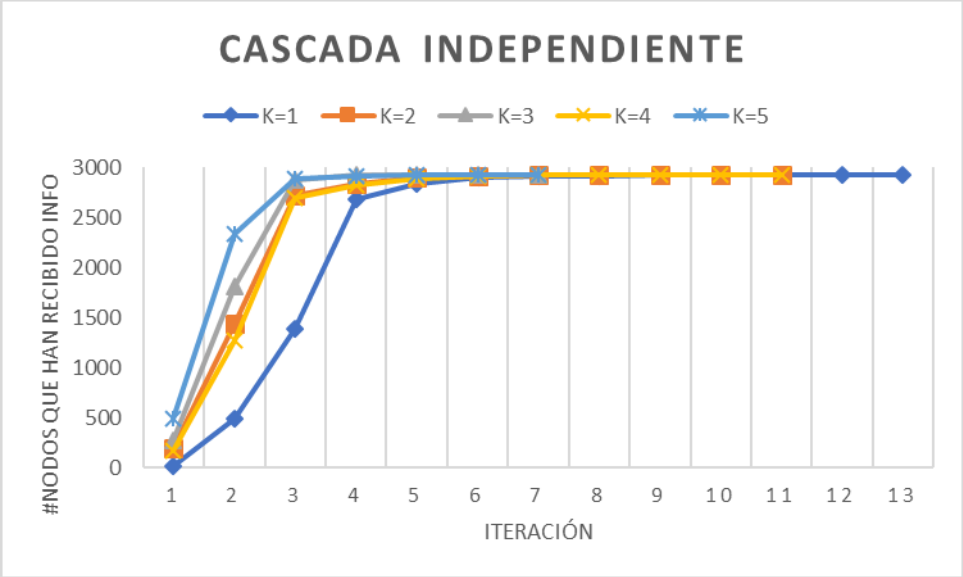


Figura A-1: Gráfica de la velocidad para ICM.

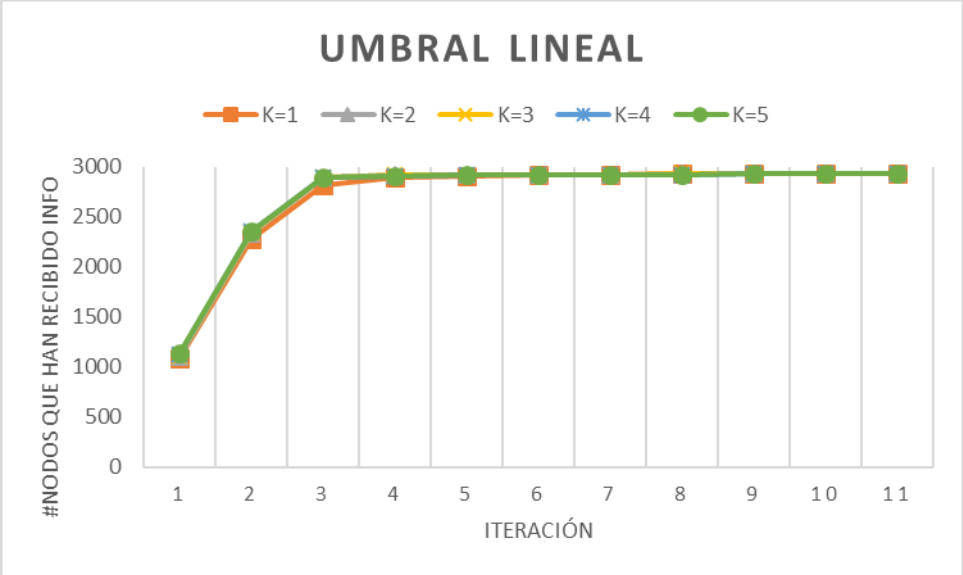


Figura A-2: Gráfica de la velocidad para LTM.

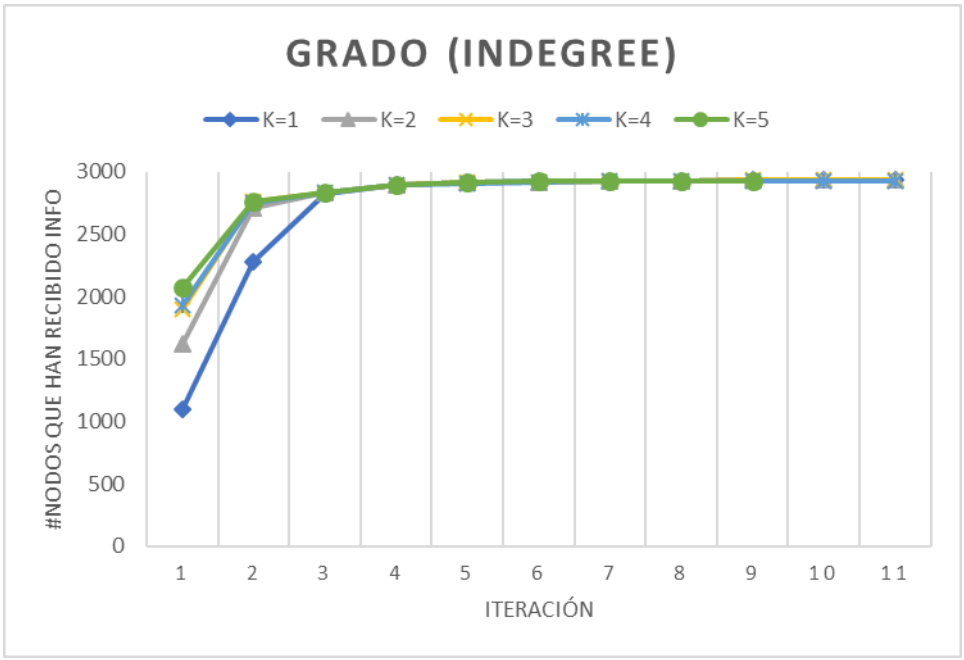


Figura A-3: Gráfica de la velocidad para indegree.

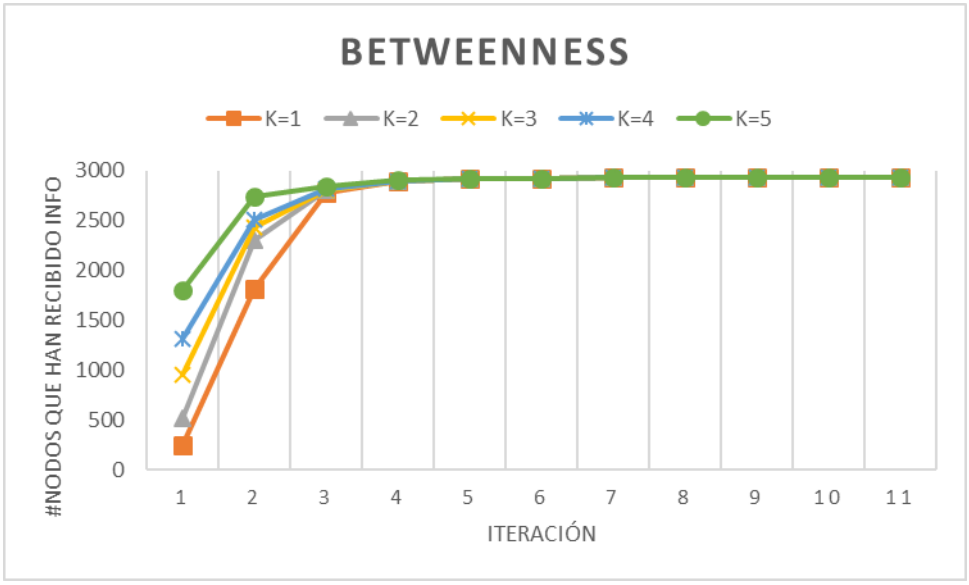


Figura A-4: Gráfica de la velocidad para betweenness.

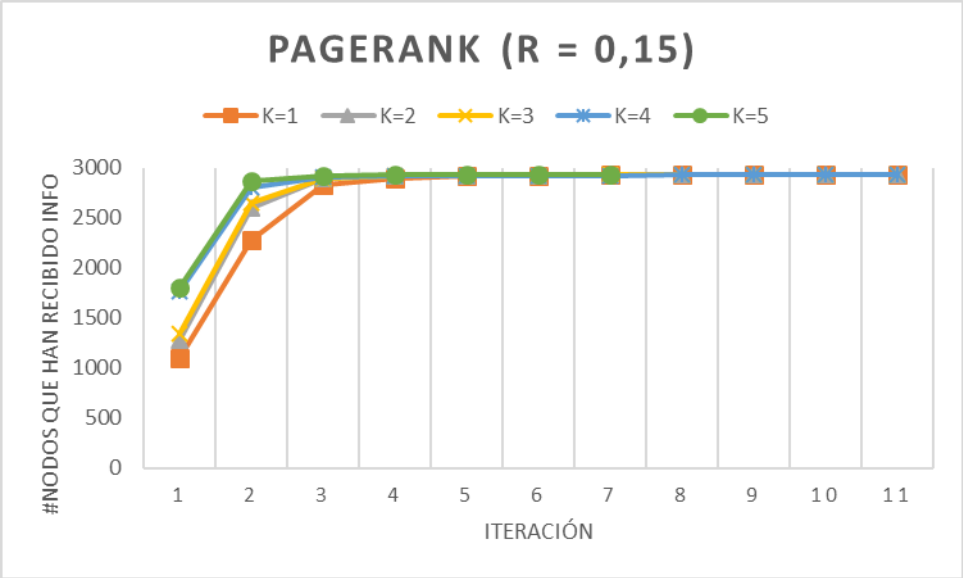


Figura A-5: Gráfica de la velocidad para PageRank.

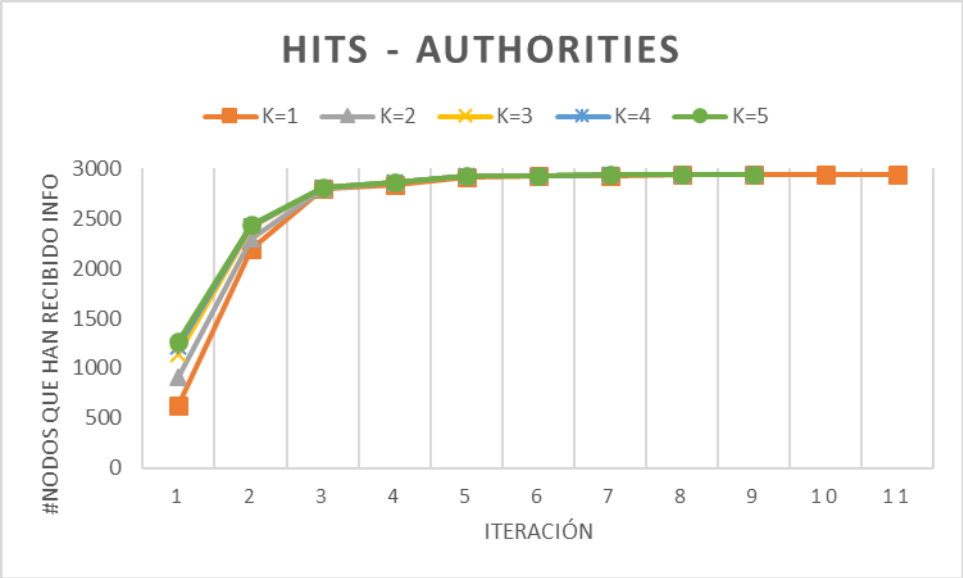


Figura A-6: Gráfica de la velocidad para HITS (autoridades).

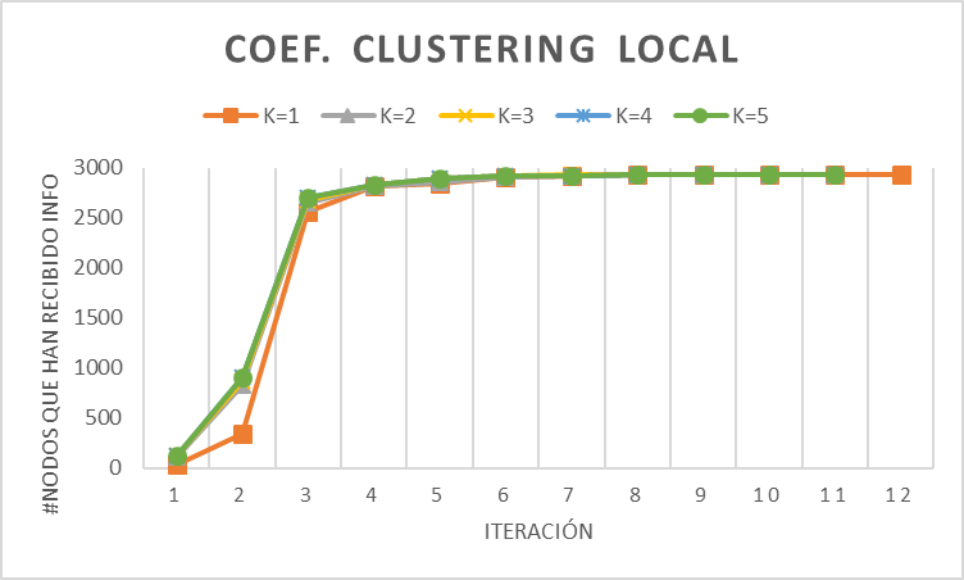


Figura A-7: Gráfica de la velocidad para coef. clustering local.

Figura A-8: Gráfica de la velocidad para el algoritmo aleatorio.

Gráficas de la propagación (suma acumulativa) por algoritmo. K significa el número de semillas utilizado.

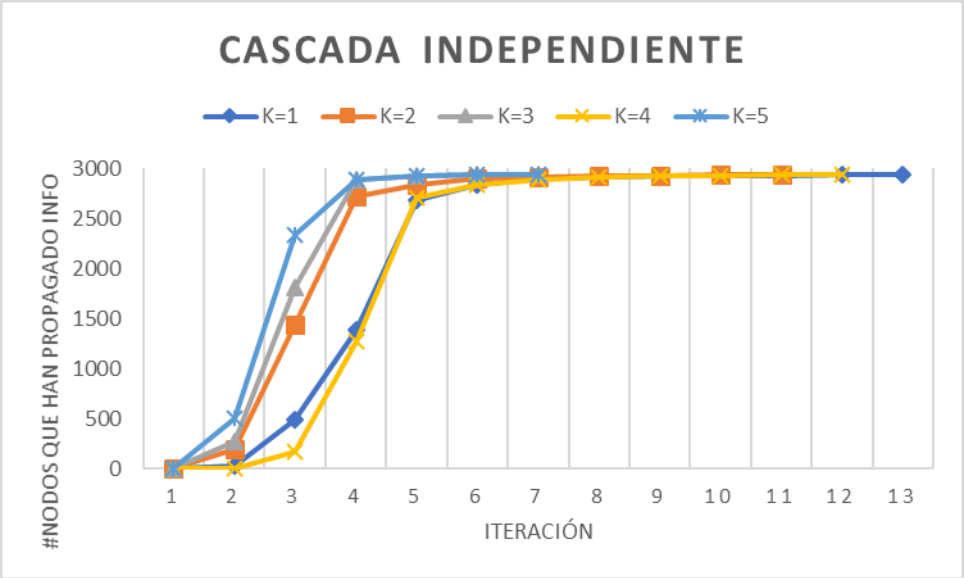


Figura A-9: Gráfica de la propagación para ICM.

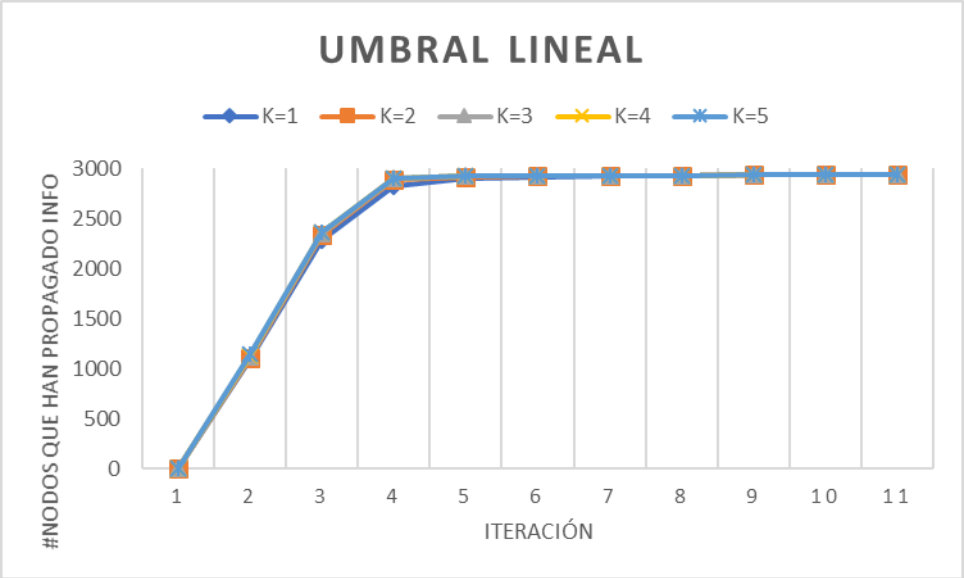


Figura A-10: Gráfica de la propagación para LTM.

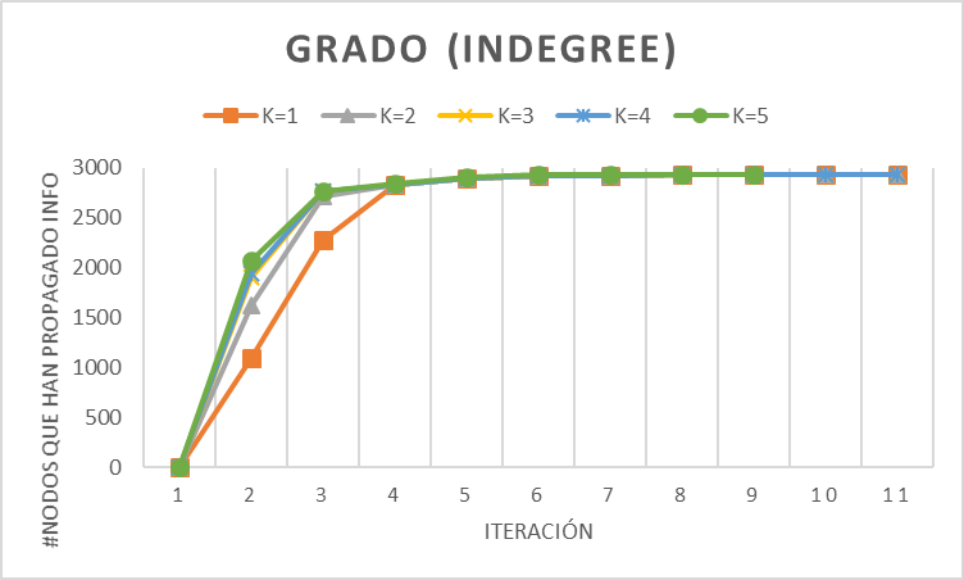


Figura A-11: Gráfica de la propagación para indegree.

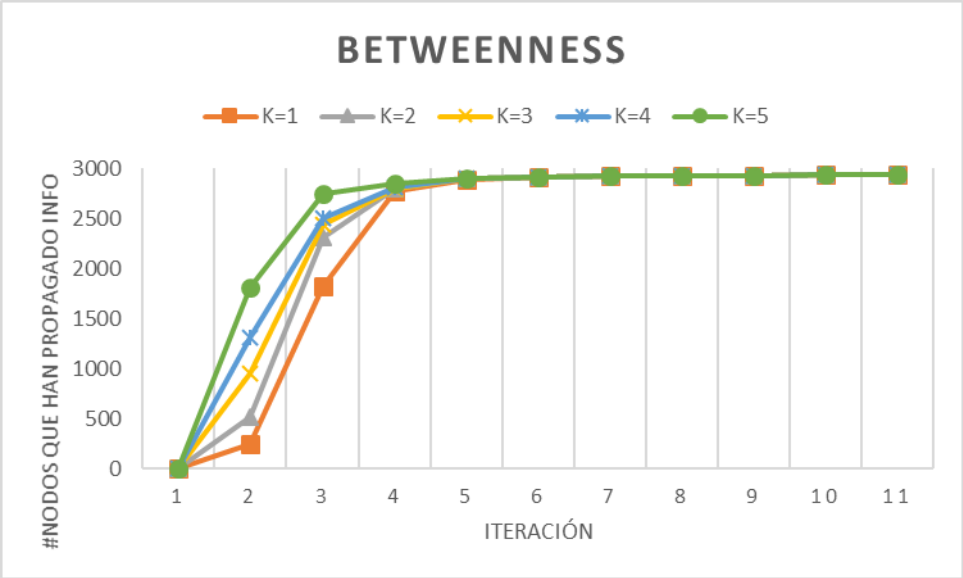


Figura A-12: Gráfica de la propagación para betweenness.

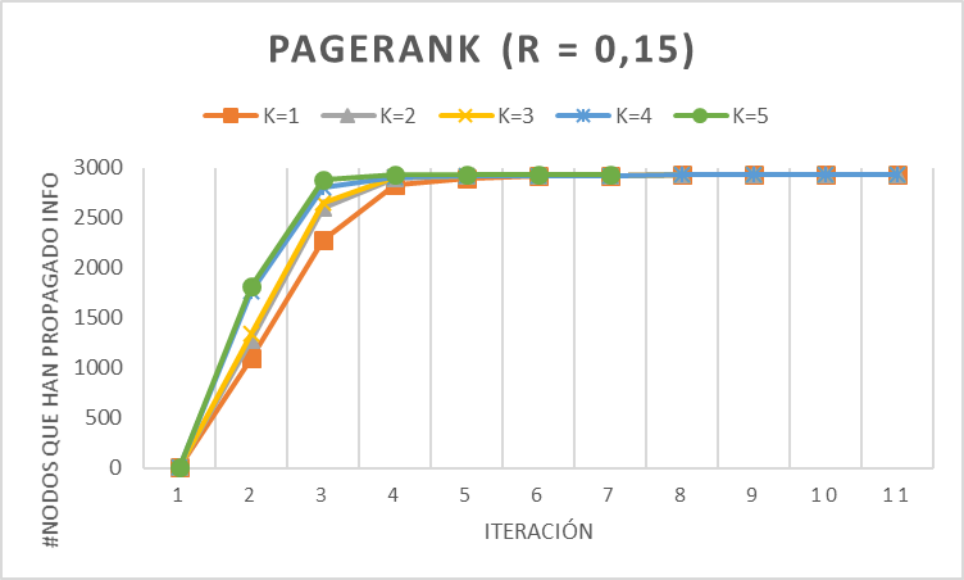


Figura A-13: Gráfica de la propagación para PageRank.

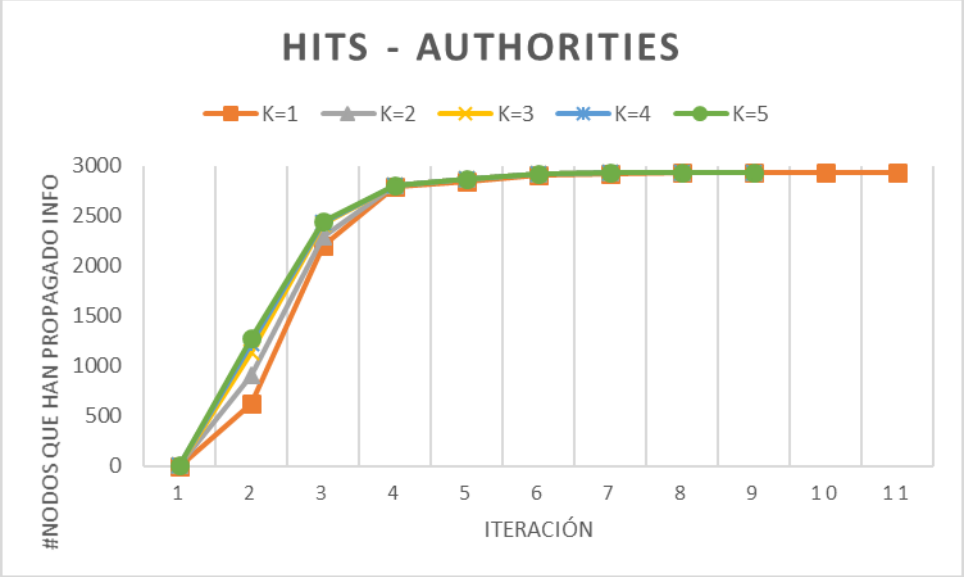


Figura A-14: Gráfica de la propagación para HITS (autoridades).

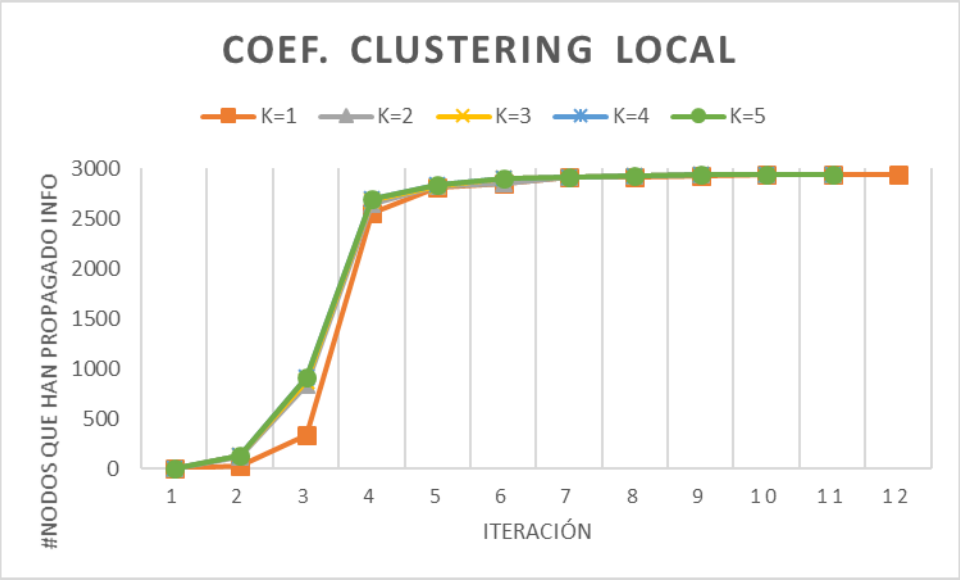


Figura A-15: Gráfica de la propagación para coef. clustering local.

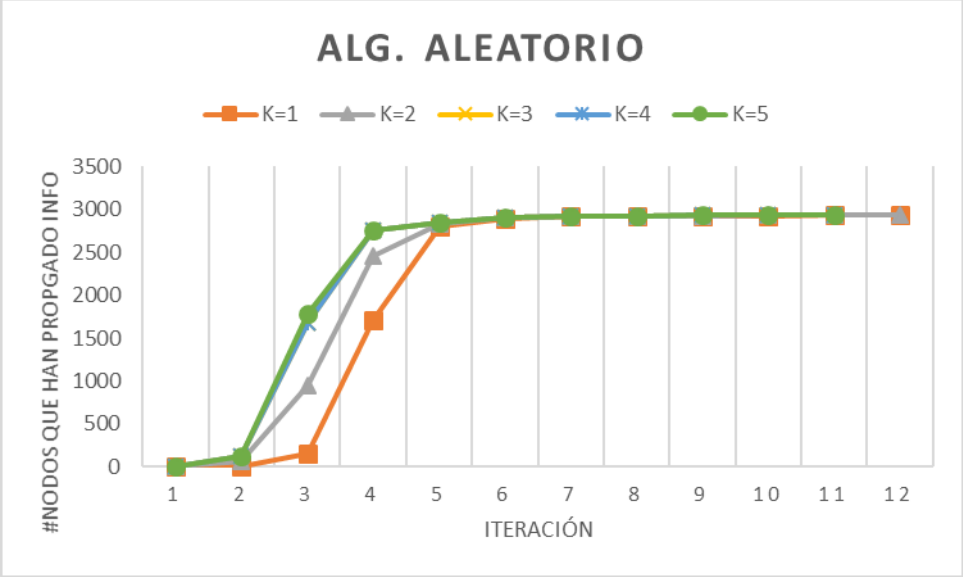


Figura A-16: Gráfica de la propagación para el algoritmo aleatorio.

B Gráficas de simulaciones con recomendación no personalizada

K es el número de usuarios recomendados a cada usuario en esa simulación.

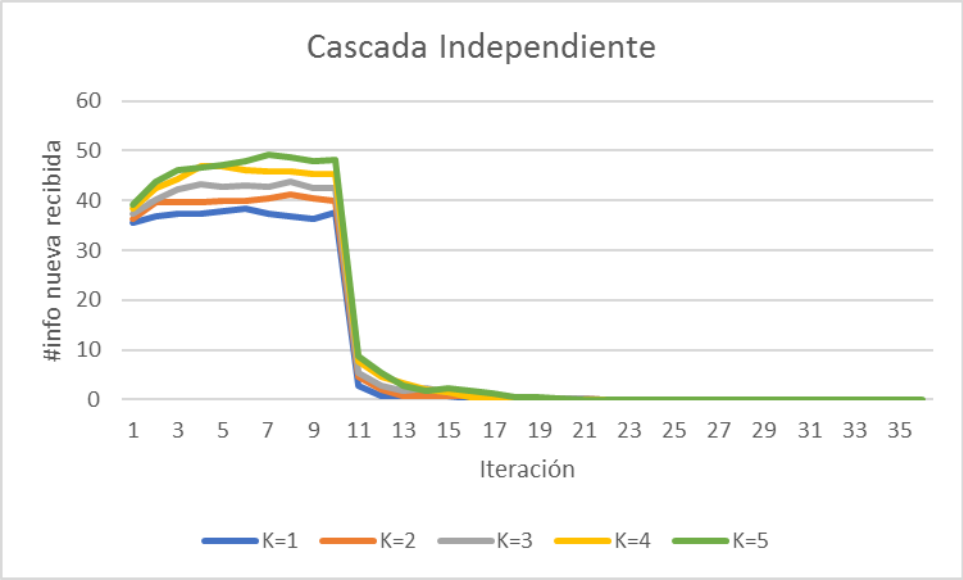


Figura B-1: Nueva información con ICM.

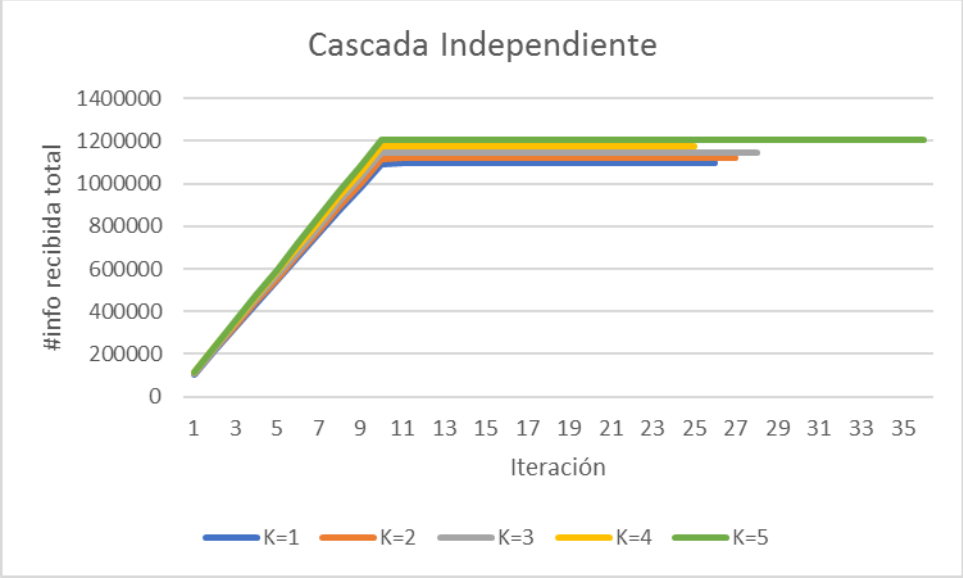


Figura B-2: Velocidad con ICM.

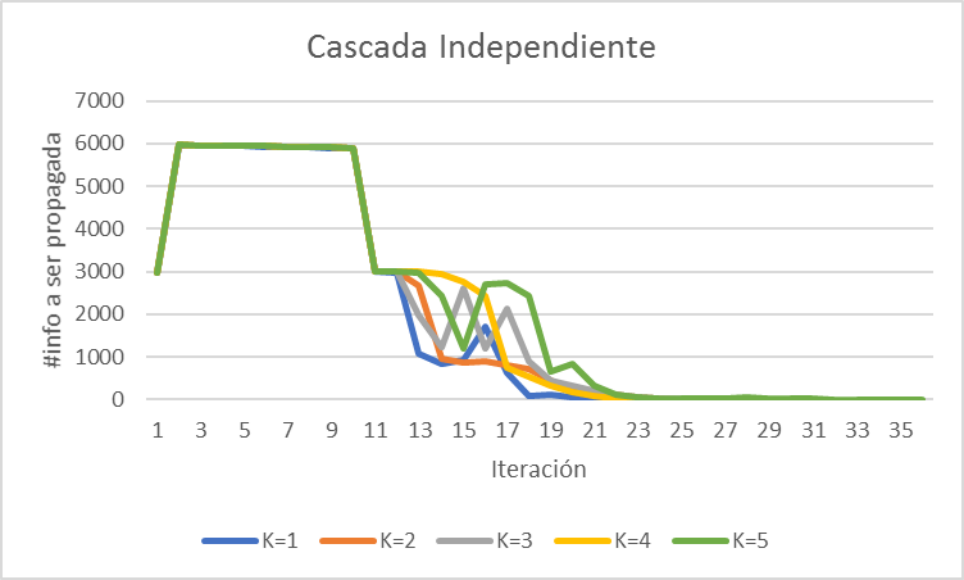


Figura B-3: Propagación con ICM.

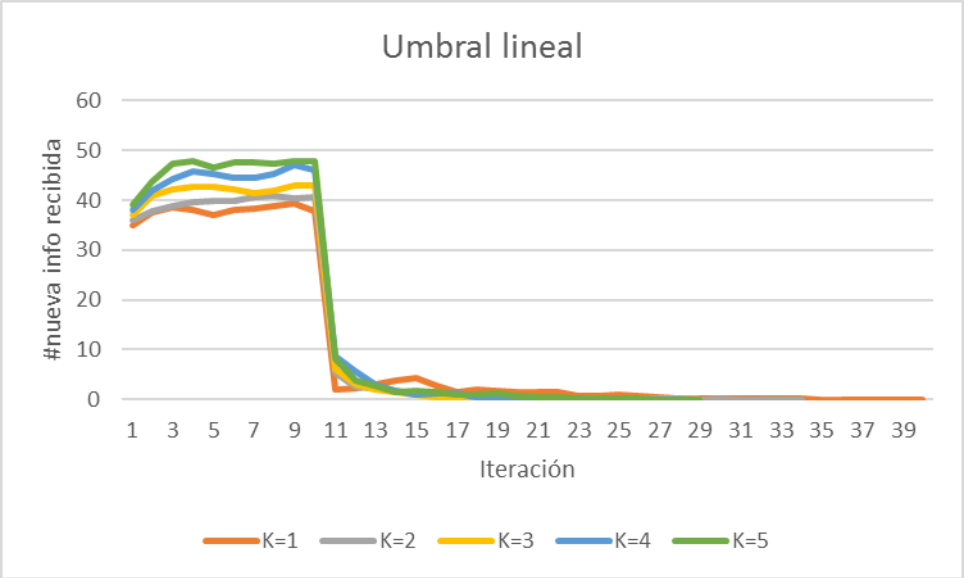


Figura B-4: Nueva información con LTM

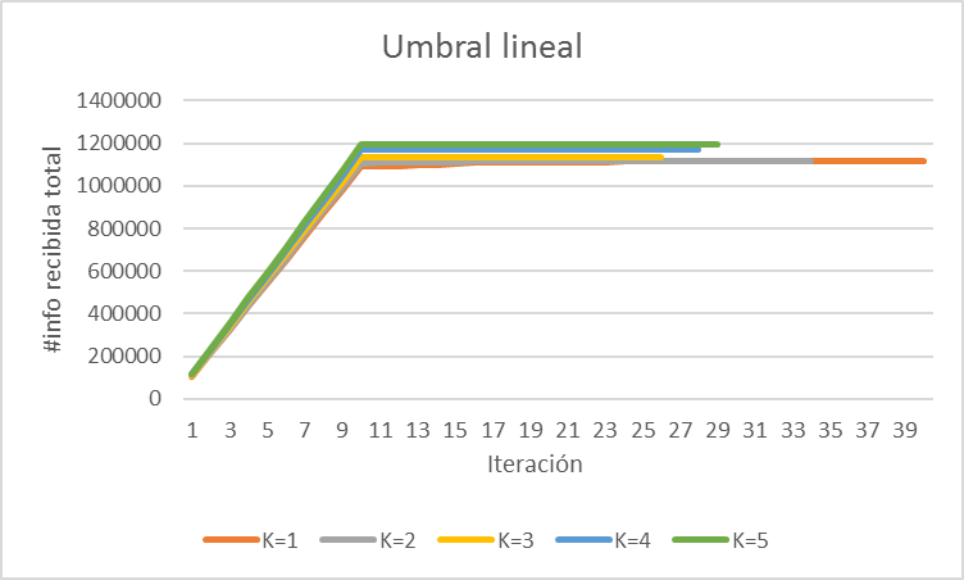


Figura B-5: Velocidad con LTM.

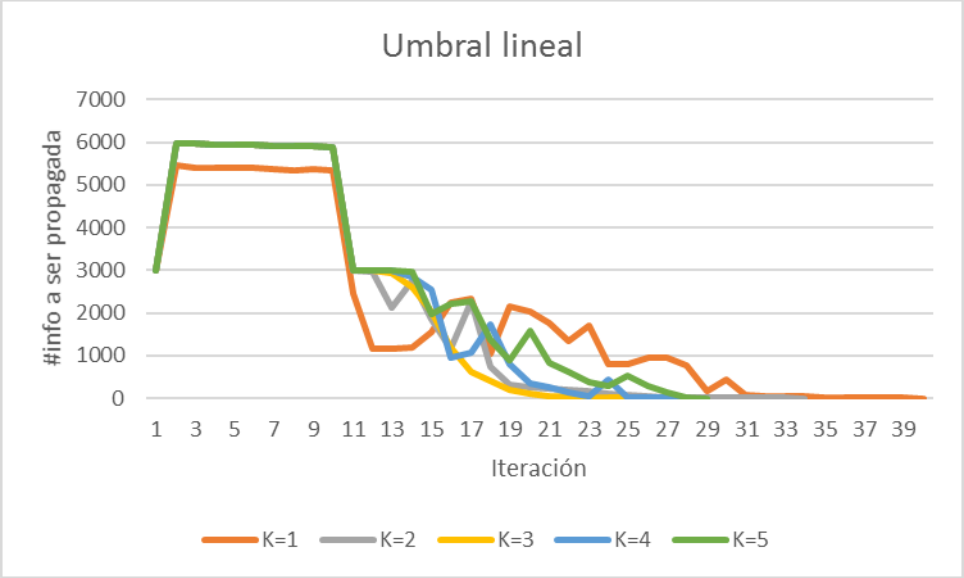


Figura B-6: Propagación con LTM.

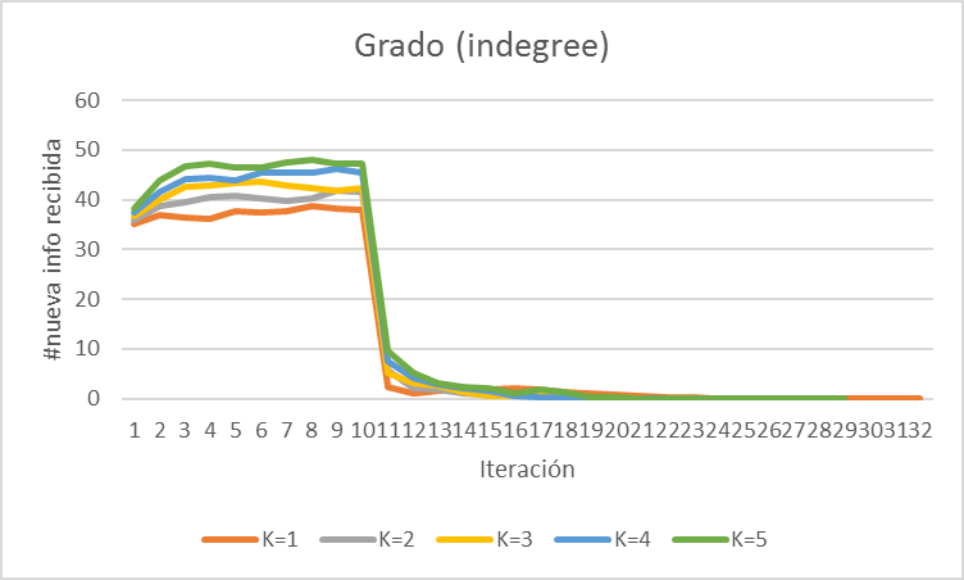


Figura B-7: Nueva información con indegree.

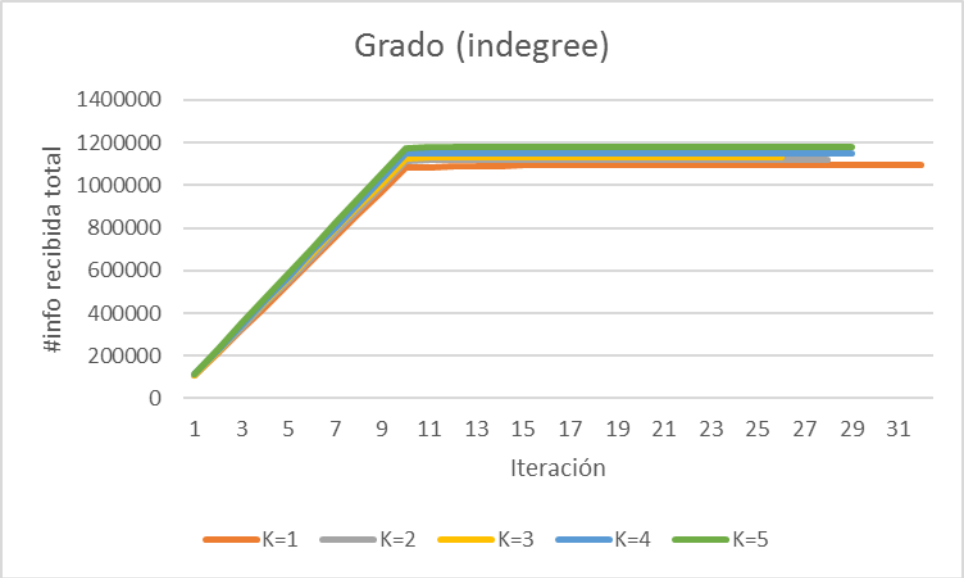


Figura B-8: Velocidad con indegree.

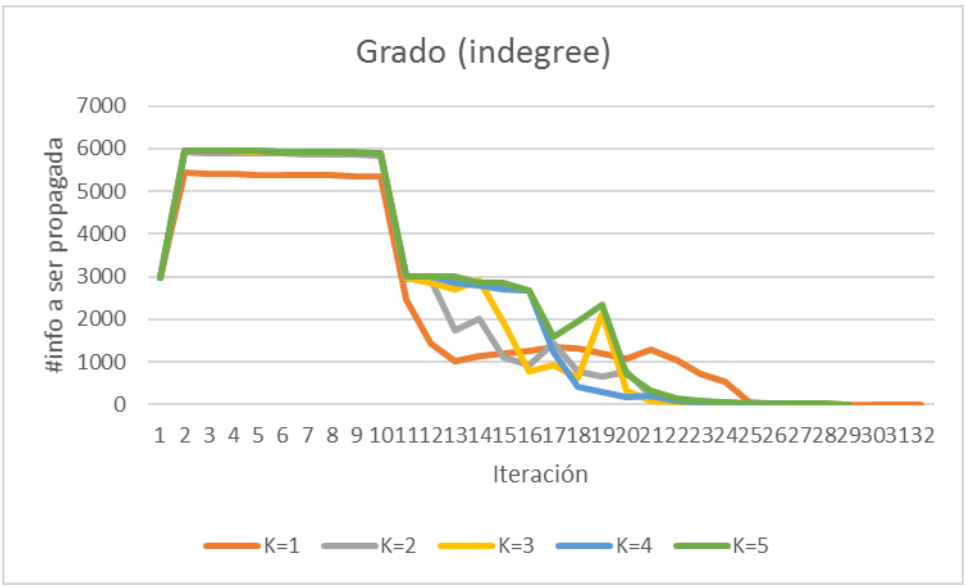


Figura B-9: Propagación con indegree.

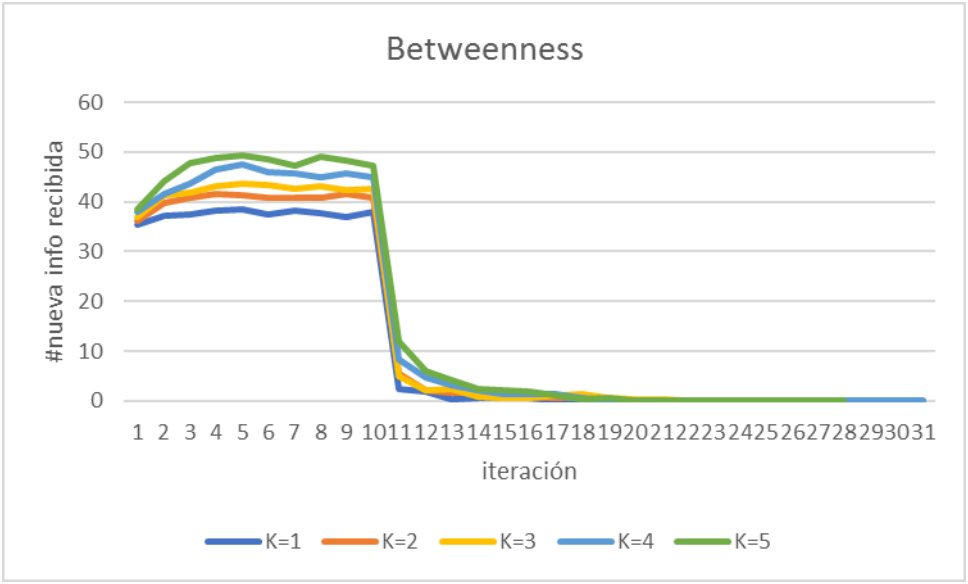


Figura B-10: Nueva información con betweenness.

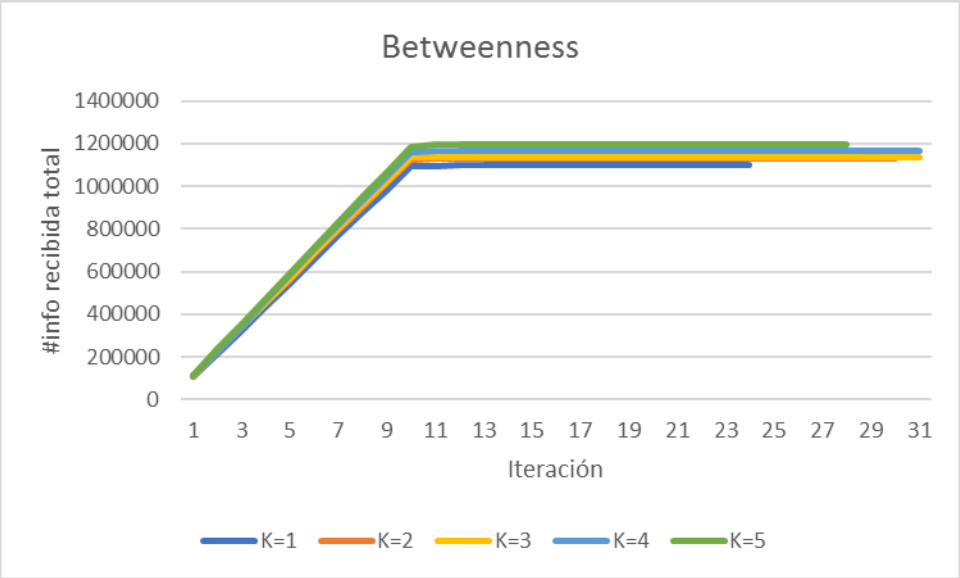


Figura B-11: Velocidad con betweenness.

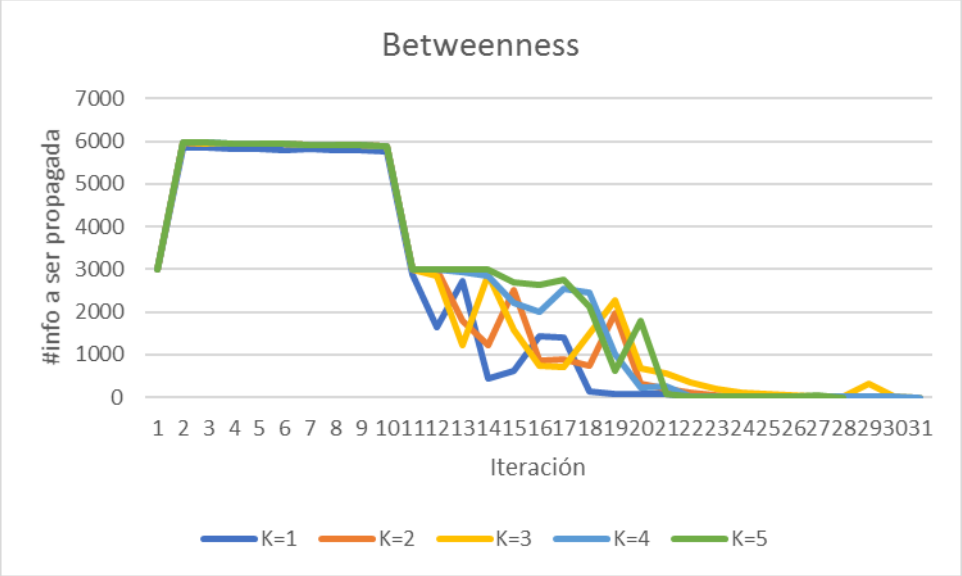


Figura B-12: Propagación con betweenness.

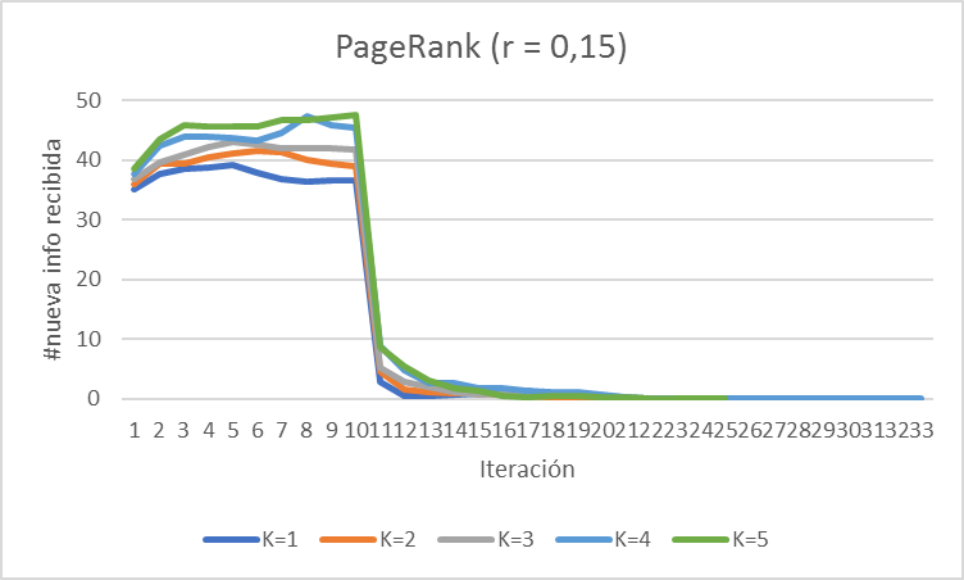


Figura B-13: Nueva información con PageRank.

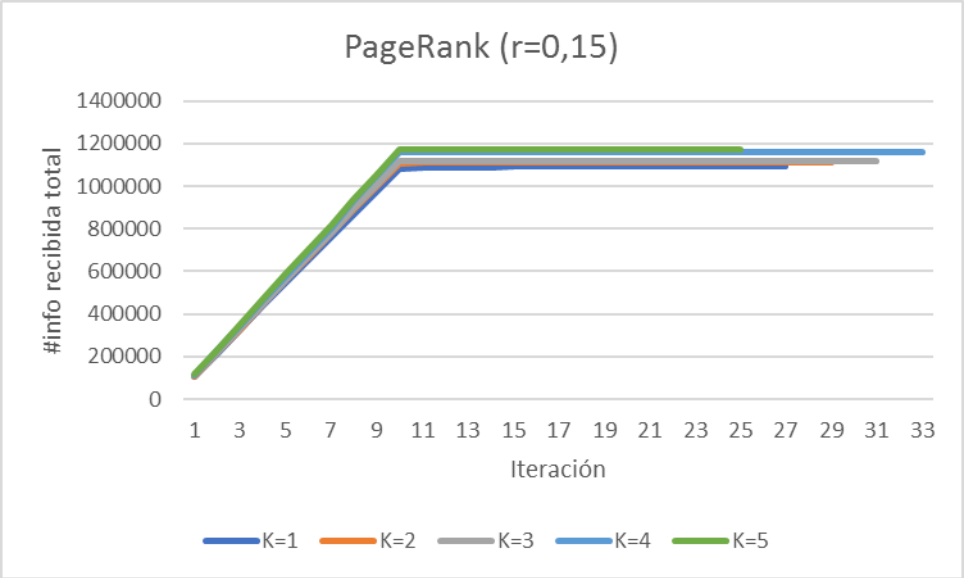


Figura B-14: Velocidad con PageRank.

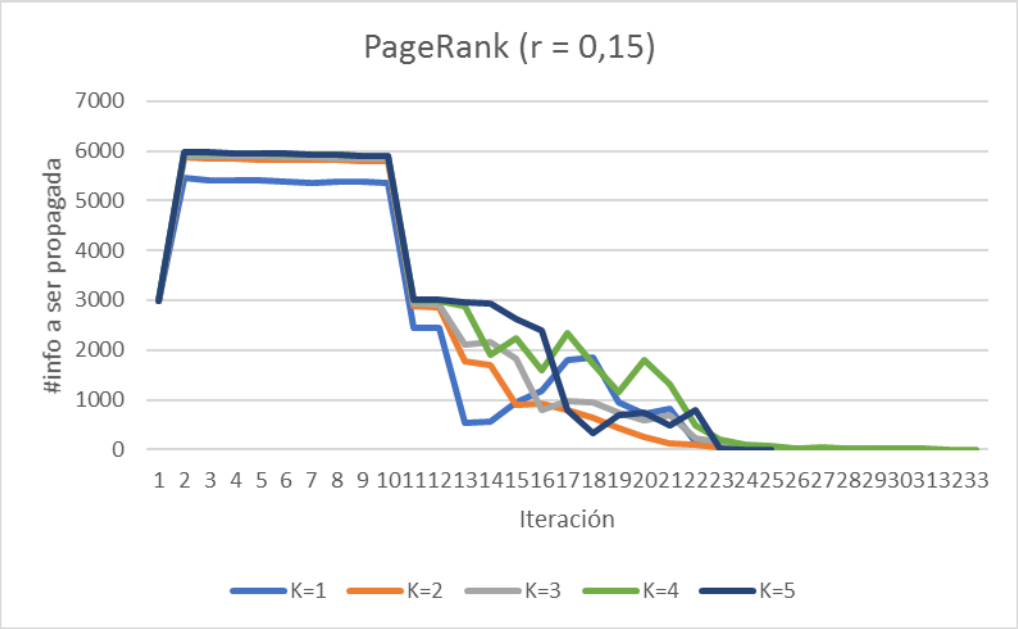


Figura B-15: Propagación con PageRank.

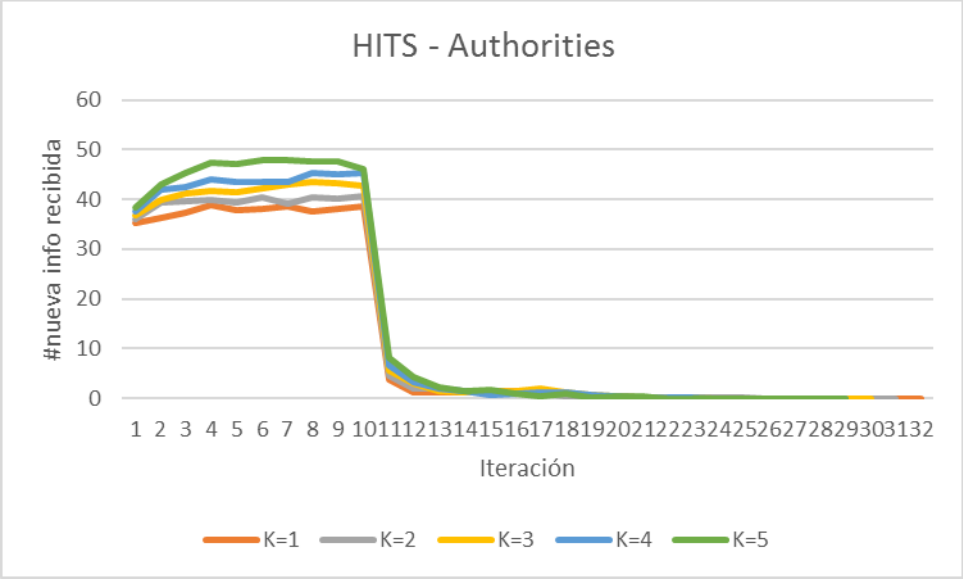


Figura B-16: Nueva información con HITS (autoridades).

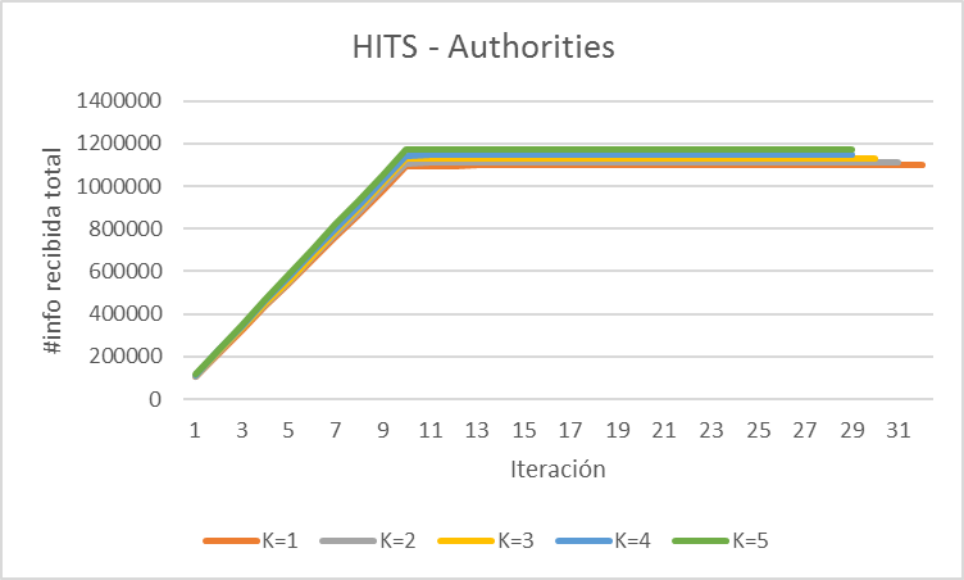


Figura B-17: Velocidad con HITS (autoridades).

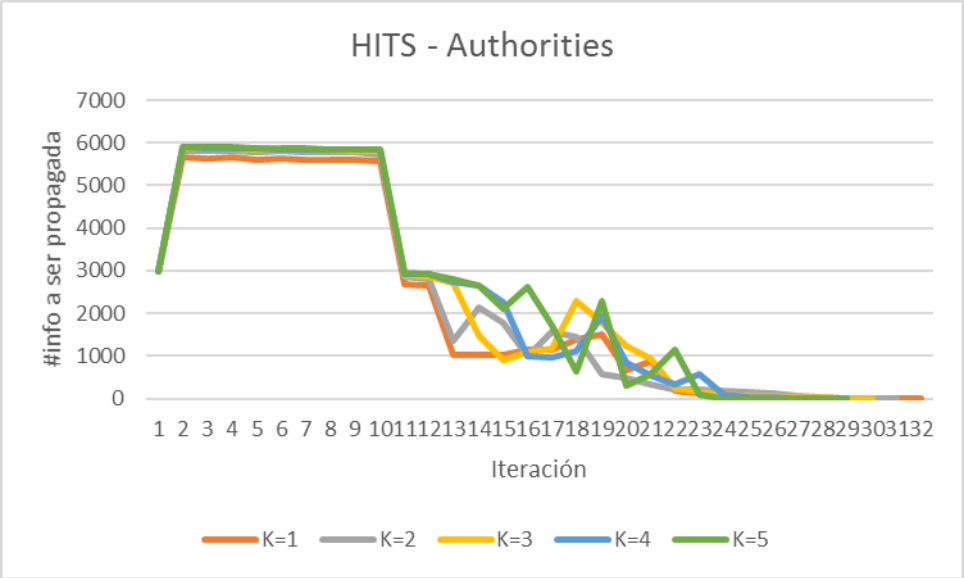


Figura B-18: Propagación con HITS (autoridades).

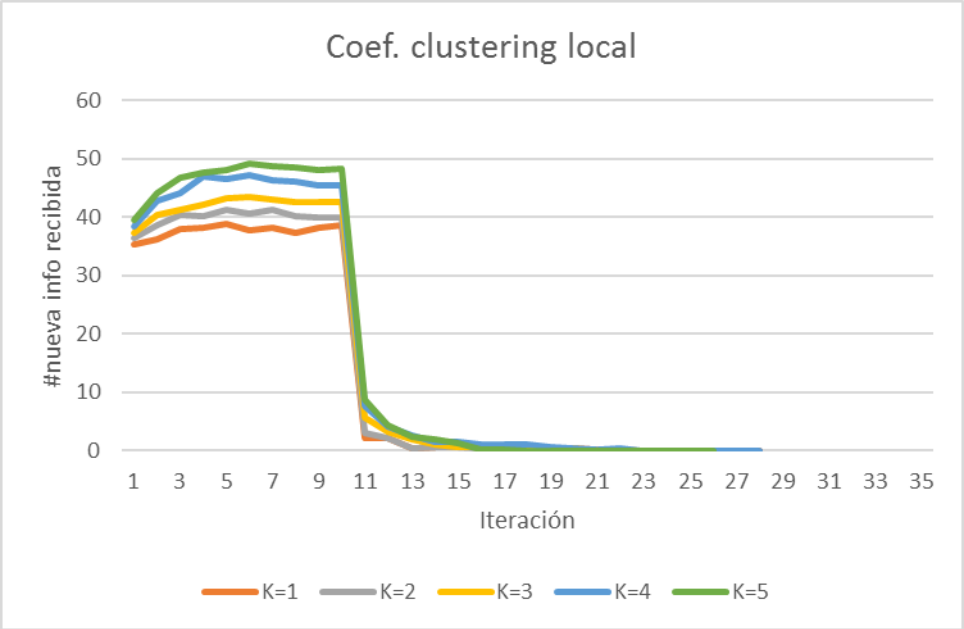


Figura B-19: Nueva información con coef. clustering local.

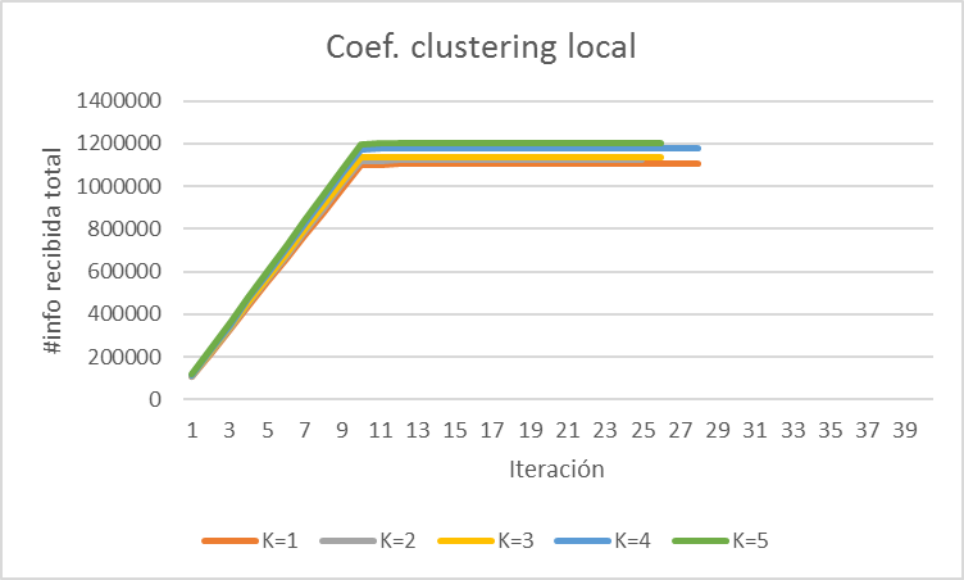


Figura B-20: Velocidad con coef. clustering local.

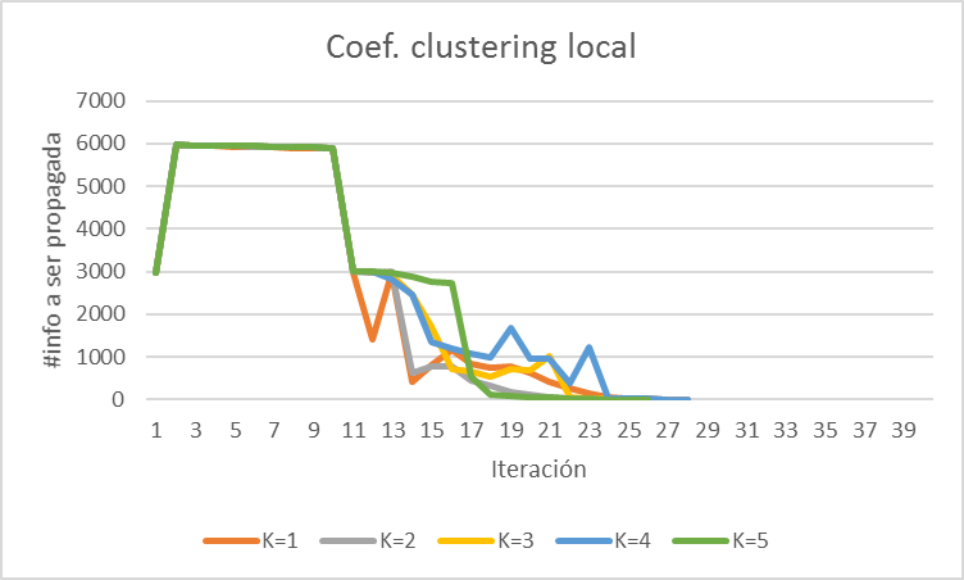


Figura B-21: Propagación con coef. clustering local.

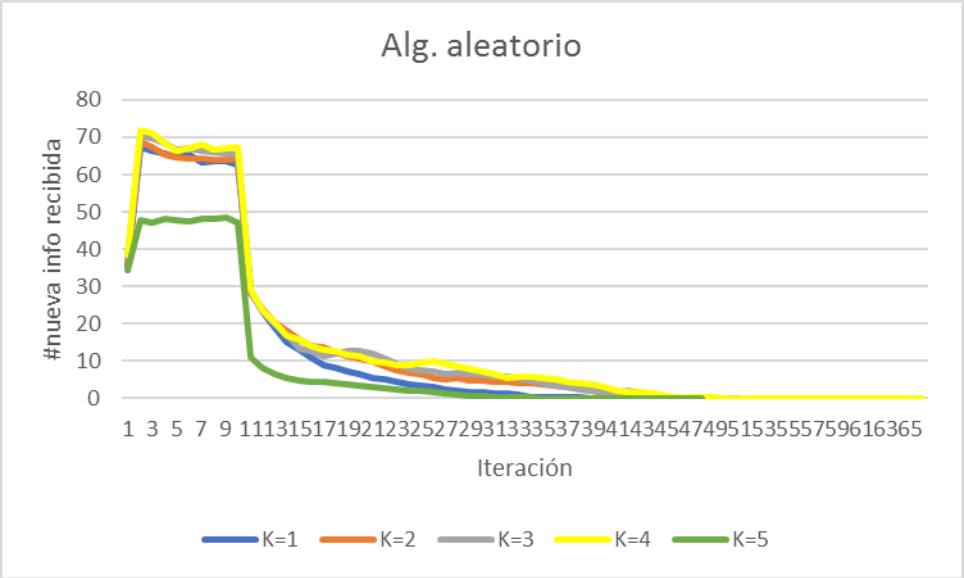


Figura B-22: Nueva información con el algoritmo aleatorio.

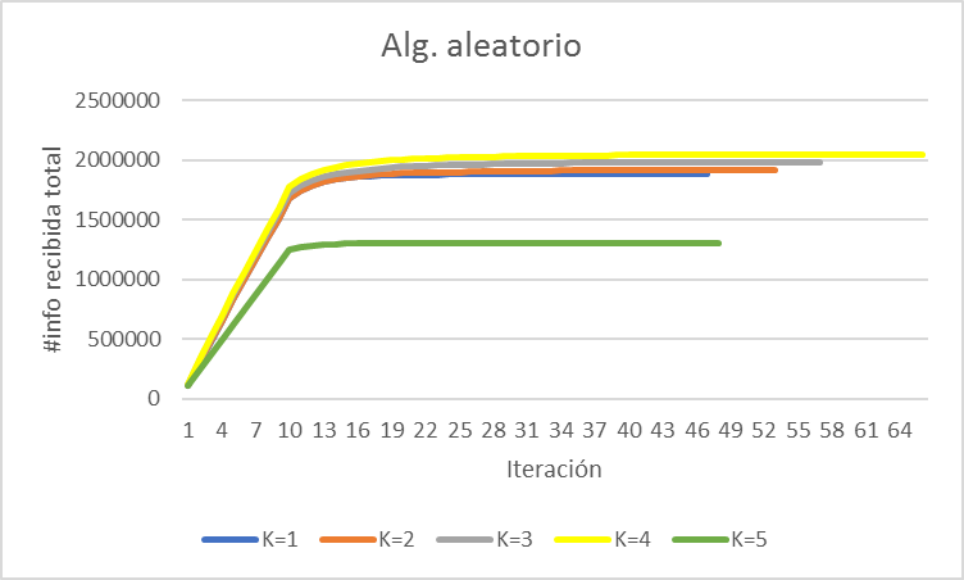


Figura B-23: Velocidad con el algoritmo aleatorio.

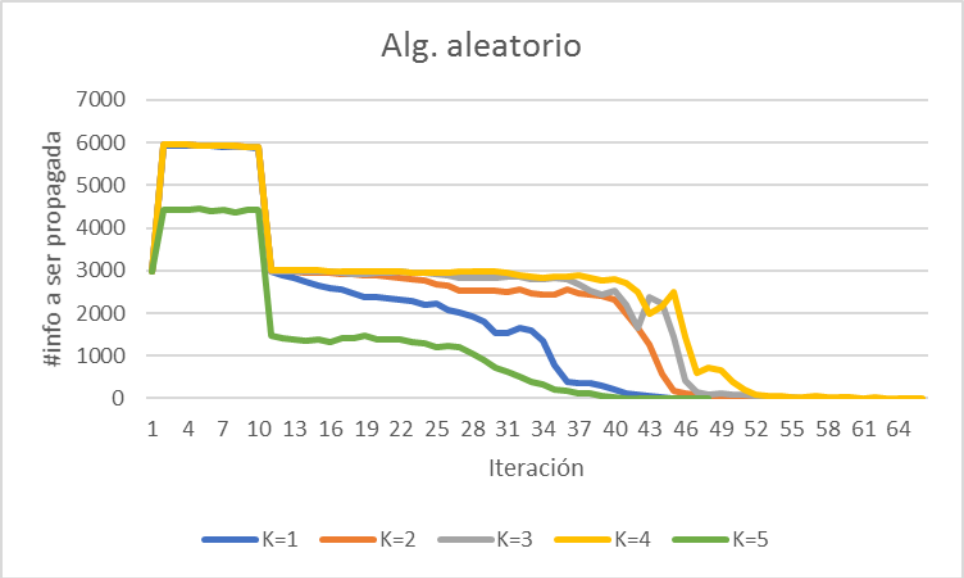


Figura B-24: Propagación con el algoritmo aleatorio.

C Gráficas de las simulaciones con recomendación personalizada

K es el número de usuarios recomendados a cada usuario en la simulación.

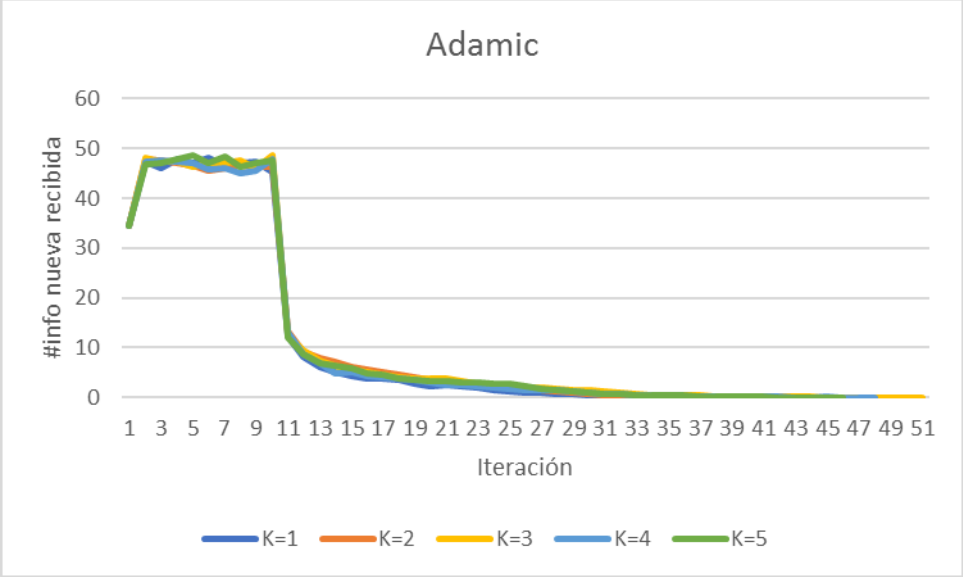


Figura C-1: Nueva información con Adamic/Adar.

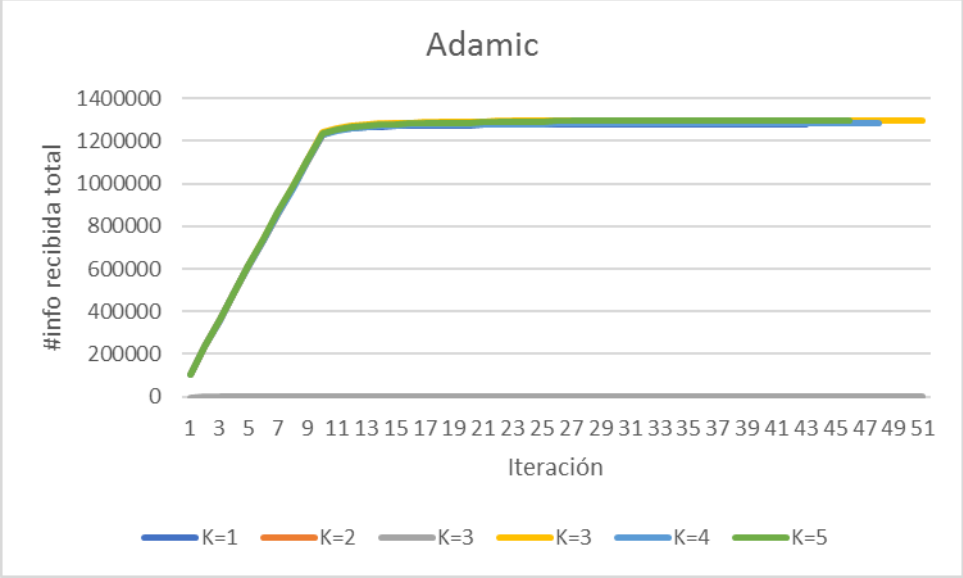


Figura C-2: Velocidad con Adamic/Adar.

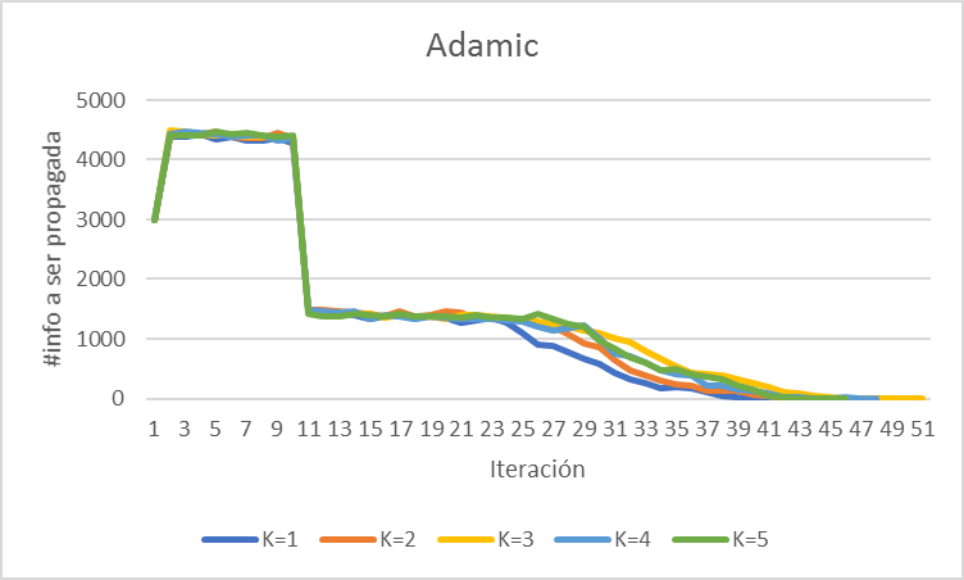


Figura C-3: Propagación con Adamic/Adar.

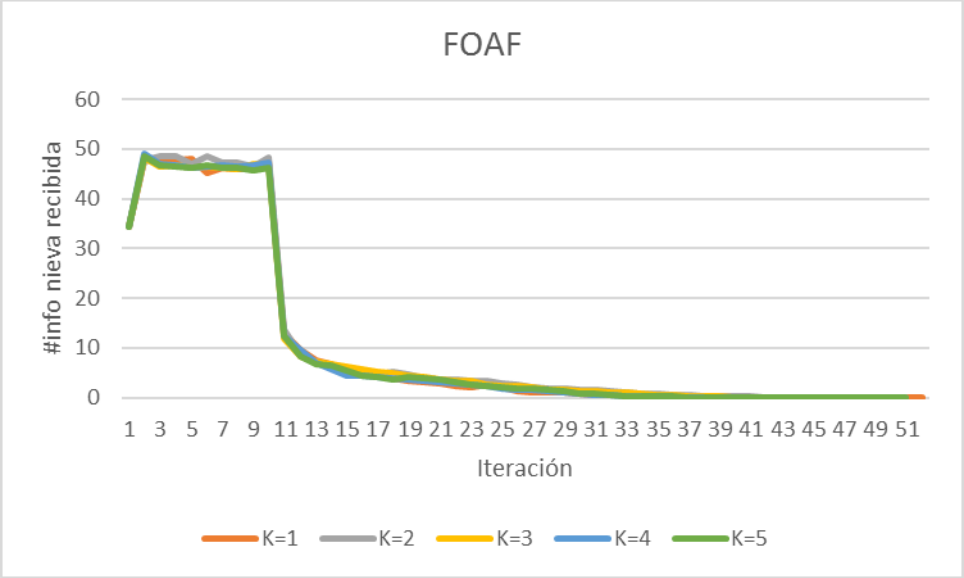


Figura C-4: Nueva información con FOAF.

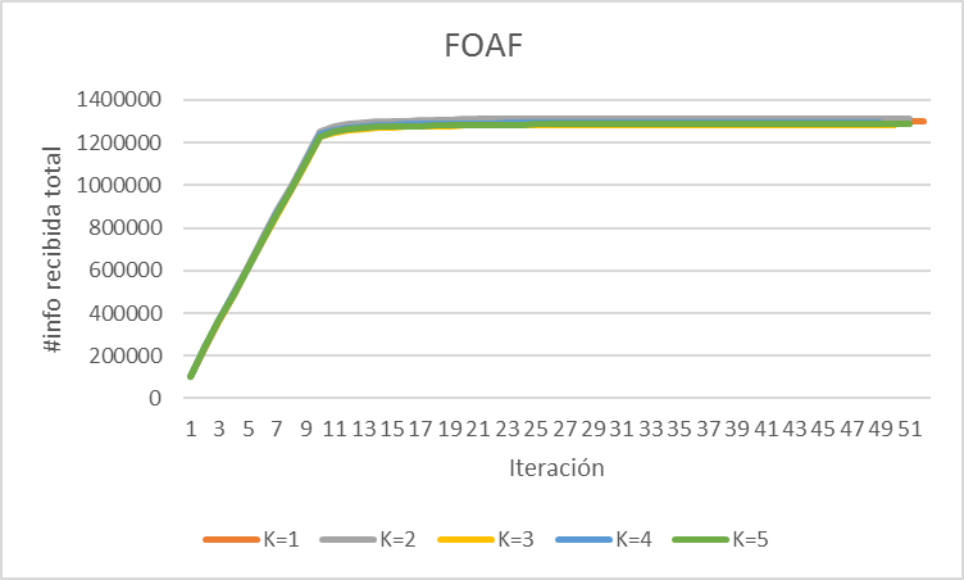


Figura C-5: Velocidad con FOAF.

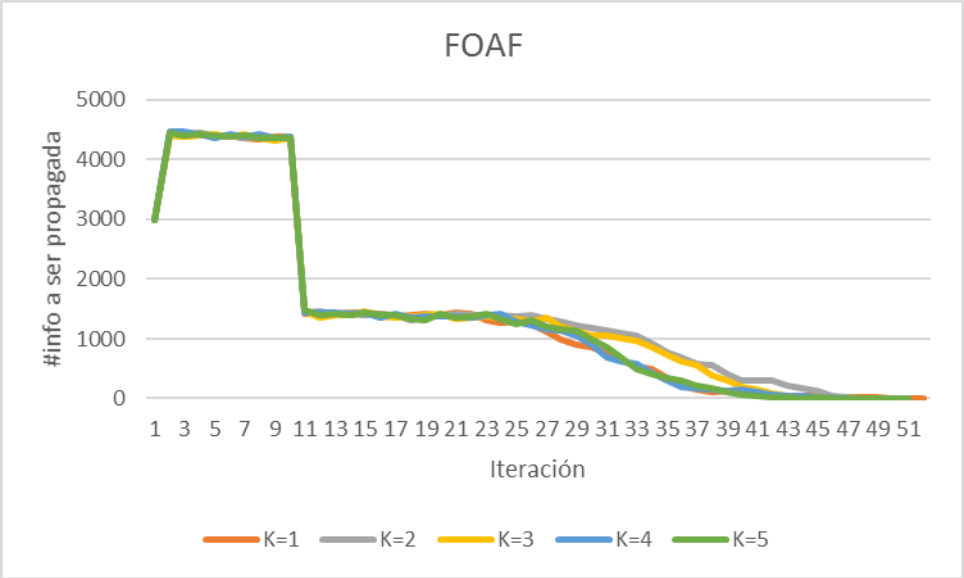


Figura C-6: Propagación con FOAF.

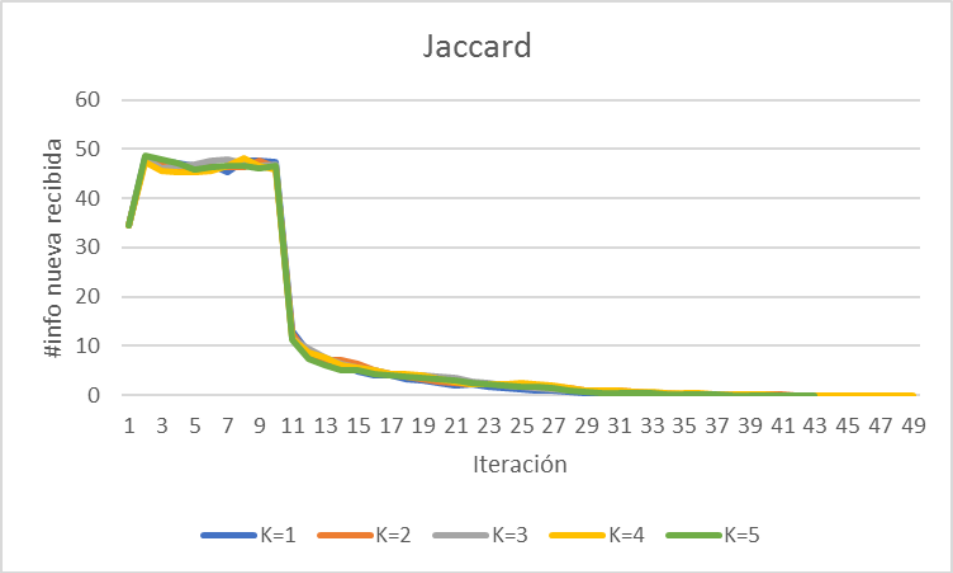


Figura C-7: Nueva información con Jaccard.

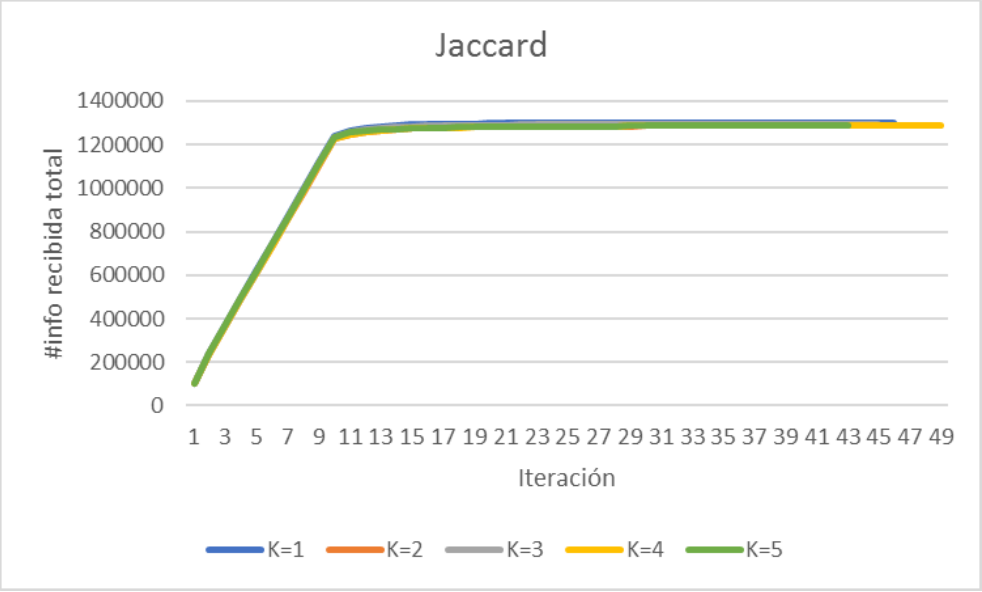


Figura C-8: Velocidad con Jaccard.

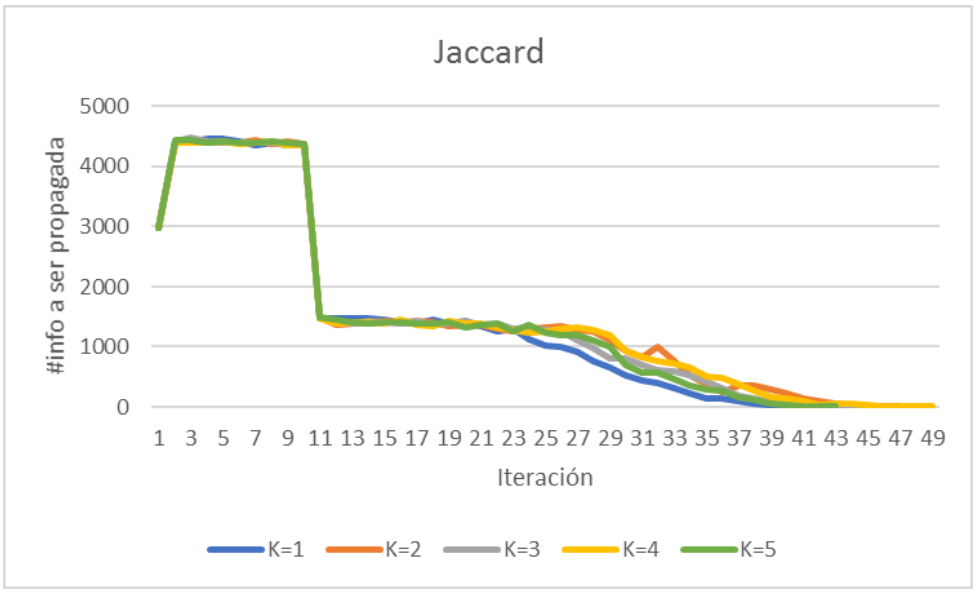


Figura C-9: Propagación con Jaccard.