UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD I, CIDI

SISTEMA PARA LA DETECCIÓN DE ROLES SOBRE COLECCIONES DE TUITS

TRABAJO FINAL PRESENTADO EN OPCIÓN AL TÍTULO DE
MÁSTER EN INFORMÁTICA AVANZADA

AUTOR: JOSÉ GABRIEL ESPINOSA RAMIREZ

Tutores:
Dr. José Ortíz Rojas
MSc. Delly Lien González Hernández

2018

La Habana, Cuba

Agradezco a *la Revolución y a la Universidad de las Ciencias Informáticas, por permitirme formarme como profesional y realizar la presente investigación. A mis tutores por su guía y apoyo. A Waldo por su valiosa colaboración. A mi familia por su constancia.*

Dedicado a *mi amada esposa y a mi querido hijo José Alejandro...*

Declaración jurada de autoría

Declaro por este medio que yo José Gabriel Espinosa Ramirez, con carnet de identidad 86020912429, soy el autor principal del trabajo final de maestría que lleva por título “Sistema para la detección de roles sobre colecciones de tuits”, desarrollado como parte de la Maestría de Informática Avanzada y que autorizo a la Universidad de las Ciencias Informáticas a hacer uso de la misma en su beneficio, así como los derechos patrimoniales con carácter exclusivo. Finalmente declaro que todo lo anteriormente expuesto se ajusta a la verdad y asumo total responsabilidad moral y jurídica que se derive de este juramento profesional. Y para que así conste, firmo la presente declaración jurada de autoría en La Habana a los _____ días del mes _____ del año _____.

Firma del maestrante Firma de los tutores

RESUMEN

Los vínculos entre las personas sirven de base para el crecimiento de la red social de cada individuo. En las relaciones con nuestros semejantes cada uno de nosotros ocupa diversos roles que varían en el decursar del tiempo en dependencia de diversas condiciones. Esto también es cierto para los sitios de redes sociales en Internet. Entre estos se distingue Twitter por su gran dinamismo y por haber jugado un papel central en acontecimientos recientes a escala global. Por lo que el estudio del mismo se ha convertido en un tema de interés para la comunidad científica, en la búsqueda de una mejor comprensión de sus dinámicas internas. Tomando como punto de partida a la unidad comunicacional de este sitio de redes sociales, el tuit, en la presente investigación se aplica el análisis de redes sociales centrado en el actor y elementos de la teoría de grafos para determinar el comportamiento de los usuarios dentro de los roles de sociable, popular, intermediario y aglutinador. El comportamiento de los usuarios en el marco de cada uno de estos roles puede ser cuantificado mediante el cálculo de métricas de centralidad. Con este objetivo fue desarrollado un sistema informático que abarca la recuperación y procesamiento de los tuits bajo demanda, así como el modelado y persistencia de las redes sociales derivadas de las relaciones de mención, cita, retuit, colaboración y réplica. Para la validación del resultado obtenido se emplearon técnicas y métodos de naturaleza cuantitativa y cualitativa.

Palabras claves: Análisis de redes sociales, detección de roles, métricas de centralidad, Twitter.

Abstract

The links between people serve as the basis for the growth of the social network of each individual. In relationships with our peers each of us occupies different roles that vary in the course of time depending on various conditions. This is also true for social networking sites on the Internet. Among these, Twitter stands out for its great dynamism and for having played a central role in recent events on a global scale. So the study of it has become a topic of interest to the scientific community, in the search for a better understanding of its internal dynamics. Taking as a starting point the communication unit of this social networking site, the tweet, in the present research it is applied social network analysis centered on the actor and elements of graph theory to determine the behavior of users within the roles of sociable, popular, intermediary and agglutinating. The behavior of users in each of these roles can be quantified by calculating centrality metrics. With this objective, a computer system was developed that covers the recovery and processing of tweets on demand, as well as the modeling and persistence of social networks derived from the relations of mention, appointment, retuit, collaboration and replication. For the validation of the obtained result, techniques and methods of a quantitative and qualitative nature were used.

Key words: Social network analysis, role dectection, central metrics, Twitter.

Índice general

Introducción **1**

1 Fundamentos teóricos sobre la detección de roles en colecciones de tuits **8**

1.1 Introducción . . . 8
1.2 Análisis de redes sociales . . . 8
1.3 Las relaciones entre los usuarios y la estructura del objeto tuit . . . 10
1.4 Detección de roles en el análisis de redes sociales . . . 12
1.4.1 Rol de popular . . . 13
1.4.2 Rol de sociable . . . 14
1.4.3 Rol de intermediario . . . 14
1.4.4 Rol de aglutinador . . . 14
1.5 Teoría de grafos aplicada a redes sociales . . . 14
1.5.1 Determinación del camino mínimo . . . 15
1.6 Algoritmo de Brandes para el cálculo de métricas de centralidad basadas en los caminos más cortos . . . 15
1.7 Métricas de centralidad . . . 16
1.7.1 Centralidad de grado . . . 16
1.7.2 Excentricidad . . . 18
1.7.3 Intermediación . . . 18
1.7.4 Centralidad por cercanía . . . 19
1.7.5 Centralidad por media armónica . . . 19
1.7.6 Centralidad por eigenvector . . . 20
1.7.7 Algoritmo HITS . . . 20
1.7.8 PageRank . . . 20
1.8 Soluciones existentes . . . 21
1.8.1 UCINET . . . 21
1.8.2 NetMiner . . . 21
1.8.3 Arcade Analytics . . . 22
1.8.4 NodeXL . . . 22

1.8.5 Social Network Visualizer . . . 22
1.8.6 Pajek . . . 23
1.8.7 Gephi . . . 23
1.8.8 Análisis de las soluciones existentes . . . 23
1.9 Conclusiones parciales . . . 24

2 Implementación de los componentes para la captura de tuits y la detección de roles 26
2.1 Introducción . . . 26
2.2 Arquitectura del sistema . . . 26
2.2.1 Integración e intercambio de mensajes . . . 27
2.3 Gestión de los requerimientos para la ejecución de estudios sobre colecciones de tuits . 28
2.4 La recuperación, procesamiento y la persistencia de la información contenida en los tuits 30
2.4.1 La recuperación de los tuits . . . 31
2.4.2 El procesamiento del tuit y la detección de relaciones . . . 32
2.4.3 Conformación de la colección de tuits . . . 35
2.4.4 Conformación del grafo de relaciones derivadas del objeto tuit . . . 36
2.5 Cálculo de métricas de centralidad sobre colecciones de tuits . . . 36
2.6 Conclusiones parciales . . . 38

3 Análisis de los resultados 39
3.1 Introducción . . . 39
3.2 Descripción de la validación de la hipótesis . . . 39
3.3 Cuasi - experimento . . . 40
3.4 Resultados de la técnica de Iadov . . . 42
3.5 Entrevista en profundidad . . . 44
3.6 Grupos focales . . . 45
3.7 Triangulación metodológica . . . 46
3.8 Conclusiones parciales . . . 47

Conclusiones 49

Recomendaciones 50

Bibliografía 52

Índice de figuras

2.1 Vista general del sistema. 27
2.2 Aplicaciones para la captura, procesamiento y persistencia de los tuits. 31
2.3 Vista del flujo de trabajo del procesador de tuits. 33
2.4 Vista de las relaciones entre dos usuarios en Neo4j. 36
2.5 Componentes para el cálculo de las métricas de centralidad. 37

Índice de cuadros

1.1 Roles y métricas de centralidad. 17
1.2 Soluciones analizadas. 24

2.1 Esquema de intercambio de mensajes entre las aplicaciones haciendo uso de RabbitMQ. 28
2.2 Atributos del objeto tuit que contienen información sobre las relaciones de réplica. 34
2.3 Atributos del objeto tuit que contienen información sobre las relaciones de cita. 35
2.4 Atributos del objeto tuit que contienen información sobre las relaciones de retuit. 35

3.1 Elementos cuantitativos para la descripción de segmentos de usuarios de la red social de Twitter, a disposición de DOWAI, previo al sistema para la detección de roles. 41
3.2 Elementos cuantitativos para la descripción de segmentos de usuarios de la red social de Twitter, a disposición de DOWAI, posterior al sistema para la detección de roles. 41
3.3 Cuadro de ladov utilizado. 42
3.4 Escala del índice de satisfacción individual. 42
3.5 Escala del índice de satisfacción grupal. 43
3.6 Objetivos a evaluar y métodos empleados en la triangulación metodológica. 47
3.7 Comparación entre los resultados obtenidos en cada técnica 47
8 Especificación de los requisitos funcionales del sistema. 51

Introducción

El ser humano es un ente social por su propia naturaleza, para su desarrollo necesita establecer relaciones con sus semejantes. Cada persona a lo largo de su vida va construyendo relaciones, de diversa índole, con quienes le rodean, y estos a su vez establecen relaciones con otros. Basándose en esto es que John Arundel Barnes introdujo el concepto de red social, al estudiar la organización social de una parroquia en Noruega. Barnes planteó el modelado de estas relaciones como una red en la que las personas o grupos de estas se representan como puntos unidos por líneas, que indican la interacción entre un par de personas o grupos determinados (Barnes, 1954). De esta manera se obtiene un modelo para la representación del tejido social, lo que facilita su interpretación y análisis.

Las redes sociales han acompañado siempre al ser humano desde su aparición en la Tierra, pero no fue hasta tiempos recientes cuando estas se hicieron más tangibles gracias a la aparición de los sitios de redes sociales en Internet. Estos surgieron en el marco de la evolución que ha tenido lugar en el entorno de las aplicaciones para la Web. El desarrollo de nuevas tecnologías propiciaron la transformación de la Web en un entorno más participativo y accesible para el público en general. Parte del éxito alcanzado por los sitios de redes sociales se debe al abaratamiento del costo de los dispositivos para el uso de aplicaciones en línea y de la conectividad, lo que ha propiciado a escala global el acceso masivo de las personas a Internet (ITU, 2011).

La presente investigación se acoge a la definición de sitio de redes sociales elaborada por Boyd y Elison. En la misma se plantea que un sitio de redes sociales es un “servicio sobre la Web que permite a las personas construir un perfil público o semipúblico dentro de un sistema cerrado, articular una lista de otros usuarios con los que se comparte una conexión, observar y navegar a través de sus listas de conexiones así como por aquellas hechas por otros dentro del sistema” (Boyd y Ellison, 2007). Tomando como base esta definición se puede afirmar que los sitios de redes sociales se distinguen por su capacidad para modelar y sistematizar las redes sociales de sus usuarios.

En el año 2006 surge el sitio de redes sociales Twitter, el cual desde sus inicios estuvo orientado hacia el servicio de mensajes cortos de la telefonía móvil. El tuit constituye su unidad básica de comunicación y permite el establecimiento de relaciones de seguimiento, mediante la cual tiene lugar la suscripción a los tuits generados por otros usuarios (Twitter, 2018a).

Twitter se ha expandido en el transcurso del tiempo hasta alcanzar más de trescientos trece millones de usuarios activos cada mes (Statista, 2018) (Twitter, 2018b). Los usuarios suelen utilizarlo para comunicarse con sus amigos, generar tuits relacionados con sus actividades cotidianas y mantenerse informados sobre los temas que sean de su interés. Por lo que ha ganado relevancia como herramienta para la divulgación y socialización de contenidos, para esto cuenta con la fortaleza de un ecosistema de

aplicaciones que interactúan con las funcionalidades del sitio y extendien su alcance (Barrera Martínez, 2017).

La capacidad, con que cuenta Twitter, de llegar directamente hasta el individuo ha sido aprovechada por las empresas. El uso de este sitio de redes sociales, a nivel empresarial puede apreciarse en dos niveles diferentes. En el primer nivel, se reconoce como un nuevo canal de comunicación que puede ser utilizado para el marketing, la asistencia al cliente y otros propósitos comerciales. En un segundo nivel, la información contenida en Twitter es apreciada como una fuente de datos importante para la toma de decisiones. Puesto que la información contenida en los tuits se puede emplear para explicar o predecir los movimientos del mercado (Xiong et al., 2016).

Twitter le posibilita a las empresas la divulgación de sus productos y servicios de persona a pesona. Esta transformación de los potenciales consumidores en promotores, ocurre bajo patrones de comportamiento que son objeto de estudio para su mejor entendimiento y aprovechamiento (Zhang et al., 2011). En la divulgación de la información es conocida la existencia de ciertos individuos que tienen una mayor capacidad de influencia que el resto. Estos, desde la posición privilegiada que tienen en la red pueden iniciar o controlar la difusión de la información (Pei y Makse, 2013) (Pei et al., 2017). La identificación de estos usuarios es un elemento de interés para potenciar la diseminación de la información. La ejecución de campañas de publicitarias en Twitter tienen un costo para las empresas, por lo que es necesario determinar previamente hacia donde dirigir los recursos y esfuerzos (Paniagua y Sapena, 2014).

No solo en el entorno comercial se puede apreciar el impacto de este sitio de redes sociales, en el entorno político también se ha hecho patente su importancia. Durante los sucesos acaecidos en el marco de la denominada primavera árabe durante el 2011 o en las protestas de Londres del 2010, dicha plataforma jugó un rol fundamental en la evolución de los acontecimientos. Es por esta razón que diversas empresas del mundo de la informática, las instituciones del gobierno y las fuerzas armadas de países como los Estados Unidos han decidido financiar el desarrollo de investigaciones sobre la dinámica del flujo de la información en el marco de un segmento de usuarios o tema de interés (Bhatt et al., 2012) (Hebenthal et al., 2012) (Chen et al., 2016).

Twitter se caracteriza por ser una vía de comunicación con una alta inmediatez ante acontecimientos sociales relevantes o situaciones extraordinarias, debido a que sus usuarios se comportan como una red de sensores sociales (Sakaki et al., 2010). Esto se ha evidenciado durante las campañas electorales en Estados Unidos, Alemania y otros países, en las cuales ha sido utilizado como una vía legítima y directa para conectar a los candidatos con sus electores. Ya que provee una retroalimentación inmediata sobre las reacciones del electorado durante el desarrollo de las campañas, por lo que ha jugado un importante papel en la estimación de los resultados finales de las contiendas electorales (Tumasjan et al., 2010) (Wang et al., 2016) (Shin et al., 2016) (Martínez et al., 2017).

Para diseminar la información de cada uno de los candidatos en Twitter, estos no solo se valen de sus cuentas oficiales, sino que se basan en el apoyo de múltiples usuarios que disponen de una audiencia

numerosa y permeable al mensaje que se transmite. Durante la campaña electoral del 2016 en Estados Unidos el Partido Republicano, que resultó vencedor, supo sacar buen provecho de este sitio de redes sociales. La adecuada selección de los mensajes y los usuarios encargados de transmitirlos les permitió sobrepasar la capacidad de divulgación del bando del Partido Demócrata en temas puntuales durante los debates que tuvieron lugar en Twitter (Espinosa et al., 2016) (Jin et al., 2017).

Si bien todos los usuarios son libres de seguir o retuitear a los usuarios que deseen, para la mayoría ser referenciado o retuiteado por un usuario relevante es algo muy significativo. Poniéndose de manifiesto la asimetría en la propagación de la información dentro de la red, la cual constituye un factor relevante para el éxito en la socialización de los contenidos en este sitio de redes sociales (Amor et al., 2016). La forma en que tienen lugar las interacciones entre los usuarios determina el comportamiento de estos como conductores de la información en un marco temporalmente determinado (Bruns et al., 2013) (Schroeder et al., 2017). Por lo que la extracción de la información en detalle sobre la estructura del grafo social de Twitter, es consecuentemente un paso fundamental para comprender la evolución del debate sobre un tema particular (Beguerisse-Díaz et al., 2014) (Amor et al., 2016).

Además de los contenidos que se publican en Twitter, el mismo ofrece otras informaciones que no están explícitas y que se encuentran relacionadas con la estructura y la dinámica de la red social sobre la que se sustenta el mismo. Para la extracción de este conocimiento implícito es necesario utilizar técnicas del análisis de redes sociales, el cual se define como el proceso de investigar las estructuras sociales mediante el uso de la teoría de grafos (Otte y Rousseau, 2002).

El éxito alcanzado por Twitter, como plataforma de divulgación y debate sobre tópicos de diversa naturaleza, lo ha convertido en el centro de atención para la realización de estudios tomando como base las interacciones entre los usuarios. Siguiendo lo planteado por Steinert-Threlkeld, el uso de Twitter como fuente de datos para la ejecución de estudios sobre eventos específicos, se basa en cuatro aspectos principales (Steinert-Threlkeld, 2017):

- Es uno de los sitios de redes sociales más utilizados a nivel global.
- Es muy utilizado durante eventos de crisis para diseminar información, incluso durante las manifestaciones y desastres.
- Aunque es utilizado para discutir sobre política, también se usa para dialogar sobre sucesos cotidianos, el clima, celebridades y otros temas de diversa naturaleza.
- Twitter cuenta con las facilidades necesarias para recuperar grandes cantidades de datos mediante interfaces de programación.

Para utilizar de manera eficiente a Twitter, es importante entender cómo la información se disemina entre los usuarios y el rol de cada uno en el debate. Con este fin, el Centro de Ideoinformática, perteneciente

a la Facultad 1 de la Universidad de las Ciencias Informáticas, cuenta con el Departamento de Operaciones Web y Análisis de Información (DOWAI). El cual tiene entre sus áreas de trabajo el estudio de Twitter, de acuerdo a las necesidades de sus clientes. Una vez formalizados y establecidos los criterios para acotar el estudio sobre la base de los requerimientos previamente establecidos, se determina el segmento de interés y luego se procede a la recopilación de los datos que permitan la descripción del segmento seleccionado.

Para realizar la descripción de las interacciones de los usuarios enmarcados en el segmento de interés de la red social se toman en cuenta la cantidad de retuits, de respuestas, de tuits marcados como favoritos, la proporción entre usuarios seguidos y seguidores, los usuarios que cuentan con una mayor cantidad de menciones y las etiquetas más utilizadas por estos, así como la relevancia de los usuarios involucrados. En la ejecución de esta labor se utilizan herramientas informáticas de terceros disponibles en Internet. Estas herramientas no permiten analizar en detalle de las interacciones entre los usuarios que tienen lugar en el marco del segmento de la red social que es objeto del estudio. Estas interacciones pueden ocurrir mediante las relaciones de mención, réplica, cita, retuit, y colaboración.

En Twitter los usuarios pueden expresarse de diversas formas, más allá del contenido plasmado en el propio tuit. Las interacciones entre los usuarios constituyen también una forma de expresión e impactan el funcionamiento de la propia red social como un todo. Tomando como partida a cada uno de estos tipos de relaciones, es posible determinar el rol que juega cada usuario en el marco del segmento de la red social estudiado. Esta información contribuye a una descripción acotada temporalmente y enmarcada en un conjunto de usuarios que participan de forma activa en el debate de un tema determinado.

Dados los elementos anteriores se puede afirmar que:

- Al carecer el departamento de un mecanismo propio para la construcción de colecciones de tuits, de acuerdo a los requerimientos de los estudios solicitados por los clientes para la descripción de un segmento de la red social, este se ve limitado en su capacidad para el análisis más allá de la descripción de los elementos cuantitativos y cualitativos obtenidos mediante herramientas informáticas de terceros.

- No se profundiza en el significado, desde un punto de vista cualitativo, de la implicación de los usuarios en los diferentes tipos de relaciones establecidas dentro de la estructura de la red social, como un todo.

- La descripción resultante se ve sesgada al solo basarse en la actividad del usuario como unidad, sin tener presente la información implícita en la interacción de este con otros usuarios del segmento de la red social objeto del estudio.

Por lo antes planteado se establece como problema científico: ¿Cómo contribuir a elevar la capacidad para describir un segmento de usuarios de la red social del sitio de redes sociales Twitter por parte del Departamento de Operaciones Web y Análisis de Información, del Centro de Ideoinformática?

El objetivo general planteado es: Desarrollar un sistema informático para la detección de roles de un segmento de usuarios de Twitter tomando como base a las interacciones de estos.
La investigación tiene como objeto de estudio al análisis de redes sociales, y como campo de acción se plantea la detección de roles sobre segmentos de usuarios del sitio de redes sociales Twitter.
Para dar cumplimiento al objetivo general se trazaron los siguientes objetivos específicos:

- Elaborar el marco teórico-referencial de la investigación relacionado con la detección de roles aplicada al sitio de redes sociales Twitter.
- Definir el esquema de roles para la descripción de un segmento de usuarios de la red social del sitio de redes sociales Twitter.
- Implementar los componentes para la recolección de tuits en tiempo real de acuerdo a los criterios de selección aportados.
- Implementar los componentes necesarios para la detección de roles sobre colecciones de tuits.
- Evaluar experimentalmente los resultados obtenidos mediante la ejecución de los componentes para la detección de roles sobre colecciones de tuits.

Teniendo en cuenta los elementos antes expuestos se plantea como hipótesis de la inviestigación: El sistema para la detección de roles sobre segmentos de usuarios de la red social Twitter incrementará la capacidad de descripción de estos segmentos por parte del Departamento de Operaciones Web y Análisis de Información.
Se identifica como variable independiente: La detección de roles sobre segmentos de usuarios de la red social Twitter, y como variable dependiente: la capacidad de descripción de los segmentos de usuarios de la red social Twitter por parte del Departamento de Operaciones Web y Análisis de Información.
Para el desarrollo de la investigación se emplearon métodos teóricos y empíricos. Los métodos teóricos son aquellos que se basan en la utilización del pensamiento en sus funciones de deducción, análisis y síntesis (Atalis Santa Cruz, 2015). A continuación se enumeran los métodos teóricos utilizados:

- Analítico - sintético: Para el estudio los elementos de la teoría de grafos y el análisis de redes sociales sobre los que se sustenta la detección de roles.
- Histórico - Lógico: Para el estudio de las acciones realizadas por un conjunto de usuarios, delimitado por un tópico específico y un intervalo de tiempo determinado, y el impacto de estas en el papel de cada usuario dentro de la red social.
- Inductivo - deductivo: Para el estudio de las principales métricas aplicables a la detección de roles y cuáles son viables para incorporar a la investigación.

- Análisis documental: Para la consulta de la literatura especializada y la extracción de los referentes teóricos para la presente investigación.
- Modelación: Para la generación de un modelo basado en la teoría de grafos mediante el cual se representen a los usuarios, las relaciones entre estos y los datos asociados, de tal forma que sea posible un análisis cuantitativo del segmento de la red social estudiado.
- Sistémico estructural: Para la descripción de la arquitectura del sistema informático desarrollado, los protocolos y tecnologías implicadas en el funcionamiento de la misma, así como la relación entre los componentes de esta y los servicios ofrecidos por la plataforma de Twitter.
- Hipotético - deductivo: Para la conformación de la hipótesis y su contrastación.

Para el desarrollo de la investigación se emplearon los siguientes métodos empíricos:

- Cuasi experimento: Para la validación del sistema mediante la utilización de un caso práctico en un entorno controlado.
- Medición: Para determinar cuantitativamente el desempeño de cada usuario en cada rol, sobre la base de los distintos tipos de relaciones que se establezcan entre estos.

Para la validación del resultado obtenido se utilizaron las técnicas:

- Grupo focal: Para la participación de especialistas en el tema tratado, con el objetivo de incorporar elementos para incrementar la calidad de la solución. Además de contribuir a la validación del resultado obtenido, como método cualitativo.
- Iadov: Para determinar el nivel de satisfacción de los especialistas del Departamento de Operaciones Web y Análisis de Información con la aplicación informática desarrollada.
- Entrevista en profundidad: Para conocer en detalle la apreciación del resultado de la investigación por parte de los usuarios finales, así como otros datos que los entrevistados puedan ofrecer desde la perspectiva de su experiencia.
- Triangulación metodológica: Para contrastar los resultados obtenidos de la aplicación de las técnicas de naturaleza cualitativa y cuantitativa.

La novedad y aporte práctico de la solución radican en el desarrollo de un sistema informático para la detección de roles sobre las interacciones entre los usuarios de Twitter, delimitadas por la presencia de un conjunto de términos y temporalmente definidas. Mediante este sistema informático se dota al Departamento de Operaciones Web y Análisis de Información de una herramienta para estimar el desempeño de los usuarios en cada rol, lo que permite incrementar la capacidad del departamento para describir un segmento de la red social de interés.

En la presente investigación se realiza una sistematización de los roles de los usuarios sobre la base del cálculo de las métricas de centralidad correspondientes. De esta manera se obtiene una medida cuantificable del desempeño de cada usuario dentro de cada rol, para los tipos de relaciones estudiados. Desde un punto de vista social, mediante el sistema informático obtenido, al determinarse el papel desempeñado por cada usuario dentro de la dinámica del flujo de la información en un segmento de la red social Twitter, se pueden identificar los actores hacia los cuales dirigir los esfuerzos para mejorar el resultado de las campañas de divulgación en este sitio. Además, se cuenta con la posibilidad de observar, medir y evaluar el desarrollo de las estrategias de divulgación de terceros en la red.
De esta manera, constituye una herramienta de apoyo a la toma de decisiones, en tiempo real o cercano a este, durante la ejecución de eventos de interés que se vean reflejados en la red. Las facilidades para la detección de roles en el sitio de Twitter pueden servir de base para la creación de nuevos productos y servicios para incrementar la cartera comercial del Centro de Ideoinformática.
El informe de la investigación está estructurado de la siguiente manera:

- Capítulo I. “Fundamentos teóricos sobre la detección de roles en colecciones de tuits”: Con el objetivo de elaborar el marco teórico, se exponen los principales elementos relacionados con la teoría sobre el análisis de redes sociales y la detección de roles. Se abordan los antecedentes de la investigación, y se realiza una descripción de los algoritmos y las métricas de centralidad utilizadas para la determinación de los roles.

- Capítulo II. “Implementación de los componentes para la captura de tuits y la detección de roles”: Se describen las técnicas de programación utilizadas, así como la arquitectura del sistema. Haciéndose énfasis en la integración entre los componentes del sistema y de este con los servicios de Twitter.

- Capítulo III. “Análisis de los resultados”: Se hace una descripción de los resultados obtenidos durante la aplicación de los métodos y técnicas para la comprobación de la hipótesis. Se exponen las valoraciones sobre la aplicabilidad, pertinencia y aceptación del sistema desarrollado.

- Conclusiones: Se exponen las conclusiones finales a las que se han arribado como resultado de la investigación.

- Recomendaciones: Se enuncian elementos que pueden ser tomados como punto de partida para el desarrollo de futuras investigaciones vinculadas con la detección de roles y el análisis de redes sociales.

- Anexo: Se describen los requisitos funcionales abordados por el sistema implementado.

- Bibliografía: Se enumeran las diferentes fuentes documentales que han sido referenciadas a lo largo del informe de la investigación.

Capítulo 1.

Fundamentos teóricos sobre la detección de roles en colecciones de tuits

1.1. Introducción

En el presente capítulo se abordan los fundamentos teóricos del análisis de redes sociales y de la detección de roles sobre los que se sustenta la investigación. Se realiza un análisis del objeto tuit y de las distintas relaciones que se derivan de la estructura del mismo.

Se describen los elementos de la teoría de grafos utilizados para el modelado de la red social. Se exponen las métricas de centralidad utilizadas para cuantificar el comportamiento de los usuarios de Twitter que conforman el segmento de la red objeto del estudio. A la vez que se realiza una descripción de varias soluciones informáticas para el análisis de redes sociales.

1.2. Análisis de redes sociales

El análisis de redes sociales comprende un conjunto de técnicas para la cuantificación e interpretación de las relaciones entre las entidades sociales. Estas entidades pueden representar a individuos o grupos de individuos tales como: organizaciones, comunidades, equipos, o alianzas. El enfoque basado en las entidades sociales es el reflejo del hecho de que el análisis de redes sociales tiene su origen en el campo de la sociología y la antropología, para el estudio de las interacciones sociales de una forma práctica. Una de las características del análisis de redes sociales es la visualización de las relaciones de forma simultánea tanto a nivel individual como global (Cronin et al., 2016).

Al alcanzarse un mejor entendimiento sobre las dinámicas de las relaciones entre los actores de la red, es posible cuantificar, monitorear y evaluar el flujo de la información. Lo que puede ser empleado para mejorar el desempeño de las organizaciones dentro de la red social. Según Serrat (Serrat, 2017) el resultado del análisis de redes sociales puede ser aplicado, entre otros objetivos, para:

- Identificar los individuos, grupos y unidades que juegan un rol relevante dentro del funcionamiento de la red.
- Descubrir cuellos de botella, segmentos aislados, y otras barreras para el flujo de la información.
- Identificar oportunidades para acelerar la difusión de la información.
- Fortalecer la eficiencia y efectividad de los canales para la comunicación.

Según Wasserman y Faust, en la base del análisis de redes sociales se encuentran los conceptos del actor, el enlace, la pareja, el subgrupo, el grupo, la relación y la red. El actor se define como un individuo discreto, o un colectivo de unidades sociales. La característica que define a un enlace es el

establecimiento de una relación entre un par de actores. Los enlaces pueden representar diferentes tipos de relaciones, ser unidireccionales o bidireccionales, y solo existen entre un par de actores específicos (Wasserman y Faust, 1994).

La pareja de actores se establece al concretarse, al menos, un enlace relacional entre estos, siendo este una característica inherente de la pareja. Una misma pareja de actores puede mantener más de un enlace al mismo tiempo. El subgrupo se conforma por un subconjunto de actores y todos los enlaces entre estos. Mientras que un grupo consiste en un conjunto finito de actores, sobre los que se realiza el estudio de la red; la relación se define como la colección de enlaces de un mismo tipo entre todos los miembros de un grupo (Wasserman y Faust, 1994).

Siguiendo las anteriores definiciones, y adaptándolas al sitio de redes sociales Twitter, se asume para la investigación, que:

- Los usuarios se corresponden con la definición del actor.
- Los enlaces que se pueden establecer entre los actores representan las relaciones de seguimiento, mención, retuit, réplica, cita y contribución.
- Todos los enlaces se caracterizan por ser unidireccionales.

El análisis de redes sociales se clasifica en dos tipos principales, el análisis global de la red y el análisis de redes centrado en el actor. El primer tipo se basa en las interacciones entre los actores dentro de un entorno delimitado. Esta delimitación, geográfica o social, condiciona la naturaleza del análisis a realizar. El objetivo de este tipo de estudio es recolectar información sobre cada miembro del grupo y de las interacciones entre estos (McCarty y Molina, 2015).

El análisis de redes centrado en el actor, no está geográfica o socialmente delimitado. El propósito del análisis centrado en el actor es operacionalizar el contexto social del individuo en un conjunto de variables que puedan ser utilizadas para obtener información sobre el actor. A diferencia del análisis global de la red, en el cual los actores son conocidos con antelación, en los estudios centrados en el actor, estos son los que indican cuáles son sus contactos, expandiéndose la red con estos nuevos actores (McCarty y Molina, 2015).

Siguiendo lo planteado McCarty y Molina la presente investigación se encuentra enfocada hacia el análisis de redes centrado en el actor (McCarty y Molina, 2015). El conjunto de actores de la red social que componen el segmento de interés nunca es conocido con antelación, sino que es descubierto en la medida que cada usuario se vincule al mismo, tanto por una acción propia o de forma pasiva como resultado de una mención, réplica, contribución, cita, o retuit. Los estudios sobre segmentos de la red social abordados en la investigación, parten de la necesidad de describir la red social que se construye en torno a varias etiquetas o palabras claves específicas. De esta manera se acota la dimensión del segmento de usuarios de interés, dentro del universo de usuarios de Twitter.

Los métodos de análisis de redes sociales proveen explícitamente los estamentos formales y las mediciones de las propiedades de las estructuras sociales definidas mediante términos matemáticos. De esta manera se tienen definiciones claras de los conceptos implicados en el análisis de redes sociales y se facilita el desarrollo de los modelos. Existen varias maneras de describir los datos de una red social matemáticamente, entre ellas se destacan los esquemas basados en la teoría de grafos, los sociométricos y los algebraicos (Wasserman y Faust, 1994).

El esquema basado en la teoría de grafos constituye una forma elemental de representar a los actores y las relaciones, a la vez que constituye la base para los restantes esquemas. El segundo esquema, el sociométrico, se caracteriza por la ponderación de los enlaces entre los actores de la red. El esquema de notación algebraico se emplea para representar redes compuestas por diferentes tipos de relaciones (Wasserman y Faust, 1994).

Para el desarrollo de esta investigación el esquema que más se ajusta es el algebraico, debido a la presencia de diferentes tipos de relaciones. De esta manera el modelo que se obtiene es capaz de distinguir una relación de otra por el tipo de acción que le dio origen.

1.3. Las relaciones entre los usuarios y la estructura del objeto tuit

En el caso de Twitter, las redes sociales se componen de los usuarios y de las conexiones que estos construyen con otros a través de las acciones que tienen lugar en el marco de este sitio, tales como las menciones y respuestas (Himelboim et al., 2017). Estas conexiones pueden ser de diferente naturaleza y persistir durante un intervalo temporal determinado.

En la base de la estructura social de Twitter se encuentra la relación de seguimiento. Esta relación se establece mediante la suscripción de un usuario a los tuits generados por otro. Se caracteriza por ser un vínculo unidireccional, que puede perdurar en el tiempo y suele representar el interés del usuario seguidor por los contenidos generados por el usuario seguido. La relación de seguimiento se mantiene mientras los dos usuarios que participan de esta lo deseen.

La relación de seguimiento indica de forma directa el tamaño de la audiencia con que cuenta un usuario. Pero esta relación no necesariamente es un indicador de la capacidad para atraer la atención de otros usuarios de forma activa (Cha et al., 2010).

La unidad atómica de comunicación en Twitter es el tuit, también conocido como actualización de estado. Para su representación se utiliza un objeto homónimo, que cuenta con numerosos atributos, siendo los fundamentales el identificador, la fecha, y el texto del tuit. A su vez el tuit puede encapsular a otros objetos tales como el usuario, y las entidades en sus atributos (Twitter, 2017a).

En el marco del sitio de redes sociales Twitter existen otros tipos de relaciones que se caractarizan por estar estrechamente vinculadas al contenido u origen de los tuits. Estas son las relaciones de mención, cita, retuit, réplica y contribución, las cuales por su propia naturaleza establecen vínculos direccionales entre los usuarios.

El retuit es la acción que tiene lugar cuando un usuario publica nuevamente un mensaje que ha sido creado por otro con anterioridad. De esta acción se deriva una relación entre el usuario que realiza el retuit y aquel que publicó el tuit original. Aunque no necesariamente esta acción implica que se comparte el criterio emitido en el tuit, sí constituye una evidencia de que el usuario que realiza el retuit se muestra permeable al contenido expuesto en el tuit y desea participar de forma activa en su divulgación. Los retuits son una característica del valor intrínseco del contenido del tuit (Cha et al., 2010).

La acción de citar un tuit se realiza al ejecutar un retuit y agregar un comentario propio al tuit generado. De esta acción se deriva una relación de características similares al retuit, que además incluye de forma explícita el criterio de quien realiza la acción.

La réplica es la acción de responder a un tuit anteriormente generado por cualquier otro usuario. Mediante la réplica se establecen hilos de conversaciones en los que pueden participar más de un usuario. Las réplicas realizadas por un usuario serán visibles para sus seguidores. Al responder a un tuit se suele emitir un criterio sobre este y se manifiesta interés por el contenido abordado, a la vez que se contribuye a la divulgación del mismo (Twitter, 2018c).

Cuando un usuario incluye el nombre de otro usuario, precedido del caracter "@" en el cuerpo de un tuit esto es interpretado por Twitter como una mención, lo que modifica la estructura del objeto que representa al tuit. Esta acción da lugar a la correspondiente relación de mención, la cual tiene como origen al usuario emisor del tuit y como destino al usuario que ha sido mencionado. Al igual que la réplica, la mención permite la participación en las conversaciones que tienen lugar en Twitter (Twitter, 2018c). Mediante la mención se implica al usuario mencionado con el contenido del mensaje, de tal forma que se contribuye a su divulgación. Las menciones se orientan hacia el peso específico del nombre del usuario mencionado en el marco de la red social (Cha et al., 2010).

En el marco de la presente investigación la información contenida en el objeto tuit, es la base para la extracción de las relaciones de mención, cita, retuit, réplica y contribución entre los usuarios. Para el caso de que el tuit analizado sea una respuesta, este incluirá en el campo "in_reply_to_user_id_str" el identificador del usuario que fue autor del tuit original que ha sido objeto de la réplica (Twitter, 2017a).

En el caso de los retuits el atributo "retweeted_status" contiene el objeto que representa al tuit original. De forma similar ocurre con las citas, si el tuit en cuestión es producto de una cita, este contendrá un atributo denominado "quoted_status" que tiene como función contener la representación del tuit original (Twitter, 2017a). En ambos casos se puede extraer los datos del autor del tuit original, para el establecimiento de la correspondiente relación.

Las menciones se caracterizan por ser de naturaleza multievaluadas para un mismo tuit, y su representación es abordada mediante las entidades, las cuales ofrecen información adicional sobre el contexto de este. Las menciones se encuentran representadas como un arreglo de objetos. El objeto de mención de usuario entre otros atributos cuenta con los necesarios para representar los datos del usuario mencionado, tales como su nombre e identificador (Twitter, 2017a).

El objeto tuit puede incluir un campo en el que se indique los usuarios que contribuyeron con la elaboración del tuit. Partiendo de esta información es posible establecer una relación, en la que el usuario contribuyente constituye el origen y el usuario que emite el tuit el destino. La contribución en la elaboración de tuits no suele ocurrir con frecuencia.

1.4. Detección de roles en el análisis de redes sociales

El análisis de redes sociales se extiende más allá de la evaluación puntual de cada actor, identificando la función o posición de estos dentro de la red. La identificación del rol del actor es uno de los temas primarios del análisis de redes sociales (Scott y Carrington, 2011). Los roles como concepto han sido ampliamente investigados en diversos campos de la ciencia, tales como la antropología, la sociología y la psicología, dando lugar a tres enfoques principales para su conceptualización: el funcional, el basado en las interacciones, y el sistémico.

Desde una perspectiva funcional el rol se define como el comportamiento de un individuo desde su estatus en el marco de una estructura social determinada. De esta manera el rol se enfoca de forma externa definido por normas sociales impuestas al individuo (Newcomb, 1950) (Benamar et al., 2017).

Desde la perspectiva de las interacciones el rol es situacional, y se basa en las interacciones entre los usuarios. Desde esta perspectiva el rol se considera como una reacción al comportamiento de otros individuos, y no tiene lugar fuera de estas interacciones (Mead, 1967) (Benamar et al., 2017).

Teniendo en cuenta las anteriores conceptualizaciones, se planteó un tercer enfoque basado en la teoría de las acciones. Desde este punto de vista los roles surgen de las interacciones, pero estas interacciones toman forma de acuerdo al sistema estructurado en el cual ocurren. El rol es entonces definido como un patrón de comportamiento, relacionado con una posición particular del individuo dentro de un entorno en el cual interactúa (Füller et al., 2014) (Benamar et al., 2017).

En el marco de esta investigación el enfoque funcional del rol no se considera el más apropiado, debido a que plantea el predominio de las normas sociales sobre el individuo. En el sitio de redes sociales Twitter los usuarios cuentan con un alto grado de autonomía para construir sus redes sociales sobre la base de las acciones de mención, réplica, cita, retuit y colaboración.

En el caso de la perspectiva de las interacciones, esta cuenta con un enfoque más cercano a la presente investigación, aunque la perspectiva sistémica es la que más se ajusta. Si bien las relaciones sobre las que se construye el segmento de la red social que es analizado constituyen interacciones entre los usuarios, estas quedan dentro de las fronteras del sitio de Twitter. Por lo que quedan acotadas por los diferentes tipos de relaciones, que se admiten en este sitio de redes sociales, y sus correspondientes reglas.

Desde este punto de vista los actores pueden desempeñarse en varios roles al unísono, correspondiéndose cada rol con una característica del comportamiento de estos, y de las interacciones derivadas, en la red social. Mediante la cuantificación de los roles haciendo uso de elementos de la teoría de grafos

es posible establecer comparaciones que permiten determinar cuáles usuarios resultan más relevantes para un rol determinado, con respecto al resto del colectivo.

Twitter permite a sus usuarios interactuar de diversas maneras. De estas interacciones se derivan diferentes relaciones que pueden ser objeto del análisis de redes sociales. En esta investigación se abordan las relaciones que tiene su origen en los tuits generados por los usuarios, específicamente las relaciones de mención, cita, retuit, réplica y colaboración. Estas relaciones son fruto de las acciones del usuario en este sitio de redes sociales, y lo vinculan al debate o divulgación de un tema específico de forma explícita. De esta manera es posible conducir el análisis de redes sociales centrado en el actor sobre un tema determinado en Twitter, abarcando solamente a los usuarios que han manifestado interés en el mismo; lo que no es posible al hacer uso exclusivo de la relación de seguimiento. Mediante cada una de estas relaciones es posible construir una representación de la red social sobre la cual determinar el grado de desenvolvimiento de cada usuario.

Existe una gran diversidad de criterios sobre las clasifiaciones para los usuarios de Twitter, que varian según el propósito cada estudio. Entre ellas se encuentran las de: líder de opinión, mentor, embajador, influenciador, entusiasta, celebridad, diseminador, conector, comentador, curador entre muchas otras. Estos roles de usuarios están vinculados a características específicas de naturaleza cualitativa y cuantitativa de las interacciones en la red social. Hasta el momento no existe una taxonomía de roles que cuente con un consenso general, en su lugar se tienen múltiples formas para cuantificar el comportamiento de los usuarios en Twitter (Huang et al., 2014) (Riquelme y González-Cantergiani, 2016).

En su mayoría estos roles se fundamentan en el cálculo de métricas derivadas de la centralidad de grado, PageRank, HITS, eigenvector, intermediación, cercanía, y otras que están relacionadas con la posición del nodo estudiado dentro de la red social. En general en el análisis de redes sociales son aceptados los roles de aglutinador, intermediario, popular, y sociable; que están asociados con las métricas anteriormente mencionadas (Knoke y Burt, 1983) (Freeman, 1978) (Wasserman y Faust, 1994) (Huang et al., 2014) (Riquelme y González-Cantergiani, 2016) (Benamar et al., 2017).

1.4.1. Rol de popular

El rol de popular se corresponde con aquellos actores que suelen ocupar un lugar prominente en la red, y con los que otros usuarios tratan de establecer algún tipo de relación. La popularidad de un usuario está vinculada con su reconocimiento por otros usuarios de la red. Un usuario popular no necesariamente tiene que ser activo en la red social, pero alcanza notoriedad con el mero hecho de su presencia (Riquelme y González-Cantergiani, 2016).

El rol de popular en Twitter, es comunmente asociado con la cantidad de seguidores. Pero en el marco de esta investigación se asocia con la cantidad de usuarios que mencionan, retuitean, citan, contestan o realizan contribuciones con el usuario analizado; quedando vinculado con la métrica de centralidad de grado de entrada.

1.4.2. Rol de sociable

El rol de sociable se corresponde con el comportamiento de aquellos actores que realizan un esfuerzo activo por establecer tantos enlaces con otros actores de la red como les sea posible. La sociabilidad de un usuario está directamente vinculada con su interés de ser reconocido por otros en la red social. Un usuario sociable se caracteriza por estar involucrado de forma activa en la red (Wasserman y Faust, 1994) (Benamar et al., 2017).

Este rol se encuentra vinculado con la métrica de centralidad de grado de salida. Para las relaciones abordadas en esta investigación, un alto desempeño en este rol pone de manifiesto el interés del usuario por llamar la atención del resto.

1.4.3. Rol de intermediario

El actor dentro del rol de intermediario o puente realiza la función de proveer conexiones entre los diferentes actores de la red. Los actores que se desempeñan en este rol se caracterizan por aumentar la distancia entre los demás actores de la red al ser retirados. En el caso de que el actor estudiado participe en el único enlace del subgrafo o agrupamiento con el resto de la red, este gana en importancia al tener la capacidad de controlar el flujo. Sin embargo, el uso de un solo indicador para determinar la importancia de un actor en la red no es suficiente (Long et al., 2013) (Huang et al., 2014). En el marco de esta investigación el rol de intermediario es abordado mediante la métrica de intermediación.

1.4.4. Rol de aglutinador

El actor que juega el rol de aglutinador mantiene una posición central en el grupo. Este rol también ha recibido las denominaciones de núcleo, líder, estrella y concentrador. Se caracteriza por tener una gran influencia sobre los demás. La influencia se define como un fenómeno social que los actores pueden experimentar o ejercer, mediante la manifestación del hecho de que un actor puede inducir a sus vecinos a comportarse de manera similar a la suya (Huang et al., 2014).

Una expresión de la influencia de forma explícita es mediante la relación de retuit (Guille et al., 2013). Las métricas de excentricidad, cercanía, media armónica, eigenvector, PageRank y HITS ofrecen una descripción cuantificada del comportamiento del actor estudiado en el marco del rol de aglutinador (Dubois y Gaffney, 2014) (Bodrunova et al., 2017).

1.5. Teoría de grafos aplicada a redes sociales

La teoría de grafos constituye una herramienta para la modelación de una red social, mediante la representación de los actores como nodos y los enlaces entre estos como aristas. La representación matricial del grafo facilita las labores de cálculo. Para esto se genera una matriz en la que el valor de una celda representa la existencia, o no, de un enlace entre los actores correspondientes a la fila y la columna (Wasserman y Faust, 1994).

Un segmento de una red social puede ser modelado como un grafo G, el cual se compone por un

conjunto de vértices V y un conjunto de enlaces E entre los distintos pares de vértices v, tales que $v \in V$. Un camino se constituye por una secuencia alterna $(v_0, e_1, v_1, ..., e_n, v_n)$ de vértices v y enlaces $e \in E$ de tal forma que el enlace e_i conecta al vértice en la posición $i-1$ con el vértice en la posición i. La longitud del camino está determinada por la cantidad de enlaces que lo componen. El conjunto de vecinos para un vértice v_0 se define como $\partial\{v_0\} = v_n \in V \mid \{v_0, v_n\} \in E$ (Berge, 1985) (Wuchty y Stadler, 2003).

En el análisis de redes sociales la determinación de la distancia entre dos vértices de un grafo es muy importante, ya que constituye la base para el cálculo de las métricas de centralidad. La distancia $d(v_i, v_j)$ entre dos vértices es la longitud del camino más corto entre estos. En el caso de no existir un camino viable entre los vértices v_i y v_j la distancia se denota, por convención, como infinita: $d(v_i, v_j) = \infty$. Un grafo G se considera conexo si y solo si para todo par de vértices v_i y v_j, cualesquiera que estos sean, se cumple que al menos exista un camino viable desde el vértice v_i hasta el vértice v_j, tal que: $d(v_i, v_j) \neq \infty$ con $i \neq j$.

Tomando como punto de partida la distancia entre los nodos, es posible determinar el camino mínimo entre un par de nodos del grafo cualesquiera que estos sean. La determinación del camino mínimo, es un paso necesario para el cálculo de las métricas de centralidad utilizadas para describir el comportamiento de los usuarios en el marco de cada rol.

1.5.1. Determinación del camino mínimo

En teoría de grafos, la determinación del camino más corto es considerado uno de los problemas fundamentales de la optimización de redes. El mismo consiste en encontrar un camino entre un par de vértices, cualesquiera que estos sean, de tal forma que la distancia entre estos sea la menor posible (Cherkassky et al., 1996). En los grafos ponderados esta distancia está determinada por la suma del peso de cada una de las aristas que esté presente en el camino. En el caso de los grafos no ponderados se asume que cada arista tiene un peso unitario.

Entre los algoritmos más utilizados para el cálculo del camino mínimo entre un par de vértices de un grafo se encuentra el desarrollado por Edsger W. Dijkstra en 1956 y publicado unos años más tarde (Dijkstra, 1959). Este algoritmo inicialmente fue creado para determinar el camino mínimo entre un par de vértices de un grafo, pero posteriormente fue ajustado para calcular los caminos mínimos desde un nodo hacia los restantes, generando de forma iterativa el árbol de caminos mínimos del grafo.

1.6. Algoritmo de Brandes para el cálculo de métricas de centralidad basadas en los caminos más cortos

El cálculo de las métricas de centralidad basadas en la determinación de los caminos más cortos, implica un alto costo en recursos de cómputo y espacio. Esto que trae consigo limitaciones en las dimensiones de la red para la ejecución de estudios de tal naturaleza. En el año 2001, Ulrik Brandes presentó un nuevo algoritmo para el cálculo de la métrica de intermediación, que puede ser ajustado

para el cálculo de otras métricas basadas en los caminos más cortos de una forma más eficiente. La versión desarollada por Brandes, teniendo a m como la cantidad de enlaces y a n como la cantidad de actores, requiere un espacio en memoria de $\mathcal{O}(n+m)$ y se ejecuta en intervalos de tiempo de $\mathcal{O}(nm)$ para grafos con aristas no ponderadas y de $\mathcal{O}(nm + n^2 \log n)$ en el caso de los grafos ponderados (Brandes, 2001).

Branders demostró que es posible reducir el tiempo de ejecución al eliminar el proceso de iteración sobre el conjunto conformado por cada par de vértices, para la suma explícita de los valores necesarios para el cálculo de una métrica de centralidad determinada. En su lugar, partiendo de la observación de que las sumas parciales obedecen a una relación recursiva, se realiza el proceso de cálculo mediante una técnica de acumulación basada en la dependencia recursiva de un vértice en relación con cualquier otro (Brandes, 2001).

1.7. Métricas de centralidad

La centralidad de los nodos, o la identificación de cuáles nodos son más centrales que otros es un elemento clave del análisis de redes sociales. Esto es siguiendo el principio de que el nodo con una mayor centralidad cuenta con una ventaja sobre el resto en lo referente a la capacidad de influir sobre la dinámica de la red. Si se toma como base a una red ideal, este nodo tiene una mayor cantidad de enlaces o forma parte de enlaces relevantes en el grafo, por lo que partiendo de este se puede alcanzar al resto de los nodos con mayor facilidad, controlando el flujo de la red (Opsahl et al., 2010).

En el marco de la teoría de grafos y el análisis de redes sociales se cuenta con varias métricas de centralidad para un nodo perteneciente a un grafo conexo, que aportan información sobre la importancia del nodo con relación al grafo. Existen varias métricas de centralidad que son ampliamente utilizadas en el análisis de redes sociales: centralidad de grado, intermediación, cercanía, excentricidad, media armónica, PageRank, HITS y eigenvector (Wasserman y Faust, 1994) (Kleinberg, 1999) (Opsahl et al., 2010) (Hautz et al., 2016) (Huang et al., 2014) .

Estas métricas sirven de base para la determinación del rol que desempeña cada actor en el marco de la red social. En dependencia de los valores que se calculen para cada métrica el actor que es objeto del estudio se estima más o menos relevante dentro de la dinámica de la red, de acuerdo al objetivo de una métrica dada. La relación entre los roles y las métricas abarcadas en la investigación se expone en el cuadro 1.1

1.7.1. Centralidad de grado

La centralidad de grado se calcula como el total de nodos que son directamente adyacentes al nodo estudiado (Wasserman y Faust, 1994). En el caso de que las relaciones sean direccionadas esta métrica puede subdividirse en dos: grado de entrada y grado de salida. El grado de entrada representa la cantidad de nodos que mantienen un enlace cuyo destino es el nodo estudiado. Mediante esta métrica se puede estimar la “popularidad” del actor dentro de la red social. En el caso del grado de salida se

Cuadro 1.1: Roles y métricas de centralidad.

Roles	Descripción	Métricas
Intermediario	Suele controlar el flujo en la red, actúa como un filtro. Al ser retirado de la red aumenta la distancia entre los nodos.	• Intermediación
Popular	Se distingue por tener un alto reconocimiento, por parte del resto de los usuarios.	• Grado de entrada.
Socialble	Su participación es muy activa, e intenta llegar a tantos usuarios como le sea posible.	• Grado de salida.
Aglutinador	Mantiene una posición central en la red, se considera como el usuario más influyente del grupo.	• PageRank. • HITS. • Eigenvector. • Excentricidad. • Media armónica. • Cercanía.

refiere al total de nodos que son el destino de los enlaces que tienen como origen al nodo estudiado. El grado de salida representa la "sociabilidad" del actor analizado (Passmore, 2011).

Esta métrica se distingue por ofrecer una valoración sobre cuán involucrado se encuentra un nodo en la red. Su simplicidad radica en el hecho de que solo es necesario conocer las estructuras locales en las cercanías inmediatas al nodo estudiado. Sin embargo, tiene la limitación de no tomar en cuenta la estructura global de la red. Aunque un nodo puede presentar una alta centralidad de grado, estando conectado directamente con muchos otros nodos de la red, puede que no se encuentre en una posición desde la cual se pueda acceder rápidamente a otros nodos (Opsahl et al., 2010).

Para un grafo dado $G := (V, E)$ donde el conjunto V de vértices representa a los actores y el conjunto de aristas E a los enlaces entre los actores. Siendo n el cardinal del conjunto de vértices, la centralidad de grado de salida $C_{Out}(v)$ para un vértice v se define en la ecuación 1.1 como la media del grado de salida del vértice, estando esta última definida en la ecuación 1.2 (Berge, 1985).

$$C_{Out}(v) = \frac{deg_{Out}(v)}{n-1} \tag{1.1}$$

$$deg_{Out}(v) = \mid \{y \in V \mid \{v, y\} \in E\} \mid \tag{1.2}$$

Siguiendo el mismo principio, la centralidad de grado de entrada $C_{In}(v)$ se define en la ecuación1.3 como la media del grado de entrada del vértice para la cantidad de nodos del grafo. La definición del grado de entrada del vértice se muestra en la ecuación 1.4 (Berge, 1985).

$$C_{In}(v) = \frac{deg_{In}(v)}{n-1} \tag{1.3}$$

$$deg_{In}(v) = \mid \{y \in V \mid \{y, v\} \in E\} \mid \tag{1.4}$$

1.7.2. Excentricidad

La excentricidad $C_e(v)$, definida en la ecuación 1.5, es un índice de centralidad del nodo. Para su determinación es necesario calcular los caminos más cortos desde el nodo estudiado, v, hasta cualquier otro nodo, u en el grafo. De estos caminos más cortos se toma el de mayor longitud, y se determina el recíproco de este valor. Así, una excentricidad con un alto valor permite asumir un resultado favorable sobre la proximidad del nodo. De hecho si la excentricidad de un nodo es elevada, esto significa que todos los restantes nodos están en su proximidad. En contraste si la excentricidad es baja, significa que existe al menos un nodo que está distante del nodo estudiado (KRNC, 2015) (Center for BioMedical Computing, 2017).

$$C_e(v) = \frac{n-1}{max_{u \in V}(v, u)} \tag{1.5}$$

1.7.3. Intermediación

La centralidad es uno de los principales conceptos del análisis de redes sociales, y la centralidad por intermediación es una de las métricas de centralidad más notables. Esta métrica ofrece una medida del grado en que un nodo se desempeña como corredor o mediador, adicionando los segmentos de caminos más cortos entre todos los pares de vértices que pasan a través del nodo estudiado (Brandes, 2008).

La centralidad por intermediación se basa en la cantidad de veces que cualquier actor de la red necesita de algún otro actor para alcanzar a un tercer actor, cualquiera que este sea. Dado el conjunto de caminos más cortos entre todos los actores de la red, sea g_{ij} la cantidad de caminos más cortos desde el nodo i hasta el nodo j, y sea g_{ikj} el número de caminos más cortos tomando como origen al nodo i y como destino al nodo j que incluyan en su recorrido al nodo k. Siguiendo lo planteado por (Freeman, 1977) y (Borgatti y Everett, 2006) la centralidad por intermediación para el nodo k, C_k^{BET}, se define en la ecuación 1.6 como:

$$C_k^{BET} = \sum_i \sum_j \frac{g_{ikj}}{g_{ij}} \tag{1.6}$$

Un nodo con una alta intermediación tiene la capacidad de controlar el flujo en la red. La centralidad por intermediación puede ser determinada tanto en grafos direccionados como en no direccionados, totalmente conexos o no conexos (White y Borgatti, 1994) (Opsahl et al., 2010).

1.7.4. Centralidad por cercanía

La centralidad por cercanía se basa en la distancia de los caminos de un nodo hasta cualquier otro, y se define como el recíproco de la suma de la distancia de los caminos mínimos que tienen como origen al vértice estudiado y como fin cualquier otro vértice del grafo (Sabidussi, 1966) (Opsahl et al., 2010). Mediante esta métrica se puede estimar cuánto tiempo tomará para diseminar la información desde un nodo hasta los restantes nodos de la red, que de alguna manera estén conectados con el primero (Passmore, 2011). La centralidad por cercanía presenta limitaciones para su aplicación en grafos que no sean totalmente conexos; puesto que para la representación de la distancia entre dos elementos pertenecientes a dos subgrafos inconexos entre sí se suele utilizar el valor infinito, por lo que no es posible determinar la suma total de las distancias desde el nodo estudiado hasta los restantes nodos del grafo (Opsahl et al., 2010). El cálculo de esta métrica queda limitado solamente al subgrafo con mayor cantidad de nodos, omitiéndose el aporte de los nodos pertenecientes a los otros subgrafos (Rochat, 2009). Entre las opciones para solventar esta situación se encuentra la centralidad por media armónica.

Siguiendo lo anteriormente expuesto la centralidad por cercanía C_k^{CLN}, tomando a d_{ki} como el valor de la distancia del camino mínimo desde el nodo k hasta el i-ésimo nodo del mismo, se define en la ecuación 1.7 como:

$$C_k^{CLN} = \frac{1}{\sum_i d_{ki}} \tag{1.7}$$

1.7.5. Centralidad por media armónica

Como se expresó con aterioridad la centralidad por cercanía presenta dificultades para abordar grafos que no sean completamente conexos. Como en estos casos el uso por convención del valor infinito para representar la inexistencia de un enlace distorciona el cálculo, se hace necesario utilizar otro mecanismo equivalente. Rochat propone la utilización de la suma de los inversos de las distancias en lugar del uso del recíproco de la suma de las distancias (Rochat, 2009). Mediante la media armónica se evita las distorciones producto del uso de las distancias infinitas y se puede calcular esta métrica como equivalente de la centralidad por cercanía (Boldi y Vigna, 2014) (Marchiori y Latora, 2000).

Teniendo en cuenta que $\lim_{x\to\infty} \frac{1}{x} = 0$, se asume que $\frac{1}{\infty} = 0$; la centralidad por media armónica se define en la ecuación 1.8 como:

$$C_k^{CMA} = \sum_i \frac{1}{d_{ki}} \tag{1.8}$$

1.7.6. Centralidad por eigenvector

Mediante la métrica de centralidad por eigenvector se determina la relevancia de un nodo de acuerdo a los enlaces que este mantiene con otros. A diferencia de la centralidad de grado en la cual se le asigna a todos los contactos el mismo peso, al utilizar la centralidad por eigenvector se ponderan los nodos de acuerdo a los valores de centralidad de los restantes nodos. Esta métrica de centralidad se caracteriza por la suma ponderada de todas las conexiones con el nodo estudiado, sean directas o indirectas (Bonacich, 1972) (Bonacich, 2007).

Siendo A la matriz de adyacencia de un grafo $G := (V, E)$ que comprende n nodos tal que $A_{ij} = 1$ si existe un enlace directo desde el nodo i hasta el nodo j, y $A_{ij} = 0$ en caso contrario. Teniendo a λ como el valor normal más alto de A se define la centralidad por eigenvector C^{CEV} en la ecuación 1.9 como:

$$\lambda C_i^{CEV} = \sum_{i \neq j=1}^{n} A_{ij} C_j^{CEV} \tag{1.9}$$

1.7.7. Algoritmo HITS

El algoritmo Hypertext Induced Topic Selection (HITS), fue desarrollado por John Kelinberg a finales de la última década del siglo XX. El mismo se diseñó con el objetivo de incrementar la precisión en la recuperación de páginas web en Internet. Este se basa en las categorías de concentrador y de autoridades (Kleinberg, 1999).

La categoría de concentrador está relacionada con los actores a los que hace referencia el actor estudiado, mientras que el índice de autoridad se relaciona con los enlaces en los cuales el actor estudiado es el referenciado. Todos los actores de la red cuentan con sendos índices para estas categorías (Kleinberg, 1999).

Este algoritmo se caracteriza por ser iterativo. En cada una de las iteraciones se realizan dos pasos básicos: la actualización del índice de concentrador y del índice de autoridad. El índice de concentración se determina como la suma del índice de autoridad para todos los elementos a los que se puede llegar directamente desde el elemento estudiado. El índice de autoridad se calcula mediante la suma del índice de concentración para cada uno de los elementos que mantienen un enlace teniendo como punto de llegada al elemento estudiado. Al calcularse los nuevos valores estos son normalizados (Kleinberg, 1999).

1.7.8. PageRank

El algoritmo de PageRank fue desarrollado por Serguey Brin y Lawrence Page en 1998. Fue creado originalmente para determinar la relevancia de los recursos disponibles en la Web indexados por el motor de búsqueda de Google. Sus fundamentos se encuentran en la estructura de las relaciones entre las citas académicas de la literatura científica, mediante la cual se puede obtener una apreciación de la calidad e importancia de un documento. “PageRank extiende esta idea al no contabilizar los enlaces

desde todas las páginas por igual, en su lugar estos son normalizados por el número de enlaces a la página" (Brin y Page, 1998).

Se asume que para la página A existen $T1...n$ páginas que apuntan a esta. El parámetro d es un factor arbitrario de amortiguación que puede tomar valores entre 0 y 1. Mientras que $C(A)$ es definido como la cantidad de páginas referenciadas desde la página A. El PageRank $P(A)$ de una página A se calcula mediante iteraciones y está determinado, como se muestra en la ecuación 1.10 (Brin y Page, 1998).

$$P(A) = (1 - d) + d \sum_{i=1}^{n} (P(Ti)/C(Ti)) \tag{1.10}$$

1.8. Soluciones existentes

Para el análisis de redes sociales se han implementado diversas soluciones, algunas están disponibles desde la Web, mientras que otras pueden funcionar de forma local. En esta área se encuentran disponibles tanto aplicaciones gratuitas y de código abierto, como privativas. Entre las aplicaciones de código abierto y gratuitas se destacan Social Network Visualizer, Pajek y Gephi. Entre las privativas y de pago se encuentran las aplicaciones UCINET, NetMiner, Arcade Analitycs y NodeXL.

1.8.1. UCINET

"UCINET 6 para Windows es un paquete de software para el análisis de redes sociales, desarrollado por Lin Freeman, Martin Everett y Steve Borgatti" (UCINET, 2017). Es una aplicación privativa de código cerrado. Para su explotación es necesario contar con una licencia, aunque es posible utilizarla por un período de pruebas de 90 días. Este software incorpora la herramienta de visualización NetDraw.

UCINET requiere del sistema opeartivo Windows Vista o superior. La versión estándar es para 32 bits, aunque también existe una para 64 bits, que no incluye todas las funcionalidades. Puede ejecutar el cálculo de la métrica de centralidad de intermediación para un grafo de aproximadamente 10 000 nodos en un tiempo relativamente corto (UCINET, 2017).

1.8.2. NetMiner

NetMiner es una aplicación de escritorio para el análisis exploratorio y la visualización de datos de una red social, desarrollada por la empresa coreana CYRAM. Se caracteriza por ser una aplicación de pago y código cerrado. Requiere para su funcionamiento de alguno de los sistemas operativos de la familia de Microsoft Windows. Cuenta con una versión de prueba con una duración de dos semanas y capacidad de procesamiento hasta cinco mil nodos (CYRAM, 2018).

A través de NetMiner es posible detectar patrones e interactuar con la estructura de la red de forma dinámica. Esta aplicación incluye un entorno integrado de análisis. Cuenta con la capacidad de trabajar con los datos recuperados de sitios de redes sociales tales como Facebook y Twitter, así como blogs y foros para determinar la influencia de un usuario y el impacto de las publicaciones de este. Para esto se basan en los patrones de diseminación de la información y en el cálculo de las métricas de centralidad.

Para su integración con los sitios de redes sociales es necesario adquirir la correspondiente extensión (CYRAM, 2018).

1.8.3. Arcade Analytics

Arcade Analytics es una aplicación web para la visulización de grafos que le permite a los usuarios un mejor control sobre los datos. Cuenta con la capacidad de integrarse con los sistemas de gestión de bases de datos basados en grafos más utilizados en la actualidad. Mediante esta aplicación los usuarios pueden realizar consultas a las bases de datos y visualizar los resultados como grafos (Arcade Analytics LTD, 2018).

Inluye opciones de personalización para la visualización de los grafos y facilidades para su exportación a diversos formatos. Entre sus funcionalidades se encuentran el cálculo de la centralidad de grado y el PageRank, así como del camino mínimo entre cualquier par de nodos (Arcade Analytics LTD, 2018). No cuenta con las funcionalidades necesarias para integrarse con el sitio de Twitter de forma directa.

1.8.4. NodeXL

NodeXL es una plantilla avanzada para Microsoft Excel (2007, 2010, 2013 y 2016) en Windows (XP, Vista, 7, 8, 10) que permite la generación de grafos y reportes asociados, tomando como base a una hoja de cálculo que contenga una lista de enlaces. Cuenta con dos versiones NodeXL Basic y NodeXL Pro (Social Media Research Foundation, 2018).

NodeXL Basic es una herramienta para la visulización de los grafos contenidos en los ficheros generados mediante NodeXL Pro. Esta última es una aplicación de pago que debe renovar su licencia cada doce meses. Al hacer uso de NodeXL Pro es posible calcular métricas de centralidad, exportar e importar ficheros de grafos en diversos formatos, hacer uso de conexiones a los sitios de redes sociales tales como Facebook y Twitter entre otros así como a cuentas de correo electrónico (Social Media Research Foundation, 2018).

1.8.5. Social Network Visualizer

Social Network Visualizer es una aplicación multiplataforma para el análisis de redes sociales y la visualización de estas. Se caracteriza por ser gratuita y distribuirse bajo licencia GNU-GPL (Free Software Foundation, 2007) en su tercera versión. Permite interactuar dinámicamente con la red social que esté siendo analizada (Kalamaras, 2018).

Mediante Social Network Visualizer es posible calcular las métricas de centralidad y autoridad para una red social, tales como: cercanía, intermediación, PageRank, entre otras. Cuenta con la capacidad de rastrear la Web para detectar vínculos entre páginas, a la vez que admite como entrada archivos en diversos formatos que contengan la descripción de las redes sociales a analizar (Kalamaras, 2018).

Aunque esta aplicación no se puede acoplar directamente con los servicios de Twitter, puede procesar los datos de la red social previamente transformados en alguno de los formatos de ficheros que acepta

como entrada. La captura de los datos y la ejecución del modelo de la red debe ser realizado utilizando otras herramientas.

1.8.6. Pajek

Pajek es un software para el análisis y visualización de redes sociales. El mismo fue desarrollado en la universidad de Ljubljana, en Eslovenia por Vladimir Batagelj, Andrej Mrvar y con la contribución de Matjaž Zaveršnik (Ruiz y Jung, 2013). El software puede ser descargado desde la Web y utilizarse libremente para un uso no comercial.

Está diseñado para correr sobre sistemas operativos de la familia de Windows, aunque puede ser emulado en Linux. Mediante el uso de Pajek es posible analizar redes de gran tamaño. Como entrada Pajek recibe un fichero de texto con la descripción de la red a analizar. Este software cuenta con la funcionalidad de graficar los grafos que se han recibido como entrada.

Permite el cálculo de las métricas de centralidad de grado, intermediación, y cercanía entre otras. Además cuenta con la funcionalidad de manejar grafos que pueden variar en el tiempo, aunque carece de un mecanismo propio para la integración directa con el sitio de redes sociales Twitter.

1.8.7. Gephi

Gephi es una aplicación basada en software libre para la visualización interactiva y exploración de todo tipo de grafos, y sistemas complejos. Se encuentra disponible de forma gratuita y es compatible con los sistemas operativos Windows, Mac OS X, y Linux. A través de Gephi el usuario puede visualizar las estructuras de los grafos dirigidos, no dirigidos y mixtos. Soporta la visualización de hasta un millón de nodos y aristas (Gephi, 2017).

Este sistema incluye facilidades para la interacción con la red que se está analizando, permitiendo el filtrado en tiempo real y la generación de nuevas redes sobre el resultado de consultas. Gephi está preparado para recibir como entrada archivos en la mayoría de los formatos para grafos, así como en formato CSV. Además puede interactuar con sistemas de gestión de bases de datos relacionales y no relacionales (Gephi, 2017).

Aunque Gephi se centra en la visualización de grafos también incluye facilidades para la realización de análisis y cuenta con la posibilidad de extender sus funcionalidades mediante la inclusión de plugins. Permite el cálculo de todas las formas de centralidad de grado, la centralidad por eigenvector, intermediación, y PageRank, entre otras (Fagan, 2017).

1.8.8. Análisis de las soluciones existentes

Estas aplicaciones brindan las facilidades necesarias para el manejo de redes de grandes dimensiones. Cuentan con un modelo basado en la teoría de grafos, y sobre este ejecutan el cálculo de las métricas de centralidad. Entre estas se destacan las de excentricidad, media armónica, cercanía, PageRank, eigenvector, HITS, centralidad de grado de entrada y salida.

Para la ejecución del análisis de redes sociales en Twitter, la capacidad de integración de las aplicaciones con los servicios de esta plataforma es una característica deseable. Como se muestra en el cuadro 1.2, NetMiner, NodeXL y Gephi cuentan con esta funcionalidad. Tanto para NetMiner como NodeXL es necesario el pago correspondiente a sus licencias, las cuales deben ser renovadas periódicamente, y el acceso a su código fuente está restringido.

Cuadro 1.2: Soluciones analizadas.

Solución	Conexión con Twitter	Tipo de licencia	Extensible
UCINET	No disponible.	Licencia de pago.	No.
NetMiner	Mediante un plugin.	Licencia de pago.	No.
Arcade Analytics	No disponible.	Licencia de pago.	No.
NodeXL	Directamente.	Licencia de pago.	No.
Social Network Visualizer	No disponible.	GNU-GPL v3.	El código fuente está disponible y puede ser extendido.
Pajek	No disponible.	Licencia gratuita.	No.
Gephi	Mediante un plugin.	GNU-GPL v3.	El código fuente está disponible, y puede ser extendido mediante plugins.

Mientras que Gephi al ser de código abierto y estar licenciado bajo GNU-GPL en su versión 3, ofrece un alto grado de flexibilidad para su uso y la extensión de sus funcionalidades. El desarrollo de Gephi sigue una arquitectura modular, basada en componentes con un bajo acomplamiento; lo que permite la construcción de aplicaciones complejas y sostenibles (Gephi, 2018a). Los módulos esenciales de Gephi se encuentran empaquetados en el Gephi Toolkit (Gephi, 2018b), que puede ser utilizado para la integración de sus funcionalidades en nuevas aplicaciones. Por lo anterior, Gephi es la solución precedente que más se ajusta al propósito de la investigación, seleccionándose el Gephi Toolkit como motor para el cálculo de las métricas de centralidad para la detección de roles.

1.9. Conclusiones parciales

Las soluciones que anteceden a esta investigación tienen en común el uso de un modelo basado en la teoría de grafos para la representación de las redes sociales. Tomando como base este modelo calculan diferentes métricas de centralidad para determinar la relevancia de los actores de la red social. Siendo las más comunes la centralidad por intermediación, la centralidad de grado de entrada y salida, así como el eigenvector y PageRank. De las aplicaciones analizadas Gephi se distingue por contar con interfaces de programación que le confieren una gran flexibilidad, y facilidades para ser extendida mediante plugins.

Los grafos que se obtendrán como modelos de la red social a estudiar no serán necesariamente conexos, por lo que esta característica impacta en la selección de las métricas de centralidad utilizadas para determinar la participación del actor en la red. Las métricas de centralidad de grado y la centralidad por intermediación son compatibles con los grafos no completamente conexos, por lo que son seleccionadas para su implementación. Teniendo presentes las limitaciones de la métrica de centralidad por cercanía para el manejo de los grafos no completamente conexos se toma como alternativa para el cálculo de la centralidad por media armónica, la cual según la bibliografía analizada, ha demostrado una alta correlación con la centralidad por cercanía.

Capítulo 2.

Implementación de los componentes para la captura de tuits y la detección de roles

2.1. Introducción

En el presente capítulo se realiza un análisis de las herramientas y técnicas utilizadas para la implementación del sistema; así como, una descripción de su arquitectura. El sistema se corresponde con la especificación de requisitos expuesta en el cuadro número 8 del anexo I. Atendiendo a su naturaleza, los requisitos funcionales fueron agrupados en tres conjuntos fundamentales:

- La gestión de los requerimientos para la ejecución de estudios sobre colecciones de tuits, que abarca los requisitos funcionales desde el primero al séptimo.
- La recuperación bajo demanda, el procesamiento y el almacenamiento de los tuits generados desde Twitter; que agrupa a los requisitos funcionales desde el octavo al decimoséptimo.
- La ejecución del análisis basado en el cálculo de métricas de centralidad sobre los contenidos recuperados de Twitter, que incluye desde el décimoctavo requisito hasta el vigésimo quinto.

El primer grupo de requerimientos se refiere a la gestión de los términos e intervalos de tiempo que delimitan el alcance de la colección de tuits que es de interés para la ejecución de un estudio determinado. En el segundo grupo se aborda la recuperación de los tuits que coincidan con las necesidades de los estudios para su almacenaje y preprocesamiento, de tal forma que se facilite el cumplimiento de los requerimientos agrupados en el tercer punto. El último grupo se refiere al procesamiento de las relaciones establecidas entre los usuarios, mediante el cálculo de las métricas de centralidad sobre el grafo construido como resultado de la agregación de estas relaciones.

2.2. Arquitectura del sistema

El sistema para la detección de roles sobre colecciones de tuits está compuesto por cuatro aplicaciones que interactúan entre sí, utilizando como mecanismo de integración a un sistema de gestión de colas para el intercambio de mensajes. Cada una de estas aplicaciones tiene como propósito la especialización en la ejecución de una tarea específica.

En la figura 2.1 se muestra una vista general del sistema. Los cuatro componentes principales del sistema son:

- Aplicación web: Cumple la función de ser la interfaz del sistema, mediante la cual los usuarios finales pueden gestionar y observar los resultados de los estudios de detección de roles.

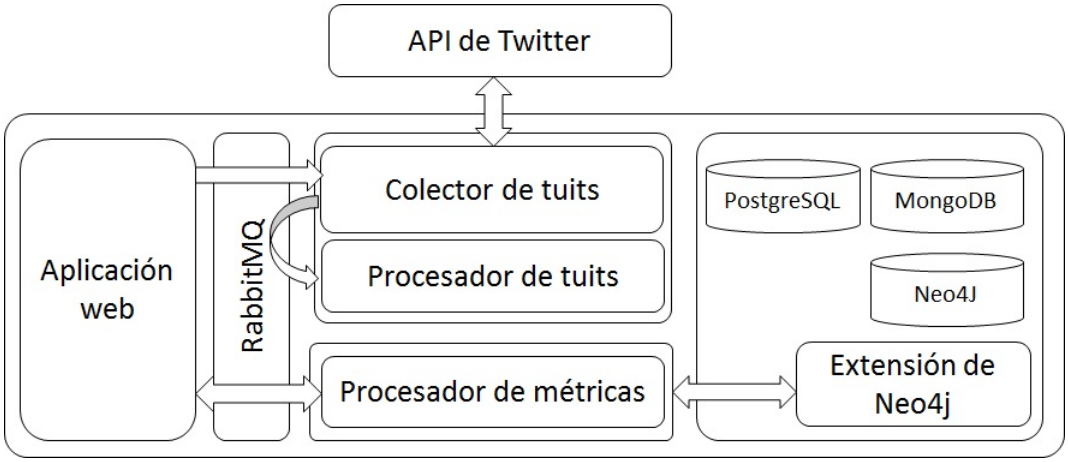

Figura 2.1: Vista general del sistema.

- Colector de tuits: Es la aplicación encargada de interactuar directamente con los servicios de Twitter para la recuperación de los tuits, de acuerdo con los criterios de los estudios de detección de roles, que se encuentren activos.

- Procesador de tuits: Se encarga del procesamiento asíncrono de los tuits que han sido recuperados por el recolector. Tiene entre sus funciones la extracción de los usuarios contenidos en el tuit y de las relaciones que se establecen entre estos. Así como, la persistencia del tuit y del modelo de relaciones obtenido.

- Procesador de métricas: Realiza el cálculo de las métricas de centralidad para la detección de roles, tomando como base al grafo generado para un estudio determinado.

Además de las aplicaciones anteriores, el sistema hace uso de tres sistemas de gestión de bases de datos, PostgeSQL, MongoDB y Neo4J. Para este último fue necesario implementar una extensión, que expone los grafos, vinculados con los estudios de detección de roles, como un flujo de datos. De esta manera se obtiene un bajo acoplamiento entre la aplicación encargada de procesar las métricas y el sistema gestor de bases de datos, a la vez que se obtiene un punto de acceso para la integración con otras aplicaciones externas al sistema.

2.2.1. Integración e intercambio de mensajes

Para el funcionamiento del sistema desarrollado es necesario orquestar la comunicación entre las aplicaciones que lo componen. Esta comunicación debe suceder de tal forma que los eventos relevantes acaecidos dentro de la frontera de cada una de las aplicaciones puedan ser notificados al resto de las aplicaciones, para que en el caso que corresponda se realicen las acciones necesarias.

Como mecanismo de integración fue seleccionado el ofrecido por las colas de mensajería; las cuales se definen como un sistema para la comunicación entre procesos mediante el intercambio de mensajes. El funcionamiento de las colas de mensajería se basa en el patrón productor - consumidor. Siguen el

esquema en el cual el emisor del mensaje realiza el envío de este hacia una cola que puede direccionarlo hacia uno o varios consumidores. El mensaje es el vehículo utilizado para la transmisión de la información, la cual se encuentra contenida en el cuerpo de este (Pivotal, 2017a).

Para la gestión del intercambio de mensajes se utilizó a RabbitMQ, un sistema para el intercambio de mensajes entre procesos gestionado por la compañía Pivotal. RabbitMQ está basado en código abierto, y se encuentra disponible de forma gratuita bajo licencia *Mozilla Public License* (Pivotal, 2017b).

Aunque para la implementación de cada aplicación se utilizaron diferentes tecnologías y lenguajes de programación, la interacción entre estas es posible gracias a las facilidades ofrecidas por RabbitMQ. El uso de colas para el intercambio de mensajes permite la emisión y recepción de notificaciones con un bajo acoplamiento entre el emisor y receptor al hacer uso del estándar *Advanced Message Queuing Protocol*, AMQP (OASIS, 2017). En el cuadro 2.1 aparece una descripción del esquema de intercambios de mensajes que sirve de base para la integración de las aplicaciones que componen el sistema.

El desarrollo de las aplicaciones siguiendo el diseño propuesto por RabbitMQ dota al sistema de un alto grado de flexibilidad, al utilizar a las colas de RabbitMQ como un punto de intercambio y almacenamiento temporal. Las aplicaciones que componen el sistema pueden ser instanciadas más de una vez para ajustar la capacidad de procesamiento en dependencia de la demanda que se tenga en un momento dado.

Cuadro 2.1: Esquema de intercambio de mensajes entre las aplicaciones haciendo uso de RabbitMQ.

Productor	Consumidor	Descripción
Aplicación Web	Colector de tuits	Notificación de la variación en los parámetros para la recuperación de los tuits desde los servicios de Twitter. Tiene lugar mediante un intercambio asíncrono haciendo uso de una sola cola de mensajes.
Aplicación Web	Colector de tuits	Solicitud de cálculo de las métricas de centralidad. Tiene lugar mediante un intercambio asíncrono haciendo uso del patrón *Remote Procedure Call*. El intercambio de mensajes es bidireccional por lo que es necesario emplear dos colas de mensajería.
Colector de tuits	Procesador de tuits	Solicitud de procesamiento del tuit recuperado desde los servicios de Twitter. Tiene lugar mediante un intercambio asíncrono haciendo uso de una sola cola de mensajes.

2.3. Gestión de los requerimientos para la ejecución de estudios sobre colecciones de tuits

Para la ejecución de un estudio sobre un tema o acontecimiento específico que sea de interés en Twitter es necesario conocer con antelación los términos que lo delimitan, así como el intervalo temporal que se desea abarcar. Estos parámetros acotan el alcance del estudio, y al mismo tiempo sientan las bases para la recuperación de los tuits mediante los servicios que Twitter ha puesto a disposición de

la comunidad de desarrolladores, a través de las interfaces de programación de aplicaciones (API, del inglés *application programming interface*).

La agrupación conformada por el conjunto de los términos y el intervalo temporal correspondiente constituyen la unidad básica a partir de la cual se realiza la conformación de una colección de tuits. Los términos puden ser nombres de usuarios, etiquetas o cadenas de texto, que están estrechamente vinculados con el tema abordado en el estudio. Los términos y el intervalo temporal permiten diferenciar cuáles tuits deben ser tomados en cuenta para un estudio específico, en función de los contenidos de los mismos y del momento en que fueron generados. En lo adelante se denominará a esta agrupación de términos e intervalo como "requerimientos para la ejecución del estudio".

Con el objetivo de gestionar los requerimientos para la ejecución del estudio se desarrolló una aplicación web, utilizando el lenguaje de programación orientado a objetos PHP en su versión 7.0. Con el objetivo de agilizar el proceso de desarrollo y hacer uso de las mejores prácticas plasmadas en los marcos de trabajo, se seleccionó a Symfony en su versión 3.4.8.

El acceso al sistema de gestión de bases de datos se implementó haciendo uso de la capa de abstracción para la ejecución del mapeo objeto - relación ofrecida por Doctrine (Doctrine-Project, 2018). Esto favorece el desarrollo de las aplicaciones manteniendo un bajo acoplamiento con el sistema de gestión de base de datos seleccionado, siempre que este se base en el modelo relacional. Como sistema de gestión de bases de datos se seleccionó a PostgreSQL en su versión 9.4.14, aunque al utilizar las mejores prácticas recomendadas por Doctrine este puede ser sustituido por otros, tales como: MySQL, SQL Server o SQLite.

Siguiendo las prácticas recomendadas por el marco de trabajo de Symfony se tiene que el bundle es la unidad básica de agrupación de componentes del mismo. Un bundle es un conjunto estructurado de archivos dentro de un directorio que implementan una sola característica del sistema. Todos los elementos de la implementación de esta característica residen en el interior del bundle (SensioLabs, 2017).

Los componentes de la aplicación fueron implementados dentro de un único bundle encargado de la gestión de la entidad que representa a los requerimientos para la ejecución del estudio. Además el bundle cuenta con la funcionalidad de notificar a la aplicación encargada de la recuperación de los tuits desde el API de Twitter ante la adición de un nuevo requerimiento para la ejecución del estudio o la variación del intervalo de tiempo de uno previamente existente. La emisión de esta notificación se realiza mediante el envío de un mensaje hacia una cola de trabajo de RabbitMQ. En el cuerpo de este mensaje se incluyen los datos de los requerimientos para la ejecución del estudio que temporalmente se encuentren activos.

Las notificaciones son emitidas bajo dos condiciones:

- El usuario realiza de forma directa una modificación o creación de un requerimiento para la ejecución del estudio.

- Se ejecuta como una tarea programada el comando implementado para notificar la variación de alguno de los parámetros para la recuperación de los tuits.

La ejecución de este comando como tarea programada, responde al hecho de que los requerimientos para la ejecución de estudios solo se encuentran activos para la recolección de los tuits durante un período finito de tiempo. Una vez culminado este intervalo temporal se hace necesario notificar a la aplicación encargada de la recolección de los tuits para que realice la correspondiente variación de los parámetros de filtrado. De esta manera se mantiene controlado el flujo de la comunicación con los servicios de Twitter evitando solicitar contenidos que no se correspondan con los estudios que estén activos.

2.4. La recuperación, procesamiento y la persistencia de la información contenida en los tuits

El proceso de interacción con las interfaces de programación del sitio de redes sociales Twitter fue implementado mediante dos aplicaciones. La primera es la encargada de interactuar con los servicios de Twitter directamente en la recuperación de los tuits; mientras que la segunda se ocupa del preprocesamiento de los tuits recuperados, la extracción de la información sobre las relaciones establecidas entre los usuarios y su posterior persistencia.

Ambas aplicaciones fueron implementadas utilizando el lenguaje de programación orientado a objetos JAVA en su versión 8. Las aplicaciones fueron desarrolladas sobre la base de un cliente para RabbitMQ, que se encuentra a la espera de la emisión de una notificación con los parámetros necesarios para la ejecución de la tarea específica que deben acometer.

La separación en dos aplicaciones de las tareas de recuperación de los tuits y su procesamiento se debe al hecho de que bajo ciertas circunstancias puede ocurrir una demora en el procesamiento, por ejemplo ante un incremento en el flujo de datos que se recibe. Esta demora puede tener implicaciones para el mantenimiento de la conexión con los servicios del sitio de Twitter, por lo que debe ser evitada. Con este fin, la aplicación encargada de la recolección de los tuits se especializa en su recepción desde el servicio del sitio de Twitter y realiza el envío de los mismos hacia una cola de mensajería para su posterior procesamiento y persistencia de forma asíncrona, como se muestra en la figura 2.2.

Además del incremento de la cohesión de los componentes en torno a una tarea más específica, se logra una mayor flexibilidad para el manejo de los recursos de la plataforma de cómputo ante situaciones puntuales. En caso de que ocurra un incremento en el volumen de datos a procesar, se puede solventar esta situación mediante la puesta en ejecución de un mayor número de instancias de la aplicación para el procesamiento y persistencia de los tuits, como se muestra en la figura 2.2.

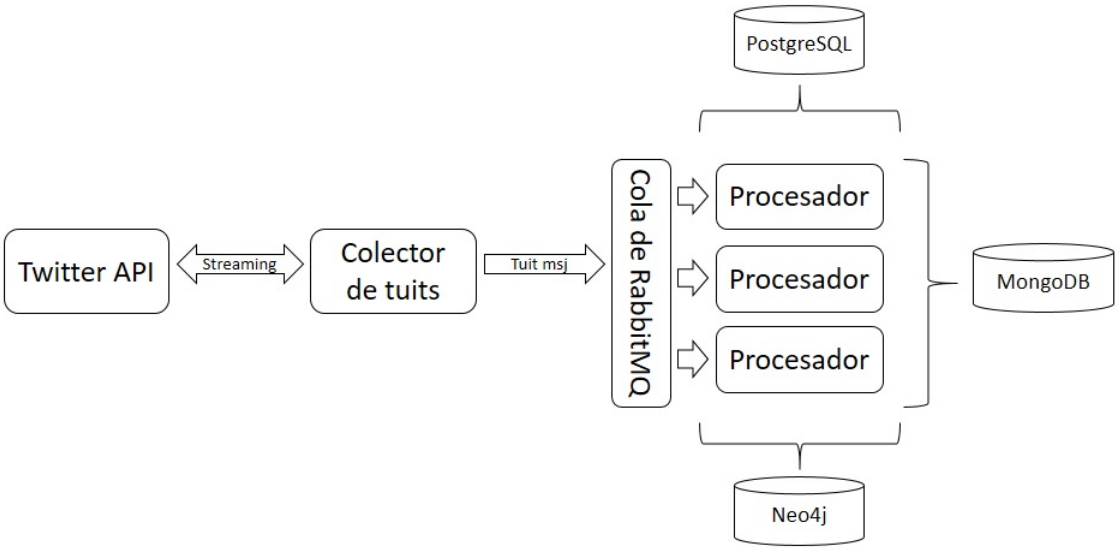

Figura 2.2: Aplicaciones para la captura, procesamiento y persistencia de los tuits.

2.4.1. La recuperación de los tuits

Una vez que son conocidos los requerimientos para la ejecución del estudio es posible proceder a la construcción de la colección de tuits que será el cuerpo del análisis de redes sociales. Con este propósito el sistema cuenta con una aplicación, encargada de interactuar de forma ágil con los servicios de Twitter.

La integración con los servicios de Twitter se implementó sobre la base de la biblioteca Twitter4J; la cual se basa en código abierto y se encuentra disponible para la comunidad de desarrolladores de forma gratuita bajo licencia Apache 2.0 (Twitter4j, 2017).

El sitio de redes sociales Twitter ha puesto a disposición de la comunidad de desarrolladores un conjunto de funcionalidades a través de servicios que permiten la interacción con aplicaciones de terceros. Entre estas se destacan el Filter Tweets API y el PowerTrack API. Ambas se caracterizan por ofrecer las facilidades necesarias para la recuperación de tuits en tiempo real (Twitter, 2017b).

El PowerTrack API es un servicio pensado para el entorno empresarial y para acceder al mismo es necesario contar con ciertos permisos que no están a disposición del público general, los cuales deben ser tramitados ante el departamento de ventas de Twitter. Mientras que el Filter Tweets API es un servicio que puede ser accedido de forma gratuita, aunque no cuenta con las mismas prestaciones y facilidades para el filtrado de los tuits que el PowerTrack API.

El Filter Tweets API permite la recuperación de los tuits, en tiempo real, que coincidan con uno o más de los parámetros de filtrado. Es posible especificar múltiples parámetros para la recuperación de los tuits en una sola petición. Estos parámetros pueden ser (Twitter, 2017c):

- De seguimiento: una lista de identificadores de usuarios separados por coma, para la recuperación de los tuits generados por estos.

- De rastreo: una lista de frases o palabras claves separadas por coma.

- De ubicación geográfica: un conjunto de cuadrantes definidos por coordenadas geográficas.

La longitud de los parámetros aceptados para el filtrado por el Filter Tweets API está limitada hasta 400 palabras claves, 5000 identificadores de usuarios y 25 ubicaciones geográficas (Twitter, 2017b).

La aplicación está desarrollada sobre la base de un cliente para RabbitMQ, que se encuentra a la espera de la emisión de una notificación con los parámetros para la recuperación de los tuits. Una vez que se ha recibido la notificación, se procede a la extracción de los parámetros del cuerpo del mensaje para conformar la consulta que se enviará al servicio de Twitter. La integración con los servicios del sitio de redes sociales Twitter se realiza mediante el Filter Tweets API.

El proceso de conexión con el servicio Filter Tweets API se inicia al conformarse la consulta. Cuando se ha establecido la conexión se procede a la recepción de los tuits. El procesamiento de los tuits, en el caso de que el flujo de arribo sea alto, puede representar una dificultad para el mantenimiento de la conexión, por lo que los tuits una vez recibidos son enviados hacia una cola de trabajo de RabbitMQ para su procesamiento de forma asíncrona. Los tuits son procesados con el objetivo de identificar con cuáles requerimientos para la ejecución del estudio están relacionados, así como para la extracción de los usuarios y las relaciones entre estos que estén presentes en el tuit.

Para acceder al Filter Tweets API es necesario contar con las credenciales de una aplicación registrada en el sitio de redes sociales Twitter. Al intentar establecer una nueva conexión con el Filter Tweets API utilizando unas credenciales que estén siendo empleadas por otra conexión previa, el servicio de Twitter procederá a desconectar la conexión existente. Esta característica del funcionamiento del API de Twitter es utilizada para actualizar los parámetros de filtrado, en tiempo de ejecución. Con este objetivo siempre deben ser ejecutadas un par de instancias de la aplicación para la recolección de los tuits por cada credencial de acceso disponible.

2.4.2. El procesamiento del tuit y la detección de relaciones

Para el procesamiento asíncrono del tuit el sistema cuenta con una aplicación diseñada para detectar los usuarios y la naturaleza de las relaciones que se derivan del tuit, cuyo flujo de trabajo básico puede apreciarse en la figura 2.3. Esta aplicación funciona sobre la base de un cliente de RabbitMQ que se encuentra a la espera del arribo de un mensaje que incluya en su cuerpo el objeto tuit.

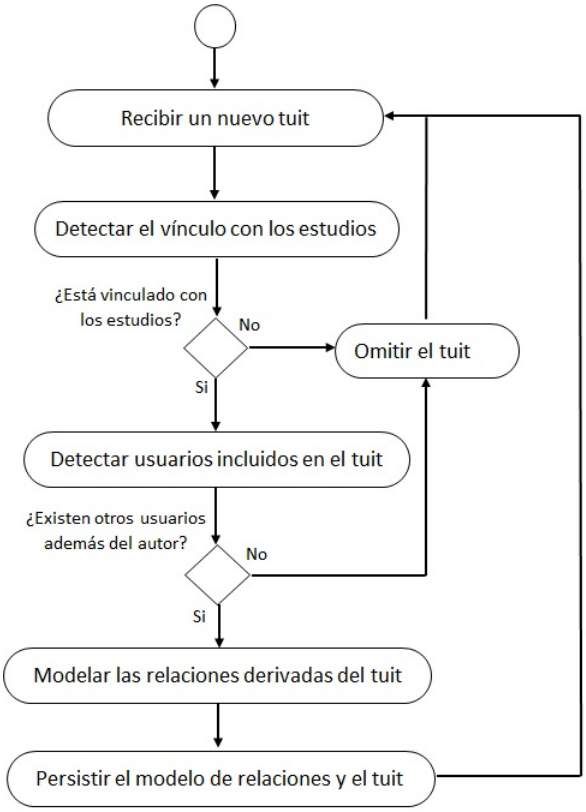

Figura 2.3: Vista del flujo de trabajo del procesador de tuits.

Cada objeto tuit cuenta con un identificador único, el cual consiste en un número entero de 64 bits de longitud. El tuit incluye los atributos “id” y “id_str” para la representación de este valor como un valor numérico y una cadena de texto respectivamente. Siguiendo la recomendación de Twitter durante la implementación de la aplicación se hizo uso del atributo id_str (Twitter, 2017a).

Otro dato de interés que es extraído de la estructura del objeto tuit es la fecha de emisión, la cual se encuentra descrita en el atributo “created_at”. La información del usuario emisor del tuit es extraída del objeto contenido en el atributo “user”. Utilizando este objeto se extraen el nombre de usuario y su número de identificador del usuario. El usuario emisor del tuit cumple el rol de origen para las relaciones de mención, retuit, cita y réplica.

Además del usuario emisor del tuit, otros usuarios pueden estar involucrados como parte de las relaciones implícitas en el mismo. El procesamiento del tuit se realiza con el objetivo de identificar las relaciones de mención, contribución, réplica, cita y retuit, incluidas en la información contenida por los atributos del tuit. Estas relaciones se caracterizan por ser direccionales.

Las relaciones de contribución se derivan del atributo “contributors”. El mismo se compone por una lista de números de identificación, mediante la cual se enumeran los usuarios señalados como contribuyentes con el tuit analizado. A diferencia de las otras relaciones, en esta el usuario emisor del tuit asume el rol de destino en el enlace que se establece. En estas relaciones cada uno de los usuarios enumerados en el atributo “contributors” asume la función del origen.

El tuit incluye un objeto de tipo “entities”, dentro del cual se encuentra el atributo “user_mentions”. Este atributo consiste en una lista con los identificadores de los usuarios que han sido mencionados en el tuit que está siendo procesado. Para la detección de las relaciones de mención se toma como punto de partida al usuario definido como emisor del tuit y como punto de llegada a cada uno de los usuarios enumerados en la lista contenida dentro del atributo “user_mentions” del objeto entidades.

En su estructura el objeto tuit incluye la información relacionada con las réplicas en los atributos descritos en el cuadro 2.2. Para aquellos tuits en que no sean emitidos como una réplica estos atributos tendrán valor nulo. En los tuits que sean una réplica el atributo “in_reply_to_user_id_str” contendrá el identificador del usuario al que se le está replicando. Esta información sirve como base para establecer una relación de réplica en la cual el usuario emisor del tuit que está siendo procesado es tomado como origen y el usuario al que se le ha replicado se toma como destino de la relación.

Cuadro 2.2: Atributos del objeto tuit que contienen información sobre las relaciones de réplica.

Atributo	Descripción
in_reply_to_status_id	En caso de que el tuit sea una réplica contiene el id del primer tuit como un número entero.
in_reply_to_status_id_str	En caso de que el tuit sea una réplica, contiene el id del primer tuit como cadena de texto.
in_reply_to_user_id	En caso de que el tuit sea una réplica, contiene el id del usuario al que se réplica como un número entero.
in_reply_to_user_id_str	En caso de que el tuit sea una réplica, contiene el id del usuario al que se réplica como una cadena de texto.
in_reply_to_screen_name	En caso de que el tuit sea una réplica, contiene el nombre del usuario que generó el tuit original.
reply_count	Indica la cantidad de veces que el tuit ha sido replicado.

Para determinar si un tuit es una cita se cuenta con el atributo “is_quote_status”. Este atributo es de tipo lógico y toma valor verdadero cuando el tuit es producto de una cita y falso en caso contrario. El objeto que representa al tuit original se encuentra descrito en el atributo “quoted_status”. Mediante este objeto es posible extraer el identificador del usuario que emitió el tuit que está siendo citado, a través de la consulta del atributo “user” del mismo. En el cuadro 2.3 se exponen los campos del objeto tuit que contienen información sobre la cita.

En las relaciones de cita el usuario emisor del tuit original, que está siendo citado, le es asignado el rol de receptor. Mientras que al usuario que realiza la cita asume el rol de origen para la relación.

Para conocer si el tuit es producto de una acción de retuit, el objeto tuit incorpora el atributo “retweeted”. Este atributo es de tipo lógico, teniendo valor verdadero para el caso en el cual se haya producido el retuit y falso para el caso contrario. El atributo “retweeted_status” contiene el objeto que representa al tuit que ha sido retuiteado. Este objeto incluye en su atributo “user” los datos del objeto que representa al usuario emisor del tuit original. De estos se extraen el nombre de usuario, contenido en el atributo “screen_name”, y el número identificador, disponible en el atributo “id_str”.

Cuadro 2.3: Atributos del objeto tuit que contienen información sobre las relaciones de cita.

Atributo	Descripción
quoted_status_id	En caso de que el tuit sea una cita, representa el identificador del tuit citado como un número entero.
quoted_status_id_str	En caso de que el tuit sea una cita, representa el identificador del tuit citado como una cadena de texto.
is_quote_status	Indica si el tuit es una cita.
quoted_status	Solo se incluye en los tuits que son citas. Contiene el tuit original.
quote_count	Indica la cantidad de veces, aproximadamente, que el tuit ha sido citado.

Las relaciones de retuit son modeladas tomando como nodo de llegada al usuario emisor del tuit original, y como punto de partida al usuario que es autor del retuit. En el cuadro 2.4 se exponen los atributos del tuit que están vinculados con las relaciones de retuit.

Cuadro 2.4: Atributos del objeto tuit que contienen información sobre las relaciones de retuit.

Atributo	Descripción
retweeted	Indica si el tuit es producto de una acción de retuit.
retweeted_status	Contiene la representación del tuit original que fue retuiteado.
retuit_count	Indica la cantidad de veces que el tuit ha sido retuiteado.

2.4.3. Conformación de la colección de tuits

Los sistemas de gestión de bases de datos no SQL abarcan una amplia variedad de tecnologías, desarrolladas en respuesta a las necesidades de las aplicaciones modernas de una mayor flexibilidad en el manejo de los datos. Entre los sistemas de gestión de bases de datos no SQL sobresalen, desde la perspectiva de la presente investigación, los orientados a documentos y los orientados a grafos. Los primeros se basan en una estructura de datos compleja, denominada documento. Un documento puede contener diferentes pares de llave y valor, pudiendo ser el valor un documento anidado, un arreglo u otro tipo de datos (MongoDB, 2017a).

Por su parte, los sistemas de gestión de bases de datos orientados a grafos, se basan en la interconexión de entidades, o nodos, cada una de las cuales puede contener diversos atributos. Las relaciones que se establecen entre los nodos proveen conexiones direccionadas y semánticamente relevantes, que al igual que los nodos pueden tener atributos (Neo4J, 2017).

En el marco de la presente investigación se seleccionó el sistema de gestión de bases de datos no SQL, orientado a documentos, MongoDB para el almacenamiento íntegro de los tuits y el sistema de gestión de bases de datos no SQL, orientado a grafos, Neo4j para la persistencia de los usuarios y las relaciones extraídas de la estructura de los tuits.

MongoDB es un sistema de gestión de bases de datos no SQL orientado a documentos que está basado en software libre y disponible bajo las licencias *Affero General Public License* y *Apache License*

(MongoDB, 2017b). MongoDB almacena los documentos en formato JSON, lo cual simplifica el trabajo con los datos recibidos directamente desde el servicio de Twitter en este formato. Para la creación del documento de MongoDB además de incluir la representación del objeto tuit, se adiciona un atributo que contiene un listado de los identificadores de los requerimientos para la ejecución del estudio, que se corresponden con el tuit.

2.4.4. Conformación del grafo de relaciones derivadas del objeto tuit

Neo4j es un sistema de gestión de bases de datos no SQL orientado a grafos desarrollado por Neo Technology. Su desarrollo está basado en el código abierto y se encuentra disponible en dos variantes: para la comunidad y empresarial. Para el uso de la versión empresarial es necesario contar con una licencia comercial, mientras que la versión para la comunidad cuenta con una licencia dual, *General Public License v3* y *Affero General Public License* (Neo4J, 2017).

Figura 2.4: Vista de las relaciones entre dos usuarios en Neo4j.

Al hacer uso de Neo4j es posible modelar los usuarios y las relaciones contenidas en el tuit como un grafo para su persistencia en la base de datos. Con este fin se almacenan los datos de los usuarios como atributos de los nodos y las relaciones como aristas entre estos. En Neo4j las relaciones también pueden contener atributos, lo que es aprovechado para almacenar el tipo de relación, el identificador del tuit y los identificadores de los estudios sobre colecciones de tuits que están vinculados con el tuit del cual se extrajo la relación en cuestión. De esta manera se garantiza la recuperación de los grafos para un tipo de relación y estudio específicos. Durante el proceso de persistencia se toman las precauciones necesarias para evitar redundancia en los datos que se almacenan en Neo4j.

2.5. Cálculo de métricas de centralidad sobre colecciones de tuits

El modelo, basado en grafos, del segmento de la red social que fue obtenido durante el procesamiento del tuit, constituye la base para la detección de roles que se realiza mediante el procesador de métricas. Las relaciones entre los componentes de esta aplicación se muestran en la figura 2.5.

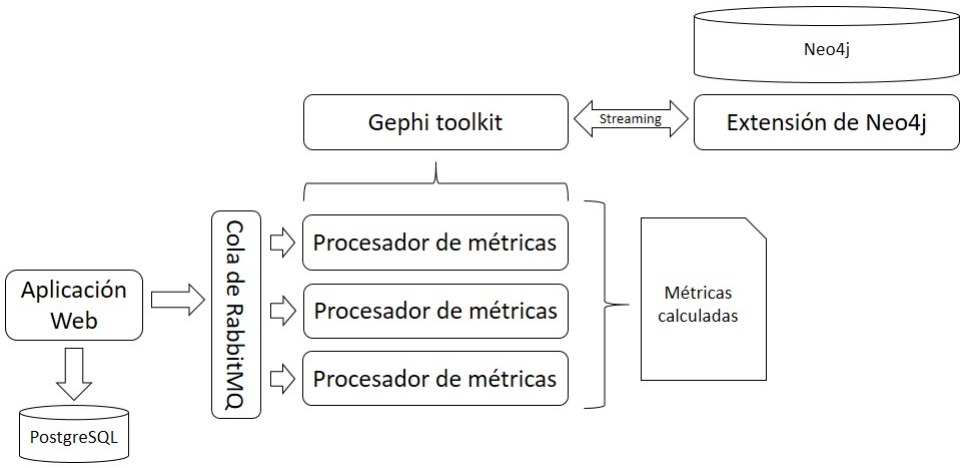

Figura 2.5: Componentes para el cálculo de las métricas de centralidad.

La extensión de Neo4j se realizó con el objetivo de facilitar el acceso a los grafos generados a partir de las relaciones obtenidas del procesamiento de los tuits. Mediante la extensión se cuenta con un mecanismo de abstracción para el acceso a los datos, de tal forma que no es necesario elaborar, ni ejecutar, una consulta de forma explícita en el lenguaje declarativo Cypher para obtener los datos almacenados en Neo4j.

La extensión desarrollada acepta como parámetros el identificador del estudio, y el tipo de la relación. Partiendo de los datos suministrados, se genera una representación de los datos recuperados en formato JSON que es enviada como respuesta, en la forma de un flujo de datos.

El mecanismo de cálculo es desatado por el usuario al realizar una petición desde la aplicación de gestión de los requerimientos para la ejecución del estudio, solicitando los resultados para un estudio específico. La aplicación web envía un mensaje que incluye el número de identificación del requerimiento para la ejecución del estudio hacia una cola de trabajo de RabbitMQ, para el cálculo de las métricas de centralidad y espera por el resultado del procedimiento.

La aplicación encargada del cálculo de las métricas de centralidad se basa en un cliente de RabbitMQ que hace uso del patrón *Remote Procedure Call*. Esta aplicación se comporta como un consumidor de RabbitMQ que se encuentra a la espera de un mensaje con el identificador de la instancia del requerimiento para la ejecución del estudio que se desea procesar. Para cada uno de los tipos de las relaciones se invoca la extensión de Neo4j, incluyendo como parámetros de la petición el identidicador del requerimiento para la ejecución del estudio y el tipo de relación específico que determina el grafo que se desea recuperar.

Para el cálculo de las métricas de centralidad se hizo uso de las funciones incluidas en la librería Gephi Toolkit en su versión 0.9.2. El proyecto Gephi Toolkit empaqueta los módulos “Graph, Layout, IO, Filters y otros” como una biblioteca de Java, de tal forma que puede ser reutilizado dentro de otros proyectos. Estas herramientas permiten extender las funcionalidades de las aplicaciones haciendo uso de las

funciones que Gephi incluye y que pueden ser invocadas desde una línea de comando (Gephi, 2018b). Los resultdos obtenidos del cálculo de las métricas de centralidad son retornados a la aplicación web como un flujo de datos. Los datos recepcionados son procesados para su presentación al usuarios en la vista de resultados de la solicitud de estudio correspondiente en la aplicación web.

2.6. Conclusiones parciales

La arquitectura del sistema comprende cuatro aplicaciones desarrolladas utilizando diferentes tecnologías. Estas aplicaciones, para su funcionamiento necesitan del intercambio de información entre sí. El estándar ofrecido por AMQP brinda una solución para la comunicación entre procesos, los cuales pueden funcionar de forma asíncrona o síncrona, ajustándose a las necesidades del sistema desarrollado.
Al hacer uso de RabbitMQ se cuenta con un mecanismo para el intercambio de mensajes, independiente de las tecnologías sobre las que se implementan el productor y el consumidor, del mensaje. Al construir las aplicaciones altamente cohesionadas en torno a las funciones específicas de cada una, se obtiene un bajo nivel de acoplamiento a la vez que se cuenta con la flexibilidad necesaria para la extensión de las funcionalidades en caso de ser necesario, o la sustitución de una aplicación por otra desarrollada con una tecnología diferente.
Al desarrollarse una aplicación web para la gestión de los requerimientos para la ejecución del estudio se facilita la interacción del usuario con el sistema. La recolección y procesamiento de los tuits, así como el cálculo de las métricas de centralidad se delegan como aplicaciones asíncronas en un segundo plano. Aunque, por la naturaleza de la información manejada, se utilizan varios sistemas de gestión de base de datos la visibilidad de estos queda acotada para cada aplicación.

Capítulo 3.

Análisis de los resultados

3.1. Introducción

En este capítulo se describe la validación del resultado de la presente investigación. Para esto se hizo uso de diferentes métodos, a continuación se expone la forma en que estos fueron aplicados y su propósito específico en la validación. Teniendo en cuenta las características de la investigación se aplicaron métodos de naturaleza cualitativa y cuantitativa, que fueron triangulados para asegurar una mayor precisión y objetividad en los resultados.

Para la validación del resultado, los métodos utilizados fueron:

1. Cuasi experimento: Con el objetivo de comprobar en qué medida la detección de roles sobre colecciones de tuit incrementa la capacidad de descripción de segmentos de usuarios de Twitter por parte del Departamento de Operaciones Web y Análisis de Información.

2. Técnica de Iadov: Aplicada a una muestra de profesores y especialistas del Departamento de Operaciones Web y Análisis de Información para medir mediante el índice de satisfacción grupal del nivel de aceptación del sistema desarrollado, por parte de los usuarios finales.

3. Entrevista en profundidad: Con el objetivo de verificar la posibilidad de aplicar el sistema desarrollado en el marco del Departamento de Operaciones Web y Análisis de Información, desde el punto de vista de la experiencia de sus especialistas.

4. Grupo focal: Con la participación de especialistas y profesores vinculados al análisis de redes sociales en el Departamento de Operaciones Web y Análisis de Información, específicamente en el trabajo con Twitter. Se diseñó como una guía de preguntas con el objetivo de validar que el sistema para la detección de roles sobre colecciones de tuit, permite elevar la capacidad de descripción de segmentos de usuarios de este sitio de redes sociales.

5. Triangulación intermétodo simultánea: Para la contrastación de los resultados obtenidos de la aplicación de las técnicas anteriormente enunciadas.

3.2. Descripción de la validación de la hipótesis

Para la investigación se tomó como hipótesis que: “El sistema para la detección de roles sobre segmentos de usuarios de la red social Twitter incrementará la capacidad de descripción de estos segmentos por parte del Departamento de Operaciones Web y Análisis de Información”. Dicha hipótesis pleanteada se caracteriza por ser bivariada.

Se identificó como variable independiente: “La detección de roles sobre segmentos de usuarios de la red social Twitter” y se toma como dominio de valores de la misma a los siguientes:

- Nivel nulo: Caracterizado por la inexistencia de herramientas propias para la detección de roles sobre segmentos de usuarios de la red social de Twitter.
- Nivel A: Caracterizado por la existencia de las facilidades necesarias para la detección de roles sobre segmentos de usuarios de la red social de Twitter.

Como variable dependiente se tiene: “la capacidad de descripción de los segmentos de usuarios de la red social Twitter por parte del Departamento de Operaciones Web y Análisis de Información” y se toma como dominio de valores de la misma los siguientes:

- Básico: Solo es posible describir los segmentos de usuarios en base a la centralidad de grado de entrada y salida para las relaciones de mención y retuit, haciendo uso de herramientas de terceros.
- Mejorado: Es posible describir los segmentos de usuarios en base al valor de las métricas de centralidad de intermediación, PageRank, Hits, eigenvector, media armónica, excentricidad, cercanía, grado de entrada y grado de salida, para las relaciones de mención, retuit, réplica, cita y contribución.

La naturaleza de estas variables fue tomada en cuenta para el diseño de los métodos de validación aplicados. Al culminarse el proceso de validación los resultados obtenidos son utilizados para determinar el incremento o no de la capacidad de descripción de segmentos de usuarios por parte del departamento.

3.3. Cuasi - experimento

La investigación cuasi - experimental proviene del ámbito educativo y de la psicología, donde la investigación de ciertos fenómenos no podía desarrollarse siguiendo los procedimientos experimentales. El término cuasi - experimento se refiere a diseños experimentales en los cuales los sujetos de estudio no están asignados de forma aleatoria (Campbell y Riecken, 1968) (Manterola y Otzen, 2015).

En el marco de la investigación se desarrolló un cuasi - experimento del tipo de diseño con preprueba - postprueba con grupos intactos. Para esto se tomó como grupo intacto a los especialistas y profesores vinculados al Departamento de Operaciones Web y Análisis de Información, encargados de realizar los estudios sobre Twitter. Para la realización de las pruebas se puso en ejecución una instancia del sistema, en un entorno controlado y solo accesible desde la subred del laboratorio del departamento.

El estudio se dividió en dos momentos, anterior y posterior al uso del sistema desarrollado para la detección de roles. Primeramente se recopilaron los componentes empleados para la descripción cuantititativa de un segmento de usuarios de la red social de Twitter, como se muestra en el cuadro 3.1. Posteriormente se les explicó a los especialistas las características del sistema desarrollado y como

acceder a las funcionalidades que el mismo ofrece. Seguidamente los especialistas interactuaron con el sistema e insertaron en el mismo varios requerimientos para la ejecución de estudios que contenían etiquetas y palabras claves relacionadas con temas de interés en Twitter.

Cuadro 3.1: Elementos cuantitativos para la descripción de segmentos de usuarios de la red social de Twitter, a disposición de DOWAI, previo al sistema para la detección de roles.

Relaciones	Seguimiento	Mención	Cita	Retuit	Contribución	Réplica
Grado de entrada	Si	Si	Si	Si	No	No
Gado de salida	Si	Si	Si	Si	No	No
Intermediación	No	No	No	No	No	No
Vector normal	No	No	No	No	No	No
HIT	No	No	No	No	No	No
PageRank	No	No	No	No	No	No
Media armónica	No	No	No	No	No	No
Excentricidad	No	No	No	No	No	No
Cercanía	No	No	No	No	No	No

Una vez introducidos los requerimientos para la ejecución de estudios, el sistema procedió a recolectar los tuits. Los tuits recibidos fueron procesados y almacenados para la extracción de las relaciones entre los usuarios que se encuentran implícitas en su estructura. Estas relaciones constituyen la base del grafo sobre el que se determinan las métricas de centralidad para cada nodo. Para cada uno de los requerimientos para la ejecución de estudios se calcularon estas métricas en tiempo cercano al real, siendo los resultados expuestos en forma de tabla en la vista correspondiente de la aplicación web.

Los resultados ofrecidos por el sistema de detección de roles fueron aceptados como positivos por los especialistas del departamento, quienes en momentos posteriores participaron en la aplicación de otras técnicas como la entrevista en profundidad, el grupo focal y la técnica de Iadov. Al finalizar el segundo momento del cuasi - experimento, se comprobó que los resultados obtenidos mediante el cálculo de las métricas de centralidad para la determinación de los roles de los usuarios, incrementan la capacidad de descripción de segmentos de usuarios de Twitter por parte del departamento, como se muestra en el cuadro 3.2.

Cuadro 3.2: Elementos cuantitativos para la descripción de segmentos de usuarios de la red social de Twitter, a disposición de DOWAI, posterior al sistema para la detección de roles.

Relaciones	Seguimiento	Mención	Cita	Retuit	Contribución	Réplica
Grado de entrada	Si	Si	Si	Si	Si	Si
Gado de salida	Si	Si	Si	Si	Si	Si
Intermediación	No	Si	Si	Si	Si	Si
Vector normal	No	Si	Si	Si	Si	Si
HIT	No	Si	Si	Si	Si	Si
PageRank	No	Si	Si	Si	Si	Si
Media armónica	No	Si	Si	Si	Si	Si
Excentricidad	No	Si	Si	Si	Si	Si
Cercanía	No	Si	Si	Si	Si	Si

3.4. Resultados de la técnica de Iadov

La técnica de Iadov fue creada para el estudio de la satisfacción en las carreras pedagógicas (Kuzmina, 1970), aunque ha demostrado su utilidad en otras áreas de las ciencias. Mediante esta técnica se puede determinar, de forma indirecta, el grado de satisfacción de los individuos involucrados en el proceso que está siendo objeto de análisis. La técnica de Iadov se basa en un cuestionario conformado por cinco preguntas, de las cuales tres son cerradas y dos abiertas. Las preguntas cerradas guardan una relación entre sí, que previamente no es de conocimiento por parte del sujeto al que se le aplica la técnica.
Las tres perguntas cerradas se relacionan mediante el cuadro lógico de Iadov. El cuadro lógico utilizado en la investigación se muestra en el cuadro 3.3. Las respuestas a cada una de estas preguntas permiten determinar la posición de cada sujeto en la escala de satisfacción que toma valores desde 1 hasta 6.

Cuadro 3.3: Cuadro de Iadov utilizado.

	¿Estima provechosa la herramienta desarrollada para el trabajo que se realiza en el departamento?								
	Si			No se			No		
¿Qué tan satisfecho se siente con los resultados de la herramienta desarrollada para la detección de roles de usuarios en Twitter?	¿Utilizaría la herramienta desarrollada para la detección de roles en Twitter para estudiar las interacciones entre los usuarios?								
	Si	No se	No	Si	No se	No	Si	No se	No
Me satisface	1	2	6	2	2	6	6	6	6
Me resulta más satisfactorio que insatisfactorio	2	2	3	2	3	3	6	3	6
Me son indiferentes	3	3	3	3	3	3	3	3	3
Me resulta más insatisfactorio que satisfactorio	6	3	6	3	4	4	3	4	4
No me satisfacen en lo absoluto	6	6	6	6	4	4	6	4	5
No se que decir	2	3	6	3	3	3	6	3	4

Para medir el grado de satisfacción de los usuarios respecto al sistema desarrollado, se tomó como muestra a ochos especialistas y profesores vinculados a DOWAI. La selección se realizó teniendo en cuenta la experiencia como analista en el trabajo con el sitio de redes sociales Twitter, y el dominio de técnicas de análisis de redes sociales, entre otros aspectos. Los resultados obtenidos para la satisfacción de forma individual se exponen en el cuadro 3.4.

Cuadro 3.4: Escala del índice de satisfacción individual.

Satisfacción	Escala	Participantes en la escala
Clara satisfacción	1	4
Más satisfecho que insatisfecho	2	3
No definido	3	1
Más insatisfecho que satisfecho	4	0
Clara insatisfacción	5	0
Contradictorio	6	0

Esta técnica permite determinar el índice de satisfacción grupal (ISG), que representa los niveles de

satisfacción en una escala numérica que abarca el intervalo desde -1 hasta 1, como se muestra en el cuadro 3.5.

Cuadro 3.5: Escala del índice de satisfacción grupal.

Resultado	Desde	Hasta
Satisfacción	0.50	1.00
Contradicción	0.49	-0.49
Insatisfacción	-0.50	-1.00

Para la determinación del índice de satisfacción grupal se utiliza la fórmula expuesta en la ecuación 3.1. Las variables representan las cantidades de participantes agrupados por las escalas del índice de satisfacción individual. La cantidad de participantes que expresaron tener una clara satisfacción son representados por a, la cantidad que se sienten más satisfechos que insatisfechos se expresan mediante b, los que evidencian contradicción mediante c, los que se sienten más insatisfechos que satisfechos mediante d, y e es la cantidad de participantes que expresan una clara insatisfacción. El valor de n representa el total de participantes.

$$ISG = \frac{(a * 1{,}0) + (b * 0{,}5) + (c * 0) + (d * -0{,}5) + (e * -1{,}0)}{n} \tag{3.1}$$

El cálculo del índice de satisfacción grupal arrojó un resultado de 0.69, lo que evidencia que existe satisfacción con el sistema desarrollado y se reconoce su utilidad para incrementar la capacidad de descripción de segmentos de usuarios del sitio de redes sociales Twitter por parte del Departamento de Operaciones Web y Análisis de Información.
El formulario presentado a los participantes incluyó dos preguntas abiertas, mostradas a continuación:

- ¿Qué importancia le concede al conocimiento de las interacciones entre los usuarios en Twitter?
- ¿A su juicio cuál es el impacto que tienen las métricas de centralidad en su apreciación del comportamiento de los usuarios en Twitter?

Sobre la primera pregunta, los participantes manifestaron la relevancia del conocimiento de las interacciones entre los usuarios en Twitter para sacar un mejor provecho de la red social. Para esto se citó como ejemplo la estructuración de campañas de comunicación con una mayor efectividad. También manifestaron su utilidad para el estudio de la expresión de los fenómenos sociales en este sitio de redes sociales, así como la evolución de un usuario específico en su relación con el entorno. Fue expuesta la importancia de las acciones realizadas por el usuario de forma activa sobre los tuits, como una expresión del interés o la sensibilidad del usuario ante la información implícita en el mismo.
Sobre la segunda pregunta los participantes estuvieron de acuerdo en la utilidad de la detección de roles basada en métricas de centralidad para el análisis de redes sociales. A la vez que reconocieron

que de esta forma se obtiene una perspectiva de la estructura de la red social que permite evaluar el comportamiento de cada actor, en su relación con el colectivo, sobre la base de las relaciones más allá de la cantidad de seguidores.

Se planteó que tomando en cuenta la información aportada por el cálculo de las métricas de centralidad, es posible realizar acciones con el objetivo de optimizar el flujo de contenidos dentro de un segmento de usuarios de Twitter. Contar con las métricas de centralidad como forma cuantitativa de medir el comportamiento de los usuarios permite identificar aquellos que son relevantes para la dinámica del segmento de la red social que esté siendo estudiado.

3.5. Entrevista en profundidad

La entrevista en profundidad es un método de investigación cualitativo que se basa en uno o varios encuentros personales entre el entrevistador y el informante que siguen un modelo de conversación entre iguales en un ambiente de confianza (Taylor y Bogdan, 2008).

La entrevista en profundidad se realizó con el objetivo de conocer, desde la experiencia acumulada por los entrevistados, la aplicabilidad del sistema para la detección de roles sobre colecciones de tuits por parte del Departamento de Operaciones Web y Análisis de Información. En calidad de entrevistados fueron seleccionados dos especialistas de dicho departamento que trabajan directamente con el sitio de redes sociales de Twitter. Las preguntas sobre las que versó la entrevista fueron:

- ¿Estima conveniente conocer el rol que desempeñan los usuarios dentro de un segmento de la red social de Twitter?
- ¿Considera que además del seguimiento, las relaciones de mención, retuit, réplica, cita y contribución, contienen información relevante sobre la dinámica de la red social de Twitter?
- ¿La capacidad para describir a un segmento de usuarios de la red social de Twitter dado un tema específico se vería incrementada al utilizar la detección de roles, basada en métricas de centralidad?
- ¿Estima factible y conveniente utilizar el sistema para la detección de roles sobre colecciones de tuit por parte del Departamento de Operaciones Web y Análisis de Información?

Las entrevistas se realizaron en distintos momentos y por separado. A las preguntas realizadas los entrevistados respondieron positivamente. En el caso de la primera pregunta se manifestó que el conocimiento del rol de cada actor en la red social permite tener una visión más clara del alcance del intercambio de contenidos en la red. Uno de los entrevistados puso como ejemplo que al visualizarse el grafo generado para un tema específico se identifican las estructuras de las relaciones entre los participantes y que esta es una información valiosa para intencionar el flujo de información en la red social.

Sobre la segunda pregunta los entrevistados manifestaron, que si bien el seguimiento es un factor para evaluar la importancia de una cuenta en la red social, existen otros elementos que aportan información sobre su relevancia en un momento dado. Las relaciones de seguimiento suelen mantenerse por un largo período de tiempo, en comparación con otras que tienen solo un carácter circunstancial pero que implican acciones directas por parte del usuario en Twitter. Este tipo de relaciones permiten contextualizar la dinámica de la red para un intervalo temporal determinado.

En el caso de la tercera pregunta se manifestó que contar con el sistema implementado permite enriquecer las descripciones resultantes de los estudios realizados en Twitter sobre temas específicos. La descripción de los roles sobre la base de las métricas de centralidad para los segmentos de usuarios de Twitter, aporta una valoración cuantitativa que incrementa las posibilidades de una representación detallada de un segmento de usuarios temáticamente determinado.

Al abordar la última pregunta, en las respuestas se puso de manifiesto la conveniencia de contar con un sistema elaborado en el propio centro que permita visibilizar y evaluar el rol que desempeñan los usuarios en el sitio de redes sociales Twitter sobre un tema específico, con el fin de optimizar la proyección del departamento en esta red social. También se señaló la importancia del almacenamiento de los tuits y las relaciones entre los usuarios para conducir otros tipos de estudios por parte del departamento.

3.6. Grupos focales

La técnica de grupos focales es una forma de entrevista grupal que utiliza la comunicación entre el investigador y los participantes para explorar los conocimientos, experiencias y puntos de vista de los participantes (Hernández Sampieri et al., 2010) (Hamui-Sutton y Varela-Ruiz, 2013). Los grupos focales se centran en una temática específica, se desarrollan siguiendo una guía de preguntas abiertas, siendo el intercambio entre los participantes dirigido por un moderador, quien debe inspirar confianza entre los participantes y abstenerse de tomar partido durante el intercambio.

Para la ejecución del grupo focal fueron seleccionados cinco especialistas y profesores con experiencia en el análisis de redes sociales vinculados con el trabajo desarrollado por el Departamento de Operaciones Web y Análisis de Información. Los participantes son graduados de carreras de las ciencias informáticas y de las ciencias humanísticas, dos de los cuales cuentan con el grado científico de máster. La actividad se desarrolló siguiendo un protocolo mediante el cual se previó mediante la realización de una entrevista abierta y estructurada, en la cual todos los participantes tuvieran espacio para expresar sus criterios sobre el tema tratado.

El debate estuvo dirigido por el moderador en base a una guía de preguntas, confeccionada con el objetivo de conocer sus posiciones sobre la pertinencia del sistema desarrollado para la detección de roles sobre colecciones de tuits y el impacto del mismo en la capacidad de descripción de segmentos de usuarios de Twitter por parte del Departamento de Operaciones Web y Análisis de Información.

Durante el debate se evidenció un consenso en torno de la pertinencia del sistema desarrollado. Se

emitieron valoraciones positivas sobre la influencia de este en la capacidad de descripción del comportamiento de segmentos de usuarios de Twitter en relación con un tema específico. Se planteó la posibilidad de elaborar un índice de relevancia para los usuarios que tome como base y unifique los valores de las métricas de centralidad calculadas. Este planteamiento fue tomado e incluido en las recomendaciones, puesto que el mismo se encuentra más allá del alcance de esta investigación.
De los cinco participantes en el estudio tres no expresaron interés por la inclusión de alguna nueva característica, argumentando que el sistema desarrollado les ofrece información de provecho para una mejor comprensión de las dinámicas de las relaciones y acciones entre los usuarios sobre un tema determinado. Mientras que dos participantes expresaron sus recomendaciones consistentes en:

- Que el sistema permita realizar análisis sobre temas y eventos que tuvieron lugar en intervalos temporales previos.
- Que el sistema permita realizar análisis por intervalos de tiempo y usuarios, de tal forma que se pueda observar en el tiempo la evolución de un usuario determinado en el marco de una métrica y relación puntual.

En el caso de la primera recomendación para su implementación es necesario utilizar otros servicios del API de Twitter, diferentes al utilizado en el sistema desarrollado. Los servicios de Twitter que se encuentran abiertos, de forma gratuita, para el público en general no están diseñados para este fin. El servicio que más se acerca a lo solicitado solo realiza búsquedas sobre una muestra de tuits enmarcada en los últimos 7 días, y se centra en la relevancia no en la integridad de los datos (Twitter, 2017d).
Sobre la segunda recomendación, para su implementación es necesario utilizar otros sistemas de gestión de bases de datos basados en series temporales u otras estructuras de datos sobre sistemas relacionales que están más allá del alcance de la presente investigación. No obstante ambas contribuciones se tienen en cuenta como recomendaciones para la continuidad de la investigación.

3.7. Triangulación metodológica

La triangulación es un término utilizado originalmente en el ámbito de la navegación, para localizar una posición desconocida tomando como referencia múltiples puntos. En 1959 Campbell y Fiske son los primeros en utilizar la triangulación en la investigación (Valencia, 2000).
La triangulación metodológica puede ser secuencial o simultánea. La triangulación metodológica secuencial tiene lugar cuando el resultado de un método es esencial para la planificación del siguiente método. La triangulacion metodológica simultánea tienen lugar al utilizar métodos cualitativos y cuantitativos al mismo tiempo (Morse, 1991).
La triangulación metodológica intermétodos se refiere a la aplicación de diversos métodos, en el marco de una misma investigación para recaudar información contrastando los resultados de cada uno de estos, analizando coincidencias y diferencias. Mediante este tipo de triangulación es posible combinar

métodos de naturaleza cuantitativa y cualitativa, con el objetivo de verificar si se arriban a las mismas conclusiones (Hussein, 2009) (Aguilar Gavira y Barroso Osuna, 2015).

En la investigación se utilizó la triangulación intermétodos simultánea, tomando como base el empleo de técnicas de naturaleza cualitativa y cuantitativa para validar el resultado obtenido, tal y como se muestra en el cuadro 3.6. Como se muestra en el cuadro 3.7 al analizar los resultados obtenidos mediante las técnicas cuantitativas y cualitativas, se puede apreciar que existe una convergencia que nos permite afirmar la validez de la hipótesis planteada al inicio de la investigación.

Cuadro 3.6: Objetivos a evaluar y métodos empleados en la triangulación metodológica.

	De naturaleza	
Objetivo	cuantitativa	cualitativa
Incrementar la capacidad de descripción de segmentos de usuarios de Twitter por parte de DOWAI.	Técnica de Iadov: Nivel de satisfacción de los especialistas y profesores del departamento con el sistema desarrollado.	Grupo focal: Para evaluar la aplicabilidad del sistema desarrollado y la valoración del aporte del mismo para la descripción de segmentos de usuarios de Twitter, por parte de los especialistas y profesores del departamento.
	Cuasi-experimento: Para comprobar el impacto del sistema desarrollado en la capacidad del departamento para la descripción de segmentos de usuarios de Twitter.	Entrevista en profundidad: Para conocer, tomando como base la experiencia de los especialistas y profesores de DOWAI, el impacto de los resultados ofrecidos por el sistema desarrollado en la descripción de segmentos de usuarios de Twitter.

Cuadro 3.7: Comparación entre los resultados obtenidos en cada técnica

Técnicas			
Iadov	Cuasi-experimento	Entrevista en profundidad	Grupo focal
Existe satisfacción con el sistema desarrollado.	Se evidenció un incremento en la capacidad de descripción de segmentos de usuarios de Twitter por parte de DOWAI.	Los resultados ofrecidos por el sistema se valoran positivamente por los especialistas y profesores de DOWAI.	Los participantes manifestaron que el sistema desarrollado es aplicable al departamento y contribuye a la descripción de segmentos de usuarios de Twitter.

3.8. Conclusiones parciales

Mediante la ejecución del cuasi - experimento, se comprobó que el sistema desarrollado contribuye al incremento de la capacidad de descripción de segmentos de usuarios del sitio de redes sociales de Twitter. Los resultados obtenidos de la aplicación de la técnica de Iadov permitieron determinar la

aceptación del sistema desarrollado por los usuarios finales del mismo.

Al aplicar las técnicas de entrevistas en profundidad y grupo focal se pudo comprobar que el uso de las métricas de centralidad sobre las relaciones de retuit, réplica, cita, mención y colaboración, es valorado de forma positiva por parte de los especialistas y profesores de DOWAI.

Al aplicar la triangulación metodológica intermétodo simultánea a los resultados obtenidos del cuasi - experimento, la técnica de Iadov, la entrevista en profundidad y el grupo focal, se observa una coincidencia que permite afirmar que el sistema desarrollado incrementa la capacidad de descripción de los segmentos de usuarios de Twitter por parte del departamento.

Conclusiones

Como resultado de la investigación se arribó a las siguientes conclusiones finales:

- Los roles desempeñados por los usuarios de Twitter, se caracterizan por ser variables en el tiempo, y estar enmarcados en un contexto determinado; sobre la base de los diferentes tipos de relaciones que pueden ser establecidos en este sitio de redes sociales.
- La ejecución del análisis de redes sociales centrado en el actor es viable mediante el estudio de las relaciones de cita, mención, retuit, réplica y colaboración.
- Como resultado del estudio realizado se comprobó que los roles de popular, sociable, intermediario y aglutinador permiten describir cuantitativamente el comportamiento de los usuarios para la ejecución de estudios sobre segmentos de la red social que estén temática y temporalmente acotados.
- El desarrollo del sistema siguiendo una arquitectura distribiuda, lo dota de un alto grado de flexibilidad y bajo acoplamiento; ganando en escalabilidad de acuerdo a la demanda de procesamiento para cada aplicación de forma puntual.
- El sistema desarrollado incrementa la capacidad de descripción del Departamento de Operaciones Web y Análisis de Información, al incorporar nuevos descriptores, extendiendo los estudios más allá de la vecindad del actor, y permitiendo el descubrimiento de nuevos usuarios relevantes de acuerdo al contexto de los estudios.

Recomendaciones

Como resultado de la investigación, y en especial de la interacción de los usuarios finales con el sistema desarrollado, se derivaron las siguientes recomendaciones:

- Generar un índice de relevancia global para los usuarios tomando como base los resultados obtenidos de las métricas de centralidad.
- Extender el sistema para la ejecución de análisis por intervalos de tiempo y usuarios, de tal forma que se pueda apreciar la evolución de un usuario determinado en el marco de una métrica y relación puntual.

Anexo I

Cuadro 8: Especificación de los requisitos funcionales del sistema.

Requisito	Descripción
RF 1	Listar los requerimientos para la ejecución de los estudios.
RF 2	Crear nuevos requerimientos para la ejecución del estudio.
RF 3	Modificar el intervalo temporal de un requerimiento para la ejecución del estudio.
RF 4	Eliminar un requerimiento para la ejecución del estudio.
RF 5	Notificar la creación de un requerimiento para la ejecución del estudio a otras aplicaciones.
RF 6	Notificar la variación del intervalo temporal de un requerimiento para la ejecución del estudio a otras aplicaciones.
RF 7	Mostrar los resultados de las métricas de centralidad calculadas dados el estudio y un tipo de relación.
RF 8	Recuperar los tuits generados desde el sitio de redes sociales Twitter que se correspondan con los requerimientos para la ejecución del estudio.
RF 9	Recibir notificaciones sobre las variaciones de los requerimientos para la ejecución del estudio.
RF 10	Mantener la conexión con el sitio de redes sociales Twitter ante una modificación de los parámetros de filtrado.
RF 11	Extraer los datos de los usuarios implicados en el tuit analizado.
RF 12	Extraer las relaciones de mención.
RF 13	Extraer las relaciones de retuit.
RF 13	Extraer las relaciones de cita.
RF 14	Extraer las relaciones de réplica.
RF 15	Extraer las relaciones de coautoría.
RF 16	Almacenar los usuarios identificados y las relaciones establecidas entre estos.
RF 17	Almacenar integramente el tuit.
RF 18	Calcular la intermediación para cada usuario dados una relación y un estudio.
RF 19	Calcular la centralidad de grado de entrada para cada usuario dados una relación y un estudio.
RF 20	Calcular la centralidad de grado de salida para cada usuario dados una relación y un estudio.
RF 21	Calcular la métrica de HITS para cada usuario dados una relación y un estudio.
RF 22	Calcular la métrica de eigenvector para cada usuario dados una relación y un estudio.
RF 23	Calcular la métrica de media armónica para cada usuario dados una relación y un estudio.
RF 24	Calcular la excentricidad para cada usuario dados una relación y un estudio.
RF 25	Calcular la métrica PageRank para cada usuario dados una relación y un estudio.

Bibliografía

John Arundel Barnes. Class and committees in a norwegian island parish. *Human relations*, 7(1):39–58, 1954.

Statistics ITU. Key ict indicators for developed and developing countries and the world (totals and penetration rates). *URL: http://www. itu. int/ITUD/ict/statistics/at_glance/KeyTelecom. html*, 29:2012, 2011.

Danah M. Boyd y Nicole B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007. ISSN 1083-6101. doi: 10.1111/j.1083-6101.2007.00393.x. URL http://dx.doi.org/10.1111/j.1083-6101.2007.00393.x.

Twitter. Preguntas frecuentes sobre el seguimiento, 2018a. URL https://help.twitter.com/es/using-twitter/following-faqs.

Statista. Twitter: number of active users 2010-2018, 2018. URL https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

Twitter. About, 2018b. URL https://about.twitter.com/en_us/company.html.

Waldo Barrera Martínez. Redes sociales de internet: una visión histórica de su impacto en los movimientos sociales contemporáneos (1995-2016). Master's thesis, Uninversidad de La Habana, 2017.

Feng Xiong, Acklesh Prasad, y Larelle June Chapple. The economic consequences of corporate financial reporting on twitter. In *7th Conference on Financial Markets and Corporate Governance Conference*, Melbourne, Vic, March 2016. URL http://eprints.qut.edu.au/93196/.

Mimi Zhang, Bernard J. Jansen, y Abdur Chowdhury. Business engagement on twitter: a path analysis. *Electronic Markets*, 21(3):161, 2011. ISSN 1422-8890. doi: 10.1007/s12525-011-0065-z. URL http://dx.doi.org/10.1007/s12525-011-0065-z.

Sen Pei y Hernán A Makse. Spreading dynamics in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(12):P12002, 2013.

Sen Pei, Flaviano Morone, y Hernán A Makse. Theories for influencer identification in complex networks. *arXiv preprint arXiv:1707.01594*, 2017.

Jordi Paniagua y Juan Sapena. Business performance and social media: Love or hate? *Business horizons*, 57(6):719–728, 2014.

R.P. Bhatt, R. Rastogi, V.S. Chaoji, y S. Ranu. Method and system for maximizing content spread in social network, October 11 2012. URL https://www.google.com/patents/US20120259915. US Patent App. 13/080,661.

D.C. Hebenthal, C.J. Saretto, K.P. Mulcahy, y J.E. Allard. Following online social behavior to enhance search experience, April 19 2012. URL https://www.google.com/patents/US20120095976. US Patent App. 12/903,865.

J. Chen, K. Lee, y J.U. Mahmud. Selecting strangers for information spreading on a social network, February 2 2016. URL https://www.google.com/patents/US9251475. US Patent 9,251,475.

Takeshi Sakaki, Makoto Okazaki, y Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772777. URL http://doi.acm.org/10.1145/1772690.1772777.

Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, y Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10(1):178–185, 2010.

Yu Wang, Yang Feng, Xiyang Zhang, Richard Niemi, y Jiebo Luo. Will sanders supporters jump ship for trump? fine-grained analysis of twitter followers. *CoRR*, abs/1605.09473, 2016. URL http://arxiv.org/abs/1605.09473.

Jieun Shin, Lian Jian, Kevin Driscoll, y François Bar. Political rumoring on twitter during the 2012 us presidential election: Rumor diffusion and correction. *New Media & Society*, 0(0):1461444816634054, 2016. doi: 10.1177/1461444816634054. URL http://dx.doi.org/10.1177/1461444816634054.

Víctor Manuel Pérez Martínez, María Dolores Rodríguez González, y María Tobajas Gracia. Movilización y participación en twitter. estudio de caso del hashtag# supertuesday en las primarias presidenciales de eeuu 2016. *Revista Latina de comunicación social*, (72):679–703, 2017.

José Gabriel Espinosa, Katherine Tarazona, y Miguel Ormanza. Roles en comunidades de interés sobre el servicio de redes sociales de twitter. *Congreso Ciencia y Tecnología*, (11):188–194, 2016. ISSN 1390-4663.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, y Jiebo Luo. Rumor detection on twitter pertaining to the 2016 U.S. presidential election. *CoRR*, abs/1701.06250, 2017. URL http://arxiv.org/abs/1701.06250.

B Amor, S Vuik, Ryan Callahan, Ara Darzi, Sophia N Yaliraki, y Mauricio Barahona. Community detection and role identification in directed networks: understanding the twitter network of the care data debate. *Dynamic Networks and Cyber-Security*, 1:111, 2016.

Axel Bruns, Tim Highfield, y Jean Burgess. The arab spring and social media audiences. *American Behavioral Scientist*, 57(7):871–898, 2013. doi: 10.1177/0002764213479374. URL http://dx.doi.org/10.1177/0002764213479374.

Rob Schroeder, Sean Everton, y Russell Shepherd. Mining twitter data from the arab spring. 2017.

Mariano Beguerisse-Díaz, Guillermo Garduno-Hernández, Borislav Vangelov, Sophia N Yaliraki, y Mauricio Barahona. Interest communities and flow roles in directed networks: the twitter network of the uk riots. *Journal of the Royal Society Interface*, 11(101):20140940, 2014.

Evelien Otte y Ronald Rousseau. Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6):441–453, 2002. doi: 10.1177/016555150202800601. URL http://dx.doi.org/10.1177/016555150202800601.

Zachary C Steinert-Threlkeld. Spontaneous collective action: Peripheral mobilization during the arab spring. *American Political Science Review*, 111(2):379–403, 2017.

Anaris Atalis Santa Cruz. El método histórico lógico en la investigación educacional de posgrado, 2015. URL http://biblioteca.uniss.edu.cu/sites/default/files/CD/2015%20Universidad%202016/personal-injury/c2/c1.pdf.

Bruce Cronin et al. Social network analysis. 2016.

Olivier Serrat. Social network analysis. In *Knowledge solutions*, pages 39–43. Springer, 2017.

Stanley Wasserman y Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

Christopher McCarty y José Luis Molina. Social network analysis. *Handbook of methods in cultural anthropology. Second edition. Rowman and Littlefield, Lanham, Maryland, USA*, pages 631–657, 2015.

Itai Himelboim, Marc A Smith, Lee Rainie, Ben Shneiderman, y Camila Espina. Classifying twitter topic-networks using social network analysis. *Social Media+ Society*, 3(1):2056305117691545, 2017.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, P Krishna Gummadi, et al. Measuring user influence in twitter: The million follower fallacy. *Icwsm*, 10(10-17):30, 2010.

Twitter. Tweet object, 2017a. URL https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object.

Twitter. About replies and mentions, 2018c. URL https://help.twitter.com/en/using-twitter/mentions-and-replies.

John Scott y Peter J Carrington. *The SAGE handbook of social network analysis*. SAGE publications, 2011.

Theodore M Newcomb. Role behaviors in the study of individual personality and of groups. *Journal of personality*, 18(3):273–289, 1950.

Lamya Benamar, Christine Balagué, y Mohamad Ghassany. The identification and influence of social roles in a social media product community. *Journal of Computer-Mediated Communication*, 22(6): 337–362, 2017. URL http://dx.doi.org/10.1111/jcc4.12195.

George Herbert Mead. Mind, self, and society: From the standpoint of a social behaviorist (works of george herbert mead, vol. 1), 1967.

Johann Füller, Katja Hutter, Julia Hautz, y Kurt Matzler. User roles and contributions in innovation-contest communities. *Journal of Management Information Systems*, 31(1):273–308, 2014.

Shaobin Huang, Tianyang Lv, Xizhe Zhang, Yange Yang, Weimin Zheng, y Chao Wen. Identifying node role in social network based on multiple indicators. *PLOS ONE*, 9(8):1–16, 08 2014. doi: 10.1371/journal.pone.0103733. URL http://dx.doi.org/10.1371%2Fjournal.pone.0103733.

Fabián Riquelme y Pablo] González-Cantergiani. Measuring user influence on twitter: A survey. *Information Processing & Management*, 52(5):949–975, 2016.

David Knoke y Ronald S Burt. Prominence. applied network analysis: a methodological introduction, 1983.

Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

Janet C Long, Frances C Cunningham, y Jeffrey Braithwaite. Bridges, brokers and boundary spanners in collaborative networks: a systematic review. *BMC health services research*, 13(1):158, 2013.

Adrien Guille, Hakim Hacid, Cecile Favre, y Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(2):17–28, 2013.

Elizabeth Dubois y Devin Gaffney. The multiple facets of influence: Identifying political influentials and opinion leaders on twitter. *American Behavioral Scientist*, 58(10):1260–1277, 2014.

Svetlana S Bodrunova, Anna A Litvinenko, y Ivan S Blekanov. Comparing influencers: Activity vs. connectivity measures in defining key actors in twitter ad hoc discussions on migrants in germany and russia. In *International Conference on Social Informatics*, pages 360–376. Springer, 2017.

Claude Berge. *Graphs*, volume 6. North-Holland, 1985.

Stefan Wuchty y Peter F Stadler. Centers of complex networks. *Journal of Theoretical Biology*, 223(1): 45–53, 2003.

Boris V Cherkassky, Andrew V Goldberg, y Tomasz Radzik. Shortest paths algorithms: Theory and experimental evaluation. *Mathematical programming*, 73(2):129–174, 1996.

E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959. ISSN 0945-3245. doi: 10.1007/BF01386390. URL http://dx.doi.org/10.1007/BF01386390.

Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2): 163–177, 2001.

Tore Opsahl, Filip Agneessens, y John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, 32(3):245– –251, 2010.

Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46 (5):604–632, 1999.

Wolf E Hautz, Gert Krummrey, Aristomenis Exadaktylos, y Stefanie C Hautz. Six degrees of separation: the small world of medical education. *Medical education*, 50(12):1274–1279, 2016.

David L Passmore. Social network analysis: Theory and applications. *Institute for Research in Training & Development–IRTD*, 2011.

MATJAŽ KRNC. Centrality measures of large networks. 2015.

Center for BioMedical Computing. Eccentricity, 2017. URL http://www.cbmc.it/fastcent/doc/Eccentricity.htm.

Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008.

Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

Stephen P Borgatti y Martin G Everett. A graph-theoretic perspective on centrality. *Social networks*, 28 (4):466–484, 2006.

Douglas R White y Stephen P Borgatti. Betweenness centrality measures for directed graphs. *Social Networks*, 16(4):335–346, 1994.

Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

Yannick Rochat. Closeness centrality extended to unconnected graphs: The harmonic centrality index. In *ASNA*, number EPFL-CONF-200525, 2009.

Paolo Boldi y Sebastiano Vigna. Axioms for centrality. *Internet Mathematics*, 10(3-4):222–262, 2014.

Massimo Marchiori y Vito Latora. Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications*, 285(3):539–546, 2000.

Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972.

Phillip Bonacich. Some unique properties of eigenvector centrality. *Social networks*, 29(4):555–564, 2007.

Sergey Brin y Lawrence Page. The anatomy of a large-scale hypertextual web search engine in: Seventh international world-wide web conference (www 1998), april 14-18, 1998, brisbane, australia. *Brisbane, Australia*, 1998.

UCINET. Ucinet software, 2017. URL https://sites.google.com/site/ucinetsoftware/home.

CYRAM. Netminer - social network analysis software, 2018. URL http://www.netminer.com/.

Arcade Analytics LTD. Arcade - the best graph visualization tool for rdbms and graphdb, 2018. URL https://arcadeanalytics.com.

Social Media Research Foundation. Nodexl graph gallery: About nodexl, 2018. URL http://nodexlgraphgallery.org/Pages/AboutNodeXL.aspx.

Free Software Foundation. Gnu general public license, 2007. URL https://www.gnu.org/licenses/gpl-3.0.html.

Dimitris Kalamaras. Social network analysis and visualization software, 2018. URL http://socnetv.org/.

Alejandro Ruiz y Nina Jung. Visualización con 'pajek'. *Recuperado el*, 15, 2013. URL http://mrvar.fdv.uni-lj.si/pajek/spanish/Spanish.pdf.

Gephi. Features, 2017. URL https://gephi.org/features.

Jesse Fagan. Introduction to gephi. 2017.

Gephi. Developers introduction, 2018a. URL https://gephi.org/developers/intro/.

Gephi. Gephi toolkit, 2018b. URL https://gephi.org/toolkit/.

Pivotal. Amqp 0-9-1 model explained, 2017a. URL https://www.rabbitmq.com/tutorials/amqp-concepts.html.

Pivotal. Amqp 0-9-1 model explained, 2017b. URL https://www.rabbitmq.com/mpl.html.

OASIS. Amqp is the internet protocol for business messaging, 2017. URL https://www.amqp.org/.

Doctrine-Project. Object relational mapper - doctrine - php database tools, 2018. URL https://www.doctrine-project.org/projects/orm.html.

SensioLabs. The bundle system, 2017. URL http://symfony.com/doc/current/bundles.html.

Twitter4j. Twitter4j - a java library for the twitter api, 2017. URL http://twitter4j.org/en/index.html.

Twitter. Publish and manage tweets, and analyze tweet data, 2017b. URL https://developer.twitter.com/en/products/tweets.

Twitter. Filter realtime tweets, 2017c. URL https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html.

MongoDB. Nosql databases explained, 2017a. URL https://www.mongodb.com/nosql-explained.

Neo4J. What is a graph database?, 2017. URL https://neo4j.com/developer/graph-database/.

MongoDB. Reinventando la gestión de datos, 2017b. URL https://www.mongodb.com/es.

Donald T Campbell y HW Riecken. Quasi-experimental design. *International encyclopedia of the social sciences*, 5:259–263, 1968.

Carlos Manterola y Tamara Otzen. Estudios experimentales 2 parte: Estudios cuasi-experimentales. *International Journal of Morphology*, 33(1):382–387, 2015.

NV Kuzmina. Metódicas investigativas de la actividad pedagógica. *Editorial Leningrado*, 1970.

SJ Taylor y R Bogdan. La entrevista en profundidad. *MÉTODOS CUANTITATIVOS APLICADOS 2*, page 194, 2008.

Roberto Hernández Sampieri, Carlos Fernández Collado, y Pilar Baptista Lucio. Metodología de la investigación . méxico, df, 2010.

Alicia Hamui-Sutton y Margarita Varela-Ruiz. La técnica de grupos focales. *Investigación en educación médica*, 2(5):55–60, 2013.

Twitter. Tweet object, 2017d. URL https://developer.twitter.com/en/docs/tweets/search/overview.

María Mercedes Arias Valencia. La triangulación metodológica: sus principios, alcances y limitaciones. *Investigación y educación en enfermería*, 18(1):13–26, 2000.

Janice M Morse. Approaches to qualitative-quantitative methodological triangulation. *Nursing research*, 40(2):120–123, 1991.

Ashatu Hussein. The use of triangulation in social sciences research: Can qualitative and quantitative methods be combined? *Journal of comparative social work*, 4(1), 2009.

Sonia Aguilar Gavira y Julio Barroso Osuna. La triangulación de datos como estrategia en investigación educativa. *Pixel-bit. Revista de medios y educación*, (47), 2015.