Detección y Refuerzo de Comunidades de Celíacos en Twitter Argentina

Andrés Giordano, Santiago Banchero, Natacha Cerny, Mauricio De Marzi, and Gabriel Tolosa

Departamento de Ciencias Básicas, Universidad Nacional de Luján, Argentina
{agiordano, sbanchero, ncerny, mdemarzi, tolosoft}@unlu.edu.ar

Abstract. Las redes sociales han mostrado un gran crecimiento en cuanto a la cantidad de usuarios y contenido generado. Por ejemplo, Twitter es utilizado como medio para juntar apoyos, expresar ideas y opiniones sobre diversos temas o relacionarse con usuarios similares.
En este último caso, aparece la idea de la formación de comunidades, es decir, grupos de usuarios que se encuentran más densamente vinculados entre sí respecto al resto de los nodos.
En este trabajo se propone la detección de la comunidad de usuarios de Argentina interesados en la enfermedad celíaca. Se aplican una serie de técnicas con el objeto de su detección y caracterización. Además, se propone el uso de una metodología para la detección de nodos más influyentes y activos, mostrando cómo se puede reforzar la comunidad a partir de la sugerencia de enlaces.

1 Introducción

Las redes sociales han mostrado un gran crecimiento en cuanto a la cantidad de usuarios y contenido generado, principalmente en los últimos años. Un ejemplo claro es Twitter, en la cual no solamente los usuarios publican sus actividades sino que, en algunos casos, se utiliza como medio para juntar apoyos, expresar ideas y opiniones sobre diversos temas o relacionarse con usuarios similares.

A partir de esta dinámica, las formas de comunicación se han ampliado, generando patrones de unión y comportamiento entre usuarios que poseen propiedades emergentes que resultan de interés conocer para comprender su alcance y efectividad. Estas relaciones, que ocurren tanto en la naturaleza como en fenómenos sociales, pueden ser representadas y analizadas en términos de una red, o formalmente, un grafo. En general, a una escala macroscópica, estas redes ofrecen cierto grado de organización [26].

Uno de estos fenómenos es la formación de comunidades en redes sociales. Las personas tienden a agruparse instintivamente en el mundo digital como en el real, con otros con quienes comparten ideas, gustos, hobbies, etc., lo que facilita la comunicación. Si bien no existe una definición global y única sobre qué es una comunidad, se la puede definir como un conjunto de personas que interactúan en el tiempo con un objetivo, interés o necesidad [31]. En cuanto al análisis de la red subyacente, se trata de grupos de nodos que se encuentran más densamente vinculados entre ellos que respecto el resto de los nodos.

Existen comunidades implícitas y explícitas [28]. Las primeras se forman por las interacciones diarias de un grupo de usuarios, las cuales no son siempre vistas por todos (por ejemplo, las publicaciones de usuarios sobre un tema en Twitter, con su grupo de seguidores). Por otro lado, las comunidades explícitas son aquellas en las cuales los usuarios toman una decisión consciente de participar de un grupo, pueden conocer el conjunto de miembros del mismo y el alcance de sus publicaciones (por ejemplo, un grupo cerrado de Facebook sobre algún tema particular).

En este último caso, la comunidad está claramente delimitada y resulta relativamente sencillo el análisis de las interacciones. Sin embargo, la identificación de comunidades implícitas en redes sociales es una tarea un poco más compleja, cuyo resultado no es exacto y que puede brindar información útil acerca de la dinámica y comportamiento de grupos de usuarios con intereses comunes con diversos objetivos, por ejemplo, proveer servicios relacionados con su interés.

Los patrones de comunicación tienden a ser más intensos entre miembros de un mismo grupo, respecto de los demás. Estos siguen el principio sociológico conocido como *homofilia* que propone que las personas tienen a relacionarse en mayor medida con pares similares por alguna característica (edad, educación, religión, entre otras). Otra relación importante que aparece es la *influencia*, en la cual algunos miembros de un grupo desarrollan ideas o visiones similares sobre algún concepto siguiendo la opinión de uno o varios de sus miembros [13].

La detección de comunidades es un problema relevante en el mundo del análisis de redes sociales o, más ampliamente, en Ciencia de las Redes[1]. Por un lado, permite la identificación de relaciones no triviales entre integrantes de la red y su auto-organización y, por el otro, ayuda a comprender los procesos que tienen lugar para su formación y dinámica [28, 31].

Además, no todos los usuarios que se conectan entre si comparten todos los mismos intereses [18], por lo que considerar sólo la estructura de enlaces en la red puede ser un criterio incompleto o que no aplica para todos los casos. Por ello, la similitud entre usuarios a través del contenido de sus publicaciones y otros datos como localización, sexo o edad pueden también ayudar a determinar la pertenencia a un mismo grupo o comunidad.

La habilidad para detectar comunidades en una red social tiene implicancias prácticas en múltiples dominios. En este trabajo se propone la detección de una comunidad particular relacionada con usuarios de Argentina interesados con la enfermedad celíaca, como complemento de estudios epidemiológicos[2] [7].

La enfermedad celíaca es un desorden autoinmune complejo generado por la ingesta de gluten, una proteína que se encuentra en ciertos cereales. Esta enfermedad interfiere con la absorción de nutrientes al dañar parte del intestino delgado y se la encuentra vinculada con otras patologías como tiroidismo, osteoporosis, infertilidad, diabetes, entre otras. Actualmente, se estima que su incidencia en Argentina es del 1.0%[3]. Esta patología tiene impacto en la vida de las personas, incluso en su vida social, principalmente en la formación de capital social informal, esto es, el contacto con amigos, familiares, colegas [36]. Las redes sociales colaboran con el mantenimiento de cierto capital social digital. Por un lado, facilitando y ampliando la comunicación con otras personas y, por otro, permitiendo el intercambio de información. En este caso, se vuelven herramientas poderosas para obtener, generar y propagar información sensible relacionada con la enfermedad celíaca que puede ser de ayuda a otros, desde indicaciones para obtener alimentos libres de gluten, recetas hasta discusiones acerca de signos, síntomas y diagnósticos en cada caso. Muchas veces, compartir experiencias abre nuevas perspectivas a aquellos quienes padecen enfermedades.

[1] La Ciencia de las Redes (*Network Science*) es un campo de investigación relativamente nuevo que estudia sistemas complejos y su representación como redes tanto de fenómenos naturales como sociales, intentando obtener modelos predictivos del comportamiento de sus actores.

[2] Este trabajo se relaciona con un proyecto interdisciplinario cuyo objetivo principal es caracterizar la incidencia de la enfermedad celíaca y su relación con patologías relacionadas.

[3] https://www.argentina.gob.ar/salud/glosario/enfermedadceliaca

1.1 Motivación y Objetivos

La relación entre redes sociales y comportamientos patológicos en grupos de personas es un tema de interés [37]. Sin embargo, no se han encontrados trabajos relacionados con la enfermedad celíaca y su repercusión en una red social. Teniendo en cuenta el crecimiento de las redes sociales, la intensidad de participación de sus usuarios y las posibilidades que brinda el poder estudiar masivamente grupos de usuarios casi en tiempo real, resulta de especial interés generar metodologías y estudios específicos que apoyen a otras disciplinas en vías de caracterizar desde otra perspectiva un fenómeno humano, como una patología particular.

El objetivo principal de este trabajo es detectar potenciales usuarios argentinos con interés en la enfermedad celíaca (paciente/familiar/amigo) y sugerir vínculos que permitan reforzar comunidades de acuerdo a este interés para facilitar el intercambio de información valiosa en el contexto. En particular, los aportes de este trabajo son:

- La detección de comunidades de usuarios interesados en la enfermedad celíaca combinando tanto las relaciones (*links*) como el contenido de sus publicaciones. Se combinan ambos enfoques mostrando una mejora en la precisión final.
- El uso de una técnica de clustering combinada con la búsqueda de *ciertos* usuarios de interés para determinar el cluster que representa a la comunidad. Se muestra cómo varía la precisión en la comunidad identificada.
- La identificación de los usuarios más influyentes y activos en la comunidad y se utiliza esta métrica para recomendar enlaces a subgrupos de usuarios. Se muestra cómo la comunidad se vuelve más densa conforme se aceptan enlaces, lo que refuerza la propagación de información interna.

El resto del trabajo se encuentra organizado de la siguiente manera: en la Sección 3 se introducen los conceptos básicos necesarios del contexto del trabajo. Luego, se propone una metodología en la Sección 4 que determina los experimentos (y sus resultados) de la Sección 5. Finalmente, se presentan las conclusiones y los trabajos futuros (Sección 6).

2 Trabajos Relacionados

Existen varios trabajos que abordan la problemática de búsquedas de comunidades en redes sociales [3, 11, 29] haciendo uso de las conexiones existentes en la red.

Por el contrario, la búsqueda vertical de comunidades es una tarea compleja dado que no basta con explotar las conexiones existentes sino que es necesario explorar el contenido de las publicaciones para determinar los temas que son de interés para los usuarios. Algunos abordajes para detección de comunidades orientadas a temas combinan técnicas de agrupamiento de objetos sociales con el análisis de enlaces [40,41]. La técnica consiste en agrupar, a partir de características denominadas objetos sociales, en grupos temáticos utilizando el algoritmo *Entropy Weighting K-Means* [16]. Dentro de cada uno de los grupos temáticos se realiza un análisis de enlaces utilizando la modularidad para buscar las potenciales comunidades presentes por cada tema.

La búsqueda semántica de comunidades es otro abordaje al problema, existen técnicas que utilizan el contenido de la red como Latent Dirichlet Allocation (LDA) [4] para realizar agrupamientos por temas. El modelo *Link-Block-Topic*, utiliza LDA y realiza detección de comunidades temáticas sin necesidad de indicar la cantidad de comunidades a buscar ni el tamaño de las comunidades [39].

Otros trabajos, basados en aproximaciones locales vinculan temáticas a partir de identificar usuarios con gran cantidad de seguidores, considerando que los usuarios seleccionados son representativos de una categoría de interés que es en la que realizan mayor cantidad de publicaciones [21]. Estas técnicas utilizan luego un cálculo de solapamiento, entre los seguidores de referentes y las comunidades de la red a través de *Clique Percolation Method* (CPM) [9].

Complementariamente, Yang y otros introducen CESNA [38] (*Communities from Edge Structure and Node Attributes*), un método que utiliza un modelo de probabilidad basado en distribuciones de Bernoulli donde se combinan la pertenencia a una comunidad, la estructura de la red y los atributos de los nodos a partir del modelo. Esta solución se apoya en el supuesto de que los vértices son más propensos a ser vecinos cuanto más comunidades los comparten. Si bien el algoritmo CESNA tiene un *runtime* lineal con el tamaño de la red la interpretación de los resultados no es buena [6].

3 Preliminares

El modelo subyacente de una red social corresponde a un grafo $G = \langle V, E \rangle$, donde V es el conjunto de nodos o vértices que representan a los usuarios de la red social y E es el conjunto de aristas que representan las relaciones entre los usuarios. Por ejemplo, en Facebook, si el usuario A es amigo de B entonces existe una arista $(A, B) \in E$, mientras que en el caso de Twitter (A, B) representa la relación "A sigue a B" pero no a la inversa (*followers* vs *followings*). Por lo tanto, si B sigue a A entonces existe $(B, A) \in E$ lo cual denota un grafo dirigido. A su vez, la intensidad de la relación corresponde al peso o ponderación de la arista en base a alguna métrica calculada para ambos usuarios (por ejemplo, la cantidad de *retweets* que realiza un usuario).

3.1 Comunidades

Como se mencionó, no existe una definición única para el concepto de *comunidad*, pero existe una característica en común a todas y es que se componen de usuarios que tienen un tema o tópico de interés común. Existen comunidades que se componen de usuarios que periódicamente publican noticias, anécdotas y generan charlas o discusiones sobre algún tema en específico de su interés, es decir, los usuarios de la comunidad tienen un alto grado de interacción entre sí. Por otro lado, existen comunidades que se distinguen claramente por sus "relaciones físicas" dentro de la red, es decir, "followers" y "followings" en el caso de Twitter o amistades en Facebook. Aquí puede o no existir alta interacción entre los usuarios pero si debe existir una alta densidad en los vínculos intra-comunidad.

Incluso, existen comunidades cuyo único vínculo es el tópico de interés común, sin la existencia de relación física o de interacción dentro de la red social. Para detectarlas, se recurre al análisis del contenido de las publicaciones en busca de características que lo acerquen a un tópico específico.

3.2 Detección de Comunidades

Existen diferentes métodos para la detección de comunidades, los cuales resultan más o menos adecuados de acuerdo al tipo de comunidad buscada, su patrón de interacción o a la porción de la red social explorada. Por ejemplo, en el caso de las comunidades interactivas como Twitter es necesario recolectar todo tipo de interacción entre los usuarios (*retweets*, menciones y comentarios). Lim [22] propone generar un grafo donde las relaciones son las menciones entre los usuarios y aplicar detección de comunidades sobre la estructura generada. En general, los algoritmos de detección de comunidades se clasifican en:

- **Basados en la Topología:** estos métodos se basan sólo en el grafo subyacente a la red, es decir, los usuarios y sus relaciones [5, 21, 32]. Si bien los algoritmos que aplican este enfoque son eficaces suelen agrupar usuarios que tratan de tópicos diferentes aunque densamente conectados (carecen de alta precisión).
- **Basados en el Contenido:** este enfoque explora el contenido de las publicaciones de los usuarios y no considera la información estructural de la red como lo es la densidad de las conexiones que puede existir en un conjunto de usuarios. En Twitter, por ejemplo, esto se refiere al contenido de los tweets separando texto libre de *hashtags*, urls y menciones [22].
- **Híbridos:** en estos métodos se utilizan los dos enfoques anteriores en conjunto [17, 33, 35, 40]. Básicamente, se construye el grafo de relaciones estructurales y se agregan características como la similitud de contenido, aplicada como peso o importancia de la relación entre un par de usuarios. Una vez generada esta estructura se aplica algún algoritmo de detección de comunidades conocido que use el peso de las aristas.

3.3 Algoritmos

A continuación, se describen dos algoritmos específicos para detección de comunidades utilizados en este trabajo. El primero, conocido como método de Louvain[4] [5], se basa en la optimización de la modularidad de las particiones a medida que el algoritmo progresa en su ejecución de forma *greedy*. Por otro lado, el método Infomap [32], que se basa en la teoría de la información para representar las comunidades.

Método Louvain: este método busca maximizar la modularidad del grafo a medida que se agrupan los nodos en comunidades. Es robusto y eficiente, ya que ha sido usado y revisado en varios trabajos [21, 22] y nuevos algoritmos de detección de comunidades se basan en éste [8, 12, 25, 30]. En términos de tiempo de cómputo, corre en $O(n \cdot \log n)$.

La modularidad se establece con el fin de juzgar la calidad de las particiones de las comunidades formadas [26] y ha sido ampliamente utilizada con este fin [3, 34] como calibre de la calidad de las comunidades. Esta métrica se define como:

$$Q(G) = \frac{1}{2m} \sum_{l=1}^{K} \left(\sum_{i \in C_l, j \in C_l} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \right) \tag{1}$$

donde:

K es la cantidad de comunidades,
A_{ij}, el peso de la arista entre i y j,
k_i corresponde a la suma de los pesos de las aristas adjuntas a i,
C_l es la comunidad a la cual i y j están asignados,
$m = \frac{1}{2} \sum_{i,j \in V} A_{ij}$
$\frac{1}{2m}$ normaliza el resultado entre -1 y 1.

[4] La afiliación de los autores (Universidad Católica de Louvain, Bégica) da nombre al método.

Luego, el algoritmo agrupa los nodos de G en dos pasos que son repetidos en cada iteración:

1. **Optimización de modularidad:**
   (a) Asignar cada nodo a una comunidad diferente.
   (b) Por cada nodo i, procesar todos sus vecinos j calculando la ganancia de modularidad en mover el nodo i a la comunidad de j. Luego, i es movido a la comunidad cuya ganancia es la máxima si y sólo si la ganancia es positiva.
   (c) Repetir paso (b) hasta llegar a un máximo de modularidad local, es decir, cuando ya no hay movimientos entre comunidades.
2. **Agregación de comunidades:** se forma una nueva red donde cada nodo es una de las comunidades anteriormente formadas. Se agrega un link entre cada par de nodos nuevos (links entre comunidades) si existe al menos un par de nodos que las una y es ponderado con la suma de los pesos de los links existentes entre las dos comunidades. Los nodos intra-comunidad se representan como loops. Este paso permite establecer un parámetro de corte para encontrar comunidades más grandes o más chicas.

Método Infomap: se basa en la teoría de la información para representar las comunidades con el código de menor longitud posible. Básicamente, propone representar una caminata aleatoria sobre un grafo de forma efectiva y compacta. Se utilizan dos niveles de descripción basados en códigos de Huffman [15]:

1. el primer nivel establece un código unívoco para cada nodo intra-comunidad cuya longitud es inversamente proporcional a la cantidad de veces que ese nodo es visitado en la caminata.
2. el segundo nivel define códigos de la misma manera pero para identificar las diferentes comunidades.

Entonces, el problema de encontrar la mejor partición del grafo en grupos o comunidades de usuarios se expresa como encontrar la mínima cantidad de información necesaria para representar la caminata aleatoria usando los niveles de descripción anteriormente planteados.

El código de Huffman es diseñado para asignar códigos cortos a los símbolos con mas frecuencia en un lenguaje determinado y viceversa. Se espera que el caminante se mantenga por un tiempo prolongado dentro de cada comunidad visitando varias veces los mismos nodos ya que la cantidad de links intra-comunidad es mayor que los que unen nodos en comunidades diferentes. De este modo, dentro de cada comunidad se puede generar un código óptimo para representar cada uno de sus nodos y solo es necesario un par de códigos extra para indicar que el caminante entró o salio de una comunidad determinada logrando así expresar todo el trayecto recorrido en la mínima cantidad de código posible. Este método ha sido ampliamente utilizado [21,22], incluso aportando en otras áreas como la biología [10].

Clustering clásico: otra técnica para la detección de comunidades es aplicar algoritmos clásicos de clustering en base a características de los usuarios, como por ejemplo, *Kmeans*. Éste método es utilizado ampliamente por la comunidad científica en diversas áreas de la computación como procesamiento de imágenes satelitales, data mining, entre otras [14, 24]. El enfoque de este algoritmo es identificar K clusters asignando cada ejemplar (usuario) al cluster cuyo centroide (centro de masa) se encuentre mas cerca. Para ello, se requiere representar a cada individuo mediante un vector de características. Por ejemplo, mediante el procesamiento de sus publicaciones y construyendo la distribución de frecuencias de los términos que utiliza. Opcionalmente, se puede aplicar un algoritmo de detección de tópicos como LDA (Latent Dirichlet Allocation) con lo cual se pasa a acumular la frecuencia con la que se publica sobre cierto tópico.

3.4 Fortalecimiento de la Comunidad

Las estrategias para el fortalecimiento de una comunidad tienen por objetivo establecer mayor cantidad de vínculos intra-comunidad, resultando éstas en una estructura más densa. La idea principal se basa en la sugerencia (o recomendación) de enlaces[5].

En general, estos métodos intentan estimar la probabilidad que cierto vínculo se establezca en un futuro y se recomiendan aquellos que maximizan tal métrica. Algunos se basan en modelos de aprendizaje [1] y otros de proximidad [19].

En este trabajo se propone combinar dos características de los nodos: su influencia y su actividad (Sección 4.5). Es decir, dado el objetivo de la comunidad, se prefiere usuarios que puedan diseminar información rápidamente (influyentes) pero lo hagan periódicamente (activos).

3.5 Métricas

En esta sección se describren algunas métricas utilizadas para el análisis que, básicamente, son medidas sobre el grafo $G = \langle V, E \rangle$ o sus nodos.

Diámetro $(D(G))$**:** Se define la distancia entre dos vértices $(u, v \in G)$ como la longitud del camino mas corto entre ellos. Luego, el diámetro de G es la distancia máxima entre todos los pares de nodos.

Closeness $(C(u))$**:** La métrica *Closeness* de un nodo cualquiera $u \in G$, intenta cuantificar qué tan cerca se encuentra u de los demás nodos de G. Se define como la inversa de la suma de las distancias de u a todos los demás vértices v, $C(u) = \frac{1}{\sum_{v \in V} d(v,u)}$.

Coeficiente de Clustering (CC)**:** El CC de un vértice $u \in G$ cuantifica qué tanto está de agrupado o interconectado con sus vecinos. Corresponde a la proporción entre los enlaces conectados con sus vecinos (e_{ij}) y el número de enlaces existentes en un clique (conectividad máxima). Se define como $C_i = \frac{|e_{ij}|}{k_i(k_i-1)}$. Luego el CC Promedio (CCP) de G resulta $\frac{1}{n}\sum_i^n C_i$.

4 Metodología

Con el objetivo de identificar la comunidad objetivo se parte de un enfoque basado en la topología de la red, a partir del muestreo de las publicaciones (*tweets*) de usuarios utilizando la API pública de Twitter[6]. Luego, se construye un grafo en base a los vínculos entre los usuarios y se analiza su pertenencia (o no) a la comunidad buscada.

4.1 Recolección de Datos

Los tweets fueron recolectados entre el 20 de Abril y el 02 de Julio de 2017 (74 días). Para la identificación positiva de las publicaciones se utilizan palabras clave relacionadas con el tema: *celiaco, celíaco, celiac, coeliac, celiaquia, celiaquía, sintacc, tacc, gluten, "libre de gluten", glutenfree*[7]. La cantidad de publicaciones recolectadas es de 131.550 con un total de 76.233 usuarios únicos.

Filtrado por ubicación del usuario: Dado que el objetivo es la detección de una comunidad de celíacos en Twitter en Argentina, se filtraron los tweets para obtener sólo aquellos publicados por usuarios argentinos. Esta tarea puede hacerse de dos maneras diferentes: (1) si el tweet está

[5] Por ejemplo, en Facebook se ofrece una lista de "Gente que Tal Vez Conozca"

[6] Se utilizó la API de Streaming usando el parámetro "track".

[7] El uso de algunas palabras en inglés responde a que se detectó que eran usadas en algunos *hashtags*.

geolocalizado se toman las coordenadas del campo `'coordinates'` del tweet y se realiza una resolución reversa en un servicio de mapas, (2) en caso contrario se analiza el campo `'location'` del usuario y se lo compara con una lista de localidades y provincias de Argentina.

4.2 Generación del grafo

Para la generación del grafo se toman como nodos únicamente las publicaciones de usuarios argentinos (U_{arg}). Por cada usuario $u \in U_{arg}$ se obtiene el conjunto de usuarios a los que sigue (*friends*) y se agrega la arista correspondiente. La estructura resultante es un grafo $G = \langle V, E \rangle$ dirigido donde cada arista $(u_1, u_2) \in E$ representa la relación u_1 "sigue a" u_2.

Los nodos aislados son usuarios que no tienen conexión con el resto del grafo, es decir, no existe una arista que parta o llegue a éste. Por ello son eliminados, resultando en un grafo de un total de 2.068 nodos y 20.675 aristas. Este constituye el "grafo base", G_{base}.

4.3 Identificación de la Comunidad

Para la identificación inicial se utiliza el método de Louvain y luego se aplica una técnica similar a la utilizada por Lim [20] donde el interés de los usuarios en cierto tópico es detectado a través del concepto de "celebridades". Las celebridades son usuarios con mas de n seguidores (donde n siempre es un número elevado respecto del resto de los usuarios) y de los cuales se sabe a priori que tienen interés en el tópico en cuestión (aunque podría no ser el único interés de dicha celebridad). En dicho trabajo se obtiene el conjunto de usuarios que siguen a la totalidad de las celebridades sobre el que luego se aplica la detección de comunidades, verificando que los usuarios tienden a seguir a referentes en los tópicos que es de su interés.

Sin embargo, la exploración de celebridades en el mundo de la celiaquía en Argentina muestra números muy bajos. Dado que no existen usuarios argentinos interesados en este tópico con una alta cantidad de seguidores, se suprimió este requisito. Cinco de las seis cuentas de Twitter que fueron seleccionadas (Tabla 1) han sido creadas específicamente con el fin de compartir novedades o información acerca del tema. El usuario restante corresponde a una persona que se describe como celíaca en su perfil.

Us uario	Descripción
@asoc_celiaca_ar	*Pionera en América Latina. Brinda apoyo a quienes deben seguir una dieta libre de gluten*
@CeliacoCom	*Sos Celiaco? acá todo lo que necesitas. recetas, vídeos, info de interés, donde comer, comprar y mucho más !!!*
@cocinaceliaca	*Soy chef de alta cocina especialista en cocina apta celiacos.*
@SoyCeliacoNoET	*Recetas, experiencias, consejos e información sobre celiaquía y dieta libre de gluten. #SinTACC #SinGluten#GlutenFree #Food*
@rom_kari	*Hija de Celiaca, Celiaca, Mamá de Celiaco*
@rojasglutenfree	*Supermercado Exclusivo para Celíacos.*

Tabla 1. Celebridades Identificadas en el Tema "Celiaquía"

Validación de la Comunidad: Luego de la ejecución de cada método se realiza una validación de la comunidad encontrada. En cada caso, se solicita un grupo de usuarios voluntarios que, a partir de observar el perfil público de cada usuario, juzguen si corresponde a uno interesado en el tema celiaquía (o no).

4.4 Similitud entre Usuarios

Los métodos basados en contenido requieren de alguna técnica para relacionar a los usuarios (en vez de utilizar los enlaces). Una posibilidad comúnmente utilizada es calcular alguna medida de similitud entre los usuarios tomando sus publicaciones como representativas de sus intereses.

Para ello, se genera un documento por usuario formado por la concatenación de los últimos n tweets[8] [40]. Por cada tweet se eliminan stopwords, URLs, números, signos de puntuación, emoticones, flechas y todo aquel token que excede los 30 caracteres. Luego, el cálculo de la similitud entre usuarios se realiza en base al modelo vectorial usando la fórmula de similitud por coseno, clásica en el área de Recuperación de Información [23], definida como:

$$score(d_u, d_s) = \frac{\overrightarrow{V}(d_u) \cdot \overrightarrow{V}(d_s)}{\left|\overrightarrow{V}(d_u)\right| \left|\overrightarrow{V}(d_s)\right|} \tag{2}$$

donde $\overrightarrow{V}(d_n)$ es el vector de pesos correspondiente a cada documento (que corresponden a los usuarios u, s, respectivamente). El denominador corresponde a producto de la norma de ambos vectores y tiene el propósito de normalizar el largo de los documentos para la comparación. Para la ponderación de los términos en cada $\overrightarrow{V}(d_n)$ se utiliza TF/IDF [2] donde el valor de TF representa la frecuencia normalizada del término i en el documento j del usuario, $TF = \frac{frec(i,j)}{maxfrec(j)}$. El valor IDF corresponde a la inversa de la frecuencia en documentos del término en la colección, $IDF(t) = log(\frac{N}{df})$, donde N es la cantidad de documentos (en este caso, la cantidad de usuarios) y *df* es la sumatoria de las frecuencias del término en cada documento. Una vez calculada la similitud entre cada par de usuarios, ésta es utilizada como peso o importancia de la relación entre ambos.

4.5 Usuarios Influyentes y Activos

Para la búsqueda de usuarios influyentes y activos en el tema se genera un ranking de usuarios para cada una de estas características y luego se toman aquellos que se encuentren a un porcentaje p en ambas listas. El objetivo de esto es obtener un conjunto de usuarios a recomendar al resto de los interesados en la enfermedad.

Para clasificar los usuarios según su influencia se utiliza el método sugerido por [13] donde se genera un grafo dirigido (G_{infl}) cuyos nodos son usuarios, las aristas representan las relaciones A "retwitea a" B y/o A "menciona a" B y el peso de la relación se encuentra dado por la cantidad de veces que ocurre cada una de las relaciones. Luego, se calcula sobre G_{infl} alguna métrica clásica de importancia de los nodos (por ejemplo, Hubs o PageRank). En este caso se usa PageRank [27].

Para generar el ranking de activos se parte de los últimos n tweets del usuario y se calcula el porcentaje de términos publicados referidos a la enfermedad. Para esto se utilizan las palabras clave *celiaco, celiaquia, sintacc, tacc, gluten, glutenfree, #celiaco, #celiaquia, #sintacc, #tacc, #gluten, #glutenfree* sobre las cuales se cuenta la frecuencia de publicación. Luego se calcula el porcentaje respecto de todos los términos publicados por el usuario en los n tweets.

[8] En este caso se utilizaron los últimos 3200 tweets ya que existe un limite de recuperación en la API.

5 Experimentos y Resultados

Para los experimentos de detección de comunidades de celíacos se parte del grafo base G_{base} y se generan dos nuevas versiones de éste ponderando las aristas de acuerdo al criterio de similitud entre usuarios (Sección 4.4): G_{base_w} y $G_{base\_nd\_w}$. Este último, además, asume las aristas como no dirigidas, reflejando con más peso la simetría en el cálculo de la similitud entre usuarios.

5.1 Detección de Comunidades

Utilizando el método Louvain sobre los tres grafos, se varía el parámetro de corte en el rango $[0,1;1]$ aumentando de a $0,1$ en cada paso. Esto se puede pensar como la "altura" a la que se corta un dendrograma. Si el valor de corte se acerca a 1 se obtienen comunidades más grandes (menor resolución) y si se acerca a 0 las comunidades formadas serán más pequeñas (mayor resolución). Esto se relaciona con el Coeficiente de Clustering (CC) obtenido, siguiendo la idea [34] que las redes con comunidades subyacentes tienden a tener un valor de CC promedio (CCP) mucho más alto que redes aleatorias con la misma cantidad de aristas y nodos.

Finalmente, el valor de corte seleccionado se determina por el mayor CC alcanzado en la comunidad celíaca encontrada. Luego, dicha comunidad se valida según se especifica en la sección 4.3. La tabla 2 muestra el resultado para cada valor de corte en cada uno de los grafos. Los valores de corte finales usados fueron $0,3$, $0,2$ y $0,1$ para los grafos G_{base}, G_{base_w} y $G_{base\_nd\_w}$, respectivamente. En los tres casos las comunidades detectadas (U_{com}) cuentan con un elevado porcentaje de usuarios interesados superando el 65% sobre la cantidad total de individuos que componen el grupo. Particularmente, sobre el grafo $G_{base\_nd\_w}$ se logra una mayor precisión (74,6%) al costo de una baja en la cantidad de recuperados del 7,79% (Tabla 3).

De forma complementaria, se ejecuta el método Infomap. Para evaluar la variabilidad de la comunidad de celíacos formada utilizando este método se calcula la intersección sobre la unión del conjunto de usuarios en 10 diferentes ejecuciones. Los resultados muestran que la comunidad varía en solo el 1% verificando la consistencia de este algoritmo a pesar de aplicar una técnica aleatoria relativa al grado de los nodos.

La tabla 4 muestra la cantidad y el porcentaje de usuarios interesados en el tema dentro de la comunidad celíaca encontrada (U_{com}). La precisión alcanzada sobre G_{base_w} y $G_{base\_nd\_w}$ alcanza el 77% en ambos casos a una diferencia de 8,73% de usuarios menos recuperados en el peor de los casos respecto de G_{base}.

Corte	CC Promedio		
	G_{base}	G_{base_w}	$G_{base\_nd\_w}$
0,1	0,276	0,298	**0,523**
0,2	0,334	**0,299**	0,521
0,3	**0,358**	0,254	0,484
0,4	0,355	0,251	0,469
0,5	0,343	0,254	0,43
0,6	0,346	0,217	0,424
0,7	0,281	0,185	0,470
0,8	0,324	0,276	0,455
0,9	0,325	0,204	0,319
1,0	0,166	0,233	0,360

Tabla 2. Valor de corte vs CCP para cada grafo (en negrita el valor mas alto)

Grafo	$\|U_{com}\|$	Interesados	% Interesados
G_{base}	104	68	65,4%
G_{base_w}	82	57	69,5%
$G_{base\_nd\_w}$	71	53	74,6%

Tabla 3. Celíacos/Interesados en la comunidad detectada U_{com} (Louvain) para los tres grafos utilizados.

Grafo	$\|U_{com}\|$	Interesados	% Interesados
G_{base}	91	63	69,2%
G_{base_w}	71	55	77,4%
$G_{base\_nd\_w}$	74	57	77%

Tabla 4. Celíacos/Interesados en la comunidad detectada U_{com}(Infomap) para los tres grafos utilizados.

5.2 Comunidades Mediante Clustering

Como se mencionó anteriormente, una posibilidad es tratar la formación de comunidades como un problema de clustering de documentos. Aquí se prueba con el ampliamente utilizado método KMeans (con inicialización KMeans++), para el cual se requiere representar a cada individuo mediante un vector de características. Este vector se obtiene a partir del vector documento $\overrightarrow{V}(d_n)$ (Sección 4.4) donde cada término es una característica del usuario.

Para ello, se prueba el algoritmo variando el valor de K en el rango $[5-95]$ (con pasos de 5). Luego, para cada valor de K se identifica y valida la comunidad celíaca (Sección 4.3). A partir de los k cluster, se determina que aquel en el cual se encuentran los 6 usuarios anteriormente detectados como celebridades corresponde a la comunidad. La figura 1 muestra, por un lado, la relación entre el valor de K y el tamaño de la comunidad de usuarios interesados en la enfermedad celíaca. Por el otro, se muestra la precisión lograda en el cluster comunidad.

Cabe destacar que para ciertos valores de K no se identifica la comunidad de celíacos al no haber un cluster que contenga los 6 usuarios mencionados. Se puede ver que, conforme aumenta K, la cantidad total de usuarios que componen la comunidad es menor haciendo que el método se vuelva mas preciso. A partir de $k = 35$ la cantidad de usuarios totales y de interesados se mantiene estable.

La ventaja de usar este método para detección de comunidades es que permite calcular fácilmente la pertenencia o no de un nuevo usuario a la comunidad encontrada. Para ello, simplemente se calcula la distancia del usuario a todos los clusters formados y se lo asigna al que más cerca se encuentre (resemblando otra iteración de KMeans). Si dicho cluster corresponde a la comunidad celíaca entonces se lo considera interesado.

La selección de k es crucial para obtener buenos resultados en la clasificación. Un método válido consiste en seleccionar el k que haya formado la comunidad destino con la mayor precisión. Si se cuenta con una comunidad con un porcentaje alto de interesados entonces hay una alta probabilidad de clasificar correctamente el nuevo usuario como tal.

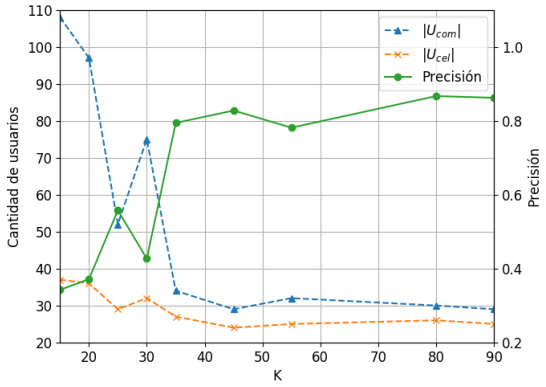

Fig. 1. Tamaño de la comunidad $|U_{com}|$, cantidad de usuarios interesados $|U_{cel}|$ y precisión obtenida por KMeans en la comunidad encontrada vs cantidad de clusters (K).

5.3 Recomendación de Usuarios

El objetivo de este experimento es evaluar el cambio estructural que sufre la comunidad ante un proceso de recomendación de usuarios referentes en la red social sobre celiaquía. En esta instancia solamente se simula el proceso sin intervención en la misma. Se parte del conjunto de usuarios de la comunidad celíaca encontrada por el método Louvain sobre G_{base} ($U_{com\_l\_base}$) y se sigue el método de selección de usuarios a recomendar (Sección 4.5). La idea subyacente es seleccionar recomendaciones basadas en dos atributos de los usuarios: su influencia y su actividad.

Selección de usuarios: Inicialmente, se obtienen los usuarios influyentes mediante PageRank (G_{inf}). La figura 2 muestra el grafo G_{inf} con el tamaño de los nodos proporcional a su puntaje. Aquí se puede apreciar un *hub* (*"asoc_celiaca_ar"*) como el más influyente.

Usuario	Score
SoyCeliacoNoET	12,350
AlimentoSinTacc	9,333
Paulidd	8,916
goutcafe1	8,472
GlutenFreeArg	7,859
TaccAway	7,744
sansglutenmdp	7,633
celigourmet	7,078
Cocelia1	6,023
rojasglutenfree	5,901

Tabla 5. Usuarios activos en la comunidad celíaca (Top-10).

A continuación, se obtiene la lista de usuarios activos en el tema (Top-10 en la Tabla 5) y, finalmente, se calcula la intersección entre ambos rankings. La Figura 3 muestra cómo evoluciona

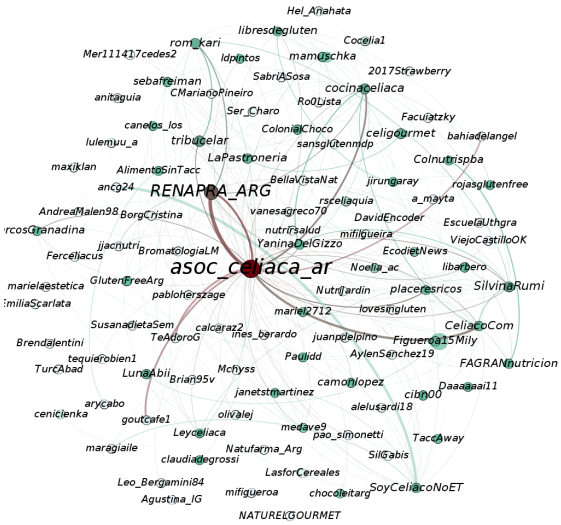

Fig. 2. Grafo de usuarios más influyentes (por PageRank).

la intersección de ambas listas de acuerdo a diferentes proporciones de corte (el mismo en ambas listas).

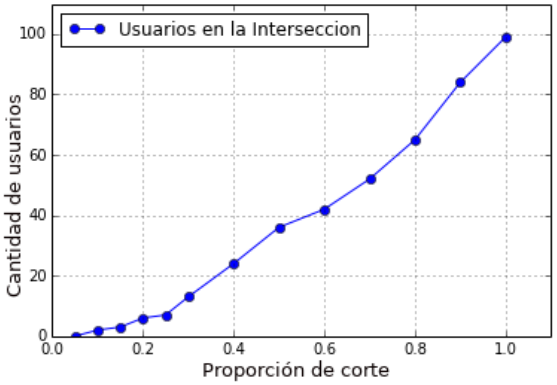

Fig. 3. Tamaño de la intersección entre el ranking de influyentes y de activos.

Proceso de recomendación: Para obtener el conjunto de usuarios a recomendar se define un porcentaje arbitrario p (en este caso $p = 0,2$) sobre las listas anteriores y se calcula la intersección obteniendo el conjunto de usuarios a recomendar U_{rec}. La simulación de recomendaciones se realiza tomando cada usuario $u_{rec} \in U_{rec}$ y por cada usuario $u_{com\_l\_base} \in U_{com\_l\_base} : u_{com\_l\_base} \neq u_{rec}$, se evalúa si existe un link $(u_{com\_l\_base}, u_{rec})$. En caso negativo, se agrega con una probabilidad de aceptación $P(a)$. Se evalúa luego el Coeficiente de Clustering, diámetro y promedio de Closeness Centrality para los nodos recomendados antes y después de la simulación para evaluar

las modificaciones estructurales que sufre la red, lo que puede beneficiar/perjudicar el flujo de información.

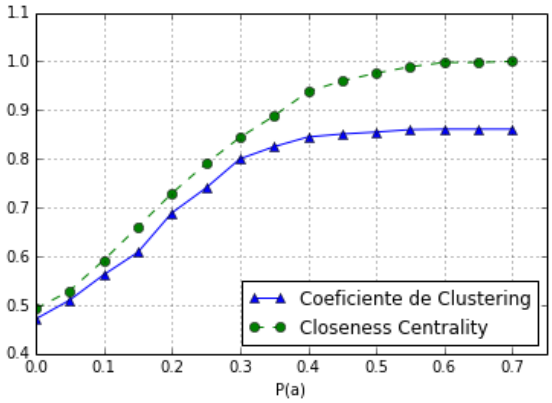

Fig. 4. CC y Closeness de U_{rec} vs Probabilidad de aceptación de la recomendación.

En la Figura 4 se puede apreciar que con $P(a) = 0,35$ el Coeficiente de Clustering alcanza un valor de $0,825$ y la serie cambia la pendiente, es decir, con una probabilidad relativamente baja la red se vuelve rápidamente más densa y mejor conexa. De igual forma, la métrica Closeness Centrality promedio para los usuarios recomendados aumenta de $0,492$ a $0,888$ ($+80\%$). Esta medida permite evaluar la rapidez con la que aumenta la capacidad de estos usuarios referentes en celiaquía para divulgar información. Finalmente, el diámetro de la red en este punto decrece de 5 a 3 lo cual reduce la cantidad de saltos necesarios entre los dos nodos más alejados (se reduce casi a la mitad, Figura 5).

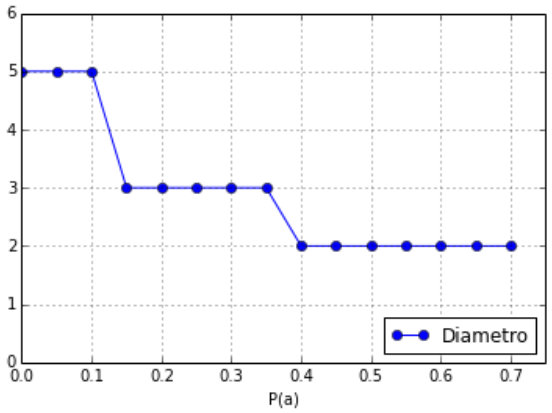

Fig. 5. Diámetro de G vs Probabilidad de aceptación de la recomendación (*link*).

6 Conclusiones y Trabajos Futuros

La formación de comunidades en redes sociales digitales es un fenómeno de interés desde múltiples puntos de vista. Como estructura subyacente, las comunidades presentan características particulares como su densidad, mientras que teniendo en cuenta los usuarios y sus interacciones aparecen diferentes comportamientos de acuerdo a la naturaleza de la comunidad (compartir ideas, gustos, hobbies, etc.).

En este trabajo se aborda el problema de la detección y fortalecimiento de una comunidad de usuarios de Twitter interesados en la enfermedad celíaca, particularmente en Argentina, complementando estudios médicos y biológicos de campo.

Aplicando combinaciones de técnicas, se detecta una comunidad limitada en cantidad de usuarios sobre la cual se identifican usuarios altamente influyentes. Si bien con la estructura del grafo se alcanza un 65% de precisión, ésta métrica mejora al ponderar las aristas por un criterio de similitud entre los usuarios (hasta un 77%). En cuanto a la utilización de KMeans combinado con el criterio de las celebridades, se muestra que se puede obtener una precisión cercana al 80% con $K = 35$, no aumentando significativamente esta métrica con una valor mayor de K.

Finalmente, la estrategia de recomendación de usuarios basada en influyentes y activos muestra que, seleccionando solamente un grupo pequeño de usuarios y con una probabilidad relativamente baja de aceptación de las recomendaciones, la red se vuelve rápidamente más densa y mejor conectada, lo que permite una mejor difusión de información valiosa respecto de la enfermedad celíaca entre los interesados.

Como trabajos futuros, se pretende ampliar el estudio considerando la evolución de la comunidad en el tiempo, y proponiendo una estrategia para la inclusión en la misma de usuarios que participan de varias comunidades, lo que dificulta su identificación, o aquellos parcialmente interesados en el tema. Complementariamente, se propone comparar las comunidades con otras áreas geográficas para las cuales existan estudios de campo actuales.

References

1. Backstrom, L., Leskovec, J.: Supervised random walks: Predicting and recommending links in social networks. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. pp. 635–644. WSDM '11, ACM, New York, NY, USA (2011)
2. Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1999)
3. Bedi, P., Sharma, C.: Community detection in social networks. Wiley Int. Rev. Data Min. and Knowl. Disc. 6(3), 115–135 (May 2016)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
5. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (2008)
6. Bothorel, C., Cruz, J.D., Magnani, M., Micenkova, B.: Clustering attributed graphs: models, measures and methods. Network Science 3(3), 408–444 (2015)
7. Cerny, N., Tamborenea, M.I., Cánepa, A., Cimarelli, M., Tolosa, G., Zunino, S., Ghiglieri, R., Gretel, H., Emilio, M., Rubén, I., Mauricio, D.M.: Epidemiological study of celiac disease in chivilcoy, buenos aires. IV International Congres in Translational Medicine. School of Pharmacy and Biochemistry of Universidad de Buenos Aires (2018)
8. Darmaillac, Y., Loustau, S.: MCMC louvain for online community detection. CoRR abs/1612.01489 (2016)
9. Derényi, I., Palla, G., Vicsek, T.: Clique percolation in random networks. Physical review letters 94(16), 160202 (2005)

10. Edler, D., Guedes, T., Zizka, A., Rosvall, M., Antonelli, A.: Infomap bioregions: Interactive mapping of biogeographical regions from species distributions. Systematic Biology 66(2), 197–204 (2017)
11. Fortunato, S., Castellano, C.: Community structure in graphs. In: Computational Complexity, pp. 490–512. Springer (2012)
12. Gach, O., Hao, J.K.: Improving the louvain algorithm for community detection with modularity maximization. In: Legrand, P., Corsini, M.M., Hao, J.K., Monmarché, N., Lutton, E., Schoenauer, M. (eds.) Artificial Evolution. pp. 145–156. Springer International Publishing, Cham (2014)
13. Gianan, O.: Finding influencers on twitter (2016), https://nycdatascience.com/blog/student-works/finding-influencers-twitter/
14. Gurusamy, V.: Mining the attitude of social network users using k-means clustering. International Journal of Advance Research in Computer Science and Software Engineering 7, 226–230 (05 2017)
15. Huffman, D.A.: A method for the construction of minimum-redundancy codes. Proceedings of the IRE 40(9), 1098–1101 (Sept 1952)
16. Jing, L., Ng, M.K., Huang, J.Z.: An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Transactions on knowledge and data engineering 19(8) (2007)
17. Kewalramani, M.N.: COMMUNITY DETECTION IN TWITTER.pdf. Ph.D. thesis, University of Maryland Baltimore County (2011)
18. Kiciman, E., De Choudhury, M., Counts, S., Gamon, M., Thiesson, B.: Analyzing social media relationships in context with discussion graphs. Eleventh Workshop on Mining and Learning with Graphs (2013)
19. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management. pp. 556–559. CIKM '03, ACM, New York, NY, USA (2003)
20. Lim, K.H., Datta, A.: Following the follower: Detecting communities with common interests on twitter. In: Proceedings of the 23rd ACM Conference on Hypertext and Social Media. pp. 317–318. HT '12, ACM, New York, NY, USA (2012)
21. Lim, K.H., Datta, A.: A topological approach for detecting twitter communities with common interests. In: Atzmueller, M., Chin, A., Helic, D., Hotho, A. (eds.) Ubiquitous Social Media Analysis. pp. 23–43. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
22. Lim, K.H., Datta, A.: An interaction-based approach to detecting highly interactive twitter communities using tweeting links. Web Intelligence 14(1), 1–15 (2016)
23. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
24. Mathur, G., Purohit, D.H.: Performance analysis of color image segmentation using k-means clustering algorithm in different color spaces. IOSR Journal of VLSI and Signal Processing 4, 01–04 (12 2014)
25. Meo, P.D., Ferrara, E., Fiumara, G., Provetti, A.: Generalized louvain method for community detection in large networks. CoRR abs/1108.1502 (2011)
26. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113 (Feb 2004)
27. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference. pp. 161–172 (1998), citeseer.nj.nec.com/page98pagerank.html
28. Papadopoulos, S., Kompatsiaris, Y., Vakali, A., Spyridonos, P.: Community detection in social media. Data Mining and Knowledge Discovery 24(3), 515–554 (2012)
29. Plantié, M., Crampes, M.: Survey on social community detection. In: Social media retrieval, pp. 65–85. Springer (2013)
30. Que, X., Checconi, F., Petrini, F., Gunnels, J.A.: Scalable community detection with the louvain algorithm. In: 2015 IEEE International Parallel and Distributed Processing Symposium. pp. 28–37 (May 2015)
31. Ren, Y., Kraut, R., Kiesler, S.: Applying common identity and bond theory to design of online communities. Organization studies 28(3), 377–408 (2017)

32. Rosvall, M., Axelsson, D., Bergstrom, C.T.: The map equation. The European Physical Journal Special Topics 178(1), 13–23 (Nov 2009)
33. Ruan, Y., Fuhry, D., Parthasarathy, S.: Efficient community detection in large networks using content and links. CoRR abs/1212.0146 (2012)
34. Tang, L., Liu, H.: Community Detection and Mining in Social Media. Morgan and Claypool Publishers, 1st edn. (2010)
35. Vathi, E., Siolas, G., Stafylopatis, A.: Mining and categorizing interesting topics in twitter communities. Journal of Intelligent and Fuzzy Systems 32(2), 1265–1275 (2017)
36. Vis, E., Scheepers, P.: Social implications of celiac disease or non-celiac gluten sensitivity. International Journal of Celiac Disease 5(4), 133–139 (2017)
37. Wang, T., Brede, M., Ianni, A., Mentzakis, E.: Detecting and characterizing eating-disorder communities on social media. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. pp. 91–100. WSDM '17, ACM, New York, NY, USA (2017)
38. Yang, J., McAuley, J., Leskovec, J.: Community detection in networks with node attributes. In: Data Mining (ICDM), 2013 IEEE 13th international conference on. pp. 1151–1156. IEEE (2013)
39. Yu, X., Yang, J., Xie, Z.Q.: A semantic overlapping community detection algorithm based on field sampling. Expert Systems with Applications 42(1), 366–375 (2015)
40. Zhang, Y., Wu, Y., Yang, Q.: Community discovery in twitter based on user interests. Journal of Computational Information Systems (2012)
41. Zhao, Z., Feng, S., Wang, Q., Huang, J.Z., Williams, G.J., Fan, J.: Topic oriented community detection through social objects and link analysis in social networks. Knowledge-Based Systems 26, 164–173 (2012)