Wikipedia: estudio de comunidades a través del tiempo

Leizerovich, M., Goren, G.

13 de diciembre de 2018

Resumen

Se presenta un estudio de la estructura de categorías de la Wikipedia en inglés y su evolución en el período 2015-2018. Para esto, se seleccionaron dos categorías, cada una junto con sus categorías "descendientes", para las cuales se esperaba observar un crecimiento apreciable durante dicho período y una intercomunicación fuerte (correspondientes a las temáticas Estadística y *Machine learning*). El análisis se basó en una comparación entre particiones del conjunto de artículos obtenidas a partir de las categorías y a partir del algoritmo *Infomap*. Este último, a su vez, se aplicó sobre dos tipos de grafo diferentes: la red de hipervínculos de Wikipedia y grafos semánticos obtenidos mediante la aplicación de LSA o Análisis Semántico Latente. Los resultados indican una evolución positiva en el acuerdo entre la estructura de categorías de la Wikipedia y las comunidades detectadas de forma no supervisada. Se concluyó que la estructura de categorías de Wikipedia es adecuada en el sentido de que, al incorporarse información a los artículos, esta es consistente con las categorías y permite mejorar su detección automática. Por otro lado, las comunidades del grafo semántico predicen las categorías significativamente mejor que las comunidades del grafo de hipervínculos, lo cual avala la efectividad del método de análisis semántico. Por último, se observó que la similaridad entre los grafos semántico y de hipervínculos no crece monótonamente, concluyéndose que la detección de comunidades a través de ambos caminos es cualitativamente diferente.

1. Introducción

En el presente trabajo, se describe un estudio de realizado sobre una región relativamente pequeña de la enciclopedia online Wikipedia en inglés, en el cual se buscó caracterizar la evolución en el tiempo de la estructura de categorías. Para esto, se realizó un análisis de *clusterización* o detección de comunidades complementado con un análisis semántico de texto.

El cuerpo de datos empleado consiste en información sobre 3667 entradas de la Wikipedia en inglés correspondiente a cuatro instantes de tiempo en el período 2015-2018. La información fue mediante un proceso de búsqueda en anchura sobre el grafo de categorías de Wikipedia que se describe en la Sección 2. El proceso de búsqueda comenzó a partir de dos categorías "semilla" o "raíz", las cuales se eligieron como *Machine learning* (Aprendizaje de máquina) y *Statistics* (Estadística). Definimos como *macrocategoría* al conjunto de todos los artículos a los que se accedió partiendo de una de estas categorías semilla, e indicamos a las macrocategorías correspondientes a *Machine learning* y *Statistics* por ML y ST respectivamente.

La elección de ST y ML como macrocategorías respondió a un doble criterio: por un lado, se esperaba observar un crecimiento apreciable de las mismas durante el período estudiado (fundamentalmente en ML). Por otro lado, se esperaba observar una intercomunicación fuerte entre las mismas, cuya evolución temporal sería susceptible de ser analizada.

Tal como ya fue mencionado, una de las herramientas principales para el análisis de los datos obtenidos fue la detección no supervisada de comunidades. Para ello, se construyeron dos grafos diferentes para cada

instantánea, reflejando distintos aspectos de los datos. En ambos grafos los nodos son las entradas de la enciclopedia. A partir de ellos se construye en primer lugar el grafo de hipervínculos, en el cual asigna una arista dirigida de i a j si en el i-ésimo artículo existe un hipervínculo al artículo j; esta estructura es la que usualmente se interpreta como "la red de Wikipedia". En segundo lugar, se construye un grafo a partir del análisis semántico del texto de las entradas, el cual denominamos *grafo semántico*. En este grafo se asigna una arista *no* dirigida a un par de nodos si su distancia semántica es menor a un determinado valor de umbral (ver Sección 3). En ambos casos, el algoritmo de detección de comunidades empleado fue *Infomap* (ver Secciones 4.1.1 y 4.2.1).

El objetivo de este trabajo fue, en primer lugar, determinar hasta qué punto es posible reconstruir las estructura de categorías de Wikipedia a partir de la detección no supervisada de comunidades, aplicada tanto sobre el grafo de hipervínculos como sobre el grafo semántico. En un segundo lugar, observar estas comunidades en función del tiempo e intentar extraer conclusiones sobre la evolución de los tópicos seleccionados en la enciclopedia. En particular, es de interés determinar si la categorización de Wikipedia se vuelve más consistente con las comunidades detectadas con el paso del tiempo.

2. Adquisición de datos

Debido a la especificidad de la información buscada (en especial, el requerimiento de obtener toda la información para distintos instantes de tiempo), no fue posible encontrar una interfase de alto nivel adecuada para la adquisición de la misma. Es por esto que se desarrolló un código dedicado en el lenguaje Python para interactuar directamente con la API de Wikipedia. Esto implicó familiarizarse con las distintas opciones y parámetros de llamada que ofrece la API y diseñar una manera de obtener los datos minimizando el número de llamadas[1].

Para especificar con precisión el subconjunto de artículos recolectado, resulta económica una descripción del algoritmo empleado. Este consiste fundamentalmente en una búsqueda en anchura, no sobre la red de hipervínculos de la Wikipedia sino sobre la estructura de categorías de la misma. En Wikipedia, cada artículo está etiquetado por una serie de categorías, a cada una de las cuales pertenece. Por otro lado, las categorías se relacionan entre sí mediante una relación asimétrica de inclusión, según la cual una categoría puede ser subcategoría de otra. Esto define un grafo dirigido en el cual los vértices son categorías y las aristas inciden desde una categoría A a una categoría B si B es subcategoría de A. Definimos que A es *ancestro* de B si es posible llegar de A hasta B siguiendo los enlaces dirigidos. Equivalentemente, diremos en ese caso que B es *descendiente* de A. La búsqueda en anchura consiste en partir de una categoría "semilla", *visitarla* (es decir, adquirir la información de los artículos que pertenecen a la misma) y almacenar en una cola de espera una referencia a todas las subcategorías de la misma. Una vez hecho esto, se procede a visitar cada categoría en la lista de espera, agregando a la cola de espera todas las subcategorías de cada una en la medida en que son visitadas (se ignoran las subcategorías que ya fueron visitadas o ya están en la cola de espera). El proceso termina cuando no hay más categorías en la cola de espera o bien se adquiere una determinada cantidad de datos.

Este proceso se realizó partiendo de las dos categorías semilla, *Machine learning* y *Statistics*, obteniéndose las macrocategorías ML y ST. La búsqueda partiendo de *Machine learning* se ejecutó por completo, obteniéndose 1180 artículos. Es decir que ML contiene la totalidad de las entradas categorizadas bajo *Machine learning* o categorías descendientes. Por otro lado, la búsqueda partiendo de *Statistics* fue interrumpida antes de su finalización, adquiriéndose 2859 artículos. En este punto, es importante notar que las macrocategorías

[1]Los detalles de la implementación, así como todo el código empleado para este trabajo, pueden ser consultados en `https://github.com/GRGab/wikipedia-proyecto`.

no son disjuntas, sino que 372 artículos pertenecen a ambas simultáneamente, configurando un total de 3667 artículos adquiridos.

Para cada entrada de las categorías visitadas, se almacenó el código HTML de la misma, la lista de hipervínculos presentes y la lista de categorías a las que el artículo pertenece. A su vez, esta información se adquirió para cuatro instantes de tiempo correspondientes a octubre de 2015, 2016, 2017 y 2018, configurando cuatro instantáneas de la región de Wikipedia bajo estudio.

2.1. Estructura de categorías

En la Figura 1, se muestra una visualización del grafo de categorías recorrido, en la cual se señalan con flechas las categorías semilla. En total fueron visitadas 361 categorías. El grafo es conexo, pero se observa inmediatamente que no es un árbol, en el sentido de que contiene múltiples ciclos no dirigidos (ciclos obtenidos ignorando la dirección de las flechas).

A partir de la estructura de categorías, se buscó definir una partición categórica de los artículos, i.e. una regla que a cada artículo le asigne exactamente una categoría. Esto permitió comparar directamente las categorías con particiones detectadas de forma automática.

Figura 1: Grafo de categorías de Wikipedia recorrido durante la adquisición de datos. El diámetro de los nodos refleja su grado de salida, y se señalan con flechas las categorías semilla. En rojo, se resaltan las categorías semilla y sus subcategorías (descendientes directos).

Para ello, fue necesario tener en cuenta que un artículo puede pertenecer a una categoría C sin pertenecer a los ancestros de C. Por ejemplo, solo 11 artículos pertenecen a la categoría *Statisticians* (estadísticos y estadísticas), pero 526 pertenecen a dicha categoría o categorías descendientes (tales como, por ejemplo, categorías correspondientes a investigadores de nacionalidades específicas). En contraposición, la noción misma de subcategoría contiene implícitamente la hipótesis semántica de que si un artículo contiene información sobre el tópico correspondiente a una categoría dada, entonces también contiene información respecto de sus categorías ancestro.

Considerando esta observación, se adoptó la siguiente estrategia para definir una partición categorial. En primer lugar, se eligió un *nivel* dentro del grafo, definido como el subconjunto de categorías que se encuentran a una distancia i de por lo menos una de las dos semillas. A continuación, se construyó un mapeo entre categorías, que envía a cada categoría de nivel $j \geq i$ a su ancestro en el nivel i (el cual puede ser ella misma).

En este paso es esencial recordar que el grafo de categorías no es un árbol, por lo que una categoría dada puede tener múltiples ancestros en el nivel i (incluso si la categoría misma pertenece a dicho nivel). Para resolver esta ambigüedad, se desarrolló un *helper* o script asistente que permite a un usuario humano decidir caso por caso a qué categoría de nivel i debe asignarse la categoría en cuestión. El criterio adoptado fue ciertamente subjetivo y puede considerarse arbitrario, si bien respondió a un intento de reducir la cantidad total de categorías (en los casos en que la decisión implicaba eliminar una categoría de nivel i) y de asignar las categorías siempre a la opción más general disponible.

Una vez obtenido el mapeo de categorías, cada artículo fue asignado a una categoría de la siguiente manera: si el artículo pertenecía a una categoría que es mapeada a la categoría C, entonces el artículo fue asignado a C. Si el artículo pertenecía a distintas categorías, las cuales son mapeadas a categorías $\{C_1, C_2, ..., C_n\}$, $n > 1$, entonces se asignó el artículo a alguna de ellas de manera arbitraria. En el caso en que el artículo no pertenecía a ninguna categoría que sea mapeada a categorías del nivel i, se asignó el artículo a una de tres categorías definidas *ad hoc*: "General_ML" si el artículo pertenecía a ML y no a ST, "General_ST" ysi pertenecía a ST y no a ML, y "General_shared" si pertenecía tanto a ML como a ST.

Para realizar el procedimiento, se eligió el nivel $i = 1$, obteniéndose 36 categorías en el mismo luego del proceso de desambiguación, las cuales se ven marcadas en rojo junto con las categorías semilla en la Figura 1. En la figura 2 se observan las categorías porcentualmente más importantes.

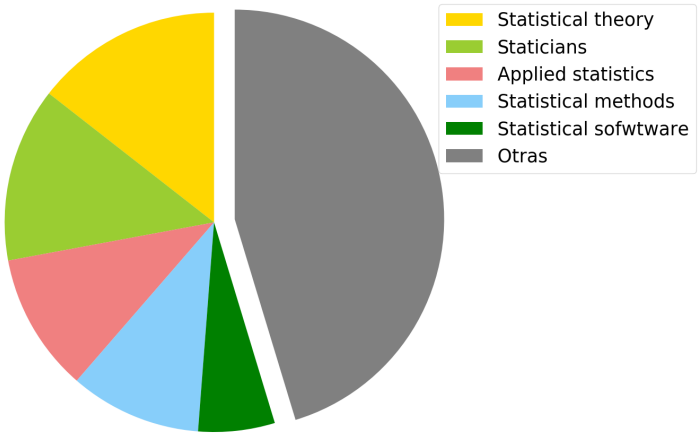


Figura 2: Las cinco categorías más importantes asignadas a los artículos de Wikipedia. Decimos que una categoría es más importante si es asignada a un número mayor de artículos.

3. Análisis semántico

El objetivo del análisis semántico fue crear grafos sobre los cuales aplicar un algoritmo de detección de comunidades. Tal como se mencionó antes, estos grafos tienen los mismos nodos que las redes de hipervínculos correspondientes, pero difieren en sus enlaces. La idea general es asignar el texto de cada artículo a un punto en un espacio vectorial, definir una distancia entre estos puntos y asignar un enlace entre dos artículos si la distancia medida entre sus textos es inferior a un determinado umbral.

El corpus de textos sobre el cual se trabajó fue el texto plano extraído de los HTML de cada artículo (y para cada instantánea de tiempo). Esto incluye no solamente el texto del cuerpo del artículo, sino también barras laterales, referencias, etc.

En primer lugar, a partir del corpus se confecciona un vocabulario o lista de términos presentes. En este

vocabulario, se excluyen palabras comunes (en este caso, del idioma inglés) que no aportan información sobre el tópico de los textos, tales como artículos, preposiciones, conectores, etc., denominadas colectivamente *stop-words*. Por otro lado, se incluyen no solamente palabras individuales sino también diagramas (combinaciones de dos palabras sucesivas). Una vez hecho esto, se asocia cada texto a un vector de longitud n igual a la longitud del vocabulario, en el cual cada entrada es el número de veces que un cierto término del vocabulario aparece en ese texto. Agrupando estos vectores en una matriz $m \times n$, donde m es el número de textos en el corpus, se obtiene una *matriz de conteos*.

A continuación, se obtiene una nueva matriz mediante la denominada *transformación TF-IDF*, la cual recibe su nombre de la expresión inglesa *term frequency - inverse document frequency* (frecuencia de términos - frecuencia inversa de documentos). La transformación realiza una ponderación de los distintos elementos de la matriz de conteos, asignando una mayor importancia dentro de cierto artículo a los términos que aparecen con una frecuencia alta en el mismo pero que aparecen en relativamente pocos artículos (es decir, con una baja frecuencia de documentos). De esta manera, se disminuye el impacto de términos que aparecen frecuentemente en el corpus y por lo tanto son menos informativos a la hora de clasificar los artículos. En particular, en la implementación utilizada, la matriz de TF-IDF Y se define a partir de la matriz de conteos X según

$$Y_{ij} = X_{ij} \times \log\left(\frac{1+m}{1+DF_{ij}}\right) + 1$$
$$DF_{ij} = \sum_i (1 - \delta(X_{ij}, 0))$$

donde $\delta(X_{ij}, 0) = 1$ si $X_{ij} = 0$ y 0 si no. Finalmente, cada vector-artículo $\bar{Y}^i = (Y_{i1}, ..., Y_{in})$ es normalizado según la norma euclídea.

Una vez realizada la transformación TF-IDF, se aplica la técnica de reducción de dimensionalidad denominada Análisis Semántico Latente (LSA, por sus siglas en inglés). Esta técnica, introducida por Deerwester et al. en 1990 [2] en el contexto de sistemas de búsqueda de información, puede resumirse en nuestro contexto como un intento de extraer una cierta cantidad predefinida de tópicos a partir de combinaciones lineales de términos presentes en el corpus. La teoría detrás de este método es rica y tiene implicaciones en la psicología del lenguaje [3].

Esencialmente, LSA consiste en aproximar la matriz original por una matriz de menor rango. Esto se logra realizando sobre Y una descomposición en valores singulares (análoga a una diagonalización en el caso de matrices cuadradas) y truncando los valores singulares (análogos a autovalores), i.e. preservando los k valores singulares más grandes y reemplazando los restantes por ceros. De esta manera, todos los vectores-artículo resultantes quedan contenidos dentro de un subespacio de dimensión k. Cada una de estas k dimensiones corresponde a una combinación lineal de términos del vocabulario, los cuales pueden interpretarse como tópicos en un contexto de clasificación de documentos. Este proceso de *granularización* o *coarse-graining* desdibuja los límites entre los términos y permite, por ejemplo, identificar similaridad entre textos que usan palabras sinónimas para hablar de las mismas cosas.

Como paso final, se empleó la métrica denominada "distancia coseno" para determinar la cercanía semántica entre vectores-artículo, la cual se define para vectores u y v como $d(u, v) = 1 - \cos(\theta)$ donde θ es el ángulo entre u y v. Abusando de la notación, y simbolizando a los nodos o artículos de igual modo que sus representaciones vectoriales, se construye el grafo semántico asignando una arista entre los nodos u y v sii $d(u, v) \leq \alpha$ para cierto valor de umbral α.

En el proceso descrito se realizó una vez para cada instantánea. Para ello, dos parámetros debieron ser ajustados previamente: k y α. Dado que α determina la densidad del grafo semántico resultante, se eligió

de forma tal que dicha densidad coincida con la densidad de la red de hipervínculos. Por otro lado, para la determinación de k se realizó un barrido sobre $k \leq 40$, aplicando en cada caso el algoritmo de detección de comunidades *Infomap* sobre el grafo resultante y comparando las particiones obtenidas con las correspondientes a aplicar *Infomap* sobre la red de hipervínculos. De esta manera, se buscó el valor de k que maximizara la información mutua normalizada entre ambas particiones[2], obteniéndose $k = 26$.

4. Resultados

4.1. Red de hipervínculos

En esta sección, se presentan las redes obtenidas y sus caracterizaciones preliminares. En la Figura 3 se observan las cuatro redes de hipervínculos, en sucesión temporal. En cada instantánea, se marcan con rojo los nodos y enlaces nuevos respecto de la instantánea anterior. Algunos datos sumarios sobre estas redes pueden verse en la Tabla 1.

Figura 3: Redes de hipervínculos para distintos instantes de tiempo. Los nodos y enlaces rojos son los que se agregan en relación a la instantánea anterior.

Durante el período de tiempo estudiado, la cantidad de nodos presentes en la red aumentó monótonamente, como era de esperar, pasando de 3115 nodos en 2015 a 3667 en 2018. La mayor parte de este crecimiento ocurrió entre 2017 y 2018, con casi 300 artículos nuevos versus poco más de 100 para los otros dos intervalos anuales. La densidad de enlaces es de apenas el 0.5 % y se mantiene así durante todo el período. Los grados máximos de entrada y salida crecen monótonamente, indicando que los hubs crecen junto con el resto de la red, mientras que el grado medio crece hasta 2017 y baja en 2018. Algo similar ocurre con las medidas de transitividad (proporción de triángulos) y coeficiente medio de clustering. Se puede inferir entonces que gran parte del crecimiento ocurrido entre 2017 y 2018 se debe a la adición de artículos de grado muy bajo y con poca transitividad.

Por otro lado, con respecto a la distinción entre macrocategorías ML y ST, se observa que la cantidad de artículos en ML aumenta en aproximadamente 60 entradas anuales. El crecimiento más grande de la red se observa en ST, donde los incrementos son, en orden cronológico, de 99, 72 y 251 artículos. El solapamiento entre ambas macrocategorías se mantiene más o menos constante en un 10 % del tamaño total de la red, agregándose tan solo entre 10 y 20 artículos en común a ambas categorías por año. En ese sentido, dado que ya

[2] La información mutua normalizada entre dos particiones se define como

$$IM = \frac{\sum_{C_1,C_2} p(C_1,C_2) \log\left(\frac{p(C_1,C_2)}{p(C_1)\cdot p(C_2)}\right)}{\frac{1}{2}(H(p(C_1)) + H(p(C_2))}$$

donde $p(C_1)$ es la probabilidad de que un nodo elegido de manera uniformemente aleatoria pertenezca al cluster C_1 de la primera partición, mientras que $p(C_2)$ es el equivalente correspondiente a la segunda. La notación $p(C_1, C_2)$ corresponde a la distribución de probabilidad conjunta para ambas particiones, mientras que $H(q)$ es la entropía de la distribución de probabilidad q, dada por $H(q) = \sum_x q(x) \log(q(x))$.

en 2015 las macrocategorías compartían más de 300 artículos, puede decirse que durante este período no hubo movimientos de fusión o diferenciación de las macrocategorías ni una expansión de una sobre los artículos de la otra.

Para complementar esta descripción del solapamiento entre macrocategorías, se estudió también la interconexión entre los artículos que las macrocategorías *no* tienen en común, pero pueden referenciarse mutuamente. Para ello, se calculó para cada instante de tiempo la modularidad[3] del subgrafo compuesto por todos los artículos que pertenecen a una única macrocategoría, particionado justamente por macrocategoría. Se observó que la modularidad crece monótonamente de 0.158 en 2015 hasta 0.195 en 2018, lo cual indica que la intercomunicación entre las macrocategorías *disminuye* en el tiempo en comparación con la comunicación interna dentro de cada una de ellas.

Año	N	$\langle k \rangle$	k^{in}_{max}	k^{out}_{max}	$\langle C_i \rangle$	C_Δ
2015	3117	16.022	954	1269	0.366	0.458
2016	3257	16.633	983	1289	0.377	0.555
2017	3372	17.117	1010	1300	0.378	0.554
2018	3668	16.862	1045	1304	0.376	0.490

Tabla 1: Datos sumarios sobre la red de hipervinculos, para cada año del período analizado. N es el número de nodos de la red, $\langle C_i \rangle$ es su coeficiente medio de clustering, C_Δ es su transitividad, $\langle k \rangle$ es el grado medio, y k^{in}_{max} y k^{out}_{max} son los grados máximos de entrada y salida respectivamente.

4.1.1. Particionamiento de la red de hipervínculos

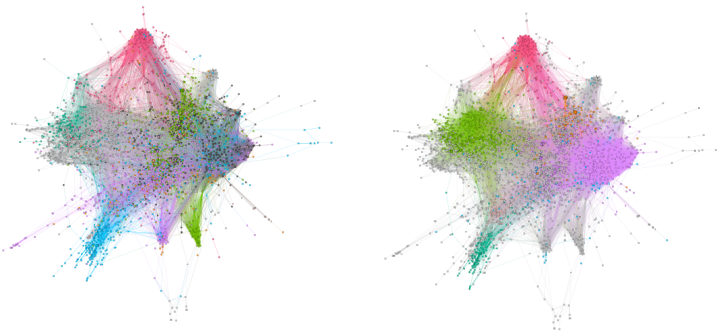
Como se mencionó previamente, el algoritmo de detección de comunidades empleado fue *Infomap*. Este método, así llamado por sus autores, fue presentado en 2009 por Rosvall y Bergstrom[1]. La idea general consiste en generar caminatas al azar a lo largo del grafo, las cuales son consideradas como secuencias finitas de nodos, y buscar la partición que minimice la longitud de descripción de las mismas, usando una codificación Huffman de dos niveles (un código indica las transiciones entre comunidades a lo largo de la caminata, mientras que un segundo código indexa a cada nodo dentro de una misma comunidad). A través de la comparación entre la partición obtenida por *Infomap* y la partición categorial, buscamos determinar si la estructura de categorías resuena con la estructura topológica de la red de hipervínculos.

En la figura 4 se muestran dos coloraciones de la red de hipervínculos de 2018: una según la partición categorial y otra según la partición por *Infomap*. Las visualizaciones para años anteriores son similares. Por otro lado, en la Figura 5 se muestra la proporción de particiones más importantes según *Infomap* para 2018, notando con leyendas aquellas que corresponden más claramente con categorías de Wikipedia: *Applied Statistics, Statistical Software* y *Artificial Neural Networks*.

[3] La modularidad de una partición en comunidades $\{c_i\}_{i \in I}$ en un grafo G se define según

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - k_i \frac{k_j}{2m} \right) \delta(c_i, c_j)$$

donde m es el número de enlaces de la red, A es la matriz de adyacencia de la red, k_i es el grado del i-ésimo nodo y c_i es la comunidad a la que pertenece. La notación $\delta(c_i, c_j)$ debe interpretarse como igual a 1 si los nodos i y j pertenecen a la misma comunidad, y 0 en caso contrario.



(a) (b)

Figura 4: Grafos de hipervínculos particionados según las categorías a las que pertenecen (a) y el algoritmo *Infomap* (b).

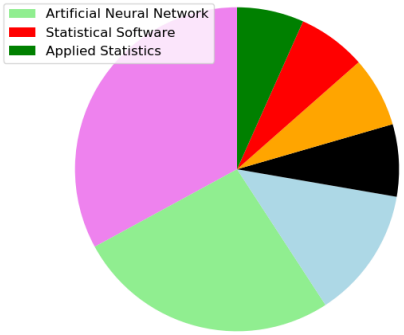


Figura 5: Se grafican las proporciones de las comunidades más importantes para la red de hipervínculos según *Infomap* para el año 2018, donde el color corresponde a los de las comunidades de la Figura 4b. Se señalan con leyendas las categorías que se identifican claramente con dichas particiones.

4.2. Grafo semántico

Los grafos semánticos obtenidos para cada instantánea se presentan en la Figura 6. Sobre ellos, se realizó un análisis análogo al de la red de hipervínculos. En la tabla 2 se detallan los datos sumarios sobre el grafo en cada período.

Como se puede observar, el grado medio de la red aumentó alrededor de un 17 % entre el 2015 y el 2018. A su vez, el grado máximo casi se duplicó en el mismo período. La transitividad de la red es monótonamente creciente, con un aumento del 13 % en el tiempo, pero el coeficiente de clustering alcanza un máximo en 2016, para luego descender en los años siguientes.

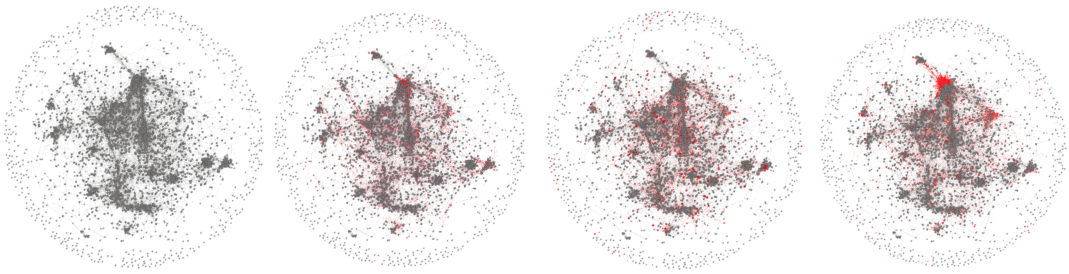


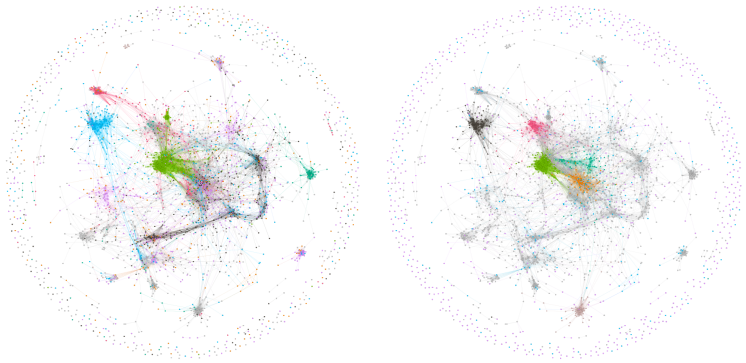
Figura 6: Grafos semánticos para distintos instantes de tiempo. Los nodos y enlaces rojos son los que se agregan en relación al año anterior.

	Año	N	$\langle k \rangle$	k_{max}	$\langle C_i \rangle$	C_Δ
0	2015	3117	18.145	115	0.464	0.640
1	2016	3257	18.883	135	0.467	0.655
2	2017	3372	19.347	153	0.455	0.663
3	2018	3668	21.558	214	0.433	0.727

Tabla 2: Datos sumarios sobre los grafos semánticos, para cada año del período analizado. N es el número de nodos de la red, $\langle C_i \rangle$ es su coeficiente medio de clustering, C_Δ es su transitividad, $\langle k \rangle$ es el grado medio, y k_{max} es el grados máximo alcanzado.

4.2.1. Particionamiento del grafo semántico

Se aplicó el algoritmo *Infomap* sobre los 4 grafos semánticos para luego compararlos con la partición categorial (ver Sección 5). En la Figura 7 se observan los grafos coloreados por las particiones de *Infomap* y por categorías. Por otro lado, en la Figura 8 se observan la proporción entre las comunidades más grandes detectadas. Las categorías que más resuenan en la topología de la red en este caso son *Statisticians* y *Applied statistics*.



(a) (b)

Figura 7: Grafos LSA particionados según las categorías a las que pertenecen y el algoritmo *Infomap*, respectivamente.

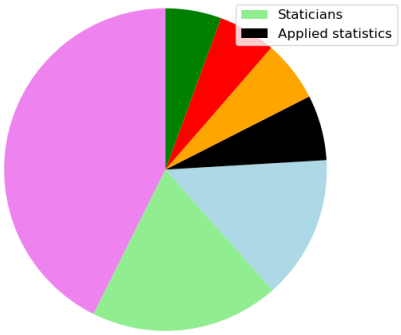


Figura 8: Se muestra la proporción entre las comunidades más grandes para el grafo semántico según *Infomap* para el año 2018, donde el color corresponde a los de las redes 7b. Se señalan con leyendas las categorías que se identifican claramente con dichas particiones.

5. Análisis de acuerdo en función del tiempo

En esta sección, buscamos determinar si la categorización de Wikipedia se vuelve más consistente con las comunidades detectadas con el paso del tiempo. Para cuantificar este acuerdo entre particiones categorial e *Infomap*, se calculó la información mutua normalizada entre ambas, para cada tipo de grafo y cada instante de tiempo. Los resultados se muestran en la tabla 3.

Se observa que el acuerdo entre las particiones crece monótonamente en el tiempo tanto para la red de hipervínculos como el grafo semántico. Esta observación está en acuerdo con nuestra hipótesis de que las categorías de Wikipedia reflejan mejor la estructura de la red a medida que ésta evoluciona en el tiempo. Esto podría indicar que la estructura de categorías fue avanzando hacia una mejor caracterización de los tópicos abordados por las entradas de la enciclopedia; sin embargo también es posible que sea el texto e hipervínculos de los artículos lo que mejora en el tiempo, haciendo que sea cada vez más fácil predecir su pertenencia categorial. Si bien es necesario un mayor análisis para responder definitivamente esta pregunta, el hecho de no haber observado grandes cambios en la estructura de categorías durante el período estudiado sugiere que se trata de lo segundo. De ser ese el caso, el hecho de que el acuerdo crezca en el tiempo implicaría que la estructura de categorías es "adecuada", en el sentido de que, al agregarse información a los artículos individuales (texto e hipervínculos), estos reflejan cada vez mejor las categorías a las que fueron asignados.

Otra observación interesante es el hecho de que, para cada instante analizado, el grafo semántico contó con un valor de información mutua mayor a el de la red de hipervínculos. Esto puede ser explicado por el hecho de que páginas que "hablan de lo mismo", probablemente conectadas en el grafo semántico, suelen compartir las categorías a las que pertenecen. Sin embargo, páginas que están relacionadas a partir de un hipervínculo entre sí no necesariamente refieren al mismo tópico. En otras palabras, el grafo semántico contiene evidencia sobre un posible tópico común al nivel de los enlaces aislados, mientras que las redes de hipervínculos solo aportan información sobre posibles tópicos en el nivel mesoscópico de patrones de interconectividad heterogéneos. La verificación de esta hipótesis aporta confianza en el método de análisis semántico empleado.

Año	Hip	LSA
2015	0.086	0.441
2016	0.084	0.455
2017	0.080	0.466
2018	0.064	0.483

Tabla 3: Se muestran los valores de información mutua normalizada calculados entre la partición categorial y la partición detectada por *Infomap*, para la red de hipervínculos y el grafo semántico en función del tiempo.

Para cuantificar el acuerdo entre la red de hipervínculos y el grafo semántico, se calculó la cantidad de enlaces en común que contienen. Para normalizar esta cantidad por una magnitud representativa del tamaño de la red, se la dividió por el numero total de enlaces del grafo semántico. Esta proporción se calculó para todos los períodos de tiempo analizados. Los resultado se muestran en la Tabla 4.

Como se puede observar, la proporción no crece monótonamente como se esperaba, sino que existe un máximo de similaridad entre ambas redes en Octubre de 2016. Más allá de los motivos por los que esto ocurre, puede inferirse de este resultado que la capacidad del análisis semántico de texto de identificar las categorías de la enciclopedia no depende de imitar la estructura de hipervínculos, sino que ambos aspectos de lo que constituye a una entrada de Wikipedia contribuyen información en cierta medida independiente.

Año	Proporción
2015	0.086
2016	0.084
2016	0.080
2017	0.064

Tabla 4: Proporción de enlaces en común entre la red de hipervínculos y el grafo semántico.

6. Conclusiones

En este trabajo, se extrajo información en función del tiempo correspondiente a una sección de Wikipedia mediante una búsqueda sobre la estructura de categorías, y con ella se construyeron redes de hipervínculos y grafos semánticos para el estudio de la estructura de categorías de dicha sección de la enciclopedia. Esta sección estuvo compuesta por dos macrocategorías, ST y ML, que se mantuvieron en una relación relativamente estable durante el período 2015-2018, con un solapamiento constante del 10 %, y se observó a través de un análisis de modularidad que la interconexión entre las macrocategorías ST y ML disminuye de forma monótona en contraposición con lo que se esperaba.

Utilizando el algoritmo *Infomap*, se encontraron particiones de las redes estudiadas y se las compararon con la estructura de categorías propia de la Wikipedia. De esta forma, se encontraron categorías que efectivamente resuenan con la topología de la red al estar en acuerdo con las particiones de *Infomap*.

El análisis de información mutua entre las particiones encontradas por el algoritmo de *Infomap* y las categorías asociadas a los artículos relevados indica una evolución positiva en el acuerdo entre la estructura de categorías de la Wikipedia y las comunidades detectadas de forma no supervisada.

El hecho de no haber observado grandes cambios en la estructura de categorías durante el período estudiado

sugiere que son las entradas de la enciclopedia las que, al ser mejoradas en el tiempo, van permitiendo una mejor detección automática de sus categorías. De ser ese el caso, podría inferirse que la estructura de categorías es "adecuada", dado que los artículos se vuelven más consistentes con ella a medida que son completados y mejorados a un nivel individual.

Por otro lado, se observó que la información mutua entre categorías y comunidades de *Infomap* siempre es mayor para el grafo semántico que para la red de hipervínculos, lo cual es consistente con lo esperado dado que un enlace en el grafo semántico, considerado aisladamente, ya proporciona evidencia sobre la posible existencia de una categoría en común, mientras que en la red de hipervínculos la información sobre posibles tópicos se encuentra en el nivel de organización mesoscópico. La verificación de esta hipótesis brinda confianza en el método de análisis semántico empleado.

Por último, se observó que la similaridad entre los grafos semántico y de hipervínculos no crece monótonamente, lo cual sugiere que la capacidad del análisis semántico de texto de identificar las categorías de la enciclopedia no depende de imitar la estructura de hipervínculos, sino que ambos caminos de análisis contribuyen información independiente. Como perspectiva a futuro, esto sugiere combinar la información presente en ambos tipos de grafo para lograr una mejor predicción de la partición categorial. Esto podría lograrse, por ejemplo, mediante un grafo pesado en el cual el peso de una arista esté dado por la suma de una similaridad semántica derivada de LSA y un coeficiente de acoplamiento bibliográfico.

Referencias

[1] Rosvall, M. y Bergstrom, C.T. (2008) Maps of random walks on complex networks reveal community structure. PNAS 105(4): 1118-1123. doi:10.1073/pnas.0706851105

[2] Deerwester, S. et al. (1990) Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6): 391-407.

[3] Landauer, T.K. y Dumais, S.T. (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2): 211-240.